Statistics of Income Sales of Capital Assets Sample Redesign for Tax Year 2007

by Yan K. Liu, Jana Scali, Michael Strudler, and Janette Wilson, Internal Revenue Service

The Statistics of Income (SOI) of the IRS selects a cross-sectional stratified random sample of individual returns from the population of all U.S. individual tax returns filed with the IRS every year. This yearly sample is used for various studies, including the study of Form 1040 items of Sales of Capital Assets (SOCA), such as the total amount of Sales Price, Basis, and Net Gain/Loss. However, the individual return sample provides SOCA data only at the return level, not at the capital asset transaction level because of the high processing cost associated with editing finer-level data.

To study SOCA at the transaction level, a smaller representative sample was selected from the Tax Year 1999 individual return sample, called the SOCA Cross-Sectional Sample. The same sample design as the 1999 individual return sample was used, and weights were adjusted accordingly. Further, from this 1999 cross-sectional SOCA sample file, a subsample was selected to serve as the base year for a SOCA panel sample, in which returns were followed in subsequent tax years.

The SOCA panel is also a stratified random sample, but the stratum definition is different from that of the SOCA cross-sectional sample and individual return sample. Due to various resource and planning constraints, no refreshment sample has been added to this panel sample since that tax year. Subsequently, the SOCA panel sample has drifted and is no longer representative of the current-year population. Also, 1999 was the last year that SOI had a SOCA cross-sectional file. Therefore, a new cross-sectional SOCA sample is needed for Tax Year 2007 and a new panel sample will be developed from it.

Since there is a close relationship among the individual return sample, SOCA cross-sectional sample, and SOCA panel sample, it is important to understand how these samples are related. In Tax Year 1999, the individual return sample of 176,966 returns was drawn from the population of 127,321,626 returns; the SOCA cross-sectional sample was a subsample of 121,053 returns of the 176,966 individual sample returns; and the SOCA panel sample of 83,432 returns was a subsample of the SOCA cross-sectional sample. The stratum boundaries of the SOCA cross-sectional sample followed the same boundaries used in the individual sample, but the SOCA panel sample used different stratum boundaries. Details are given below.

The individual return sample is a stratified random sample (Testa and Scali, 2005). The stratification is achieved by the return type code, as shown in Table 1, and income code, as shown in Table 2. The income code is determined by the income classification and the "degree of interest" for the modeling purpose. It is a four-level categorical variable where "1" is assigned to returns that are least interesting and "4" to those most interesting. The final stratification is achieved by combining return type code and income code, as summarized in Table 3.

Each sample code identifies a stratum. As shown in Table 3, returns with a return type code of 1 or 2 indicate returns with high nontaxable income or large business receipts respectively sampled with certainty, regardless of the income amount. The rest of the returns are divided into 24 income classes within each tax return type.

The sample consists of two parts: a Bernoulli sample and a CWHS (Continuous Work History Sample) (Weber, 2001). A Bernoulli sample is selected independently from each sample code, with rates ranging from 0.1 percent to 100 percent. The sample selection utilizes a permanent random number that is an integer function of the primary taxpayer's Social Security Number, called the Transformed Taxpayer Identification Number (TTIN). The last five digits of the TTIN is a pseudorandom number. A return for which the pseudorandom number is less than the sampling rate multiplied by 100,000 is selected in the sample.

This paper was originally presented at the 2008 June Statistical Meeting of the American Statistical Association in Denver, CO.

The selection criteria, which are given in Table 4, show that a same sampling rate is used for sample codes with the same income code except for sample code 101-124 and 201-204 in which all returns are taken with certainty. For example, a sampling rate of 33.4 percent is used for sample codes 003, 303, 403, 503, 603, 703 and 803. In other words, population returns with the last five digits of the TTIN smaller than 33,400 in those sample codes are selected. In addition to returns selected using the pseudorandom number, returns having one of the specific final four digits in the taxpayer's SSN are also selected.

Returns that have one of the specific final four digits in the taxpayer's SSN form a special subsample, called the Continuous Work History Sample (CWHS).[1] Before 2005, there were five specific final four digits used for CWHS, which represent 23 percent of the individual return sample. Starting from 2005, ten specific final four digits have been used, which represent 46 percent of the individual return sample. Note that some returns selected by TTIN may also be part of the CWHS.

The 1999 SOCA cross-sectional sample was a subsample selected from the 1999 individual return sample

Table 1. Return Type Code

Return Type Code	Special Category
1	High-Income Nontaxable Returns
2	Large Business Receipts
3	Form 2555 (Foreign Earned Income)
4	Form 1116 & Schedule C or F
5	Form 1116 (Foreign Tax Credit)
6	Schedule C & Schedule F
7	Schedule C (Nonfarm Sole Proprietors)
8	Schedule F (Farm Sole Proprietors)
0	All Others

Table 2. Income Code

Income Code	Income Range	Degree of Interest
	NEGATIVE INCOME	
1	$10,000,000 or more	All
2	$5,000,000 - under $10,000,000	All
3	$2,000,000 - under $5,000,000	All
4	$1,000,000 - under $2,000,000	All
5	$500,000 - under $1,000,000	All
6	$250,000 - under $500,000	All
7	$120,000 - under $250,000	All
8	$60,000 - under $120,000	All
9	Under $60,000	All
	POSITIVE INCOME	
10	Under $30,000	1
11	Under $30,000	2
12	Under $30,000	3-4
13	$30,000 - under $60,000	1-2
14	$30,000 - under $60,000	3-4
15	$60,000 - under $120,000	1-3
16	$60,000 - under $120,000	4
17	$120,000 - under $250,000	1-3
18	$120,000 - under $250,000	4
19	$250,000 - under $500,000	All
20	$500,000 - under $1,000,000	All
21	$1,000,000 - under $2,000,000	All
22	$2,000,000 - under $5,000,000	All
23	$5,000,000 - under $10,000,000	All
24	$10,000,000 or more	All

[1] CWHS returns are considered as randomly selected since the SSN endings are approximately random.

Table 3. Sample Code (Stratum)

Sample Code	Return Type Code	Income Code	# Of Strata
101-124	1	all	1
201-224	2	all	1
301-324	3	1-24	24
401-424	4	1-24	24
501-524	5	1-24	24
601-624	6	1-1	24
701-724	7	1-24	24
801-824	8	1-24	24
001-024	0	1-24	24

Table 4. 2005 Individual Return Sample Random Selection Criteria

Income Code	Sample Code	Cutoff of the Last Five Digits of TTIN*
All	101 – 124	All
All	201 – 224	All
1	301 – 801	All
2	302 – 802	All
3	303 - 803	33,399
4	304 – 804	15,999
5	305 - 805	3,309
6	306 - 806	894
7	307 – 807	413
8	308 – 808	211
9	309 – 809	86
10	310 – 810	0
11	311 – 811	0
12	312 - 812	53
13	313 – 813	0
14	314 – 814	57
15	315 – 815	0
16	316 - 816	50
17	317 – 817	95
18	318 – 818	234
19	319 – 819	619
20	320 – 820	2,379
21	321 – 821	12,099
22	322 - 822	32,399
23	323 – 823	All
24	324 – 824	All

**Sampling rate = last five-digit /100,000. A '0' Cutoff means no return is selected by TTIN and only the CWHS returns are included.*

using the same stratum boundaries with smaller sampling rates in some strata.

The 1999 SOCA panel sample was a subsample selected from the 1999 SOCA cross-sectional sample.[2] However, the strata were defined differently in that the return type was not used and only income code was used. For example, strata 003, 103, 203, 303, 403, 503, 603, 703 and 803 are pooled into one stratum that have the same income code of "03." Further, strata with income codes "01" and "24" were broken into two each by the income amount, as shown in Table 5. The panel sample includes all returns that were randomly selected using the pseudorandom number and additional returns containing any of five CWHS ending digits. The approximate sampling rates and TTIN cutoffs are also given in Table 5.

▶ Designing the 2007 Cross-Sectional SOCA Sample

The 2007 SOCA cross-sectional sample is a subsample of the individual return sample and should include the 2007 SOCA panel sample, which will serve as the base-year panel sample for coming years. It was decided that the 2007 panel sample will have the same stratum boundaries as the 1999 SOCA panel design and include at least returns selected using the criteria of the 1999 SOCA panel design, defined in Table 5. In other words, it should start with at least returns satisfying the

Table 5. 1999 SOCA Panel Sample Design

SOCA Panel Stratum ID	Income Code	Specified Sampling Rate (%)*	Cutoff of the Last Five Digits of TTIN
0	01 (income>=20,000,000)	100	ALL
1	01 (income<20,000,000)	48.47	48,444
2	2	22.05	22,011
3	3	4.2	4,152
4	4	1.42	1,371
5	5	0.58	530
6	6	0.12	70
7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18	07, 08, 09, 10, 11, 12, 13, 14, 15, 16, 17, 18	0.05	0
19	19	0.18	130
20	20	0.59	540
21	21	1.72	1,671
22	22	5.73	5,683
23	23	18.88	18,839
24	24 (income<20,000,000)	57.62	57,599
25	24 (income>=20,000,000)	100	ALL

**Including CWHS returns*

[2] The 1999 panel sample was designed to represent all Tax Year 1999 returns, including late returns, while the 1999 individual return sample and 1999 SOCA cross-sectional sample were designed to represent all returns filed in Calendar Year 2000. Therefore, the 1999 panel sample was drawn from the 1999 SOCA cross-sectional sample and supplemented with the 2000 and 2001 individual return samples in order to include returns that were filed up to 2 years late.

same selection criteria in Table 5 and add more returns to some strata, as appropriate. This is because we want to have a 2007 SOCA panel at least as large as the 1999 SOCA panel in each stratum.

In designing the 2007 SOCA cross-sectional sample, we needed to determine the stratum boundaries and sample size allocation across strata. In terms of stratum boundaries, we employed the same boundaries as for the 1999 SOCA panel sample, instead of using the same stratum boundaries of the individual return sample, for two reasons: (1) the return type (Table 1) is not considered to be related to the SOCA analysis; and (2) it is consistent with the new panel sample design. In terms of sample size allocation, we made use of the available information from Tax Year 2005 data to balance the variance and the processing cost. Details are given below.

To determine the final sample size allocation of the 2007 SOCA cross-sectional sample, we used the variance information from the most recent available 2005 individual return sample data and processing cost information from the most recent panel sample data for Tax Year 2005.[3] Although the SOCA file is used to mainly estimate the totals of some variables by asset type, it is impossible to have a sample that is optimum for each of the 22 asset types. Thus, our design target was based on the precision levels of estimates for the totals of the three key variables: Sales Price (E21550), Net Short Term Gain/Loss (E22250), and Net Long Term Gain/Loss (E23250).

We first calculated the optimum sample size allocation using Neyman allocation (Cochran, 1977), then adjusted the sample sizes for some constraints on the lower and upper bounds. Therefore, the final stratum sample sizes were not strictly obtained by Neyman optimum allocation. Instead, Neyman allocation was used as a starting step of the sample size allocation process.

For a given sample size n, the sample size proportion for stratum h by Neyman optimum allocation is:

$$p_h = \frac{n_h}{n} = \frac{N_h S_h \big/ \sqrt{c_h}}{\sum_h N_h S_h \big/ \sqrt{c_h}}, \qquad (1)$$

where N_h, S_h, and c_h are the population size, standard deviation, and cost per return for stratum h; and n_h/n is the sample size allocation across strata. The population size N_h is known.

To use the Neyman allocation equation (1), we need information for S_h and c_h. Here, c_h is the average cost for SOI to edit each return because the individual return sample consists of both SOCA returns and non-SOCA returns. Thus, for our design purpose, the processing cost of non-SOCA returns was treated as zero, and the processing cost per SOCA return was from the 2005 panel sample.

The average cost per return c_h was obtained by multiplying the processing cost per SOCA return that was obtained from the 2005 panel sample and the percentage of SOCA returns that was calculated from the 2005 individual return sample. The reason that processing cost per SOCA return was obtained from the 2005 panel data and not from the individual return sample was that returns are processed at the tax form line level for the individual return sample, while returns are processed at the transaction level for the SOCA cross-sectional sample.[4] For example, if a taxpayer had 100 different short-term stock transactions, SOI would edit only the total sales and total net income/loss from the combination of these transactions for the individual file. However, for SOCA, each sale would be processed separately. The resulting cost information is given in Table 6.

Because of the relatively low cost for returns in strata 10-17, it was decided to include all sampled individual returns in the SOCA cross-sectional sample. Strata 0, 1, 24, and 25 are certainty strata, and all their returns are taken in the SOCA cross-sectional sample

[3] The 1999 SOCA panel sample was followed in each tax year. The most recent year was 2005.
[4] The last SOCA cross-sectional sample was done in 1999. The most recent cost estimates at the transaction level are from the 2005 panel sample. Therefore, cost information from the 2005 panel sample was used.

Table 6. Processing Cost Per Return by Stratum

Stratum h	2005 Population Size N_h	2005 Individual Return Sample		Cost Per SOCA return (Minutes)	Average Cost Per Individual Return (Minutes) c_h
		Sample Size	% SOCA returns		
0	850	850	82.60%	86.3	71.299
1	1,019	1,019	95.20%	86.3	82.18
2	2,865	2,865	92.60%	81.9	75.903
3	11,583	3,921	92.90%	79.9	74.283
4	24,668	4,051	91.10%	54.5	49.618
5	62,671	2,322	89.80%	46.1	41.369
6	145,074	1,684	87.50%	26.6	23.27
7	304,998	1,700	81.50%	24.9	20.308
8	426,292	1,362	73.00%	17.9	13.079
9	1,394,836	2,700	59.20%	25.3	14.979
10	30,444,834	30,396	0.10%	2.9	0.003
11	28,944,931	28,868	5.70%	1.9	0.106
12	10,232,344	15,703	19.70%	3.3	0.649
13	23,743,039	23,823	11.30%	2.6	0.296
14	10,255,177	16,198	26.20%	2.7	0.704
15	13,842,711	13,790	27.00%	3.8	1.032
16	6,346,609	9,607	42.70%	3.6	1.515
17	1,746,471	5,880	53.70%	3.5	1.873
18	4,089,699	16,085	61.10%	22.9	13.976
19	1,628,792	15,441	73.90%	21.3	15.736
20	551,000	15,084	83.70%	27.8	23.296
21	185,095	23,086	90.80%	47.9	43.454
22	78,029	25,543	94.70%	72	68.226
23	19,107	19,107	97.20%	105.7	102.666
24	7,572	7,572	98.20%	120.8	118.611

as well. For the rest of the strata, the standard deviation S_h was calculated.

In calculating S_h, some returns are excluded so that all returns used in the standard deviation calculation have the same weights. These excluded returns are from sample codes 101–124 and 201-224 and would not have been selected if using the selection criteria of other sample codes (see Table 4). For example, stratum 3 consists of returns from sample codes 003, 103, 203, 303–803, and CWHS returns. All returns from sample codes 103 and 203 were selected, and only returns with a TTIN smaller than 33,400 were selected for sample codes for 003 and 303–803 (see Table 4). Therefore, non-CWHS returns that had a TTIN greater than 33,399 were excluded. Basically, these excluded returns were from sample codes 103 and 203 and have zero probability of being selected in the SOCA cross-sectional sample.

Further, for the 13 returns that had an original income code different from the edited income code and could have large impact on the variance, their strata were adjusted using the edited income code (instead of the original amount used for stratification). For example, one return had the original income code of "03" and the edited income code of "01" because the edited income was larger than $10,000,000. Leaving it to its original income code would inflate the standard deviation of stratum 3. Therefore, it was moved to stratum 1. The standard deviation S_h of each key variable was calculated using return-level data where a non-SOCA return was assigned a value of zero. Table 7 gives the standard deviation estimates for three key variables.

We then calculated sample size allocation percentages across strata using Neyman optimum allocation (1) for each of the three key variables, denoted as p_{h1}, p_{h2}, and p_{h3} for each stratum h and then taking the average of the three. That is, for a given sample size n, the stratum sample size is $n_h = n(p_{h1} + p_{h2} + p_{h3})/3$. The sample size n_h was further adjusted by lower end L_h and upper end U_h, i.e., $L_h \leq n_h \leq U_h$ for all h. The lower end L_h was decided by the selection criteria of 1999 panel sample (Table 5) to ensure that the new panel sample was a subsample of the 2007 SOCA cross-sectional sample and, thus, satisfied at least the selection criteria of the 1999 panel. The upper end U_h was the stratum sample size of the individual return sample after removing the excluded returns because the SOCA cross-sectional sample will be selected from the individual return sample.[5] Therefore, if the calculated n_h was smaller than L_h, it was forced to be equal to L_h; if the calculated n_h was larger than U_h, it was reduced to be the same as U_h.

After evaluating some options of sample size and processing cost, the final choice is summarized in Table 8. Based on the 2005 population, the projected cost and Coefficient of Variation (CV) for the three key variables are given in Table 9. Here, the extra cost is

Table 7. Data Summary for Sample Size Allocation Summary

Stratum	Population Size	Standard Deviation S_h			Average Cost Per Return c_h	Sample Size Low End L_h	Sample Size High End U_h
		Sales Price (E21550)	Net Short-Term Gain or Loss (E22250)	Net Long-Term Gain or Loss (E23250)			
2	2,865	46,422,576	2,910,246	10,828,250	75.903	613	2,865
3	11,583	23,511,291	1,271,933	1,700,676	74.283	496	3,808
4	24,668	16,613,886	603,837	720,621	49.618	357	3,890
5	62,671	22,123,386	314,883	360,997	41.369	372	2,038
6	145,074	5,208,675	156,964	181,711	23.27	244	1,395
7	304,998	2,612,603	74,011	89,332	20.308	306	1,577
8	426,292	2,191,341	33,792	48,225	13.079	444	1,333
9	1,394,836	1,067,603	9,776	18,056	14.979	1,441	2,676
18	4,089,699	761,388	16,084	46,303	13.976	3,946	13,581
19	1,628,792	2,029,207	30,672	105,765	15.736	3,743	11,683
20	551,000	3,221,822	68,763	243,875	23.296	3,544	13,670
21	185,095	6,217,282	146,646	565,720	43.454	3,324	22,558
22	78,029	11,987,771	329,655	1,414,260	68.226	4,632	25,325
23	19,107	16,748,060	784,437	4,461,375	102.666	3,600	19,107

[5] The excluded returns are from sample codes 101–124 and 201-224 and would not have been selected if using the selection criteria of other sample codes.

the total cost, excluding the cost for returns that also fall in the 1999 panel sample. Also note that CVs here are for estimates of the overall totals. However, the SOCA estimates are also broken by asset type, which can result in much higher CVs for some asset types. Finally, Table 10 gives cost estimates by Electronic Filing Status and Service Center, which were used for budget allocation purposes.

▶ References

Cochran, W. G. (1977), Sampling Techniques, Wiley.

Testa, V. and Scali, J (2005), "Description of the Sample," *Statistics of Income–2005, Individual Income Tax Returns, Internal Revenue Service, Washington, DC.*

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

Table 8. Selection Criteria of 2007 SOCA Cross-Sectional Sample

Stratum	Selection Criteria		Overall Sampling Proportion (Random selection and CWHSI) (%)	Based on 2005 Population**	
	Cutoff of random selection (TTIN)	CWHSI*		Sample Size	# SOCA Returns
0	99999	1, 2	100	850	702
1	99999	1, 2	100	1019	970
2	74808	1, 2	74.83	2144	1,995
3	23337	1, 2	23.41	2712	2,513
4	14459	1, 2	14.55	3588	3,258
5	3155	1, 2	3.25	2038	1,805
6	862	1, 2	0.96	1395	1,188
7	417	1, 2	0.52	1577	1,268
8	213	1, 2	0.31	1333	967
9	92	1, 2	0.19	2676	1,578
10	0	1, 2	0.1	30396	34
11	0	1, 2	0.1	28832	1,628
12	53	1, 2	0.15	15660	3,050
13	0	1, 2	0.1	23811	2,685
14	58	1, 2	0.16	16151	4,196
15	0	1, 2	0.1	13774	3,713
16	51	1, 2	0.15	9560	4,065
17	100	1, 2	0.2	3490	1,682
18	232	1, 2	0.33	13581	8,212
19	618	1, 2	0.72	11683	8,625
20	2383	1, 2	2.48	13670	11,456
21	5970	1, 2	6.06	11224	10,187
22	10887	1, 2	10.98	8565	8,077
23	22411	1, 2	22.49	4297	4,183
24	99999	1, 2	100	7572	7,432
25	99999	1, 2	100	4180	4,104

**CWHSI is the indicator for CWHS status. A return with a CWHSI value of 1 and 2 falls in the 10 CWHS endings.*

***The sample size and the number of SOCA returns based on 2007 population are expected to be larger.*

Table 9. The Projected Cost and CV from the 2007 SOCA Cross-Sectional Sample

Sample Size (# returns)	Total Cost (years)	Extra Cost (years)	CV		
			Sales Price (E21550)	Net Short-Term Gain or Loss (E22250)	Net Long-Term Gain or Loss (E23250)
235,778	36.35	28.85	4.72%	-1.76%	0.89%

Table 10. Cost Estimate by Electronic Filing Status and Service Center for the 2007 SOCA Cross-Sectional Sample (Projection Based on 2005 Population)

Electronic Filing	Service Center	Number of Returns	Number of SOCA Returns	Total cost (Years)	Extra cost (Years)
No	Atlanta (7)	24,133	13,573	5.54	4.46
No	Andover (8)	19,380	11,267	5.18	3.83
No	Kansas City (9)	24,404	12,057	4.61	3.58
No	Cincinnati (17)	29	29	0.03	0.01
No	Austin (18)	21,151	10,988	4.31	3.51
No	Philadelphia (28)	11,929	5,720	2.14	1.69
No	Fresno (89)	31,125	16,787	7.31	5.7
Subtotal		**132,151**	**70,421**	**29.12**	**22.77**
Yes	Andover (8)	24,067	7,445	1.96	1.65
Yes	Kansas City (9)	19,183	5,256	1.12	0.93
Yes	Cincinnati (17)	1	1	0	0
Yes	Austin (18)	19,740	4,275	0.87	0.73
Yes	Philadelphia (28)	16,054	3,149	0.66	0.57
Yes	Fresno (89)	24,582	9,026	2.62	2.2
Subtotal		**103,627**	**29,152**	**7.23**	**6.07**