AN ESTIMATE OF THE EFFICIENCY OF RAKING RATIO ESTIMATORS UNDER SIMPLE RANDOM SAMPLING

M.D. Bankier[1]

Raking ratio estimators give estimates of the population values of characteristics examined on a sample basis utilizing the row and column totals of a contingency table of characteristics examined on a 100% basis. In this paper, the asymptotic variance of the maximum likelihood estimator of a sample characteristic subject to the marginal constraints of the above contingency table is derived. From this, we are able to compute the loss in efficiency of the raking ratio estimators relative to the maximum likelihood estimator in an empirical study.

1. INTRODUCTION

Raking ratio estimators given estimates of the population values of characteristics examined on a sample basis utilizing the row and column totals of a contingency table of characteristics examined on a 100% basis. Arora and Brackstone [1] described the use of RRE in the 1971 and 1976 Canadian Censuses of Population and Housing. They also derived formulae for the asymptotic variance of RRE under simple random sampling without replacement (s.r.s.w.o.r.) and presented an empirical study. Rao [3] found the asymptotic variance-covariance (V-C) matrix of the maximum likelihood estimators (MLE) for those variables examined on a 100% basis where the estimators were subject to the marginal constraints of the above contingency table. He restricted himself to the special case of estimators of frequency counts.

In this paper, Rao's results are generalized to variables examined on a sample basis. The results are derived under the assumption of simple random sampling with replacement (s.r.s.w.r.) since assuming sampling without replacement makes the problem much more complex. Because the

[1] M.D. Bankier, Census Survey Methods Division, Statistics Canada.

MLE are asymptotically most efficient, we can compute the loss in efficiency of the RRE relative to the MLE. This has been done in an empirical study presented in Section 4.

Variables examined on a 100% basis and variables examined on a sample basis will be known as 2A-categories and 2B-categories respectively.

2. DERIVATION OF THE ASYMPTOTIC VARIANCE OF THE MLE

Suppose there are t 2B-categories and let

$$N_{ijk} \qquad \begin{array}{l} i = 1, 2, \ldots, r \\ j = 1, 2, \ldots, s \\ k = 1, 2, \ldots, t \end{array}$$

be the number of individuals in the population that belong to the kth 2B-category and fall in the (i,j)th cell of the 2A-category cross-classification table. Define

$$\pi_{ijk} = \frac{N_{ijk}}{N}$$

so that

$$\sum_j \sum_k \pi_{ijk} = \pi_{i..} \qquad i = 1, \ldots, r-1$$

$$\sum_i \sum_k \pi_{ijk} = \pi_{.j.} \qquad j = 1, \ldots, s-1 \tag{2.1}$$

$$\sum_i \sum_j \sum_k \pi_{ijk} = 1$$

where $\pi_{i..} = \frac{N_{i..}}{N}$ and $\pi_{.j.} = \frac{N_{.j.}}{N}$ are the known marginal proportions.

In this section the asymptotic variance of the MLE π^*_{ijk} of π_{ijk} is found subject to the constraints (2.1).

Let us assume that we have taken an s.r.s.w.r. Then the likelihood of the sample frequencies n_{ijk}'s is given by

$$L \propto \prod_{i,j,k} (\pi_{ijk})^{n_{ijk}} . \qquad (2.2)$$

We maximize

$$\ln L = \text{constant} + \sum_i \sum_j \sum_k n_{ijk} \ln \pi_{ijk} \qquad (2.3)$$

subject to the constraints (2.1) to find the MLE π^*_{ijk}. To do this we would have to solve a system of non-linear equations iteratively.

Silvey [4] has given a general method for finding the asymptotic V-C matrix of the MLE. Let

$$\underset{\sim}{B}(\underset{\sim}{\pi}) = \frac{1}{n} \left(E \left(- \frac{\partial^2 \ln L}{(\partial \pi_{ijk})(\partial \pi_{i'j'k'})}\right)\right) \qquad (2.4)$$

denote the (rst) x (rst) information matrix. Let $\underset{\sim}{H}(\underset{\sim}{\pi})$ be the (rst) x (r + s - 1) matrix of derivatives of (2.1) with respect to the π_{ijk}'s. The derivatives are in the order that the π_{ijk}'s fall in $\underset{\sim}{\pi}$ where

$$\underset{\sim}{\pi}' = (\underset{\sim}{\pi}'_{1..} | \underset{\sim}{\pi}'_{2..} | \ldots | \underset{\sim}{\pi}'_{r-1,..} | \underset{\sim}{\pi}'_{.s.} | \underset{\sim}{\pi}'_{r..} | \underset{\sim}{\pi}'_{rs.}) \qquad (2.5)$$

$$\pi'_{i..} = (\underset{\sim}{\pi}'_{i1.} | \underset{\sim}{\pi}'_{i2.} | \ldots | \underset{\sim}{\pi}'_{i,s-1,.}) \qquad (2.6)$$

$$\underset{\sim}{\pi}'_{.j.} = (\underset{\sim}{\pi}'_{1j.}|\underset{\sim}{\pi}'_{2j.}|\cdots|\underset{\sim}{\pi}'_{r-1,j.}) \tag{2.7}$$

and

$$\underset{\sim}{\pi}'_{ij.} = (\pi_{ij1}, \pi_{ij2}, \ldots, \pi_{ijt}). \tag{2.8}$$

The asymptotic V-C matrix of the π^*_{ijk} is given by $\frac{1}{n}\underset{\sim}{D}(\underset{\sim}{\pi})$ where

$$\left[\begin{array}{c|c} \underset{\sim}{B}(\underset{\sim}{\pi}) & \underset{\sim}{H}(\underset{\sim}{\pi}) \\ \hline \underset{\sim}{H}'(\underset{\sim}{\pi}) & \underset{\sim}{0}_{(r+s-1),(r+s-1)} \end{array}\right]^{-1} = \left[\begin{array}{c|c} \underset{\sim}{D}(\underset{\sim}{\pi}) & \underset{\sim}{Q}(\underset{\sim}{\pi}) \\ \hline \underset{\sim}{Q}'(\underset{\sim}{\pi}) & \underset{\sim}{R}(\underset{\sim}{\pi}) \end{array}\right] \tag{2.9}$$

and $\underset{\sim}{0}_{(r+s-1),(r+s-1)}$ is a matrix of zeroes of order $(r+s-1) \times (r+s-1)$.

Using the formula for the inverse of a partitioned matrix, we find

$$\begin{aligned} \underset{\sim}{D}(\underset{\sim}{\pi}) &= \underset{\sim}{B}^{-1}(\underset{\sim}{\pi}) - \underset{\sim}{B}^{-1}(\underset{\sim}{\pi})\ \underset{\sim}{H}\ (\underset{\sim}{\pi})\ [\underset{\sim}{H}'(\underset{\sim}{\pi})\ \underset{\sim}{B}^{-1}(\underset{\sim}{\pi})\ \underset{\sim}{H}(\underset{\sim}{\pi})\ ]^{-1}\ \underset{\sim}{H}'(\underset{\sim}{\pi})\underset{\sim}{B}^{-1}(\underset{\sim}{\pi}) \\ &= \underset{\sim}{B}^{-1}(\underset{\sim}{\pi}) - \underset{\sim}{A}\ \underset{\sim}{F}^{-1}\ \underset{\sim}{A}' \end{aligned} \tag{2.10}$$

where

$$\underset{\sim}{A} = \underset{\sim}{B}^{-1}(\underset{\sim}{\pi})\underset{\sim}{H}(\underset{\sim}{\pi}) \text{ and } \underset{\sim}{F} = \underset{\sim}{H}'(\underset{\sim}{\pi})\ \underset{\sim}{B}^{-1}(\underset{\sim}{\pi})\ \underset{\sim}{H}(\underset{\sim}{\pi}).$$

Define

$$\underset{\sim}{c}' = \underbrace{[(\underset{\sim}{e}_t^k)'|(\underset{\sim}{e}_t^k)'|\ldots|(\underset{\sim}{e}_t^k)']}_{\text{rs submatrices}} \tag{2.11}$$

where

$$(\underset{\sim}{e}_t^k)' = \overbrace{[0, \ldots, 0, 1, 0, \ldots, 0]}^{\text{t columns}} \tag{2.12}$$

and the one is in the kth column.

It can be shown that the asymptotic variance of $\pi^*_{..k}$ is

$$V(\pi^*_{..k}) = \sum_i \sum_j \sum_{i'} \sum_{j'} C(\pi^*_{ijk}, \pi^*_{i'j'k})$$

$$= \frac{\underset{\sim}{c}'\ (\underset{\sim}{B}^{-1}(\underset{\sim}{\pi}) - \underset{\sim}{A}\underset{\sim}{F}^{-1}\ \underset{\sim}{A}')\ \underset{\sim}{c}}{n}$$

$$= \frac{1}{n}\,(\pi_{..k} - \underset{\sim}{a}'\ \underset{\sim}{F}^{-1}\ \underset{\sim}{a}) \tag{2.13}$$

where $\underset{\sim}{a}' = (\pi_{1.k}, \pi_{2.k}, \ldots, \pi_{r-1,.k}, \pi_{.1k}, \pi_{.2k}, \ldots, \pi_{.,s-1,k}, \pi_{..k})$ (2.14)

$$\underset{\sim}{F} = \left[\begin{array}{c|c} \underset{\sim}{F}_{11} & \underset{\sim}{f}_{12} \\ \hline \underset{\sim}{f}'_{12} & 1 \end{array}\right] \tag{2.15}$$

$$\underset{\sim}{f}'_{12} = (\pi_{1..},\pi_{2..},\ldots,\pi_{r-1,..},\pi_{.1.},\pi_{.2.},\ldots,\pi_{.,s-1,.}) \tag{2.16}$$

$$\underset{\sim}{F}_{11} = \left[\begin{array}{c|c} \underset{\sim}{E}_1 & \underset{\sim}{E}_2 \\ \hline \underset{\sim}{E}'_2 & \underset{\sim}{E}_3 \end{array}\right] \tag{2.17}$$

$$\underset{\sim}{E}_1 = \text{diag }(\pi_{1..},\pi_{2..},\ldots,\pi_{r-1,..}) \tag{2.18}$$

$$\underset{\sim}{E}_3 = \text{diag }(\pi_{.1.},\pi_{.2.},\ldots,\pi_{.,s-1,.}) \tag{2.19}$$

and

$$\underset{\sim}{E}_2 = \begin{bmatrix} \pi_{11.} & \pi_{12.} & \cdots & \pi_{1,s-1,.} \\ \pi_{21.} & \pi_{22.} & \cdots & \pi_{2,s-1,.} \\ \vdots & & & \vdots \\ \pi_{r-1,1.} & \pi_{r-1,2.} & \cdots & \pi_{r-1,s-1.} \end{bmatrix} . \tag{2.20}$$

Rao [3] demonstrated that the inversion of $\underset{\sim}{F}$ can be reduced to the inversion of $\underset{\sim}{E}_3 - \underset{\sim}{E}'_2\,\underset{\sim}{E}_1^{-1}\,\underset{\sim}{E}_2$ which is a (s-1) x (s-1) matrix.

3. AN EMPIRICAL STUDY

This empirical study uses the same data and categories as the paper by Arora and Brackstone [1]. The data came from one Electoral District (ED) that contained 15 Weighting Areas (WAs) and the data were gathered in the 1974 Canadian Test Census.

In the previous section, the asymptotic variance of the MLE $\pi^*_{..k}$ of $\pi_{..k}$ was derived under s.r.s.w.r. To allow efficiency comparisons with the RRE, the large-sample variances of the RRE without the f.p.c. are used. The estimate SE^*_0 of the asymptotic standard error of $N\pi^*_{..k}$ is calculated using zero iteration raking ratio estimates $\hat{\pi}^{(0)}_{ijk} = \frac{n_{ijk}}{n}$ for the π_{ijk}'s in (2.13). Because the zero iteration estimator of π_{ijk} is unbiased, it is felt that it is better to use it rather than the fourth iteration raking ratio estimate $\hat{\pi}^{(4)}_{ijk}$. SE^*_0 expressed as a percentage of SE_p (the estimated pth iteration RRE standard error under s.r.s.w.r.) will be denoted by RE^0_p. The table at the end of this paper gives at the ED level SE_p, RE^0_p (p=0, 1, 2, 3, 4) and SE^*_0. The categories examined are:

Class A(2B-Categories):

- A1 Households with Employed Heads
- A2 Households with Unemployed Heads
- A3 Households with Heads Not in Labour Force
- A4 Household with Heads Not Moved in 5 Years
- A5 Households with Heads Moved in Last 5 Years in Same Municipality
- A6 Highest Grade of Head is 1 to 10
- A7 Heads with Bachelor Degree or Higher

Class B(2A-Categories)

- B1 Households with 3 or 4 Persons
- B2 Age of Head is Less Than 25
- B3 Age Of Head is 25 to 34
- B4 Head Who is Widowed, Divorced, or Separated

Class C(2A-Categories): C1 Households with 2 or Fewer Persons
C2 Age of Head is 65 or More
C3 Owned Dwellings
C4 Rented Apartments

4. ANALYSIS OF THE RESULTS

For all 2B-categories, it can be seen that $RE_2^0 \geq 98.6\%$ and $RE_4^0 \geq 99.8\%$. These results indicate that the fourth iteration RRE is almost as efficient as the MLE and that the RRE do not gain much in efficiency after the second iteration.

The results are similar for the 2A-categories with the exception of C3 and C4. For C3, the RRE at the third iteration and the MLE both have zero standard errors while for C4, $RE_3^0 = 99.9\%$. Thus for the 2A-categories there is near equality between the standard errors of the RRE at either the third or fourth iteration and the MLE.

5. CONCLUSIONS

The empirical study, which assumes s.r.s.w.r. and that n is large, indicates that the RRE by the second iteration is almost as efficient as the MLE for most categories and that generally only small gains in efficiency are made at the third and fourth iteration.

6. ACKNOWLEDGEMENTS

This problem was suggested to me by Professor J.N.K. Rao for my Master's Thesis [2]. I would like to thank him for his time spent and for his advice. I would also like to express my appreciation to H.R. Arora and G.J. Brackstone who had many discussions with me about their work.

ESTIMATED EFFICIENCY OF THE RRE VERSUS THE MLE

UNDER S.R.S.W.R. AT THE ED LEVEL

	Iteration 0		Iteration 1		Iteration 2		Iteration 3		Iteration 4		MLE
Category	SE_0	RE_0^0	SE_1	RE_1^0	SE_2	RE_2^0	SE_3	RE_3^0	SE_4	RE_4^0	SE_0^*
A1	102.9	77.4	101.7	78.3	79.8	99.8	79.9	99.8	79.7	99.9	79.7
A2	39.7	98.6	39.5	99.0	39.2	99.9	39.2	99.9	39.2	99.9	39.1
A3	99.1	75.1	98.2	75.8	74.6	99.8	74.7	99.7	74.5	99.9	74.5
A4	116.0	83.3	104.4	92.6	98.0	98.6	96.8	99.8	96.8	99.8	96.6
A5	103.6	94.2	100.8	96.9	98.0	99.6	97.8	99.9	97.8	99.9	97.7
A6	115.1	94.9	113.9	95.9	109.4	99.8	109.4	99.8	109.3	99.9	109.2
A7	59.6	97.3	59.2	97.9	58.1	99.8	58.1	99.9	58.0	99.9	58.0
B1	114.4	84.1	113.2	85.1	96.4	99.9	96.6	99.6	96.4	99.9	96.3
B2	61.1	83.9	57.7	88.8	51.5	99.5	51.5	99.6	51.3	99.9	51.3
B3	98.0	51.1	96.0	52.1	50.2	99.7	51.1	98.0	50.1	99.9	50.1
B4	87.8	67.4	86.3	68.6	59.2	99.9	59.6	99.3	59.2	99.9	59.2
C1	112.5	40.4	106.7	42.7	45.6	99.8	47.7	95.4	45.5	99.9	45.5
C2	82.8	25.0	81.7	25.4	20.8	99.7	21.4	96.8	20.7	99.9	20.7
C3	109.4	0.0	0.0	-	41.9	0.0	0.0	-	11.2	0.0	0.0
C4	80.1	17.6	14.7	96.4	34.7	40.7	14.2	99.9	17.0	83.4	14.1

RESUME

Les estimateurs d'échantillon en formation donnent des estimations de la valeur, dans la population, des caractéristiques qui ont été étudiées à partir d'un échantillon, en utilisant les totaux des rangées et des colonnes d'un tableau de contingence des caractéristiques qui ont été étudiées pour toutes les unités de la population. Dans cet article, on donne la variance asymptotique de l'estimateur du maximum de vraisemblance d'une caractéristique échantillonné, soumise aux contraintes marginales dudit tableau de contingence. A partir de cette variance on peut calculer, dans une étude empirique, la diminution de l'efficacité des estimateurs d'échantillon en formation relatifs à l'estimateur du maximum de vraisemblance.

REFERENCES

[1] Arora, H.R. and Brackstone, G.J. (1977), "An Investigation of the Properties of Raking Ratio Estimators With Simple Random Sampling", Survey Methodology, Vol. 3, No. 1, 62-83.

[2] Bankier, M.D. (1977), "Properties of Raking Ratio Estimators", Master's Thesis, Carleton University.

[3] Rao, J.N.K. (1974), "Raking Ratio Estimator - III", Technical Report, Statistics Canada.

[4] Silvey, S.D. (1959(, "The Lagrangian Multiplier Test", Ann. Math. Statist., 30, 389-407.