EVALUATING SAMPLE DESIGN MODIFICATIONS: BALANCING MULTIPLE OBJECTIVES

Susan Hinkins and Fritz Scheuren, Internal Revenue Service

Since 1951, the U.S. Internal Revenue Service has been sampling corporation tax returns to produce annual estimates of economic and tax variables. (Prior to 1951, the published information was based on all filed returns, rather than a sample.) Every year the population being estimated is the population of corporate returns filed for that tax year, and the primary objective of the sample design is to make accurate annual estimates. Associated with each tax return is a corporate entity, and many of these corporations continue in existence for many years. Therefore, there is also an interest in measuring changes over time; i.e., year-to-year and longitudinal estimates.

There has been an almost continual decline in sampling rates over the years, due to the combination of a constant growth in the number of organizations filing corporate tax returns and a practical limit on the number of returns which we can process. These declining sampling rates not only adversely affect the annual estimates, but also make it more difficult to keep corporations in the sample over a period of years, in turn, hampering accurate measures of change.

This paper focuses on design modifications to improve estimates of year-to-year change and to enhance the longitudinal composition of the sample, without jeopardizing the cross-sectional estimates of interest. Some such design features are already in place, and other options are being considered for the future.

BACKGROUND

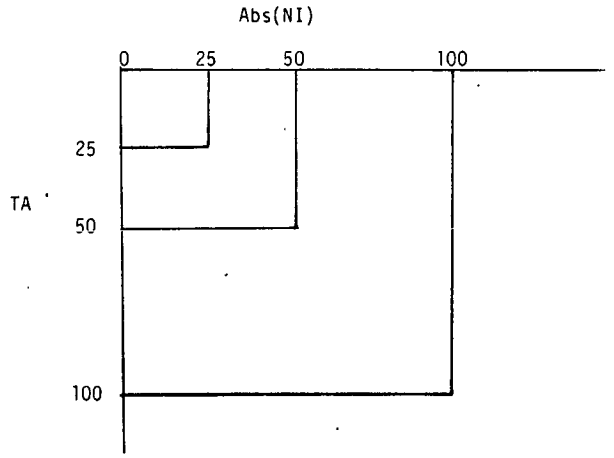
The population of corporate returns is highly skewed, with a relatively few large corporations accounting for well over half of the assets and income. In 1984, for example, the smallest 56% of all corporations accounted for only 0.5% of the U.S. total assets, while the top 0.11% of the corporations accounted for 75% of the U.S. total assets. The sample design is, therefore, stratified by "size," and all very large corporations are selected with certainty.

Size has been defined in terms of two items: Total Assets (TA) and Net Income or Deficit (NI). Ten size strata are defined using a "corner" shape. Figure 1 shows the definitions for the first three size strata; note that the absolute value of NI is used (Westat, Inc., 1974).

In addition to size, the sample is also stratified to select returns with certain items or characteristics of special interest, such as specific Principal Business Activities. These "special interest" strata represent only about 1% of the population and they are selected at relatively high sampling rates. So, in order to simplify our study, we consider only that part of the population sampled on the basis of size alone.

Before the sample is drawn, only estimates of the population sizes are available. Therefore,

Figure 1.--Size Stratification, in $1000



the sample is a stratified Bernoulli sample, where the sampling rates are determined using Neyman allocation. The within-strata population standard deviation is assumed to be proportional to the range of values in each stratum, assuming uniform distribution within strata. By the end of the sampling process, the population counts are known and can, then, be used for post-stratified estimation, conditioned on the achieved sample rates. Further descriptions of the design and estimation methods can be found in Harte (1982), Jones and McMahon (1984), Oh and Scheuren (1987), Jones (1988), and Mulrow (1989).

Over the years, the growth in the number of returns in the population and the need to sample all "large" returns have resulted in dramatic reductions in the sampling rates for the smaller size classes. Also, the amount of information (the number of items) being retrieved from each return has been increasing over the years. This expansion in complexity raises the cost of collecting and cleaning the data and adds to the problem of delivering estimates in a timely manner. The increase in cost, in terms of both time and money, is especially high for the large returns, which are so important to the total estimates. With a fixed budget, the net effect of this can be to further reduce the sample sizes in the strata for the smaller returns.

ESTIMATING POPULATION DYNAMICS

The current sample design already addresses the problem of estimating year-to-year change by including some features to assure a high sample overlap from year to year. The present study is to evaluate the longitudinal aspects of the design and consider alternatives for improvement in this area. Specifically, we are interested in year-to-year dynamics: the population movement from year to year and the sample characteristics from year to year.

To look at these trends, we want to move the

Figure 2.-- Matrix of Change Probabilities by Size Strata

			1985 Strata									
		Deaths	1	2	3	4	5	6	7	8	9	10
	Births		.3000	.1800	.1400	.1200	.0900	.0800	.0600	.0500	.0500	.0400
	1	.2300	.6607	.0848	.0173	.0050	.0009	.0006	.0004	.0001	.0001	.0001
	2	.1200	.1554	.5456	.1593	.0172	.0009	.0008	.0005	.0001	.0001	.0001
	3	.0900	.0171	.1039	.6661	.1124	.0084	.0011	.0005	.0002	.0002	.0001
1984	4	.0800	.0110	.0132	.1193	.6401	.1236	.0108	.0014	.0003	.0002	.0002
Strata	5	.0600	.0042	.0046	.0124	.1093	.6824	.1205	.0052	.0006	.0005	.0003
	6	.0500	.0010	.0015	.0026	.0108	.0967	.7407	.0868	.0078	.0016	.0006
	7	.0400	.0006	.0011	.0023	.0091	.0154	.1160	.6699	.1315	.0119	.0021
	8	.0300	.0002	.0009	.0020	.0038	.0121	.0173	.1135	.6761	.1350	.0091
	9	.0300	.0001	.0001	.0009	.0020	.0043	.0066	.0128	.0769	.7523	.1140
	10	.0200	.0001	.0001	.0003	.0009	.0013	.0031	.0066	.0106	.0367	.9203

population and sample from one year to the next. This simulation can be carried out using an estimate or model of the matrix of change probabilities. For example, if we are interested in the sampling strata, we use a matrix of change probabilities as shown in Figure 2. The rows show how the corporations in a given stratum in 1984 move or change strata in 1985. The first column indicates deaths, i.e., corporations no longer in business in 1985, or corporations that have been absorbed through mergers. This is determined by the presence or absence of the corporation's unique identifier--the Employer Identification Number (EIN)--on the file for the sample year. (When an EIN is no longer present on the file it is called a death. A new EIN is called a birth.) The first row indicates births, or new corporations in 1985, as a percentage of the 1984 stratum total. Usually mergers continue to use an existing identifier, but some births are the result of mergers of older, established entities. The center of the table shows the strata changes for corporations in the population in both years. Cells off the diagonal represent corporations that changed strata between 1984 and 1985. Take, for example, those corporations in stratum 2 in 1984: we estimate that 12% will not be in the population in 1985, 15.5% will have dropped to stratum 1, and 54.6% will remain in stratum 2 in 1985, etc.

The matrix was contructed in two parts. The interior of the table shows the movement between strata for corporations existing in both years. These probabilities were estimated from sample data, smoothing the tails where the probabilities are small and the sample is sparse. The primary data base is the sample of 1985 corporations, restricted to corporations filing on Forms 1120, 1120A or 1120S, with no "special" properties. This category of corporations accounts for approximately 99% of the corporations in the population and approximately 89% of the sample. From that population, "final year" or part-year returns were excluded, as they present a special problem and will have to be dealt with separately. This excludes only another 5% of the population.

If the corporation was also selected in the 1984 sample, then the 1984 TA and NI and the 1984 sampling rate are also available. From these records, we estimate the population sizes in change cells (or, equivalently, the change probabilities), by weighting up the sample counts by the inverse of the probability of being in both samples. This gives estimates ofthe population dynamics for the interior of the table in Figure 2. Later, we will use the 1986 sample file to verify this model.

The sample provides little usable information about births and it provides no information about deaths. A corporation may exist in both years but not be in both years' samples. Therefore, birth and death rates were modelled based on prior knowledge. They were also initially set so that the number of corporations grows by 6% from year to year, the current average.

For deaths, only estimates of the proportion of deaths by 1984 sample strata were needed. For births, we needed estimates of the number of new corporations by 1985 strata, as well as the distribution of births by variables of interest, TA and NI, for example. To estimate these, we do have some additional information from the 1985 sample file, which contains records that were not in the 1984 sample. These records could occur in several ways:

- new corporations that did not exist in 1984;
- corporations that existed in 1984, but were not sampled because, in that year, they were sampled with a smaller sampling rate than in 1985; or
- corporations that existed in 1984 and should have been in the 1984 sample, but were not sampled because of errors in the data at the time of selection.

We cannot distinguish between these groups with complete accuracy. However, if we ignore the latter group, we can take as a subset of births those records in the 1985 sample but not in the 1984, that would have been selected into the 1984 sample at the very smallest 1984 sampling rate. We use these records to estimate the distribution of births within strata.

The birth and death rates for large corporations may look a little high, but recall that "death" and "birth," here, do not include only

corporations going out of business or corporations just starting up. As already noted, some of these "birth" and "deaths" of EIN's are generated by mergers or splits. Corporations may also change EIN's for other reasons, as well.

SAMPLE OVERLAP

If drawn independently, the effective sampling rate for selecting a corporation into the samplein both years would be the product of the two years' sampling rates. Therefore if left to chance, for corporations existing in both years there would be very little overlap in the sample from year to year, except for large static corporations, which were selected at a 100% rate in both years.

The corporate sample design addresses the objective of estimating change by assuring a much larger sample overlap from year to year. In general terms, random sampling is done using a pseudo-random number generator (uniform distribution), and a return is selected if the generated random number is less than the designated sampling rate. Overlap in the sample is achieved by using the corporation's unique EIN as the seed to the generator in both years. Therefore, if the corporation is selected on one occasion, it will be selected again if the selection rate is at least as high. (This type of procedure is discussed in Harte, 1986. See, also, Sunter, 1986.) Using the EIN, the effective selection rate for a corporation being in both samples is the minimum of the two years' sampling rates.

This procedure significantly increases overlap and retains the representative sampling of new corporations. However, it still results in most of the sample overlap being on or above the diagonal. No emphasis is placed on corporations that change and corporations that shrink in size are penalized, to some extent.

Cells above the diagonal represent corporations that grew from 1984 to 1985. The sample overlap, here, is small because the sampling rates in 1984 were smaller. To increase overlap we would need to predict, in 1984, which corporations would grow in the future. It is doubtful whether we will ever be able to do this effectively.

Cells below the diagonal represent corporations that got "smaller" in 1985, so the 1985 selection rate is smaller than that for 1984. Therefore, many of these corporations would be in the 1984 sample but not in the 1985 sample. We could improve the overlap, here, by looking back to 1984 results before sampling in 1985.

In the last several years, stratifying variables have been added to the design to increase the number of corporations in both samples by "looking back" in this way. A recent design change was to replace the stratifying variable Total Assets by the maximum of Total Assets and Beginning Assets. This attempts to "look back" to 1984, because the 1985 Beginning Assets should equal the 1984 Total Assets. Also the stratifying variable abs(NI) was replaced by the maximum of the abs(NI) and the absolute value of cash flow, where cash flow = NI + Depreciation + Depletion. Both of these modifications should increase the sample overlap below the diagonal. However, because budget considerations demand that the cost of sampling remain essentially the same, if we want to increase the sample size below the diagonal, we have to reduce it somewhere else.

We are looking at different options for doing this by investigating how designs work over several years. By simulating the year-to-year dynamics using a matrix of estimated change probabilities, we can model how the population changes over several years, and we can follow a sample design over several years. This was done with the following assumptions:

- the change matrix is the same each year; and
- changes are independent from year to year.

ORIGINAL CROSS-SECTIONAL DESIGN

We first looked at the corner design in Figure 1 based on TA and abs(NI) alone. We assumed the sample size was fixed at 85,000 each year, strata definitions did not change, and the largest stratum was sampled at 100%. The matrix of change probabilities in Figure 2 was used to project the population and the sample from year to year. Neyman allocation was used each year to determine sampling rates.

Using the change matrix, we projected the 1984 sample of 85,000 returns out for five years. In this way, we estimate how many of the corporations in the 1984 sample will still be available for sampling after one year or after five years. In particular, we followed two subsets of the 1984 sample:

- the 20,927 corporations in the top stratum, sampled with certainty in 1984; and
- the 64,073 corporations that were sampled with sampling rates less than 1.0.

The estimated movement after one year is summarized in the second column of Figure 3. Of the 85,000 corporations in the 1984 sample, 4,536 (4,117 + 419) are no longer in the population. Of the 64,073 corporations selected in 1984 with sampling rates less than one (size strata 1-9), 59,956 are available in 1985; 44,488 of these are in the same stratum in both years; etc.

Figure 3.--Year-to-Year Sample Overlap by Change in Strata: 1984-1985

Status in 1985	Original 1984 Sample	Design Based on Size	Adding Change Class
	Sample Selected at Rates Less Than 100% in 1984		
Deaths	4,117	----	----
Available	59,956	50,921	40,536
No Strata Change	44,488	39,982	27,694
Moved Up	7,903	7,903	7,568
Moved Down	7,565	3,036	5,274
	Sample Selected with Certainty in 1984		
Deaths	419	----	----
Available	20,508	19,852	20,462
No Strata Change	19,259	19,259	19,259
Moved Down	1,249	593	1,203

To evaluate the sample design based on TA and NI, we estimated how many of the corporations from the 1984 sample were still retained in the sample after one year and after five years, using this design. The third column in Figure 3 shows the estimated sample overlap after one year, using this design, compared to the sample available. As noted earlier, the design results in a large sample overlap. For corporations selected in 1984 with sampling rates less than one, 50,921 would be selected again in 1985, out of a possible 59,956 (85%). But corporations that move to a lower stratum (decrease in size) are under-represented; less than half of those available would be selected again in 1985.

Figure 4 shows the results after five years. These are estimates of the number of corporations for which we would have data for all six years, 1984-1989. Rather than try to track the strata movement up and down, we only report two categories: corporations that never change strata and corporations that moved to another stratum at least once in five years. Over half the corporations available in all six years are selected using this design. However, to estimate change, we might be especially interested in corporations that changed through the years-- i.e., corporations that changed strata. Figure 4 indicates, again, that the sample design based on TA and NI does not reflect this objective very well, especially for corporations that moved out of the 100% class.

Figure 4.--Longitudinal Sample Overlap: 1984-1989

Status in 1989	Original 1984 Sample	Design Based on Size
	Sample Selected at Rates Less Than 100% in 1984	
Deaths	17,303	---
Available	46,770	24,086
No Strata Change	10,850	6,156
Changed	35,920	17,930
	Sample Selected with Certainty in 1984	
Deaths	2,168	---
Available	18,759	15,161
No Strata Change	13,815	13,815
Moved Down	4,944	1,346

FIRST ATTEMPT AT DESIGN MODIFICATION

We first considered an "optimal" design, assuming optimal--but not necessarily realistic--conditions. Since the primary objective is still accurate cross-sectional estimation, modifications for improving estimates of change are made within the cross-sectional designs. For 1985, the cross-sectional design based on TA and NI consists of 10 size strata. Sampling rates are set using Neyman allocation, with the sample size set at 85,000.

To stratify on year-to-year change, the following change classes were defined. For a corporation existing in both years, the absolute difference between TA in 1984 and TA in 1985 was one classification variable; the absolute difference between 1984 NI and 1985 NI was the second. Using these two variables, ten change classes were defined, based on the same class limits as the corner design (Figure 1), replacing TA by the absolute change in TA between 1984 and 1985 and replacing abs(NI) by the absolute change in NI. New corporations in 1985 make up a separate change class. In this way, each corporation in the 1985 population can be cross-classified by these two stratifying techniques: 1) into a size class, using 1985 TA and NI; and 2) into a change class, defined by the change in TA and NI from 1984 to 1985. Take, for example, a corporation with the following values:

Variable	1984	1985	Absolute Change
TA	90,000	68,000	22,000
NI	30,000	60,000	30,000

Referring to Figure 1, in 1985 this corporation would be in size stratum 3 (TA and NI) and change class 2 (absolute change in TA and NI).

Using this two-way cross-classification, each 1985 size stratum is subdivided into the 11 possible change classes (10 change classes and a class for births), as shown in Figure 5. The sampling rate at the bottom of each column is the 1985 rate for the usual cross-sectional design, based on 1985 size stratum only. This represents a predicted sample size for each stratum. The 1984-1985 change classes are added within the structure of the cross-sectional design by fixing each cross-sectional (column) sample size, and allocating this sample among the change classes within that column. Again, Neyman allocation is used. The only additional constraint is that in each column the sampling rate for births is kept equal to the original cross-sectional rate, to keep a representative sample of births.

If we could select the 1985 sample in this manner and retrieve the 1984 data for all sampled returns, this should be an optimal design, constrained by the cross-sectional design, for estimating year-to-year change. Unfortunately, we cannot necessarily go back in 1985 and pick up 1984 information for corporations not selected in the 1984 sample. We are, therefore, interested again in the estimated sample overlap using this design, and that is shown in the last column of Figure 3. For estimating year-to-year change, this is probably a preferable design--certainly better than the design based on size alone--because it captures significantly more of the corporations that changed strata. However, for corporations selected with rates less than one in 1984, the overall number of corporations in the two-year overlap is much smaller (40,536 vs 50,921), and, hence, the longitudinal composition of the sample overlap--say over 5 years--would be much reduced using this design. Therefore, this variation on the size design will not work.

Figure 5.--Sample Design Adding Change Classes

		1985 Strata						
		1	2	3	....	8	9	10
Change Classes 1984-1985	Births	.003	.005	.008	....	.235	.672	1.00
	1	.003	.004	.006	....	.010	.024	1.00
	2	.005	.006	.009		.015	.036	1.00
	3	.008	.011	.015	....	.025	.060	1.00
	4	.018	.023	.033		.055	.132	1.00
	5	.032	.043	.060	....	.101	.240	1.00
	6	.080	.107	.150		.252	.600	1.00
	7	.112	.149	.210	....	.353	.839	1.00
	8	.223	.298	.420		.706	1.000	1.00
	9	.638	.852	1.000	....	1.000	1.000	1.00
	10	1.000	1.000	1.000		1.000	1.000	1.00
1985 Cross-Sectional Sampling Rates		.003	.005	.008	....	.235	.672	1.00

Greater overlap might be achieved by using "change in strata" rather than actual dollar changes as a measure of year-to-year change. This is being considered. (See below.)

FUTURE PLANS

This is another in a series of still preliminary reports. We have updated earlier description of the model and estimation procedures used in this evolving project. As noted, the primary issue is improving the longitudinal sample composition with minimum effect on cross-sectional estimates. Designs still to be considered currently include:

- other designs adding a specific "change" variable to the original size design;
- the "new" design, using the maximum of TA and Beginning Assets (instead of TA), and the maximum of abs(NI) and abs(cash flow) (instead of the abs(NI)); and
- designs that include indexing the strata definitions.

Comparisons will be in terms of estimated variances for cross-sectional estimates and for estimates of year-to-year change. The study will also look at each design in terms of the longitudinal sample composition over a five-year period.

The discussant correctly points out the difficulty of controlling the percentage of overlap with these schemes. On a trial basis, we are implementing one of his suggestions; we are including an imbedded panel in the sample design. One possible disadvantage of including a panel is that we may have to reduce the sampling rates for the cross-sectional estimates, due to budgetary constraints. We will be evaluating the costs and benefits of this design.

ACKNOWLEDGMENTS

The authors would like to thank Jeri Mulrow for many helpful discussions and Richard Collins for providing the necessary data base. We also wish to thank Wendy Alvey and Beth Kilss for their editorial assistance throughout, and Bettye Jamerson for help in typing the manuscript.

REFERENCES

HARTE, J. M. (1982). Post-Stratification Approaches in the Corporation Statistics of Income Program, Proceedings of the American Statistical Association, Section on Survey Research Methods, 250-253.

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 American Statistical Association, Section on Survey Research Methods, 603-608.

JONES, H. (1988), A Description of the SOI Corporate Sample, Working Paper, Statistics of Income Division, Internal Revenue Service.

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. (1989). Description of the Sample and Limitations of the Data, Statistics of Income ...1986, Corporation Income Tax Returns, Publ. 16, 9-15.

OH, H. L., and SCHEUREN, F. J. (1987). Modified Raking Ratio Estimation, Survey Methodology Journal, vol. 13, no. 2, Statistics Canada.

SUNTER, A. B. (1986). Implicit Longitudinal Sampling from Administrative Files: A Useful Technique Journal of Official Statistics, 2, 2, Statistics Sweden: Stockholm.

WESTAT, INC. (1974). Results of a Study to Improve Sampling Efficiency of Statistics of Corporation Income, Working Paper, Bethesda, MD.