# AN INVESTIGATION OF STRATIFICATION ERRORS

Jeri Mulrow and Louise Woodburn, Internal Revenue Service
P.O. Box 2608, Washington DC 20013-2608

**KEY WORDS:** Stratified Random Sampling, Simulation Study, Bias Estimation.

In stratified random sampling, the initial values of the stratifying variables are very important to the sampling and estimation process. Population counts and sample selection depend on this information. Errors in the stratifying variables can cause problems with the whole sampling and estimation process. This paper presents a simulation study using corporation income tax data to show the effect of stratification errors on estimates. Several techniques which can be used to adjust for such errors are considered.

## INTRODUCTION

Stratified random sampling is often used when the population of interest can be divided into several distinct subpopulations. In many cases a measure of the size of the sampling unit is used to divide the population. For example when sampling corporate income tax returns, size could be measured by using total assets. In a study of wheat production, the acreage of a farm may be used to stratify the sample.

In stratified sampling, the population of N units is divided into L distinct nonoverlapping subpopulations $N_1, N_2, \ldots, N_L$ called strata. Once the strata have been set, then a sample is drawn from each. If a simple random sample is taken from each stratum then the procedure is known as stratified random sampling.

Typically the values of the $N_i$, i = 1, 2, ..., L are known and a set sample size of $n_i$ is taken from each stratum with an overall sample size of $n = \Sigma n_i$, i = 1, 2, ..., L. Proportionate sampling or optimum allocation techniques can be used to determine the $n_i$. Estimates of population means and totals can be calculated following standard formulas given in many statistical texts.

The two greatest advantages of using stratified random sampling over simple random sampling are a gain in precision of the estimates and a reduction in the overall cost of obtaining the sample. However, several factors can affect these benefits significantly. Many of these aspects are discussed in Cochran (1977), and the reader is referred there for further discussions.

This paper addresses the effects of misclassification on the estimates. Data from the 1985 and 1986 Statistics of Income annual Corporate Income Tax studies will be used as an illustration. In this particular study the $N_i$ , i = 1, 2, ..., L are not known until the end of the sampling process. Approximately three and one-half percent of the sample is known to be misclassified each year. The sample and the population are classified at the same time, so it is presumed that about three and one-half percent of the population is also misclassified. For a detailed description of the sampling procedure and further processing, see Mulrow (1990).

Misclassified sampled units can be easily detected and reclassified using additional information collected during the data gathering phase. The question that remains is how to treat the misclassified population units. According to Deming (1960), "A real universe is dynamic, not static, and the information that is used for classification is always to some extent out of date. ... Misclassification ... is thus expected as the natural course of events, " thus suggesting that nothing needs to be done about misclassification. Cochran (1977) shows, though, that the estimates are biased when the weights, $W_i = N_i / n_i$, are not known exactly.

Since, in the Corporate Income Tax study, the population and sample strata sizes include misclassified units, problems with the estimates may exist. Three methods to handle misclassification in the population are proposed in this paper. Simulation studies are used to evaluate the different methods and the effects on the estimates.

This paper is organized into seven sections, including the Introduction section. The Background section gives a short description of the Corporate Income Tax study. The three proposed methods to handle misclassification are presented in the Problem Description section and a numerical example follows in the next section. An outline of the simulation study is given next. The Results and Conclusions section includes several figures showing results of the simulation study. Finally, a Future Work section is included which discusses further work needed.

## BACKGROUND

The population of interest in the Corporate Income Tax study consists of over 3.5 million corporations filing U.S. income tax forms. The distribution of the population is heavily skewed with only a very small percentage of the corporations accounting for more than three quarters of the total assets and income. In the 1986 study, the top 0.13% or 4471 corporations accounted for over 78% of the U.S. total assets while the lower 51% or 1,736,486 corporations accounted for less that 1/2 of 1% of the total assets.

A stratified probability sample of approximately 85,000 corporate tax returns is taken yearly with the largest corporations selected with certainty. In the 1986 study the sampling rates ranged from 0.33% to 100%. The two principal variables used for stratification were total assets and proceeds, where proceeds is defined as the larger of the absolute values of net income or absolute value of net income plus depreciation plus depletion.

The causes of misclassification in the study have been traced to problems in the initial values of the stratifying variables. In particular, it has been found that over 75% of the misclassifications arise from converting dollars and cents to dollars or converting dollars to dollars and cents during data transcription. Other causes of misclassification included substitution of one number for another number and data transcription errors. For a more detailed analysis of the misclassification errors in the corporate study see Mulrow and Jones (1989).

## PROBLEM DESCRIPTION

Some notation will be useful in describing the problem. Let:

$n_i$ = sample size of stratum i;
$N_i$ = population size of stratum i;
$w_i = N_i / n_i$ = weight of stratum i;
$f_{ih}$ = number of sample units misclassified in stratum i that belong to stratum h;
$p_{ih} = f_{ih} / n_i$ = proportion of sample units misclassified in stratum i that belong to stratum h;
$n_i'$ = adjusted sample size of stratum i;
$N_i'$ = adjusted pop. size of stratum i;
$w_i' = N_i' / n_i'$ = adjusted weight of stratum i.

The problem can then be simply stated as: Find $w_i' = N_i' / n_i'$ for i = 1, 2, ..., L such that the estimates derived from the sampled data are the "best" estimates. Best is defined later.

In this paper, three different methods are proposed to determine the adjusted $n_i'$ and $N_i'$. The three proposed methods are now discussed below.

The first method, which is used as a basis for comparison, ignores any misclassification that may have occurred. That is, $n_i' = n_i$, $N_i' = N_i$ and $w_i' = w_i$. However, the values of the misclassified variables are corrected. This approach will be referred to as the Basic approach.

The second proposed method uses information about the misclassification in the sample to adjust both the sample and population sizes. The approach is to adjust the population stratum sizes by the same values by which the sample stratum sizes are adjusted. This method takes into account all of the misclassification seen in the sample without any extrapolation to other misclassified population units not seen in the sample. It will be referred to as the Unweighted approach. A numerical example of this method is given below.

The last proposed method also uses the information in the sample but tries to extrapolate to the unseen misclassified population units. The approach uses a proportionate or weighted adjustment to the population stratum sizes based on the sample misclassifications. Using this method, if 3% of the sampled units were misclassified into other strata then a corresponding 3% of the population units in the initial stratum would be reclassified into the other strata. This approach will be referred to as the Weighted approach.

## EXAMPLE

In this very simplified example, the population is stratified into three strata. The sample size is 40 and the population size is 400 giving an overall sampling rate of 10%. Twenty of the sampled units are misclassified, as shown in Figure 1. Diagonal cells in Figure 1 represent the number of sampled units that were initially classified correctly. Off-diagonal cells represent the number of sampled units that were incorrectly classified and the movement after reclassification. For example, $f_{12} = 5$ is the number of sampled units that were initially misclassified in stratum 1 that should be reclassified to stratum 2.

Figure 1. -- Misclassified Sample Units

ADJUSTED STRATUM

| ORIGINAL STRATUM i \ h | 1 | 2 | 3 | Original Sample |
|---|---|---|---|---|
| 1 | 10 | 5 | 5 | 20 |
| 2 | 2 | 5 | 3 | 10 |
| 3 | 3 | 5 | 2 | 10 |
| Corrected Sample | 15 | 15 | 10 | 40 Overall Total |

The original population strata sizes are : $N_1 = 220$, $N_2 = 120$, and $N_3 = 60$. Figure 2 gives the adjusted population and sample strata sizes using all three proposed methods. For the Basic approach, the original sample and population sizes are used to calculate the $w_i$. That is, no adjustment for misclassification is made to either the sample or the population sizes under the Basic method. The values of the misclassified variables are corrected but the sampled units are left in the original sampling stratum.

Under the Unweighted approach, the adjusted population sizes would be $N_1' = 220 - (20 - 15) = 215$, $N_2' = 120 - (10 - 15) = 125$ and $N_3' = 60 - (10 - 10) = 60$. The population stratum sizes are adjusted simply by the total number of sampled units that moved out of or into that particular stratum. That is, the misclassified sampled units are moved into the appropriate stratum and the population and sample sizes are adjusted accordingly.

Figure 2. -- Adjusted Population and Sample Sizes

| h | BASIC APPROACH | | UNWEIGHTED APPROACH | | WEIGHTED APPROACH | |
|---|---|---|---|---|---|---|
| | $n_h$ | $N_h$ | $n_h$ | $N_h$ | $n_h$ | $N_h$ |
| 1 | 20 | 220 | 15 | 215 | 15 | 152 |
| 2 | 10 | 120 | 15 | 125 | 15 | 145 |
| 3 | 10 | 60 | 10 | 60 | 10 | 103 |
| TOTAL | 40 | 400 | 40 | 400 | 40 | 400 |

The adjusted population sizes under the Weighted approach can be found using Figure 3. Each cell in the table is equal to the number of originally classified population stratum units times the proportion, $p_{ij}$, of sampled units that changed stratum. In the example, $p_{23} = 3/10$, that is 3 out of the original 10 sampled units in stratum 2 are reclassified into stratum 3. So, under the Weighted approach, $N_2 * p_{23} = 120 * (3/10) = 36$ is the number of population units that are reclassified from stratum 2 to stratum 3. The row totals are the original population sizes and the column totals are the adjusted population sizes. For the example, these sizes turn out to be $N_1' = 152$, $N_2' = 145$, and $N_3' = 103$.

Figure 3. -- Weighted Population Sizes

ADJUSTED STRATUM

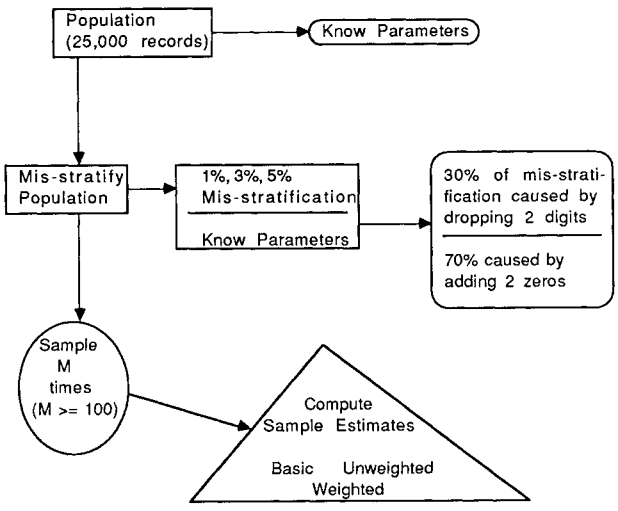
| ORIGINAL STRATUM i \ h | 1 | 2 | 3 | Original Population |
|---|---|---|---|---|
| 1 | $\frac{10}{20}(220) = 110$ | $\frac{5}{20}(220) = 55$ | $\frac{5}{20}(220) = 55$ | 220 |
| 2 | $\frac{2}{10}(120) = 24$ | $\frac{5}{10}(120) = 60$ | $\frac{3}{10}(120) = 36$ | 120 |
| 3 | $\frac{3}{10}(60) = 18$ | $\frac{5}{10}(60) = 30$ | $\frac{2}{10}(60) = 12$ | 60 |
| Adjusted Population | 152 | 145 | 103 | 400 Overall Population Total |

## SIMULATION STUDY

Simulation studies are used to evaluate the estimates produced from using the three different methods outlined above. The flow chart in Figure 4 shows the evaluation process. To begin, a 'Minipopulation' of 25,000 records was created which resembles the Corporate Income Tax study population of over 3.5 million records. The distribution of size units, based on size of total assets, is preserved in the minipopulation. Overall and within strata population parameters are known for the minipopulation and can be used for comparison with the estimates obtained using the three methods.

Misclassification in the minipopulation is created by either adding two zeros at the end of the original total asset amount or by dropping off the last two digits of the total asset amount. Thirty percent of the misclassifications are caused by dropping digits, the other seventy percent are caused by adding two zeros. These types and levels of misclassification are chosen to mimic those seen in the Corporate Income Tax study. Other causes of misclassification are not studied in this paper. Overall, 1%, 3%, and 5% of the population are misclassified in this manner.

**Figure 4. -- Simulation Study Flow Diagram**




Next, stratified random samples are drawn and population estimates are calculated for each sample using the Basic, Unweighted, and Weighted approaches. The strata boundaries, based on size of total assets, sampling rates and population sizes for the eleven strata are given in Figure 5.

Population estimates from 100 samples are used to approximate the expected value of the estimates under each method. Comparisons between the simulated expected values and the true population parameters are presented in terms of percent bias. Precision estimates are not available at the time of printing but will be forthcoming. 'Best' will be defined in terms of percent bias and precision at that time.

## RESULTS AND CONCLUSIONS

Since precision estimates are not available at the time of printing, the results from the simulation experiments will be discussed in terms of percent bias. One of the main objectives of the Corporate Income Tax study is to produce good overall population estimates for hundreds of tax variables including size estimates. In this paper, five tax variables, Total Assets, Net Income or Deficit, Accounts Receivable, Inventories and Taxes Paid, are studied along with size. Size is considered in this list, since the total misclassification in the population is unknown and thus the population sizes gathered at the time of sample selection are not the true values.

Total assets is the main stratifying variable in the minipopulation (and the Corporate study); thus, studying the estimates from the three methods on this variable will be important. Net Income or Deficit is another stratifying variable in the Corporate study, although less than five percent of the sampled units are actually classified by their income amount. It is not used to stratify units in the minipopulation, but estimates from this variable are nonetheless of interest. Accounts Receivable was chosen as a variable to study because it has a fairly high correlation of $r = .85$ with Total Assets. The other two variables, Inventories and Taxes Paid, have correlations of $r = .48$ and $r = .27$, respectively, with Total Assets.

Figure 6 shows the true population sizes by stratum of the minipopulation before misclassification. It also presents the difference in the estimated population sizes, using the three proposed methods, from the true values. As the amount of misclassification in the population increases, the estimated sizes using the Unweighted method grow farther and farther away from the true values. On the other hand, both the Basic and Weighted methods appear to give similar results over the varying levels of misclassification, and both are very accurate for the overall size. Since population stratum sizes are of interest in the Corporate Income Tax study, the Unweighted method does not appear to be a good procedure to use based on these results.

Although the level of misclassification had a large effect on the size estimates from the three proposed methods, it did not have much of an effect on the variable estimates in terms of bias. The same patterns arose concerning the three methods no matter what the level of misclassification in the minipopulation. Thus, only results from the 3% misclassified minipopulation are presented here.

The percent bias, in all three proposed methods, for total assets is presented in Figure 7. As expected, the Unweighted method does the poorest, giving large, almost 80%, biases in the upper strata. The Unweighted methods will not be studied further and is not included in the rest of the analyses.

Figure 8 presents the same data as Figure 7 with the Unweighted method estimates eliminated. The From the simulation results, the Basic method appears to provide less bias estimates than the Weighted method, in general, for total assets. In Figures 9 through 12, the percent bias in the estimates for the other four variables is presented. In these cases, it also appears that the Basic method is providing the least biased estimates overall, although, as the correlation of the variable with total assets decreases, the bias in the Basic estimates is more noticeable. The overall totals for the population and the three proposed methods are given in Figure 13.

Based on the results presented in terms of percent bias, the Basic method seems to give the better estimates. The Weighted method is giving slightly larger percent biases in the estimates in almost every case and in some cases is biased in the opposite

Figure 5. --Minipopulation Stratum Sizes and Sampling Rates

| STRATA BOUNDARIES TOTAL ASSETS ( $000'S) | STRATUM | MINI-POPULATION SIZE | SAMPLING RATES |
|---|---|---|---|
| 0 - < 50 | 1 | 10,594 | 0.0036 |
| 50 - < 100 | 2 | 3,526 | 0.0056 |
| 100 - < 250 | 3 | 4,378 | 0.0091 |
| 250 - < 500 | 4 | 2,591 | 0.0197 |
| 500 - < 1,000 | 5 | 1,713 | 0.0340 |
| 1,000 - < 2,500 | 6 | 1,242 | 0.0765 |
| 2,500 - < 5,000 | 7 | 494 | 0.1040 |
| 5,000 - < 10,000 | 8 | 223 | 0.2041 |
| 10,000 - < 25,000 | 9 | 140 | 0.3649 |
| 25,000 - < 50,000 | 10 | 37 | 0.4774 |
| > = 50,000 | 11 | 72 | 1.0000 |
| TOTAL | | 25,010 | |

Figure 6. -- Differences Between Estimated and True Population Sizes

| STRATUM | POPULATION SIZE | 1% MIS-STRATIFICATION | | | 3% MIS-STRATIFICATION | | | 5% MIS-STRATIFICATION | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | BASIC | UNWEIGHTED | WEIGHTED | BASIC | UNWEIGHTED | WEIGHTED | BASIC | UNWEIGHTED | WEIGHTED |
| 1 | 10,594 | -6 | -20 | -6 | 4 | -59 | 4 | 7 | -158 | 7 |
| 2 | 3,526 | 2 | -30 | 1 | -10 | -79 | -10 | 11 | -137 | 11 |
| 3 | 4,378 | 2 | -31 | 2 | -5 | -89 | -5 | 5 | -125 | 5 |
| 4 | 2,591 | -4 | -12 | -4 | -3 | -30 | -3 | -15 | -57 | -16 |
| 5 | 1,713 | -3 | 7 | -3 | 2 | 21 | 2 | -4 | 24 | -4 |
| 6 | 1,242 | 1 | 9 | 1 | 8 | 38 | 8 | -2 | 95 | -2 |
| 7 | 494 | 9 | 24 | 9 | 3 | 53 | 3 | 1 | 122 | 1 |
| 8 | 223 | -1 | 22 | -1 | 2 | 54 | 2 | -6 | 8 | -6 |
| 9 | 140 | 0 | 22 | 0 | 0 | 58 | 0 | 3 | 88 | 3 |
| 10 | 37 | 0 | 9 | 1 | 0 | 29 | 0 | 0 | 51 | 0 |
| 11 | 72 | 0 | -1 | 0 | 0 | -1 | 0 | 0 | -2 | 0 |
| TOTAL | 25,010 | 0 | -1 | 0 | 1 | -5 | 1 | 0 | -91 | -1 |

Figure 7. -- Percent Bias: Total Assets All Methods

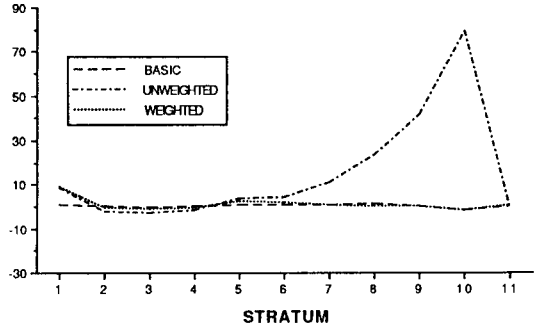



Figure 8. -- Percent Bias: Total Assets

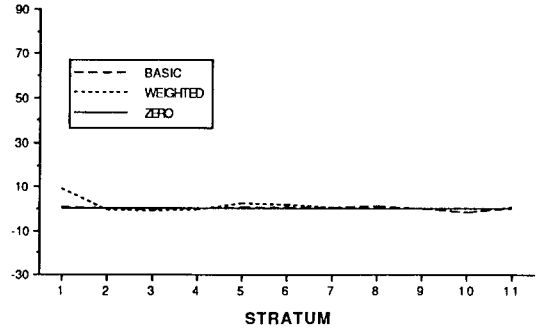



Figure 9. -- Percent Bias: Net Income

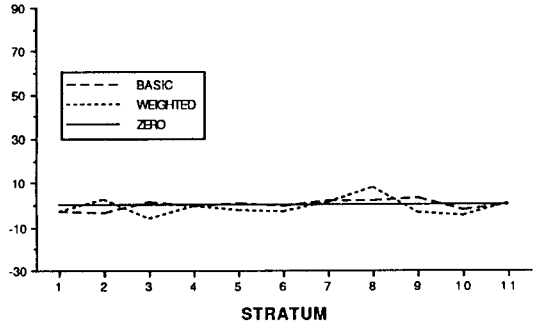



Figure 10. -- Percent Bias: Accounts Receivable

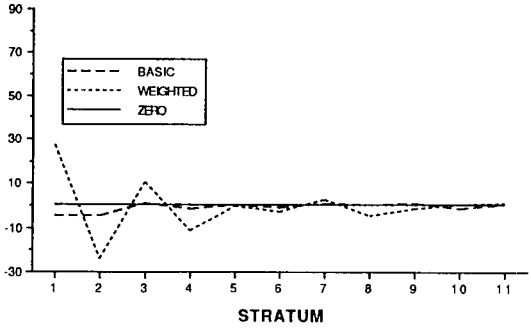


**Figure 11. -- Percent Bias Inventories**

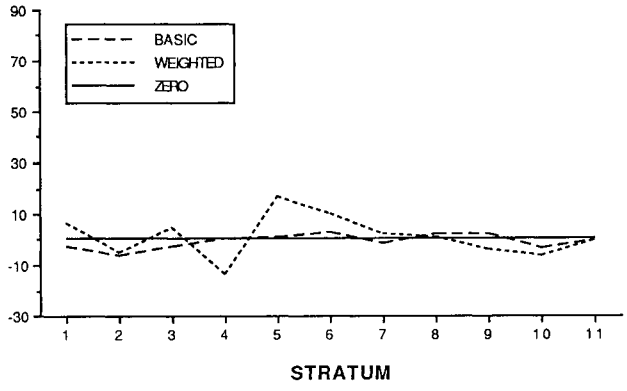



**Figure 12. -- Percent Bias Taxes Paid**

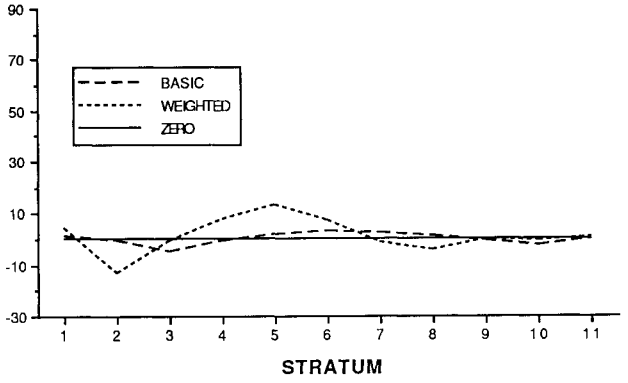



**Figure 13. -- Overall Totals for Population and Proposed Methods**

| | Population | Basic | Unweighted | Weighted |
|---|---|---|---|---|
| **Total Assets** | 61,761,311,258 | 61,745,694,900 | 64,178,190,100 | 62,163,979,400 |
| **Net Income** | 3,789,444,223 | 3,794,427,550 | 3,978,125,450 | 3,787,520,630 |
| **Accounts Receivable** | 17,433,084,860 | 17,410,465,500 | 18,073,833,800 | 17,492,156,300 |
| **Inventory** | 4,610,810,891 | 4,608,547,440 | 4,958,671,520 | 4,647,600,600 |
| **Taxes Paid** | 2,786,116,612 | 2,785,277,800 | 2,879,495,830 | 2,807,599,990 |

direction of the Basic method estimates. At the time of this paper, it is not known why this is occurring and further study is needed. It is clear, however, from these preliminary results that the Unweighted method is not giving good estimates and probably should not be considered as a method for dealing with population misclassification.

## FUTURE WORK

The results presented above are based on preliminary findings about percent bias in the estimates from the three proposed methods to handle misclassification in the sample and the population. Future work will expand the analysis to include variance estimates from the different methods. The real advantage (or disadvantage) of the Weighted method should show up in this latter analysis. Final conclusions concerning the three methods will be made taking into account both bias and precision.

More levels of misclassification in the population will be studied. It is of interest to know if, at some particular level of misclassification, the Basic method estimates become more biased than the Weighted method estimates. Currently it is planned to study additional levels of 10%, 15% and 25%.

Also the amount of misclassification in the samples will be forced to achieve a particular level, especially in the upper strata. It is known in the Corporate Income Tax study, that a larger number of misclassifications occur in the upper strata due to the addition of two zeros to the stratifying variables. In the simulation studies conducted above, the misclassifications in the minipopulation were left to a random process to achieve the overall rate. This will be changed to force more misclassifications to occur in the upper strata.

## ACKNOWLEDGMENTS

The authors would like to express their thanks to the many people who reviewed the paper and, in particular, to Beth Kilss for her help in organizing the ASA posters and this paper.

## REFERENCES

Cochran, W. G. (3rd ed., 1977), Sampling Techniques, John Wiley & Sons, New York.

Deming, W. E., (1960), Sample Design in Business Research, John Wiley & Sons, New York.

Jones, H. W., and McMahon, P. (1984), Sampling Corporation Income Tax Returns for Statistics of Income, 1951 to Present, Proceedings of the American Statistical Association, Section on Survey Research Methods 437- 442.

Mulrow, J. (1990), Description of the Sample and Limitations of the Data, Statistics of Income 1987, Corporation Income Tax Returns, Publ. 16, 9-15.

Mulrow, J., and Jones, H. W. (1989) Sampling Administrative Records: Detection and Correction of Stratification Error, Statistics of Income and Related Administrative Record Research: 1989-1990.