STATISTICA, anno LXXII, n. 1, 2012

ESTIMATION OF FINITE POPULATION MEAN IN TWO-PHASE SAMPLING WITH KNOWN COEFFICIENT OF VARIATION OF AN AUXILIARY CHARACTER

H.P. Singh, R. Tailor, R. Tailor

1. INTRODUCTION

In sample surveys we often use an auxiliary variable x to construct more précised estimates of the population mean $\bar{Y}$ or total $Y(=N\bar{Y})$ of the study variable y where N (finite) is the population size. A large number of estimators of the population mean $\bar{Y}$ using information on the population mean $\bar{X}$ of the auxiliary variable x have been proposed by various authors for instance see (Singh, 1986), (Singh and Upadhyaya, 1986), (Singh, 2003), (Singh *et al.*, 2004) and the references cited therein. Using information on the population mean $\bar{X}$ and the coefficient of variation C_x of the auxiliary variable x motivated by (Singh and Ruiz Espejo, 2003) and Singh and Tailor, 2005) suggested the following ratio-cum-product estimator for the population mean $\bar{Y}$ as

$$t=\bar{y}\left[\alpha\left(\frac{\bar{X}+C_x}{\bar{x}+C_x}\right)+(1-\alpha)\left(\frac{\bar{x}+C_x}{\bar{X}+C_x}\right)\right], \tag{1}$$

where $C_x=\dfrac{S_x}{\bar{X}}$ is the coefficient of variation of x with $S_x=\sqrt{\sum_{i=1}^{N}(x_i-\bar{X})^2\big/(N-1)}$

and α is suitably chosen scalar; $\bar{y}=(1/n)\sum_{i=1}^{n}y_i$ and $\bar{x}=(1/n)\sum_{i=1}^{n}x_i$ are sample means of y and x respectively based on n observations drawn by using simple random sampling without replacement (SRSWOR) from the population of size N.

In many situations of practical importance it may happen that the population mean $\bar{X}$ of the auxiliary variable x is not known before start of the survey. It is well known that if the necessary auxiliary information is not readily available for the population before sampling it might pay to collect such information for a large preliminary sample and then collect more precise information for the variable under study on a final or second phase sample. This technique known as two phase sampling is very much in use in the practice see (Adhvaryu and Gupta, 1983, p. 223). More precisely the double sampling model is as follows:

(i) The first phase sample 's' of fixed size $n_1\,(n_1 < N)$ is drawn to observe the auxiliary variable x in order to furnish an estimate of the population mean $\bar{X}$

(ii) The second phase sample 'r' of fixed size $n\;(n < n_1)$ is drawn to observe the study variable y only in either of the following manners,

Case I: "as a sub sample from the first phase sample"

Case II: "independently to the first phase sample"

When the population mean $\bar{X}$ of x is not known using the two-phase sampling procedure as described above we define the classical ratio and product estimators for the population mean $\bar{Y}$ respectively as

$$t_{Rd} = \bar{y}\frac{\bar{x}_1}{\bar{x}}, \tag{2}$$

and

$$t_{Pd} = \bar{y}\frac{\bar{x}}{\bar{x}_1}, \tag{3}$$

where $\bar{y}$ and $\bar{x}$ are the sample means based on second phase sample of size n and $\bar{x}_1 = (1/n_1)\sum_{i=1}^{n_1} x_i$ is the first phase sample mean of x based on n_1 observations.

In this paper we have suggested the two-phase (or double) sampling version of the estimator due to (Singh and Tailor, 2005) and its properties are studied under large sample approximation. An empirical study is carried out to judge the merits of the proposed estimator over other competitors.

We shall use the SRSWOR sampling scheme through out the paper.

2. SUGGESTED CLASS OF RATIO-CUM-PRODUCT ESTIMATORS

Replacing $\bar{X}$ by $\bar{x}_1$ in (1) we define the double sampling version of the (Singh and Tailor, 2005) estimator t as

$$t_d = \bar{y}\left[\alpha\left(\frac{\bar{x}_1 + C_x}{\bar{x} + C_x}\right) + (1-\alpha)\left(\frac{\bar{x} + C_x}{\bar{x}_1 + C_x}\right)\right], \tag{4}$$

where notations have the same meaning as described in Section 1. For $\alpha = 1$ t_d reduces to the (Kawathekar and Ajagaonkar's, 1984) estimator

$$t_d^{(1)} = \bar{y}\left(\frac{\bar{x}_1 + C_x}{\bar{x} + C_x}\right). \tag{5}$$

while for $\alpha = 0$ it boils down to the product-type estimator

$$t_d^{(2)} = \bar{y}\left(\frac{\bar{x} + C_x}{\bar{x}_1 + C_x}\right). \tag{6}$$

To obtain the variance of t_d we write

$\bar{y} = \bar{Y}(1+e_0)$, $\bar{x} = \bar{X}(1+e_1)$, and $\bar{x}_1 = \bar{X}(1+e_1^{'})$,

$\mathrm{E}(e_0) = \mathrm{E}(e_1) = \mathrm{E}(e_1^{'}) = 0$.

in both the cases I and II. The other expected values ignoring finite population correction (fpc) terms in case I and case II are given by,

Case I: $\mathrm{E}(e_0^2) = C_y^2/n$, $\mathrm{E}(e_1^2) = C_x^2/n$, $\mathrm{E}(e_1^{'2}) = C_x^2/n_1$,
$\mathrm{E}(e_0 e_1) = CC_x^2/n$, $\mathrm{E}(e_0 e_1^{'}) = CC_x^2/n_1$, $\mathrm{E}(e_1 e_1^{'}) = C_x^2/n_1$.

Case II: $\mathrm{E}(e_0^2) = C_y^2/n$, $\mathrm{E}(e_1^2) = C_x^2/n$, $\mathrm{E}(e_1^{'2}) = C_x^2/n_1$,
$\mathrm{E}(e_0 e_1) = CC_x^2/n$, $\mathrm{E}(e_0 e_1^{'}) = \mathrm{E}(e_1 e_1^{'}) = 0$,

where $C_y = S_y/\bar{Y}$, $C_x = S_x/\bar{X}$, $C = \rho C_y/C_x$, $\rho = S_{xy}/S_x S_y$,

$$S_y^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2/(N-1),\ S_{xy} = \sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})/(N-1).$$

Expressing (4) in terms of $e's$ we have

$$t_d = \bar{Y}(1+e_0)\left[\alpha\left(\frac{\bar{X}(1+e_1^{'}) + C_x}{\bar{X}(1+e_1) + C_x}\right) + (1-\alpha)\left(\frac{\bar{X}(1+e_1) + C_x}{\bar{X}(1+e_1^{'}) + C_x}\right)\right],$$

$$= \bar{Y}(1+e_0)[\alpha(1+\theta e_1^{'})(1+\theta e_1)^{-1} + (1-\alpha)(1+\theta e_1)(1+\theta e_1^{'})^{-1}],$$

where $\theta = \bar{X}/(\bar{X} + C_x)$.

Expanding, multiplying and neglecting terms of $e's$ having power greater than 'unity' in the right hand side of the above expression we have

$$(t_d - \bar{Y}) \cong \bar{Y}[e_0 + (1-2\alpha)\theta(e_1 - e_1^{'})].$$

Squaring both the sides of the above expression we have

$$(t_d - \bar{Y})^2 = \bar{Y}^2[e_0^2 + (1-2\alpha)^2\theta^2(e_1^2 + e_1^{'2} - 2e_1e_1^{'}) + 2(1-2\alpha)\theta(e_0e_1 - e_0e_1^{'})]. \quad (7)$$

Taking expectation of both the sides of (7) and using the results in case I and case II we get the variance of t_d to the first degree of approximation in case I and case II respectively as

$$V(t_d)_I = \bar{Y}^2\left[\left(\frac{1}{n}\right)C_y^2 + \theta(1-2\alpha)\left(\frac{1}{n} - \frac{1}{n_1}\right)\{(1-2\alpha)\theta + 2C\}\right], \quad (8)$$

$$V(t_d)_{II} = \bar{Y}^2\left[\left(\frac{1}{n}\right)C_y^2 + \theta(1-2\alpha)\left\{\theta(1-2\alpha)\left(\frac{1}{n} + \frac{1}{n_1}\right) + \frac{2C}{n}\right\}\right]. \quad (9)$$

3. OPTIMUM CHOICE OF THE SCALAR 'α'

Minimization of $V(t_d)_I$ and $V(t_d)_{II}$ with respect to α yields the optimum values of α in case I and II respectively as

$$\alpha = \frac{(\theta + C)}{2\theta} = \alpha_{(opt)I}, \quad (10)$$

and

$$\alpha = \frac{1}{2}\left[1 + \frac{n_1}{n + n_1}\frac{C}{\theta}\right] = \alpha_{(opt)II}. \quad (11)$$

Substitutions of $\alpha_{(opt)I}$ and $\alpha_{(opt)II}$ in place of α in (4) respectively yield the optimum estimators in the case I and case II as

$$t_{d(opt)I} = \frac{\bar{y}}{2\theta}\left[(\theta + C)\left(\frac{\bar{x}_1 + C_x}{\bar{x} + C_x}\right) + (\theta - C)\left(\frac{\bar{x} + C_x}{\bar{x}_1 + C_x}\right)\right], \quad (12)$$

$$t_{d(opt)II} = \frac{\bar{y}}{2}\left[\left\{1 + \frac{C}{\theta}\frac{n_1}{(n + n_1)}\right\}\left(\frac{\bar{x}_1 + C_x}{\bar{x} + C_x}\right) + \left\{1 - \frac{C}{\theta}\frac{n_1}{(n + n_1)}\right\}\left(\frac{\bar{x} + C_x}{\bar{x}_1 + C_x}\right)\right]. \quad (13)$$

It is observed from (12) and (13) that the optimum estimators $t_{d(opt)I}$ and $t_{d(opt)II}$ depend upon the unknown population parameters such as $\bar{X}$, ρ, C_y, C

and θ which lacks the practical utility of these estimators. For the application of such estimators one has to use the close guessed values of these parameters obtained from the past studies or with the familiarly of the experimental material. (Das and Tripathi, 1978) have illustrated that the estimators based on guessed values are better than the conventional estimators. On the other hand if the guessed values close enough are not available then it is worth advisable to replace the unknown population parameters by their consistent estimators. Thus replacing $\bar{X}$, ρ and C_y by their consistent estimates $\bar{x}_1$, $\hat{\rho} = s_{xy}/s_x s_y$ and $\hat{C}_y = s_y/\bar{y}$ respectively in (12) and (13) we get the estimator based on 'estimated optimum' as

$$\hat{t}_{d(opt)I} = \frac{\bar{y}}{2\hat{\theta}}\left[(\hat{\theta}+\hat{C})\left(\frac{\bar{x}_1+C_x}{\bar{x}+C_x}\right)+(\hat{\theta}-\hat{C})\left(\frac{\bar{x}+C_x}{\bar{x}_1+C_x}\right)\right], \tag{14}$$

and

$$\hat{t}_{d(opt)II} = \frac{\bar{y}}{2}\left[\left\{1+\frac{\hat{C}}{\hat{\theta}}\frac{n_1}{(n+n_1)}\right\}\left(\frac{\bar{x}_1+C_x}{\bar{x}+C_x}\right)+\left\{1-\frac{\hat{C}}{\hat{\theta}}\frac{n_1}{(n+n_1)}\right\}\left(\frac{\bar{x}+C_x}{\bar{x}_1+C_x}\right)\right], \tag{15}$$

where $\hat{\theta} = \bar{x}_1/(\bar{x}_1+C_x)$ and $\hat{C} = s_{xy}/(\bar{y}s_x C_x)$ are the consistent estimator of θ and C respectively.

Using the standard technique it can be shown to the first degree of approximation that

$$\min.V(t_d)_I = V(t_{d(opt)I}) = V(\hat{t}_{d(opt)I}) = \frac{S_y^2}{n}\left[1-\frac{(n_1-n)}{n_1}\rho^2\right], \tag{16}$$

$$\min.V(t_d)_{II} = V(t_{d(opt)II}) = V(\hat{t}_{d(opt)II}) = \frac{S_y^2}{n}\left[1-\frac{n_1}{(n\ +n_1)}\rho^2\right], \tag{17}$$

where $\min.V(t_d)_I$ and $\min.V(t_d)_{II}$ stand for the minimum variance of the proposed class of estimators in case I and case II respectively.

It can be easily seen from (16) and (17) that the difference $V(t_{d(opt)I}\ or\ \hat{t}_{d(opt)I}) - V(t_{d(opt)II}\ or\ \hat{t}_{d(opt)II})$ is always positive which follows that the optimum estimator (or estimator based on estimated 'optimum') in case II is more efficient than the optimum estimator (or estimator based on estimated 'optimum') in case I.

4. EFFICIENCY COMPARISON OF t_d WITH OTHER ESTIMATORS

It is well known under SRSWOR sampling scheme (ignoring fpc) that

$$V(\bar{y}) = \frac{S_y^2}{n} \tag{18}$$

This can be expressed as

$$V(\bar{y}) = \frac{\bar{Y}C_y^2}{n}. \tag{19}$$

For the purpose of comparison we write the variance expression of the estimators $t_{\mathrm{R}d}$, t_{Pd}, $t_d^{(1)}$ and $t_d^{(2)}$ to the first degree of approximation in case I and case II respectively as:

$$V(t_{\mathrm{R}d})_I = \bar{Y}^2\left[\left(\frac{1}{n}\right)C_y^2 + \left(\frac{1}{n} - \frac{1}{n_1}\right)C_x^2(1-2C)\right], \tag{20}$$

$$V(t_{Pd})_I = \bar{Y}^2\left[\left(\frac{1}{n}\right)C_y^2 + \left(\frac{1}{n} - \frac{1}{n_1}\right)C_x^2(1+2C)\right], \tag{21}$$

$$V(t_d^{(1)})_I = \bar{Y}^2\left[\left(\frac{1}{n}\right)C_y^2 + \theta\left(\frac{1}{n} - \frac{1}{n_1}\right)C_x^2(\theta-2C)\right], \tag{22}$$

$$V(t_d^{(2)})_I = \bar{Y}^2\left[\left(\frac{1}{n}\right)C_y^2 + \theta\left(\frac{1}{n} - \frac{1}{n_1}\right)C_x^2(\theta+2C)\right], \tag{23}$$

$$V(t_{\mathrm{R}d})_{II} = \bar{Y}^2\left[\left(\frac{1}{n}\right)C_y^2 + C_x^2\left\{\left(\frac{1}{n} - \frac{1}{n_1}\right) - \frac{2C}{n}\right\}\right], \tag{24}$$

$$V(t_{Pd})_{II} = \bar{Y}^2\left[\left(\frac{1}{n}\right)C_y^2 + C_x^2\left\{\left(\frac{1}{n} + \frac{1}{n_1}\right) + \frac{2C}{n}\right\}\right], \tag{25}$$

$$V(t_d^{(1)})_{II} = \bar{Y}^2\left[\left(\frac{1}{n}\right)C_y^2 + \theta C_x^2\left\{\left(\frac{1}{n} + \frac{1}{n_1}\right)\theta - \frac{2C}{n}\right\}\right], \tag{26}$$

$$V(t_d^{(2)})_{II} = \bar{Y}^2\left[\left(\frac{1}{n}\right)C_y^2 + \theta C_x^2\left\{\left(\frac{1}{n} + \frac{1}{n_1}\right)\theta + \frac{2C}{n}\right\}\right]. \tag{27}$$

4.1 *Efficiency Comparison of t_d With Other Estimators*

From (8) (19) (20) (21) (22) and (23) we note that the suggested estimator t_d is more efficient (under case I) than:

(i) sample mean $\bar{y}$ if

$$\begin{cases} \textit{either} & \frac{1}{2} < \alpha < \left(\frac{1}{2} + \frac{C}{\theta}\right) \\ \textit{or} & \left(\frac{1}{2} + \frac{C}{\theta}\right) < \alpha < \frac{1}{2} \end{cases}$$

or equivalently

$$\min.\left\{\frac{1}{2}, \left(\frac{1}{2} + \frac{C}{\theta}\right)\right\} < \alpha < \max.\left\{\frac{1}{2}, \left(\frac{1}{2} + \frac{C}{\theta}\right)\right\}, \tag{28}$$

(ii) usual two-phase sampling ratio estimator t_{Rd} if

$$\begin{cases} \textit{either} & \frac{(1+\theta)}{2\theta} < \alpha < \left(\frac{(\theta + 2C - 1)}{2\theta}\right) \\ \textit{or} & \left(\frac{(\theta + 2C - 1)}{2\theta}\right) < \alpha < \frac{(1+\theta)}{2\theta} \end{cases}$$

or equivalently

$$\min.\left\{\frac{(1+\theta)}{2\theta}, \frac{(\theta + 2C - 1)}{2\theta}\right\} < \alpha < \max.\left\{\frac{(1+\theta)}{2\theta}, \frac{(\theta + 2C - 1)}{2\theta}\right\}, \tag{29}$$

(iii) usual two-phase sampling product estimator t_{Pd} if

$$\begin{cases} \textit{either} & \left(\frac{(\theta + 2C + 1)}{2\theta}\right) < \alpha < \frac{(\theta - 1)}{2\theta} \\ \textit{or} & \frac{(\theta - 1)}{2\theta} < \alpha < \left(\frac{(\theta + 2C + 1)}{2\theta}\right) \end{cases}$$

or equivalently

$$\min.\left\{\frac{(\theta-1)}{2\theta}, \frac{(\theta + 2C - 1)}{2\theta}\right\} < \alpha < \max.\left\{\frac{(\theta-1)}{2\theta}, \frac{(\theta + 2C + 1)}{2\theta}\right\}. \tag{30}$$

(iv) Kawathekar and Ajagaonkar's, 1984 estimator $t_d^{(1)}$ if

$$\left\{\begin{array}{ll} \textit{either} & \frac{C}{\theta} < \alpha < 1 \\ \textit{or} & 1 < \alpha < \frac{C}{\theta} \end{array}\right.$$

or equivalently

$$\min.\left(1\ ,\frac{C}{\theta}\right) < \alpha < \max.\left(1\ ,\frac{C}{\theta}\right), \tag{31}$$

(v) $t_d^{(2)}$ if

$$\left\{\begin{array}{ll} \textit{either} & 0 < \alpha < \left(1+\frac{C}{\theta}\right) \\ \textit{or} & \left(1+\frac{C}{\theta}\right) < \alpha < 0 \end{array}\right.$$

or equivalently

$$\min.\left\{0\ ,\left(1+\frac{C}{\theta}\right)\right\} < \alpha < \max.\left\{0\ ,\left(1+\frac{C}{\theta}\right)\right\}. \tag{32}$$

We note from (9) (19) (24) (25) (26) and (27) that the proposed estimator t_d dominates over the estimator (in case II),

(i) sample mean $\bar{y}$ if

$$\left\{\begin{array}{ll} \textit{either} & \frac{1}{2} < \alpha < \frac{1}{2}\left[1+\frac{n_1}{(n+n_1)}\frac{C}{\theta}\right] \\ \textit{or} & \frac{1}{2}\left[1+\frac{n_1}{(n+n_1)}\frac{C}{\theta}\right] < \alpha < \frac{1}{2} \end{array}\right.$$

or equivalently

$$\min.\left\{\frac{1}{2},\frac{1}{2}\left(1+\frac{n_1C}{(n+n_1)\theta}\right)\right\} < \alpha < \max.\left\{\frac{1}{2}\ ,\frac{1}{2}\left(1+\frac{n_1C}{(n+n_1)\theta}\right)\right\}, \tag{33}$$

(ii) two-phase sampling ratio estimator t_{Rd} if

$$\left\{\begin{array}{ll} \textit{either} & \dfrac{(1+\theta)}{2\theta} < \alpha < \left[\dfrac{(\theta-1)}{2\theta} + \dfrac{n_1}{(n+n_1)}\dfrac{C}{\theta}\right] \\ \textit{or} & \left[\dfrac{(\theta-1)}{2\theta} + \dfrac{n_1}{(n+n_1)}\dfrac{C}{\theta}\right] < \alpha < \dfrac{(1+\theta)}{2\theta} \end{array}\right.$$

or equivalently

$$\min.\left\{\frac{(1+\theta)}{2\theta}, \left(\frac{(\theta-1)}{2\theta} + \frac{n_1 C}{(n+n_1)\theta}\right)\right\} < \alpha < \max.\left\{\frac{(1+\theta)}{2\theta},\ \left(\frac{\theta-1}{2\theta} + \frac{n_1 C}{(n+n_1\theta)}\right)\right\}, \tag{34}$$

(iii) two-phase sampling product estimator t_{Pd} if

$$\left\{\begin{array}{ll} \textit{either} & \left[\dfrac{(1+\theta)}{2\theta} + \dfrac{n_1}{(n+n_1)}\dfrac{C}{\theta}\right] < \alpha < \dfrac{(\theta-1)}{2\theta} \\ \textit{or} & \dfrac{(\theta-1)}{2\theta} < \alpha < \left[\dfrac{(1+\theta)}{2\theta} + \dfrac{n_1}{(n+n_1)}\dfrac{C}{\theta}\right] \end{array}\right.$$

or equivalently

$$\min.\left\{\frac{(\theta-1)}{2\theta}, \left(\frac{(1+\theta)}{2\theta} + \frac{n_1 C}{(n+n_1)\theta}\right)\right\} < \alpha < \max.\left\{\frac{(\theta-1)}{2\theta},\ \left(\frac{1+\theta}{2\theta} + \frac{n_1 C}{(n+n_1)\theta}\right)\right\}, \tag{35}$$

(iv) Kawathekar and Ajagaonkar's, 1984 estimator $t_d^{(1)}$ if

$$\left\{\begin{array}{ll} \textit{either} & \dfrac{n_1}{(n+n_1)}\dfrac{C}{\theta} < \alpha < 1 \\ \textit{or} & 1 < \alpha < \dfrac{n_1}{(n+n_1)}\dfrac{C}{\theta} \end{array}\right.$$

or equivalently

$$\min.\left\{1, \frac{n_1 C}{(n+n_1)\theta}\right\} < \alpha < \max.\left\{1, \frac{n_1 C}{(n+n_1)\theta}\right\}, \tag{36}$$

(v) $t_d^{(2)}$ if

$$\begin{cases} \textit{either} \qquad 0 < \alpha < \left[1 + \frac{n_1}{(n_1+n)}\frac{C}{\theta}\right] \\ \textit{or} \qquad \left[1 + \frac{n_1}{(n_1+n)}\frac{C}{\theta}\right] < \alpha < 0 \end{cases}$$

or equivalently

$$\min.\left\{0, \frac{n_1 C}{(n+n_1)\theta}\right\} < \alpha < \max.\left\{0, \frac{n_1 C}{(n+n_1)\theta}\right\}. \tag{37}$$

5. COMPARISON WITH SINGLE – PHASE SAMPLING

Case I

For case I we consider the cost function

$$C_0 = C_1 n_1 + C_2 n\,, \tag{38}$$

where C_0 is the total cost C_1 and C_2 are cost per unit of the first and the second phase samples.

The variance of t_d at (8) is written as,

$$V(t_d)_I = \frac{V_1}{n} + \frac{(V_2 - V_1)}{n_1} \tag{39}$$

where

$$V_1 = \bar{Y}^2[C_y^2 + \theta(1-2\alpha)C_x^2\{\theta(1-2\alpha) + 2C\}]\,,$$

$$V_2 = \bar{Y}^2 C_y^2\,.$$

The optimum values of n and n_1 for fixed cost C_0 which minimizes the variance of t_d are given by

$$\left.\begin{aligned} n_{opt} &= \frac{C_0\sqrt{V_1/C_2}}{\sqrt{C_2V_1} + \sqrt{C_1(V_2-V_1)}} \\ n_{1opt} &= \frac{C_0\sqrt{(V_2-V_1)/C_1}}{\sqrt{C_2V_1} + \sqrt{C_1(V_2-V_1)}} \end{aligned}\right\}. \tag{40}$$

The variance of t_d corresponding to optimal two-phase sampling strategy is

$$V_{opt}(t_d)_I = \frac{1}{C_0}\left[\sqrt{C_2 V_1} + \sqrt{C_1(V_2 - V_1)}\right]^2. \tag{41}$$

If all resources are devoted to a single sample on which only study variate is measured then

$$C_0 = nC_2 \quad \text{and}$$

$$V(\bar{y})_{opt} = \frac{C_2}{C_0}\bar{Y}^2 C_y^2 = \frac{C_2}{C_0}V_2. \tag{42}$$

Thus the two-phase sampling estimator t_d would be beneficial so long as $V(t_d)_I < V(\bar{y})_{opt}$

$$\text{i.e. } \frac{C_1}{C_2} < \left[\frac{\sqrt{V_2} - \sqrt{V_1}}{\sqrt{(V_2 - V_1)}}\right]^2. \tag{43}$$

In particular if we use $t_{d(opt)I}\ (or\ \hat{t}_{d(opt)I})$ then (43) reduces to

$$\frac{C_1}{C_2} < \frac{\left[1 - \sqrt{(1-\rho^2)}\right]^2}{\rho^2}. \tag{44}$$

Case II

We note from (16) that the variance of the optimum estimators $t_{d(opt)II}$ (or the variance of the estimators $\hat{t}_{d(opt)II}$ based on the estimated 'optimum' values) is

$$V(\hat{t}_{d(opt)II}) = \frac{S_y^2}{n}\left[1 - \frac{n_1}{(n+n_1)}\rho^2\right] = V(t_{d(opt)II}) \tag{45}$$

Under this case let the auxiliary variate x be measured on $(n+n_1)$ units and the study variate y on n units.

We consider the cost function of the following form

$$C_0^* = C_1^*(n+n_1) + nC_2^* \tag{46}$$

[for instance see Srivastava, 1970)] where C_1^* and C_2^* are cost per unit of observing x and y respectively. The optimal selection of sample sizes n and n_1 subject to the condition that (45) is minimum under the fixed cost C_0^* is such that

$$\frac{n}{n+n_1}=\left[\frac{C_1^*}{C_2^*}\frac{(1-\rho^2)}{\rho^2}\right]^{1/2}. \tag{47}$$

This equation with (47) determines optimum values of n and n_1 and hence of $(n+n_1)$.

The resulting variance corresponding to these optimum values of n and n_1 is given by

$$V(\hat{t}_{d(opt)II})_{opt}=V(\hat{t}_{d(opt)I})=\frac{S_y^2}{C_0^*}\left[\sqrt{(1-\rho^2)C_2^*}+\rho\sqrt{C_1^*}\right]^2. \tag{48}$$

If the sample mean $\overline{y}$ is to be used then its variance corresponding to optimal sampling strategy is

$$V(\overline{y})_{opt}=\frac{C_2^*}{C_0^*}S_y^2. \tag{49}$$

From (48) and (49) it is obtained that the two phase sampling estimator $(t_{d(opt)II})$ or $(\hat{t}_{d(opt)II}$) yields fewer variance than that of $\overline{y}$ for the same fixed cost if

$$\rho^2>\frac{4C_2^*C_1^*}{(C_1^*+C_2^*)^2}. \tag{50}$$

6. EMPIRICAL STUDY

To examine the merits of the suggested estimator we have considered five natural population data sets. The description of the populations are given below.

<u>Population – I</u>: Murthy (1967 p. 228)

N= 80 y: Output
n' = 30` x: Fixed Capital

$\overline{Y}=51.8264$, $\overline{X}=11.2646$, $C_y=0.3542$,
$C_x=0.7507$, $\rho=0.9413$, $C=0.4441$ $\theta=0.9375$.
n=10.

<u>Population – II</u>: Murthy (1967 p. 228)

N= 80 y: Output
n' = 30` x: Number of Workers

$\bar{Y} = 51.8264$, $\bar{X} = 2.8513$, $C_y = 0.3542$,

$C_x = 0.9484$, $\rho = 0.9150$, $C = 0.3417$ $\theta = 0.7504$.

n=10.

Population – III, Das (1988)

N= 278	y: Number of agricultural laborers for 1971
n' = 50	x: Number of agricultural laborers for 1961

$\bar{Y} = 39.0680$, $\bar{X} = 25.1110$, $C_y = 1.4451$,

$C_x = 1.6198$, $\rho = 0.7213$, $C = 0.6435$ $\theta = 0.9394$.

n=25.

Population – IV, Steel and Torrie (1960 p. 282)

N= 30	y: Log of leaf burn in secs
n' = 12	x: Clorine percentage

$\bar{Y} = 0.6860$, $\bar{X} = 0.8077$, $C_y = 0.700123$, $C_x = 0.7493$,

$\rho = -0.4996$, $C = -0.3202$, $\theta = 0.5188$

n=6.

Population – V, Maddala (1977)

N= 16	y: Consumption per capita
n' = 8	x: Deflated prices of veal

$\bar{Y} = 7.6375$, $\bar{X} = 75.4313$, $C_y = 0.2278$, $C_x = 0.0986$,

$\rho = -0.6823$, $C = -1.5761$, $\theta = 0.9987$

n=4.

We have computed the ranges of α for which the proposed estimator t_d is better than $\bar{y}$, t_{Rd}, t_{Pd}, $t_d^{(1)}$ and $t_d^{(2)}$. The optimum value of α and the common range of α in which t_d is better than $\bar{y}$, t_{Rd}, t_{Pd}, $t_d^{(1)}$ and $t_d^{(2)}$ have also been computed. Findings are displayed in Table 1. The percent relative efficiencies (PREs) of $t_{d(opt)I}$ $(or\ \hat{t}_{d(opt)I})$ and $t_{d(opt)II}$ $(or\ \hat{t}_{d(opt)II})$ with respect to $\bar{y}$, t_{Rd}, t_{Pd}, $t_d^{(1)}$ and $t_d^{(2)}$ have been computed and results are shown in Table 2.

TABLE 1

Range of α in which t_d is better than $\overline{y}$, t_{Rd}, t_{Pd}, $t_d^{(1)}$ and $t_d^{(2)}$ in case I and II

Population	case	Range of α in which t_d is better than the estimator					Optimum value of α	Common range of α in which t_d is better than $\overline{y}$ t_{Rd} t_{Pd} $t_d^{(1)}$ and $t_d^{(2)}$
		$\overline{y}$	t_{Rd}	t_{Pd}	$t_d^{(1)}$	$t_d^{(2)}$	α_0	
I	I	(0.50, 0.9737)	(0.4404, 1.0333)	(- 0.0333, 1.507)	(0.4737, 1.00)	(0.00, 1.4737)	0.7368479	(0.50, 0.9737)
$n'=30, n=10$	II	(0.50, 0.6776)	(0.3219, 1.0333)	(-0.0333, 1.3886)	(0.3553, 1.00)	(0.00, 1.3553)	0.6776359	(0.50, 0.6776360)
II	I	(0.50, 0.9554)	(0.2891, 1.0166)	(-0.1663, 1.6217)	(0.4554, 1.00)	(0.00, 1.4554)	0.7276785	(0.50, 0.9554)
$n'=30, n=10$	II	(0.50, 0.67075)	(0.1752, 101663)	(-0.1663, 1.5078)	(0.3415, 1.00)	(0.00, 1.3415)	0.6707589	(0.50, 0.6707590)
III	I	(0.50, 1.1850)	(0.6528, 1.0323)	(-0.0323, 1.7173)	(0.6850, 1.00)	(0.00, 1.6850)	0.8425058	(0.6850, 1.00)
$n'=50, n=25$	II	(0.50, 0.7283)	(0.4244, 1.0322)	(-0.0322, 1.4889)	(0.4567, 1.00)	(0.00, 1.4567)	0.7283372	(0.50, 1.00)
IV	I	(-0.1172, 0.50)	(-1.0810, 1.4638)	(-0.4638, 0.8466)	(-0.6172, 1.00)	(0.00, 0.3828)	0.1914032	(0.00, 0.50)
$n'=12, n=6$	II	(0.2942, 0.50)	(-0.8752, 1.4638)	(-0.4638, 1.0523)	(-0.4115, 1.00)	(0.00, 0.5885)	0.2942688	(0.29427, 0.5885)
V	I	(-1.0782, 0.50)	(-1.5788, 1.0007)	(-0.5775, 0.0007)	(-1.5782, 1.00)	(-0.5782, 0.00)	0.2890758	(-0.5775, 0.00)
$n'=8, n=4$	II	(-0.02605, 0.50)	(-1.0527, 1.0006)	(-0.0514, 0.0006)	(-1.0521, 1.00)	(-0.0521, 0.00)	0.0260505	(-0.0514, 0.00)

TABLE 2

Percent relative efficiencies (PREs) of $t_{d(opt)I}$ (or $\hat{t}_{d(opt)I}$) and $t_{d(opt)II}$ (or $\hat{t}_{d(opt)II}$) with respect to $\overline{y}$, t_{Rd}, t_{Pd}, $t_d^{(1)}$ and $t_d^{(2)}$

Population	Estimator	Percent relative efficiencies (PREs) of (.)				
		w.r.t. $\overline{y}$	w.r.t. t_{Rd}	w.r.t. t_{Pd}	w.r.t. $t_d^{(1)}$	w.r.t. $t_d^{(2)}$
I	$t_{d(opt)I}$ (or $\hat{t}_{d(opt)I}$)	244.32	326.15	1625.82	278.15	1496.65
$n'=30, n=10$	$t_{d(opt)II}$ (or $\hat{t}_{d(opt)II}$)	298.10	894.16	3272.85	752.33	2982.41
II	$t_{d(opt)I}$ (or $\hat{t}_{d(opt)I}$)	226.32	568.79	2047.30	280.71	1390.17
$n'=30, n=10$	$t_{d(opt)II}$ (or $\hat{t}_{d(opt)II}$)	268.76	1521.09	4154.73	727.31	2703.59
III	$t_{d(opt)I}$ (or $\hat{t}_{d(opt)I}$)	135.16	110.79	329.34	107.43	312.74
$n'=50, n=25$	$t_{d(opt)II}$ (or $\hat{t}_{d(opt)II}$)	153.10	194.07	689.19	175.16	640.28
IV	$t_{d(opt)I}$ (or $\hat{t}_{d(opt)I}$)	114.26	221.60	137.79	153.61	110.13
$n'=12, n=6$	$t_{d(opt)II}$ (or $\hat{t}_{d(opt)II}$)	119.96	414.06	238.07	221.08	129.78
V	$t_{d(opt)I}$ (or $\hat{t}_{d(opt)I}$)	130.34	181.03	104.06	180.95	104.08
$n'=8, n=4$	$t_{d(opt)II}$ (or $\hat{t}_{d(opt)II}$)	145.00	271.38	100.11	271.16	100.12

Table 1 indicates that there is enough scope of choosing the scalar α involved in the suggested estimator t_d to obtain better estimators than $\overline{y}$, t_{Rd}, t_{Pd}, $t_d^{(1)}$ *and* $t_d^{(2)}$. It is observed from Table 2 that the optimum estimators $t_{d(opt)I}$ *and*

$t_{d(opt)II}$ or estimators based on 'estimated optimum' values $\hat{t}_{d(opt)I}$ and $\hat{t}_{d(opt)II}$ are better than $\bar{y}$, t_{Rd}, t_{Pd}, $t_d^{(1)}$ and $t_d^{(2)}$. It is further observed from Table 2 that the performance of the estimator $t_{d(opt)I}$ (or $\hat{t}_{d(opt)}$) in case I is better than case II. Larger gain in efficiency is observed by using proposed estimators over other estimators except in few cases where the gain is marginal (or the estimators are almost equally efficient). Thus we recommend the use of the proposed estimators t_d for its use in practice.

7. CONCLUSION

This paper deals with the problem of estimating population mean $\bar{Y}$ of the study variable y using two-phase (or double) sampling procedure. The double sampling version of the class of estimators envisaged by Singh and Tailor (2005) has been suggested and its properties are studied under two well known cases designated as case I and case II. Optimum estimators in the proposed class have been identified in both the cases alongwith their approximate variance formulae. Estimators based on estimated optimum values are also derived along with their approximate variance formulae. It is interesting to mention that the estimators based on 'optimum value' and 'estimated optimum value' have the same approximate variance formula which follows that the proposed estimator can be used fruitfully even if the optimum values of the constants involved in the estimator are not known.

An empirical study is carried out to throw light on the merits of the proposed estimator over other existing competitors.

Theoretically and empirically it has been demonstrated that the proposed optimum estimators (or estimators based on estimated optimum values) in case I is more efficient than in case II. Results of this paper are quite illuminating and useful to the practitioners

School of Studies in Statistics, Vikram University
Ujjain- 456010, M.P. India
HOUSILA P. SINGH

Wood Properties and Uses Division,
Institute of Wood Science and Technology,
Bangalore-560003, Karnataka, India
RITESH TAILOR

School of Studies in Statistics, Vikram University
Ujjain- 456010, M.P. India
RAJESH TAILOR

ACKNOWLEDGEMENT

Authors are thankful to referee for his valuable suggestions regarding the improvement of the paper.

REFERENCES

A.K. DAS (1988), *Contribution to the theory of sampling strategies based on auxiliary information.* Ph.D. thesis submitted to BCKV; Mohanpur Nadia West Bengal India.

A.K. DAS, T.P. TRIPATHI (1978), *Use of auxiliary information in estimating the finite population variance.* "Sankhya" C 40, pp. 139-148.

D. ADHVARYU, GUPTA P.C. (1983), *On some alternative strategies using auxiliary information.* "Metrika" 30, pp. 217-226.

D.M. KAWATHEKAR, S.G.P AJAGAONKAR (1984), *A modified ratio estimator based on the coefficient of variation in double sampling.* "Journal of Indian Statistical Association." 36 (2) pp. 47-50.

G.S. MADDALA (1977), *Econometrics* "McGraw Hills pub.Co." New York.

H. P. SINGH, L.N. UPADHYAYA (1986), *A dual to modified ratio estimator using coefficient of variation of auxiliary variable.* "Proceedings National Academy of Sciences" (India) 56, A IV, pp. 336-340.

H.P. SINGH (1986), *A generalized class of estimators of ratio product and mean using supplementary information on an auxiliary character in PPSWR sampling scheme.*"Gujarat Statistical Review" 13(2), pp. 1-30.

H.P. SINGH, L.N. UPADHYAYA, P. CHANDRA (2004), *A general family of estimators for estimating population mean using two auxiliary variables in two-phase sampling.* "Statistics in Transition" 6 (7), pp. 1055-1077.

H.P. SINGH, M.R. ESPEJO (2003), *On linear regression and ratio – product estimation of a finite population mean.* "Statistician" 52(1), pp. 59-67.

H.P. SINGH, R. TAILOR (2005), *Estimation of finite population mean with coefficient of variation of an auxiliary character.* "Statistica" 65 (3), pp. 407-413.

M.N. MURTHY (1967), *Sampling theory and methods* Statistical Publishing Society Calcutta (India), pp. 1-1220 (Vol. I and Vol. II) Kluwer Academic Publishers The Netherlands.

R.G.D. STEEL, J. H. TORRIE (1960), *Principles and procedures of Statistics* McGraw Hill Book Co.

S. SINGH (2003), *Advanced sampling theory with applications, How Michel "Selected Amy.*

S.K. SRIVASTAVA (1970), *A two phase sampling estimator in sample surveys.* "Australian Journal of Statistics", 12, pp. 23-27.

V. N. REDDY (1978), *A study on the use of prior knowledge on certain population parameters in estimation.* "Sankhya" C, 40, pp. 29-37.

SUMMARY

Estimation of finite population mean in two-phase sampling with known coefficient of variation of an auxiliary character

In this paper a double (or two-phase) sampling version of (Singh and Tailor, 2005) estimator has been suggested along with its properties under large sample approximation. It is shown that the estimator due to (Kawathekar and Ajgaonkar, 1984) is a member of the proposed class of estimators. Realistic conditions have been obtained under which the proposed estimator is better than usual unbiased estimator $\overline{y}$ usual double sampling ratio (t_{Rd}) product (t_{Pd}) estimators and (Kawathekar and Ajgaonkar, 1984) estimator. This fact has been shown also through an empirical study.