Selecting the SOI Individual Tax Return Sample

Tracy Haines, Valerie Testa
Internal Revenue Service, 77 K Street NE, Washington, DC 20002

Abstract

The Statistics of Income (SOI) Division of the Internal Revenue Service is responsible for conducting the Individual Tax Return Sample Study. SOI's annual Individual Sample is selected using a Stratified Bernoulli Sample. In return, the sample is used for cross-sectional estimation of information reported on the Tax Form 1040. This paper details changes made to the sample design as a result of analysis to evaluate potential improvements. Changes to SOI's Individual Tax Return Sample over time are necessary for better accuracy and effectiveness.

Keywords: Bernoulli Sample, cross-sectional estimation

1. Introduction

1.1 SOI Individual Tax Return Sample

The Internal Revenue Service's Statistics of Income (SOI) Division selects a sample of individual income tax returns. It was last redesigned for tax year 1991. The sample design was recently reexamined, and we document some of the changes resulting from the review here. The Statistics of Income sample is designed to support annual estimation for Individual tax return information. It provides a sample for revenue estimation conducted by Treasury's Office of Tax Analysis (OTA) and the Congressional Joint Committee on Taxation (JCT).

The target population for the Individual Cross-Sectional Sample includes all pre-audited individual Income Tax Returns, for a certain tax year, from all forms 1040, 1040A and 1040EZ. The target population does not include tentative or amended returns.

The sampling frame for a certain tax year consists of tax returns posted to the Individual Master File (IMF) during the following calendar year. For example, the tax year 2013 sample corresponds to income incurred from January through December 2013, and reported to the IRS in 2014. Amended or tentative returns, also known as out-of-scope returns, are excluded from sampling.

The cross-sectional sample is selected using a stratified Bernoulli sampling design. We set all sample selection rates in advance of collecting the sample, based on predictions of Form 1040 filings for a given calendar year. Each return is independently selected for the sample using Bernoulli sampling with probabilities of selection that vary across the strata. Returns are sampled daily as the IRS processes the returns for tax administration purposes.

1.2 Definition of Income

Income is measured by taking the larger of gross positive income and gross negative income (Czajka, 2014). All components included in this calculation are items taken from Form 1040 and supplemental schedules C, D, E, and F.

The concept of income is more than just the total income components included in Adjusted Gross Income (AGI). By using the component item amounts rather than the net computed AGI, the gross positive and gross negative income amounts can be larger than the AGI itself. Returns with higher component amounts will ultimately be selected at a higher sampling rate than returns with lower component item amounts. These returns have more use for tax policy analysis purposes.

1.3 Data Processing

Individual tax returns are filed and reviewed for tax liability purposes at IRS processing centers throughout the country. All tax data is transmitted and updated on a daily basis to the individual master file (IMF) system. Data processing for the SOI sample begins with loading data into a database, including information that has already been extracted for administrative tax purposes. This database controls information for each record designated for the sample.

All data is subjected to testing for consistency and identified for manual correction of errors when needed. Once reviewed, validated, tested, and balanced, computer adjustments and prior year imputations are used for consistency. Sample weights are then assigned to each record in the sample data file. Data is then tabulated for publication. All statistics are also reviewed for accuracy.

Sample weights are calculated by dividing the population count of returns in a stratum by the number of sample returns for that stratum. Weights are then adjusted for any misclassified returns.

1.4 Data Customers and Related Samples

SOI's Individual Tax Return Sample has a wide variety of uses. The Revenue Act of 1916 mandated the annual publication of statistics related to "the operations of the internal revenue laws". SOI fulfills this requirement by using this sample in the publication of such statistics as those laws impact individuals. The data is used by Treasury's Office of Tax Analysis (OTA), the Congressional Joint Committee on Taxation (JCT), and others to meet a wide range of analyses, including revenue estimation and gauging impacts of tax law changes.

A public use file is created from the individual tax return sample after rigorous data disclosure methods are applied. This is designed to tabulate and present statistical information for the Individual Tax Returns filed in a given tax year. This Individual Tax File is designed for making national level estimates (Bryant, 2012). Also, coefficients of variation (CVs) are computed using the sample data to gauge the impact of sampling errors.

The Sales of Capital Assets (SOCA) cross-sectional sample is selected during daily processing throughout each year. The study is not used yearly, but is repeated periodically. SOCA strata are determined by the mode of filing (paper or electronic) and

by Indexed Total Positive Income and Indexed Total Negative Income. Each SOCA return is processed at the transaction level, while all returns in the cross-sectional sample are processed at the return level. The SOCA cross-sectional sample is used to study items on Form 1040 such as the total amount of Sales Price, Basis, and Net Gain/Loss.

SOI's Individual Income Tax Return Sample is also used to create the Edited Panel Sample. This Panel Sample is used to study the Sales of Capital Assets. Sales of capital assets include such things as stocks, bonds, mutual funds, property, as well as other assets. While the yearly SOI cross-sectional sample produces adequate estimates of capital gains and losses on a tax return basis, the panel study provides much more detailed information about each financial transaction reported on Schedule D as well as other forms (Liu, 2009). The panel sample is used for these cross-sectional estimates in addition to the longitudinal analysis.

2. Sample Selection

The cross-sectional estimation is a stratified Bernoulli sample. All sample selection rates are set in advance of collecting the sample based on predictions of Form 1040 filings for a given calendar year. The sample is selected as data are collected through the IRS Individual Master File (IMF) as tax returns are filed and entered into the IRS system. Under Bernoulli sampling, this results in a random sample size.

2.1 Strata Attributes

Strata are formed based on having high income and no tax liability or high business and farm receipts, the presence or absence of certain forms, attributes of interest to OTA for modeling purposes, and size and composition of income. Each stratum is identified by a three-digit stratum identification code, called the sample code.

The following attributes are used for stratification:

1. Adjusted Gross Income or Expanded Income of $200,000 or more and no tax liability, including no alternative minimum tax. These returns are referred to as "High Income Non-Taxable" or HINT returns. Expanded Income is defined as the sum of Adjusted Gross Income, Intangible Drilling Costs, Incentive Stock Options, Tax Exempt Interest, Depletion, Nontaxable Social Security, and the Foreign Earned Income Exclusion less any associated deductions.

2. Combined business and farm receipt of $50,000,000 or more.

3. Presence or absence of special forms or schedules of special interest (Form 2555 (Foreign Earned Income), Form 1116 (Foreign Tax Credit), Form 1040 Schedule C (Profit or Loss from Business), and Form 1040 Schedule F (Profit or Loss from Farming)).

4. Interest in the return for revenue estimation and econometric modeling. This attribute is used to ensure that the sample will contain sufficient numbers of returns with less common sources of income and filing situations to allow for reliable modeling of the effects of tax law and policy changes.

5. Totals of positive or negative income items reported on the return and total business receipts or inventories for returns with at least one Schedule C (Profit or Loss from Business) attached.

All returns are stratified using an indexed Total Positive Income or indexed Total Negative Income. Returns with Schedule C attached may be assigned to a stratum with a higher probability of selection because of Combined Gross Receipts or Ending Inventory from Schedule C. The index comes from the Chain-Type Price Index for Gross Domestic Product with a base year of 1991.

2.2 Total Positive and Total Negative Income

There are 35 components that contribute to Total Positive Income and Total Negative Income. They are all found on an Individual Tax Return form. There are nine components of income that are strictly positive. These consist of Wage Amount, Tested Tax Exempt Interest, Taxable Dividends, Alimony Received, Tested Pension Amount, Taxable IRA Distribution, Tested Unemployment Compensation, Tested Social Security, and State Income Tax Refund. There are four other items that are added if they are strictly gains. These are Total Rental Payments Amount, Total Royalty Payments Amount, Partnership and/or S-Corporation Income, and Estate/Trust Income.

Like positive income, negative income has strictly loss items and deduction or adjustments amounts. The eight strictly loss components are Partnership, S Corporation Loss, Estate and Trust Loss, Total Expenses All Property Amount, Total Depreciation All Property Amount, Alimony Paid, Form 3903 Moving Expense Amount, and Business at Home Expense. The three Deduction or Adjustment items that are added to negative income are Total Deductions, Total Farm Expenses, and Negative Income Adjustment.

Depending on the sign of the item, some income components could be categorized into either positive or negative income. Within this list are the following items: Schedule C-1 Gross Profit/Loss, Schedule C-2 Gross Profit/Loss, Schedule C-3 Gross Profit/Loss, Schedule F-1 Gross Income/Loss, Schedule F-2 Gross Income/Loss, Supplemental Gains/Losses, Other Income Amount, Farm/Rent Income/Loss, Taxable Interest Income, Net Short Term Gain/Loss Amount, and Net Long Term Gain/Loss Amount.

2.3 Sample Code

Each stratum is identified with a three-digit code, called the sample code. The return type becomes the first digit of the sample code, which is used to assign the return to a sampling stratum. The return types shown in Table 1 are assigned hierarchically. Table 1, located in the Appendix, shows the return type code and definition of what is included within that return type.

Income class code is a two-digit code that becomes the second and third digits of the sample code. If Total Positive Income is greater than or equal to Total Negative Income then the income class is determined by Indexed Total Positive Income. If Total Negative Income is greater than Total Positive Income then the income class is determined by Indexed Total Negative Income.

Indexing of positive and negative income is done by dividing each by the ratio of the Chain-Type Price Index for the Gross Domestic Product for the fourth quarter of previous tax year to the fourth quarter of the base year 1991. The indices were calculated using the Gross Domestic Product (GDP) Chain-type Price Index (U.S. Bureau of Economic Analysis).

Table 2, also located in the Appendix, is the basis for determining the income class code. It shows the boundaries for both negative and positive income, the income class codes, and the transformed taxpayer identification number (TTIN) cutoffs for each stratum.

To ultimately assign the sample code, the return type and income class codes are concatenated. The return type is the leftmost position while the two rightmost positions are the income class code.

2.4 Components of Sample

The individual cross-sectional sample is the union of two independently drawn samples. Within each stratum, there are two methods of selecting a return. These two selection methods are the Continuous Work History Study (CWHS) and a random number selection.

The continuous work history study, or CWHS, is a simple random sample based on the last 4 digits of the SSN. The CWHS utilizes a feature of the SSN numbering system where the last four digits of the number have the properties of a random number. Thus, by sampling on the last four digits, a random sample can be obtained (Weber, 2001). There are 10 possible 4-digit SSN ending combinations that make up the continuous work history study. Both primary and secondary SSNs are used for the full CWHS sample, but only the 10 primary SSN ending combinations are embedded in the annual cross-sectional sample. Once a return is selected for a given tax year, that specific SSN will always be selected annually in order to provide some year-to-year stability.

The second part of the annual cross-sectional sample is the random number selection. The primary taxpayer's social security number (SSN) is used to create a permanent random number for sample selection purposes. The primary SSN is a seed to an algorithm that generates an 11-digit number, the transformed taxpayer identification number (TTIN). The diagram in Figure 1 shows this transformation.

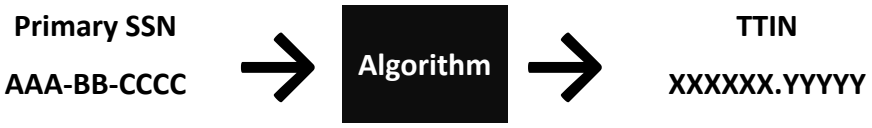

Figure 1: Calculating the TTIN

All returns that are not included in certainty strata are selected by comparing the last 5 digits of this transformed SSN to the range of all possible values chosen to determine the desired probability of selection for that stratum. Because the same algorithm is used every year, and the SSN is used as the base for the algorithm, this will always be a permanent random number. If the last five digits of the TTIN, shown above as YYYYY,

are less than the predetermined sampling rate for the stratum multiplied by 100,000, then the associated return is selected for the sample. Any return selected into an SOI sample in a given year will be selected again the next year, provided that the return is filed using the same identifier in the two years and it falls into a stratum with the same or higher sampling rate.

3. Sample Design Changes

3.1 Why change?

The SOI Individual Tax Return sample has not undergone a major redesign in over 25 years. Although it still serves the needs of its principal customers well, there are certain aspects that could be revisited to better suit the customers' needs. Customers' needs have changed. The population has grown. The technology has changed. By revisiting the design, SOI can better meet the customers' needs while also doing so more efficiently.

Due to significant income growth at the upper end of the sample distribution, an increased fraction of the population is being sampled with certainty and provides substantially more precise estimates than it was designed to provide. While SOI's unit editing costs have declined, reallocating some of the resources away from the Individual sample should be addressed (Czajka, 2014).

3.2 Improvement Suggestions

After thorough examination, an assessment was done to communicate the redesign needs for the SOI Individual Tax Return Sample. Some recommendations were to retain the current methods and procedures such as handling missing returns, misclassification errors, and prior year returns, as well as maintaining equal selection rates for electronic and paper returns. Other recommendations were to add additional components to the definition of income, reallocation of the sample by changing the stratum boundaries, and replacing the current index with one based on personal income. Also, it was suggested to eliminate sub stratification by degree of interest within the sample.

4. Timeline for Improvements

4.1 What has been changed?

The Individual Tax Return Sample Study has already implemented some of these changes, effective processing year 2016. Positive income is now revised to include State Income Tax Refund amount. There were concerns that at times, AGI could be larger than positive income. By implementing this change, it will prevent AGI from being greater than the positive income amount.

Another suggestion that has been implemented is the elimination of sub-stratification by degree of interest. This sub-stratification is no longer useful. As a result, the income class codes are now condensed. You can see these changes by looking at Table 2's "NEW" column in the Appendix.

Return type amounts and sample codes have also been collapsed. By collapsing the return types, this eliminated some post-stratification and allows it to occur during the sample

selection. Looking at the Appendix, Table 1, one can see the newly implemented collapsing.

4.2 What will still be changed?

Within the near future, more changes are set to be made. These include reallocating the income stratum boundaries, reevaluating the current index, and adjusting the sample rates for better allocation. The latter suggestions would reduce the sample size. It is suggested to replace the current index, which based on GDP, with one that is based on personal income. Making these changes would better reflect both inflation and real income growth. Research will need to be completed before implementing these changes to ensure no major negative impact will occur to the sample.

5. Conclusions

The Statistics of Income (SOI) sample is designed to support estimation for Individual tax return information. Data is collected through the IRS Individual Master File (IMF) as tax returns are filed and entered into the IRS system. Returns are sampled through daily processing cycles throughout the year. The cross-sectional sample is selected using a stratified Bernoulli sampling design. SOI's Individual Tax Return Sample has a wide variety of uses, and needs to adapt to suit all customers' needs. The basic design of the sample still meets the customers' needs in an efficient manner. With some changes that have recently been made to the design, and some that will still be implemented in the near future, this will ensure better accuracy and effectiveness for the customers' changing needs.

References

Bryant, V. (2012), "General Description Booklet for the 2008 Public Use Tax File." *Individual Statistics Branch, Statistics of Income Division IRS*.

Czajka, J.L., Amang, S., and Brendan, K. (2014), "An Assessment of the Need for a Redesign of the Statistics of Income Individual Tax Sample." Mathematica Policy Research.

Liu, Y. K., Auten, G., Testa, V.L., and Strudler, M. (2009), "Redesign of SOI's Individual Tax Return Edited Panel Sample." Proceedings of the Joint Statistical Meetings, Section on Government Statistics, American Statistical Association, 3129-3143.

U.S. Bureau of Economic analysis, "Price Indexes for Gross Domestic Product," [http://www.bea.gov/] (accessed December 5, 2013).

Weber, M. (2001), "The Statistics of Income 1979-2002 Continuous Work History Sample Individual Income Tax Return Panel," Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association.

Appendix

Below are the tables that are referenced throughout the paper that are used in defining the sample code.

Table 1: Return Type Codes

Return Type		
OLD	**NEW**	**Definition**
1	1	High Income Nontaxable
2	2	Business High Receipts
3	3	Foreign Earned Income (Form 2555)
4	4	Foreign Tax Credit, Farm or Business (Form 1116 and Schedule C or F)
5	4	Foreign Tax Credit, Nonbusiness, Nonfarm (Form 1116)
6	5	Business and Farm (Schedule C and F)
7	5	Nonfarm Business (Schedule C)
8	6	Nonbusiness Farm (Schedule F)
0	0	Nonbusiness, Nonfarm

Table 2: Income Class Codes

Income Class			
Negative Income			
OLD	**NEW**	**Boundary**	**TTIN Cutoff**
01	01	$10,000,000 or more	99,999
02	02	$5,000,000 under $10,000,000	99,999
03	03	$2,000,000 under $5,000,000	33,999
04	04	$1,000,000 under $2,000,000	15,999
05	05	$500,000 under $1,000,000	3,309
06	06	$250,000 under $500,000	894
07	07	$120,000 under $250,000	413
08	08	$60,000 under $120,000	211
09	09	under $60,000	86
Positive Income			
OLD	**NEW**	**Boundary**	**TTIN Cutoff**
10-12	10	under $30,000	CWHS only
13-14	11	$30,000 under $60,000	CWHS only
15-16	12	$60,000 under $120,000	CWHS only
17-18	13	$120,000 under $250,000	234
19	14	$250,000 under $500,000	619
20	15	$500,000 under $1,000,000	2,379
21	16	$1,000,000 under $2,000,000	12,099
22	17	$2,000,000 under $5,000,000	32,399
23	18	$5,000,000 under $10,000,000	99,999
24	19	$10,000,000 or more	99,999