SAMPLING CORPORATION INCOME TAX RETURNS FOR STATISTICS OF INCOME, 1951 TO PRESENT

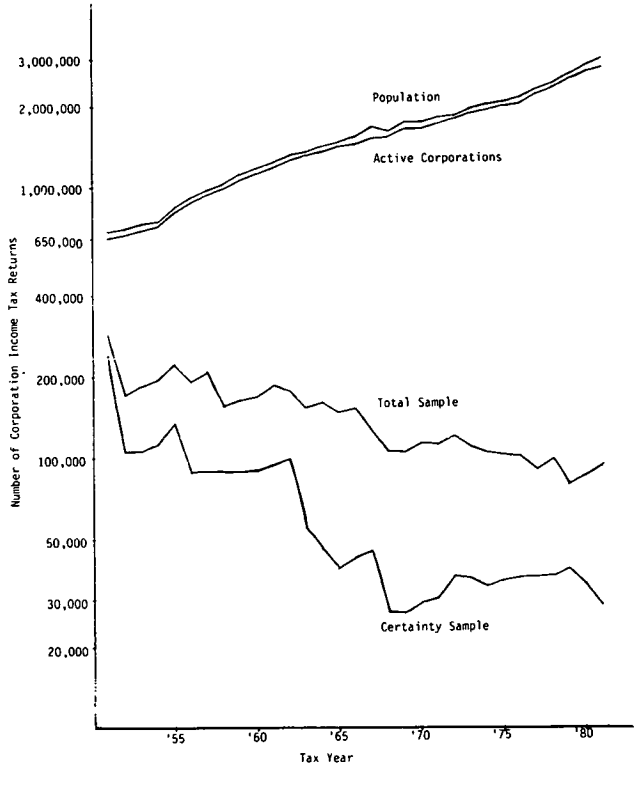
Homer W. Jones and Paul B. McMahon, Internal Revenue Service

The Internal Revenue Service has published economic data based on Corporation Income Tax Returns for each year since 1913. The aggregate estimates are published in the Statistics of Income, Corporation Income Tax Returns [1] volumes, with more exhaustive estimates presented in the Source Book [2].

Sampling techniques had been employed by the Service for three decades [3] before their introduction to the main Statistics of Income (SOI) Corporation Studies (although several special projects, such as the 1949 Federal Trade Commission Study [4] had used sampling). In part this delay was due to the diverse natures of the various corporation subpopulations. Mainly, however, the delay was due to a reluctance on the part of the major users (Congress, the Department of the Treasury and the Bureau of Economic Analysis) to base important decisions on sample data. By 1951, though, the cost of surveying the entire corporation population outweighed that reluctance.

The first section of this paper describes the changes in the sample design since 1951. The following sections present an evaluation of the accuracy of the resultant estimates and the methods used to compute those estimates. The final section briefly outlines possible directions of future changes.

Figure A : Number of Corporations in the Population and Sample 1951 - 1981



EVOLUTION OF THE SAMPLE DESIGN

Over the four decades, 1951-1984, that sampling has been used for SOI Corporation Statistics, two major forces have affected the evolution of the sample designs; a shrinking proportion of the population in the sample, and changes in the administrative processes of the Internal Revenue Service.

The decreasing proportion of the population used in the sample is illustrated in Figure A. The sample for Tax Year 1951, the initial sample, contained 41.5 percent of the population, 285,000 returns of the 687,000 filed. For 1981, this proportion had decreased to 3.1 percent, or 93,000 returns from a population of over 3,000,000. Most of the decrease in the sample proportion has been due to the growth of the population. This rate of growth has increased significantly since 1976, which will undoubtedly lead to still further sample design changes [5].

In fact, there has been only one decrease in the size of the population subject to sampling. That came in 1968 - the year that the Internal Revenue Service converted to computer processing. This decrease was artificial, a consequence of the delays common to such conversions. Indeed, the estimated population of active corporations (the population of interest to the users of the data) showed a modest increase that year.

Another factor in the decreasing sample proportion is a decline in the size of the sample. The smaller sample sizes were the result of budget reductions and the growing costs for increasingly complex data abstraction procedures [6].

The decrease in the sample size has generally been accomplished by reducing the size of the certainty classes (the strata where all returns are included in the sample). The means of reducing the size of the certainty classes was the raising of the threshold for those classes. For 1951, that threshold, which was based on size of Total Assets, was $250,000. The 240,000 returns with more than that amount were included in the sample. The next year that threshold was raised to $500,000, reducing the certainty sample to 107,000 returns (and incidentally creating a new stratum for returns with Total Assets between $250,000 and $500,000). This pattern of raising the threshold continues, although additional constraints have been added such as the corporation's industry. (The highest Total Assets threshold in use today is $50,000,000 for Financial Returns.)

The other major factor which influenced the sample design of Corporation Returns was the administrative environment. One of the interesting aspects of the initial design (for

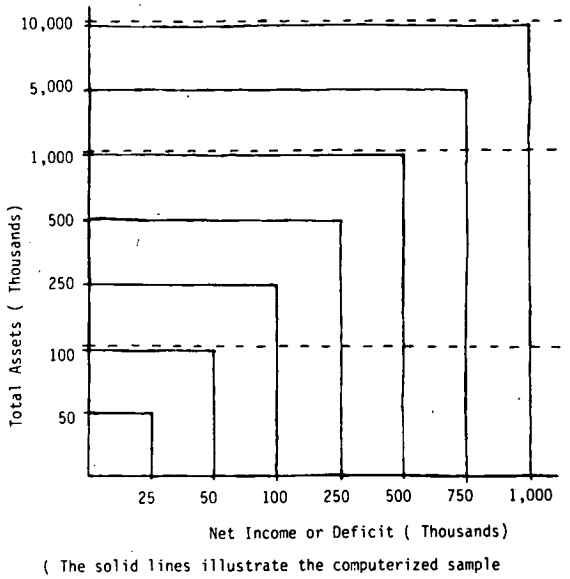
Tax Year 1951) is the relative absence of restrictions imposed by the administrative processes. This absence was due to the centralization of the sampling operations in Washington, D.C. This centralization allowed the sample design to employ hundreds of strata, although only two sampling rates were used. The stratifiers included State, Industry, Taxability, Total Assets and other criteria. The most important criterion in determining the sampling rate to which a return would be subjected was size of Total Assets.

The centralized sampling procedure, however, entailed a high cost in shipping hundreds of thousands of returns. Thus, in a cost reduction, the sampling process was decentralized to the various District Offices (where the returns were originally filed). One consequence of this move was that the sample design became dependent on the existent administrative classification system. This classification system sorted returns into groups designed to aid in audit and collection activities. Fortunately, the audit categories were not very different from those used for sampling in 1951. Thus, the sample design required only minimal changes for Tax Year 1952.

For 1953 Corporation Returns, the administrative sort changed. The most important revision was that Total Assets was replaced with an income stratifier (Gross Sales, Gross Receipts or Total Income). The Corporation Sample Design, now firmly tied to the adminstrative sorting, had to follow suit.

Thus, the major stratifier between 1953 and 1958 became the largest of Gross Sales, Gross Receipts or Total Income. For Tax Year 1959, this income classification system was itself replaced, and Total Assets again served as the basis for the sort. As before, the Corporation Returns sample design changed accordingly.

Figure B : Comparison of Manual and Computerized Sample Designs, 1968



(The solid lines illustrate the computerized sample design, while the dashes show the manual design.)

In 1968, the Internal Revenue Service began computerizing its manual administrative processes. In this process, the filing and storage of the various types of the tax returns was centralized at regional "service centers" (there are now ten of these centers). One important advantage of automation was that the sample selection process could be computerized.

Not all of the Tax Year 1968's Corporation Returns were subjected to sampling by the new computer system. Some were manually stratified and sampled. This leads to a natural comparison of the two designs employed.

Both designs set aside strata for rare subpopulations of particular interest to the users (this practice continues today), so we will not consider these, but only the basic designs [7]. The manually selected sample was stratified into classes based on the size of Total Assets. These classes are shown by the dotted lines in Figure B. The larger the size of Total Assets a return showed, the higher the sampling class and sampling rate. A consequence of this design was that returns with large amounts of income but small Total Assets were selected infrequently. This had the effect of lessening the accuracy for the resultant estimates of income items.

The design on which the computerized sample selection program was based addressed this problem through the creation of a nested box design. The new design also employed Total Assets as a major stratifier, but size of Net Income or Deficit was added. The solid lines in Figure B illustrate this design. In addition to the inclusion of Net Income or Deficit, the computerized sampling also allowed the number of strata to be greatly increased [8].

Another design change made possible through the use of computers is less obvious than those above. Under the manual sampling procedures, selection depended upon a sequence number assigned to each return as they were received. Selection by computer, however, used the ending digits of Employer Identification Numbers, which is permanently and uniquely assigned to each corporation. Because a corporation uses the same Employer Identification Number each year, use of this identifier to select the sample over several years will tend to include the same corporations' returns over those years. The advantage here lies in the reduction of variance for estimates of year to year change. Clearly, such a panel structure could not be obtained with the manual sequential selection process.

This longitudinal aspect of the sample was enhanced for 1978 with the introduction of the Employer Identification Number Transform [9]. This transform generates a random number from the Employer Identification Number through the use of very large prime numbers and modular arithmetic. The use of this transform has allowed the longitudinal aspect of the sample to be extended readily across strata boundaries. Further, the transform allows longitudinal

sampling at very small sampling rates, a significant advantage in light of the ever decreasing proportion of the population included in the sample.

While the selection procedures based on the Employer Identification Number and the transform both created a panel structure, they also provided for the selection of returns from new corporations. This happens because only the ending digits of the Employer Identification Number or the transform's random number are used to determine selection (or non-selection).

For 1981, the administrative environment again caused a change in the sampling process. A shortage of computer resources caused the Internal Revenue Service to once again shift the site of sample selection. The selection point is now the National Computer Center, where the master file list of all corporation Employer Identification Numbers is maintained. This shift of site has many advantages such as: tighter control of the sampling process, yet with more flexibility; validation of each Employer Identification Number before sample selection; and access to additional population data.

The validation of the identifier has marginally improved the longitudinal selection. More importantly, though, is the longitudinal data capture that can be made. Currently, only a few prior year items are carried forward at the moment (as part of a research effort), but this aspect of the sample is being expanded.

The access to additional population data has already had a great effect on the estimation procedures, leading to use of raking ratio estimation [10,11].

ACCURACY OF THE ESTIMATES

The data published for Tax Years before 1951 were not subjected to sampling error, as all returns in the population was included in the tabulations. With over half a million returns, however, it was impossible to ensure that the manually abstracted and tallied data from each return were correct [12]. In fact, most "error resolution" or testing involved tabulated data. The tables were "balanced" so that the various row or column details added to the totals and subtotals.

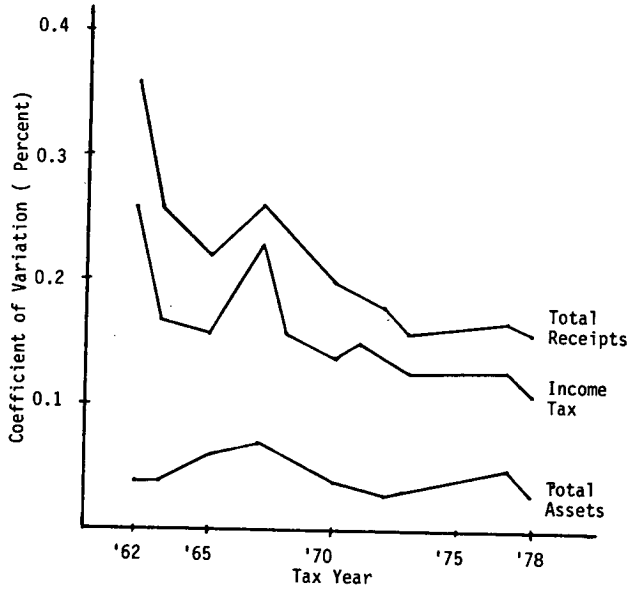
Over the years there has been an increasing number of controls introduced to reduce the nonsampling errors. One of the early controls involved computer generating the tables to avoid manual arithmetic errors. More recently, corporation returns selected for the sample and used in the tabulations have all been subjected to numerous intricate consistency tests by computer. Many of these tests are based on tax law restrictions; others on the structure of the tax return which has many interrelated data items; and still other tests are statistical in nature - identifying outliers for verification and possible correction.

Sampling the returns reduces the cost of collecting the data, and also makes the review process more feasible, but at the cost of introducing imprecision to the estimates. This imprecision can be quantified. The measure of the imprecision for an estimate is the coefficient of variation for that estimate.

Coefficients of variation have been published for various estimates for most years since 1962. We would expect, from all the sample size reductions, that the trend would show decreasing accuracy. After all, the sample has fallen from 170,000 returns for 1962, to 93,000 for 1981.

A comparison of the coefficients of variation for the national estimates of Total Assets, Total Receipts and Income Tax over time, as shown in Figure C, shows a different tendency. The coefficients for Total Assets show that the accuracy of these estimates has remained essentially stable. Perhaps this should be expected, for Total Assets was the major stratifier for the entire period. Total Receipts and Income Tax, however, actually show improvements.

Figure C : Comparison of Coefficients of Variation for Selected Estimates



These improvements are the result of increased stratification and the inclusion of Net Income as a stratifier. For 1962 Corporation Returns, only three strata were used. These strata were defined by size of Total Assets with breaks at $100,000 and $1,000,000. Compare this to the 1981 sample design with twelve sampling classes. Although, most of the additional strata resulted from the redefining of the certainty strata, five strata now divide the population with Total Assets under $1,000,000 whereas only two were used for the 1962 Tax Year design.

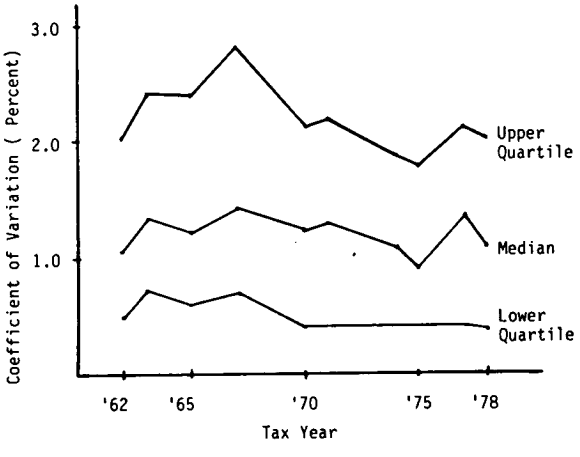
The above comparisons used national, all industries estimates. Many of the users of corporation statistics, however, are interested in industry groups, asset categories or other

subpopulations. This demand for increasing detail is a direct outgrowth of the increasing availability of computers and software packages for analyzing the basic data.

We cannot explore the accuracy of each subpopulation estimate, for there are simply too many. Indeed, a detailed analysis of the fifty-eight major industry classifications would require much more space than is allotted here. We have, though, summarized the major industry coefficients of variation for the Total Assets estimates by reviewing the median, upper and lower quartiles for those coefficients of variation. This is shown in Figure D.

It appears from these data that the estimates of Total Assets for the various industries have maintained essentially the same level of accuracy since 1962, again despite decreasing sample size. Since we are examining the accuracy of the Total Assets estimates we might expect that certain industries might come to dominate that lower quartile, and as the sample is heavily stratified on that variable, to be predominant in the sample as well.

Figure D : Distribution of Coefficients of Variation for Total Assets Across Major Industries

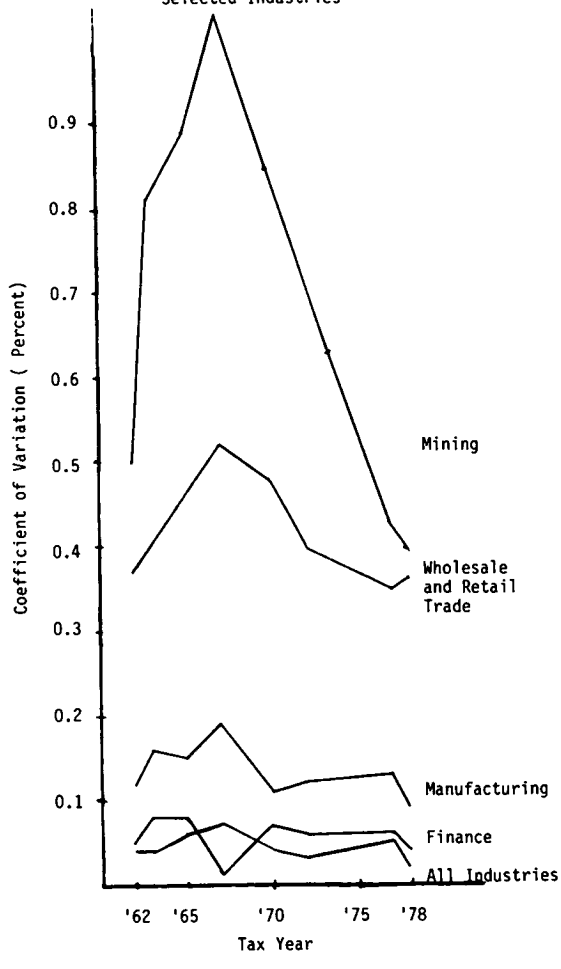


An examination of the twelve industrial divisions present some candidates. Four of these are shown in Figure E below. As might be expected, the coefficients for divisions "Finance" and "Manufacturing" are quite close to the coefficients for the all industries (national) total. Also, these industries, due to their asset intensive natures, are disproportionately represented in the sample. This is not to suggest that disproportionate representation of subpopulations is undesirable. Many designs, including this one, incorporate this feature as a means of improving critical estimates [13].

ESTIMATION PROCEDURES

The procedure employed for the 1951 Corporation data used weights calculated by dividing the number of returns counted in a sampling class by the number actually selected

Figure E : Coefficients of Variation for Total Assets for Selected Industries



for the sample from that class. This straightforward system was used until 1973, when 'integer weighting' was introduced.

The integer weighting procedure started with a weight computed from the population and sample counts, as above, to two decimal places. Integer weights of values w and w+1 were then assigned to the sample records so that the average weight equaled the preliminary (two digit) weight. For example, if a weight of 4.45 were computed for a given sampling class, then 45 percent of the returns in that class would have received an integer weight of 5. The balance of the returns in that class (55 percent) would receive a weight of 4. Clearly this system does not improve the accuracy of the estimates [14, 15]. This was not its function; its purpose was to reduce the cost of manually reviewing each table to ensure that the totals and subtotals balanced. With newer computer systems and software packages this problem vanished, and the integer weighting process was discarded for the 1980 statistics.

As noted above, there is a growing demand for accurate subpopulation estimates from the Corporation sample. The most obvious subpopulations are those of various industries. A pilot study of raking ratio estimation using Tax

Year 1979 data showed that, with some modifications, this procedure was feasible and did lead to an overall improvement in the subpopulation estimates [11]. The procedure, which was used in computing the Tax Year 1980 estimates, used a 580 cell matrix (58 major industries by 10 weighting classes) for each of four time periods (for the sampling rates varied over the two year sampling program). The first step in the procedure was to calculate as a weight the inverse of the probability of selection adjusted for missing returns for each cell of the matrix. Those cells containing more than 199 sample returns were then excluded from further weight calculations, for they contained a sufficient sample for accurate parameter estimation.

The remaining cells were then weighted and iteratively scaled to the row and column constants until the resultant weighted totals changed less than a specified tolerance.

Two other constraints were placed on this process of weight computation. This first was that no weight would be less than 1.00. The second constrained the "raked" weight to fall within a narrow range. The upper limit of the range was 125 percent of the weight computed for the sampling class (the "national weight") while the lower boundary was 80 percent of the national weight.

A comparison of estimates obtained using the "raked weights" and the "national weights" shows that, while many differences exist in the subpopulation estimates, the marginal totals (across industries) are very close. The standard errors for the raking estimates, however, have been reduced considerably [11].

FUTURE DIRECTIONS

We noted above that the growth of the Corporation population presages future sample design changes. One direction this change may take is the continued raising of the threshold of the certainty class. Other tactics are also under investigation, such as increased use of industry stratification. The main thrust, however, is to maintain the accuracy of the resultant economic estimates, without increasing the cost of the data collection. This is also being accomplished through increased use of external data and through imputation [16].

Aside from maintaining the accuracy of the current estimates, we are seeking methods for improved longitudinal estimates. Some of these methods involve computer editing of prior years' data onto current year records. We are also seeking a method which will routinely provide reliable variance estimates for the particular version of raking ratio estimation now being used. So far these efforts have met with only limited success.

ACKNOWLEDGMENTS

The authors would like to thank Robert E. Fay III for his comments, Dan Rosa and Keith Gilmour for their review of this paper, and Dawn Alexander who typed the document. We also acknowledge the contributions of the predecessors of the current Statistics of Income Division upon whose work the paper is based.

NOTES AND REFERENCES

[1] Statistics of Income - Corporation Income Tax Returns, (Various Years), Publication Number 16, Department of the Treasury, Internal Revenue Service, U.S. Government Printing Office, Washington, D.C.

[2] Source Book, Statistics of Income, Corporation Income Tax Returns, (Various Years), Publication Number 1053, Department of the Treasury, Internal Revenue Service, U.S. Government Printing Office, Washington, D.C.

[3] Paris, David and Gilmour, Keith, "70th Year of Statistics of Income," 1984 American Statistical Association Proceedings, Section on Survey Research Methods.

[4] Rosander, A.C., Blythe, R.H. and Johnson, D.W., "Sampling 1949 Corporation Income Tax Returns," Journal of the American Statistical Association, June 1951, pages 233-241. It should be noted that this study also served as the pilot for the initial, 1951, Statistics of Income Corporation sample.

[5] Clickner, R.P., Galfond, G.J. and Thibodeau, L.A., "Redesign of the IRS Corporate SOI Sample," 1984 American Statistical Association Proceedings, Section on Survey Research Methods.

[6] For Tax Year 1951 Corporations, estimates were published for about 70 variables. By comparison, over 400 data items were edited for 1981 Corporations.

[7] An example of a special subpopulation is the group of returns claiming a U.S. Possessions Tax Credit.

[8] The strata definitions for returns with less than $1,000,000 in Total Assets and Net Income (or Deficit) less that $500,000 are unchanged since 1968. The other boundaries, however, have been revised, although the pattern shown in Figure B has been retained.

[9] Westat, Inc., Results of a Study to Improve Sampling Efficiency of Statistics of Corporation Income, Bethesda, Maryland, January 1974 (Unpublished).

[10] Harte, James M., "Post-Stratification Approaches in the Corporation Program," 1982 American Statistical Association Proceedings, Section on Survey Research Methods, pgs. 250-253.

[11] Leszcz, Michael R., Oh, H. Lock and Scheuren, Fritz J., "Modified Raking Estimation in the Corporate SOI Program," 1983 American Statistical Association Proceedings, Section on Survey Research Methods, pgs. 434-438.

[12] During this period all data were tallied using punch card technology, hence only limited record - by - record checking was possible. In 1954 the statistical processing became computerized. [3]

[13] Due to the conditional nature of the coefficients of variation when post-stratification is used, and problems with the computational methodology, no coefficients of variation for years after 1978 were included in the discussion of the accuracy of the estimates. Some estimates are available however. See [11].

[14] Hansen, Morris H., Hurwitz. William N., and Madow, William G., Sample Survey Methods and Theory, John Wiley and Sons, Inc., New York, 1953, Volume 2 pgs. 139-141.

[15] Oh, H.L.. and Scheuren, F.J., "Weighting Adjustments for Unit Nonresponse," Incomplete Data: The Theory of Current Practice, National Academy of Sciences, Panel on Incomplete Data, 1984.

[16] Hinkins, Susan, "Matrix Sampling and the Effects of Using Hot Deck Imputation," 1984 American Statistical Association Proceedings, Section on Survey Research Methods.