AN EVALUATION OF BOUNDED RAKING RATIO ESTIMATION IN THE STATISTICS OF INCOME CORPORATE PROGRAMS

Jeri Mulrow, H. Lock Oh and Richard Collins
Internal Revenue Service (R:S:P), P.O. Box 2608, Washington, DC 20013-2608

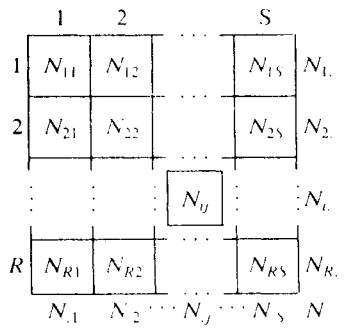
Raking ratio estimation, or more simply raking, is a common technique used in sample surveys. This paper looks at a modification of raking called bounded raking ratio estimation as applied to the Internal Revenue Service's (IRS) Statistics of Income (SOI) corporate programs. Bounded raking ratio estimation was introduced by Oh and Scheuren in 1987 . This paper explores in more depth one of the constraints placed on raking, namely bounding, in the 1987 paper.

The paper begins with a brief introduction of raking ratio estimation. Then the SOI corporate programs are described. Next is an explanation of how bounded raking ratio estimation is used, followed by how it is evaluated using bootstrapping techniques. Finally, a results and conclusions section is presented with a brief note about future research.

INTRODUCTION

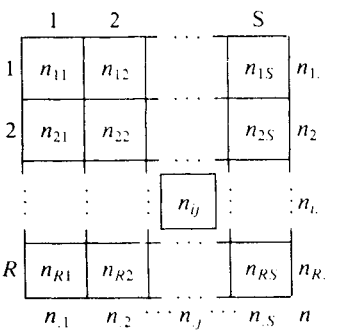
Raking ratio estimation, as named, was first proposed in a paper by Deming and Stephen (1940), as a way of assuring consistency between complete count and sample data from the 1940 U.S. Census of Population. Since then, advances and modifications have been numerous. A reasonably complete bibliography of the statistical research on raking prior to 1987 can be found in Oh and Scheuren (1987).

Raking ratio estimation usually assumes that two (or more) marginal population totals, say $N_{i.}$ and $N_{.j}$ are known, but that the interior of the table N_{ij} can only be estimated from the sample by, say n_{ij}, where graphically (Deming, 1943) we have



with I = 1, ..., R and J = 1, ..., S.

The corresponding sample count table is



with i = 1, ..., R and j = 1, ..., S.

In simple random sampling, the raking algorithm begins by setting

$$N_{ij} = (N/n)\, n_{ij}$$

and then proceeds by proportionately scaling the N_{ij}, such that the relations

$$\Sigma N_{ij} = N_{i.} \text{ and } \Sigma N_{ij} = N_{.j}$$

are satisfied in turn. Each step in the algorithm begins with the results of the previous step, with the N_{ij} continuing to change. The process terminates either after a fixed number of steps or when both summations are simultaneously satisfied to the closeness desired.

In the SOI corporate programs, the interior of the table, the N_{ij}, is actual population values. The raking process is employed as a systematic way to handle cells in the table where the n_{ij} are small. Conventional collapsing alternatives are available, e.g., Cochran (1977) and Fuller (1966), but did not seem adequate. (See Oh and Scheuren, 1987.)

Under fairly general conditions, raking ratio estimation in contingency tables is optimal for estimating N_{ij}, given just $N_{i.}$ and $N_{.j}$. Unfortunately, its use in adjusting sample weights, W_{ij} , where $N_{ij} = W_{ij}\, n_{ij}$, is not always successful. One of the reasons for this is that if the variables used in the raking are not highly correlated with all of the variables in the sample, the weighting adjustments may lead to some degradation in variance for uncorrelated variables. Minimizing any potential detrimental impacts from raking is a primary concern in the corporate program and motivated the modifications presented in Oh and Scheuren (1987).

One of these modifications to the raking process

was to constrain the raking adjustments so that they fall within a relatively narrow range. This approach is referred to as bounded raking ratio estimation and has often been employed in simple ratio estimation. (See Hanson, 1978.) Studying these bounds will be the focus of this paper.

THE SOI CORPORATE PROGRAMS

The Internal Revenue Service has produced statistics on corporate tax returns annually since 1916. The major publications of these data include the annual publications, *Statistics of Income Corporation Income Tax Returns* (U.S. Department of Treasury, 1990) and the *Source Book of Statistics of Income -- Corporation Income Tax Returns* (U.S. Department of Treasury, 1990). The *Source Book* contains detailed tabulations of data available from SOI, featuring income and balance sheet data classified by industry type and size of total assets. The broadest industry level in the *Source Book* is the twelve industrial divisions, and the lowest level is the 185 minor industries.

Up through 1951, corporate statistics were based on a complete census of the returns filed. Since then, stratified probability sampling has been employed, with the current sample sizes averaging about 82,500 returns annually from a population of nearly four million returns. Measures of assets and income are the principal stratifying variables. (See Jones and McMahon, 1984, and Mulrow, 1990.)

Since 1980, a bounded raking ratio estimation approach to post-stratification by industry has been employed to produce the estimates. During the '80's, several research papers, including Leszcz, Oh and Scheuren (1983) and Oh and Scheuren (1987) were written. They describe the raking, modified raking and bounded raking techniques considered and used.

BOUNDED RAKING RATIO ESTIMATION IN SOI CORPORATE FILE

To study the effects of different bounding limits on raking ratio estimation, simulation studies using 1987 SOI corporate data are employed. The coefficients of variation (CV) for three selected variables are computed utilizing bootstrapping techniques.

The data consist of 77,393 records containing three variables -- Business Receipts, Net Income, and Total Assets -- collected from tax returns filed with the IRS. The data are initially stratified into 10 size strata and post-stratified into 47 industry classes. To ensure stability in the raking process, the 47 industry classes were collapsed from the original 58 classes of interest in the *Source Book* . The use of different collapsing techniques will be the subject of another paper and still requires further investigation.

The data are classified into the 10 x 47 = 470 cells for raking and estimation. Bootstrap samples (Efron and Tibshirani, 1990) from each cell are taken and estimates for the three named variables are calculated for each estimator listed in Figure 1. The raking is complete when the raking equations are within a tolerance of 0.00001. During the first phase of the simulations only ten bootstrap samples are taken. Then, for the second phase, the results are based on 100 bootstrap samples from each of the 470 cells. The results and conclusions from the two phases are presented in the next section.

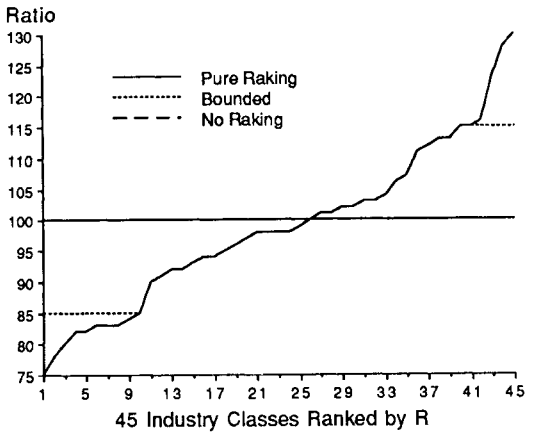
Figure 1. - - Description of Estimators used in Raking

Estimators	Labels
Original Weight (Normal Stratified Sample Estimation based on size strata only)	Original
Pure Raking Ratio Estimation	Pure
Pure Raking Ratio Estimation Excluding cells with 500 or more observations*	Pure - 500
Bounded Raking Ratio Estimation Excluding cells with 500 or more observations* (1st phase)	(0.85, 1.15) ($\sqrt{2/3}$, $\sqrt{3/2}$) (0.75, 1.25)
Bounded Raking Ratio Estimation Excluding cells with 500 or more observations* (2nd phase)	(090, 1.10) ($\sqrt{2/3}$, $\sqrt{3/2}$) (0.70, 1.30)
Averaged Bounded Raking Ratio Estimation Excluding cells with 500 or more observations* (1st phase)	A(0.85, 1.15) A($\sqrt{2/3}$,$\sqrt{3/2}$) A(0.70, 1.30)
Averaged Bounded Raking Ratio Estimation Excluding cells with 500 or more observations* (1st phase)	Averaged

* For cells with 500 or more observations, simple ratio estimates are used.

In bounded raking ratio estimation, bounds are placed on the ratio, R, equal to the raked weight from each cell divided by the original weight for the size category. The bounding assures that the post-stratified weighting adjustments are not wildly different from the original weighting adjustments but allows for difference due to post-stratification. Figure 2 shows this concept pictorially for one simulation.

Figure 2. -- Illustration of Bounded Raking



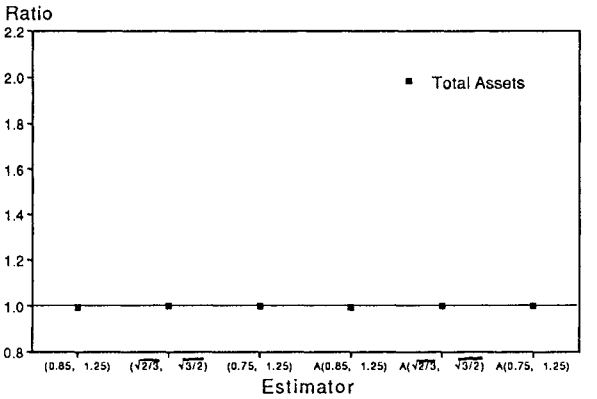
In averaged bounded raking ratio estimation, the raking process is carried out as usual, but afterwards the weights from the raking adjustments are ranked. The ranked weights are then averaged over adjoining cells, so that the cell sizes of the averaged adjustments are at least twenty-five. If the cell size is originally twenty-five, it may be left alone or combined with an adjacent cell that is smaller than twenty-five. Averaging provides stability in calculating variance estimates.

RESULTS AND CONCLUSIONS

Preliminary results for the first phase, the ten bootstrap samples, are shown in Figure 3 for bounded raking adjustments and averaged bounded raking adjustments . The results are presented in terms of ratios to the bounded adjustments using the ($\sqrt{2/3}$, $\sqrt{3/2}$) bounds. Because of the small number of simulations, the results in Figure 3 are uninteresting; there is not much difference between the estimates using the different bounding schemes. That is, all the ratios lie near one. Thus, for the second, more expensive stage of 100 bootstrap samples, the bounding limits were broadened. Also, only one averaged bounded raking adjustment was calculated using the ($\sqrt{2/3}$, $\sqrt{3/2}$) bounds in the second phase.

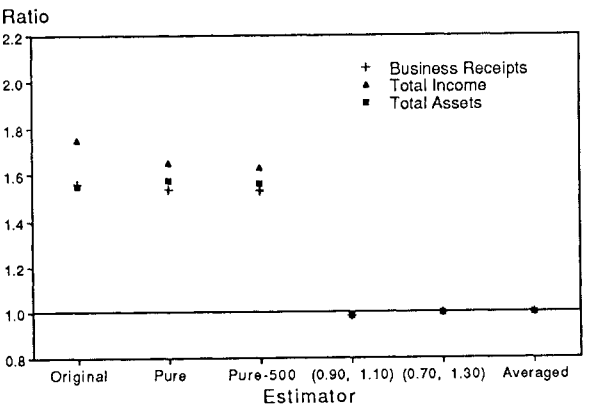
The results from the second phase are presented in Figures 4 - 8 for selected major industries and one size category. All results are presented in terms of ratios of CV's from each estimator to the ($\sqrt{2/3}$, $\sqrt{3/2}$) bounded estimator -- the latter being the one of most interest in the SOI corporate programs and the one that has been in use for several years. Ratios greater than one indicate that the CV for that estimator was greater than the CV from the ($\sqrt{2/3}$, $\sqrt{3/2}$) bounded estimator. Ratios smaller than one indicate that the estimator has a smaller CV and may be considered for future use in the programs.

Figure 3. -- Raking Ratio Estimation Preliminary Bounded Results (Phase 1)



The ratios for all three variables, Business Receipts, Total Income, and Total Assets are presented in Figure 4. The same general pattern can be seen for all the variables in the Manufacturing: Chemical & Allied Products Industry. This result holds over all of the other industries. Even though the levels of the ratios differ for the three variables, the same pattern is evident. Therefore, the presented results in Figures 5 - 8 are given for only one variable at a time and make for simpler charts.

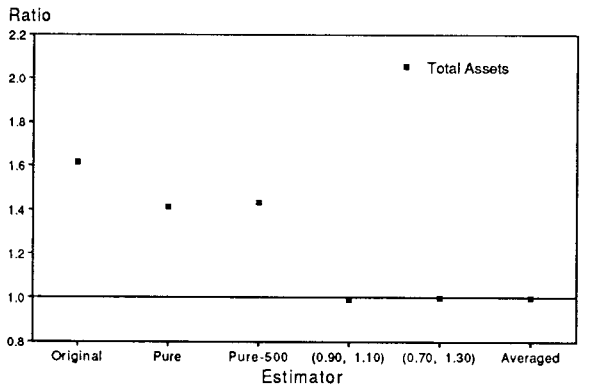
Figure 4. -- Raking Ratio Estimation for Manufacturing: Chemical & Allied Products



Typical results from three different industries are shown in Figures 4 - 6. In these cases, the original estimator produces the largest CV. Pure raking and

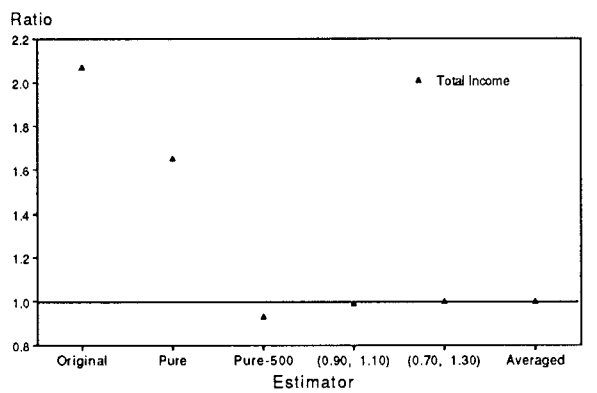
pure raking excluding cells with sample sizes over 500, Pure - 500, produce estimates with smaller CV's than the original estimator, but the bounded raking and bounded averaged raking produce the smallest CV's in general. Sometimes, in industries where there are many cells with 500 or more observations, such as Wholesale and Retail Trade, removing these cells and using simple ratio estimates decreases the CV's.

Figure 5. -- Raking Ratio Estimation for Electric, Gas and Sanitary Services



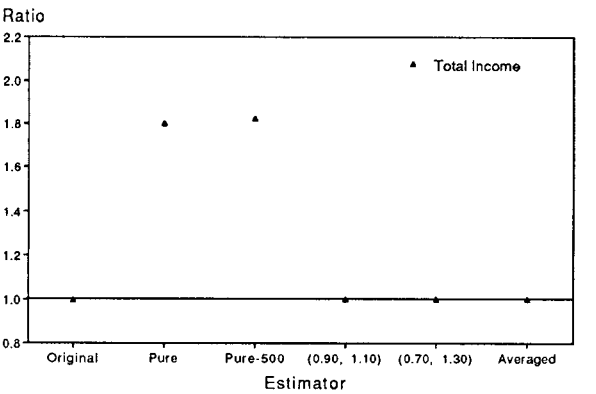
Since the ratios from the bounded adjustments are very close to one, there is little evidence to indicate that the specific choice of bounding limits makes a difference in the SOI corporate programs. However, it would seem reasonable that at some point when the bounds are either very restrictive or very loose, the bounding will cease to be an advantage. This may suggest to the reader that one should investigate one's own data carefully before determining bounding limits to be used in bounded raking ratio estimation.

Figure 6. -- Raking Ratio Estimation for Wholesale and Retail Trade



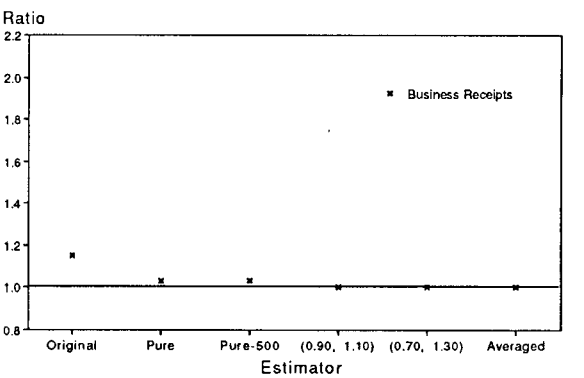
The ratios shown in Figure 7 are only representative of a small number of industries in which there are a small number of corporations in the population. In these cases, the Pure and Pure - 500 raking adjustments cause the CV's to be larger than the Original. The bounded raking ratio process, however, brings the estimates back in line with the Original estimates.

Figure 7. -- Raking Ratio Estimation for Manufacturing: Textile Mill Products



One desired property of the bounded raking ratio procedure is that it not hurt the estimates if post-stratification is not used. The results in Figure 8 show how the estimates compare over one original size strata. It turns out that the bounded raking ratio adjustments do at least as well, if not better, than the Original weighting adjustments. Thus, the bounded raking process can be used whether post-stratification is desired or not.

Figure 8. -- Raking Ratio Estimation for Sample Strata 1



SUMMARY OF FINDINGS

To summarize, in general, the post-stratification process provides for smaller coefficients of variation than the Original non-post-stratified process. For a small number of industries, such as Manufacturing and Textile Products, the pure raking can cause the estimates to have larger CV's. Fortunately, estimates using bounded raking and averaged bounded raking

adjustments are at least as good if not better than using pure raking. The use of different bounding limits did not seem to have much of an effect on the estimates nor did averaging. Thus, the bounding limits of ($\sqrt{2/3}$, $\sqrt{3/2}$) will continue to be used in the SOI corporate programs. Finally, the bounded adjustments did not adversely effect the estimates if post-stratification is not desired.

NEXT STEPS

There are still several questions that are left unanswered by this paper. Some of them to be considered in future research include:

- What is the optimal collapsing scheme?
- What is the optimal cell size limitation?
- How can methods of variance estimation be improved further?
- What would the effect be of incorporating information from prior years into the raking process?

The most likely next steps in this evaluation process will be to consider the first two questions above and attempt to answer the last question.

ACKNOWLEDGMENTS

The authors would like to extend their thanks to Wendy Alvey and Beth Kilss for their assistance with the oral and written presentations of this paper.

REFERENCES

COCHRAN, W.G. (1977). *Sampling Techniques, (3rd ed.).* New York: John Wiley.

DEMING, W.E. (1943). *Statistical Adjustment of Data.* New York: Dover.

DEMING, W.E., and STEPHEN, F.F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known, *Annals of Mathematical Statistics,* 11, 427-444.

EFRON, B., and TIBSHIRANI, R.J (1990). An Introduction to the Bootstrap, seminar presented for the Washington Statistical Society and the Center for Computational Statistics, George Mason University, Fairfax, VA.

FULLER, W.A. (1966). Estimation Employing Post Strata, *Journal of the American Statistical Association,* 61, 1172-1183.

HANSON, R. (1978). The Current Population Survey: Design and Methodology, Technical Paper No. 40, U.S. Bureau of the Census.

JONES, H., and MCMAHON, P.B. (1984). Sampling Corporation Income Tax Returns for Statistics of Income, 1951 to Present, *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 269-274.

LESZCZ, M.; OH, H.L.; and SCHEUREN, F. (1983). Modified Raking Estimation in the Corporate SOI Program, *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 107-111.

MULROW, J. (1990). Description of the Sample and Limitations of the Data, *Statistics of Income - 1987 Corporation Income Tax Returns,* Publication 16, Internal Revenue Service, U.S. Department of Treasury.

OH, H.L., and SCHEUREN, F. (1987). Modified Raking Ratio Estimation, *Survey Methodology,* 13, 2., 209-219.

U.S. DEPARTMENT OF TREASURY (1990). *Statistics of Income - 1987 Corporation Income Tax Returns,* Publication 16, Internal Revenue Service.

U.S. DEPARTMENT OF TREASURY (1990). *1987 Source Book of Statistics of Income -- Corporation Income Tax Returns,* Internal Revenue Service.