# MULTIVARIATE PRODUCT METHOD OF ESTIMATION FOR FINITE POPULATIONS

By M. P. Singh

*Indian Statistical Institute, Calcutta*

## 1. Introduction

Whenever there is supplementary information available, the sampler tends to utilise it in the methods of estimation which gives maximum efficiency. The ordinary ratio method of estimation makes use of just one supplementary variable and gives more efficient estimator than simple unbiased estimator provided the variable under study is highly positively correlated with the supplementary variable. Quite often information on many supplementary variables are available in the survey which can be utilised to increase the precision of the estimate. Olkin (1958) has considered the use of multi-supplementary variables in building up ratio estimator and this estimator has been found to be efficient under much the same conditions. However, when the supplementary variables are negatively correlated with the study variable, ratio method of estimation cannot be efficiently used. In such cases, for the use of single supplementary variable, Murthy (1964) has considered the complementary situation of improving upon the unbiased estimator by considering a suitable product estimator. The present paper is an extension of the usual product estimator to the multi-variate case. The estimator has been introduced in a general form. In Section 2, exact expressions for bias and mean square error have been obtained and then comparisons have been made between different types of estimators in Section 3. The estimator has been extended to two-phase sampling scheme also. An empirical study is also included for illustration.

## 2. Multi-variate Product Method of Estimation

Let there be $k$ supplementary variables $(x_1)$, $(x_2)$, ...$(x_k)$ information on which is available for each unit of the population. Let $y$ and $x_i$ be the unbiased estimators of the parameters $Y$ and $X_i$

corresponding to the variable under study and the $i$-th supplementary variable ($i=1, 2, \ldots k$), based on any probability sample design and let $X_i$ be known in advance. Then the multi-variate product estimator may be defined as

$$\hat{Y}_p = \sum_{i=1}^{k} \frac{w_i p_i}{X_i} \qquad \ldots(2.1)$$

where $p_i = y\, x_i$

and weights $w_i$ 's are such that

$$\sum_{i=1}^{k} w_i = 1.$$

Now we shall obtain the exact expressions for the bias and mean square error of this estimator, though as usual, for the sake of comparisons in the next section, an approximation of the order $n^{-1}$ has been considered.

Let us write

$$x_i = X_i\,(1+e_i),\ i = 1, 2, \ldots, k$$

and $y = Y\,(1+e_o)$

where $E\,(e_i)=0$ for all $i$, $i=0, 1, 2, \ldots, k$.

The subscript 0 will indicate the variable $y$, such that $\rho_{oi}$ will mean the correlation coefficient between $y$ and $x_i$ and so on. Then we get,

$$E\,(\hat{Y}_p) = Y \sum_i w_i\, E\left(\frac{p_i}{P_i}\right) \qquad \ldots(2.2)$$

and

$$M\,(\hat{Y}_p) = Y^2 \sum_i \sum_j w_i w_j \operatorname{cov}\,(p_i, p_j)/P_i P_j \ \ldots(2.3)$$

where $P_i = YX_i$.

We note that

$$p_i = P_i\,(1+e_o)\,(1+e_i)$$
$$= P_i\,(1+\alpha_i+\beta_i)$$

where $\alpha_i = (e_o+e_i)$, $\beta_i = (e_o e_i)$

which gives

$$E\left(\frac{p_i}{P_i}\right) = 1 + E\,(\beta_i)$$
$$= 1 + b_i$$

and

$$\frac{\operatorname{cov}\,(p_i, p_j)}{P_i P_j} = E\,\alpha_i\,(\alpha_j+\beta_j) + E(\alpha_j\beta_i) + E(\beta_i\beta_j)$$
$$= a_{ij} + b_{ij} + c_{ij}$$

where
$$b_i = V^{11}_{oi}$$
$$a_{ij} = \left(V^{2}_{o} + V^{11}_{oi} + V^{11}_{oj} + V^{11}_{ij}\right)$$
$$b_{ij} = \left(V^{21}_{oi} + 2V^{111}_{oij} + V^{21}_{oj}\right)$$
$$c_{ij} = \left(V^{211}_{oij}\right)$$

and
$$V^{rst}_{oij} = \frac{E[(y-Y)^r (x_i - X_i)^s (x_j - X_j)^t]}{Y^r X^s_i X^t_j}.$$

Now substituting these values of
$$E\left(\frac{p_i}{P_i}\right) \text{ and } \frac{\text{cov }(p_i, p_j)}{P_i P_j}$$
in equation (2.2) and (2.3) we get the exact expression for the bias and mean square error of $\hat{Y}_p$ as

$$B\,(\hat{Y}_p) = Ywb' \qquad \text{...(2.4)}$$
$$M\,(\hat{Y}_p) = Y^2 w\,(A+B+C)w' \qquad \text{...(2.5)}$$

where
$$A = (a_{ij}),\ B = (b_{ij}),\ C = (c_{ij})$$
are matrices of order $k \times k$ each, $b$ and $w$ are vectors represented by $(b_1, b_2 \ldots b_k)$ and $(w_1 w_2 \ldots w_k)$ respectively. $b'$ and $w'$ are the transpose of $b$ and $w$. It may be noted that the matrices $A$, $B$ and $C$ are at least semi positive definite. Obviously the estimator will be unbiased only if $b' = 0$ which will happen if the correlation between $y$ and $x_i$ is equal to zero.

In case of sampling schemes such as simple random sampling or varying probability with replacement or any other sampling scheme involving selection of independent sub-samples,

$$E\left(\frac{p_i}{P_i}\right) \text{ and } \frac{\text{cov }(p_i, p_j)}{P_i P_j} \qquad \text{take the form}$$
$$E\left(\frac{p_i}{P_i}\right) = 1 + \frac{b_i}{n}$$

and
$$\frac{\text{cov }(p_i\, p_j)}{P_i\, P_j} = \frac{a_{ij}}{n} + \frac{b_{ij}}{n^2} + \frac{c_{ij}}{n^3}$$

where $n$ is the sample size or number of sub-samples.

This shows that the contribution of the terms involving $n^{-2}$ and $n^{-3}$ in the mean square error may be neglected for large values of $n$ in which case (2.5) reduces to

$$M(\hat{Y}_p)=Y^2 w\ Aw'. \qquad \ldots(2.6)$$

Following the procedure used by Olkin (1958) for determination of optimum weights it can easily be established that

$$w_i=\frac{\text{sum of the elements in } i^{th} \text{ column of } A^{-1},}{\text{sum of all } k^2 \text{ elements in } A^{-1}}$$

*i.e.*

$$w=\frac{e\,A^{-1}}{e\,A^{-1}\,e'},$$

where $e=(1,\ 1,\ \ldots\ 1)\ k\times 1$ and $A^{-1}$ is the matrix inverse of $A$.

Assuming that weights for all the supplementary variables are uniform which will happen only when the sums of each column of matrix $A$ are equal the optimum $\hat{w}$ is then given by $(e/k)$ and corresponding bias and mean square by

$$B(\hat{Y}_p)=Y\frac{eb'}{k} \text{ and } M(\hat{Y}_p)=Y^2\,\frac{d}{k}$$

where $d$ is the scalar such that $eA=ed$, $(d\neq 0)$ and for $d=0$ implies that $A$ is singular.

As an example of uniform weights let us suppose

$$C_i=C,\ \rho_{oi}=\rho_o$$

and

$$\rho_{ij}=\rho\ (i\neq j) \text{ for all } i=1,\ 1,\ 2\ldots k\ldots \qquad (2.7)$$

which gives the bias and mean square error as

$$B(\hat{Y}_p)=Y\,\rho_o C_o\,C \qquad \ldots(2.8)$$

and

$$M(\hat{Y}_p)=\frac{Y^2}{k}\left[C^2\,(1-\rho)+k\left(C_o^2+2\rho_o C_o C+\rho C^2\right)\right] \qquad \ldots(2.9)$$

where $C_{ij}=\overset{11}{\underset{ij}{V}}=\rho_{ij}\,C_iC_j$, $C_o$, $C_i$ and $C_j$ being the coefficients of variation of the estimates $y$, $x_i$ and $x_j$ respectively.

If in addition

$$\rho=\rho_o \quad \text{and} \quad C_o=C$$

then $$B(\hat{Y}_p)=Y\rho C^2$$

and

$$M(\hat{Y}_p)=\frac{Y^2}{k}(k+1-\rho\overline{1-3k})\,C^2.$$

## 3. Comparison of the Estimators

### 3·1. Unbiased estimator and multivariate product estimator

A comparison can be made when no supplementary information is employed *i.e.* simple unbiased estimator with the ordinary product estimator which utilise just one supplementary variable. It will be noticed that use of this product estimator will be justified if $\rho<-\frac{1}{2}(C/C_o)$. Similar criterion can easily be obtained by comparing mean square error of the present estimator given in equation (2.9) with the variance of unbiased estimator $y$ given by

$$V(y)=Y^2C_o^2.$$

We get the condition as

$$\frac{k\rho_o}{1+(k-1)\rho}<-\frac{1}{2}(C/C_o)$$

for the efficient use of $k$-variate product estimator. If $\rho_o=o$, $C_o=C$, the condition takes the form

$$\rho<-\frac{1}{3k-1}.$$

In this connection it may be mentioned that the multi-variate ratio estimator $\hat{Y}_r=\sum_{i=1}^{k} w_1 r_i X_i$ is more efficient than usual unbiased estimator $y$ under the condition (2.7) of uniform weights, when

$$\frac{k\rho_o}{1+(k-1)\rho}>\frac{1}{2}(C/C_o)$$

where $r_i$ is ratio of $y$ to $x_i$.

Hence the estimator to be used in a particular situation when $k$-supplementary variables are utilised in the above forms is the

Product estimator, $\Sigma\frac{w_i p_i}{X_i}$, if $-1\leqslant\frac{k\rho_o}{1+(k-1)\rho}<-\frac{1}{2}(C/C_o)$

Unbiased estimator, $y$, if $-\frac{1}{2}\,(C/C_o) \leqslant \frac{k\,\rho_o}{1+(k-1)\,\rho} \leqslant +\frac{1}{2}(C/C_o)$

Ratio estimator, $\Sigma\, w_i r_i X_i$, if $+\frac{1}{2}\,(C/C_o) < \frac{k\,\rho_o}{1+(k-1)\,\rho} \leqslant +1$

provided $Y$ and $X_1$'s are all positive. In case otherwise, the conditions will get changed accordingly. Now putting $k=1$, we get the conditions obtained by Murthy (1964).

### 3·2. Uni-variate versus multi-variate product estimator

*Theorem.* Let $\hat{Y}_{pq}$ and $\hat{Y}_{pk}$ denote respectively the multi-variate product estimators of $Y$ with optimum weights based on the sets of supplementary variables $(x_1)$, $(x_2)$,...$x_{(q)}$ and $(x_1)$, $(x_2)$...$x_{(k)}$ where $k$ is greater than $q$. Then

$$M\,(\hat{Y}_{pq}) \geqslant M(\hat{Y}_{pk}).$$

Proof of the theorem is similar to Olkin (1958) and is omitted here. If $k=1$, $\hat{Y}_{p1}$ denotes the uni-variate product estimator. As a particular case of the above theorem, when the weights are uniform and the condition (2·7) is satisfied, we get,

$$M(\hat{Y}_{pq})-M(\hat{Y}_{pk})=Y^2C^2\,(1-\rho)\left(\frac{k-q}{kq}\right) > 0.$$

Application of above results can be studied for any particular sampling scheme. The expression in case of simple random sampling can be easily obtained with the help of results derived by Sukhatme (1953).

## 4. An Empirical Study

For the purpose of the present study reference is made to an investigation undertaken by the Biometry Research Unit of the Indian Statistical Institute in connection with multi-variate investigation of blood chemistry. Such investigation entails the collection of multiple measurements on each individual examined to study the blood chemistry. Data were collected on 32 variables for three groups of individuals and details of finding are given by Das (1966). For the purpose of present illustration we consider the estimation of 'eosinophil' (one of the 32 variables) content, based on data for Group C collected on 69 individuals, and compare the unbiased, uni-variate product and two-variate product estimator, considering 'height' and 'weight' as the two supplementary variables, for a simple random sample of size $n$.

For two-variate product estimator, we have $M\hat{Y}_{p2})=\bar{Y}^2\ w\ A\ w'$ where $w=(w_1, w_2)$, $w'=\left(w_1/w_2\right)$ and

$$A=(a_{ij})_{2x2}=\begin{pmatrix} C_0^2+2C_{01}+C_1^2 & C_0^2+C_{01}+C_{02}+C_{12} \\ C_0^2+C_{02}+C_{01}+C_{21} & C_0^2+2C_{02}+C_2^2 \end{pmatrix}.$$

The optimum weight

$$w_1=\frac{2C+3C_{o2}+C_2^2+C_{01}+C_{12}}{4\ (C_0^2+C_{01}+C_{02})+C_1^2+C_2^2+2C_{12}}=1-w_2.$$

It may be mentioned that $C_i$ and $\rho_{ij}$ $(i\neq j+0, 1, 2)$ are the coefficient of variation and corr. coeff. respectively for the sample means. For the present scheme $C_i^2=RC_i'^2$, where $R=\dfrac{N-n}{(N-1)n}$ and $C_i'$ are defined for the corresponding variables.

We have the following values for this population

| | |
|---|---|
| $C_0'=0.60$ | $\rho_{01}=-0.1752$ |
| $C_1'=0.033$ | $\rho_{02}=-0.2505$ |
| $C_2'=0.28$ | $\rho_{12}=0.0099$ |

which gives $w=_1 0.48$ and $w_2=0.52$ and the corresponding m.s.e. as

$$M(\hat{\bar{Y}}_{p1})=\bar{Y}^2R(0.3541) \text{ and } M(\hat{\bar{Y}}_{p2})=\bar{Y}^2R(0.3343)$$

showing that the relative efficiency of unbiased, uni-variate product and two-variate product estimator as 100, 104 and 110 respectively It may be noted that this example is being given here by way of illustration and not suggesting the use of product etimator invariably for such studies as the gain in efficiency is not much.

## 5. Two-Phase Sampling Procedure

We shall suppose that information on the variables $(x_1)$, $(x_2)$, ..., $(x_k)$ is not readily available but could be collected rather inexpensively from a fairly large sample. In this case the preliminary sample of size $n_1$ is taken in which only $x_i$'s are observed. The second-phase sample of size $n_2$ in which $y$ alone is observed may be drawn in two ways, namely, (*i*) a sub-sample of the preliminary sample. (*ii*) independent sub-sample.

Assuming that a simple random sample (without replacement) is drawn at both the phases, we get the following estimators. Case (*i*). An estimator of the population mean can be defined as

$$\hat{\bar{Y}}_{pt}=\Sigma_i \, w_i \, p_i / \overline{X_i}' \qquad \ldots(5{\cdot}1)$$

where $p_i=\bar{y}x_i$ ; $\bar{y}$ and $\bar{x}_i$ being sample means based on the second phase sample and $\overline{X_i}'$ is the mean based on the preliminary sample. Here $\overline{X_i}'$ which is estimator of $\overline{X}_i$ is also subject to the sampling error. We get,

$$\begin{aligned} V(\hat{\bar{Y}}_{pt}) &= \overline{Y}^2 \; \Sigma_i \, \Sigma_j \, w_i \, w_j \, \text{cov}\,(p_i, p_j)/P_i'P_j' \\ &= \overline{Y}^2 \; \Sigma_i \, \Sigma_j \, w_i \, w_j \, d_{ij} \\ &= \overline{Y}^2 \; wDw' \qquad \ldots(5{\cdot}2) \end{aligned}$$

where $d_{ij}=\frac{1}{n_2}(1-f_2)C_0^{\,2}+\frac{1}{n_2}\;(1-f_1)(C_iC_j\,\rho_{ij}+\rho_{oi}C_0C_i+\rho_{oj}C_oC_j)$

and $D=(d_{ij})k\times k$ and $P_i'=\overline{Y}\,.\,\overline{X}_i'$.

The optimum $w_i$'s can be defined in the similar way replacing matrix $A$ by $D$. The mean square error of product estimator when only one supplementary information is used will be given as [assuming the weights $w_i{}'^s$ are uniform and (2·7) is satisfied],

$$M(\hat{\bar{Y}}_{pt})=\frac{\overline{Y}^2}{n_2}[(1-f_2)\;C_0^{\,2}+(1-f_1)(C^2+2\;\rho_oC_oC)]$$

where $f_1\left(=\frac{n_2}{n_1}\right)$ and $f_2\left(=\frac{n_2}{N}\right)$ are f. p.c.

Ignoring $f_2$.

$$M(\hat{Y}_{pt})=\overline{Y}^2\left[\frac{C_0^{\,2}+C^2+2\rho_0C_0C}{n_2}-\frac{C^2+2\rho_0C_0C}{n_1}\right] \qquad \ldots(5{\cdot}3)$$

**Case** (*ii*). The estimator of the mean in case of independent sub-sample would be of the same form as given by equation (5·1). Its mean square error in this case would be

$$V(\hat{\bar{Y}}_{pt})=\bar{Y}^2\ wGw' \qquad \ldots(5{\cdot}4)$$

where

$$G=(g_{ij})k\times k,$$

$$g_{ij}=\frac{1}{n_2}\ (1-f_2)\ (C_0^2+\rho_{oi}C_oC_i+\rho_{oj}C_oC_j+\rho_{ij}C_iC_j)$$

$$+\frac{1}{n_1}\left(1-f_1'\right)\rho_{ij}C_iC_j$$

and

$$f_1'=\frac{n_1}{N}.$$

If only one supplementary variable is used, the mean square error of the product estimator (assuming the weights to be uniform) is

$$M(\hat{\bar{Y}}_{pt})=\bar{Y}^2\left[\left(\frac{C_0^2+2\rho_0C_0C+C^2}{n_2}\right)+\frac{C^2}{n_1}\right] \qquad \ldots(5{\cdot}5)$$

Both the equations (5·3) and (5·5) are similar to the expressions for ratio estimator under similar conditions (Cochran 1963).

The author is grateful to Professor D. B. Lahiri and Professor M. N. Murthy for help and encouragement in the preparation of this paper. Thanks are also due to Prof. B. C. Das for providing the necessary data for the empirical study.

## Summary

A multi-variate product estimator of the population total, similar to Olkin's multi-variate ratio estimator, using data on two or more supplementary variables, has been suggested here. Exact expressions for the bias and the mean square error of the estimator are derived and a comparison with the unbiased and the multi-variate ratio estimators has been made for any sampling design, giving the specific situations under which either of them may be efficiently used. The estimator has also been extended to two-phase sampling where the data on the supplementary variables are not at hand but are collected from a large first-phase sample. An empirical study is also included for illustration.

## REFERENCES

1. Cochran, W. G. (1963) Sampling Techniques, Second Edition, New York ; John Wiley and Sons.

2. Das, B. C. (1966) Multi-variate investigations of blood chemistry and morphology, proceedings of the Symposium on human adaptability to environments and physical fitness, Defence Institute of Physiology and Allied Sciences, Madras-3.

3. Goodman, L. A. (1960) 'On the exact variance of products', J. Amer, Stat. Ass., 33, 708-713.

4. Murthy, M. N. (1964) 'Product method of estimation'. Sankhya Series A, 26, 69-74.

5. Olkin, I. (1958) Multi-variate ratio estimation for finite population, Biometrika 43, 154-63.

6. Sukhatme, P.V. (1953) 'Sampling theory of surveys with applications' Indian Society of Agricultural Statistics, India.