

RATIO ESTIMATION IN POSTSTRATIFIED SAMPLING OVER TWO OCCASIONS

A. C. ONYEKA

(Received 30 August 2002; Revision accepted 15 October, 2002).

ABSTRACT

The theory of ratio estimation in a one-time poststratified sampling is extended to ratio estimation in poststratified sampling over two occasions. Two ratio-type double sampling estimators are proposed for the estimation of the second occasion population ratio (R_2) in poststratified sampling over two occasions. One of the proposed estimators, e_1 , assumes knowledge of the first occasion population ratio (R_1) while the second estimator, e_2 , does not make any such assumption. Both estimators are based on a partial replacement of sample units on the second occasion. Conditions under-which one estimator is to be preferred to the other are obtained for repeated samples of fixed sizes.

Keywords: Poststratification, successive sampling, ratio estimation, partial replacement, repeated samples.

INTRODUCTION

A number of surveys of practical importance sometimes centre on the estimation of the population ratio of two characters of study. Examples of such ratios are the ratio of total crop yield to the size of the entire farm, the ratio of men to women workers in a given locality, the ratio of employed to unemployed graduates in a country and the ratio of public to private servants in a region, just to mention a few. In surveys where stratification is used, it may be worth-while estimating the ratio of the character of study to the stratification variate.

Estimation of population ratio in a one-time survey abounds in literature. However, Rao (1957) was the first to consider ratio estimation in repetitive sampling. His estimator of the second occasion population ratio (R_2) is of the form:

$$\hat{R}_2 = \frac{\gamma_2}{\gamma_1} R_1 \quad (1.1)$$

where $\gamma_2 = \bar{y}_2/\bar{x}_2$, $\gamma_1 = \bar{y}_1/\bar{x}_1$, $R_2 = \bar{Y}_2/\bar{X}_2$, $R_1 = \bar{Y}_1/\bar{X}_1$, and y_{ij} (x_{ij}) is the i^{th} observation of the study characters on the j^{th} occasion ($i = 1, 2, \dots, N$; and $j = 1, 2$). Equation (1.1) is a ratio-type estimator based on a complete matching of sample units on the second occasion, and assuming knowledge of the first occasion population ratio (R_1).

Rao and Pereira (1968), following after Rao (1957), proposed a product-type estimator of R_2 based on a complete matching of sample units on the second occasion, and assuming knowledge of the first occasion population ratio, R_1 . But, Tripathi and Sinha (1976) introduced the use of partial matching of sample units, for the estimation of population ratio in successive sampling. Their estimator of R_2 is a linear (composite) function of estimates of R_2 based on both matched and unmatched sample units. Again, unlike the estimators proposed by Rao (1957) and Rao and Pereira (1968), the estimator proposed by Tripathi and Sinha (1976) does not assume knowledge of the first occasion population ratio, R_1 .

Rao (1957), Rao and Pereira (1968), and Tripathi and Sinha (1976) all considered ratio estimation when using a single cluster sampling scheme on successive occasions. The present study, however, focuses on ratio estimation when using poststratified sampling (PSS) scheme on successive occasions. Authors like Holt and Smith (1979), Ige (1984) and Onyeka (2001) have severally highlighted the importance of Poststratification in sample survey.

The present study is actually an extension of ratio estimation in a one-time poststratified sampling (PSS) to PSS over two occasions. Two ratio-type estimators are proposed here for the purpose of estimating the

second occasion population ratio, R_2 , in PSS over two occasions. The first estimator, e_1 , like those of Rao (1957) and Rao and Pereira (1968), assumes knowledge of the first occasion population ratio, R_1 . The second estimator, e_2 , like the estimator proposed by Tripathi and Sinha (1976), does not assume knowledge of R_1 . Furthermore, the two proposed estimators, e_1 and e_2 , are based on a partial replacement of sample units on the second occasion. Properties of the proposed estimators are obtained for repeated samples of fixed sizes, and conditions under-which one estimator is to be preferred to the other are also obtained.

The Proposed Estimators

Consider the following sampling design for poststratified sampling over two occasions.

A random sample of size n is drawn from a population of N units using simple random sampling without replacement (SRSWOR) method on the first occasion. The sampled units are allocated to their respective strata where n_{1h} is the number of units that fall into the h^{th} stratum such that $\sum_h n_{1h} = n$, ($h = 1, 2, \dots, L$).

It is assumed that n is large enough such that $\text{Prob}(n_{1h} = 0) = 0$ for all h . On the second occasion, $m_h = \lambda n_{1h}$ units of the first occasion sample are retained in the h^{th} stratum, $\sum_h m_h = m = \lambda n$, ($0 < \lambda < 1$). The

remaining $u_{1h} = n_{1h} - m_h = n_{1h} - \lambda n_{1h} = \mu n_{1h}$ units are discarded, $\sum_h u_{1h} = u = \mu n$, and $\mu + \lambda = 1$. Then, the

matched sample of size m is supplemented with a fresh (unmatched) sample of u units drawn independently from the entire population, again using SRSWOR method. The u sampled units are allocated to their respective strata where u_{2h} is the number of units that fall into the h^{th} stratum such that

$\sum_h u_{2h} = u$ ($= \sum_h u_{1h}$). Again, it is assumed that u is large enough such that $\text{Prob}(u_{2h} = 0) = 0$ for all h .

Let y_{jhi} and x_{jhi} denote observations on the i^{th} unit of the two characters of study in the h^{th} stratum on the j^{th} occasion, $i = 1, 2, \dots, N$; $h = 1, 2, \dots, L$ and $j = 1, 2$. The variate, x , in some cases can be the stratification variate. But generally, the variates y and x are to be taken as any two characters of study.

Let $R_2 = \bar{Y}_2/\bar{X}_2$ and $R_1 = \bar{Y}_1/\bar{X}_1$ respectively denote the second and first occasion population ratios of the two characters of study. We propose the following two estimators of R_2 in PSS over two occasions.

$$e_1 = \theta_1 \left(\frac{\gamma_{2m}}{\gamma_{1m}} R_1 \right) + (1 - \theta_1) \gamma_{2u} \quad (2.1)$$

$$\text{and } e_2 = \theta_2 \left(\frac{\gamma_{2m}}{\gamma_{1m}} \gamma_{1u} \right) + (1 - \theta_2) \gamma_{2u} \quad (2.2)$$

where

$$\gamma_{2m} = \frac{\sum_h W_h \bar{y}'_{2h}}{\sum_h W_h \bar{x}'_{2h}}, \quad \gamma_{1m} = \frac{\sum_h W_h \bar{y}'_{1h}}{\sum_h W_h \bar{x}'_{1h}}, \quad \gamma_{1u} = \frac{\sum_h W_h \bar{y}_{1h}}{\sum_h W_h \bar{x}_{1h}}, \quad \gamma_{2u} = \frac{\sum_h W_h \bar{y}''_{2h}}{\sum_h W_h \bar{x}''_{2h}} \quad (2.3)$$

\bar{y}_{1h} , \bar{x}_{1h} are sample means based on the entire first occasion sample of size n_{1h}

\bar{y}'_{2h} , \bar{x}'_{2h} , \bar{y}'_{1h} , \bar{x}'_{1h} are sample means based on the matched sample of size m_h

\bar{y}''_{2h} , \bar{x}''_{2h} are sample means based on the second occasion unmatched sample of size u_{2h}

and

θ_1 and θ_2 are constant (weighting) fractions of the matched and unmatched parts of the estimators e_1 and e_2 .

Properties Of The Proposed Estimators

Let S_{yh}^2 and S_{xh}^2 respectively denote the population variances of the variates y and x on both occasions in the h^{th} stratum. Also, let S_{xyh} denote the covariance of y and x on both first and second occasions in the h^{th} stratum. Theorem 1 gives the properties of the estimator e_1 , while Theorem 2 gives the properties of the estimator e_2 .

Theorem 1

The proposed estimator, $e_1 = \theta_1 \left(\frac{\gamma_{2m}}{\gamma_{1m}} R_1 \right) + (1 - \theta_1) \gamma_{2u}$, is biased for the second occasion population ratio (R_2) in poststratified sampling over two occasions. For repeated samples of fixed sizes n, m and u , the optimum value of the weighting fraction, θ_1 , and the associated mean square error of e_1 are respectively given by:

$$\theta_{01} = \frac{\lambda \sum_h W_h \sigma_{2h}}{\sum_h W_h \sigma_{2h} + \mu R_y^2 \sum_h W_h \sigma_{1h} - 2 \mu R_y \sum_h W_h \sigma_{21h}} \tag{3.1}$$

and

$$MSE(e_1) = \frac{\sum_h W_h \sigma_{2h} + R_y^2 \sum_h W_h \sigma_{1h} - 2 R_y \sum_h W_h \sigma_{21h}}{\sum_h W_h \sigma_{2h} + \mu R_y^2 \sum_h W_h \sigma_{1h} - 2 \mu R_y \sum_h W_h \sigma_{21h}} \cdot \frac{\sum_h W_h \sigma_{2h}}{n \bar{X}_2^2} \tag{3.2}$$

where

$$\left. \begin{aligned} \sigma_{2h} &= S_{yh}^2 + R_2^2 S_{xh}^2 - 2R_2 S_{xyh} & \sigma_{1h} &= S_{yh}^2 + R_1^2 S_{xh}^2 - 2R_1 S_{xyh} \\ \sigma_{21h} &= S_{yh}^2 + R_2 R_1 S_{xh}^2 - (R_2 - R_1) S_{xyh} \end{aligned} \right\} \tag{3.3}$$

and $R_2 = \bar{Y}_2 / \bar{X}_2, R_1 = \bar{Y}_1 / \bar{X}_1, R_y = \bar{Y}_2 / \bar{Y}_1$

Proof

The proposed estimator, e_1 , can be re-written as

$$e_1 = \theta_1 e_{1m} + (1 - \theta_1) \gamma_{2u} \tag{3.4}$$

where $e_{1m} = \frac{\gamma_{2m}}{\gamma_{1m}} R_1$ (3.5)

The estimator, e_{1m} , is a ratio-type double sampling estimator based on the matched sample, while the estimator, γ_{2u} , is the estimator based on the unmatched sample. Thus, the proposed estimator, e_1 , is a linear combination of the estimators e_{1m} and γ_{2u} based on the matched and unmatched samples, respectively. The weighting (constant) fraction, θ_1 , of the estimators, e_{1m} , and γ_{2u} , should be chosen so as to minimize the mean square error of the estimator, e_1 . Using the least squares method, the optimum value of θ_1 , and the associated mean square error of e_1 are respectively obtained as

$$\theta_{01} = \frac{MSE(\gamma_{2u})}{MSE(\gamma_{2u}) + MSE(e_{1m})} \tag{3.6}$$

and $MSE(e_1) = \frac{MSE(\gamma_{2u})MSE(\gamma_{2m})}{MSE(\gamma_{2u}) + MSE(e_{1m})}$ (3.7)

noting that the covariance of e_{1m} and γ_{2u} is zero since they are based on entirely independent matched and unmatched samples. To obtain the mean square error of the unmatched estimator, γ_{2u} , we write

$$\gamma_{2u} = \frac{\sum_h W_h \bar{y}_{2h}''}{\sum_h W_h \bar{x}_{2h}''} = \frac{\bar{y}_{2p}''}{\bar{x}_{2p}''} = R_2 (1 + \delta \bar{y}_{2p}'') (1 + \delta \bar{x}_{2p}'')^{-1} \quad (3.8)$$

where $\delta \bar{y}_{2p}'' = \frac{\bar{y}_{2p}'' - \bar{Y}_2}{\bar{Y}_2}$ and $\delta \bar{x}_{2p}'' = \frac{\bar{x}_{2p}'' - \bar{X}_2}{\bar{X}_2}$.

Expanding equation (3.8) in Taylor's series up to terms of n^{-1} in expected value gives

$$(\gamma_{2u} - R_2) = R_2 (\delta \bar{y}_{2p}'' - \delta \bar{x}_{2p}'' - \delta \bar{y}_{2p}'' \delta \bar{x}_{2p}'' + \delta^2 \bar{x}_{2p}'') \quad (3.9)$$

$$\text{and } (\gamma_{2u} - R_2)^2 = R_2^2 (\delta^2 \bar{y}_{2p}'' + \delta^2 \bar{x}_{2p}'' - 2 \delta \bar{y}_{2p}'' \delta \bar{x}_{2p}'') \quad (3.10)$$

Taking the conditional expectation (E_2) of equation (3.10) for an achieved sample configuration, $\underline{u} = (u_{21}, u_{22}, \dots, u_{2L})$, gives the conditional mean square error of γ_{2u} as

$$\text{MSE}_2(\gamma_{2u}) = R_2^2 \left[(\bar{Y}_2^2)^{-1} V_2(\bar{y}_{2p}'') + (\bar{X}_2^2)^{-1} V_2(\bar{x}_{2p}'') - 2 (\bar{Y}_2 \bar{X}_2)^{-1} C_2(\bar{y}_{2p}'', \bar{x}_{2p}'') \right] \quad (3.11)$$

where V_2 and C_2 are conditional variance and covariance given an achieved sample configuration, $\underline{u} = (u_{21}, u_{22}, \dots, u_{2L})$.

Following Ige (1984) and Onyeka (2001), we have $V_2(\bar{y}_{2p}'') = \sum_h W_h^2 u_{2h}^{-1} S_{yh}^2$,

$V_2(\bar{x}_{2p}'') = \sum_h W_h^2 u_{2h}^{-1} S_{xh}^2$ and $C_2(\bar{y}_{2p}'', \bar{x}_{2p}'') = \sum_h W_h^2 u_{2h}^{-1} S_{xyh}$. Making the necessary substitutions in equation

(3.11) gives the conditional mean square error of γ_{2u} as

$$\text{MSE}_2(\gamma_{2u}) = (\bar{X}_2^2)^{-1} \sum_h W_h^2 u_{2h}^{-1} \sigma_{2h} \quad (3.12)$$

where σ_{2h} is as given in equation (3.3). Taking the unconditional expectation (E_1) of equation (3.12) for repeated samples of fixed size u , and approximating up to terms of order n^{-1} gives the unconditional mean square error of γ_{2u} as

$$\text{MSE}_2(\gamma_{2u}) = (\mu n \bar{X}_2^2)^{-1} \sum_h W_h^2 \sigma_{2h} \quad (3.13)$$

noting that $E_1(u_{2h}^{-1}) = (\mu n W_h)^{-1}$. Similarly, the unconditional mean square errors and covariance of the estimators γ_{2m} and γ_{1m} are obtained as

$$\text{MSE}(\gamma_{2m}) = (\lambda n \bar{X}_2^2)^{-1} \sum_h W_h \sigma_{2h} \quad (3.14)$$

$$\text{MSE}(\gamma_{1m}) = (\lambda n \bar{X}_1^2)^{-1} \sum_h W_h \sigma_{1h} \quad (3.15)$$

$$\text{Cov}(\gamma_{2m}, \gamma_{1m}) = (\lambda n \bar{X}_2 \bar{X}_1)^{-1} \sum_h W_h \sigma_{21h} \quad (3.16)$$

Now, to obtain the mean square error of the matched estimator, e_{1m} , we write

$$e_{1m} = \frac{Y_{2m}}{Y_{1m}} R_1 = R_2(1 + \delta\gamma_{2m})(1 + \delta\gamma_{1m})^{-1} \tag{3.17}$$

Using Taylor's series to expand equation (3.17) up to terms of n^{-1} in expected value gives

$$(e_{1m} - R_2) = R_2(\delta\gamma_{2m} - \delta\gamma_{1m} - \delta\gamma_{2m}\delta\gamma_{1m} + \delta^2\gamma_{1m})$$

and $(e_{1m} - R_2)^2 = R_2^2(\delta^2\gamma_{2m} + \delta^2\gamma_{1m} - 2\delta\gamma_{2m}\delta\gamma_{1m})$ (3.18)

Taking the conditional and unconditional expectations of equation (3.18) in sequence gives the unconditional mean square error of e_{1m} as:

$$MSE(e_{1m}) = R_2^2 \left[(R_2^2)^{-1} MSE(\gamma_{2m}) + (R_1^2)^{-1} MSE(\gamma_{1m}) - 2(R_2 R_1)^{-1} Cov(\gamma_{2m}, \gamma_{1m}) \right] \tag{3.19}$$

And, using equations (3.14), (3.15) and (3.16) to make the necessary substitutions in equation (3.19) gives

$$MSE(e_{1m}) = (\lambda n \bar{X}_2^2)^{-1} \left[\sum_h W_h \sigma_{2h} + R_y^2 \sum_h W_h \sigma_{1h} - 2R_y \sum_h W_h \sigma_{21h} \right] \tag{3.20}$$

Finally, using equations (3.13) and (3.20) to make the necessary substitutions in equations (3.6) and (3.7) give the optimum weighting fraction, θ_{01} , and the associated mean square error of the proposed estimator, e_1 , as given in the theorem.

This completes the proof.

Theorem 2

The proposed estimator, $e_2 = \theta_2 \left(\frac{Y_{2m}}{Y_{1m}} \gamma_{1u} \right) + (1 - \theta_2) \gamma_{2u}$ is biased for the second occasion population ratio (R_2) in poststratified sampling over two occasions. For repeated samples of sizes n , m and u , the optimum weighting fraction, θ_{02} , and the associated mean square error of e_2 are respectively given by

$$\theta_{02} = \frac{\lambda \sum_h W_h \sigma_{2h}}{\sum_h W_h \sigma_{2h} + \mu^2 R_y^2 \sum_h W_h \sigma_{1h} - 2\mu^2 R_y \sum_h W_h \sigma_{21h}} \tag{3.21}$$

and $MSE(e_2) = \frac{\sum_h W_h \sigma_{2h} + \mu R_y^2 \sum_h W_h \sigma_{1h} - 2\mu R_y \sum_h W_h \sigma_{21h}}{\sum_h W_h \sigma_{2h} + \mu^2 R_y^2 \sum_h W_h \sigma_{1h} - 2\mu^2 R_y \sum_h W_h \sigma_{21h}} \cdot \frac{\sum_h W_h \sigma_{2h}}{n \bar{X}_2^2}$ (3.22)

where σ_{2h} , σ_{1h} , σ_{21h} and R_y are as previously defined in Theorem 1. The proof of Theorem 2 is similar to that of Theorem 1.

Comparison Of Proposed Estimators

The two proposed estimators, e_1 and e_2 , are both biased for the second occasion population ratio (R_2) in poststratified sampling over two occasions. Theorem 3 compares the performance of both estimators.

Theorem 3

The proposed estimator, e_1 , would perform better than the proposed estimator, e_2 , in terms of having a

smaller mean square error if $\frac{\sum_h W_h \sigma_{21h}}{\sum_h W_h \sigma_{1h}} > \frac{1}{2} R_y$ (4.1)

Conversely, the proposed estimator, e_2 , would perform better than the proposed estimator, e_1 , in terms of

having a smaller mean square error if
$$\frac{\sum_h W_h \sigma_{21h}}{\sum_h W_h \sigma_{1h}} < \frac{1}{2} R_y \quad (4.2)$$

Proof

From Theorems 1 and 2, the unconditional mean square errors of the proposed estimators, e_1 and e_2 , are respectively given by

$$\text{MSE}(e_1) = \frac{\sum_h W_h \sigma_{2h} + R_y^2 \sum_h W_h \sigma_{1h} - 2R_y \sum_h W_h \sigma_{21h}}{\sum_h W_h \sigma_{2h} + \mu R_y^2 \sum_h W_h \sigma_{1h} - 2\mu R_y \sum_h W_h \sigma_{21h}} \cdot \frac{\sum_h W_h \sigma_{2h}}{n\bar{X}_2^2} \quad (4.3)$$

and

$$\text{MSE}(e_2) = \frac{\sum_h W_h \sigma_{2h} + \mu R_y^2 \sum_h W_h \sigma_{1h} - 2\mu R_y \sum_h W_h \sigma_{21h}}{\sum_h W_h \sigma_{2h} + \mu^2 R_y^2 \sum_h W_h \sigma_{1h} - 2\mu^2 R_y \sum_h W_h \sigma_{21h}} \cdot \frac{\sum_h W_h \sigma_{2h}}{n\bar{X}_2^2} \quad (4.4)$$

Subtracting equation (4.4) from equation (4.3) gives

$$\begin{aligned} \text{MSE}(e_1) - \text{MSE}(e_2) &= \left[\frac{AC - B^2}{BC} \right] \frac{\sum_h W_h \sigma_{2h}}{n\bar{X}_2^2} \\ &= \left[\frac{R_y^2 \sum_h W_h \sigma_{1h} - 2R_y \sum_h W_h \sigma_{21h}}{BC} \right] \frac{\sum_h W_h \sigma_{2h}}{n\bar{X}_2^2} \end{aligned} \quad (4.5)$$

where

$$A = \sum_h W_h \sigma_{2h} + R_y^2 \sum_h W_h \sigma_{1h} - 2R_y \sum_h W_h \sigma_{21h}$$

$$B = \sum_h W_h \sigma_{2h} + \mu R_y^2 \sum_h W_h \sigma_{1h} - 2\mu R_y \sum_h W_h \sigma_{21h}$$

$$\text{and } C = \sum_h W_h \sigma_{2h} + \mu^2 R_y^2 \sum_h W_h \sigma_{1h} - 2\mu^2 R_y \sum_h W_h \sigma_{21h}$$

From equation (3.20), we have $A > 0$. It is trivial to verify that $B > 0$ and $C > 0$, so that the product $BC > 0$. Thus, equation (4.5) is less than zero if and only if:

$$R_y^2 \sum_h W_h \sigma_{1h} - 2R_y \sum_h W_h \sigma_{21h} < 0 \quad \text{or} \quad \frac{\sum_h W_h \sigma_{21h}}{\sum_h W_h \sigma_{1h}} > \frac{1}{2} R_y$$

This proves the first part of the theorem. The second part of the theorem follows directly from the first part. And this completes the proof.

DISCUSSION OF RESULTS

Theorem 3 shows that the proposed estimator e_1 would be preferred to the proposed estimator e_2 if equation (4.1) holds. Now, the left-hand side of equation (4.1) is an expression for the regression coefficient of $(y_{2hi} - R_2 x_{2hi})$ on $(y_{1hi} - R_1 x_{1hi})$, which is ultimately a function of the regression coefficient of second occasion observations on first occasion observations of the study characters.

Again, the quantity $R_y = \bar{Y}_2 / \bar{Y}_1$ on the right-hand side of equation (4.1) is very close to the rate of change of the variate y on the second occasion. Where there is only a little change in the study characters on the

second occasion, as it is often the case, the quantity R_y would approximate to unity. Consequently, equation (4.1) would reduce to:

$$\frac{\sum_h W_h \sigma_{21h}}{\sum_h W_h \sigma_{1h}} > \frac{1}{2} \quad (4.6)$$

Equation (4.6), therefore, suggests that the proposed estimator e_1 would be preferred to the proposed estimator e_2 if the regression coefficient of second occasion observations on first occasion observations is greater than 0.5. That is, if there is high correlation between the second and first occasion observations of the study variates. Noting that this condition is likely to be satisfied in many practical surveys, it then means that equation (4.1) is more likely to be satisfied than equation (4.2) would be satisfied. That is, the number of cases where the estimator e_1 would perform better than the estimator e_2 is more likely to be greater than the number of cases where e_2 would perform better than e_1 .

In summary, therefore, the proposed estimator e_1 is expected to perform better than the proposed estimator e_2 in most cases. But the estimator e_1 assumes knowledge of the first occasion population ratio, R_1 , while the estimator e_2 does not require any such assumption. And, we know there are quite a number of situations where information on R_1 might not be readily available or very highly expensive to obtain. In such cases, the estimator e_2 would obviously be preferred to the estimator e_1 since the estimator e_2 does not assume knowledge of the first occasion population ratio, R_1 .

REFERENCES

- Holt, D and Smith, T.M.F., 1979. Post Stratification. *J.Roy. Stat. Soc., A*, 142: 33-46.
- Ige, A. F., 1984. Contributions to post stratified sampling and double sampling for stratification. A Ph.D. thesis submitted to the Faculty of Science, University of Ibadan, Nigeria.
- Onyeka, A. C., 2001. Univariate estimators of the population mean in poststratified sampling over two occasions. *J. Nig. Stat. Asso.* 14: 26-33.
- Rao, J. N. K., 1957. Double ratio estimate in forest surveys. *J. Ind. Soc. Agric. Stat.*, 9: 191 – 204.
- Rao, J.N.K. and Pereira, N.P., 1968. On double ratio estimators. *Sankhya, A*, 30: 83 – 90.
- Tripathi, T. P. and Sinha, S. K. P., 1976. Estimation of ratio on successive occasions. Proc of Conference on, Recent development in Survey Sampling, held in ISI Calcutta.