A NOTE ON OPTIMUM WEIGHTS IN MULTIVARIATE RATIO, PRODUCT AND REGRESSION ESTIMATORS

By

T. P. Tripathi

Indian Statistical Institute, Calcutta

(Received : September, 1975)

1. Summary And Introduction

An impetus to the work on the use of multivariate auxiliary information in forming the estimators for finite population mean (or total) was given by Olkin's paper in 1958 [*e.g.* Raj (1965), Srivastava (1966*a*) ; Singh (1967), Khan and Tripathi (1967) and Tripathi (1970-76), to cite a few). Let $U=\{U_1,\ldots,U_N\}$ be a finite population of N (given) units and let $y_0, y_1,\ldots y_p$ be $(p+1)$ variates defined on U. Considering simple random sampling, Olkin (1958) defined a multivariate ratio estimator for $\bar{Y}_0$, the population mean of a charactor y_0, as

$$\bar{Y}_{0y}=w'\alpha \qquad \ldots(1{\cdot}1)$$

using the supplementary information on $y_1,\ldots,y_p$, where $w'=(w_1, \ldots,w_p)$ are weights such that $w'e=1$, e being p-dimensional unit coloumn vector and $\alpha'=(\alpha_1,\ldots, \alpha_p)$, $\alpha_i=(\bar{Y}_0/\bar{Y}_i)\,\bar{y}_i$ and $\bar{Y}_i$ being the sample mean and population mean respectively of character y_i $(i=0, 1,\ldots,p)$.

The bias and mean square errors (MSE) of $\bar{Y}_{or}$ to terms of order n^{-1} (n being sample size) are given by

$$B\,(\bar{Y}_{0r})=K^*\bar{Y}_0\,w'b$$

and

$$M\,(\bar{Y}_{0r})=K^*\bar{Y}_0^2\;w'Aw$$

where $b'=(b_1,\ldots,b_p)$, $b_i=C_i^2 \;-\rho_{oi}\,C_o\,C_i, K^*=\left(\frac{1}{n}-\frac{1}{N}\right)$; $A=(a_{ik})$; $a_{ik}=C_0^2 \;-\rho_{oi}\,C_oC_i-\rho_{ok}\,C_oC_k+\rho_{ik}\,C_iC_i$; $i, k=1, \ldots, p$; C_i being the coefficient of variation of y_i and ρ_{ik} the correlation coefficient between y_i and y_k $(i, k=0, 1, \ldots, p)$.

He showed that the weight vector w which minimizes $M(\bar{Y}_{or})$ is given by

$$w_o = A^{-1}e/e'\,A^{-1}e \qquad \ldots(1{\cdot}2)$$

and then the resulting bias and MSE would be

$$B_o(\bar{Y}_{or}) = K^*\bar{Y}_o\,(e'\,A^{-1}/e'\,A^{-1}\,e)\,b$$

$$M_o\,(\bar{Y}_{or}) = K^*\bar{Y}_0^{\,2}\,(e'A^{-1}e)^{-1} = K^*\bar{Y}_0^{\,2}\left(\sum_{i=1}^{p}\sum_{k=1}^{p} a^{ik}\right)^{-1}$$

where a^{ik} is the (i, k)th element of A^{-1}.

The multivariate product and regression estimators for $\bar{Y}_o$ defined by Singh (1967) and Srivastava (1966*a*), in case of simple random sampling, are

$$\bar{Y}_{op} = w'\alpha \text{ with } \alpha_i = \bar{y}_o\bar{y}_i/\bar{Y}_i \qquad \ldots(1{\cdot}3)$$

and

$$\bar{Y}_{og} = w'\alpha \text{ with } \alpha_i = \bar{y} - b_i\,(\bar{y}_i - \bar{Y}_i)$$

respectively, where b_i is sample regression coefficient of y_o on y_i.

To the terms of order n^{-1}, bias, MSE and optimum weights (hence resulting bias and MSE also) for $\bar{Y}_{op}$ would be given by the same expressions as for $\bar{Y}_{or}$ with a change that in this case b_i and a_{ik} would be replaced by

$$b_i^* = \rho_{oi}\,C_o\,C_i$$

and

$$a_{ik}^* = C_0^2 + \rho_{oi}\,C_o\,C_i + \rho_{ok}\,C_o\,C_k + \rho_{ik}\,C_i\,C_k$$

respectively ($i, k = 1, \ldots, p$). Also, to the terms of order n^{-1}, a_{ik} in case of $\bar{Y}_{og}$ would be

$$a_{ik}^* = C_0^2\,\{1 - \rho_{0i}^2 - \rho_{0k}^2 + \rho_{oi}\,\rho_{0k}\,\rho_{ik}\}$$

We observe that the optimum weights in (1.2) depend upon the unknown parameters (C_o, C_i, ρ_{oi} and ρ_{yk}) and thus, in a sense, the estimators $\bar{Y}_{or}$, $\bar{Y}_{op}$ and $\bar{Y}_{og}$ are not well defined. If w_o were estimated by the sample at hand and used in defining $\bar{Y}_{or}$, $\bar{Y}_{op}$ and $\bar{Y}_{og}$ then w's in (1.1) and (1.3) would no longer be constants and complexity would arise. Further there would be deviations of the resulting MSE from optimum mean square errors $M_o\,(\bar{Y}_{or})$, $M_o(\bar{Y}_{op})$ and $M_o(\bar{Y}_{og})$.

For $p=2$, Srivastava (1966*b*) showed that if an auxiliary character y_2 is to be used, in addition to y_1, for defining $\bar{Y}_{or}$ and the optimum weights are not exactly known then in order to increase

the precision over that of using y_1 alone, the weight w_2 should satisfy

$$0 \leqslant w_2 \leqslant 2\, w_{02} \qquad \ldots(1.4)$$

where $w_{02}=(a_{11}-a_{12})\ (a_{11}+a_{22}-2a_{12})^{-1}$. (It is to be noted that this criterion would be true for every estimator of the form $\Sigma w_i \alpha_i$, in which case one would have $a_{ik}=E(\alpha_i-E\alpha_i)\ (\alpha_k-E\alpha_k)$ (Tripathi, 1970). But there is a trouble with this criterion also. How would one know that w_2 satisfies (1.4) ? It needs some approximate knowledge about w_{o2}.

In this note we give alternative weight vector in which coefficient of variations C_i of auxiliary characters and correlation coefficients ρ_{ik} $(i\neq k=1,\ldots,p)$ between them are assumed to be known. No knowledge about ρ_{oi} and C_o is needed, unlike to the case with w_o. At first we deal with the case $p=2$ and then with the general case of $p>2$ considering ratio, product and regression estimators. By an empirical study we demonstrate that there is no appreciable loss in precision by using the proposed weights, in the ratio estimator, compared to the optimum weights. We compare the efficiency of the modified ratio estimators too.

2. Modified Ratio and Product Estimators with Two Auxiliary Variates

Let the quantities C_1, C_2, $\bar{Y}_1$, $\bar{Y}_2$ and ρ_{12} be known. We define an Olkin-type ratio estimator for y_o using information on y_1 and y_2 as

$$\bar{Y}_r=\sum_{i=1}^{2} w_i\ (\bar{y}_o/\bar{y}_i)\ \bar{Y}_i \qquad \ldots(2.1)$$

where $w_1=(C_2^2-2\rho_{12}C_1C_2)/d=1-w_2;$

$$d=C_1^2+C_2^2-2\rho_{12}\ C_1\ C_2 \qquad \ldots(2.2)$$

We would have

$$B(\bar{Y}_r)\ =K^*\ \bar{Y}_0 \sum_{i=1}^{2} w_i\ b_i$$

$$M(\bar{Y}_r)=K^*\ \bar{Y}_0^2\left[\ w_1^2\ a_{11}+2w_1\ w_2\ a_{12}+w_2^2\ a_{22}\ \right]$$

It is easily seen that

$$B\ (\bar{Y}_r)-B_o\ (\bar{Y}_{or})=K^*\ \bar{Y}_o\ (a_1/d)\left[\ C_1^2-C_2^2+a\ \right]$$

and $M(\bar{Y}_r)-B_o(\bar{Y}_{or})=K^* \bar{Y}_0^2 (a^2/d)$

where $a=\rho_{o2} C_o C_2-\rho_{o1} C_o C_1.$

It is to be noted that if $\rho_{o2} C_2=\rho_{o1} C_1$, the weights in (2.1) would be same as the optimum weights in (1.2) for $p=2$; so there would be no loss in precision by using $\bar{Y}_r$ when exact optimum weights are not available which is usual in practice.

Let $\bar{Y}_{r1}=(\bar{y}_o/\bar{y}_1)\bar{Y}_1$ and $\bar{Y}_{r2}=(\bar{y}_o/\bar{y}_2)\bar{Y}_2$

We find that

$$B(\bar{Y}_{r1})-B(\bar{Y}_r)=K^* \bar{Y}_o w_2 \left[C_1^2-C_2^2+a \right]$$

$$M(\bar{Y}_{r1})-M(\bar{Y}_r)=K^* \bar{Y}_0^2 (dw_2)[w_2+2a/d] \quad \ldots(2.3)$$

$$B(\bar{Y}_{r2}-B(\bar{Y}_r)=K^* \bar{Y}_0 w_1 \left[C_2^2 - C_1^2-a \right]$$

$$M(\bar{Y}_{r2})-M(\bar{Y}_r)=K^* \bar{Y}_0^2 (dw_1)[w_1-2a/d]$$

From (2.3) we get that use of information on auxiliary character y_2 in addition to y_1, would result in increased precision over that of using y_1 alone

if $1-\rho_{12}(C_2/C_1)-2a/C_1^2>0$ in case $\rho_{12}C_2/C_1<1$ ...(2.4)

and if $1-\rho_{12}(C_2/C_1)+2a/C_1^2<0$ in case $\rho_{12} C_2/C_1>1$

Similarly $\bar{Y}_r$ would be better than $\bar{Y}_{r2}$

if $1-\rho_{12} C_1/C_2-2a/C_2^2>0$ in case $\rho_{12} C_1/C_2<1$

and if $1-\rho_{12} C_1/C_2-2a/C_2^2<0$ in case $\rho_{12} C_1/C_2>1$ ...(2.5)

If $\rho_{o2} C_2=\rho_{01} C_1$ then the estimetor $\bar{Y}_r$ would always be better than $\bar{Y}_{r1}$ and $\bar{Y}_{r2}$.

In case of product and regression esttmators we propose to use the same weights as in $\bar{Y}_r$. Thus we define

$$\bar{Y}_p=\sum_{i=1}^{2} w_i \bar{Y}_{pi};\ \bar{Y}_{pi}=\bar{y}_o \bar{y}_i/\bar{Y}_i$$

and $$\overline{Y}_g=\sum_{i=1}^{2} w_i\,\overline{Y}_{gi}\;;\;\overline{Y}_{gi}=\bar{y}_o-b_i\;(\bar{y}_i-\overline{Y}_i)$$

with w_i 's given by (2.2).

Let $$u=\rho_{01}^2-\rho_{01}\,\rho_{02}\,\rho_{12}$$

and $$v=\rho_{01}^2+\rho_{02}^2-2\rho_{01}\,\rho_{02}\,\rho_{12}$$

We would have

$$M(\overline{Y}_g)-M_0(\overline{Y}_{0g})=K^*\overline{Y}_0^2\quad C_0^2\;[(w_1v-u)^2/v]$$

$$M(\overline{Y}_{g1})-M(\overline{Y}_g)=K^*\overline{Y}_0^2\quad C_0^2\quad w_2^2\;[(2/w_2)(v-u)-v]$$

$$M(\overline{Y}_{g2})-M(\overline{Y}_g)=K^*\,\overline{Y}_0^2\quad C_0^2\quad w_1^2\;[(2/w_1)u-v]$$

and

$$B(\overline{Y}_p)-B_0(\overline{Y}_{0p})=K^*\overline{Y}_0(a^2/d)$$

$$M(\overline{Y}_p)-M_0\;(\overline{Y}_{0p})=K^*\overline{Y}_0^2\;\;(a^2/q)$$

$$B(\overline{Y}_{p1})-B(\overline{Y}_p)=K^*\overline{Y}_0w_2\;(-a/d)$$

$$B(\overline{Y}_{p2})-B(\overline{Y}_p)=K^*\overline{Y}_0w_1\;(a/d)$$

$$M(\overline{Y}_{p1})-M(\overline{Y}_p)+K^*\,\overline{Y}_2^2\;\;(dw_2)[w_2-2a/d]$$

$$M(\overline{Y}_{p2})-M\;(\overline{Y}_p)=K^*\overline{Y}_0^2\;\;(dw_1)[w_1+2a/d]$$

We note that $\overline{Y}_p$ would be better than $\overline{Y}_{p1}$ if (2·4) holds with (a) replaced by $(-a)$. Similarly $\overline{Y}_p$ would be better than $\overline{Y}_{q2}$ if (2.5) with (a) replaced by $(-a)$ holds true. If $\rho_{02}\,C_2=\rho_{01}\,C_1$, the estimator $\overline{Y}_p$ would always be better than $\overline{Y}_{p1}$ and Y_{p2}.

3. Estimators with Several Auxiliary Variaties

In case information on p-auxiliary characters $y_1,\ldots,y_p$ is available, we propose to use the modified multivariate ratio, product and regression estimators (as the situation be) defined by

$$\overline{Y}_r=w'\alpha,\;\alpha_i=(\bar{y}_0/\bar{y}_i)\;\overline{Y}_i\;;\;\;i=1,\ldots,p \qquad \ldots(3.1)$$

$$\overline{Y}_p=w'\alpha^*,\;\;\alpha_i^*\;=\bar{y}_0\;\bar{y}_i/\overline{Y}_i$$

and $$\overline{Y}_g=w'\alpha^0,\;\alpha^0\;=\bar{y}_0-b_i(\bar{y}_i-\overline{Y}_i)$$

respectively, where

$$w'=e'\;D^{-1}/e'\;D^{-1}\;e,\;D=(dik)\;i,\;\;k=1,\ldots,p \qquad \ldots(3.2)$$

$dik=\rho_{ik}\ C_i\ C_k$ and $e=(1,\ldots,1)'$ is p-dimensional unit column vector. The matrix D is assumed to be positive definite.

It may be easily verified that for $p=2$ the weights in (3.2) reduce to those given by (2.2).

In case

$$C_i=C \text{ and } \rho_{ik}=\rho(t\neq k)\ .\ i,\ k=1,\ldots,p \qquad \ldots(3\cdot3)$$

from (3.2) we get $w'=e'/p$ giving $w_i=1/p$ for all $i=1,\ldots,p$. Also under the conditions (3.3) we get,

$$M(\bar{Y}_r/t)=M(\bar{Y}_r/s)=\left(\ K^*\bar{Y}_0^2\ /ts\ \right)[(s-t)C^2(1-p)$$

$$+2C\ C_0\left\{\ s\ \sum_1^t \rho_{01}-t\sum_1^s \rho_{0i}\ \right\}\Big]$$

$$M(\bar{Y}_p/t)-M(\bar{Y}_p/s)=\left(K^*\bar{Y}_0^2\ /ts\right)[(s-t)\ C^2(1-p)$$

$$+2C\ C_0\left\{\ s\sum_1^r \rho_{0i}-t\sum_1^s \rho_{0i}\ \right\}\Big]$$

and

$$M(\bar{Y}_g/t)-M(\bar{Y}_g/s)=\left(K^*\bar{Y}_0^2\ \ C_0^2\ /t^2s^2\ \right)$$

$$\Big[(1-\rho)\Big\{s^2\sum_1^t \rho_{0i}^2-t^2\sum_1^s \rho_{0i}^2\ \Big\}$$

$$+\rho\Big\{s^2\left(\sum_1^t \rho_{0i}\ \right)^2-t^2\left(\sum_1^s \rho_{0i}\ \right)^2\Big\}$$

$$-2ts\left\{\ s\sum_1^t \rho_{0i}^2-t\sum_1^s \rho_{0i}^2\ \right\}\Big] \qquad \ldots(3.4)$$

where t in $(\bar{Y}_r/t)$ indicates that information on t auxiliary characters is being used in defining the estimator $\bar{Y}_r$. Thus use of auxiliary characters $y_1,\ldots,\ y_t\ldots y_s$ would increase the precision of $\bar{Y}_r$ over the use of $y_1,\ldots,\ y_t$ alone if

$$(s-t)\ C^2\ (1-\rho)>2\ C\ C_0\left\{\ s\sum_1^t \rho_{0i}-t\sum_1^s \rho_{0i}\ \right\} \qquad \ldots(3.5)$$

In case of $\bar{Y}_p$, we would get $2C\,C_0\left\{ t \sum_{1}^{s} \rho_{0i} - s \sum_{1}^{t} \rho_{0i} \right\}$

in right hand side of (3·5). In particular, from (3·5) we note that if ρ_{0i} are same, say ρ, for all $i=1,\ldots,p$ and $\rho \neq 1$, then inclusion of extra auxiliary characters always results in increased precision of the estimators $\bar{Y}_r$ and $\bar{Y}_p$. The same is true in case of the estimator $\bar{Y}_g$ also.

Further under the conditions (3·3) we get that $\bar{Y}_r$ and $\bar{Y}_p$ would be better than simple unbiased estimetor $\bar{y}_0$ if

$$2(C_0/C) \sum_{i=1}^{p} \rho_{0i} > (p-1)\,\rho + 1 \qquad \ldots(3.6)$$

and

$$-2\,(C_0/C) \sum_{i=1}^{p} \rho_{0i} > 1 + (p-1)\,\rho \qquad \ldots(3.7)$$

hold respectively. If $C_0 = C$ and $\rho_{0i} = \rho$, the conditions (3·6) and (3·7) reduce into $\rho > 1/(p+1)$ and $\rho < -1/(3p-1)$ as obtained by Olkin (1958) and Singh (1967) in respective cases. Further $\bar{Y}_g$ would be better than $\bar{y}_0$, under (3·3),

if

$$(2p-1) \sum_{i=1}^{p} \rho_{0i}^2 > \rho \left\{ \left(\sum_{i=1}^{p} \rho_{0i} \right)^2 - \sum_{i=1}^{p} \rho_{0i}^2 \right\}$$

which always holds in case ρ_{0i} are same for all $i=1,\ldots,p$.

4. An Empirical Study

We consider the population used by Olkin (1958). Here y_0, y_1 and y_2 are number of inhabitants in the cities under consideration in 1950, 1940 and 1930 respectively and we want to estimate $\bar{Y}_0$ using information on y_1 and y_2. For this population

$$C = (C_{ik}) = \begin{pmatrix} 1.049 & 1.059 & 1.056 \\ \ldots & 1.098 & 1.108 \\ \ldots & \ldots & 1.131 \end{pmatrix} \quad C_{ik} = \rho_{ik}\, C_i\, C_k \quad i, k = 0, 1, 2$$

$\bar{Y}_0 = 169900$ $\qquad$ $\rho_{01} = 0.987$

$\bar{Y}_1 = 148200$ $\qquad$ $\rho_{02} = 0.970$

$\bar{Y}_2 = 142000$ $\qquad$ $\rho_{12} = 0.995$

Table 4.1, below, gives the percent relative efficiency of $\bar{Y}_r$ compared to $\bar{Y}_{0r}$, $\bar{Y}_{r1}$, $\bar{Y}_{r2}$, $\bar{Y}_{10}$, $\bar{Y}_{02}$, $\bar{Y}_{03}$ $\bar{Y}_{0g}$, $\bar{Y}_g$, $\bar{Y}_{g1}$ and $\bar{Y}_{g2}$ where

$$\bar{Y}_{01}=\bar{y}_0\,(\bar{Y}_1/\bar{y}_1)\,(\bar{y}_2/\bar{Y}_2)\,;$$

$$\bar{Y}_{02}=\bar{y}_0\,(\bar{Y}_1/\rho_1)/(\bar{Y}_2/\bar{y}_2)$$

and $$\bar{Y}_{03}=\rho_0\,(\bar{Y}_1/\rho_1)/(\rho_2/\bar{Y}_2)$$

are the estimators considered by Singh (1965, 1967, 1969) with

$$M(\bar{Y}_{03})=K^*\,\bar{Y}_0^2\,[C_0^2+C_1^2-2\,C_{01}+\alpha^2\,C_2^2+2\,\alpha\,C_{02}-2\,\alpha\,C_{12}]$$

and the value of α which minimizes $M(\bar{Y}_{03})$ being given by $\alpha^*=(C_{12}-C_{02})/C_2^2$. The estimators $(\bar{Y}_{r1}, \bar{Y}_{r2}, \bar{Y}_{g1}$ and $\bar{Y}_{g2}$ are defined in Section 2.

For the given population we have

$$w_1=1.769,$$

$$w_2=-0.769,$$

$$w_{01}=2,$$

$$w_{02}=-1,$$

$$\alpha^*=0.046.$$

TABLE 4.1

Relative Efficiency of the Estimator $\bar{Y}_r$

Compared to Estimator	$\bar{Y}_r$	$\bar{Y}_{0r}$	$\bar{Y}_{r1}$	$\bar{Y}_{r2}$	$\bar{Y}_{01}$	$\bar{Y}_{02}$	$\bar{Y}_{03}(^*)$
% Relative Efficiency	100	94.1	170.6	400	6111.8	7435 3	158.8
Compared to Estimator			$\bar{Y}_0$	$\bar{Y}_{0g}$	$\bar{Y}_g$	$\bar{Y}_{g1}$	$\bar{Y}_{g2}$
% Relative Efficiency			6281.4	63.5	86.2	162.3	371.3

From the table we note that the percent relative loss in precision of $\bar{Y}_r$ compared to $\bar{Y}_{0r}$ is 5.9 which is moderate. However, though the percent relative loss in precision of $\bar{Y}_r$ compared to $\bar{Y}_g$ is not appreciable (13.8), it is high compared to $\bar{Y}_{0g}$ (36.5). The percent relative loss in precision of $\bar{Y}_g$ compared with $\bar{Y}_{0g}$ is 26.3. The percent relative gains in precision of $\bar{Y}_r$ over $\bar{Y}_{r1}$, $\bar{Y}_{r2}$, $\bar{Y}_{03}$, $\bar{Y}_{g1}$ and $\bar{Y}_{g2}$ are 70.6, 300, 58.8, 62.3 and 271.3 respectively. The estimators $\bar{Y}_{01}$, $\bar{Y}_{02}$ and $\bar{Y}_0$ are obviously very poor in this case.

ACKNOWLEDGEMENT

The author is thankful to the referee whose comments have led to the further improvement of the paper.

REFERENCES

[1] Khan, S. and Tripathi, T.P. (1967) : The use of multivariate auxiliary information in double sampling. *Jour. Ind. Stat. Assoc.*, Vol. 5, 43-48.

[2] Olkin, I. (1958) : Multivariate ratio estimation for finite populations. *Biometrika*, 45, 154-165.

[3] Raj, D. (1965) : On a method of using multi-auxiliary information in sample surveys. *J. Amer, Stat. Assoc.*, 60, 270-77.

[4] Singh, M.P. (1965) : On the estimation of ratio and product of the population parameters. *Sankhya*, B, 27, 321-28.

[5] Singh, M.P. (1967*a*) : Ratio cum product method of estimation. *Metrika*, 12, 34-42.

[6] Singh, M.P. (1967*b*) : Multivariate product method of estimation for finite populations. *Jour. Ind. Soc. Agri. Statist.*, Vol. XIX, 1-10.

[7] Singh, M.P. (1969) : Comparison of some ratio-cum-product estimators. *Sankhya*, B, 31, 375-78.

[8] Srivastava, S.K. (1966*a*) : On ratio and linear regression methods of estimation with several auxiliary variables. *Jour. Ind. Stat. Assoc.*, Vol. 4, 66-72.

[9] Srivastava, S.K. (1966*b*) : A note on Olkin's multivariate ratio estimator. *Jour. Ind. Stat. Assoc.*, Vol. 4, 202-208.

[10] Tripathi, T.P. (1970) : Contributions to the sampling theory using multivariate information. Ph. D. Thesis, Punjab University, Patiala.

[11] ——— (1976) : On double sampling for multivariate ratio and different methods of estimation. *Jour. Ind. Soc. Agri. Statist.*, Vol. XXVIII, 33-54.