# SAMPLING ADMINISTRATIVE RECORDS: DETECTION AND CORRECTION OF STRATIFICATION ERROR

Jeri M. Mulrow and Homer W. Jones, Jr., Internal Revenue Service

## INTRODUCTION

The U.S. Internal Revenue Service (IRS) is required by law [Sec. 6108 of the IR Code] to publish annual statistics with respect to the income tax laws. The Statistics of Income Division and its predecessors have been publishing statistics on both individual and corporate income tax returns since 1913. This paper will concentrate on the corporate income tax sample. Corporation statistics were first based on a tabulation of all corporate tax returns; then, after 1950, on a sample of these returns. To represent the diversity of organizations filing corporation income tax returns, the current samples are stratified by form type, certain industry characteristics, size of total assets and size of income.

Due to the large number of returns being processed throughout the program, there is a good chance that some of the returns will be misclassified, just as in any other operation which depends on clerical assistance. Several causes of misclassification or mis-stratification can be identified and measured in the sample. This paper focuses on the detection and correction of stratification errors in the corporate sample. The 1985 income year corporation sample will be used for illustration. This year refers to corporations which had an ending accounting period in the time frame July 1985 through June 1986. The topics discussed briefly include relevant features of the corporate sample design, cause of, detection of, and assumptions needed to correct for mis-stratification.

## BACKGROUND

The main purpose of the corporate sample is to produce estimates of economic and tax variables for the primary users of the data--the Bureau of Economic Analysis of the Commerce Department, the Office of Tax Analysis of the Treasury Department, and the Joint Congressional Committee on Taxation. For example, estimates for 1985 corporations show that:

- total assets were equal to $12.8 trillion;
- total deductions for companies with assets $250 million or more were equal to $4 trillion; and
- total liabilities, including equity for construction companies with total assets less than $100 thousand, were equal to $5.4 billion.

Initially, a census of all returns was used to produce these and other statistics. Over the years, the population grew so large that it was no longer practical to do this because of budgetary constraints. In 1951, stratified sampling was introduced and the Statistics of Income (or SOI) sample for corporations was born, in which over 40% of the population was sampled. The sampling rate has decreased over the years and, for the 1985 SOI income year, less than 3% (or 94,150 returns) of the 3,569,609 corporate tax returns were sampled.

A relatively small number of corporations account for more than half of the total assets and income. In 1985, the top 0.12% (or 4052) active corporations accounted for 77% of the U.S. total assets, while the bottom 56% (or 1,833,451) active corporations accounted for less than 1/2 of 1 percent of the total assets. Therefore, the sample is stratified to handle this skewness, and the largest corporations are sampled with certainty. The rates range from 0.35% to 100%, as shown in Table 1.

Table 1.--1985 SAMPLE RATES BY STRATUM

| Stratum* | Sample Rates |
|---|---|
| 1 | 0.0035 |
| 2 | 0.0050 |
| 3 | 0.0085 |
| 4 | 0.0180 |
| 5 | 0.0320 |
| 6 | 0.0760 |
| 7 | 0.1050 |
| 8 | 0.2050 |
| 9 | 0.3800 |
| 10 | 1.0000 |

*Stratum allocation is based on the amount of total assets and absolute value of Net income/deficit reported by a business in a given tax year.

The two principal variables used to stratify the 1985 sample were total assets and absolute value of net income/deficit [1]. These variables were chosen to represent the balance sheet and the income statement on a corporate return. They summarize these statements well and, thus, seem to be appropriate to use for stratification. Other choices are possible, such as total receipts or business receipts, but these may change dramatically from year to year, and so are not presently used.

There are several different types of corporate income tax forms that may be filed. The majority of companies file on Forms 1120, 1120A, or 1120S. The number filing on these returns make up 3,540,113 (or 99.2%) of the 1985 population. Thus, the paper will concentrate on this group of corporations and the others will not be discussed here.

The strata are corner shaped (see Figure A) to handle the two-dimensionality of total assets and income as stratifiers. A corporate tax Form 1120, 1120A, or 1120S is placed into a particular stratum based on the larger of total assets or income. The sampling rates are determined by a Neyman allocation. For a more detailed description of the sample design see Internal Revenue Service (1987).

## SAMPLE SELECTION PROCESS

Each filing corporation has a unique Employer Identification Number (EIN), which is similar to an individual's social security number, for identification purposes. This EIN is transformed into a pseudo-random number and a company's return is selected to be in the sample based on its transformed number and its stratum classification. If the transformed number is less than the sampling rate for a particular stratum, then the company's return is included in the sample, otherwise it is not.

Sample selection for the SOI corporation sample is done from the IRS's Business Master File (BMF). The BMF contains a limited amount of tax and administrative information about all tax filing companies and is maintained by the Tax Processing Systems Division of the IRS. Data are not used directly from that file, because the information on the BMF is maintained for tax collection and auditing purposes and, hence, is not well suited for economic data analysis.

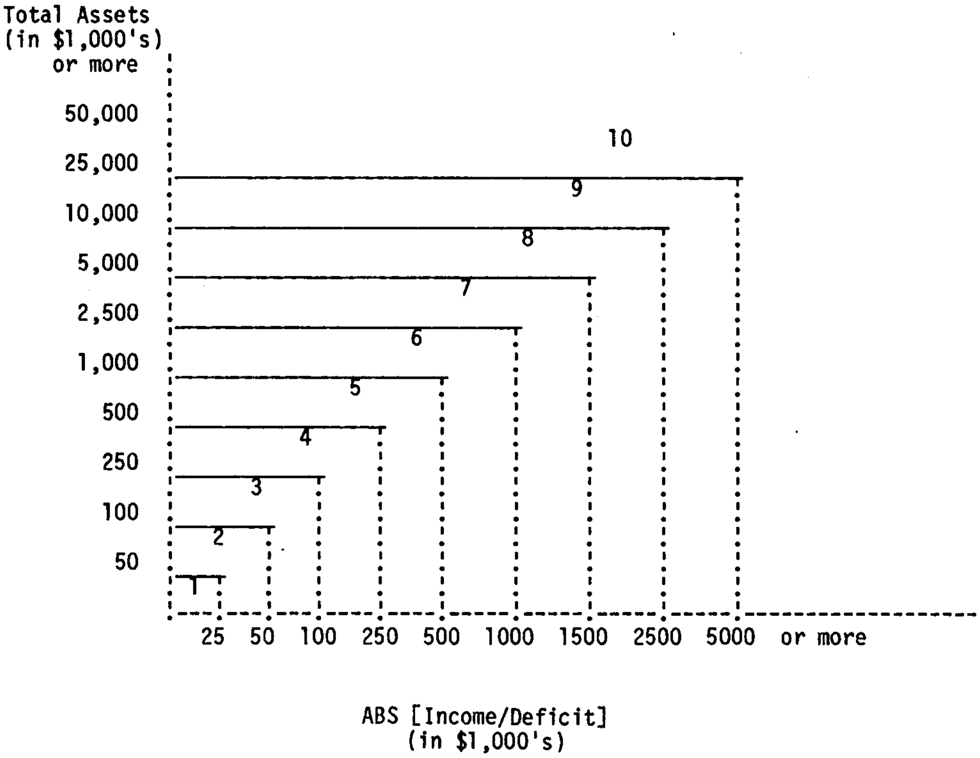
Instead, for SOI purposes, additional tax information is collected directly from the original tax return. Thus, if a corporation is designated to be in the sample, then its return is physically retrieved by SOI from one of the ten IRS Service Centers, where all IRS returns are received and processed for revenue purposes. For a more detailed description of the data collection and abstraction processes see Internal Revenue Service (1989).

Selected tax information is verified and validated before it is entered onto the BMF (i.e., before sample selection). However, errors in the stratifying variables may still exist at this stage, particularly since these amounts often can be verified and validated only when additional tax information is entered onto the file during SOI processing.

Two types of problems can arise from errors in the stratifying variables. First, a corporation may be in the sample when it should not. Its transformed number might be less than the sampling number (10,000 x sampling rate) for its selected stratum but not for its true stratum. This case occurs if large amounts are substituted for smaller amounts, e.g., extra digits are mistakenly added to one or more of the stratifying variables. On the other hand, a company may be omitted from the sample because its transformed number is too large for a particular stratum, but not too large for its true stratum. This case occurs if small amounts are substituted for larger amounts, e.g., digits

Figure A.--1985 STRATUM ALLOCATION SCHEME FOR 1985 STRATUM CLASSES

Regular Fiscal Tax Returns



are mistakenly dropped in one or more of the stratifying variables or the stratifying variable is blank (missing).

## CAUSES OF MIS-STRATIFICATION

There are several reasons why a return may be mis-stratified. Since the information from a corporation tax return must be entered manually by input processors, the data are subject to input error.

There are, at present, eight different causes of mis-stratification recorded by SOI Division. Some of these are unique to the Statistics of Income file, while others may occur in any type of survey sample. This paper will consider only the causes which may occur in any type of sample. They are:

- added digits in a stratifying variable;
- dropped or missing digits in a stratifying variable; and
- one variable amount substituted for another or an editing change to a stratifying variable.

The IRS does perform data input verification at the time of initial processing but, due to the large number of tax returns being processed in a limited amount of time, problems with the data may still be present. Recently, the most common cause of mis-stratification in the sample has been the presence of added digits in the stratifying amounts. Most often this is caused by entering dollars and cents where only dollars were to have been entered. The tax instructions may contribute to this, because a taxpayer is given the option of using whole dollar accounting or the more complete dollars and cents accounting. Table 2 gives tabulations of mis-stratification causes by stratum.

Table 2.--1985 MIS-STRATIFICATION COUNTS BY CAUSE AND STRATUM

| Stratum | Total | Causes | | | |
|---|---|---|---|---|---|
| | | Added digits | Dropped digits | Subs/ changes | Other |
| 1 | 28 | 0 | 5 | 3 | 20 |
| 2 | 21 | 3 | 2 | 6 | 10 |
| 3 | 35 | 6 | 5 | 6 | 18 |
| 4 | 58 | 10 | 10 | 15 | 23 |
| 5 | 56 | 16 | 5 | 10 | 25 |
| 6 | 137 | 79 | 9 | 13 | 36 |
| 7 | 120 | 76 | 5 | 8 | 31 |
| 8 | 318 | 218 | 12 | 26 | 62 |
| 9 | 570 | 469 | 8 | 22 | 71 |
| 10 | 2106 | 1695 | 46 | 169 | 196 |
| TOTAL | 3449 | 2572 | 107 | 278 | 492 |

## DETECTION OF MIS-STRATIFICATION

The SOI corporate sample is subject to many types of "consistency tests" before the data are used to produce any estimates of income and taxation. During this stage, it is possible to cross-check entries on a tax return to determine if a return has been correctly stratified. The first stage in detection is done by computer. Inconsistencies found in the first stage are then manually examined for possible mis-stratification.

The extra information added by SOI includes entries from the various tax schedules filed along with the return. From this added information, net income/deficit and total assets can be computed. These computed fields are used in the consistency tests. If the stratum for the corporation using the computed fields is different than the original stratum assigned the corporation, then the data are subject to a manual examination to determine the cause of mis-stratification. Many times, the problem is easily explained, as in the case of missing or added digits in total assets.

The following three examples illustrate how mis-stratification might be detected in the SOI corporation sample. They show how various items relating to the stratifying variables can be cross-checked to determine the correct stratum.

- Example 1.--Digits are Dropped from the Stratifying Variable

| Element | Assume a corporation return has the following items available: | |
|---|---|---|
| 1 | Total assets after SOI editing ............... | $24,985,000 |
| 2 | Total assets as read from BMF file ......... | 249,850 |
| 3 | Initial stored entry of Element 2 ............. | 24,985,000 |
| 4 | Prior year's total assets................. | 18,423,186 |

The corporation would be originally stratified according to the value of $249,850 total assets from the BMF file, but after evaluation of the added information from Element 4, the return would be stratified according to the value of $24,985,000 total assets.

- Example 2.--Digits are Added to the Stratifying Variable

| Element | Assume a corporation return has the following items available: | |
|---|---|---|
| 1 | Total assets after SOI editing ........... | $330,019 |
| 2 | Total assets as read from BMF file.......... | 3,300,195 |
| 3 | Initial stored entry of Element 2........... | 330,019 |
| 4 | Prior year's total assets................. | not available |

The corporation would be originally

stratified according to the value of $3,300,195 total assets from the BMF file, but after evaluation of added information from Elements 1 and 3, the return would be stratified according to the value of $330,019 total assets. In this case, the digits being added to the stratifying variable are probably caused by dollars and cents being recorded as dollars.

- Example 3.--One Variable Amount is Substituted for Another

| Element | Assume a corporation return has the following items available: | |
|---|---|---|
| 5 | Net income after SOI editing.......... | $65,008 |
| 6 | Net income as read from BMF file........ | 45,500 |
| 7 | Initial stored entry of Element 5......... | 45,500 |
| 8 | Prior year's net income........... | not available |

The corporation would be stratified according to the value of $45,500 net income from the BMF file rather than by total assets, since none are available. After evaluation of the added information, the return would be stratified according to the value of $65,008 net income.

## CORRECTION OF MIS-STRATIFICATION

The sample strata counts are easily corrected during the process of strata verification. Sample counts are changed by one each time a corporation changes strata. For example, a corporation originally assigned to stratum 2 is correctly assigned to stratum 5 after verification. Then the sample stratum 2 count is adjusted by subtracting one (-1) from its count and the sample stratum 5 count is adjusted by adding one (+1) to its count.

The population strata counts have not been changed at all up to this stage. Three different methods for handling the population counts are described below. The first method is to leave the population counts as they are, not making any correction for mis-stratification (Method 1). The second method is to adjust the population counts exactly as the sample counts are corrected for mis-stratification (Method 2). That is, if 40 companies moved out of a particular stratum in the sample, then 40 would be subtracted from the population count for that stratum. The last method involves using a weight to adjust the population counts (Method 3). The three methods are discussed in more detail below.

Method 1 has the advantage of requiring no action. It may be of value when the populations are very large relative to the sample size and any change would affect only the third, fourth or fifth decimal place in the sampling weight N(h)/n(h). Sampling weights are normally only taken to two decimals in the SOI corporation program.

Two major adverse consequences can occur in this situation. First, the sample estimate can be biased if the population counts are inaccurate for the stratum and, second, the usual estimate of the sample standard deviation can underestimate the true standard deviation.

The second method corrects for known cases of mis-stratification in the population, namely those seen in the sample. This method may still have the same consequences as Method 1, since nothing has been done about any mis-stratifications that go undetected in the population counts.

Method 3 is an attempt to alleviate the problems associated with Methods 1 and 2. This method, however, requires some type of assumption about the way mis-stratification occurs in the sample. For the purposes of this paper, mis-stratification will be considered to occur randomly throughout the population. That is, each return being processed has an equal chance of being mis-stratified during processing. Since the sample can be considered to be a stratified random sample, then the proportion of mis-stratifications seen in the sample strata should be the same as in the population. The sampling rates differ for the strata, however, and any correction for population counts must take this into account.

The following example illustrates Method 3. Assume two corporations were mis-stratified. The first one was found in Stratum 1, which has a sampling rate of .0035, and the other was found in Stratum 5, which has a sampling rate of .032, as shown in Table 1. Also, assume the company incorrectly sampled in Stratum 1 should have been in Stratum 5, and the company incorrectly sampled in Stratum 5 should have been in Stratum 1. Then there is no adjustment to the sample counts, since one is being added and subtracted from each stratum. But under Method 3, the population counts do change. For the company moving from Stratum 1 to 5 the population count in Stratum 1 is adjusted by (-1 x 1/.0035) = -286, and the population count in Stratum 5 is adjusted by +286. Now for the company moving from Stratum 5 to 1, the population count in Stratum 5 is adjusted by (-1 x 1/.032) = -31, and the count in Stratum 1 is adjusted by +31. The net adjustment in Stratum 1 is -255, while the net adjustment in Stratum 5 is +255.

Since the sampling rate is smaller in Stratum 1 than in Stratum 5, a mis-stratification seen in the smaller strata has more "weight" than a mis-stratification seen in a larger one. Table 3 gives the original and final (adjusted) sample counts using Method 3 for the 10 strata. Table 4 gives the original and final (adjusted) population counts for the 10 strata for the 1985 SOI corporation file.

Table 3.--1985 SAMPLE COUNTS AND MIS-STRATIFICATION ADJUSTMENTS BY STRATUM

| Stratum | Orignial count | Mis-stratifications moved | | Total changes | Final count | % Inc/ Dec. |
|---|---|---|---|---|---|---|
| | | In | Out | | | |
| 1 | 4955 | 190 | 28 | 162 | 5117 | 3.3 |
| 2 | 2809 | 242 | 21 | 221 | 3030 | 7.9 |
| 3 | 5370 | 533 | 35 | 498 | 5868 | 9.2 |
| 4 | 6866 | 719 | 58 | 661 | 7527 | 9.6 |
| 5 | 7570 | 533 | 56 | 477 | 8047 | 6.3 |
| 6 | 12,409 | 399 | 137 | 262 | 12,671 | 2.1 |
| 7 | 6200 | 202 | 120 | 82 | 6282 | 1.3 |
| 8 | 6639 | 122 | 318 | -196 | 6443 | -3.0 |
| 9 | 8534 | 144 | 570 | -426 | 8108 | -5.0 |
| 10 | 22,197 | 415 | 2106 | -1691 | 20,506 | -7.6 |
| Total | 83,549 | 3499 | 3449 | 50* | 83,599 | |

*Changes from tax forms besides Forms 1120, 1120A, and 1120S.

Table 4.--1985 POPULATION COUNTS WITH ADJUSTMENTS FOR MIS-STRATIFICATION BY STRATUM

| Stratum | Original count | Weighted adjustment | Estimated final count | Percent Increase/ Decrease |
|---|---|---|---|---|
| 1 | 1,432,110 | -4055 | 1,428,055 | -0.3 |
| 2 | 543,315 | +1081 | 544,396 | 0.2 |
| 3 | 645,033 | +2718 | 647,751 | 0.4 |
| 4 | 380,427 | +1536 | 381,963 | 0.4 |
| 5 | 236,238 | + 665 | 236,903 | 0.3 |
| 6 | 164,897 | + 734 | 165,631 | 0.4 |
| 7 | 60,244 | + 141 | 60,385 | 0.2 |
| 8 | 32,527 | - 966 | 31,561 | -3.0 |
| 9 | 23,196 | - 974 | 22,222 | -4.2 |
| 10 | 22,126 | - 493 | 21,633 | -2.2 |
| Total | 3,540,113 | + 387* | 3,540,500 | 0.01 |

*Filing tax forms other than Forms 1120, 1120A, or 1120S.

## SUMMARY

This paper illustrated a system to correct input mis-stratification in sampled administrative records, using corporation tax returns as an example. Typically, in sampling, the data are changed but the stratification is not. Under the proposed plan, both the data and the stratification are changed.

A small amount of mis-stratification is tolerable, but when the actual extent of mis-stratification is not negligible, then problems can occur in the estimation procedure. Oftentimes the population strata counts are known only at the time of sampling, whether correct or incorrect. Under this plan, the underlying population and sample counts are changed, which yield new weights for the strata. The method provides a way to correct for mis-stratification in the population without having to rework the entire population (a costly process).

In the 1985 corporation sample, there were 2,106 mis-stratifications in Stratum 10. Most of these corporations would not have been in the sample originally if they had been correctly stratified, but they are reclassified into their proper strata and kept in the sample. This does not hurt the estimates to have extra units in the sample and all large errors are found, because Stratum 10 is sampled at 100%.

The problem occurs when companies are mis-classified into the lower strata. We assume that these companies are indicative of others in the population which the sample does not reach, because of the low sampling rates of these strata. The companies are moved into their correct strata and the population counts are adjusted for the ones not seen in the sample. This procedure has the benefit of eliminating what would have been known previously as outliers in these lower strata; so, outliers no longer have to be handled separately. The weights now reflect these larger corporations that were missed in the sample due to mis-classification.

Future work is planned to study the effects of the three methods on the estimation of population total. Simulation methods will be used in the evaluation.

## EXPLANATION OF TABLES

Several interesting points about the sample can be seen by looking at the tables. The sampling rates used in the 1985 SOI corporation sample are given in Table 1. The rates range from a low of .0035 to a high of 1.00. Stratum 10 represents the "giant" corporations--those with total assets of $50 million or more--and is sampled with certainty. Stratum 1 has an extremely small sampling rate since the largest number of companies fall into this stratum. Also, the variability of this class is relatively small.

Table 2 is a tabulation of the counts of misclassification in the sample by cause and stratum. There were 3,449 corporations mis-stratified in the 1985 sample. Of the 3,449, nearly 75%, or 2572 of the mis-stratifications were caused by added digits in the total assets variable field. In comparison to the numbers for this cause, the others are quite small. These tables point out a problem which might be addressed by using better quality control methods in the data gathering and entry processes.

Original and corrected sample strata counts are given in Table 3. It can be seen that the upper three strata (8-10) decreased in size, while all the others increased in size. These results are to be expected after seeing that the majority of mis-stratifications were caused by added digits in the total assets variable field. The correction process moves these companies back into the smaller strata, where they should have been originally. Since these companies have already been processed for the file, they are kept in the sample, even if they many not have originally been selected based on the true amounts of the stratifying variables.

Original and estimated population strata counts are given in Table 4. Notice that the largest weighted adjustment occurs in Stratum 1. Because the weight for Stratum 1 is 1/.0035 = 285.7 (Table 1), only a small number of companies moving out causes a large negative adjustment. Likewise, notice that even though the most mis-stratifications occurred in Stratum 10, only a small negative weighted adjustment is made to that stratum. Corporations moving into and out of this stratum carry a weight of 1.00 because of the nature of the processing. SOI accounts for all of these corporations in the file.

## FOOTNOTE

[1] SOI corporations processed after 1986 were stratified according to the maximum of end-of-year and beginning-of-year total assets and cash flow rather than end-of-year assets and absolute value of net income/deficit. Cash flow is defined as deprectiation + depletion + net income.

## REFERENCES

COCHRAN, W.G. (1977). Sampling Techniques, (3rd ed.) New York: John Wiley, 85.

HARTE, J.M. (1986). Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS, Proceedings of the Section on Survey Research Methods, American Statistical Association, 603-608.

INTERNAL REVENUE SERVICE (1987). Description of the Sample and Limitations of the Data, Statistics of Income - 1984, Corporation Income Tax Returns, Publ. 16, Washington, D.C., 7-14.

INTERNAL REVENUE SERVICE (1989). Description of the Sample and Limitations of the Data, Statistics of Income - 1986, Corporation Income Tax Returns, Publ. 16, Washington, D.C., 9-18.

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