Metrika, Band 24, 1977, Seite 203–208. Physica-Verlag, Wien.

Estimating the Variance of the Ratio Estimator for the *Midzuno-Sen* Sampling Scheme

By *T.J. Rao*, Melbourne[1])

Summary: The problem of estimating the variance of the ratio estimator for the *Midzuno-Sen* sampling scheme is further studied in this paper. Sufficient conditions are derived for which the suggested variance estimator is always positive definite.

1. Introduction

Consider a finite population of N units

$$\mathbf{U} : (U_1, U_2, \ldots, U_N). \tag{1.1}$$

Let **Y** and **X** be real valued characteristics taking values Y_i and X_i respectively on U_i. The values X_i are assumed to be positive, $i = 1, 2, \ldots, N$. Let $\hat{R} = \sum_{i \in s} y_i / \sum_{i \in s} x_i$ be the ratio estimator for the estimation of the population ratio $R = Y/X$ where $Y = \sum_{i=1}^{N} Y_i$ and $X = \sum_{i=1}^{N} X_i$, based on a sample s of size n from (1.1). In general, $\hat{R}$ is biased for R and *Midzuno* [1952] and *Sen* [1952] have independently given a very simple procedure which makes the ratio estimator unbiased. Their method consists in drawing the first unit with probability proportional to size (the X-characteristic) and the rest of the $(n-1)$ units by Simple Random Sampling With Out Replacement (SRSWOR) from the remaining $(N-1)$ units of the population. It can be easily derived that the probability of selection of the sample under this scheme is given by

$$P_s = \sum_{i \in s} x_i \Big/ \binom{N-1}{n-1} X \tag{1.2}$$

An exact expression for the variance of the ratio estimator $\hat{R}$ under the *Midzuno-Sen* sampling scheme has been derived in *Rao* [1966] and further results on the properties of the coefficients which occur in the variance expression and the problem of estimation of this variance have been studied later in *Rao* [1967, 1972]. We have from these results that

[1]) *T.J. Rao*, CSIRO Division of Mathematics and Statistics, Melbourne, P.O. Box 56, Highett, Vic. 3190, Melbourne. On leave from Indian Statistical Institute, Calcutta.

$$V(\hat{R}) = \sum_{i=1}^{N} \lambda_i Y_i^2 + \sum\sum_{i \neq j}^{NN} \lambda_{ij} Y_i Y_j \tag{1.3}$$

where

$$\lambda_i = \left\{ \left(1 \Big/ \binom{N-1}{n-1} X \right) \sum_{\lambda} (X_i + X_\lambda^i)^{-1} \right\} - X^{-2}, \tag{1.4}$$

X_λ^i being the sum of the λth set of $(n-1)$ distinct X's other than X_i and the summation (over λ) being taken over all such $\binom{N-1}{n-1}$ sets; and

$$\lambda_{ij} = \left\{ \left(1 \Big/ \binom{N-1}{n-1} X \right) \sum_{\lambda} (X_i + X_j + X_\lambda^{ij})^{-1} \right\} - X^{-2}, \tag{1.5}$$

X_λ^{ij} being the sum of the λth set of $(n-2)$ distinct X's other than X_i and X_j and the summation (over λ) being taken over all such $\binom{N-2}{n-2}$ sets. Next the problem of estimation of $V(\hat{R})$ has been discussed in *Rao* [1967, 1972] and an unbiased estimator suggested therein is given by

$$\hat{V}(\hat{R}) = \sum_{i \in s} \frac{\lambda_i y_i^2}{\pi_i^1} + \sum\sum_{i \neq j \in s} \frac{\lambda_{ij} y_i y_j}{\pi_{ij}^1} \tag{1.6}$$

where

$$\pi_i^1 = \frac{n-1}{N-1} + \frac{N-n}{N-1} \frac{X_i}{X} \text{ and}$$

$$\pi_{ij}^1 = \frac{n-1}{N-1} \frac{n-2}{N-2} + \frac{n-1}{N-1} \frac{N-n}{N-2} \left(\frac{X_i}{X} + \frac{X_j}{X} \right). \tag{1.7}$$

Furthermore, it is suggested there that a sufficient condition for $\hat{V}(\hat{R})$ to be non-negative is that $\lambda_{ij} \geqslant 0$ for all i, j. *Chaudhuri* [1975] assumes that the characteristic $\mathbf{Y}$ can take negative values in which case the above sufficient condition is not valid. *Chaudhuri* therefore, suggests an alternative unbiased estimator by writing $V(\hat{R})$ in a different form, namely

$$\begin{aligned} V(\hat{R}) &= \sum_{i=1}^{N} \lambda_i Y_i^2 + \sum\sum_{i \neq j}^{NN} \lambda_{ij} Y_i Y_j \\ &= \frac{1}{X^2} \left\{ \sum_{i=1}^{N} (T_i - 1) Y_i^2 + \sum\sum_{i \neq j}^{NN} (T_{ij} - 1) Y_i Y_j \right\}, \end{aligned}$$

where

$$T_i = \lambda_i X^2 + 1, T_{ij} = \lambda_{ij} X^2 + 1, \text{ so that, if } t_i = Y_i / X, \text{ we have}$$

$$V(\hat{R}) = \sum_{i=1}^{N} (T_i - 1) t_i^2 + \sum\sum_{i \neq j}^{NN} (T_{ij} - 1) t_i t_j$$

$$= \sum_{i=1}^{N} t_i^2 (n T_i - N) + \sum\sum_{i<j}^{NN} (1 - T_{ij}) (t_i - t_j)^2 , \qquad (1.8)$$

using *Rao*'s [1972] results of Theorem 3.1 (i.e. $\sum_{j \neq i}^{N} T_{ij} = (n-1) T_i$).

He then proposed the unbiased estimator

$$\hat{V}_c(\hat{R}) = \sum_{i \in s} t_i^2 \frac{(nT_i - N)}{\pi_i^1} + \sum\sum_{(i<j) \in s} \frac{(1 - T_{ij})}{\pi_{ij}^1} (t_i - t_j)^2 \qquad (1.9)$$

where π_i^1 and π_{ij}^1 are defined in (1.7).

This estimator can be used even when Y_i's (equivalently t_i's) are negative. Sufficient conditions given by *Chaudhuri* for $\hat{V}_c(\hat{R})$ to be positive definite are that

$$T_{ij} \leqslant 1 \text{ for all } i \neq j \text{ and}$$

$$T_i \geqslant \frac{N}{n} \text{ for all } i. \qquad (1.10)$$

In the next sections we improve upon the sufficient conditions proposed by *Chaudhuri* and comment on various other alternative estimators.

2. Non-Negativity of the Variance Estimator

Rewrite the expression for $V(\hat{R})$ in the form:

$$2\, V(\hat{R}) = \sum\sum_{i \neq j}^{NN} \left[\frac{(T_i - 1)}{N-1} t_i^2 + 2(T_{ij} - 1) t_i t_j + \frac{(T_j - 1)}{N-1} t_j^2 \right] . \qquad (2.1)$$

Sufficient conditions for the positive-definiteness of the quadratic form Q_{ij} within the parenthesis of (2.1) are given by

$$T_i > 1 \qquad (2.2a)$$

and $$(T_{ij} - 1)^2 - (T_i - 1)(T_j - 1) / (N-1)^2 < 0. \qquad (2.2b)$$

It is proved in *Rao* [1972] that (2.2a) is always true and the only condition to be verified is (2.2b) which involves a certain amount of calculation and does not explicitly specify the bounds for T_{ij}. However, when the conditions are satisfied one could consider a non-negative unbiased estimator given by

$$\hat{V}'(\hat{R}) = \frac{1}{2} \sum\sum_{i \neq j \in s} Q_{ij} / \pi^1_{ij} . \tag{2.3}$$

It is easy to see that the expression for $V(\hat{R})$ can be thrown into an alternative form:

$$2\,V(\hat{R}) = \sum_{i \neq j}^{N}\sum^{N} \left(\frac{T_{ij}}{n-1} - \frac{1}{N-1}\right) t_i^2 + 2\,(T_{ij} - 1)\, t_i\, t_j + \left(\frac{T_{ij}}{n-1} - \frac{1}{N-1}\right) t_j^2 . \tag{2.4}$$

using the fact that $\sum_{j \neq i}^{N} T_{ij} = (n-1)\,T_i$ [*Rao*, 1972]. It is now immediate that the sufficient conditions for the positive definiteness of the quadratic form of (2.4) are given by

$$T_{ij} > \frac{n-1}{N-1} \tag{2.5a}$$

$$\text{and } (T_{ij} - 1)^2 - \left(\frac{T_{ij}}{n-1} - \frac{1}{N-1}\right)^2 < 0 \ . \tag{2.5b}$$

Again we have from *Rao* [1972] that $T_{ij} > \left(\frac{n-1}{N-1}\right)^2 \frac{1}{\pi^1_{ij}}$ which shows that (2.5a) is satisfied. (2.5b) is true if, and only if

$$T_{ij}^2 \frac{n\,(n-2)}{(n-1)^2} - T_{ij} \frac{2nN - 2N - 2n}{(n-1)(N-1)} + \frac{N\,(N-2)}{(N-1)^2} < 0$$

$$\text{i.e.} \quad T_{ij}^2 - T_{ij} \frac{(2nN - 2N - 2n)(n-1)}{n\,(n-2)(N-1)} + \frac{N\,(N-2)}{(N-1)^2} \frac{(n-1)^2}{n\,(n-2)} < 0, \text{ if } n > 2$$

$$\text{and} \quad T_{ij} > \frac{N}{2\,(N-1)} \quad \text{if } n = 2$$

$$\text{or} \quad \left[T_{ij} - \frac{N\,(n-1)}{n\,(N-1)} \right] \left[T_{ij} - \frac{N-2}{n-2} \frac{n-1}{N-1} \right] < 0, \text{ if } n > 2$$

$$\text{and} \quad T_{ij} > \frac{N}{2\,(N-1)} \quad \text{if } n = 2 \tag{2.6}$$

which gives the bounds for T_{ij} as

$$T_{ij} > \frac{N}{2\,(N-1)}, \text{ if } n = 2 \text{ and}$$

$$\frac{N}{n}\frac{n-1}{N-1} < T_{ij} \leq \frac{N-2}{n-2}\frac{n-1}{N-1} \quad . \tag{2.7}$$

Notice here that $\frac{N-2}{n-2}\frac{n-1}{N-1} > 1$, while the bound given by *Chaudhuri* is 1.

Thus for a given sample it is easy to verify whether the T_{ij}'s lie in the interval and then use the estimator

$$\hat{V}''(\hat{R}) = \frac{1}{2} \sum\sum_{i \neq j \in s} \left[\left(\frac{T_{ij}}{n-1} - \frac{1}{N-1} \right) t_i^2 + 2(T_{ij} - 1) t_i t_j + \right.$$
$$\left. + \left(\frac{T_{ij}}{n-1} - \frac{1}{N-1} \right) t_j^2 \right] \Big/ \pi_{ij}^1 \tag{2.8}$$

as an unbiased and non-negative estimator of the variance of R, whatever be the sign of t_i's (equivalently Y_i's).

3. Remarks

The motivation for considering $\hat{V}_c(\hat{R})$ by *Chaudhuri* is that although in survey-sampling problems the characteristics studied are usually positive valued, it is not difficult to find instances when they may assume negative values as well. In such cases, one could consider $Y_i^1 = Y_i + CX_i$, where C is a constant chosen in such a way that $Y_i^1 > 0$ for all $i = 1, 2, \ldots, N$.

Then

$$\hat{R}^1 = \frac{\sum_{i \in s} y_i^1}{\sum_{i \in s} x_i} = \frac{\sum_{i \in s} (y_i + Cx_i)}{\sum_{i \in s} x_i} = R + C \tag{3.1}$$

and $V(\hat{R}) = V(\hat{R}^1)$ (3.2)

and $V(\hat{R}^1)$ can be estimated unbiasedly by

$$\hat{V}(\hat{R}^1) = \sum_{i \in s} \lambda_i {t_i^1}^2 / \pi_i^1 + \sum\sum_{i \neq j \in s} \lambda_{ij} t_i^1 t_j^1 / \pi_{ij}^1 \tag{3.3}$$

where $t_i^1 = y_i^1 / X$.

Once again, the sufficient conditions for non-negativity of $\hat{V}(\hat{R}^1)$ are that $\lambda_{ij} > 0$. However, since the conditions $\lambda_{ij} > 0$ may not be satisfied for all sampled pairs, we use the approach given in section 2 above.

Another alternative attempt when the characteristic Y takes negative values (being the difference of two positive valued y-variates, as for example, increase in yield of a crop in one year over that in a preceding base year where decrease in value is treated as negative increase) would be to write

$$V(\hat{R}) = V\left(\frac{\sum\limits_{i\in s}(y_i^1 - y_i^0)}{\sum\limits_{i\in s} x_i}\right)$$

$$= V\left(\frac{\Sigma y_i^1}{\Sigma x_i}\right) + V\left(\frac{\Sigma y_i^0}{\Sigma x_i}\right) - 2\,\mathrm{Cov.}\left(\frac{\Sigma y_i^1}{\Sigma x_i}, \frac{\Sigma y_i^0}{\Sigma x_i}\right) \tag{3.4}$$

(where the notation is self-explanatory) and estimate the variances and covariances separately [cf. *Cochran*, 1963]. Notice here that an expression for covariance can be obtained as for the variance. However, it would not be easy to obtain conditions under which $\hat{V}(\hat{R})$ would be non-negative in this case.

References

Chaudhuri, A.: On some inference problems with finite populations and related topics in survey sampling theory. Economic Statistics Papers. No. 10, 1975. University of Sydney.

Cochran, W.G.: Sampling Techniques. New York 1963, 181–184.

Midzuno, H.: On the sampling system with probability proportional to the sum of sizes. Ann. Inst. Stat. Math. 3, 1952, 99–108.

Rao, T.J.: On the variance of the ratio estimator for Midzuno-Sen sampling scheme. Metrika 10, 1966, 89–91.

–: Contributions to the theory of sampling strategies, Ph. D. Thesis submitted to the Indian Statistical Institute, 1967.

–: On the variance of the ratio estimator. Metrika 18, 1972, 209–215.

Sen, A.R.: Present status of probability sampling and its use in estimation of farm characteristics (abstract). Econometrica 20, 1952, 103.

Eingegangen am 19. Mai 1976