J. Ind. Soc. Agril. Statist.
59(2), 2005 : 97–103

Multivariate Indirect Methods of Estimation in Successive Sampling

Artes Rodriguez, Eva M. and Gracia Luengo, Amelia V.
Carretera de Sacramento s/n. Edifificio CITE III, University of Almeria, Spain
(Received : February, 2002)

SUMMARY

The problem of estimation of a finite population mean for the current occasion based on the samples selected over two occasions has been considered. For the case when several auxiliary variables are negatively correlated with the main variable, a double-sampling multivariate product estimate from the matched portion of the sample is presented. Expressions for optimum estimator and its error have been derived. The gain in efficiency of the combined estimate over the direct estimate using no information gathered on the first occasion is computed.

A comparison with the univariate product estimator has been made, giving the specific situations under which either of them may be efficiently used. An empirical study is also included for illustration.

Key words : Product estimator, Successive sampling, Gain in efficiency, Matching fraction.

1. INTRODUCTION

Successive sampling has been extensively used in applied sciences and the environment to provide more efficient estimates of population characteristics such as means. The problem of sampling on two successive occasions with a partial replacement of sampling units was first considered by Jessen (1942) in the analysis of a survey which collected farm data.

In a later paper (Sen (1971)), this sampling plan was applied with success in designing a mail survey in Ontario of waterfowl hunters who hunted successively during 1967-68 and 1968-69. An estimate was developed for the current season (1968-69), based on the relationship between the value of a characteristic during the current season and its value during the previous season, that yielded more precise estimates of the kill of waterfowl than the usual estimates based on simple random sampling and using the hunter's current season's performance only.

Successive sampling has also been discussed in some details by Patterson (1950), Yates (1981) and others. Their discussions have, however, been confined to combining ratio or regression estimates from the matched portion of the sample with a mean per unit estimate based on the current occasion.

The use of ratio method of estimation in successive sampling was first introduced by Avdhani (1963 Ph.D. thesis) and later Sen *et al.* (1975). Gupta (1970 Ph.D. thesis) and later Artes *et al.* (1998) and Artes and Gracia (2000, 2001) have suggested the use of product method.

It has been shown, in this context, that the combined estimator that uses a double sampling ratio estimator for the matched part of the sample is more accurate than the simple estimator $\bar{y}$ when the auxiliary variable is positively related to the principal variable y, and it is verified that $\rho > \frac{1}{2}\frac{C_x}{C_y}$ (Sen *et al.* (1975)).

Frequently, the study of environmental issues involves negatively correlated characteristics. So, the product method of estimation is relevant to these cases.

If the relation between the auxiliary variable x and the study variable y is negative, it has also been proved that the optimum estimator which combines a double sampling product estimator for the matched part of the

sample and a simple sample mean for the unmatched part, has less variance than the usual estimator $\bar{y}$ provided $\rho < -\frac{1}{2}\frac{C_x}{C_y}$ (Artes *et al.* (1998)).

Because in many agricultural surveys computations involving product estimates become relatively complex, we propose to investigate in this paper some theory of successive sampling using a multivariate product estimate and examine the efficiency over the direct estimate exclusively based on the sampling units for the current occasion.

2. DEVELOPMENT OF MULTIVARIATE PRODUCT METHOD OF ESTIMATING THE MEAN ON THE SECOND OCCASION

2.1 Selection of the Sample

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 large 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. 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.

Information from the first occasion on $k\,(k \geq 1)$ auxiliary variables $x_1, \ldots, x_k$ is available, whose means are denoted by $\bar{x}_1, \ldots, \bar{x}_k$, respectively. Let y be the variable under study on the second occasion, and we suppose that it is negatively correlated with $x_1, \ldots, x_k$.

2.2 Notation Used

Let m = sample size of those questioned on both occasions (matched sample), u = n – m (unmatched sample), $\bar{x}_1^m, \ldots, \bar{x}_k^m\,(\bar{y}_m)$ = matched sample mean on the first (second) occasion estimating $\bar{X}_1, \ldots, \bar{X}_k\,(\bar{Y})$, $\bar{y}_u$ = unmatched sample mean on the second occasion estimating $\bar{Y}$, $C_0 = \frac{S_y}{\bar{Y}}, C_i = \frac{S_{x_i}}{\bar{X}_i}, i = 1,2,\ldots,k$, $\Delta_i = \frac{C_i}{C_0}, i = 1,2,\ldots,k$, ρ_{0i} =Pearson correlation coefficient between x_i and y, i=1,2,...,k, ρ_{ij} =Pearson correlation coefficient between x_i and x_j, i, j=1,2, . . .,k $(i \neq j)$, $p = \frac{m}{n}$ the matching fraction.

2.3 The Multivariate Product Method of Estimation

The unmatched (u units) and matched (m units) portions of the second occasion sample provide independent estimates ($\bar{y}_m$ and $\bar{y}_u$) of the population mean on the second occasion $\bar{Y}$. For the matched portion an estimate improved of $\bar{Y}$ may be obtained using a double sampling multivariate product estimate

$$\bar{y}'_m = \omega_1 \frac{\bar{x}_1^m}{\bar{x}_1}\bar{y}_m + \ldots + \omega_k \frac{\bar{x}_k^m}{\bar{x}_k}\bar{y}_m$$

If $W_{1\times k} = (\omega_1, \ldots, \omega_k)$ is defined, we obtain

$$V(\bar{y}'_m) = \bar{Y}^2 WDW' \qquad (1)$$

where $D = (d_{ij})_{k\times k}$ is the matrix defined by

$$d_{ij} = \frac{1}{m}C_0^2 + \left(\frac{1}{m} - \frac{1}{n}\right)(C_iC_j\rho_{ij} + C_0C_i\rho_{0i} + C_0C_j\rho_{0j}) \qquad i, j = 1,2,\ldots,k$$

and the value W is obtained by maximizing the precision of $\bar{y}'_m$.

Thus, agree with Singh (1967) and proceeding as Olkin (1958) we obtain the optimum weighting vector given by

$$\hat{W} = \frac{eD^{-1}}{eD^{-1}e'}$$

where $e_{(1\times k)} = (1, \ldots, 1)$ and D^{-1} is the inverse matrix of D. Substituting into (1) we obtain the minimum variance of the estimator

$$V(\bar{y}'_m) = \bar{Y}^2 \hat{W} D \hat{W}'$$

Suppose the weights are uniform for the auxiliary variables $x_1, \ldots, x_k$ (Singh (1967)) the optimum weighting vector is given by

$$\hat{W}_{(1\times k)} = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$$

As an example, uniform weighting is obtained when

$C_i = C,\ \rho_{0i} = \rho_0$

and

$$\rho_{ij} = \rho\,(i \neq j) \text{ for } i, j = 1, 2, \ldots, k \qquad (2)$$

Then, we find that $\Delta_i = \Delta$, and provide the variance of $\bar{y}'_m$ given by

$$V_{min}(\bar{y}'_m) = \frac{S^2_y}{m}\left(1 + \frac{u}{n}\Delta\left(\frac{1+\rho(k-1)}{k}\Delta + 2\rho_0\right)\right)$$

An estimate of the variance may be obtained by replacing the populations parameters by their appropriate sample estimates.

If, however, the direct estimate $\bar{y}_m$ based on the m sampling units, its variance would be

$$V(\bar{y}_m) = \frac{S^2_y}{m}$$

we obtain that $\bar{y}'_m$ is more efficient than $\bar{y}_m$ if

$$\frac{k\rho_0}{1+\rho(k-1)} < -\frac{1}{2}\left(\frac{C}{C_0}\right)$$

Hence, we construct an estimate of the mean of the population on the second occasion, $\bar{Y}$, by combining the two independent estimates, $\bar{y}'_m$ and $\bar{y}_u$ with weights ω and $(1-\omega)$. Thus

$$\bar{y}_{2PM} = \omega\bar{y}'_m + (1-\omega)\bar{y}_u$$

and

$$V(\bar{y}_{2PM}) = \omega^2 V(\bar{y}'_m) + (1-\omega)^2 V(\bar{y}_u)$$

The best estimate of the mean $\bar{Y}$ on the second occasion is obtained by using the values of ω that minimize $V(\bar{y}_{2PM})$

$$\omega_{opt} = \frac{V(\bar{y}_u)}{V(\bar{y}_u) + V(\bar{y}'_m)} = \frac{p}{1+(1-p)^2 Z}$$

Also, $V(\bar{y}_u)$ is given by

$$V(\bar{y}_u) = \frac{S^2_y}{u}$$

and substituting in the variance, we have

$$V_{min}(\bar{y}_{2PM}) = \frac{V(\bar{y}'_m)V(\bar{y}_u)}{V(\bar{y}'_m) + V(\bar{y}_u)} = \frac{S^2_y}{n}\frac{1+qZ}{1+q^2Z} \qquad (3)$$

where $Z = \Delta\left(2\rho_0 + \Delta\dfrac{1+\rho(k-1)}{k}\right)$

If no uniform weighting, Z is given by

$$Z = \frac{1}{k^2}[2k\rho_{01}\Delta_1 + \ldots + 2k\rho_{0k}\Delta_k + \Delta_1^2 + \ldots + \Delta_k^2 + 2\Delta_1\Delta_2\rho_{12} + \underbrace{\ldots\ldots}_{\frac{k(k-1)}{2}} + 2\Delta_i\Delta_j\rho_{ij}]$$

$$i, j = 1, 2, \ldots, k \ (i \neq j)$$

The optimum matching fraction is obtained minimizing in (4) with respect to u, and so, we have

$$p_{opt} = \frac{\left[1+\Delta\left(2\rho_0 + \Delta\dfrac{1+\rho(k-1)}{k}\right) - \sqrt{1+\Delta\left(2\rho_0 + \Delta\dfrac{1+\rho(k-1)}{k}\right)}\right]}{\Delta\left(2\rho_0 + \Delta\dfrac{1+\rho(k-1)}{k}\right)}$$

and

$$V_{opt}(\bar{y}_{2PM}) = \frac{S^2_y}{2n}\left(1 + \sqrt{1-\Delta\left(-2\rho_0 - \Delta\frac{1+\rho(k-1)}{k}\right)}\right)$$

Consider the special case where

$\rho = -\rho_0$ and $C = C_0$

which gives an expression more simple for the variance

$$V_{min}(\bar{y}_{2PM}) = \frac{S^2_y}{n}\frac{1+q\left(\dfrac{1+(k+1)\rho_0}{k}\right)}{1+q^2\left(\dfrac{1+(k+1)\rho_0}{k}\right)}$$

3. COMPARISON OF ESTIMATORS

3.1 Unbiased Estimator and Combined Multivariate Product Estimator

The simple unbiased estimator $\bar{y}$ of the population mean on the current occasion $\bar{Y}$ is exclusively based on the n sampling units for the second occasion, using no information gathered on the first occasion, its variance is given by

$$V(\bar{y})=\frac{S_y^2}{n}$$

We can compute the gain in precision G of the combined estimate $\bar{y}_{2PM}$, obtained by using a double-sampling multivariate product estimate from the matched portion of the sample on the second occasion, over the direct estimate.

$$G=\frac{V(\bar{y})-V(\bar{y}_{2PM})}{V(\bar{y}_{2PM})}=\frac{-Zp(1-p)}{1+(1-p)Z}$$

where $Z=\Delta\left(2\rho_0+\Delta\frac{1+\rho(k-1)}{k}\right)$

Necessarily $p\le 1$. 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

$$\frac{k\rho_0}{1+\rho(k-1)}<-\frac{1}{2}\left(\frac{C}{C_0}\right)$$

Further, we conclude that the gain in precision of the combined estimate, $\bar{y}_{2PM}$, over the direct estimate, $\bar{y}$, increase with increasing ρ_0 absolute value (larger dependence between the auxiliary variables $x_1,\ldots,x_k$ with the variable under study y), and decreasing ρ (smaller correlation between x_i and x_j, i, j=1, 2,., . . .k $(i\neq j)$).

If an auxiliary variable x_1 on the stage of estimation is only employed, the combined estimate obtained by using an univariate product estimate from the matched portion on the second occasion, is given by

$$\bar{y}_{2p}=\omega\frac{\bar{x}_1^m}{\bar{x}_1}\bar{y}_m+(1-\omega)\bar{y}_u$$

and the gain in precision improve the direct estimate provided

$$\rho<-\frac{1}{2}\left(\frac{C}{C_0}\right)$$ (Artes *et al.* (1998))

3.2 Combined Estimator of Univariate Product versus Multivariate

A comparison of the precision of a combined estimate of multivariate product with the univariate product estimator from the matched portion of the sample, has been made. (Agree with Artes *et al.* (1998))

$$V_{min}(\bar{y}_{2p})=\frac{S_y^2}{n}.\frac{1+q(2\rho_0+1)}{1+q^2(2\rho_0+1)}$$

However, if the provided auxiliary information by $x_1,\ldots,x_k$ is utilized on the first occasion, and a double sampling multivariate product estimate from the matched portion of the sample of the second occasion is considered, we obtain an expression for $V_{min}(\bar{y}_{2PM})$ given by (4). In this situation, when the weights are uniform and is check the condition (2), we obtain

$$V_{min}(\bar{y}_{2p})-V_{min}(\bar{y}_{2PM})\ge 0$$

when

$$\frac{(1-\rho)(k-1)}{k}>0$$

The combined estimate of multivariate product is more precise than the estimate obtained by using an univariate product estimate from the matched portion.

4. SPECIAL CASE

A particular case of special interest in the theory for two auxiliary variables which has the frequent application

$$V_{min}(\bar{y}_{2P2})=\frac{S_y^2}{n}\frac{1+q\Delta\left(2\rho_0+\Delta\frac{1+\rho}{2}\right)}{1+q^2\Delta\left(2\rho_0+\Delta\frac{1+\rho}{2}\right)}$$

where

$$p_{opt}=\frac{\Delta\left(2\rho_0+\Delta\frac{1+\rho}{2}\right)-1-\sqrt{1-\Delta\left(2\rho_0-\Delta\frac{1+\rho}{2}\right)}}{\Delta\left(2\rho_0+\Delta\frac{1+\rho}{2}\right)}$$

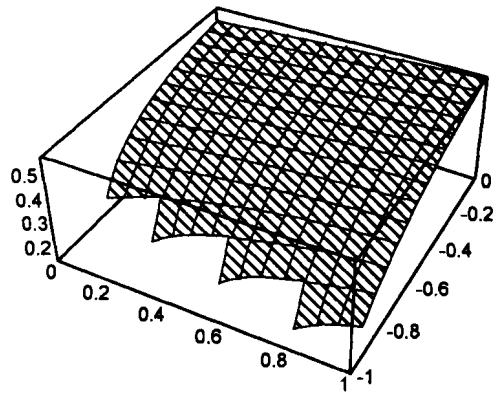

Fig.1. Optimum matching fraction for various values of ρ_0 and ρ, when $\Delta = 1$. The best percentage to match never exceeds 50% (Patterson (1950), Kulldorff (1963, Tikkiwal (1967)) and decrease steadily as ρ_0 (absolute value) increases

The gain in precision G of the combined estimate $\bar{y}_{2P2}$ obtained by using a double-sampling bivariate product estimate from the matched portion of the sample on the second occasion, over the direct estimate is given by

$$G = \frac{1 + q_{opt}^2\left(2\rho_0 + \frac{1+\rho}{2}\right)}{1 + q_{opt}\left(2\rho_0 + \frac{1+\rho}{2}\right)} - 1$$

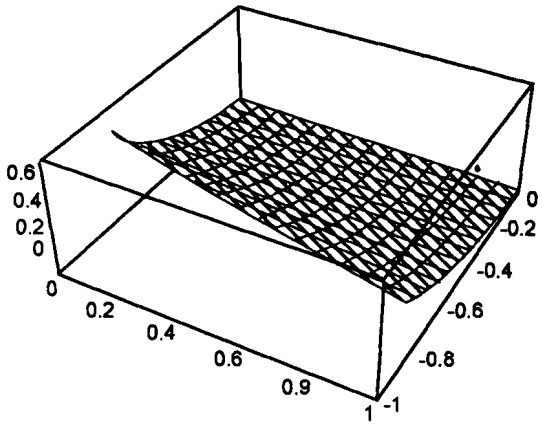

Fig.2. Gain in precision, for various values of ρ_0 and ρ, when $\Delta = 1$. The greatest attainable gain in precision is 100% when $\rho_0 = -1$. Unless ρ_0 (absolute value) is high, the gain are modest

Table 1. Percent gain in precision of a bivariate product estimate $\bar{y}_{2P2}$ over $\bar{y}$, when $\Delta = 1$

ρ_0	$\rho \downarrow p \rightarrow$	0.3	0.5
-0.9	0.7	59.55	45.24
-0.9	0.6	70.00	50.00
-0.8	0.7	33.16	30.00
-0.8	0.6	38.18	33.33
-0.8	0.5	44.07	36.95
-0.7	0.7	18.78	18.96
-0.7	0.6	21.72	21.43
-0.7	0.5	25.04	24.07

Let us consider

$$\rho = -\rho_0 \text{ and } C = C_0$$

which gives an expression more simple for the variance

$$V_{min}(\bar{y}_{2PM}) = \frac{S_y^2}{n} \frac{1 + q\left(\frac{1+3\rho_0}{2}\right)}{1 + q^2\left(\frac{1+3\rho_0}{2}\right)} \qquad (4)$$

The optimum matching fraction is obtained minimizing in (4) with respect to u, and so, we have

$$p_{opt} = \frac{1 + \frac{1+3\rho_0}{2} - \sqrt{1 + \frac{1+3\rho_0}{2}}}{\frac{1+3\rho_0}{2}}$$

The gain in precision is given by

$$G = \frac{-\frac{1+3\rho_0}{2} p(1-p)}{1 + (1-p)^{\frac{1+3\rho_0}{2}}}$$

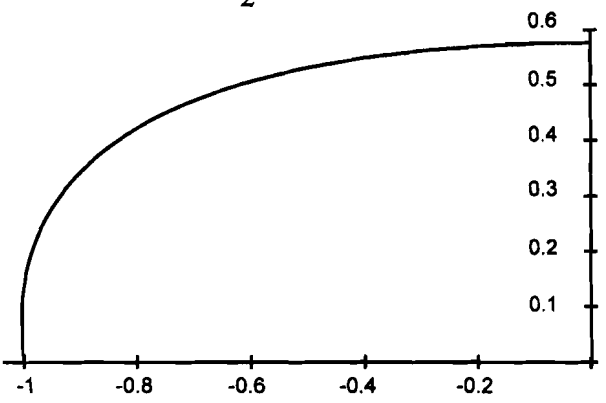

Fig.3. Optimum matching fraction, when $\rho = -\rho_0$

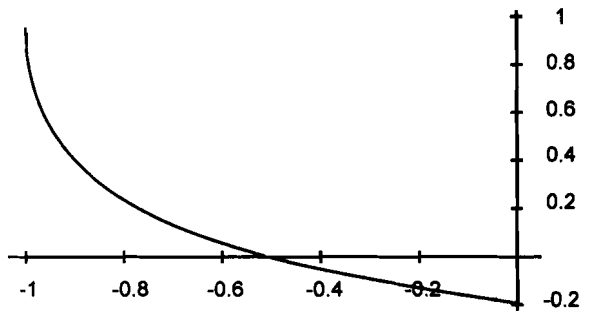

Fig.4. Gain in precision, when $\rho = -\rho_0$

4.1 Empirical Study

We have used the data collected, (Casimiro (1999)), in a survey on healthy habits and fitness level to assess the optimal operation of the proposed method. In order to achieve the targets of the study, we have considered the estimation of the add of fold (y, one of the multiple variables which affect the survey) at the second occasion, taking as auxiliary variables the arm maintained flexion (x_1) and the maximum volume of oxygen (x_2) from the first occasion.

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

First Occasion (April' 98) : Large sample n = 328, among the 2211 schoolchildren conforming the population.

Second Occasion (June' 98) : Matched sample m = 131, unmatched sample u = 197.

$\hat{C}_0 = 0.42$ $\quad \hat{\rho}_{01} = -0.50$

$\hat{C}_1 = 0.17$ $\quad \hat{\rho}_{12} = 0.49$

$\hat{C}_2 = 0.17$ $\quad \hat{\rho}_{02} = -0.50$

From these data we can state that

$$\hat{V}_{min} = (\bar{y}_{2P2}) = 0.91\frac{s_y^2}{n} < \frac{s_y^2}{n} = \hat{V}(\bar{y})$$

which means a gain in precision of 9.89% of the proposed estimator over the usual estimator.

We have also calculated the optimum matching fraction

$\hat{p}_{opt} = 45.55\%$

Moreover, we have compared the accuracy of the proposed estimator with other indirect estimators. Table 2 shows the results. As we can see, the ratio method is not efficient when the auxiliary variables are negatively correlated to the principal variable y, as the gain in accuracy over $\bar{y}$ is negative (G = –9.23%). However, the combined estimator based upon a bivariate product estimator for the matched part of the sample and a simple sample mean of the unmatched part, $\bar{y}_{2P2}$, is more accurate than the correspondent estimator which makes use of a univariate product estimator for the matched sample, $\bar{y}_{2P}$, and it even improves the accuracy of the one which makes use of regression estimator for the matched part, $\bar{y}_{2reg}$.

Table 2. Comparison of estimators

Estimators	Auxiliary variables	% Gain in precision
Direct $\bar{y}$	none	
Univariate Product $\bar{y}_{2P}$	x_1	6.72%
Bivariate Ratio $\bar{y}_{2R2}$	x_1 y x_2	-9.23%
Bivariate Regression $\bar{y}_{2reg}$	x_1 y x_2	8.89%
Bivariate Product $\bar{y}_{2P2}$	x_1 y x_2	9.89%

REFERENCES

Artes, E., Rueda, M. and Arcos, A. (1998). *Successive Sampling using a Product Estimate, Applied Sciences and the Environment.* Computational Mechanics Publications, 85-90.

Artes, E. and Garcia, A.V. (2000). A note on successive sampling using auxiliary information. *Proceedings of the 15th International Workshop on Statistical Modelling,* 376-379.

Artes, E. and Garcia, A.V. (2001). Estimating the current mean in successive sampling using a product estimate. *Conference on Agricultural and Environmental Statistical Application in Rome,* XLIII-1 - XLIII-2.

Avadhani, M.S. and Sukhatme, B.V. (1972). Estimation in sampling on two successive occasions. *Statistical Neerlandica,* **26(2),** 47-54.

Casimiro, A.J. (1999). Comparacion, evolucion y relacion de habitos saludables y nivel de condicion fisica-salud en escolares, entre final de Educacion Primaria (12 anos) y final de Educacion Secundaria Obligatoria (16 anos). Tesis Doctoral, Universidad de Granada.

Cochran, W.G. (1977). *Sampling Techniques.* Third edition. John Wiley & Sons, New York.

Garcia, A. (2001). *Mejora de Estimadores en Muestreo en Occasiones Sucesivas.* Servicio de Publicaciones de la Universidad de Almeria.

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

Kulldorff, G. (1963). Some problems of optimum allocation for sampling on two occasions. *Internat. Statist. Rev.*, **31**, 24-57.

Olkin, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, **43**, 154-165.

Patterson, H.D. (1950). Sampling on successive occasions with partial relacement of units. *J. Roy. Statist. Soc.*, **B12**, 241-255.

Sen, A.R. (1971). Increased precision in Canadian waterfowl harvest survey through successive sampling. *J. Wildl. Manag.*, **33**.

Sen, A.R. (1972). Successive sampling with two auxiliary variables. *Sankhya*, **B**, 371-378.

Sen, A.R., Sellers, S., Smith, G.E.J. (1975). The use of ratio estimate in successive sampling. *Biometrics*, **31**, 673-683.

Singh, M.P. (1967). Multivariate product method of estimation for finite populations. *J. Ind. Soc. Agril. Statist.* **19**(2), 1-10.

Singh, P. and Yadav, R.J. (1992). Generalised estimation under successive sampling. *J. Ind. Soc. Agril. Statist.*, **44**, 27-36.

Tuteja, R.K. and Bahl, S. (1991). Multivariate product estimators. *Cal. Stat. Assoc. Bull.*, **42**, 161-164.

Tikkiwal, B.D. (1951).*Theory of successive sampling*, Thesis for diploma, I.C.A.R., New Delhi.

Tikkiwal, B.D. (1953). Optimum allocation in successive samplings. *J. Ind. Soc. Agril. Statist.*, **5**, 100-102.

Tikkiwal, B.D. (1965). The theory of two stage sampling on successive occasions. *J. Ind. Statist. Assoc.*, **3**, 125-135.

Tikkiwal, B.D. (1967). Theory of multiphase sampling from a finite population of successive occasions. *Internat. Statist. Rev.*, **35(3)**, 247-263.

Yates, F. (1981). *Sampling Methods for Censuses and Surveys*. 4th ed., Griffin, London.