A NOTE ON GENERAL DOUBLE SAMPLING SCHEMES FOR MULTIVARIATE PRODUCT METHOD OF ESTIMATION

BY

S.K P. SINHA

I.I.T., Delhi

(Received : January, 1974)

1. INTRODUCTION

The problem considered is to estimate Y, the population total of a character y, using the information on a p-dimensional random vector of auxiliary characters, $X=(X_1,\ldots\ldots, X_p)$, in forming a multivariate product estimator, for any sampling design. Let X_i denote the population total of the character x_i $(i=1, 2,\ldots\ \ldots, p)$. Singh[5] gave the multivariate product estimator

$$\hat{Y}_p=\sum_{i=1}^{p} w_i\ \hat{Y}\hat{X}_i/X_i,$$

where $\hat{Y}$ and $\hat{X}_i$ are unbiased estimators of Y and X_i respectively based on any sampling design and w_i's are the weights such that

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

He extended his results to two phase sampling where, he considered equal probability selection at both the phases. In this note we extend the estimator $\hat{Y}_p$ to the general double sampling schemes and the results obtained by Singh[5] are derived as a particular case thereof. Following Tripathi[6] we present a general technique of two phase sampling as follows.

*Present address : Bihar State Dairy Corporation, Patna.

We select a preliminary large sample of n units at moderate cost according to a specified sampling design and then a subsequent sample of m units, according to a different (or same) sampling design. The second sample may either be a subsample of the first or independent of the first.

Let E_1 (.), V_1 (.), C_1(.,.) denote the unconditional expectation (mean), variance and covariance and; E_2(.), V_2(.), C_2(.,.) the conditional mean, variance and covariance over the variation in the second phase sample keeping the first phase sample fixed. Let $\hat{X}_{i(1)}$ be an unbiased estimator of X_i based on the observations of the first phase sample alone and $\hat{X}_{i(2)}$ and $\hat{Y}_{(2)}$ be the unbiased estimators of X_i and Y respectively based on the observations of the second phase sample. That is

$$E\hat{X}_{i(2)}=E\hat{X}_{i(1)}=X_i \text{ and } E\hat{Y}_{(2)}=Y \qquad \text{...(1)}$$

It is to be noted that $\hat{X}_{i(2)}$ and $\hat{Y}_{(2)}$ may or may not utilize the information in the first phase sample.

For the case of subsample we shall assume that $\hat{X}_{i(2)}$ is conditionally unbiased for $\hat{X}_{i(1)}$, that is

$$E_2\hat{X}_{i(2)}=\hat{X}_{i(1)} \qquad \text{...(2)}$$

It is to be noted that

$$C_1(\hat{X}_{i(2)}, \hat{X}_{i(1)})=C_2(\hat{Y}_{(2)}, \hat{X}_{i(1)})=0 \qquad \text{...(3)}$$

in the case of subsamples and independent samples both.

As $\hat{X}_{i(2)}$ and $\hat{Y}_{(2)}$ may utilize the information on the first phase sample,

$$\text{Cov }(\hat{Y}_{(2)}, \hat{X}_{i(1)}) \text{ and Cov }(\hat{X}_{i(2)}, \hat{X}_{i(1)})$$

would not be zero for some sampling schemes. We shall use following well-known results :

$$E(d)=E_1E_2(d)$$
$$V(d)=V_1E_2(d)+E_1V_2(d)$$

and

$$\text{Cov }(d, d^*)=C_1(E_2\, d, E_2\, a^*)+E_1\, C_2(d, d^*) \qquad \text{...(4)}$$

for any estimators d and d^*.

2. Product Method of Estimation Using a General Double Sampling Scheme :

We define the multivariate Product estimator for Y as,

$$\hat{Y}_p = \sum_{i=1}^{p} w_i\alpha_i \text{ where } \alpha_i = \hat{Y}_{(2)}\hat{X}_{i(2)}/\hat{X}_{i(1)} \qquad ...(5)$$

Let $\delta_u = (u - Eu)/Eu$ for a random variable u.

Theorem 1.

The general multivariate product estimator $\hat{Y}_p$ is in general biased. It is a consitent estimator of Y and its bias $B(\hat{Y}_p)$ and mean squares error (MSE) $M(\hat{Y}_p)$ are given by.

$$B(\hat{Y}_p) = wq' + \phi \qquad ...(6)$$

and

$$M(\hat{Y}_p) = wAw' + \phi \qquad ...(7)$$

where

$$q = (q_1, q_2, ..., q_p)$$

$$q_i = 1/X_i[R_i V(\hat{X}_{i(1)}) - \text{Cov}(\hat{Y}_{(2)}, \hat{X}_{i(1)}) - R_i \text{ Cov}(\hat{X}_{i(2)}, \hat{X}_{i(1)}) + \text{Cov}(\hat{Y}_{(2)}, \hat{X}_{i(2)})] \qquad ...(8)$$

$$A = (a_{ik}) \qquad i,k = 1,2,...,p$$

$$a_{ik} = V(\hat{Y}_{(2)}) + R_i\{\text{Cov}(\hat{Y}_{(2)}, \hat{X}_{i(2)}) - \text{Cov}(\hat{Y}_{(2)}, \hat{X}_{i(1)})\} + R_k\{\text{Cov}(\hat{Y}_{(2)}, \hat{X}_{k(2)}) - \text{Cov}(\hat{Y}_{(2)}, \hat{X}_{k(1)})\} + R_iR_k \ \{\text{Cov}(\hat{X}_{i(2)}, \hat{X}_{k(2)})\} - \text{Cov}(\hat{X}_{i(1)}, \hat{X}_{k(2)}) - \text{Cov}(\hat{X}_{i(2)}, \hat{X}_{k(1)}) + \text{Cov}(\hat{X}_{i(1)}, \hat{X}_{k(1)})\} \qquad ...(9)$$

$$R_i = Y/X_i$$

ϕ denotes the product moments (central) of third and higher order. It may be noted that for a number of sampling schemes (*e.g.*, mentioned in section 3) the terms ϕ are of the order less than or equal to the order $m^{-3/2}$ where m is the second phase sample size.

Proof :

By expanding α_i through Taylor's series expansion at the points

$$\hat{Y}_{(2)}=Y,\ \hat{X}_{i(1)}=X_i \text{ and } \hat{X}_{k(1)}=X_t$$

we have

$$\begin{aligned}
\alpha_i-Y=&(\hat{Y}_{(2)}-Y)+(\hat{X}_{i(2)}-X_i)\frac{Y}{X_i}-(\hat{X}_{i(1)}-X_i)\ \frac{Y}{X_i}\\
&+\frac{1}{2\,!}[2(\hat{X}_{i(1)}-X_i)^2\ \frac{Y}{X_i^2}\\
&+2(\hat{Y}_{(2)}-Y)(\hat{X}_{i(2)}-X_i)\ \frac{1}{X_t} \qquad \ldots(10)\\
&-2(\hat{Y}_{(2)}-Y)\ (\hat{X}_{i(1)}-X_i)\ \frac{1}{X_t}\\
&-2(\hat{X}_{t(2)}-X_i)(\hat{X}_{i(1)}-X_t)\ \frac{Y}{X_t^2}]+\ldots.
\end{aligned}$$

$$\begin{aligned}
E(\alpha_i-Y)\ =\ &\frac{1}{X_i}[R_i V(\hat{X}_{i(1)})+\text{Cov}(\hat{Y}_{(2)},\ \hat{X}_{i(2)})\\
&-\text{Cov}(\hat{Y}_{(2)},\ \hat{X}_{i(1)})-\text{Cov}(\hat{X}_{i(1)},\ \hat{X}_{i(2)})]+\phi \qquad \ldots(11)
\end{aligned}$$

Since

$$B(\hat{Y}_p)=\sum_{i=1}^{p} w_i\ E(\alpha_i-Y) \qquad \ldots(12)$$

then substituting the value from (11) in (12) we get the result (6).

Now

$$M(\hat{Y}_p)=\sum_{i=1}^{p}\ \sum_{k=1}^{p} w_i w_k\ (\alpha_i-Y)E(\alpha_k-Y) \qquad \ldots(13)$$

using (10) we find that,

$$E(\alpha_i-Y)(\alpha_k-Y)=a_{ik}+\phi \qquad \ldots(14)$$

From (13) and (14) we get the result (7).

The consistency of $\hat{Y}_p$ follows from the fact that $B(\hat{Y}_p)$ and $M(\hat{Y}_p)$ are each $0(m^{-1})$ hence as m becomes large $B(\hat{Y}_p)$ and $M(\hat{Y}_p)$ each tends to zero.

Remark :

An unbiased multivariate difference estimator for Y in general double sampling schemes is defined by Tripathi[6] as

$$\hat{Y}_d = \sum_{i=1}^{p} w_i \, \alpha_i$$

where

$$\alpha_i = \hat{Y}_{(2)} - \lambda_i(\hat{X}_{i(2)} - \hat{X}_{i(1)}) \qquad ...(15)$$

We observe that $V(\hat{Y}_d) = M_a(\hat{Y}_p)$ provided $\lambda_i = -R_i$ where $M_a(\hat{Y}_p)$ represents $M(\hat{Y}_p)$ to the terms of order (m^{-1}).

If the second phase sample is a subsample of the first then using (2), (3) and (4) the expressions (8) and (9) reduce to

$$q_i = (1/X_i)\, E_1C_2(\hat{Y}_{(2)}, \hat{X}_{i(2)}) \qquad ...(16$$

and

$$a_{ik} = V_1(\hat{Y}_{(1)}) + E_1V_2(\hat{Y}_{(2)}) + R_iE_1C_2\,(\hat{Y}_{(2)}, \hat{X}_{i(2)})$$
$$+ R_kE_1C_2(\hat{Y}_{(2)}, \hat{X}_{k(2)}) + R_iR_kE_1C_2(\check{X}_{i(2)}, \check{X}_{k(2)}) \qquad ...(17$$

where

$$\hat{Y}_{(1)} = E_2(\hat{Y}_{(2)}).$$

If the second phase sample is independent of the first the expressions (8) and (9) reduce to

$$q_i = (1/X_i)[R_i\, V(\hat{X}_{i(1)}) - C_1(E_2\hat{Y}_{(2)}, \hat{X}_{i(1)})$$
$$- R_iC_1(\hat{X}_{i(1)}, E_2(\hat{X}_{i(2)})$$
$$+ \mathrm{Cov}(\hat{Y}_{(2)}, \hat{X}_{i(2)})] \qquad ...(18)$$

and

$$a_{ik} = V\,(\hat{Y}_{(2)}) + R_i\,\{\mathrm{Cov}\,(\hat{Y}_{(2)}, \hat{X}_{i(2)}) - C_1(E_2\hat{Y}_{(2)}, \hat{X}_{i(1)})\}$$
$$+ R_k\{\mathrm{Cov}(\hat{Y}_{(2)}, \hat{X}_{k(2)}) - C_1(E_2\hat{Y}_{(2)}, \hat{X}_{k(1)})\}$$
$$+ R_iR_k\,\{\mathrm{Cov}(\hat{X}_{i(2)}, \hat{X}_{k(2)}) - C_1(E_2\hat{X}_{k(2)}, \hat{X}_{i(1)})$$
$$- C_1(E_2\hat{X}_{i(2)}, \hat{X}_{k(1)}) + \mathrm{Cov}(\hat{X}_{i(1)}, \hat{X}_{k(1)})\} \qquad ...(19$$

If the estimators $\hat{X}_{i(2)}$ and $\hat{Y}_{(2)}$ do not make use of the information on the first phase sample (in which case $E_2\hat{X}_{i(2)}=X_i$ and $E_2\hat{Y}_{(2)})=Y)$

then (18) and (19) reduce to

$$q_i=(1/X_i)[R_i\ V(\hat{X}_{i(1)})+\text{Cov}(\hat{Y}_{(2)},\ \hat{X}_{i(2)})] \qquad \text{...(20)}$$

and

$$\begin{aligned} a_{ik}=V(\hat{Y}_{(2)})+R_i\ \ &\text{Cov}\ (\hat{Y}_{(2)},\ \ \hat{X}_{i(2)})\\ &+R_k\ \text{Cov}\ (\hat{Y}_{(2)},\ \hat{X}_{k(2)})\\ &+R_iR_k\{\text{Cov}(\hat{X}_{i(2)},\ \hat{X}_{k(2)})\\ +\text{Cov}\ (\hat{X}_{i(1)},\ \hat{X}_{k(1)})\} \end{aligned} \qquad \text{...(21)}$$

If m is large enough so that the term ϕ may be neglected, following the procedure used by Olkin[4] we would have

$$\begin{aligned} w_{opt} &= eA^{-1}/eA^{-1}e'\\ B_{opt} &= (eA^{-1}/eA^{-1}e')\,q'\\ M_{opt} &= 1/eA^{-1}e' \end{aligned}$$

where e is $1\times p$ unit row vector.

3. Particular Sampling Schemes.

The general results obtained in section 2, can be used to obtain results for particular sampling schemes.

In case the samples at both phases are simple rendom samples, we may define

$$\hat{X}_{i(1)}=N\bar{x}_{in},\ \ \hat{X}_{i(2)}=N\bar{x}_{im},\ \hat{Y}_{(2)}=N\bar{y}_m \text{ where } \bar{x}_{im} \text{ and } \bar{y}_m$$

denote the sample means based on the observations for the characteristic x_i and y respectively. We can find that using the inequality,

$$Euv\leqslant(Eu^2)^{1/2}\ E(v^2)^{1/2}$$

for any variable u and v and noting that $\mu_2(\bar{y}_m)=0(m^{-1})$ and $\mu_4(\bar{y}_m)=0(m^{-2})$ (Cramer,[1]).

$$\begin{aligned}\mu_{111}(\bar{y}_m, \bar{x}_{im}, \bar{x}_{km}) &= E(\bar{y}_m-\bar{y})(\bar{x}_{im}-\bar{x}_i)(\bar{x}_{km}-\bar{x}_k)\\ &\leqslant \{E(\bar{y}_m-\bar{y})^2\}^{1/2}\ \{E(\bar{x}_{im}-\bar{x}_i)^2(\bar{x}_{km}-\bar{x}_k)^2\}^{1/2}\\ &\leqslant \{0(m^{-1})\}^{1/2}\{E\bar{x}_{im}-x_i)^4\}^{1/4}\{E(\bar{x}_{km}-\bar{x}_k)^4\}^{1/4}\\ &\leqslant \{0(m^{-1})\}^{1/2}\{0(m^{-2})\}^{1/4}\{0(m^{-2})\}^{1/4}\\ &\leqslant 0(m^{-3/2}).\end{aligned}$$

Thus for large m the terms ϕ in (6) and (7) may be neglected and then we get the results as obtained by Singh[5] for double sampling scheme.

In case the population consists of L strata, the h-th stratum consisting of N units,

$$\sum_{h=1}^{L} N_h = N$$

and simple random samples of n_h and m_h units are selected from the h th stratum at the first phase and the second phase respectively with

$$n = \sum_{h=1}^{L} n_h, \quad m = \sum_{h=1}^{L} m_h$$

we may define

$$\hat{x}_{i(1)} = \sum_{h=1}^{L} N_h \bar{x}'_{ih}, \quad \hat{x}_{i(2)} = \sum_{h=1}^{L} N_h \bar{x}_{ih},$$

$$\hat{y}_{(2)} = \sum_{h=1}^{L} N_h \bar{y}_h$$

Where $\bar{x}'_{ih}$ and $\bar{x}_{ih}$ are sample means of n_h units and m_h units respectively from stratum h. Following the above inequality we can find that in this case also the term ϕ is $0(m^{-3/2})$ and thus we get the result.

In case a first phase sample of n units is selected with equal probabilities without replacement on which z and x are observed and the second phase sample of m units is selected with ppz and with replacement.

In this case let,

$$\hat{X}_{i(1)} = N\bar{x}_{\ i(n)}\,\hat{X}_{i(2)} = N/(nm) \sum_{j=1}^{m} \ (x_{ij}/p_j),\ \ \hat{Y}_{(2)} = (N/nm) \sum_{j=1}^{m} (Y_j/p_j)$$

where

$$p_j = z_j / \sum_{j=1}^{n} z_j$$

Again following the above inequality we can find that the term ϕ is $\leqslant 0(m^{-3/2})$ and from (16), (17), (20) and (21) we get the results obtained by Des Raj[2].

Acknowledgements

The author is indebted to Dr. T.P. Tripathi for his valuable guidance and discussion. The author also wishes to thank the referee for many helpful suggestions.

References

[1] Cramer, H. (1946) : Mathematical Methods of Statistics. University Press, Princeton.

[2] Des Raj (1964) : On double sampling for pps estimation *Ann. Math. Stat.* 35 900-902.

[3] Des Raj (1965) : On sampling over the two occasions with pps, *Ann. Math. Stat.* 36, 327-30.

[4] Olkin, I (1958) : Multivariate ratio estimation for finite population *Biometriks*; 45, 154-65.

[5] Singh, M.P. (1967) : Multivariate Product method of estimation for finite population. *J. Ind. Soc. Agri. Stat.* XIX.

[6] Tripathi, T.P. (1970) : *Contributions to the sampling theory using multivariate information.* Unpublished Ph. D. thesis; Punjabi University, Patiala.