EFFICIENT ESTIMATION OF POPULATION MEAN USING INCOMPLETE SURVEY DATA ON STUDY AND AUXILIARY CHARACTERISTICS

H. Toutenburg, V.K. Srivastava

1. INTRODUCTION

Ratio and product methods are two popular and easily comprehensible techniques for the estimation of population mean in survey sampling when an auxiliary characteristic correlated with the study characteristic is available; see, e.g., Sukhatme, Sukhatme *et al.* (1984). These techniques provide generally biased but more efficient estimators in comparison to the traditional unbiased estimator, viz., the sample mean provided that the correlation between the auxiliary characteristic and the study characteristic is sufficiently positive in case of ratio method and negative in case of product method. Both the methods of estimation assume that the sample data contain no missing observation and the population mean of auxiliary characteristic is known. One or both of these specifications may not be tenable in many practical applications; see, e.g., Rubin (1987) for an excellent exposition. When no observation is missing but the population mean of auxiliary characteristic is not available, it is customary to make use of a large preliminary sample for finding an estimate of it. If the circumstances do not permit to have the preliminary sample due to some practical difficulties or otherwise, an alter native estimator for the population mean of auxiliary characteristic based on the given sample data may be utilized; see Srivastava and Bhatnagar (1981).

On the other strand, when some observations are missing but the population mean of auxiliary characteristic is available, Tracy and Osahan (1994) have considered two estimators arising from ratio method and have analyzed their efficiency properties. There appears to be no effort reported in the literature when both the assumptions are violated simultaneously, i.e., some observations are missing in the survey data and the population mean of the auxiliary characteristic is not available. Considering the missingness of few observations on both the characteristics, Toutenburg and Srivastava (1998) have discussed the estimation of the ratio of population means. Their estimators can be used immediately to formulate estimators for the population mean of study characteristic provided that the population mean of the auxiliary characteristic is known. In the absence

of such knowledge, straightforward application is not possible. This is the main concern of present investigations. The plan of paper is as follows. In Section 2, we consider the estimation of the population mean of study characteristic using sample data when some observations on both the study and auxiliary characteristics are missing. One unbiased and four biased estimators arising from the ratio and product methods of estimation are presented. Their bias properties are analyzed in Section 3 while their mean squared errors are compared in Section 4. Finally, some summarizing remarks are offered in Section 5 and derivation of results is presented in Appendix.1

2. ESTIMATORS FOR POPULATION MEAN

Let there be a finite population consisting of N distinct units with values $Y_1, Y_2, ..., Y_N$ for the study characteristic and values $X_1, X_2, ..., X_N$ for the auxiliary characteristic. It is proposed to estimate the population mean $\overline{Y}$ using the auxiliary information on the basis of a random sample of size n drawn according to the procedure of simple random sampling without replacement. When all the observations are available and the population mean $\overline{X}$ of the auxiliary characteristic is known, the ratio and product methods of estimation provides the following estimators of $\overline{Y}$:

$$\hat{\overline{Y}}_R = \frac{\bar{y}_n \overline{X}}{\bar{x}_n} \tag{1}$$

$$\hat{\overline{Y}}_P = \frac{\bar{y}_n \bar{x}_n}{\overline{X}} \tag{2}$$

where $\bar{y}_n$ and $\bar{x}_n$ are the means of sample observations on the study characteristic and auxiliary characteristic respectively.

Unlike the unbiased estimator $\bar{y}_n$, both the estimators (1) and (2) are generally biased. Comparing the estimators with respect to the criterion of mean squared error using large sample theory, $\hat{\overline{Y}}_R$ is better than $\bar{y}_n$ for ρ greater than $(\theta/2)$ while $\hat{\overline{Y}}_P$ is better than $\bar{y}_n$ for ρ less than $(-\theta/2)$ where

$$\rho = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum_{i=1}^{N}(X_i - \overline{X})^2 \sum_{i=1}^{N}(Y_i - \overline{Y})^2\right]^{\frac{1}{2}}} \tag{3}$$

$$\theta = \left(\frac{\bar{Y}}{\bar{X}}\right)\left[\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}\right]^{\frac{1}{2}} \tag{4}$$

Notice that ρ is the correlation coefficient between the auxiliary and study characteristics in the population and θ is the ratio of coefficients of variation of the auxiliary and study characteristics.

The ratio estimator $\hat{\bar{Y}}_R$ and the product estimator $\hat{\bar{Y}}_P$ cannot be used in practice when there are some missing observations in the sample data. Assuming $\bar{X}$ to be known, Tracy and Osahan (1994) have presented two ratio estimators and have compared their efficiency properties. Some more ratio estimators can be formulated from the investigations conducted by Toutenburg and Srivastava (1998) who have considered the problem of estimating the ratio of two population means. All these estimators loose their practical utility when $\bar{X}$ is not known.

Let us consider the situation where $\bar{X}$ is not available and the sample contains some missing observations. In particular we assume that only $(n - p - q - k)$ observations $(x_1, y_1),(x_2, y_2),...,(x_{n-p-q-k}, y_{n-p-q-k})$ in the sample are complete. On p sampling units, observations $x_1^*, x_2^*,..., x_p^*$ are available while the corresponding observations on study characteristic are missing. Similarly, on q sampling units, we have only the observations $y_1^{**}, y_2^{**},..., y_q^{**}$ on the study characteristic without any corresponding value of the auxiliary characteristic. Further, there are k sampling units on which observations on both the study and auxiliary characteristics are not available. The numbers p, q and k are assumed to be random.

In the presence of missing observations in the data set, a popular strategy is to discard all the $(p + q + k)$ incomplete pairs of observations and to use only the $(n - p - q - k)$ complete pairs. Accordingly, an unbiased estimator of population mean is

$$\bar{y} = \frac{1}{(n - p - q - k)}\sum y_i$$

On the other hand, if we utilize incomplete observations too and use the ratio and product methods of estimators, the following four estimators of $\bar{Y}$ in view of (1) and (2) can be formulated:

$$\hat{\bar{Y}}_1 = \bar{y}\left[\frac{(n - p - q - k)\bar{x} + p\bar{x}^*}{(n - q - k)\bar{x}}\right]$$

$$\hat{\bar{Y}}_2 = \bar{y}\left[\frac{(n-q-k)\bar{x}}{(n-p-q-k)\bar{x}+p\bar{x}^*}\right]$$

$$\hat{\bar{Y}}_3 = \frac{[(n-p-q-k)\bar{y}+q\bar{y}^{**}][(n-p-q-k)\bar{x}+p\bar{x}^*]}{(n-p-k)(n-q-k)\bar{x}}$$

$$\hat{\bar{Y}}_4 = \left[\frac{(n-p-q-k)\bar{y}+q\bar{y}^{**}}{(n-p-q-k)\bar{x}+p\bar{x}^*}\right]\left(\frac{n-q-k}{n-p-k}\right)\bar{x}$$

where

$$\bar{x}^* = \frac{1}{p}\sum x_i^*$$

$$\bar{y}^{**} = \frac{1}{q}\sum y_i^{**}$$

$$\bar{x} = \frac{1}{(n-p-q-k)}\sum x_i$$

It may be observed that these four estimators utilize all the available observations on the auxiliary characteristic. So far as the use of observations on the study characteristic is concerned, the estimators $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_2$ ignore them while the estimators $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ incorporate them.

Thus the estimator $\bar{y}$ can be regarded as representing the strategy of total discard of incomplete observations. Similarly, the strategy of partial discard and partial utilization of incomplete observations leads to the estimators $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_2$ while the strategy of full utilization of available observations provides the estimator $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$.

3. COMPARISON OF BIASES

In addition to (3) and (4), let us introduce the following notation:

$$C = \frac{1}{\bar{Y}^2(N-1)}\sum_{i=1}^{N}(Y_i-\bar{Y})^2$$

$$f_s = E_1\left(\frac{1}{n-s}\right) - \frac{1}{N}$$

where the expectation operator E_1 in f_s refers to averaging over all possible values of the non-negative integer valued random variable s.

Further, we observe that

$$f_{p+q+k} \geq f_{p+k}$$

$$f_{p+q+k} \geq f_{q+k}$$

It is easy to see that $\bar{y}$ is an unbiased estimator of $\overline{Y}$ while the estimators $\hat{\overline{Y}}_1$, $\hat{\overline{Y}}_2$, $\hat{\overline{Y}}_3$ and $\hat{\overline{Y}}_4$ are generally biased. The large sample approximations for their relative biases are derived in Appendix and are presented below.

Theorem 1. The large sample approximations for the relative biases of the estimators $\hat{\overline{Y}}_1$, $\hat{\overline{Y}}_2$, $\hat{\overline{Y}}_3$ and $\hat{\overline{Y}}_4$ are given by

$$RB(\hat{\overline{Y}}_1) = E\left(\frac{\hat{\overline{Y}}_1 - \overline{Y}}{\overline{Y}}\right) = C\theta(\theta - \rho)(f_{p+q+k} - f_{q+k})$$

$$RB(\hat{\overline{Y}}_2) = E\left(\frac{\hat{\overline{Y}}_2 - \overline{Y}}{\overline{Y}}\right) = C\theta\rho(f_{p+q+k} - f_{q+k}) \tag{5}$$

$$RB(\hat{\overline{Y}}_3) = E\left(\frac{\hat{\overline{Y}}_3 - \overline{Y}}{\overline{Y}}\right) = C\theta^2(f_{p+q+k} - f_{q+k}) \tag{6}$$

$$RB(\hat{\overline{Y}}_4) = E\left(\frac{\hat{\overline{Y}}_4 - \overline{Y}}{\overline{Y}}\right) = 0.$$

It is interesting to observe that the estimator $\hat{\overline{Y}}_4$ is nearly unbiased in the sense that its bias to order $O(n^{-1})$ vanishes. Similarly, the estimator $\hat{\overline{Y}}_1$ is also nearly unbiased provided that $\theta = \rho$. When θ and ρ are not equal, the relative bias of $\hat{\overline{Y}}_1$ is negative for θ less than ρ and positive for θ greater than ρ. In case of $\hat{\overline{Y}}_2$, the relative bias has the same sign as the correlation coefficient ρ. Interestingly enough, the relative bias of $\hat{\overline{Y}}_3$ is invariably positive and does not depend upon the correlation coefficient, at least to the order of our approximation.

Comparing the estimators with respect to the magnitude of bias to the given order of approximation, we observe that $\hat{\bar{Y}}_1$ is better than $\hat{\bar{Y}}_2$ for ρ larger than $(\theta/2)$. The opposite is true, i.e., $\hat{\bar{Y}}_2$ is better than $\hat{\bar{Y}}_1$ when ρ is negative. This result remains true for positive values of ρ provided that ρ is less than $(\theta/2)$.

Similarly, the estimator $\hat{\bar{Y}}_1$ has smaller magnitude of bias in comparison to $\hat{\bar{Y}}_3$ when

$$0 < \rho < 2\theta$$

which is always satisfied if θ exceeds 0.5. When the correlation coefficient is negative, the reverse is true, i.e., $\hat{\bar{Y}}_1$ has larger magnitude of bias than $\hat{\bar{Y}}_3$. This continues to remain true when

$$\rho > 2\theta$$

provided that θ is less than 0.5.

If we compare $\hat{\bar{Y}}_2$ and $\hat{\bar{Y}}_3$, it is observed from (5) and (6) that $\hat{\bar{Y}}_2$ has smaller magnitude of bias than $\hat{\bar{Y}}_3$ as long as θ exceeds 1. This result holds true for θ not exceeding 1 when

$$\rho^2 < \theta^2$$

On the other hand, the opposite is true, i.e., $\hat{\bar{Y}}_2$ has larger magnitude of bias than $\hat{\bar{Y}}_3$ when

$$\rho^2 > \theta^2$$

provided that θ is less than 1.

4. COMPARISON OF MEAN SQUARED ERRORS

The relative variance of the unbiased estimator $\bar{y}$ is

$$RV(\bar{y}) = \mathrm{E}\left(\frac{\bar{y}-\bar{Y}}{\bar{Y}}\right)^2 = Cf_{p+q+k} \tag{7}$$

For the remaining four biased estimators, we consider the relative mean squared errors and derive their large sample approximations in Appendix.

Theorem 2. The large sample approximations for the relative mean squared errors of the estimators $\hat{\bar{Y}}_1$, $\hat{\bar{Y}}_2$, $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ are given by

$$RMSE(\hat{\bar{Y}}_1) = \mathrm{E}\left(\frac{\hat{\bar{Y}}_1 - \bar{Y}}{\bar{Y}}\right)^2 \tag{8}$$

$$= C[f_{p+q+k} + \theta(\theta - 2\rho)(f_{p+q+k} - f_{p+k})]$$

$$RMSE(\hat{\bar{Y}}_2) = \mathrm{E}\left(\frac{\hat{\bar{Y}}_2 - \bar{Y}}{\bar{Y}}\right)^2 \tag{9}$$

$$= C[f_{p+q+k} + \theta(\theta + 2\rho)(f_{p+q+k} - f_{q+k})]$$

$$RMSE(\hat{\bar{Y}}_3) = \mathrm{E}\left(\frac{\hat{\bar{Y}}_3 - \bar{Y}}{\bar{Y}}\right)^2 \tag{10}$$

$$= C[f_{p+k} + \theta^2(f_{p+q+k} - f_{q+k})]$$

$$RMSE(\hat{\bar{Y}}_4) = \mathrm{E}\left(\frac{\hat{\bar{Y}}_4 - \bar{Y}}{\bar{Y}}\right)^2 \tag{11}$$

$$= C[f_{p+k} + \theta^2(f_{p+q+k} - f_{p+k})]$$

From (7), (8) and (9), we find that the estimator $\hat{\bar{Y}}_1$ is more efficient than $\bar{y}$ when

$$\rho > \frac{\theta}{2}; \theta < 2 \tag{12}$$

while $\hat{\bar{Y}}_2$ is more efficient than $\bar{y}$ when

$$(-\rho) > \frac{\theta}{2}; \theta < 2. \tag{13}$$

Notice that (12) is a well known condition for the superiority of ratio estimator over the sample mean when there are no missing observations in the data and $\bar{X}$ is known. Similarly, (13) is the condition under which product estimator is better than the sample mean provided that no observation is missing and $\bar{X}$ is known.

It is interesting to observe from (10) and (11) that the mean squared errors of the estimators $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ do not depend upon the correlation coefficient ρ, at least to the order of our approximation. Thus, for all values of ρ, the estimators $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ are more efficient than $\bar{y}$ when

$$\theta^2 > \left(\frac{f_{p+q+k} - f_{p+k}}{f_{p+q+k} - f_q + k} \right) \tag{14}$$

When the conditions (12), (13) and (14) hold with a reversed inequality sign, the estimator $\bar{y}$ remains unbeaten.

Next, let us compare the biased estimators.

It is seen from (8) and (9) that $\hat{\bar{Y}}_1$ is better than $\hat{\bar{Y}}_2$ for ρ greater than 0.25 while the opposite is true, i.e., $\hat{\bar{Y}}_2$ is better than $\hat{\bar{Y}}_1$ for ρ less than 0.25 which always hold true for negative correlation between the study and auxiliary characteristics.

Similarly, if we compare $\hat{\bar{Y}}_1$ with $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$, we observe that the estimator $\hat{\bar{Y}}_1$ is better than the estimators $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ when

$$\rho > \frac{(f_{p+q+k} - f_{p+k})}{2\theta(f_{p+q+k} - f_{q+k})}$$

$$\left(\frac{f_{p+q+k} - f_{p+k}}{f_{p+g+k} - f_{q+k}} \right) < 2\theta$$

The opposite is true, i.e., both the estimators $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ are better than $\hat{\bar{Y}}_1$ when

$$\rho < \frac{(f_{p+q+k} - f_{p+k})}{2\theta(f_{p+q+k} - f_{q+k})}$$

which is clearly satisfied so long as

$$\left(\frac{f_{p+q+k} - f_{p+k}}{f_{p+q+k} - f_{q+k}} \right) > 2\theta \tag{15}$$

In a similar manner, comparing $\hat{\bar{Y}}_2$ with $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$, we find that $\hat{\bar{Y}}_2$ is better than $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ when

$$(-\rho) > \frac{(f_{p+q+k} - f_{p+k})}{2\theta(f_{p+q+k} - f_{q+k})}$$

$$\left(\frac{f_{p+q+k} - f_{p+k}}{f_{p+q+k} - f_{q+k}}\right) < 2\theta$$

which requires correlation to be negative.

On the other hand, both the estimators $\hat{\overline{Y}}_3$ and $\hat{\overline{Y}}_4$ are better than $\hat{\overline{Y}}_2$ when

$$(-\rho) < \frac{(f_{p+q+k} - f_{p+k})}{2\theta(f_{p+q+k} - f_{q+k})} \tag{16}$$

which is always satisfied as long as ρ is positive. For negative correlation coefficient, again the condition (16) is satisfied provided that the inequality (15) holds good.

Finally, it is evident from (10) and (11) that $\hat{\overline{Y}}_3$ and $\hat{\overline{Y}}_4$ are equally efficient, at least to the given order of approximation.

5. SOME REMARKS

We have considered the problem of estimating the mean of a population of size N on the basis of a random sample of size n drawn according to the procedure of simple random sampling without replacement. It is assumed that some observations in the sample are missing randomly. In particular, there are only $(n-p-q-k)$ pairs of complete observations; the remaining $(p+q+k)$ pairs are incomplete. Out of these, p observations on the study characteristic and q observations on the auxiliary characteristic are missing. There are k sampling units on which observations on both the characteristic are missing. Further, $\overline{X}$ is assumed to be unknown.

In all, four estimators $\hat{\overline{Y}}_1$, $\hat{\overline{Y}}_2$, $\hat{\overline{Y}}_3$ and $\hat{\overline{Y}}_4$ of population mean $\overline{Y}$ arising from the ratio and product methods of estimation are formulated. The estimators $\hat{\overline{Y}}_1$ and $\hat{\overline{Y}}_2$ can be regarded as based on the strategy of partial discard and partial utilization of available information in the sense that they do not use the q observations on the study characteristic. The strategy of full utilization of available information provides the estimators $\hat{\overline{Y}}_3$ and $\hat{\overline{Y}}_4$. For the sake of comparison, we have also considered the estimator $\bar{y}$ as representative of the strategy of outright discard of incomplete information.

Using the large sample theory, our investigations have revealed that $\bar{y}$ is an exactly unbiased estimator of $\overline{Y}$ while $\hat{\overline{Y}}_4$ is nearly unbiased. The other estimator $\hat{\overline{Y}}_3$ representing the strategy of full utilization of available observations is always biased in the positive direction. However, the direction of relative bias in case of the estimator $\hat{\overline{Y}}_1$ depends upon the magnitude as well as the sign of correlation coefficient ρ and value of θ, the ratio of the coefficients of variation while the relative bias of $\hat{\overline{Y}}_2$ has the same sign as the correlation coefficient ρ.

Comparing with respect to the criterion of magnitude of bias, it is found that $\hat{\overline{Y}}_3$ is superior to $\hat{\overline{Y}}_1$ for all negative values of ρ such that $\rho > 2\theta$. Similarly, the estimator $\hat{\overline{Y}}_3$ is superior to $\hat{\overline{Y}}_2$ when the absolute value of ρ exceeds θ. If we compare the estimators $\hat{\overline{Y}}_1$ and $\hat{\overline{Y}}_2$ arising from the strategy of partial utilization, it is seen that $\hat{\overline{Y}}_1$ has smaller (larger) amount of bias in comparison to the estimator $\hat{\overline{Y}}_2$ when θ is smaller (larger) than 2ρ.

When we compare the performance of estimators with respect to the criterion of mean squared error to the given order of approximation, our investigations have brought out that no strategy is uniformly superior to the other. For instance, the strategy of outright discard of incomplete pairs of observations may outperform the strategies of partial and full utilization of the available observations.

It is interesting to observe that the estimators $\hat{\overline{Y}}_3$ and $\hat{\overline{Y}}_4$ have identical mean squared errors, at least to the order of our approximation. Thus the estimator $\hat{\overline{Y}}_4$ may be preferable in comparison to the estimator $\hat{\overline{Y}}_3$ by its virtue of being nearly unbiased.

Another interesting observation relates to comparison of $\bar{y}$ with $\hat{\overline{Y}}_1$ and $\hat{\overline{Y}}_2$. The biased estimators $\hat{\overline{Y}}_1$ and $\hat{\overline{Y}}_2$ are found to be superior than the unbiased estimator $\bar{y}$ precisely under the same conditions which are required for the ratio and product estimator to be better than the sample mean when no observation is missing and $\overline{X}$ is known.

Finally, it may be remarked that an appropriate choice of estimator can be made on the basis of our analysis in any given situation. This requires the knowledge of ρ and θ which are generally unknown. However, one may often have some prior information about these parameters and may use it in making a choice of estimator as pointed out by Toutenburg and Srivastava (1998).

APPENDIX

If we write

$$u = \left(\frac{\bar{x} - \bar{X}}{\bar{X}} \right)$$

$$\eta = \frac{(n - p - q - k)(\bar{x} - \bar{X}) + p(\bar{x}^{*} - \bar{X})}{(n - q - k)\bar{X}}$$

$$v = \left(\frac{\bar{y} - \bar{Y}}{\bar{Y}} \right)$$

$$\varepsilon = \frac{(n - p - q - k)(\bar{x} - \bar{X}) + q(\bar{y}^{**} - \bar{Y})}{(n - p - k)\bar{Y}}$$

it can be easily verified, following Toutenburg and Srivastava (1998), that

$$\mathrm{E}(u) = \mathrm{E}(v) = \mathrm{E}(\eta) = \mathrm{E}(\varepsilon) = 0$$
$$\mathrm{E}(u^2) = C\theta^2 f_{p+q+k}$$
$$\mathrm{E}(v^2) = C f_{p+q+k}$$
$$\mathrm{E}(\eta^2) = C\theta^2 f_{q+k}$$
$$\mathrm{E}(\varepsilon^2) = C f_{p+k}$$
$$\mathrm{E}(uv) = C\theta\rho f_{p+q+k}$$
$$\mathrm{E}(u\eta) = C\theta^2 f_{p+k}$$
$$\mathrm{E}(u\varepsilon) = C\theta\rho f_{p+k}$$
$$\mathrm{E}(\eta v) = C\theta\rho f_{q+k}$$
$$\mathrm{E}(\eta\varepsilon) = C\theta\rho f_{p+k}$$

when n is large.

Now we can express

$$\left(\frac{\hat{\bar{Y}}_1 - \bar{Y}}{\bar{Y}} \right) = [(v + \eta - u) + v\eta](1 + u)^{-1}$$

$$= (v + \eta - u) + [v\eta - (v + \eta - u)u] + \mathbf{O}_p\left(n^{-\frac{3}{2}} \right)$$

$$\left(\frac{\hat{\bar{Y}}_2-\bar{Y}}{\bar{Y}}\right)=[(v-\eta+u)+uv](1+\eta)^{-1}$$

$$=(v-\eta+u)+[uv-(v-\eta+u)\eta]+\mathrm{O}_p\left(n^{-\frac{3}{2}}\right)$$

$$\left(\frac{\hat{\bar{Y}}_3-\bar{Y}}{\bar{Y}}\right)=[(\varepsilon+\eta-u)+\varepsilon\eta](1+u)^{-1}$$

$$=(\varepsilon+\eta-u)+[\varepsilon\eta-(\varepsilon+\eta-u)u]+\mathrm{O}_p\left(n^{-\frac{3}{2}}\right)$$

$$\left(\frac{\hat{\bar{Y}}_4-\bar{Y}}{\bar{Y}}\right)=[(\varepsilon-\eta+u)+\varepsilon u](1+\eta)^{-1}$$

$$=(\varepsilon-\eta+u)+[\varepsilon u-(\varepsilon-\eta+u)\eta]+\mathrm{O}_p\left(n^{-\frac{3}{2}}\right).$$

Thus the relative biases to order $\mathrm{O}_p(n^{-1})$ are given by

$$RB(\hat{\bar{Y}}_1)=\mathrm{E}\left(v+\eta-u\right)+\mathrm{E}(v\eta-uv-u\eta+u^2)$$

$$=C\theta(\theta-\rho)(f_{p+q+k}-f_{q+k})$$

$$RB(\hat{\bar{Y}}_2)=\mathrm{E}(v-\eta+u)+\mathrm{E}(uv-v\eta+\eta^2-u\eta)$$

$$=C\theta\rho(f_{p+q+k}-f_{q+k})$$

$$RB(\hat{\bar{Y}}_3)=\mathrm{E}(\varepsilon+\eta-u)+\mathrm{E}(\varepsilon\eta-\varepsilon u-u\eta+u^2)$$

$$=C\theta^2(f_{p+q+k}-f_{q+k})$$

$$RB(\hat{\bar{Y}}_4)=\mathrm{E}(\varepsilon-\eta+u)+\mathrm{E}(\varepsilon u-\varepsilon\eta+\eta^2-u\eta)$$

$$=0$$

which provide the results stated in Theorem 1.

In a similar manner, the large sample approximations for the mean squared errors are

$$RMSE(\hat{\bar{Y}}_1)=\mathrm{E}(v+\eta-u)^2$$

$$=C[f_{p+q+k}+(f_{p+q+k}-f_{q+k})(\theta-2\rho)\theta]$$

$$RMSE(\hat{\bar{Y}}_2) = \mathrm{E}(v - \eta + u)^2$$

$$= C[f_{p+q+k} + (f_{p+q+k} - f_{q+k})(\theta + 2\rho)\theta]$$

$$RMSE(\hat{\bar{Y}}_3) = \mathrm{E}(v + \eta - u)^2$$

$$= C[f_{p+k} + (f_{p+q+k} - f_{q+k})\theta^2]$$

$$RMSE(\hat{\bar{Y}}_4) = \mathrm{E}(\varepsilon - \eta + u)^2$$

$$= C[f_{p+k} + (f_{p+q+k} - f_{q+k})\theta^2]$$

which lead to Theorem 2.

Institute of Statistics HELGE TOUTENBURG
University of Munich, Germany

Department of Statistics V.K. SRIVASTAVA
University of Lucknow, India

REFERENCES

D.B. RUBIN, (1987), *Multiple Imputation for Nonresponse in Sample Surveys*, Wiley, New York.

V.K. SRIVASTAVA, S. BHATNAGAR, (1981), *Ratio and product methods of estimation when $\overline{X}$ is not known*, "Journal of Statistical Research", 15, 29-39.

P.V. SUKHATME et al., (1984), *Sampling Theory Of Surveys With Applications*, Iowa State University Press, Iowa.

H. TOUTENBURG, V.K. SRIVATAVA, (1998), *Estimation of ratio of population means in survey sampling when some observations are missing*, "Metrika", 48, 177-187.

D.S. TRACY, S.S. OSAHAN, (1994), *Random non-response on study variable versus on study as well as auxiliary variables*, "Statistica", 54, 163-168.

RIASSUNTO

Stima efficiente della media di popolazione usando dati campionari incompleti e variabili ausiliarie

Nel lavoro viene considerato il problema della stima della media di popolazione basata sui metodi del rapporto e del prodotto in presenza di dati mancanti e quando la media di popolazione della variabile ausiliaria non è nota. Oltre ad uno stimatore corretto, costruito scartando le coppie di osservazioni incomplete, sono presentati quattro ulteriori stimatori, in genere distorti. I primi due stimatori sono costruiti tramite un uso parziale dei dati, mentre i rimanenti due usano tutta l'informazione disponibile. Viene quindi effettuato uno studio comparativo delle proprietà di efficienza degli stimatori proposti e, infine, viene discusso il problema della scelta dello stimatore.

SUMMARY

Efficient estimation of population mean using incomplete survey data on study and auxiliary characteristics

This paper considers the problem of estimating the population mean using the ratio and product methods when some observations in the sample data are missing at random and the population mean of the auxiliary characteristic is not known. Besides an unbiased estimator arising from the total discard of incomplete pairs of observations, four generally biased estimators are presented. The first two estimators arise from the partial utilization of data while the remaining two are based on full utilization. A comparative study of the efficiency properties of estimators is reported and the choice of estimators is discussed.