SOI Sampling Methodology and Data Limitations

This article discusses typical sampling procedures used in most Statistics of Income (SOI) programs. Aspects covered briefly include sampling criteria, selection techniques, methods of estimation, and sampling variability. Some of the nonsampling error limitations of the data are also described, as well as the tabular conventions employed.

Additional information on sample design and data limitations for specific SOI studies can be found in the separate SOI reports. More technical information is available, on request, by writing to the Director, Statistics of Income Division RAS:S, Internal Revenue Service, P.O. Box 2608, Washington, DC 20013-2608.

Sample Criteria and Selection of Returns

Statistics compiled for the SOI studies are generally based on stratified probability samples of income tax returns or other forms filed with the Internal Revenue Service (IRS). The statistics do not reflect any changes made by the taxpayer through an amended return or by the IRS as a result of an audit. As returns are filed and processed for tax purposes, they are assigned to sampling classes (strata) based on such criteria as: industry, presence or absence of a tax form or schedule, and various income factors or other measures of economic size (such as total assets, total receipts, size of gift, and size of estate). The samples are selected from each stratum over the appropriate filing periods. Thus, sample selection can continue for a given study for several calendar years—3 for corporations because of the incidence of fiscal (noncalendar) year reporting and extensions of filing time. Because sampling must take place before the population size is known precisely, the rates of sample selection within each stratum are fixed. This means, in practice, that both the population and the sample size can differ from those planned. However, these factors do not compromise the validity of the estimates.

The probability of a return's selection depends on its sample class or stratum and may range from a fraction of 1 percent to 100 percent. Considerations in determining the selection probability for each stratum include the number of returns in the stratum, the diversity of returns in the stratum, and interest in the stratum as a separate subject of study. All this is subject to constraints based on the estimated processing costs or the target size of the total sample for the program.

For most SOI studies, returns are designated by computer from the IRS Master Files based on the taxpayer identification number (TIN), which is either the Social Security number (SSN) or the Employer Identification Number (EIN). A fixed and essentially random number is associated with each possible TIN. If that random number falls into a range of numbers specified for a return's sample stratum, then it is selected and processed for the study. Otherwise, it is counted (for estimation purposes), but not selected. In some cases, the TIN is used directly by matching specified digits of it against a predetermined list for the sample stratum. A match is required for designation.

Under either method of selection, the TINs designated from one year's sample are, for the most part, selected for the next year's, so that a very high proportion of the returns selected in the current year's sample are from taxpayers whose previous years' returns were included in earlier samples. This longitudinal character of the sample design improves the estimates of change from one year to the next.

Method of Estimation

As noted above, the probability with which a return is selected for inclusion in a sample depends on the sampling rate prescribed for the stratum in which it is classified. "Weights" are computed by dividing the count of returns filed for a given stratum by the number of population sample returns for that same stratum. These weights are usually adjusted for unavailable returns, outliers, or trimming weights. Weights are used to adjust for the various sampling rates used, relative to the population—the lower the rate, the larger the weight. For some studies, it is possible to improve the estimates by subdividing the original sampling classes into "poststrata," based on additional criteria or refinements of those used in the original stratification. Weights are then computed for these poststrata using additional population counts. The data on each sample return in a stratum are then multiplied by that weight. To produce the tabulated estimates, the

Sample returns are designated by computer from the IRS Master Files based on the taxpayer identification number.

In transcribing and tabulating data from tax returns, checks are imposed to improve the quality of the statistics.

weighted data are summed to produce the published statistical totals.

Sampling Variability

The particular sample used in a study is only one of a large number of possible random samples that could have been selected using the same sample design. Estimates derived from the different samples usually vary. The standard error of the estimate is a measure of the variation among the estimates from all possible samples and is used to measure the precision with which an estimate from a particular sample approximates the average result of the possible samples. The sample estimate and an estimate of its standard error permit the construction of interval estimates with prescribed confidence that this interval includes the actual population value.

In SOI reports, the standard error is not directly presented. Instead, the ratio of the standard error to the estimate itself is presented in percentage form. This ratio is called the coefficient of variation (CV). The user of SOI data may multiply an estimate by its CV to recreate the standard error and to construct confidence intervals.

For example, if a sample estimate of 150,000 returns is known to have a coefficient of variation of 2 percent, then the following arithmetic procedure would be followed to construct a 68-percent confidence interval estimate:

150,000	(sample estimate)
x 0.02	(coefficient of variation)
= 3,000	(standard error of estimate)

then:

150,000	(sample estimate)
+ or - 3,000	(standard error)
= {147,000, 153,000}	(68-percent confidence interval).

Based on these data, the interval estimate is from 147 to 153 thousand returns. This means that the average estimate of the number of returns lies within an interval computed in this way. Such an estimate would be correct for approximately two-thirds of all possible samples similarly selected. To obtain this interval estimate with 95-percent confidence, the standard error should be multiplied by 2 before adding to and subtracting from the sample estimate. (In this particular case, the resulting interval would be from 144 to 156 thousand returns.)

Further details concerning sample design, sample selection, estimation method, and sampling variability for a particular SOI study may be obtained, on request, by writing to the Director, Statistics of Income Division, at the address given above.

Nonsampling Error Controls and Limitations

Although the previous discussion focuses on sampling methods and the limitations of the data caused by sampling error, there are other sources of error that may be significant in evaluating the usefulness of SOI data. These include taxpayer reporting errors and inconsistencies, processing errors, and the effects of any early cutoff of sampling. Additional information on nonsampling error as it applies to individual and corporation income tax returns is presented in the separate SOI reports on these returns.

In transcribing and tabulating the information from returns or forms selected for the sample, steps are taken to improve the quality of the resultant estimates. Tax return data may be disaggregated or recombined during the statistical abstracting and "editing" process that takes place in IRS submission processing centers. This is done to improve data consistency from return to return and to achieve definitions of the data items more in keeping with the needs of major users. In some cases, not all of the data are available from the tax return as originally filed. Sometimes, the missing data can be obtained by the Statistics of Income Division in Washington, DC, through field followup. More often, though, they are obtained through manual or computerized imputation. For this purpose, other information in the return or in accompanying schedules may be sufficient to serve as the basis for making an estimate. Prior-year data for the same taxpayer can be used for this same purpose, or comparable data from business reference books may be substituted.

Data abstracted or "edited" from returns for statistical use are subjected to a number of validation checks, including systematic verifications of a sampling of the work of each tax examiner involved in the SOI process. Data reported on sampled returns

and previously transcribed as part of processing for the IRS Master Files are subject to validation as part of the administrative process before SOI processing begins. However, during the administrative process, it is only practical to transcribe corrections to errors that have a direct bearing on the tax reported or the refund claimed. Therefore, during the SOI process, checks must also be made to correct any errors or inconsistencies left in the administrative data before the data can be accepted for the statistics.

The Statistics of Income program includes many more tax return items than are transcribed and perfected for IRS tax administration needs, especially for items reported in tax return schedules in support of the various summary totals reported on the return. Therefore, checks must also be designed to validate these additional data items and to assure that they are consistent with other data entries.

Most of the data validation checks made during the SOI process take the form of computerized tests of each record. In addition to verifying that internal consistency and proper balance and relationships among the tax return items and statistical classifications are maintained, this process is intended to check on consistency with tax law provisions, acceptable reporting practices, and generally accepted accounting principles. Most testing occurs during the data abstracting and editing operation, while the tax return source document is still on hand, although some testing for certain programs occurs later on. Records failing the tests are subjected to further review and correction.

Finally, before publication, the statistics are reviewed for accuracy and reasonableness in light of the tax law provisions, taxpayer reporting variations and other limitations, tolerances and statistical techniques allowed or employed in data processing and estimating, economic conditions, and comparability with other statistical series. However, these controls do not completely eliminate the possibility of error. When discovered, errors in *Bulletin* tables are corrected, through a published errata.

Table Conventions

Published estimates subject to excessive sampling variability are identified for most of the statistics by means of an asterisk (*) presented alongside the estimate or in place of an estimate. Presence of an asterisk means that the sampling rate was less than 100 percent of the population and that there were fewer than 10 sample observations available for estimation purposes. This method produces a rough indication of excessive sampling variability. However, the results will differ somewhat from more precise indicators of excessive sampling variability based on the standard statistical formula. For some of the statistics based on samples, asterisking was not possible because of resource and other constraints. Users should keep this limitation in mind when using these data.

A zero, in place of a frequency or an amount, in any given table cell presenting data based on an SOI sample, indicates either that (1) there were no returns in the population with the particular characteristic, or (2) because of its rarity, instances of the characteristic were not present among the sampled returns. However, for statistics based on returns selected for the sample at the 100-percent rate, a zero indicates a presumption of no returns with the particular characteristic in the population.

In addition to sampling variability, Statistics of Income is required to prevent disclosure of information about specific taxpayers or businesses in its tables. Therefore, a weighted frequency (and the associated amount, where applicable) of less than 3 is either combined with data in an adjacent cell(s) so as to meet the criteria, or deleted altogether. Similar steps are taken to prevent indirect disclosure through subtraction. However, any combined or deleted data are included in the appropriate totals. Most data on tax-exempt, nonprofit organizations are excluded from disclosure review because the Internal Revenue Code and regulations permit public access to most of the information reported by these organizations.