Section 3

Description of the Sample and Limitations of the Data

This section describes the sample design, sample selection, data capture, data cleaning, and data completion processes for the Statistics of Income (SOI) 2015 Corporation Statistics Program. It also presents the techniques used to produce estimates of the total number of corporations and associated variables as well as an assessment of the data limitations, including sampling and nonsampling errors.

Background

From Tax Years (TY) 1916 through 1950, SOI extracted data from each corporate income tax return filed. Beginning with TY 1951, however, SOI introduced stratified probability sampling. Since that time, the sample size has generally decreased while the corporate tax return population has increased. For example, for 1951, the sample accounted for 41.5 percent of the entire population, or 285,000 of the 687,000 total returns filed. For 2015, the sample accounted for about 1.81 percent of the total population of just over 6.5 million returns. This population count differs from the estimated population count cited elsewhere in this publication because the sampling frame includes out-of-scope and duplicate returns.

For 1951, SOI stratified the sample by size of total assets and industry. However, from 1952 through 1967, SOI stratified the sample by a measure of size only. The size was measured by either business volume (1953–1958) or total assets (1952 and 1959–1967). Since 1968, SOI has stratified returns by both total assets and, for Forms 1120 and 1120S, a measure of income [1].

Target Population

The target population consists of all returns of active corporations organized for profit that are required to file one of the 1120 forms included in this study.

Survey Population

The survey population includes corporate tax returns filed using one of the 1120 forms selected for the study and posted to the IRS Business Master File (BMF). Excluded are amended returns and returns for which the tax liabilities changed because of a tax audit. Figure E gives the number of corporate returns by form type that were subject to sampling during Tax Years 2012 through 2015, as well as the resulting sample sizes.

Sample Design

The current design is a probability sample stratified by form type and either by 1) size of total assets alone or 2) size of total assets and a measure of income. Form 1120 returns are stratified by size of total assets and size of "proceeds," which is the measure of income for this form. Size of proceeds is defined as the larger of the absolute value of net income (or

Figure E. Total Number of Corporation Tax Returns: Population and Sample Counts, Tax Years 2012–2015

Form type	Tax year							
	2012		2013		2014		2015	
	Population	Sample	Population	Sample	Population	Sample	Population	Sample
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1120	1,800,426	59,303	1,785,481	59,054	1,769,209	58,567	1,759,931	55,929
1120S	4,409,276	36,256	4,484,612	36,741	4,577,096	37,998	4,682,942	37,514
1120-L	657	445	600	405	581	392	540	376
1120-PC	10,218	2,456	11,721	2,669	13,264	2,920	14,598	3,146
1120-RIC	15,612	10,331	16,379	10,813	17,267	11,275	17,951	11,412
1120-REIT	2,168	1,815	2,502	2,104	2,807	2,359	3,103	2,679
1120-F	38,065	5,926	40,923	6,319	43,693	6,685	45,745	7,078
Total	**6,276,422**	**116,532**	**6,342,218**	**118,105**	**6,423,917**	**120,196**	**6,524,810**	**118,134**

Bertrand Überall, Richard Collins, and Elliot Mountjoy were responsible for the sample design and estimation of the SOI 2015 Corporation Statistics Program under the direction of Tamara Rib, Chief, SOI Program Support, Statistical Services Branch.

deficit) or the absolute value of "cash flow," which is the sum of net income, several depreciation amounts, and depletion. Form 1120S is stratified by size of total assets and size of ordinary income. SOI stratified all other 1120 forms (1120-L, 1120-PC, 1120-RIC, 1120-REIT, and 1120-F) by size of total assets only.

SOI began the design process with projected population totals derived from IRS administrative workload estimates, adjusted according to the distribution by population strata from several previous survey years. Using projected population totals by sample strata, SOI carried out an optimal allocation based on strata standard errors to assign sample sizes to each stratum such that the overall targeted sample size was approximately 118,600 returns for 2015, a slight increase from the 2014 target. Mathematical statisticians selected a Bernoulli sample independently from each stratum, with sampling rates ranging from 0.25 percent to 100 percent. The total realized sample for 2015, including inactive and noneligible corporations, is 118,134 returns.

Sample Selection

The IRS Cincinnati and Ogden Submission Processing Centers initially process all corporate returns to determine tax liability before transmitting the data daily to the BMF. After error correction, these returns are said to "post" to the BMF, which serves as the SOI sampling frame. SOI selects the sample on a weekly basis.

Sample selection for TY 2015 occurred over the 24-month period, July 2015 through June 2017. SOI requires a 24-month sampling period for two reasons. First, just over 8 percent of all corporations use noncalendar-year accounting periods. To capture these returns, the 2015 statistics include all corporations filing returns with accounting periods ending between July 2015 and June 2016. Second, many corporations, including some of the largest corporations, request 6-month filing extensions. This combination of noncalendar-year accounting periods and filing extensions means that the last TY 2015 returns the IRS received had accounting periods ending in June 2016, and therefore, had to be filed by October 2016. However, taking into account the 6-month extension, these returns could have been filed as late as March 2017 and still be considered timely. To account for the normal processing time, the sample selection process remained open for the 2015 study until the end of June 2017. However, SOI added a few very large returns to the TY 2015 sample as late as July 2017.

Each tax return in the survey population is assigned to a stratum and subject to sampling. Each filing corporation has a unique Employer Identification Number (EIN). An integer function of the EIN, called the Transformed Taxpayer Identification Number (TTIN), is computed. The number formed by the last four digits of the TTIN is a pseudo-random number. A return for which this pseudo-random number is less than the sampling rate multiplied by 10,000 is selected for the sample.

The algorithm for generating the TTIN does not change from year to year. Therefore, corporations selected for the sample in any given year may be selected the following year, providing the corporation files a return using the same EIN and it falls into a stratum with the same or higher sampling rate. If the corporation falls into a stratum with a lower rate, the probability of selection will be the ratio of the second year sampling rate to the first year sampling rate. If the corporation files with a new EIN, the probability of selection will be independent from the prior-year selection [2].

Data Capture

Data processing for SOI begins with information already extracted for IRS administrative purposes; over 100 items available from the BMF system are checked and corrected as necessary. SOI extracts some 2,500 additional data items from the corporate tax returns during processing. This data-capture process can take as little as 15 minutes for a small, single-entity corporation filing Form 1120, or up to several weeks for a large, consolidated corporation filing several hundred attachments and schedules with the return. The process is further complicated by several factors:

- Over 2,500 separate data items may be extracted from any given tax return. This often requires constructing totals from various other items elsewhere on the return.
- Each 1120 form type has a different layout with different types of schedules and attachments, making data extraction less than uniform for the various forms.
- There is no legal requirement for a corporation to meet its tax return filing requirements by filling in, line by line, the entire U.S. tax return form. Therefore, many corporate taxpayers report financial details using schedules of their own design or using commercial tax-preparation software packages.
- There is no single accepted method of corporate tax accounting in the United States, but rather, several accepted "guidelines," which can vary by geographic location. SOI staff attempt to standardize these differences during data abstraction and editing.
- Different companies may report the same data item, such as other current liabilities, on different lines of the tax form. SOI staff also attempt to standardize these differences.

To help staff overcome these complexities and differences in taxpayer reporting, for each tax year, SOI prepares detailed instructions for the editing units at the IRS Submission Processing Centers. For TY 2015, these instructions consisted of almost 1,000 pages, covering standard and straightforward procedures and instructions for addressing data exceptions.

Data Cleaning

SOI staff enter data directly into the database from the corporate tax returns selected for the sample. In this context, the term "editing" refers to the combined interactive processes of data extraction, consistency testing, and error resolution. SOI runs over 860 tests to check for inconsistencies, including the following:

- Impossible conditions, such as incorrect tax data for a particular form type;
- Internal inconsistencies, such as items not adding to totals;
- Questionable values, such as a bank with an unusually large amount reported for cost of goods sold and/or operations; and
- Improper sample class codes, such as when a return has $100 million in total assets, but was selected as though it had $1 million because the last two digits of the total assets were keyed in as cents.

Data Completion

In addition to the tests mentioned above, SOI addresses missing data items and identifies returns to be excluded from the tabulations. The data completion process focuses on these issues.

Beginning with the TY 2012 sample, the criteria for imputing balance sheets for returns with incomplete balance sheets changed significantly. Now, only the largest returns with incomplete balance sheets are subject to SOI's balance sheet imputation procedure. As a result, the number of returns with imputed balance sheets will be negligible, and SOI will perform imputation on an ad hoc basis only.

SOI uses various methods to impute data for some certainty returns unavailable for editing, depending on the information available at the time the return needs to be completed for the sample. These corporations are identified from the previous year's sample using a combination of assets and receipts. Additional corporations may be identified to ensure industry coverage. SOI uses data filed electronically for those corporate returns selected for the sample, but unavailable for statistical processing. For TY 2015, there were 48 returns that met these criteria. For some returns not selected for the sample, if the current tax return was not located and no other current tax data were available, then SOI used data from the previous year's return, with adjustments for tax law changes, if needed.

The data completion process also includes identifying returns not eligible for the sample as the BMF may have duplicate and other out-of-scope returns. These returns include those filed by nonprofit corporations, returns having neither current income nor deductions, and prior-year tax returns. Additionally, amended or tentative returns, nonresident foreign corporations having no effectively connected income with a trade or business located in the United States, fraudulent returns, and returns filed by tax-exempt corporations are not eligible for the sample. Figure F displays the number of inactive sampled returns excluded from the tabulations, as well as the percentages of the total sample size they represent for 2012 through 2015.

Figure F. Corporation Tax Returns: Number of Inactive Sampled Returns for Tax Years 2012–2015

Type of inactive return	Tax year			
	2012	2013	2014	2015
	(1)	(2)	(3)	(4)
No income or deductions	1,986	2,058	2,558	2,235
Other*	4,447	4,436	4,158	4,519
Total	**6,433**	**6,494**	**6,716**	**6,754**
Percent of sample	**5.52**	**5.51**	**5.60**	**5.73**

*Includes duplicate returns (returns that appear more than once in the sample) and prior-year returns.

Figure G provides estimates of the number of active corporations by form type for 2012 through 2015. For Forms 1120-L and 1120-PC, these estimates may differ from the population counts in Figure E due to changes made during the data capture and data cleaning processes.

Figure G. Corporation Tax Returns: Estimated Number of Active Returns for Tax Years 2012–2015

Form type	Tax year			
	2012	2013	2014	2015
	(1)	(2)	(3)	(4)
1120	1,591,973	1,582,809	1,570,796	1,578,515
1120S	4,205,452	4,257,909	4,380,125	4,487,336
1120-L	713	647	631	601
1120-PC	9,461	10,720	11,933	13,303
1120-RIC	15,484	16,297	17,200	17,914
1120-REIT	2,146	2,472	2,764	3,078
1120-F*	15,592	16,949	18,043	18,817
Total	**5,840,821**	**5,887,804**	**6,001,491**	**6,119,565**

*Foreign Insurance Companies file on Forms 1120-L and 1120-PC, but are counted in Form 1120-F, Table 10.
NOTE: Detail may not add to total due to rounding.

Estimation

SOI bases the estimates of the total number of corporations and associated variables produced in this report on weighted sample data using either a one-step or two-step process, depending on the form type filed. Under the one-step process, SOI assigns a weight for the return, which is the reciprocal of the realized sampling rate, adjusted for unavailable returns, outliers, weight trimming, and any other necessary adjustments. SOI used these weights, referred to as the "national weights," to produce the estimates published in this report for Forms 1120-F, 1120-L, 1120-PC, 1120-RIC, and 1120-REIT,

as well as Forms 1120 and 1120S returns that were sampled with certainty.

The two-step process is used to improve the estimates by industry for returns filed on either Form 1120 or Form 1120S that are not selected in self-representing strata. The first stage of the two-step process is to assign an initial weight for the return as described above. The second stage involves post-stratification by industry and sample selection class. SOI uses a bounded raking ratio estimation approach to determine the final weights because certain post-stratification cells may have small sample sizes [3]. SOI used these final weights to produce the aggregated frequency and money amount estimates that are published in this report for these forms.

Data Limitations and Measures of Variability

SOI uses several extensive quality review processes to improve data quality. This starts at the sample selection stage with weekly monitoring to ensure the proper number of returns is selected, especially in the certainty strata. These processes continue through the data collection, data cleaning, and data completion procedures with consistency testing. Part of the review process includes extensive comparisons between the sample year (2015) and prior-year (2014) data. SOI designed each processing stage to ensure data integrity.

Sampling Error

Since the TY 2015 estimates are based on a sample, they may differ from population aggregates resulting from a complete census of all corporate income tax returns. The TY 2015 sample is one of many possible samples that could have been selected under the same sample design. Estimates derived from one possible sample could differ from those derived from another and also from the population aggregates. The deviation of a sample estimate from the average of all possible similarly selected samples is called the sampling error.

The standard error (SE), a measure of the average magnitude of the sampling errors over all possible samples, can be estimated from the realized sample. The estimated standard error is usually expressed as a percentage of the value being estimated. This is called the estimated coefficient of variation (CV) of the estimate, and it can be used to assess the reliability of an estimate. The smaller the CV, the more reliable the estimate is deemed to be.

SOI calculates the estimated coefficient of variation of an estimate by dividing the estimated standard error by the estimate itself and taking the absolute value of this ratio. Table 1 (see Section 4) shows the estimated coefficients of variation by industrial groupings for the estimated number of returns as well as selected money amounts.

The estimated coefficient of variation, CV(X), can be used to construct confidence intervals for the estimate X. The estimated standard error, which is required for the confidence interval, must first be calculated. For example, the estimated number of companies in the manufacturing sector with net income and the corresponding estimated coefficient of variation can be found in Table 1 and used to calculate the estimated standard error:

$$\begin{aligned} SE(X) &= X \bullet CV(X) \\ &= 151{,}346 \text{ x } 3.54/100 \\ &= 5{,}358 \end{aligned}$$

A 95-percent confidence interval for the estimated number of returns in manufacturing is constructed as follows:

$$\begin{aligned} X \pm 2 \bullet SE(X) &= 151{,}346 \pm (2 \text{ x } 5{,}358) \\ &= 151{,}346 \pm 10{,}716 \end{aligned}$$

The interval estimate is 140,630 returns to 162,062 returns. This means that if all possible samples were selected under the same general conditions and sample design, and if an estimate and its estimated standard error were calculated from each sample, then approximately 95 percent of the intervals from two standard errors below the estimate to two standard errors above the estimate would include the average estimate derived from all possible samples. Thus, for a particular sample, it can be said with 95-percent confidence that the average of all possible samples is included in the constructed interval. This average of the estimates derived from all possible samples would be equal to or near the value obtained from a census.

Nonsampling Error

In addition to sampling error, nonsampling error can also affect the estimates. Nonsampling errors can be classified into two groups: random errors, whose effects may cancel out, and systematic errors, whose effects tend to remain somewhat fixed and result in bias.

Nonsampling errors include coverage errors, nonresponse errors, processing errors, or response errors. The inability to obtain information for all sampled returns, differing interpretations of tax concepts or taxpayer instructions, inability to provide accurate information at the time of filing (data are collected before auditing), and inability to obtain all tax schedules and attachments may cause these errors. These errors may also be caused by data recording or coding errors, data collecting or cleaning errors, estimation errors, and failure to represent all population units.

Coverage Errors: Coverage errors in the SOI corporation data can result from the difference between the time frame for sampling and the actual time needed for filing and processing the returns. Since many of the largest corporations receive filing-period extensions, they may file their returns after the closing date for Sample Selection. However, any of the largest returns found are added into the file until the final file is produced.

Coverage problems within industrial groupings in the SOI Corporation study may result from the way consolidated

returns are filed. The Internal Revenue Code permits a parent corporation to file a single return, which includes the combined financial data of the parent and all its subsidiaries. These data are not separated into the different industries but are entered into the industry with the largest receipts. Thus, there is undercoverage of financial data within certain industries and overcoverage in others. Coverage problems within industries present a limitation on any analysis of the sample results.

Nonresponse Errors: There are two types of nonresponse errors: unit and item. Unit nonresponse occurs when a sampled return is unavailable for SOI processing. For example, other areas of the IRS may have the return at the time it is needed for statistical processing. These returns are termed "unavailable returns."

Item nonresponse occurs when certain items are unavailable for a return selected for SOI processing, even if the return itself is available. An example of item nonresponse would be items missing from the balance sheet, even though other items have been reported.

Processing Errors: Errors in recording, coding, or processing the data can cause a return to be sampled in the wrong sampling class. This type of error is called a misstratification error. One example of how a return might be misstratified is the following: a corporation files a return with total assets of $100,000,023 and net income of $5,000. A processing error causes the last two digits of the total assets to be keyed in as cents, so that the return is classified according to total assets of $1,000,000.23 and net income of $5,000.00. The return would be misstratified according to the incorrect value of the total assets stratifier. To adjust for misstratification errors, only returns selected in a noncertainty stratum that really belonged in a certainty stratum were moved to this certainty stratum.

Response errors: Response errors are due to data being captured before audit. Some purely arithmetical errors made by the taxpayer are corrected during the data capture and cleaning processes. Because of time constraints, SOI does not incorporate adjustments to a return during audit into the file.

References

[1] Jones, H. W., and McMahon, P. B. (1984), "Sampling Corporation Income Tax Returns for Statistics of Income, 1951 to Present," *1984 Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 437–442.

[2] Harte, J. M. (1986), "Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS," *1986 Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 603–608.

[3] Oh, H. L., and Scheuren, F. J. (1987), "Modified Raking Ratio Estimation," *Survey Methodology,* Statistics Canada, Vol. 13, No. 2, pp. 209–219.