Section 3 Description of the Sample and Limitations of the Data

This section describes the 2010 Corporate sample design, sample selection, data capture, data cleaning, and data completion. The techniques used to produce estimates as well as an assessment of the data limitations, including sampling and non-sampling errors, are also discussed.

Background

From Tax Year 1916 through Tax Year 1950, data were extracted for the Statistics of Income (SOI) program from each corporate return filed. Stratified probability sampling was introduced for Tax Year 1951. Since that time, the sample size has generally decreased while the population has increased. For example, for Tax Year 1951 the sample comprised 41.5 percent of the entire population, or 285,000 of the 687,000 total returns filed. In comparison, for 2010, the sample proportion was about 1.74 percent of the total population of just over 6.26 million. 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, stratification was by size of total assets and industry. From 1952 through 1967, the stratification was by a measure of size only. The size was measured by volume of business (1953-1958) or total assets (1952 and 1959-1967). Since 1968, returns have been stratified by both total assets and, for Form 1120 and 1120S returns, 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 that are part of the SOI study.

Survey Population

The survey population includes the returns that filed one of the 1120 forms selected for the SOI study and posted to the IRS Business Master File (BMF). Amended returns and returns for which the tax liabilities changed because of a tax audit are excluded. Figure E gives the number of corporate returns by form type that were subject to sampling during Tax Years 2007 through 2010.

Bertrand Überall and Richard Collins were responsible for the sample design and estimation of the SOI 2010 Corporation Program under the direction of Tamara Rib, Chief, Mathematical Statistics Section, Statistical Computing Branch.

Figure E.—Population Counts by Corporate Form Type, Tax Years 2007-2010

Form Type	Tax Year			
	2007	2008	2009	2010
1120	2,151,182	2,001,930	1,927,971	1,867,941
1220-S	4,292,077	4,293,544	4,332,077	4,336,365
1120-L	1,001	891	825	748
1120-PC	7,254	7,828	8,104	8,572
1120-RIC	12,192	13,221	13,106	13,385
1120-REIT	1,664	1,679	1,672	1,798
1120-F	30,532	30,620	30,295	32,414
Total	6,495,902	6,349,713	6,314,050	6,261,223

Note: Beginning in SOI 2008, older returns with very early accounting periods are excluded from the sampling frame.

Sample Design

The current sample design is a stratified probability sample, with stratification by form type, and either size of total assets alone, or both size of total assets and a measure of income. Form 1120 is stratified by size of total assets and size of "proceeds". Size of "proceeds", the measure of income, is the larger of the absolute value of net income (or deficit) or the absolute value of "cash flow", which is the sum of net income, several depreciation amounts, and depletion. Forms 1120-F, 1120-L, 1120-PC, 1120-RIC, and 1120-REIT are each stratified by size of total assets only. Form 1120S is stratified by size of total assets and size of ordinary income.

The design process began with projected population totals that were derived from IRS administrative workload estimates, adjusted according to the distribution by strata of the population from several previous survey years. Using projected population totals by sample strata, an optimal allocation, based on stratum standard errors, was carried out to assign sample sizes to each stratum such that the overall targeted sample size was approximately 115,000. A Bernoulli sample was selected independently from each stratum with sampling rates ranging from 0.25 percent to 100 percent. Figure F on the following page shows the stratum boundaries, sampling rates, and frame population and sample counts from the BMF for each form type. This table also shows the population and sample counts after adjustments for missing returns, outliers, and weight trimming. The total realized sample for Tax Year 2010, including inactive and non-eligible corporations, is 108,763 returns.

Figure F.—Corporation Returns: Number Filed, Number in Sample, and Sampling Rates, by Selection Class

Sample class number	Description of sample selection classes		Sampling Rates (%)	Number of returns			
	Size of total assets	Size of proceeds*		BMF counts		After adjustments**	
				Population	Sample	Population	Sample
	All Returns, Total			**6,261,223**	**108,763**	**6,261,223**	**108,606**
	Form 1120 (no Form 5735 attached), Total ***			**1,862,370**	**50,182**	**1,862,372**	**50,093**
1	Under $50,000	Under $25,000	0.40	794,684	3,178	794,685	3,175
2	$50,000 - $100,000	$25,000 - $50,000	0.40	203,680	813	203,680	813
3	$100,000 - $250,000	$50,000 - $100,000	0.40	268,258	1,080	268,258	1,080
4	$250,000 - $500,000	$100,000 - $250,000	1.09	199,588	2,191	199,588	2,190
5	$500,000 - $1,000,000	$250,000 - $500,000	1.81	148,413	2,589	148,413	2,587
6	$1,000,000 - $2,500,000	$500,000 - $1,000,000	3.48	119,260	4,199	119,261	4,196
7	$2,500,000 - $5,000,000	$1,000,000 - $1,500,000	5.94	48,725	2,904	48,725	2,900
8	$5,000,000 - $10,000,000	$1,500,000 - $2,500,000	10.55	29,195	3,145	29,195	3,144
9	$10,000,000 - $25,000,000	$2,500,000 - $5,000,000	27.00	21,250	5,779	21,250	5,763
10	$25,000,000 - $50,000,000	$5,000,000 - $10,000,000	50.00	10,063	5,050	10,063	5,033
11	$50,000,000 - $100,000,000	$10,000,000 - $15,000,000	100.00	6,110	6,110	6,115	6,093
12	$100,000,000 - $250,000,000	$15,000,000 or more	100.00	6,594	6,594	6,594	6,574
13	$250,000,000 - $500,000,000		100.00	2,803	2,803	2,797	2,797
14	$500,000,000 or more		100.00	3,747	3,747	3,748	3,748
	Form 1120S, Total ***			**4,335,197**	**32,974**	**4,335,194**	**32,932**
15	Under $50,000	Under $25,000	0.25	1,717,884	4,254	1,717,884	4,246
16	$50,000 - $100,000	$25,000 - $50,000	0.25	640,630	1,630	640,630	1,625
17	$100,000 - $250,000	$50,000 - $100,000	0.25	736,474	1,817	736,474	1,817
18	$250,000 - $500,000	$100,000 - $250,000	0.31	533,169	1,684	533,169	1,681
19	$500,000 - $1,000,000	$250,000 - $500,000	0.56	307,965	1,689	307,965	1,686
20	$1,000,000 - $2,500,000	$500,000 - $1,000,000	0.99	216,530	2,115	216,530	2,112
21	$2,500,000 - $5,000,000	$1,000,000 - $1,500,000	1.56	84,536	1,290	84,536	1,290
22	$5,000,000 - $10,000,000	$1,500,000 - $2,500,000	2.52	49,799	1,285	49,799	1,284
23	$10,000,000 - $25,000,000	$2,500,000 - $5,000,000	20.00	30,132	5,940	30,132	5,936
24	$25,000,000 - $50,000,000	$5,000,000 - $10,000,000	30.00	9,646	2,838	9,644	2,834
25	$50,000,000 - $100,000,000	$10,000,000 - $15,000,000	100.00	4,209	4,209	4,209	4,205
26	$100,000,000 - $250,000,000	$15,000,000 or more	100.00	3,057	3,057	3,056	3,050
27	$250,000,000 or more		100.00	1,166	1,166	1,166	1,166
	Form 1120-L, Total			**589**	**333**	**589**	**334**
28	Under $10,000,000		43.00	415	159	414	159
29	$10,000,000 - $50,000,000		100.00	102	102	101	101
30	$50,000,000 - $250,000,000		100.00	33	33	33	33
31	$250,000,000 or more		100.00	39	39	41	41
	Form 1120-F, Total			**32,341**	**5,064**	**32,342**	**5,053**
32	Under $10,000,000		13.00	30,285	3,951	30,284	3,940
33	$10,000,000 - $50,000,000		13.00	1,072	129	1,072	129
34	$50,000,000 - $250,000,000		100.00	549	549	549	547
35	$250,000,000 or more		100.00	435	435	437	437
	Form 1120-PC, Total			**8,155**	**1,831**	**8,155**	**1,828**
36	Under $2,500,000		10.00	5,779	553	5,779	551
37	$2,500,000 - $10,000,000		25.00	1,456	358	1,456	358
38	$10,000,000 - $50,000,000		100.00	736	736	736	735
39	$50,000,000 - $250,000,000		100.00	176	176	176	176
40	$250,000,000 or more		100.00	8	8	8	8
	Form 1120-REIT, Total			**1,784**	**1,458**	**1,784**	**1,457**
41	Under $10,000,000		25.00	449	123	444	118
42	$10,000,000 - $50,000,000		100.00	360	360	361	360
43	$50,000,000 - $250,000,000		100.00	449	449	449	449
44	$250,000,000 or more		100.00	526	526	530	530
	Form 1120-RIC, Total			**13,374**	**9,508**	**13,374**	**9,505**
45	Under $10,000,000		15.00	2,641	395	2,633	387
46	$10,000,000 - $50,000,000		30.00	2,289	669	2,289	668
47	$50,000,000 - $100,000,000		100.00	1,206	1,206	1,205	1,203
48	$100,000,000 - $250,000,000		100.00	2,019	2,019	2,016	2,016
49	$250,000,000 - $500,000,000		100.00	1,577	1,577	1,578	1,578
50	$500,000,000 or more		100.00	3,642	3,642	3,653	3,653
51	**Special Studies** (All Form Types)****		100.00	**7,413**	**7,413**	**7,413**	**7,404**†

* Proceeds is defined as the larger of absolute value of net income (deficit) or absolute value of cash flow (net income + depreciation + depletion).
** Includes adjustments for missing returns, undercoverage, outliers, and weight trimming.
*** Returns were classified according to either size of total assets or size of proceeds, whichever corresponded to the higher sample class.
Example: A Form 1120 return with total assets of $750,000 and proceeds of $75,000 is in sample class 8 (based on total assets), rather than in sample class 6 (based on proceeds).
**** Includes Form 1120 returns with Form 5735 (Possessions) attached.
† The adjusted sample count is lower than the adjusted population count due to returns unavailable for processing.

Sample Selection

Corporation income tax returns are processed at the Cincinnati and Ogden IRS Submission Processing Centers. All corporate returns are processed initially to determine tax liability. Then, the tax data are transmitted and updated on a weekly basis to the IRS Business Master File (BMF) system located in Martinsburg, West Virginia. These returns are said to "post" to the BMF. This BMF database serves as the SOI sampling frame. The SOI sample is also selected on a weekly basis.

Sample selection for Tax Year 2010 occurred over the period of July 2010 through June 2012. A 24-month sampling period is needed for two reasons. First, approximately 10.4 percent of all corporations had noncalendar year accounting periods. In order to take these filings into consideration, the 2010 statistics represent all corporations filing returns with accounting periods ending between July 2010 and June 2011. Also, many corporations, including some of the largest, request six-month filing extensions. The combination of noncalendar year filing and filing extensions means that the last Tax Year 2010 returns that the IRS received (those with accounting periods ending in June 2011, which must therefore be filed by October 2011) could be timely filed as late as March 2012, taking into account the six-month extension of the October 2011 due date. Normal administrative processing time lags required that the sample selection process remain open for the 2010 study until the end of June, 2012. However, a few very large returns for Tax Year 2010 were added to the sample as late as August 2012.

Each tax return posted to the BMF and in the survey population (as defined above) is assigned to a stratum and subjected 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 in the sample.

The algorithm for generating the TTIN does not change from year to year, so any corporation selected into the sample in a given year will be selected again the next year, providing that the corporation files a return using the same EIN in the two years and that 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. Some 1,630 additional data items are extracted from the tax returns during SOI processing. The SOI data capture process can take as little time as fifteen minutes for a small, single entity corporation filing on 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 1,630 separate data items may be extracted from any given tax return, and often require totals to be constructed from various other items on other parts of 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 form types.

- There is no legal requirement that a corporation meet its tax return filing requirements by filling in, line by line, the entire U.S. tax return form. Therefore, many corporate taxpayers report many of their financial details in schedules of their own design, or using commercial tax-preparation software packages.

- There is no single accepted method of corporate tax accounting used throughout the country, but rather several accepted accounting "guidelines," many of which are unique to geographic locations. 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. Again, SOI staff attempt to standardize these differences.

To help SOI editors overcome these complexities and differences due to taxpayer reporting, SOI staff prepares detailed editing instructions for the SOI editing units at the IRS Submission Processing Centers each tax year. For Tax Year 2010, these instructions consisted of almost 1,000 pages covering standard and straightforward procedures and instructions for exceptions that might be encountered.

Data Cleaning

Statistical processing of the corporate returns is performed in an online computer environment, where the data from returns selected for the corporate sample are entered directly into the SOI corporation database. In this context, the term "editing" refers to the combined interactive processes of data extraction, consistency testing, and error resolution. There are over 860 of these tests, which look for such inconsistencies as:

- 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 mistakenly keyed in as cents.

Data Completion

In addition to the tests mentioned above, missing data problems must be addressed and returns that are to be excluded from the tabulations must be identified. The data completion process focuses on these issues.

If the missing data items are from the balance sheet, then imputation procedures are used. If data for a whole return are missing because the return is unavailable to SOI during the data capture process, imputation procedures are also used in certain cases.

A ratio-based imputation procedure is used to estimate missing balance sheet items for all 1120 forms except those with less than 12-month accounting periods. The ratios are determined using the most recent data available, either the corporation's Tax Year 2009 return if the corporation filed a return for 2009 and the balance sheet was not already imputed for 2009, or the Tax Year 2008 aggregate data for the corporation's minor industrial group, which are the most recent aggregate data available at the time that editing for Tax Year 2010 begins (which is in mid-June of Calendar Year 2011). If the reported balance sheet items do not balance (i.e., the sum of asset items does not equal the sum of liability and shareholders' equity items), then the missing items are imputed. If the total assets amount is among the missing items, this item is imputed first based on the ratio of total assets to business receipts (or total receipts) from either the corporation's Tax Year 2009 return, or the Tax Year 2008 aggregate data for the corporation's minor industry. The other missing items are then imputed based on the ratios so that the total of all asset items and the total of all liability items are both equal to the total assets amount, whether this amount was reported or imputed. A description of the balance sheet imputation process is given in reference [3]. The following chart shows the number of sampled returns that had balance sheet items imputed, as well as the percentages they represent of the total sample sizes, for Tax Years 2007 through 2010.

Returns with imputations	Tax Year			
	2007	2008	2009	2010
Number of imputed returns	42	52	63	42
Percent imputed	0.04	0.05	0.06	0.04

For Tax Year 2010, the total assets from returns which had imputed total assets represent only a negligible fraction of the total estimated assets for all active returns in the Tax Year 2010 sample.

Data for unavailable critical corporations are imputed in various ways, depending on what information is available at the time the SOI database is produced. Critical corporations are identified from the previous year's sample using a combination of assets and receipts. Supplementary critical corporations may be identified to ensure industry coverage. For critical corporations selected for the sample but unavailable for statistical processing through the regular procedures, electronically filed data are used. For Tax Year 2010, there are 42 returns that meet these criteria. For critical corporations not selected for the sample, if the current tax return is not located and no other current tax data are available, data from the previous year's return are used, with adjustments for tax law changes if needed. There are no returns derived from prior year returns in the Tax Year 2010 data.

Another part of the data cleaning process is identifying sampled returns that are not eligible for the sample. The BMF system used for sample selection can include duplicate tax returns and other out-of-scope returns, such as returns of nonprofit corporations, returns having neither current income nor deductions, prior-year tax returns, amended or tentative returns, returns of nonresident foreign corporations having no effectively connected income with a trade or business located within the United States, fraudulent returns, and returns of corporations that are exempt from taxation.

Figure G below displays the number of inactive sampled returns that were excluded from tabulations, as well as the percentages they represent of the total sample sizes, for Tax Years 2007 through 2010.

Figure G.—Number of Inactive Sampled Returns for Tax Years 2007-2010

Type of inactive return	Tax Year			
	2007	2008	2009	2010
No Income or Deductions	1,603	1,480	1,360	1,608
Other*	6,562	5,367	5,145	4,686
Total	8,165	6,847	6,505	6,294
Percent of sample	7.12	6.09	5.95	5.80

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

Estimates of the number of active corporations by form type for Tax Years 2007 through 2010 are provided in Figure H below. For Forms 1120-L and 1120-PC, these estimates may be different than the population counts in Figure E due to changes made during the data capture and data cleaning processes.

Figure H.—Estimated Number of Active Returns for Tax Years 2007-2010

Form Type	Tax Year			
	2007	2008	2009	2010
1120	1,846,134	1,762,483	1,694,869	1,649,285
1120S	3,989,893	4,049,943	4,094,562	4,127,554
1120-L	1,027	945	866	796
1120-PC	7,174	7,670	7,890	8,244
1120-RIC	12,083	13,140	13,043	13,256
1120-REIT	1,641	1,660	1,635	1,766
1120-F*	10,896	11,379	11,680	12,824
Total	5,868,849	5,847,221	5,824,545	5,813,725

*Foreign Insurance Companies file on Forms 1120-L and 1120-PC, but are counted in Form 1120-F Tables 10 and 11.
Detail may not add to total due to rounding.

Estimation

Estimates of the total number of corporations and associated variables produced in this report are based on weighted sample data. Either a one-step process or a two-step process is used to determine the weights, depending on the return's form type.

Under the one-step process, the weights are assigned as the reciprocal of the realized sampling rate, adjusted for unavailable returns, outliers, weight trimming, as well as any other adjustments that might be needed. These weights, referred to as the "national weights", are used to produce the estimates published in this report for Forms 1120-F, 1120-L, 1120-PC, 1120-RIC, 1120-REIT and Form 1120 with Form 5735 attached, as well as for Form 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 1120S that are not selected in self-representing strata. The first stage is the one-step process described above, which provides an initial weight for the return. The second stage involves post-stratification by industry and sample selection class. A bounded raking ratio estimation approach is applied in order to determine the final weight, because certain post-stratification cells may have small sample sizes [4]. These final weights are used to produce the aggregated frequency and money amount estimates that are published in this report for these forms.

Data Limitations and Measures of Variability

Several extensive quality review processes are used to improve data quality, beginning at the sample selection stage with weekly monitoring to ensure that the proper number of returns is being selected, especially in the certainty strata. They continue through the data collection, data cleaning, and data completion procedures with consistency testing. Part of the review process includes extensive comparisons between the 2010 and 2009 data. A great amount of effort is made at every stage of processing to ensure data integrity.

Sampling Error

Since the corporation estimates are based on a sample, they may differ from the population aggregates that would have been obtained if a complete census of all income tax returns had been taken. The particular sample used to produce the results in this report is one of a large number of possible samples that could have been selected under the same sample design. Estimates derived from one of the possible samples could differ from those derived from other samples and 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 judged to be.

The estimated coefficient of variation of an estimate is calculated by dividing the estimated standard error by the estimate itself and taking the absolute value of this ratio. Estimated coefficients of variation by industrial groupings for the estimated number of returns, as well as for selected money amount estimates, are shown in Table 1 of this report. For the estimated number of returns by asset size and sector, estimated coefficients of variation are given in Figure I on page 15. The corresponding estimates are in Table 4.

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:

$$SE(X) = X \bullet CV(X)$$
$$= 147{,}353 \times 3.62/100$$
$$= 5{,}334$$

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

$$X \pm 2 \bullet SE(X) = 147{,}353 \pm (2 \times 5{,}334)$$
$$= 147{,}353 \pm 10{,}668$$

The interval estimate is 136,685 returns to 158,021 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. These errors can be the result of the inability to obtain information about all returns in the sample, differing interpretations of tax concepts or instructions by the taxpayer, inability to provide accurate information at the time of filing (data are collected before auditing), inability to obtain all tax schedules and attachments, errors in recording or coding the data, errors in collecting or cleaning the data, errors made in estimating for missing data, 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 extensions to their filing periods, they may file their returns after sample selection has ended for that tax year. 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 result from the way consolidated returns may be 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: 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." In 2010, there were 150 such unavailable returns in the corporation study, which constituted about 0.14 percent of the total sample. The number of unavailable returns and their percentages of the total sample size for Tax Years 2007 through 2010 are shown in the following chart.

Unavailable returns	Tax Year			
	2007	2008	2009	2010
Number of unavailable returns	530	293	141	150
Percent unavailable	0.46	0.26	0.13	0.14

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 when items are missing on the balance sheet, even though other items are reported.

Figure I.—Coefficients of Variation (CVs) for Number of Returns, by Asset Size and Sector, for Tax Year 2010

Sector	All asset sizes	Size of total assets			
		Zero Assets	$1 under $ 500,000	$500,000 under $1,000,000	$1,000,000 under $5,000,000
	(1)	(2)	(3)	(4)	(5)
All industries[1]	**0.18**	**1.61**	**0.46**	**1.12**	**0.58**
Agriculture, forestry, fishing, and hunting	2.54	13.10	4.15	5.27	3.68
Mining	6.91	21.61	10.95	16.34	9.08
Utilities	17.94	50.41	28.88	39.74	21.00
Construction	1.04	4.85	1.67	4.33	2.32
Manufacturing	2.71	10.01	4.43	6.53	3.16
Wholesale and retail trade	1.01	4.64	1.57	2.72	1.45
Transportation and warehousing	2.47	7.87	4.08	9.35	5.53
Information	3.97	11.48	5.46	14.07	8.66
Finance and insurance	2.23	7.94	3.57	7.14	4.50
Real estate and rental and leasing	1.14	4.76	1.98	2.94	1.95
Professional, scientific, and technical services	1.10	4.10	1.59	5.92	3.76
Management of companies (holding companies)	5.86	14.16	11.32	14.89	8.99
Administrative and support and waste management and remediation services	2.69	7.18	3.54	10.77	6.25
Educational services	7.03	14.34	8.91	32.88	18.63
Health care and social assistance	1.36	7.12	1.79	6.74	5.84
Arts, entertainment, and recreation	4.11	10.87	5.52	14.62	11.46
Accommodation and food services	1.48	7.85	2.22	6.04	4.77
Other services	2.02	6.54	2.69	6.23	5.24

Sector	Size of total assets—continued				
	$5,000,000 under $10,000,000	$10,000,000 under $25,000,000	$25,000,000 under $50,000,000	$50,000,000 under $100,000,000	$100,000,000 under $250,000,000
	(6)	(7)	(8)	(9)	(10)
All industries[1]	**1.03**	**0.46**	**0.58**	**0.13**	**0.10**
Agriculture, forestry, fishing, and hunting	9.40	4.08	7.00	0.00	0.00
Mining	10.34	4.68	4.48	0.46	0.61
Utilities	32.92	13.57	11.11	0.00	0.00
Construction	3.36	1.82	2.65	0.18	0.37
Manufacturing	3.71	1.35	1.59	0.11	0.15
Wholesale and retail trade	2.64	0.93	1.42	0.12	0.10
Transportation and warehousing	9.95	3.76	4.88	0.47	0.65
Information	9.93	3.90	4.19	0.31	0.38
Finance and insurance	5.63	1.86	1.93	0.52	0.26
Real estate and rental and leasing	3.72	1.71	2.31	0.13	0.92
Professional, scientific, and technical services	5.74	2.63	2.93	0.18	0.26
Management of companies (holding companies)	9.13	3.94	3.01	0.16	0.11
Administrative and support and waste management and remediation services	10.96	6.08	6.34	0.62	0.88
Educational services	33.05	13.25	12.28	0.00	0.00
Health care and social assistance	12.42	5.50	5.32	0.68	0.86
Arts, entertainment, and recreation	17.83	7.49	7.18	0.00	0.00
Accommodation and food services	5.41	3.71	4.84	0.74	0.00
Other services	15.22	7.94	9.82	0.00	0.00

[1]Includes returns not allocable by sector.

Note: Returns with assets of $250,000,000 or more are self-representing and thus are not subject to sampling error.

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 mis-stratification error. One example of how a return might be mis-stratified 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 mis-stratified according to the incorrect value of the total assets stratifier. To adjust for mis-stratification errors, only returns selected in a non-certainty stratum which really belonged in a certainty stratum were moved to this 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, adjustments to a return during audit are not incorporated into the SOI 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] Überall, B. (1995), "Imputation of Balance Sheets for the 1992 SOI Corporate Program," *1995 Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 275-280.

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