DESIGN MODIFICATIONS FOR THE SOI CORPORATE SAMPLE: BALANCING MULTIPLE OBJECTIVES

Susan Hinkins, Homer Jones, and Fritz Scheuren, Internal Revenue Service

Since 1951, the U.S. Internal Revenue Service (IRS) has been sampling corporation tax returns to produce annual estimates of economic and tax data. (Prior to 1951, the published information was based on all filed returns, rather than a sample.) Through the years, the process for collecting information and making estimates has evolved and changed, as a result of shifts in the population of corporations (economic trends) and revisions in the tax law. Also changes in user needs and advancements in computer and statistical technology have brought about modifications to the sample design.

Unfortunately, one of the most dramatic changes is that sampling rates have gone down almost continuously over the years, due to the combination of a constant growth in the number of organizations filing corporation tax returns, and a practical limit on the number of returns which we can process. The limitation on sample size is primarily due to budgetary restrictions, and the short time period within which we are permitted to process the sampled returns. 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 and, therefore, hamper accurate measures of change from year to year.

This paper focuses on design modifications to improve estimates of year-to-year change and to enhance the longitudinal composition of the sample, without compromising the cross-sectional estimates. Some such design features are already in place, and other options are being considered for the future. Of course, as we consider such modifications, we must also look at possible design effects on cross-sectional estimates. This evaluation procedure is described and preliminary results are given. Some thoughts on future directions are raised in the concluding section.

BACKGROUND

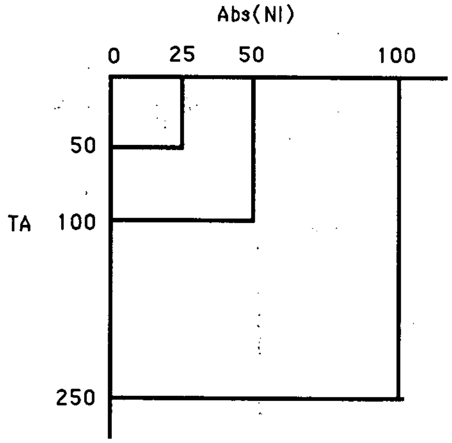
The population of corporation returns is highly skewed, with a relatively few large corporations accounting for well over half of the assets and income. In 1984, for example, the smaller corporations accounted for 56% of all corporations but only 0.5% of the U.S. total assets, while the top 0.11% of the corporations accounted for 75% of the U.S. total assets.[1] The sample design is, therefore, stratified by size, and the very large corporations are selected with certainty.

Size is defined in terms of two items: Total Assets (TA) and Net Income or Deficit (NI).[2] The former provides a measure of the level of total assets and other balance sheet items on the tax return, and the latter is used to measure the size of income statement items which make up total income and total deductions.

In addition to size, the sample is also stratified to select returns with certain Principal Business Activities or with certain items of special interest. However, these other strata criteria apply to a very small fraction of the population and they are selected at relatively high sampling rates.

Therefore, to simplify our study, we consider only that part of the population sampled on the basis of size alone -- in particular, corporations filing on Forms 1120 or 1120S, with TA under $25 million and NI under $5 million. Larger corporations are selected with certainty in all sample designs discussed here. The strata defined by size have a corner shape; Figure 1 shows the definitions for the first 3 strata. Note that the absolute value of NI is used. Table 1 in the appendix gives the definition of all size strata.

Figure 1.-- Size Stratification in $1000

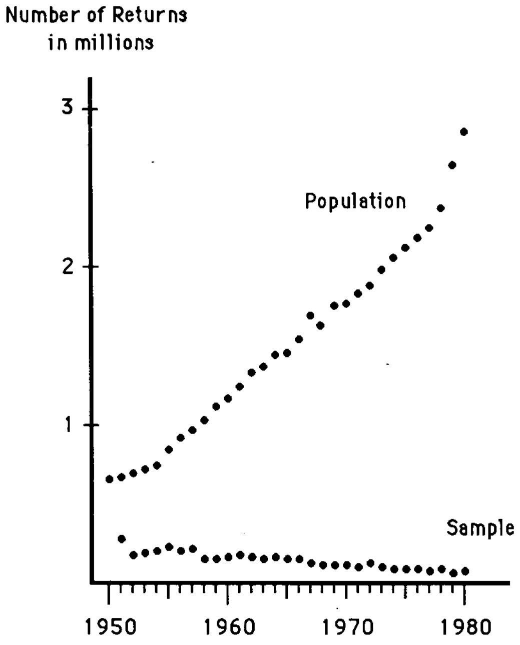


Before the sample is drawn, we only have estimates of the population sizes. Therefore the sampling rates are determined using Neyman allocation and the sample is a stratified Bernoulli sample. The within-strata population standard deviation is assumed to be proportional to the range of values in each 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.

A fairly thorough, though highly condensed, documentation of the present sample design is provided in <u>Statistics of Income -- 1984 Corporation Income Tax Returns</u>. Further descriptions of the design and estimation methods can be found in Jones (1988), Harte (1982), Jones and McMahon (1984), and Oh and Scheuren (1987).

Over the years, the population of corporate returns has grown in several ways, causing a relatively steady decline in the sampling rates. First, the number of returns being filed has increased. The sample size, however, is restricted by budget and time constraints, with results as shown in Figure 2. An essentially fixed sample size, an increasing number of returns in the population, and the need to sample all "large" returns, result in dramatic reductions in the sample rates for the smaller size classes.

Figure 2.-- Corporate Returns, 1950-1980



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. This 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 YEAR-TO-YEAR CHANGE

The features of the design described, so far, are concerned with the primary objective: making accurate annual estimates; however, with samples being taken every year, a reasonable secondary objective is to estimate change over the years. Some provision for this has been made in the sample design, and some innovations to improve these estimates are being considered. The variable or indicator we use for looking at year-to-year change is the change in "size", i.e., the strata movement from year to year. We will denote this change in size by DELTA.

To add DELTA as a proper stratifying variable, we need to assign standard deviations to change classes. The variance for the change from stratum to stratum was defined as the squared Euclidean distance between the midpoints of the strata, plus the square of an average range of change within stratum. (See Hinkins, Jones, and Scheuren, 1988.)

Figure 3 shows the dynamics of the year-to-year change from 1984 to 1985. For this purpose, the sample rates for each year are rounded to one significant digit. The first row indicates births or new corporations in 1985. (Usually mergers continue to use an existing identifier, but some births are the result of mergers of older, established entities.) The first column indicates deaths -- i.e., corporations no longer in business in 1985 or corporations that have been absorbed through mergers. The center of the table shows the change classes for corporations in the population in both years. The cells on the diagonal represent corporations that were in the same stratum from one year to the next. Cells off the diagonal represent corporations that changed strata between 1984 and 1985.

If we were designing the 1985 sample to estimate DELTA, we would want (1) a representative sample of births and (2) as much sample overlap as possible for corporations existing in both 1984 and 1985. Ideally, to measure the overlap population, we would like to select the same corporations in both years' samples. If the entire 1984 sample of continuing corporations could not be used, we would probably, at least, want to emphasize sampling off the diagonal (i.e., selecting corporations that have large changes).

The cross-sectional design results in representative sampling of births; however, 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 taken at a 100% rate. If drawn independently, the effective sampling rate for selecting a corporation into the sample in both years is the product of the two years' sampling rates. Take, as an example, the cell representing corporations in stratum 1 in 1984 and in stratum 3 in 1985. The effective rate for selecting such corporations in both years' samples would be .000012 = (.002)*(.006).

The corporate sample design addresses the objective of estimating change by assuring a much larger 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. Each corporation has a unique employer identification number (EIN). Overlap in the sample is achieved by using the 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 sample rate is at least as high. (This type of procedure is discussed in Harte, 1986.) Using the EIN, the effective selection rate for a corporation being in both samples is the minimum of the two years' sampling rates. Continuing the previous example, the effective sampling rate using the EIN would be the minimum of (.002, .006) = .002, compared to the minuscule sampling rate of .000012, if left to chance.

Figure 3.-- Cross-Tabulation for 1984 and 1985 Strata

			1985 Strata										1984 Sampling Rates
		Deaths	1	2	3	4	5	6	7	8	9	10	
	Births												n.a.
	1		····										.002
	2			····									.002
	3				····								.007
	4					····							.01
1984 Strata	5						····						.02
	6							····					.07
	7								····				.1
	8									····			.2
	9										····		.7
	10											·····	1.0
	1985 Rates	n.a.	.002	.002	.006	.01	.02	.06	.1	.2	.6	1.0	n.a.

Note: n.a. = not applicable.

Sunter (1986) calls such a procedure implicit longitudinal sampling and shows that it maximizes the overlap of units sampled on two or more occasions; but when most changes are small it still results in most of the sample overlap being on the diagonal. No emphasis is placed on corporations with great change.

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 the 1984. Therefore, many of these corporations would be in the 1984 sample but not in the 1985 sample. We can 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. For example, a recent design change was to use the maximum of Total Assets and Beginning Assets as the stratifying variable, instead of just Total Assets. This "looks back" to 1984 -- because the 1985 Beginning Assets should equal the 1984 Total Assets -- and would, therefore, 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.

AN EXAMPLE OF A DESIGN MODIFICATION

In this section, we describe one type of design modification that adds DELTA as a stratifying variable. In looking at various options (both practical and theoretical) for changing the design, we expect improvements in the estimates of year-to-year change. However, another necessary concern is the design effect on the cross-sectional (annual) estimates. What would be the increase in variance for the annual estimates of TA and NI?

To look at year-to-year change, we need a matrix of change probabilities, where the ijth element is the probability of a corporation moving from stratum i to stratum j. Unfortunately, we do not have any year-to-year tabulations for the population. Therefore these change probabilities were estimated from the sample data (Hinkins, Jones, and Scheuren, 1988). For all sample designs considered, the sample size was fixed at 85,000.

Looking again at Figure 3, columns 1-10 represent the 1985 strata for the cross-sectional design based on size. (Column 10 represents the large corporations, with TA over $25 million and/or the absolute value of NI over $5 million; these returns are selected with certainty.) In this example, we add DELTA as a stratifying variable within each size stratum. That is, we fix the sample size for each size stratum, each column, as determined by the Neyman allocation, and then allocate that sample between the change classes within the column. This gives sample sizes for each change cell. Rather than show the entire 10 x 11 table, designs are compared using the sample sizes combined into 4 categories:

- births (new corporations),
- above the diagonal (corporations increasing in size),
- below the diagonal (corporations decreasing in size), and
- on the diagonal (corporations with little change).

Using Neyman allocation within size strata in the way described would allocate the sample as shown in row 1 of Figure 4. This also assumes that we can pick the sample looking at both years' population, which is not currently feasible.

If we move the 1984 sample to 1985, using the matrix of change probabilities, the expected sample sizes available for sampling in 1985 are shown in row 2. So, while Neyman allocation would select 26,504 corporations that grew from 1984 to 1985, we only expect to have 9,155 such corporations that were sampled in 1984 and are available for sampling in 1985. The resulting sample overlap is shown in line 3, Figure 4. Comparing these sample sizes to the expected sample overlap under the current design (line 4), we see that, for estimating change, the current design over-represents the diagonal classes, at the expense of the off-diagonal cells and at the expense of the new corporations.

The design adding DELTA as a stratifier would significantly improve the estimate of year-to-year change and the longitudinal composition of the sample. However, when we looked at the design effect on the estimate of TA, we found a significant increase in the variance of the cross-sectional estimate of TA.

Therefore, the Neyman allocation was modified to increase the sample size on the diagonal, as follows. Column 1 (1985 stratum 1) has an unexpected distribution; a surprising number of units (2,691) apparently fell from stratum 10 (the largest corporations) in 1984 to stratum 1 (the smallest corporations) in 1985. Since this change cell has a very large assumed variance, Neyman allocation selects all these units in the sample. Because of the relatively small sample size for stratum 1 (3,920), this left few units available for sampling the other large cells: births and the diagonal cell. We modified the sample in this column by cutting the sampling rate in the extreme cells and increasing the sample of births and of the diagonal cell. For all other 1985 strata (2-9), we reduce the sampling rate for 1) births, 2) the cell immediately below the diagonal, and 3) the cell immediately above the diagonal, to the sampling rate of the current design. We can then increase the sampling rate on the diagonal.

The expected sample overlap using this modified design for change is shown in line 5, Figure 4. Comparing the three designs (lines 3-5), we see that this modification is a compromise between the current design and the best design for estimating change. In terms of sample overlap, the modifications are all among the births, the on-diagonal, and the below-diagonal elements. We cannot improve on the above-diagonal overlap. Looking just at the below-diagonal overlap, the modified design improves the estimate of change not only by increasing the total sample overlap compared to the current design, but by sampling more units with large change. Figure 5 shows the expected sample overlap below the diagonal by the distance from the diagonal. The first row shows the total sample overlap adding across all change cells that are immediately below the diagonal. The last row gives the sample overlap

Figure 4.-- Year-to-Year Change Sample Allocation

Sample Design	Above Diag.	Below Diag.	On Diag.	New Corps (Births)
1) Neyman Allocation for change	26,504	15,923	32,294	10,279
2) 1984 Sample in 1985 (Overlap)	9,155	13,176	57,631	NA
Sample Overlap				
3) Neyman Allocation	9,155	13,176	32,294	10,279
4) Current Design	9,155	4,068	55,186	6,384
5) Modified Change Design	9,155	7,347	48,236	6,044
Total Sample				
6) Current Design	19,362	4,068	55,186	6,384
7) Modified Change Design	21,846	8,874	48,236	6,049

Figure 5.-- Sample Overlap Below the Diagonal

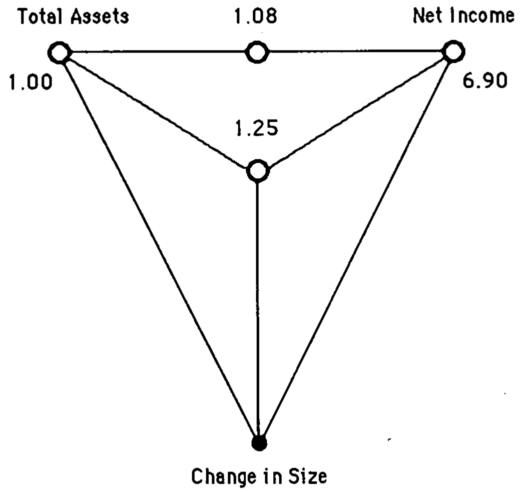
# of Cells Below Diag.	Neyman Alloc.	Current Design	Modified Design
1	7,118	3,618	3,460
2	1,059	289	1,034
3	592	90	550
4	543	41	543
5	369	13	369
6	445	8	445
7	169	1	169
8	190	1	104
9	2,691	7	673

for the most extreme change cell: corporations falling from stratum 10 in 1984 to stratum 1 in 1985. We can see here that compared to the current design, the modified design for change increases the sample in the more extreme change cells.

For many items, estimating change may not depend on having the same units in both years' samples. In Figure 4, rows 1, 6 and 7 show the 1985 total sample sizes. The pattern is the same; for estimating change, the current design over-emphasizes the diagonal cells and under-represents the off-diagonal.

Finally, the design effects are compared. We are considering three stratifying variables: 1) TA, 2) NI, and 3) DELTA (year-to-year change). The designs under consideration are: 1) the current design based on TA and NI, and 2) the current design with DELTA added. The design including change is the modification of the Neyman allocation. Figure 6 shows the design effects (deffs) for estimating TA. The optimal design for estimating TA is the design stratifying on TA alone, so it has deff equal to 1.00. All other designs are compared to this; variances are relative to the optimal variance of $\hat{TA}$. The current design has an estimated deff of 1.08, or an 8% increase over the optimal. Adding change as a stratifier, via the modified design, increases the deff to 1.25, which is a 16% increase over the current design.

Figure 6.-- Design Effects for $\hat{TA}$



Because the design including DELTA as a stratifier is built on the current design, only these 2 designs are compared for estimating DELTA. As described earlier, the estimated population movement showed a relatively large number of corporations falling from the largest size stratum to the smallest. Most of these are likely to be "final returns," i.e., corporations going out of business. We would be overstating the improvement in estimating change if we included this cell in the estimate of variance. Therefore, we computed the deff removing that cell, and found there is still a 35% decrease in the estimated variance of DELTA compared to the current design.

Hence, this modification to the current design improves the estimate of change, decreasing the variance of $\hat{DELTA}$ by 35%, with a 16% increase in the variance of TA.

AREAS OF FUTURE STUDY

This has been just one example of the design modifications currently being considered. There are other similar modifications to the current design to be investigated. For example, we could reallocate the sample between change cells within the current strata with the restriction that the variance of $\hat{TA}$ in that strata is not increased.

We are also considering another type of modification: adjusting the strata cut points for inflation. Currently, the strata boundaries are essentially fixed through the years. This is done, partially, because our annual estimates are published every year for the population classes defined by these strata. However, inflation would cause movement across these strata boundaries that is not indicative of a real change in a corporation. Therefore, another option being considered is to adjust the strata cut points so that strata more nearly represent the same part of the population from year to year and movement out of strata is more indicative of real change. Such a modification would also improve the annual estimates; what might suffer would be the published estimates of the specific subclasses.

Figure 7 shows the estimated population counts when the 1985 classes are adjusted for 6% inflation. Adjusting for inflation has not shifted more records onto the diagonal, but it has made the table more symmetric, above and below the diagonal:

Figure 7.--Strata Changes: 1984 to 1985

Diagonal	Fixed Strata Classes	Adjusted for Inflation
Above	16.2%	13.5%
On	72.7%	72.8%
Below	11.1%	13.7%

Having relatively fewer corporations moving above the diagonal should result in more sample overlap. The effects of adjusting the strata cut points will be investigated in a similar manner.

Finally, we have compared various designs using only one criterion: the design effects for estimating year-to-year change and for the cross-sectional estimates. Another comparison that needs to be included is longitudinal sample composition. Assuming a fixed sample size (85,000), a fixed growth rate (birth and death rates), and the same matrix of change probabilities, how many corporations will still be in the sample after 3 years? After 5 years? For example, taking the current design out 3 years, we estimate that the 3 year sample overlap will be 55,558, of which only 19,250 will be in off-diagonal cells, i.e. corporations that changed strata at least once in 3 years. We plan to compare the longitudinal sample composition for different sample designs.

This has been a very general description of current design modifications being considered for improving the corporate sample design. While we are still in the analysis stage and there are difficulties of application that we have not discussed here, we are optimistic that improvements can be made to the estimation of change without jeopardizing the cross-sectional estimates.

ACKNOWLEDGMENTS

The authors would like to thank Dorothy Farmer for her help in typing this manuscript and Wendy Alvey and Beth Kilss for their editorial assistance throughout this effort.

FOOTNOTES

[1] Internal Revenue Service (1987). Statistics of Income -- 1984, Corporation Income Tax Returns, Publ. 16, Washington, D.C., 7-14.

[2] For SOI Corporations processed after 1986, size is no longer defined in terms of end-of-year total assets and net income/deficit. The former is now replaced by the maximum of end-of-year and beginning-of-year total assets, and the latter is replaced by cash flow. Cash flow is the total of depreciation + depletion + net income (signed), and is applicable only where the absolute value is greater than that of the absolute value of net income/deficit.

REFERENCES

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

HARTE, J. M. (1982). Post-Stratification Approaches in the Corporation Statistics of Income Program, Proceedings of the Section on Survey Research Methods, American Statistical Association, 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 Section on Survey Research Methods, American Statistical Association, 603-608.

HINKINS, S., JONES, H., and SCHEUREN, F. (1988). Design Modifications for Estimating Change, Working Paper, Statistics of Income, IRS.

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.

JONES, H. (1988). A Description of the SOI Corporate Sample, Working Paper, Statistics of Income, IRS.

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.

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 Files: A Useful Technique, Journal of Official Statistics, vol. 2, no. 2, Statistics Sweden, 161-168. See especially p. 164.

Appendix Table 1. -- Strata Definitions

Strata	TA (Total Assets)	NI (Net Income)
1	0 under $50,000	0 under $25,000
2	$50,000 under $100,000	$25,000 under $50,000
3	$100,000 under $250,000	$50,000 under $100,000
4	$250,000 under $500,000	$100,000 under $250,000
5	$500,000 under $1,000,000	$250,000 under $500,000
6	$1,000,000 under $2,500,000	$500,000 under $1,000,000
7	$2,500,000 under $5,000,000	$1,000,000 under $1,500,000
8	$5,000,000 under $10,000,000	$1,500,000 under $2,500,000
9	$10,000,000 under $25,000,000	$2,500,000 under $5,000,000