LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK
SONDERFORSCHUNGSBEREICH 386

Toutenburg, Srivastava:

Amputation versus Imputation of Missing Values through Ratio Method in Sample Surveys

Sonderforschungsbereich 386, Paper 155 (1999)

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

Projektpartner

Amputation versus Imputation of Missing Values through Ratio Method in Sample Surveys

H. Toutenburg* V.K. Srivastava§

May 27, 1999

Abstract

In this article, we consider the estimation of population mean when some observations on the study characteristic are missing in the bivariate sample data. In all, five estimators are presented and their efficiency properties are discussed. One estimator arises from the the amputation of incomplete observations while the remaining four estimators are formulated using inputed values obtained by the ratio method of estimation.

1 Introduction

Infeasibility to have all the observations in the sample is not an uncommon aspect of data collection in many instances of sample surveys. This may occur due to a variety of reasons. For example, a break-down or some snag may arise in the instrument and/or measuring device rendering it unusable for completing the process of data collection. Subjects like patients, animals and plants may fail to survive due to factors that are unrelated to the experiment. Often typical practical difficulties are faced in the collection of data for a part of the sample. Sometimes the respondents may supply information which is inconsistent due to some inner contradictions or otherwise, and the investigator is forced to delete it.

When some observations in the sample are missing, the simplest solution is perhaps to amputate the incomplete observations and to restrict attention to complete observations only for the purpose of statistical analysis. Alternatively, one may employ some imputation method for finding the substitutes of missing observations; see, e.g., Little and Rubin (1987), Rao and Toutenburg (1995) and Rubin (1987) for an interesting account. Treating these imputed values as true observations, one may conduct the statistical analysis using the standard procedures developed for data without any missing observation. Such a practice, it is well recognized, may tend to invalidate the inferences and may often have serious consequences.

*Institute of Statistics, University of Munich, 80799 Munich, Germany
§Department of Statistics, University of Lucknow, Lucknow 226007, India

In this article, we consider the estimation of population mean on the basis of a random sample drawn according to the procedure of simple random sampling without replacement. It is assumed, following Rao and Sitter (1995) and Tracy and Osahan (1994), that some units in the sample fail to respond and the observations on the study characteristics are not available while this is not the case with the auxiliary characteristic on which all the observations in the sample are available. For the missing values on the study characteristic, the method of ratio imputation is a commonly employed procedure in sample surveys. Using it, we have considered four estimators for the population mean of study characteristic besides the conventional estimator (i.e., the mean of available observations) which amputates the incomplete observations. Comparing their efficiency properties, it is observed that outright amputation is not a good proposition and use of ratio imputation is worthwhile. It helps in improving the efficiency of estimation under some mild constraints.

The plan of this article is as follows. In Section 2, we describe the imputation procedure and present estimators for the population mean. Bias properties of these estimators are studied in Section 3. Similarly, their mean squared errors are analyzed in Section 4 and conditions for the superiority of one estimator over the other are found. Lastly, the derivation of results is provided in Appendix.

2 Estimators For Mean

Let us consider a finite population of size N with values $Y_1, Y_2, \ldots, Y_N$ of the study characteristic and values $X_1, X_2, \ldots, X_N$ of the auxiliary characteristic. For the estimation of population mean $\bar{Y}$, a random sample of size n is drawn according to the procedure of simple random sampling without replacement. Assuming the nonresponse to be random, suppose that there are $(n-p)$ complete observations (y_1, x_1), (y_2, x_2), $\ldots$, (y_{n-p}, x_{n-p}) and p incomplete observations $x_1^*, x_2^*, \ldots, x_p^*$. Thus the sample comprises two respondent sets—one of size $(n-p)$ denoted by s and the other of size p denoted by s^*.

When the incomplete observations are discarded, it is customary to estimate $\bar{Y}$ by

$$\bar{y} = \frac{1}{n-p} \sum_{i=1}^{n-p} y_i. \tag{2.1}$$

When the incomplete observations are not discarded and some imputation method is followed, the completed data set is specified by

$$z_i = \begin{cases} y_i & \text{if } i \in s \\ \tilde{y}_i & \text{if } i \in s^* \end{cases} \tag{2.2}$$

and the population mean is estimated by

$$\begin{aligned} t &= \frac{1}{n}\sum_{i=1}^{n} z_i \\ &= \frac{1}{n}\left(\sum_{i\in s} y_i + \sum_{i\in s*} \tilde{y}_i\right) \end{aligned} \tag{2.3}$$

where $\tilde{y}_i$ denotes the imputed value of the study characteristic corresponding to the observation x_i^*.

If the method of ratio imputation is employed, there are two simple choices of $\tilde{y}_i$, viz.,

$$\tilde{y}_i = \bar{y}\left(\frac{\bar{X}}{\bar{x}}\right) \tag{2.4}$$

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

where $\bar{x} = \frac{1}{n-p}\sum_{i=1}^{n-p} x_i$, $\bar{x}^* = \frac{1}{p}\sum_{i=1}^{p}$ and $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$.

In the above two formulations, it is assumed that $\bar{X}$ is known. If it is not known, we may define the imputed values as

$$\tilde{y}_i = \bar{y}\left(\frac{x_i^*}{\bar{x}}\right) \tag{2.6}$$

following Rao and Sitter (1995, p. 459).

On the same lines, we propose another set of imputed values as follows:

$$\tilde{y}_i = \bar{y}\left(\frac{nx_i^*}{(n-p)\bar{x} + p\bar{x}^*}\right). \tag{2.7}$$

Utilizing (2.4) – (2.7) in (2.3), we obtain the following four estimators of $\bar{Y}$:

$$t_1 = \bar{y}\left[\frac{(n-p)\bar{x} + p\bar{X}}{n\bar{x}}\right] \tag{2.8}$$

$$t_2 = \bar{y}\left[\frac{(n-p)^2\bar{x} + np\bar{X} + (n-p)p\bar{x}^*}{(n-p)n\bar{x} + np\bar{x}^*}\right] \tag{2.9}$$

$$t_3 = \bar{y}\left[\frac{(n-p)\bar{x} + p\bar{x}^*}{n\bar{x}}\right] \tag{2.10}$$

$$t_4 = \bar{x}\left[\frac{(n-p)^2\bar{x} + (2n-p)p\bar{x}^*}{(n-p)n\bar{x} + np\bar{x}^*}\right]. \tag{2.11}$$

Thus we have five estimators for estimating the population mean $\bar{Y}$. The estimator $\bar{y}$ is based on amputation of incomplete data while the estimators t_1, t_2, t_3 and t_4 are based on ratio imputation of missing observations. Out of these four, two estimators require the knowledge of the population mean $\bar{X}$ of the auxiliary characteristic while the remaining two estimators are free from it.

3 Comparison Of Biases

Let us write

$$\begin{aligned} S_x^2 &= \tfrac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2, \\ S_y^2 &= \tfrac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2, \\ \rho &= \tfrac{1}{S_xS_y(N-1)}\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y}), \\ \theta &= \tfrac{\bar{Y}S_x}{\bar{X}S_y}, \\ f &= \left(\tfrac{1}{n}-\tfrac{1}{N}\right)\mathrm{E}_p\left(\tfrac{p}{n}\right), \\ g &= \mathrm{E}_p\left(\tfrac{p}{n}\right)\left(\tfrac{1}{n-p}-\tfrac{1}{N}\right) \end{aligned} \tag{3.1}$$

where E_p denotes the expectation with respect to the nonnegative integer valued random variable p. Further, we assume that the correlation coefficient ρ is nonnegative which is a basic requirement for the application of ratio method.

It is easy to see that the mean $\bar{y}$ ignoring the incomplete observations is an unbiased estimator of $\bar{Y}$ while the estimators t_1, t_2, t_3 and t_4 using the ratio method of imputation for missing values are generally biased. In order to study the magnitudes and directions of their biases, we assume that p is small and $(n-p)$ is large which implies that n is large. Now let us consider the large sample approximations which are derived in Appendix following Sukhatme, Sukhatme, Sukhatme and Asok (1984).

Theorem 1 *The order $O(n^{-2})$ approximations for the biases of the estimators t_1, t_2, t_3 and t_4 are given by*

$$\begin{aligned} B(t_1) &= \mathrm{E}(t_1-\bar{Y}) \qquad (3.2)\\ &= g(\theta-\rho)\frac{S_xS_y}{\bar{X}} \\ B(t_2) &= \mathrm{E}(t_2-\bar{Y}) \qquad (3.3)\\ &= f(\theta-\rho)\frac{S_xS_y}{\bar{X}} \\ B(t_3) &= \mathrm{E}(t_3-\bar{Y}) \qquad (3.4)\\ &= g(\theta-\rho)\frac{S_xS_y}{\bar{X}} \\ B(t_4) &= \mathrm{E}(t_4-\bar{Y}) \qquad (3.5)\\ &= g\left[\left(1-\frac{f}{g}\right)\theta-\rho\right]\frac{S_xS_y}{\bar{X}}. \end{aligned}$$

From the above expressions, it is interesting to observe that all the four estimators are unbiased to order $O(n^{-1})$ and the bias precipitates in the terms of order $O(n^{-2})$. However, the estimators t_1, t_2 and t_3 are also unbiased to order $O(n^{-2})$ when $\theta=\rho$. If θ is not less than 1, these three estimators are biased in positive direction. The bias continues to remain positive so long as $\rho<\theta<1$. It changes its sign only when $\rho<\theta$. So far as the estimator t_4 is concerned, it is also unbiased to order $O(n^{-2})$ when $(f-g)\theta=g\rho$. Its bias is positive or negative according as $(f-g)\theta$ is larger or smaller than $g\rho$.

Comparing the estimators with respect to the criterion of magnitude of bias, we find that the estimators t_1 and t_3 have an equal amount of bias at least to the order of our approximation. Further, t_2 has always smaller bias than t_1 and t_3 as f cannot exceed g. Similarly, the estimator t_2 has a smaller magnitude of bias in comparison to the estimator t_4 when

$$[(g^2 - f^2)(\theta - \rho)^2 + f^2\theta^2 - 2fg\theta(\theta - \rho)] > 0 \tag{3.6}$$

while the reverse is true, i.e., t_4 is less biased in magnitude than t_2 when the inequality (3.6) holds with an opposite sign.

Observing that

$$[B(t_1)]^2 - [B(t_4)]^2 = [B(t_3)]^2 - [B(t_4)]^2 = fg\theta(\theta - 2\rho)\left(\frac{S_x S_y}{\bar{X}}\right)^2 \tag{3.7}$$

we see that t_4 has smaller magnitude of bias than t_1 and t_3 when either $\theta > 2\rho$. On the contrary, the estimators t_1 and t_3 are less biased in magnitude in comparison to t_4 when θ is less than 2ρ.

4 Comparison Of Mean Squared Errors

Recalling that $\bar{y}$ is an unbiased estimator of $\bar{Y}$, its variance is given by

$$\begin{aligned} \mathrm{V}(\bar{y}) &= \mathrm{E}(\bar{y} - \bar{Y})^2 \\ &= \left[\mathrm{E}_p\left(\frac{1}{n-p}\right) - \frac{1}{N}\right] S_y^2. \end{aligned} \tag{4.1}$$

As the estimators t_1, t_2, t_3 and t_4 are generally not unbiased, we consider their mean squared errors for the purpose of comparison. These are derived In Appendix and presented below.

Theorem 2 *To order $O(n^{-2})$, the differences between the variance of $\bar{y}$ and the mean squared errors of the estimators t_1, t_2, t_3 and t_4 are given by*

$$\begin{aligned} \Delta(\bar{y}; t_1) &= \mathrm{E}(\bar{y} - \bar{Y})^2 - \mathrm{E}(t_1 - \bar{Y})^2 \\ &= 2g\rho S_x S_y \left(\frac{\bar{Y}}{\bar{X}}\right) \end{aligned} \tag{4.2}$$

$$\begin{aligned} \Delta(\bar{y}; t_2) &= \mathrm{E}(\bar{y} - \bar{Y})^2 - \mathrm{E}(t_2 - \bar{Y})^2 \\ &= 2f\rho S_x S_y \left(\frac{\bar{Y}}{\bar{X}}\right) \end{aligned} \tag{4.3}$$

$$\begin{aligned} \Delta(\bar{y}; t_3) &= \mathrm{E}(\bar{y} - \bar{Y})^2 - \mathrm{E}(t_3 - \bar{Y})^2 \\ &= (2g\rho - f\theta) S_x S_y \left(\frac{\bar{Y}}{\bar{X}}\right) \end{aligned} \tag{4.4}$$

$$\begin{aligned} \Delta(\bar{y}; t_4) &= \mathrm{E}(\bar{y} - \bar{Y})^2 - \mathrm{E}(t_4 - \bar{Y})^2 \\ &= (2g\rho - f\theta) S_x S_y \left(\frac{\bar{Y}}{\bar{X}}\right). \end{aligned} \tag{4.5}$$

As ρ is assumed to be positive, it is clear from (4.2) and (4.3) that both the estimators t_1 and t_2 are better than $\bar{y}$ implying the superiority of imputation over amputation.

Looking at the expressions (4.4) and (4.5), we find that the estimators t_3 and t_4 are better than $\bar{y}$ when

$$2\rho > \left(\frac{f}{g}\right)\theta. \tag{4.6}$$

As $f < g$, this condition is satisfied as long as $2\rho > \theta$ which is the well-known condition for the superiority of ratio estimator over the sample mean when no observation is missing; see, e.g., Sukhatme et al. (1984, Chap. 5). Thus, so long as the favourable environment for the application of ratio method prevails (i.e., $2\rho > \theta$), the missingness of some observations on the study characteristic and their imputation by ratio method do not exert any adverse effect. In fact, the ratio imputation suceeds in widening the range of admissible values of ρ; see (4.6).

Next, let us compare the biased estimators.

When $\bar{X}$ is known, we have two estimators t_1 and t_2 out of which t_1 ignores the incomplete observations while t_2 incorparates them. Further, t_2 has always smaller magnitude of bias in comparison to t_1; see (3.2) and (3.3). If we compare their mean squared errors, it is seen from (4.2) and (4.3) that the estimator t_1 has smaller mean squared error than t_2.

When $\bar{X}$ is not known, we have again two estimators t_3 and t_4 which utilize the entire set of available observations. Further, t_3 has smaller (larger) magnitude of bias than t_4 when 2ρ is greater (less) than θ. Comparing them with respect to the criterion of mean squared error to order $O(n^{-2})$, we observe from (4.4) and (4.5) that both are equally efficient to the given order of approximation. However, the difference precipitates if we consider higher order approximations.

Theorem 3 *To order $O(n^{-3})$, we have*

$$\begin{aligned} \Delta(t_3; t_4) &= \mathrm{E}(t_3 - \bar{Y})^2 - \mathrm{E}(t_4 - \bar{Y})^2 \\ &= Q\left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{\bar{Y}}{\bar{X}}\right)^2 \mathrm{E}_p\left(\frac{p}{n}\right)^2 \end{aligned} \tag{4.7}$$

where

$$Q = \frac{1}{N-1}\sum_{i=1}^{N} X_i(X_i - \bar{X})^2. \tag{4.8}$$

We thus find that the estimator t_4 is better than t_3 when Q is positive which may generally hold good in many practical situations.

Finally, let us examine the role of knowledge of $\bar{X}$ through a comparison of estimators t_1 and t_2 with t_3 and t_4.

Let us first recall that t_1 has the same bias as t_3 but it is less biased in magnitude than t_4 for $2\rho > \theta$. Further, it is observed from (4.2), (4.4) and (4.5) that the estimator t_1 has invariably smaller mean squared error than t_3 and t_4. Similarly, we observe from (4.3), (4.4) and (4.5) that the estimator t_2 is more efficient than both the estimators t_3 and t_4. This means that the knowledge of $\bar{X}$ plays an important role in improving the efficiency of estimation when some observations are missing and the method of ratio imputation is employed for them.

APPENDIX

If we write

$$\epsilon_x = (\bar{x} - \bar{X}), \quad \epsilon_x^* = (\bar{x}^* - \bar{X}), \quad \epsilon_y = (\bar{y} - \bar{Y})$$

we observe that ϵ_x and ϵ_y are of order $O_p(n^{-\frac{1}{2}})$ while ϵ_x^* is of order $O_p(1)$. Further, we have

$$\mathrm{E}(\epsilon_x) = \mathrm{E}(\epsilon_x^*) = \mathrm{E}(\epsilon_y) = 0.$$

Now we can express

$$\begin{aligned}(t_1 - \bar{Y}) &= (\bar{y} - \bar{Y}) - \frac{p\bar{y}}{n}\left(1 - \frac{\bar{X}}{\bar{x}}\right) \\ &= \epsilon_y - \frac{p}{n\bar{X}}\epsilon_x(\bar{Y} + \epsilon_y)\left(1 + \frac{\epsilon_x}{\bar{X}}\right)^{-1}.\end{aligned}$$

Expanding and retaining terms to order $O_p(n^{-2})$, we find

$$(t_1 - \bar{Y}) = \epsilon_y - \frac{p\bar{Y}}{n\bar{X}}\epsilon_x - \frac{p}{n\bar{X}}\left(\epsilon_x\epsilon_y - \frac{\bar{Y}}{\bar{X}}\epsilon_x^2\right)$$

whence

$$\begin{aligned}\mathrm{E}(t_1 - \bar{Y}) &= \mathrm{E}(\epsilon_y) - \frac{\bar{Y}}{n\bar{X}}\mathrm{E}(p\epsilon_x) - \frac{1}{n\bar{X}}\left[\mathrm{E}(p\epsilon_x\epsilon_y) - \frac{\bar{Y}}{\bar{X}}\mathrm{E}(p\epsilon_x^2)\right] \\ &= -\frac{1}{\bar{X}}\left[g\rho S_x S_y - \frac{\bar{Y}}{\bar{X}}gS_x^2\right]\end{aligned}$$

which leads to the result (3.2) of Theorem 1.

Similarly, to order $O(n^{-2})$, we have

$$\begin{aligned}\mathrm{E}(t_1 - \bar{Y})^2 &= \mathrm{E}\left(\epsilon_y^2 - \frac{2p\bar{Y}}{n\bar{X}}\epsilon_y\epsilon_x\right) \\ &= \left[\mathrm{E}_p\left(\frac{1}{n-p}\right) - \frac{1}{N}\right]S_y^2 - 2\left(\frac{\bar{Y}}{\bar{X}}\right)g\rho S_x S_y\end{aligned}$$

giving the expression (4.2) of Theorem 2.

Next, we observe that

$$\begin{aligned}
(t_2 - \bar{Y}) &= (\bar{y} - \bar{Y}) - \frac{p\bar{y}}{n}\left[\frac{(n-p)(\bar{x}-\bar{X}) + p(\bar{x}^* - \bar{X})}{(n-p)\bar{x} + p\bar{x}^*}\right] \\
&= \epsilon_y - \frac{p}{n\bar{X}}(\bar{Y} + \epsilon_y)\left[\frac{(n-p)\epsilon_x + p\epsilon_x^*}{n}\right]\left[1 + \frac{(n-p)\epsilon_x + p\epsilon_x^*}{n\bar{X}}\right]^{-1} \\
&= \epsilon_y - \frac{p}{n\bar{X}}\left[\bar{Y}\epsilon_x + \left(\epsilon_y\epsilon_x + \frac{p\bar{Y}}{n}\epsilon_x^*\right) + \ldots\right]\left[1 - \frac{\epsilon_x}{\bar{X}} + \ldots\right] \\
&= \epsilon_y - \frac{p\bar{Y}}{n\bar{X}}\epsilon_x - \frac{p}{n\bar{X}}\left[\left(\epsilon_y - \frac{\bar{Y}}{\bar{X}}\epsilon_x\right)\epsilon_x + \frac{p\bar{y}}{n}\epsilon_x^*\right] + O_p(n^{-\frac{5}{2}}).
\end{aligned}$$

Thus the bias to order $O(n^{-2})$ is

$$\mathrm{E}(t_2 - \bar{Y}) = -\frac{p}{n\bar{X}}\left[f\rho S_x S_y - \left(\frac{\bar{Y}}{\bar{X}}\right) f S_x^2\right]$$

and the mean squared error to the same order of approximation is

$$\begin{aligned}
\mathrm{E}(t_2 - \bar{Y})^2 &= \mathrm{E}(\epsilon_y^2) - 2\left(\frac{\bar{Y}}{\bar{X}}\right)\mathrm{E}\left[\frac{p(n-p)}{n^2}\epsilon_x\epsilon_y + \frac{p^2}{n^2}\epsilon_x^*\epsilon_y\right] \\
&= \left[\mathrm{E}_p\left(\frac{1}{n-p}\right) - \frac{1}{N}\right] S_y^2 - 2\left(\frac{\bar{Y}}{\bar{X}}\right) f\rho S_x S_y.
\end{aligned}$$

These provide the result (3.3) of Theorem 1 and result (4.3) of Theorem 2.
Similarly, for the estimator t_3, we have

$$\begin{aligned}
(t_3 - \bar{Y}) &= (\bar{y} - \bar{Y}) - \frac{p\bar{y}}{n}\left(\frac{\bar{x} - \bar{x}^*}{\bar{x}}\right) \\
&= \epsilon_y - \frac{p}{n\bar{X}}(\bar{Y} + \epsilon_y)(\epsilon_x - \epsilon_x^*)\left(1 + \frac{\epsilon_x}{\bar{X}}\right)^{-1} \\
&= \epsilon_y - \frac{p}{n\bar{X}}[-\bar{Y}\epsilon_x^* + (\bar{Y}\epsilon_x - \epsilon_x^*\epsilon_y) + \epsilon_x\epsilon_y]\left(1 - \frac{\epsilon_x}{\bar{X}} + \frac{\epsilon_x^2}{\bar{X}^2}\ldots\right) \\
&= \epsilon_y + \frac{p\bar{Y}}{n\bar{X}}\epsilon_x^* - \frac{p}{n\bar{X}}\left[\bar{Y}\epsilon_x - \left(\epsilon_y - \frac{\bar{Y}}{\bar{X}}\epsilon_x\right)\epsilon_x^*\right] \\
&\quad - \frac{p}{n\bar{X}}\left(\epsilon_y - \frac{\bar{Y}}{\bar{X}}\epsilon_x\right)\left(1 + \frac{\epsilon_x^*}{\bar{X}}\right)\epsilon_x + O_p(n^{-\frac{5}{2}}).
\end{aligned}$$

Taking expectation and retaining terms upto order $O(n^{-2})$, we get

$$\mathrm{E}(t_3 - \bar{Y}) = -\frac{1}{\bar{X}}\left[g\rho S_x S_y - \left(\frac{\bar{Y}}{\bar{X}}\right) g S_x^2\right].$$

which gives the result (3.4) of Theorem 1.

Similarly, to the same order of approximation, we have

$$\begin{aligned}
\mathrm{E}(t_3 - \bar{Y})^2 &= \mathrm{E}(\epsilon_y^2) + 2\left(\frac{\bar{Y}}{n\bar{X}}\right)\mathrm{E}(p\epsilon_y\epsilon_x^*) + \left(\frac{\bar{Y}}{n\bar{X}}\right)^2 \mathrm{E}(p^2{\epsilon_x^*}^2) \\
&\quad - \frac{2}{n\bar{X}}\left[\bar{Y}\,\mathrm{E}(p\epsilon_x\epsilon_y) - \mathrm{E}(p\epsilon_x^*\epsilon_y^2) - \left(\frac{\bar{Y}}{\bar{X}}\right)\mathrm{E}(p\epsilon_x^*\epsilon_x\epsilon_y)\right] \\
&= \left[\mathrm{E}_p\left(\frac{1}{n-p}\right) - \frac{1}{N}\right] S_y^2 + \left(\frac{\bar{Y}}{\bar{X}}\right)^2 f S_x^2 - 2\left(\frac{\bar{Y}}{\bar{X}}\right) g\rho S_x S_y
\end{aligned}$$

which leads to the result (4.4) of Theorem 2.

Proceeding in the same manner, we can express

$$\begin{aligned}
(t_4 - \bar{Y}) &= (\bar{y} - \bar{Y}) - \frac{p\bar{y}}{n}\left[\frac{(n-p)(\bar{x} - \bar{x}^*)}{(n-p)\bar{x} + p\bar{x}^*}\right] \\
&= \epsilon_y - \frac{p}{n^2\bar{X}}(\bar{Y} + \epsilon_y)(n-p)(\epsilon_x - \epsilon_x^*)\left[1 + \frac{\epsilon_x}{\bar{X}} + \frac{p\epsilon_x^*}{n\bar{X}} - \frac{p\epsilon_x}{n\bar{X}}\right]^{-1} \\
&= \epsilon_y - \frac{p}{n\bar{X}}\left[-\bar{Y}\epsilon_x^* + (\bar{Y}\epsilon_x - \epsilon_y\epsilon_x^*) + \left(\epsilon_x\epsilon_y + \frac{p\bar{Y}}{n}\epsilon_x^*\right) + \ldots\right] \\
&\qquad \left[1 - \frac{\epsilon_x}{\bar{X}} + \left(\frac{\epsilon_x^2}{\bar{X}} - \frac{p\epsilon_x^*}{n\bar{X}}\right) + \ldots\right] \\
&= \epsilon_y + \frac{p\bar{Y}}{n\bar{X}}\epsilon_x^* - \frac{p}{n\bar{X}}\left[\bar{Y}\epsilon_x - \left(\epsilon_y - \frac{\bar{Y}}{\bar{X}}\right)\epsilon_x^*\right] \\
&\quad -\frac{p}{n\bar{X}}\left[\left(\epsilon_y - \frac{\bar{Y}}{\bar{X}}\epsilon_x\right)\epsilon_x + \frac{p\bar{Y}}{n}\epsilon_x^*\right]\left(1 + \frac{\epsilon_x^*}{\bar{X}}\right) + O_p(n^{-\frac{5}{2}}).
\end{aligned}$$

We thus find

$$\mathrm{E}(t_4 - \bar{Y}) = -\frac{1}{\bar{X}}\left[g\rho S_x S_y - (g-f)\left(\frac{\bar{Y}}{\bar{X}}\right)S_x^2\right]$$

which is the result (3.5) of Theorem 1.

The result (4.5) of Theorem 2 can be obtained in a similar way.

Lastly, let us consider the result stated in Theorem 3.

It is observed that

$$\begin{aligned}
\Delta(t_3; t_4) &= \mathrm{E}(t_3 - \bar{Y})^2 - \mathrm{E}(t_4 - \bar{Y})^2 \\
&= \mathrm{E}[(t_3 - \bar{Y}) + (t_4 - \bar{Y})][(t_3 - \bar{Y}) - (t_4 - \bar{Y})] \\
&= 2\,\mathrm{E}\left(\epsilon_y + \frac{p\bar{y}}{n\bar{X}}\epsilon_x^*\right)\left(1 + \frac{\epsilon_x^*}{\bar{X}}\right)\frac{p^2\bar{Y}\epsilon_y^*}{n^2\bar{X}} + O(n^{-\frac{7}{2}}) \\
&= \left(\frac{1}{n} - \frac{1}{N}\right)\left(\frac{\bar{Y}}{\bar{X}}\right)^2\left[S_x^2 + \frac{1}{\bar{X}(N-1)}\sum_{i=1}^{N}(X_i - \bar{X})^3\right]\mathrm{E}_p\left(\frac{p}{n}\right)^2 \\
&\quad + O(n^{-\frac{7}{2}})
\end{aligned}$$

which yields the desired result (4.7).

References

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.

Rao, C. R. and Toutenburg, H. (1995). *Linear Models: Least Squares and Alternatives*, Springer, New York.

Rao, J. and Sitter, R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data, *Biometrika* **82**: 453–460.

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

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. and Asok, C. (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.