Toutenburg, Srivastava:

Estimation of Ratio of Population Means in Survey Sampling When Some Observations are Missing

Sonderforschungsbereich 386, Paper 119 (1998)

Online unter: http://epub.ub.uni-muenchen.de/

Estimation of ratio of population means in survey sampling when some observations are missing

H. Toutenburg
Institut für Statistik, Universität München
80799 München, Germany

V. K. Srivastava
Department of Statistics
University of Lucknow
Lucknow 226007, India

July 8, 1998

1 Introduction

In most of the developments concerning the use of auxiliary information in the estimation of parameters in survey sampling, it is typically assumed that all the observations on selected units in the sample are available. This may not hold true in many practical situations encountered in sample surveys and some observations may be missing for various reasons such as unwillingness of some selected units to supply the desired information, accidental loss of information caused by unknown factors, failure on the part of investigator to gather correct information. In fact, missingness of observations is not an uncommon feature in opinion polls, market research surveys, mail enquiries, socio-economic investigations, medical studies and other scientific experiments. In such circumstances, the traditional procedures for deducing inferences cannot be applied straightforwardly.

Ratio of two population means in survey sampling is conventionally estimated by the ratio of corresponding sample means; see e.g., Cochran (1977) and Sukhatme, Sukhatme and Sukhatme (1984). This estimation procedure does not work when some of the observations are missing. Assuming that some observations are missing on either the study characteristic or both the study and auxiliary characteristics, Tracy and Osahan (1994) have considered the estimation of ratio of means. In this article, we consider a situation in which there are some observations missing on one of the characteristics at a time and thus the missingness phenomenon occurs for both the characteristics separately but not, of course, simultaneously. As an illustration, let the study characteristic (Y) be the consumption expenditure of a household in the current month and the auxiliary characteristic (X) be the consumption expenditure in the same month three years ago. Then there may be some households for which the values of X

are lost or could not be recorded earlier, and the households are now not able to recall the correct value of X while the corresponding values of Y are readily available. Similarly, there may be some households for which values of X are available but the corresponding values of Y cannot be obtained, for instance, when such families have gone on holidays or have permanently moved out of town.

The plan of this article is as follows. In Section 2, we describe the data framework and present four simple estimators for the ratio of population means. Large sample approximations for their relative biases and relative mean squared errors are given in Section 3 and using them a comparison of estimators is made. Some remarks are then placed in Section 4. Finally, the derivation of results is outlined in Appendix.

2 Estimation of Ratio

Consider a population of N units from which a random sample of size n is drawn according to simple random sampling without replacement procedure for the estimation of ratio $R = (\bar{Y}/\bar{X})$ of population means $\bar{X}$ and $\bar{Y}$ of two characteristics X and Y, respectively.

It is assumed that a set of $(n-p-q)$ complete observations $(x_1, y_1), (x_2, y_2), \ldots,$ (x_{n-p-q}, y_{n-p-q}) on selected units in the sample are available. In addition to these, observations $x_1^*, x_2^*, \ldots, x_p^*$ on p units in the sample are available but the corresponding observations on Y characteristic are missing. Similarly, we have a set of q observations $y_1^{**}, y_2^{**}, \ldots, y_q^{**}$ on Y characteristic in the sample but the associated values on X characteristic are missing. Further, the quantities p and q denoting the number of incomplete observations are assumed to be random following Tracy and Osahan (1994).

If we write

$$\begin{aligned} \bar{x} &= \frac{1}{n-p-q}\sum x_i \\ \bar{y} &= \frac{1}{n-p-q}\sum y_i \\ \bar{x}^* &= \frac{1}{p}\sum x_i^* \\ \bar{y}^{**} &= \frac{1}{q}\sum y_i^{**} \end{aligned} \tag{2.1}$$

the following estimators for the ratio $R = (\bar{Y}/\bar{X})$ can be formulated:

$$r_1 = \frac{\bar{y}}{\bar{x}} \tag{2.2}$$

$$r_2 = \frac{(n-q)\bar{y}}{(n-p-q)\bar{x} + p\bar{x}^*} \tag{2.3}$$

$$r_3 = \frac{(n-p-q)\bar{y} + q\bar{y}^{**}}{(n-p)\bar{x}} \tag{2.4}$$

$$r_4 = \left(\frac{n-q}{n-p}\right)\frac{(n-p-q)\bar{y} + q\bar{y}^{**}}{(n-p-q)\bar{x} + p\bar{x}^*} \tag{2.5}$$

The estimator r_1 is based on complete observations numbering $(n-p-q)$ and ignores all the incomplete pairs of observations. The estimators r_2 and r_3 make use of incomplete observations only partly. For example, the estimator r_2 utilizes the p observations on X characteristic only while r_3 uses the q observations on Y characteristic only. The estimator r_4, however, incorporates all the available observations.

3 Comparison of Estimators

For the comparison of the performance properties of the four estimators of ratio R, we introduce the following notation:

$$\begin{aligned}
C_x^2 &= \frac{1}{(N-1)\bar{X}^2}\sum^N (X_i-\bar{X})^2 \\
C_y^2 &= \frac{1}{(N-1)\bar{Y}^2}\sum^N (Y_i-\bar{Y})^2 \\
\rho &= \frac{1}{(N-1)\bar{X}\bar{Y}C_xC_y}\sum^N (X_i-\bar{X})(Y_i-\bar{Y}) \\
f_s &= \mathrm{E}\left(\frac{1}{n-s}\right)-\frac{1}{N} \qquad (3.1)
\end{aligned}$$

where expectation operator in the last quantity refers to all possible values of the non-negative integer valued random variable s.

Thus we have

$$\begin{aligned}
f_p &\le f_{(p+q)} && (3.2)\\
f_q &\le f_{(p+q)} && (3.3)\\
f_p &\lesseqgtr f_q \quad \text{if} \quad p \lesseqgtr q && (3.4)
\end{aligned}$$

We now present the results which are derived in Appendix.

Theorem I: When sample size is large, the relative biases of the four estimators of R can be approximated by

$$\begin{aligned}
\mathrm{RB}(r_1) &= \mathrm{E}\left(\frac{r_1-R}{R}\right) && (3.5)\\
&= (C_x-\rho C_y)C_xf_{p+q} \\
\mathrm{RB}(r_2) &= \mathrm{E}\left(\frac{r_2-R}{R}\right) && (3.6)\\
&= (C_x-\rho C_y)C_xf_q \\
\mathrm{RB}(r_3) &= \mathrm{E}\left(\frac{r_3-R}{R}\right) && (3.7)\\
&= (C_xf_{p+q}-\rho C_yf_p)C_x
\end{aligned}$$

$$\begin{aligned} \mathrm{RB}(r_4) &= \mathrm{E}\left(\frac{r_4 - R}{R}\right) \\ &= (C_x f_q - \rho C_y f_p) C_x \end{aligned} \tag{3.8}$$

It is clear from the above expressions that all the estimators are generally biased like the traditional ratio estimator.

For the comparison of the magnitudes of biases, we consider the squared expressions for the relative biases to the given order of approximation.

We first observe that the estimator r_1 has always larger amount of bias in comparison to r_2. The same is true for r_1 when compared with r_3 provided that the correlation coefficient ρ is negative. For positive values of ρ, this result continues to remain true provided that the following condition is satisfied:

$$\rho > 2\theta \left(1 + \frac{f_p}{f_{p+q}}\right)^{-1} \tag{3.9}$$

where $\theta = (C_x/C_y)$.

The opposite is true, i.e., r_1 has smaller bias in magnitude in comparison to r_3 when

$$0 < \rho < 2\theta \left(1 + \frac{f_p}{f_{p+q}}\right)^{-1}. \tag{3.10}$$

Similarly, if we compare r_1 with r_4, we observe from (3.5) and (3.8) that r_1 has larger bias in magnitude than r_4 when ρ is negative. The same result remains true for positive values of ρ satisfying the following constraint:

$$\left[\left(\frac{f_{p+q}^2 - f_p^2}{f_{p+q}^2 - f_p f_q}\right)\rho^2 - 2\theta\rho + \left(\frac{f_{p+q}^2 - f_q^2}{f_{p+q}^2 - f_p f_q}\right)\theta^2\right] > 0. \tag{3.11}$$

Next, comparing the estimators r_2 and r_3 which utilize the additional available observations only partially, we find the magnitude of bias of r_2 is smaller (larger) than that of r_3 when the quantity G is positive (negative) where G is defined by

$$G = (f_p^2 - f_q^2)\rho^2 - 2(f_{p+q}f_p - f_q^2)\theta\rho + (f_{p+q}^2 - f_q^2). \tag{3.12}$$

Finally, let us compare r_4 with r_2 and r_3.

From (3.6) and (3.8), we observe that r_4 has smaller magnitude of bias in comparison to r_2 when any one of the following conditions is satisfied

$$f_p > f_q \quad \text{and} \quad 0 < \rho < 2\theta \left(1 + \frac{f_p}{f_q}\right)^{-1} \tag{3.13}$$

$$f_p < f_q \quad \text{and} \quad \rho > 2\theta \left(1 + \frac{f_p}{f_q}\right)^{-1} \tag{3.14}$$

$$f_p < f_q \quad \text{and} \quad \rho < 0. \tag{3.15}$$

Similarly, it follows from (3.7) and (3.8) that r_4 has smaller magnitude of bias than r_3 when

$$\rho < \left(\frac{f_{p+q} + f_q}{2f_p}\right)\theta \tag{3.16}$$

which is obviously satisfied at least as long as the correlation coefficient is negative.

Next, let us compare the estimators with respect to the criterion of mean squared error. For this purpose, we have the following results derived in Appendix.

Theorem II: When sample size is large, the relative mean squared errors of the four estimators of R are approximated by

$$\begin{aligned}
\mathrm{RMSE}(r_1) &= \mathrm{E}\left(\frac{r_1 - R}{R}\right)^2 && (3.17)\\
&= (C_x^2 + C_y^2 - 2\rho C_x C_y) f_{p+q} \\
\mathrm{RMSE}(r_2) &= \mathrm{E}\left(\frac{r_2 - R}{R}\right)^2 && (3.18)\\
&= (C_x - 2\rho C_y) C_x f_q + C_y^2 f_{p+q} \\
\mathrm{RMSE}(r_3) &= \mathrm{E}\left(\frac{r_3 - R}{R}\right)^2 && (3.19)\\
&= (C_y - 2\rho C_x) C_y f_p + C_x^2 f_{p+q} \\
\mathrm{RMSE}(r_4) &= \mathrm{E}\left(\frac{r_4 - R}{R}\right)^2 && (3.20)\\
&= (C_y - 2\rho C_x) C_y f_p + C_x^2 f_q .
\end{aligned}$$

It is seen from (3.17), (3.18) and (3.19) that both the estimators r_2 and r_3 (which utilize the incomplete observations on either X or Y) are more efficient than r_1 (which ignores the incomplete observations all together) at least as long as the correlation coefficient ρ between X and Y is negative or zero. If ρ is positive, the estimator r_2 continues to be more efficient than r_1 provided that

$$\rho < \frac{\theta}{2} \tag{3.21}$$

while r_3 remains more efficient than r_1 when

$$\rho < \frac{1}{2\theta}. \tag{3.22}$$

We thus observe that utilizing the incomplete data partially (i.e., observations on either X or Y characteristics) is a better proposition than ignoring them completely so long as ρ is negative or zero. This result carries over for positive values of ρ too, at least as long as

$$\rho < \frac{1}{2} \min\left(\theta, \frac{1}{\theta}\right). \tag{3.23}$$

On the other hand, it may not be worthwhile using the incomplete data set so long as

$$\rho > \frac{1}{2} \max\left(\theta, \frac{1}{\theta}\right) \tag{3.24}$$

which follows from the observation that r_1 is better than r_2 in case $2\rho > \theta$ and better than r_3 in case $2\rho\theta > 1$.

Comparing r_2 and r_3, we find that r_2 is more efficient than r_3 when

$$\rho \lessgtr \Phi \quad \text{if} \quad f_q \lessgtr f_p \tag{3.25}$$

while the opposite is true, i.e., the estimator r_2 is less efficient than r_3 when

$$\rho \gtrless \Phi \quad \text{if} \quad f_q \gtrless f_p \tag{3.26}$$

where

$$\Phi = \frac{(f_{p+q} - f_p)}{2\theta(f_q - f_p)} \left[1 - \left(\frac{f_{p+q} - f_q}{f_{p+q} - f_p}\right)\theta^2\right]. \tag{3.27}$$

The above statement provides conditions under which use of incomplete observations on X characteristic is superior or inferior than the use of incomplete observations on Y characteristic.

Next, let us consider the estimator r_4 which utilizes all the available observations on both X and Y characteristics. Comparing it with r_1 which is based on complete cases only, we observe from (3.17) and (3.20) that r_4 is more efficient than r_1 when the correlation coefficient is negative or zero. This result holds for positive values of ρ also provided that

$$\rho < \frac{1}{2\theta}\left[1 + \left(\frac{f_{p+q} - f_q}{f_{p+q} - f_p}\right)\theta^2\right]. \tag{3.28}$$

When the positive value of ρ is such that it satisfies the inequality (3.28) with a reversed sign, r_4 is no more superior to r_1 meaning thereby that it is appropriate to discard the incomplete observations.

Now comparing r_4 with r_2, we find that r_4 is more efficient than r_2 when

$$\rho(f_q - f_p) < \left(\frac{f_{p+q} - f_p}{2\theta}\right). \tag{3.29}$$

This inequality is satisfied when one of the following conditions holds true

(a) $f_q = f_p$

(b) $\rho = 0$

(c) $f_q < f_p$ and $\rho > 0$

(d) $f_q < f_p$ and $\rho < 0$ but $|\rho| < \Psi$

(e) $f_q > f_p$ and $\rho < 0$

(f) $f_q > f_p$ and $0 < \rho < \Psi$

where

$$\Psi = \frac{f_{p+q} - f_p}{2\theta|f_q - f_p|}. \tag{3.30}$$

Under any one of the above conditions, we thus observe that using all the available observations on X and Y is a better strategy than using all the available observations on X only and discarding those for which Y values are available but X values are missing.

Just the opposite is true, i.e., using all those observations for which X values are available is more appropriate than using the entire set of observations when the inequality (3.29) holds true with a reversed sign. This can happen when the magnitude of ρ is greater than Ψ provided that $f_q < f_p$ for negative values of ρ and $f_q > f_p$ for positive values of ρ.

Similarly, comparing the expressions (3.19) and (3.20), it is interesting to find that r_4 is invariably more efficient than r_3 implying that it is always beneficial to use all the available observations—complete or incomplete.

4 Some Remarks

We have considered the problem of estimating the ratio of population means when observations on some selected units in the sample drawn according to simple random sampling without replacement on either X characteristic or Y characteristic but not on both of them simultaneously are missing. Accordingly, we have formulated four simple estimators for the population ratio. The first estimator is essentially the ratio of sample means employing all the complete pairs of observations and ignoring the rest. The second estimator uses the incomplete pairs, in which only X values are available, in addition to complete cases while the third estimator utilizes incomplete pairs, in which only Y values are available, besides the complete cases. The fourth estimator is based on all the complete as well as incomplete pairs of observations. Performance properties of the four estimators are analyzed with respect to the bias and mean squared error criteria using the large sample theory and the conditions are obtained for the superiority of one estimator over the other. Such an exercise has shed light on the role of missingness of observations on the ratio method of estimation.

When the population mean $\bar{X}$ is known one can straightforwardly develop estimators for the population mean $\bar{Y}$, and the role of X characteristic in improving the estimation of $\bar{Y}$ can be examined by comparing the performance properties of these estimators with the estimators $\bar{y}$ and $[(n-p-q)\bar{y} + q\bar{y}^{**}]/(n-p)$. Similar investigations can be carried out when some other procedure like the product method of estimation is followed. Extending the results for more than two characteristics with varying patterns of missingness will be an interesting exercise.

Appendix

Let us define

$$\begin{aligned} u &= \left(\frac{\bar{x}-\bar{X}}{\bar{X}}\right) \\ v &= \left(\frac{\bar{y}-\bar{Y}}{\bar{Y}}\right) \\ u^* &= \left(\frac{\bar{x}^*-\bar{X}}{\bar{X}}\right) \\ v^{**} &= \left(\frac{\bar{y}^{**}-\bar{Y}}{\bar{Y}}\right) \end{aligned}$$

so that

$$\begin{aligned} \left(\frac{r_1-R}{R}\right) &= \frac{v-u}{1+u} \qquad \text{(A.1)} \\ &= (v-u)\left(1-u+\frac{u^2}{1+u}\right) \end{aligned}$$

whence the relative bias and relative mean squared error to order $O(n^{-1})$ only are given by

$$\begin{aligned} \mathrm{RB}(r_1) &= \mathrm{E}(v)-\mathrm{E}(u)+\mathrm{E}(u^2)-\mathrm{E}(uv) \qquad \text{(A.2)} \\ \mathrm{RMSE}(r_1) &= \mathrm{E}(v^2)+\mathrm{E}(u^2)-2\,\mathrm{E}(uv). \qquad \text{(A.3)} \end{aligned}$$

Now we observe that

$$\begin{aligned} \mathrm{E}(v) &= \mathrm{E}_1\,\mathrm{E}(v|p,q) \\ &= 0 \\ \mathrm{E}(v^2) &= \mathrm{E}_1\,\mathrm{E}(v^2|p,q) \\ &= \mathrm{E}_1\left[\left(\frac{1}{n-p-q}-\frac{1}{N}\right)C_y^2\right] \\ &= C_y^2 f_{p+q}. \end{aligned}$$

Similarly, we have

$$\begin{aligned} \mathrm{E}(u) &= 0 \\ \mathrm{E}(u^2) &= C_x^2 f_{p+q} \\ \mathrm{E}(uv) &= \rho C_x C_y f_{p+q} \end{aligned}$$

Substituting these results in (A.2) and (A.3), we obtain the results (3.5) and (3.17).

Proceeding in the same manner, we can express

$$\left(\frac{r_2-R}{R}\right) = \left[v-\frac{(n-p-q)u+pu^*}{(n-q)}\right]\left[1-\frac{(n-p-q)u+pu^*}{(n-q)}\right]+O_p(n^{-\frac{3}{2}}) \qquad \text{(A.4)}$$

whence up to order $O(n^{-1})$, we have

$$\begin{aligned}
\mathrm{RB}(r_2) &= \mathrm{E}(v) - \mathrm{E}\left[\frac{(n-p-q)u+pu^*}{(n-q)}\right] + \ E\left[\frac{(n-p-q)u+pu^*}{(n-q)}\right]^2 \\
&\quad - \mathrm{E}\left[\left(1-\frac{p}{n-q}\right)uv\right] - \mathrm{E}\left[\left(\frac{p}{n-q}\right)u^*v\right] \\
&= C_x^2 f_q - \rho C_x C_y \,\mathrm{E}_1\left[\left(1-\frac{p}{n-q}\right)\left(\frac{1}{n-p-q}-\frac{1}{N}\right)\right] \\
&= C_x^2 f_q - \rho C_x C_y f_q \qquad \text{(A.5)}
\end{aligned}$$

to the order of our approximation. This leads to the result (3.6).

Similarly, from (A.4), we have

$$\begin{aligned}
\mathrm{RMSE}(r_2) &= \mathrm{E}\left[v - \left(1-\frac{p}{n-q}\right)u - \left(\frac{p}{n-q}\right)u^*\right] \\
&= \mathrm{E}(v^2) + \mathrm{E}\left[\left(1-\frac{p}{n-q}\right)^2 u^2\right] + \mathrm{E}\left[\left(\frac{p}{n-q}\right)^2 u^{*2}\right] \\
&\quad -2\,\mathrm{E}\left[\left(1-\frac{p}{n-q}\right)uv\right] - 2\,\mathrm{E}\left[\left(\frac{p}{n-q}\right)u^*v\right] \\
&\quad +2\,\mathrm{E}\left[\left(1-\frac{p}{n-q}\right)\left(\frac{p}{n-q}\right)u^*v\right] \\
&= C_y^2 f_{p+q} + C_x^2\,\mathrm{E}_1\left[\left(1-\frac{p}{n-q}\right)\left(\frac{1}{n-p-q}-\frac{1}{N}\right)\right] \\
&\quad + C_x^2\,\mathrm{E}_1\left[\left(\frac{p}{n-q}\right)^2\left(\frac{1}{p}-\frac{1}{N}\right)\right] \\
&\quad -2\rho C_x C_y\,\mathrm{E}_1\left[\left(1-\frac{p}{n-q}\right)\left(\frac{1}{n-p-q}-\frac{1}{N}\right)\right] \\
&= C_y^2 f_{p+q} + C_x^2 f_q - 2\rho C_x C_y f_q \qquad \text{(A.6)}
\end{aligned}$$

dropping the terms with higher order order of smallness than n^{-1}. This provides the result (3.18).

Next, observing that the relative estimation error of r_3 to order $O_p(n^{-1})$ is given by

$$\left(\frac{r_3-R}{R}\right) = \left[\frac{(n-p-q)v+qv^{**}}{(n-p)} - u\right](1-u) \qquad \text{(A.7)}$$

the results (3.7) and (3.19) can be easily derived in the same way as indicated in case of r_2.

Finally, let us consider the relative estimation error of r_4 to order $O_p(n^{-1})$ which can be expressed as

$$\left(\frac{r_4-R}{R}\right) = \qquad \text{(A.8)}$$

$$\left[\frac{(n-p-q)v+qv^{**}}{(n-p)} - \frac{(n-p-q)u+pu^*}{(n-q)}\right]\left[1-\frac{(n-p-q)u+pu^*}{(n-q)}\right]$$

from which the expressions (3.8) and (3.20) for the relative bias and the relative mean squared error to order $O(n^{-1})$ can be easily found.

References

Cochran, W. G. (1977). *Sampling Techniques*, Wiley, New York.

Sukhatme, P. V., Sukhatme, B. V. and Sukhatme, S. (1984). *Sampling Theory Of Surveys With Applications*, Iowa State Univercity Press, Iowa.

Tracy, D. S. and Osahan, S. S. (1994). Random non-response on study variable versus on study as well as auxiliary variables, *Statistica* **54**: 163–168.