# A GENERALIZED TWO-PHASE SAMPLING ESTIMATOR

By
SURENDRA K. SRIVASTAVA
*Punjabi University, Patiala*

(Received : May, 1979)

## SUMMARY

**In this paper a large class of ratio and product type estimators in two-phase sampling is defined. This class includes many estimators considered by various authors in recent years, as its members, *e.g.*, by Srivastava (1970) and Gupta (1978). It has been shown that the asymptotic minimum variance for any estimator of the class is equal to that of the conventional linear regression estimator for the case of two-phase sampling when second phase sample is a subsample of the first phase sample. For the case when the two samples are drawn independently, an explanation is given for the lower value of the minimum variance of the proposed class of estimators than that of the conventional linear regression estimator, as also obtained by Srivastava (1970) and Gupta (1978).**

## 1. INTRODUCTION

In survey sampling we are often concerned with estimating the mean $\bar{Y}$ of a certain characteristic y of the population of size $N$. The precision of estimators of $\bar{Y}$ can be increased by utilizing advance information about a suitable auxiliaiy characteristic $x$ correlated with $y$. Out of many the ratio method has been widely used when the correlation between $y$ and $x$ is positive. If this correlation is negative, a product estimator instead of a ratio estimator may be used. Recently, considerable attention has been given to forming ratio-type estimators and/or product-type estimators which are better than the conventional estimators. A few examples are Srivastava (1967), Walsh (1970), Reddy (1973, 1974) and Gupta (1978). These estimators require a knowledge of the population mean $\bar{X}$ of $x$. When the population mean $\bar{X}$ is not known, it is sometimes estimated from a preliminary large sample on which only the characteristic $x$ is observed. The value of $\bar{X}$ in the estimator is then replaced by this estimate. This technique has been called in literature, two-phase sampling or double sampling.

In a recent article, Gupta (1978) has considered estimators replacing $(\bar{X}/\bar{x})$ and $(\bar{x}/\bar{X})$ in the conventional ratio and product estimators by the higher order polynomials in $(\bar{X}/\bar{x})$ and $(\bar{x}/\bar{X})$ respectively. He showed that to the first degree of approximation, the optimum variance of the proposed estimators containing quadratic terms in $(\bar{X}/\bar{x})$and $(\bar{x}/\bar{X})$ is equal to that of the linear regression estimator and such estimators can not improve upon linear regression estimator even if a third degree polynomial is taken, which suggests that even if higher order polynomials are taken, the variance of the proposed estimators to the first degree of approximation, will not decrease. In fact this follows immediately from a general result given by Srivastava (1971, theorem 3.1) in the particular case of a single auxiliary variable. His result for the particular case of a single auxiliary variable is that up to terms of order $n^{-1}$, generally taken for ratio and product type estimators, estimators of the form $\bar{y}\ h\ (\bar{x}/\bar{X})$ are no more efficient than the linear regression estimator, where $h(.)$ is a function satisfying certain regularity conditions.

When $\bar{X}$ is not known, Gupta (1978) has extended the results of his proposed estimator to the two-phase sampling procedure and has shown that the proposed estimator is superior even to the two-phase sampling linear reggression estimator when the second phase sample is drawn independently of the first phase sample. The same result was obtained by Srivastava (1970) for his exponential estimator (See Srivastava, 1967).

In this paper a large class of ratio-type (product-type) estimators in two-phase sampling has been considered and which will include Gupta's (1978) estimator and many other estsmators as special cases. The results have been obtained for the general case when data is collected on more than one auxiliary character correlated with the character $y$. The case of a single auxiliary character becomes a special case of this. Throughout, samples have been drawn by the method of simple random sampling without replacement.

## The Estimator

Assume that information on $p$ auxiliary characters denoted by $x_1, \ldots, x_p$ could be collected. In two-phase sampling a first phase sample of size $n'$ is drawn from the population on which only the auxiliary characters are measured. Then a second phase sample of size $n$ $(<n')$ is drawn on which the character under study and the auxiliary characters are measured.

Let $\bar{x}'_1, \ldots, \bar{x}'_p$ denote the sample means of the characters $x_1, \ldots, x_p$ based on the first phase sample of size $n'$, and $\bar{y}, \bar{x}_1, \ldots, \bar{x}_p$ denote the sample means of the characters $y, x_1, \ldots, x_p$ based on the second phase sample of size $n$. Let

$$u_i = \frac{\bar{x}'_i}{\bar{x}_i}\ , \quad i=1, \ldots, p$$

and let $\underline{u}$ denote the column vector with elements $u_1, \ldots, u_p$ and $\underline{e}$ denote the eolumn vector of $p$ unit elements.

Whatever be the sample selected, let $\underline{u}$ assume values in a bounded closed convex subset, $D$, of the $p$-dimensional real space containing the point $\underline{e}$. Let $h\ (u_1, \ldots, u_p)=h\ (\underline{u})$ be a function of $\underline{u}$ such that it satisfies the following conditions.

1. In $D$, the function $h\ (\underline{u})$ is continuous and bounded.
2. The first and second order partial devivatives of $h\ (\underline{u})$ exist and are continuous and bounded in $D$.

Consider the class of estimators of the population mean $\bar{Y}$, defined by

$$\bar{y}_{mdh}=\bar{y}\ h\ (u_1, \ldots, u_p)=\bar{y}\ h(\underline{u}) \qquad \ldots(1)$$

Any parametric function $h\,(\underline{u})$ satisfying the above conditions can be considered as a estimator of $\bar{Y}$. The class of such estimators is very large and some members of this class for the case of a single auxiliary variable, have been studied separately by various authors. For the case of a single auxiliary character $x$, the class of estimators (1) reduce to

$$\bar{y}_{dh}=\bar{y}\ h(v) \qquad \ldots(2)$$

where $h(v)$ is a function of $v=\bar{x}'/\bar{x}$ such that $h\ (1)=1$. Taking $h\ (v)=a\ v+(1-a)v^2$, we get the estimator (20) of Gupta (1978). Taking $h\ (v)=v^{\alpha}$, we get the estimator considered by Srivastava (1970). Taking $h(v)=[1+\theta\ (v^{-1}-1)]^{-1}$, we get the two-phase sampling analogue of the estimator considered by Reddy (1974).

The bias and variance of the estimator, $\bar{y}_{mdh}$, exist since the number of possible samples is finite and we have assumed that the function is bounded. Expanding $h(u)$ about the point $\underline{u}=\underline{e}$ by a second order Taylor's series, we obtain

$$\bar{y}_{mnh}=\bar{y}\ \{h\,(\underline{e})+(\underline{u}-\underline{e})'\ h'(\underline{e})+\tfrac{1}{2}\ (\underline{u}-\underline{e})'\ h''\ (\underline{u}^*)\ (\underline{u}-\underline{e})\} \qquad \ldots(3)$$

where $\underline{u}^*=\underline{e}+\theta\,(\underline{u}-\underline{e}), 0<\theta<1$, and $h'(\underline{e})$ denotesthe column vector of first partial derivatives of $h(\underline{u})$ at the point $\underline{u}=\underline{e}$ and $h''(\underline{u}^*)$ denotes the matrix of second partial derivatives of $h(\underline{u})$ at the point $\underline{u}=\underline{u}^*$.

Defining

$$\eta=\frac{\bar{y}-\bar{Y}}{\bar{Y}},\ \varepsilon_i=\frac{\bar{x}_i-\bar{X}_i}{\bar{X}_i},\ \varepsilon_i'=\frac{\bar{x}_i'-\bar{X}_i}{\bar{X}_i},$$

and $\qquad \delta_i=u_i-1,\ i=1,\ \ldots,\ p,$

we have

$$\delta_i=\left(1+\varepsilon_i'\right)(1+\varepsilon_i)^{-1}-1$$
$$=\left(\varepsilon_i'-\varepsilon_i\right)+\left(\varepsilon_i^2-\varepsilon_i\,\varepsilon_i'\right)+\ldots.$$

Let $\qquad \delta=(\delta_1,\ \ldots,\ \delta_p)'$.

Substituting in (3) we have

$$\bar{y}_{mdh}=\bar{Y}(1+\eta)\,\{1+\underline{\delta}'\,h'\,(\underline{e})+\tfrac{1}{2}\,\underline{\delta}'\,h''(\underline{u}^*)\underline{\delta}\}$$
$$=\bar{Y}\{1+\eta+\underline{\delta}'\,h'\,(\underline{e})+\eta\underline{\delta}'\,h'\,(\underline{e})+\tfrac{1}{2}\underline{\delta}'\,h''(\underline{u}^*)\underline{\delta}+\tfrac{1}{2}\,\eta\underline{\delta}'\,h''(\underline{u}^*)\,\underline{\delta}\}\ldots(4)$$

Taking expectation in (4) and noting that

$$E(\eta)=E(\varepsilon_i)=E\left(\varepsilon_i'\right)=0,\ i=1,\ \ldots,\ p$$

and that the expectations of the second degree terms are of order $n^{-1}$, we obtain

$$E\,(\bar{y}_{mdh})=\bar{Y}+0\,(n^{-1}).$$

Thus the bias of the estimator $\bar{y}_{mdh}$, is of the order of $n^{-1}$ and hence its contribution to the mean square error will be of the order of $n^{-2}$. In what follows we assume that $n$ is large so that the bias is assumed negligible and the variance expressions are obtained up to terms of order $n^{-1}$, an approximation generally taken for ratio type of estimators.

To find the variance of $\bar{y}_{mdh}$, let

$$C_o^2\ =S_y^2\ /\bar{Y}^2 \text{ where } S_y^2\ =\frac{1}{N-1}\sum_{k=1}^{N}\ (y_k-\bar{Y})^2,$$

$$C_i^2\ =\ S_i^2\ /\bar{X}_i^2 \text{ where } S_i^2\ =\frac{1}{N-1}\ \sum_{K=1}^{N}\ (x_{ik}-\bar{X}_i)^2,$$

$$\bar{X}_i=\frac{1}{N}\ \sum_{K=1}^{N} x_{ik},$$

$y_k$ and $x_{ik}$ denoting the values of the characters $y$ and $x_i$ for the $kth$ unit of the population. Also let $\underline{b}'=(b_1,\ldots, b_p)$ be a row vector of $p$ elements and $\underline{A}=[a_{ij}]$ by a $p\times p$ matrix (assumed to be positive definite) where

$$b_i=\rho_{oi}\; C_oC_i \text{ and } a_{ij}=\rho_{ij}\,\frac{C_iC_j}{C_o^2}$$

$\rho_{ij}$ denoting the coerelation coefficient between $x_i$ and $x_j$, $i\neq j$, and $\rho_{oi}$, the correlation coefficient between $y$ and $x_i$.

Then, up to terms of order $n^{-1}$, we have

$$E(\eta\,\delta_i)=E\left(\eta\varepsilon_i' -\eta\,\varepsilon_i\right),\ i=1,\ldots,p,$$

$$E(\delta_i\delta_j)=E\left(\varepsilon_i'\;\varepsilon_j' - \varepsilon_i'\;\varepsilon_j-\varepsilon_i\,\varepsilon_j' +\epsilon_i\varepsilon_j\right),\qquad i,j=1,\ldots,p.$$

The following two cases will be considered seperately.

**Case 1**

When the second phase sample of size $n$ is a subsample of the first phase sample of size $n'$, and

**Case 2**

When the second phase sample of size $n$ is drawn independently of the first phase sample of size $n'$.

**Case 1**

In case I, we have

$$E(\eta^2)=\frac{f}{n}\,C_o^2\;,$$

$$E(\eta\,\delta_i)=-\left(\frac{f}{n}-\frac{f'}{n'}\right)\rho_{oi}\;C_oC_i,\quad i=1,\ldots,p$$

$$E(\delta_i\,\delta_j)=\left(\frac{f}{n}-\frac{f'}{n'}\right)\rho_{ij}C_iC_j,\qquad i,j=1,\ldots,p,$$

which gives

$$E(\eta\,\underline{\delta}')=-\left(\frac{f}{n}-\frac{f'}{n'}\right)C_o^2\;\underline{b}'$$

and

$$E(\underline{\delta}\;\underline{\delta}')=\left(\frac{f}{n}-\frac{f'}{n'}\right)C_o^2\;\underline{A}.$$

The variance of $\bar{y}_{mdn}$ up to the terms of order $n^{-1}$, is

$$V(\bar{y}_{mdh}) = E(\bar{y}_{mdh} - \bar{Y})^2$$

$$= \bar{Y}^2 E\{\eta^2 + 2\,\eta\underline{\delta}'\,h'(\underline{e}) + (h'(\underline{e}))/\underline{\delta}\underline{\delta}'\ h'\ (\underline{e})\} \quad \text{...(5)}$$

$$= \bar{Y}^2 C\bar{y}_o^2 \left\{\frac{f}{n} - 2\left(\frac{f}{n} - \frac{f'}{n'}\right)\underline{b}'h'(\underline{e}) + \left(\frac{f}{n} - \frac{f'}{n'}\right)(h'(\underline{e}))'\,\underline{A}\,h'(\underline{e})\right\} \quad \text{...(6)}$$

which is a function of $h'(\underline{e})$. Now $h(\underline{u})$ involves certain unknown parameters to be chosen so as to minimize the variance of $\bar{y}_{mdh}$, and $h'(\underline{e})$ also will involve these parameters. Clearly the variance at (6) is minimized for

$$h'(\underline{e}) = \underline{A}^{-1}\,\underline{b}$$

and the minimum variance is given by

$$V_{min}(\bar{y}_{mdh}) = \bar{Y}^2 C_o^2 \left\{\frac{f}{n} - \left(\frac{f}{n} - \frac{f'}{n'}\right)\underline{b}'\ \underline{A}^{-1}\ \underline{b}\right\}$$

$$= \frac{f}{n} S_y^2 (1 - R^2) + \frac{f'}{n'} S_y^2\ R^2,$$

where $R$ denotes the multiple correlation coefficient of $y$ on $x_1, \ldots, x_p$.

For the case of a single auxiliary variable $x$, the variance of $\bar{y}_{dh}$ defined at (2) is minimized for

$$h'(1) = \rho\frac{C_y}{C_x}$$

and the minimum variance is given by

$$V_{min}(\bar{y}_{dh}) = \frac{f}{n}\ S_y^2(1 - \rho^2) + \frac{f'}{n'}\ S_y^2\,\rho^2$$

which is the same as the variance of the linear regression estimator in two-phase sampling (see *e.g.* Cochran, 1963). Thus when the second phase sample is a subsample of the first phase sample, no estimator of the proposed class can have smaller variance up to the terms of order $n^{-1}$, than the linear regression estimator in two-phase sampling.

**Case 2 :**

In case II, we have

$$E(\eta^2) = \frac{f}{n}\ C_o^2$$

$$E(\eta\delta_i) = -\frac{f}{n}\rho_{oi}C_oC_i, \qquad i=1,\ldots,p$$

$$E(\delta_i\delta_j)=\left(\frac{f}{n}+\frac{f'}{n'}\right)\rho_{ij}C_iC_j, \qquad i,j=1,\cdot\,\cdot,p,$$

which gives

$$E(\eta\underline{\delta}') = -\frac{f}{n}\,C_o^2\,\underline{b}'$$

and

$$E(\underline{\delta}\,\underline{\delta}')=\left(\frac{f}{n}+\frac{f'}{n'}\right)C_o^2\,\underline{A}\ .$$

From (5), therefore, the variance of $\bar{y}_{mdh}$ up to terms of order $n^{-1}$, in this case is

$$V(\bar{y}_{mdh})=\bar{Y}^2C_o^2\left\{\frac{f}{n}-2\frac{f}{n}\underline{b}'h'(\underline{e})\right.$$
$$\left.+\left(\frac{f}{n}+\frac{f'}{n'}\right)(h'(\underline{e}))'\underline{A}\,h'(\underline{e})\right\}$$

which is minimized for

$$h'(\underline{e})=\frac{\dfrac{f}{n}}{\dfrac{f}{n}+\dfrac{f'}{n'}}\underline{A}^{-1}\,\underline{b}$$

and the minimnm variance is given by

$$V_{min}(\bar{y}_{mdh})=\frac{f}{n}\bar{Y}^2C_o^2\left(1-\frac{\dfrac{f}{n}}{\dfrac{f}{n}+\dfrac{f'}{n'}}\underline{b}'\,\underline{A}^{-1}\,\underline{b}\right)$$

$$=\frac{f}{n}S_y^2\left(1-\frac{\dfrac{f}{n}}{\dfrac{f}{n}+\dfrac{f'}{n'}}R^2\right)$$

For the case of a single auxiliary character $x$, the variance of $\bar{y}_{dh}$ is minimized for

$$h'(1)=\frac{\dfrac{f}{n}}{\dfrac{f}{n}+\dfrac{f'}{n'}}\rho\frac{C_y}{C_x},$$

and the minimum variance is given by

$$V_{min}\,(\bar{y}_{dh})=\frac{f}{n}S_y^2\left(1-\frac{\frac{f}{n}}{\frac{f}{n}+\frac{f'}{n'}}\rho^2\right) \qquad \ldots(7)$$

When the finite population correction terms $f$ and $f'$ are ignored, the miniumm variance at (7) is given by

$$V_{min}\,(\bar{y}_{dh})=\frac{1}{n}S_y^2\left(1-\frac{n'}{n+n'}\rho^2\right).$$

Up to the same order of approximation the variance of the convectional linear regression estimator in two-phase sampling, $\bar{y}_{dlr}$, in this case is given by (see *e.g.* Cochran, 1963)

$$V(\bar{y}_{dlr})=\frac{1}{n}S_y^2\left(1-\rho^2\right)+\frac{1}{n'}S_y^2\,\rho^2$$

$$=\frac{1}{n}S_y^2\left(1-\frac{n'-n}{n'}\rho^2\right).$$

And since $\frac{n'}{n+n'},>\frac{n'-n}{n'}$, the minimum variance of an estimator of the class (2) is always smaller than that of $\bar{y}_{dlr}$ in case II. This result was shown by Srivastava (1970) and Gupta (1978) for their estimators which are members of the class (2). Such a result seems to be surprising at first site since the variance of the estimators of the proposed class cannot be reduced below the variance of the linear regression eslimator for the uni-phase sampling scheme as well as for the two-phase sampling scheme in case I. This, however, happens because of the fact that the conventional linear regression estimator in two-phase sampling

$$\bar{y}_{dlr}=\bar{y}+b\,(\bar{x}'-\bar{x})$$

could be improved upon in case II of the two-phase sampling scheme by replacing $\bar{x}'$ in the estimator with a better estimator of $\bar{X}$ based on $x$-values in both the samples.

## ACKNOWLEDGEMENT

The author is thankful to a referee for helpful suggestions to improve the presentation of the paper,

## References

Cochran, W.G. (1963) : *Sampling Techniques*, (2nd Ed.) John Wiley and Sons, New York.

Gupta, P.C. (1978) : On some quadratic and higher degree ratio and product estimators. *Jour. Ind. Soc. Agri. Stat.* 30, 71-80.

Ready, V. N. (1973) : On ratio and product methods of estimation. *Sankhyā*, Ser. B. 35, 307-316.

Reddy, V.N. (1974) : On a transformed ratio method of estimation. *Sankhyā*, Ser. C. 36, 59-70.

Shukla, G.K. (1966) : An alternative multivariate ratio estimate for finite population. *Cal. Stat. Ass. Bull.* 15, 127-134.

Srivastava, S.K. (1967) : An estimator using auxillary information in sample surveys. *Cal. Stat. Ass. Bull.* 16, 121-132.

Srivastava, S.K. (1970) : A two-phase sampling estimator in sample surveys. *Aust. Jour. Stat.* 12. 23—27.

Srivastava, S.K. (1971) : A generalized estimator for the mean of a finite population using multi-auxiliary information. *Jour. Amer. Stat. Ass.* 66, 404-407.

Walsh, J.E. (1970) : Generalization of ratio estimate for population total. *Sankhya.* Ser. A. 32, 99-106.