Estimating the Variance of Between-Year Change in Domain-Level Totals

Kimberly Henry¹, Valerie Testa¹, and Richard Valliant²
¹Statistics of Income, P.O. Box 2608, Washington DC 20013-2608
²University of Michigan, 1218 Lefrak Hall, College Park MD 20742

Abstract: This paper provides the theory for estimating the variance of the difference in two years’ domain-level totals under the stratified Bernoulli sample design. Henry et al. (2008) developed an approximately design-unbiased variance estimator that used poststratification to correct for the random sample sizes created under Bernoulli sampling. We modify the Henry et al. (2008) variance estimator for the estimated change in domain-level totals. We consider both “planned domains,” domains that are related to the sample design variables, and “analysis domains,” unplanned domains of interest at the analysis stage. Our variance estimator takes into account three practical problems: a large overlap of units between two years’ samples, changing compositions of units across years that produce “stratum jumpers,” which are population and sample units that shift across strata from one year to another (Rivest, 1999), and changes in sampling rates across years. These problems affect estimating the covariance term in the variance of the difference. The variance estimator is applied to data from the Statistics of Income Division’s individual income tax return sample. Naïve variance estimates using only the separate years’ variances are compared to show the effect of ignoring the estimated covariance.

Key Words: Horvitz-Thompson estimator, stratified Bernoulli sampling, poststratification, Taylor series approximation

1. Introduction and Universe of Tax Returns for Two Years

Henry et al. (2008) provided the theoretical background to produce variance estimates of year-to-year changes between totals estimated from the Statistics of Income (SOI) Division’s Individual Tax Return sample, a stratified Bernoulli sample. We extend their theory, which is also discussed in Berger (2004), Nordberg (2000), and Wood (2008), to estimate the variance of estimated between-year change in domain-level totals. We consider this variance estimation for both “planned domains,” domains that are related to the sample design variables, and “analysis domains,” unplanned domains of interest at the analysis stage.

Henry et al. (2008) demonstrated that the post-stratification (PS) estimator produced single-year totals with lower variances than the Horvitz-Thompson estimator, the extent to which depended on the mean of the underlying variable of interest. SOI also uses the PS estimator to estimate yearly totals, so we restrict ourselves to the PS estimator (see Exp. 3.2.5 in Särndal et al., 1992). Suppose that the strata are ordered by increasing size of the sampling rate, i.e., the sampling rate for stratum 2 is greater than or equal to the rate for stratum 1, and so on. The PS estimator of year-to-year differences is affected by the location of sample units within strata in both years, so we define:

- \( U_{h0} \) = returns in stratum \( h \) that file only at time \( t_1 \) (deaths after time \( t_1 \) and before time \( t_2 \))
- \( U_{0h2} \) = returns in stratum \( h \) that file only at time \( t_2 \) (births after time \( t_1 \) and before time \( t_2 \))
- \( U_{h2} \) = returns in stratum \( h \) at time \( t_1 \) and stratum \( h_2 \) at time \( t_2 \) that file returns at both times, for \( h_1 < h_2 \) (units that move to strata with a higher sampling rate in year 2), \( h_1 = h_2 \) (units that stay in the same strata), or \( h_1 > h_2 \) (units moving to strata with lower sampling rates in year 2).

Using this notation, the two universes can be partitioned into a 2-way grid based on stratum membership at times \( t_1 \) and \( t_2 \), shown in Table 1. For sample selection purposes, the stratum \( h_1 \) and \( h_2 \) universes at times \( t_1 \) and \( t_2 \) are the union of all units (here tax returns) down column \( h_2 \) and across row \( h_1 \), respectively:
2. SOI Sample Design

The stratified Bernoulli sample design is used by most of SOI’s cross-sectional studies (IRS Winter 2008). In each study’s frame population, every unit has a unique identifier - the Social Security Number (SSN) of the primary tax filer in the Individual study and the Employer Identification Number for Corporate and Tax Exempt organizations’ tax returns. Each return’s unique identifier is used to produce a permanent random number (PRN) between 0 and 1, denoted \( r_i \). For a given year, unit \( i \) is selected for a sample if

\[
 r_i < \pi_h ,
\]  

(2.1)

where \( \pi_h \) is the pre-assigned sampling rate for stratum \( h \) that tax return \( i \) belongs to. SOI’s Individual sample consists of two parts within each stratum. First, a 0.05 percent stratified Bernoulli sample of approximately 65,000 returns is selected, called the Continuous Work History Sample (CWHS, Weber 2004). A separate Bernoulli sample is also selected independently from each stratum, with rates ranging from 0.01 to 100 percent (see Testa and Scali (2006) for details). The full sample, which itself is a Bernoulli sample, consists of the CWHS plus all additional returns selected with unequal probabilities of selection across strata. For Tax Year (TY) 2004, 200,778 returns were selected from 133,189,982. For TY 2005, the CWHS sampling rates were increased such that 292,966 returns were selected from 134,494,440. These years correspond to taxpayers’ income earned during the previous calendar year (e.g., TY 2004 represents income earned in 2004 and reported to the IRS by December 2005).

Every year, using condition (2.1) for every tax return automatically accounts for births, deaths, and the stratum jumpers in the population as follows:

- **Births**: each birth is independently assigned a PRN; if (2.1) holds, then the unit is selected for the sample. There were 19,999,605 of these returns entered the population between 2004 and 2005.
- **Deaths**: units are not present in the population file, so they are not in the sample. There were 18,695,168 of these returns departing the population between 2004 and 2005.
- **Stratum jumpers**: if, from year \( t_1 \) to \( t_2 \), a return switches from stratum \( h_1 \) to stratum \( h_2 \), then the return is in the sample in both years if \( r_i < \min(\pi_{h_1}, \pi_{h_2}) \) (i.e., if the PRN is less than the rates for both strata). There were very few (less than ten) of these returns between 2004 and 2005.

This sample selection method also ensures a large overlap between two years, since a unit is selected in both years if \( r_i \leq \pi_{h_1} \leq \pi_{h_2} \). The rotating PRN methods used in Berger (2004) and Nordberg (2000) to reduce the number of overlapping units across years due to respondent burden are not required for tax returns since the associated taxpayers are not contacted by SOI. This overlap of units across different year’s samples creates a large covariance term that must be accounted for in variance estimation of the difference between two years’ estimates. There are additional sample selection issues due to changes in population units that affect the covariance term. We use the following rules in the covariance estimation:

- **Marriages**: two “single” returns (filing either as single or married separate) in year \( t_1 \) that file as a joint married return are considered two deaths in year \( t_1 \) and a birth in year \( t_2 \).
- **Divorces**: a married joint return that becomes two single entities is considered a death in year \( t_1 \) and two births in year \( t_2 \).
- **SSN swapping**: joint married tax returns in both years are tracked and considered the same unit in both years.
Sample design changes can also result in sampling rate changes between years; our estimators account for such changes.

3. Estimators for Totals and Their Change

A Bernoulli sample is selected within each stratum as described in Section 2, where \( \pi_h \), the stratum sampling rate in a given year, is also the probability of selection for all units in stratum \( h \). Denote the sample inclusion indicators for unit \( i \) at times \( t_1 \) and \( t_2 \) by:

\[
\begin{align*}
\delta_i(t_1) &= \begin{cases} 
1 & \text{if unit } i \in s_1 \\
0 & \text{otherwise}
\end{cases} \\
\delta_i(t_2) &= \begin{cases} 
1 & \text{if unit } i \in s_2 \\
0 & \text{otherwise}
\end{cases}
\end{align*}
\]

From these expressions, the conditional and unconditional probabilities of selection by population domain can be derived. For Bernoulli sampling, the expected values and variances of the inclusion indicator for each year are

\[
\begin{align*}
E[\delta_i(t_1)] &= \pi_h, \\
Var[\delta_i(t_1)] &= \pi_h(1 - \pi_h) \\
E[\delta_i(t_2)] &= \pi_h, \\
Var[\delta_i(t_2)] &= \pi_h(1 - \pi_h).
\end{align*}
\]

Since \( \delta_i(t_2)\delta_i(t_1) = 1 \) only when a unit is in the sample for both time periods, for \( E[\delta_i(t_2)\delta_i(t_1)] - E[\delta_i(t_2)]E[\delta_i(t_1)] \), the covariance for the indicator variable for unit \( i \) in stratum \( h_1 \) at time \( t_1 \) and in stratum \( h_2 \) at time \( t_2 \) is given by:

\[
Cov[\delta_i(t_2), \delta_i(t_1)] = \min(\pi_{h_1}, \pi_{h_2}) - \pi_{h_1}\pi_{h_2} = \Delta_{h_1h_2}.
\]

The finite population totals of a study variable of interest \( y \) at times \( t_1 \) and \( t_2 \) are denoted by

\[
T(t_1) = \sum_{h_1=1}^{H_1} \sum_{i \in U_{h_1}} y_{1i}, \quad T(t_2) = \sum_{h_2=1}^{H_2} \sum_{i \in U_{h_2}} y_{2i},
\]

(3.1)

where \( y_{1i} \) and \( y_{2i} \) are the \( y \)-values (for the same variable of interest) for unit \( i \) at times \( t_1 \) and \( t_2 \).

Holt and Smith (1979) observed that, for estimation from completed samples, conditioning on an achieved post stratum sample size, as in (4.1) and (4.2), is inferentially more appropriate than averaging over all possible sample sizes. SOI uses a poststratified (PS) estimator that conditions on the number of achieved units in each stratum. This estimator, which is conditionally unbiased for the population total (Brewer et al. 1972), reduces the variability caused by the random stratum sample sizes and leads to formulae simplifications. First, the observed number of sample returns in stratum \( h \) from year \( t_1 \) is denoted by \( n_{h} \), Assuming that \( n_{h} > 0 \), it can be shown that conditional on the sample design at time \( t_1 \) is a stratified simple random sample with stratum sample sizes \( n_{1}, n_{2}, \ldots, n_{H} \), the sample design at time \( t_1 \) is a stratified simple random sample with stratum sample sizes \( n_{1}, n_{2}, \ldots, n_{H} \). Thus, for \( N_{h} \) denoting the number of population units in stratum \( h_1 \) at time \( t_1 \), the PS estimator for \( T(t_1) \) is

\[
\hat{T}(t_1|n_1) = \sum_{h_1=1}^{H_1} \frac{N_{h_1}}{N_{h_1}} \sum_{i \in U_{n_1}} \delta_i(t_1)y_{1i}.
\]

(3.2)
Similarly, for year 2, we assume that 
\( n_{h_2} = \sum_{i \in U_{h_2}} \delta_i(t_2) > 0 \) and conditional on \( n_2 = \{n_{h_1}, n_{h_2}, \ldots, n_{h_k}\} \), the time \( t_2 \) sample design is a stratified simple random sample. For \( N_{h_2} \) being the number of population elements in stratum \( h_2 \) at time \( t_2 \), the PS estimator for \( T(t_2) \) is

\[
\hat{T}(t_2|n_2) = \sum_{h_2=1}^{H_2} \frac{N_{h_2}}{n_{h_2}} \sum_{i \in U_{h_2}} \delta_i(t_2) y_{2i}.
\]

SOI uses \( \hat{T}(t_1|n_1) \) and \( \hat{T}(t_2|n_2) \) to estimate time-specific totals, which are also special forms of the PS estimator, where the poststrata are the same as the design strata.

The finite population change in level between two time points is denoted by

\[
D = T(t_2) - T(t_1).
\]

The PS estimators of time-specific totals in (3.2) and (3.3) lead to the following estimator of the (3.4) difference:

\[
\hat{D} = \hat{T}(t_2|n_2) - \hat{T}(t_1|n_1),
\]

which is conditionally unbiased for the change in level. By breaking \( \hat{T}(t_1|n_1) \) into the sum of deaths for time \( t_1 \) and units in both years’ samples summed over the year 1 strata and \( \hat{T}(t_2|n_2) \) into the sum of the births for time \( t_2 \) and the units in both samples over the year 2 strata, expression (3.5) can be rewritten as

\[
\hat{D} = \sum_{h_2=1}^{H_2} \frac{N_{h_2}}{n_{h_2}} \sum_{i \in U_{h_2}} \delta_i(t_2) y_{2i} - \sum_{h_1=1}^{H_1} \frac{N_{h_1}}{n_{h_1}} \sum_{i \in U_{h_1}} \delta_i(t_1) y_{1i} \\
+ \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} \frac{N_{h_2}}{n_{h_2}} \sum_{i \in U_{h_2}} \left[ \frac{\delta_i(t_2) y_{2i}}{n_{h_2}} - \frac{N_{h_1} \delta_i(t_1) y_{1i}}{n_{h_1}} \right].
\]

The ratio and relative differences (relative to the year 1 total) of the two years’ differences are given by

\[
R = \frac{T(t_2)}{T(t_1)} \quad \text{and} \quad RD = \frac{T(t_2) - T(t_1)}{T(t_1)} ,
\]

which are estimated by

\[
\hat{R} = \frac{\hat{T}(t_2|n_2)}{\hat{T}(t_1|n_1)} \quad \text{and} \quad \hat{RD} = \frac{\hat{T}(t_2|n_2) - \hat{T}(t_1|n_1)}{\hat{T}(t_1|n_1)}.
\]

4. Theoretical Variances

The theoretical conditional variances of the PS estimators for both years are simply the variances of a total under stratified simple random sampling:

\[
Var \left[ \hat{T}(t_1|n_1) \right] = \sum_{h_1=1}^{H_1} \frac{N_{h_1}^2}{n_{h_1}} \left( 1 - \frac{N_{h_1}}{n_{h_1}} \right) S_{h_1}^2
\]

\[
Var \left[ \hat{T}(t_2|n_2) \right] = \sum_{h_2=1}^{H_2} \frac{N_{h_2}^2}{n_{h_2}} \left( 1 - \frac{n_{h_2}}{N_{h_2}} \right) S_{h_2}^2
\]
Using linear approximations to the PS estimators, the unconditional variance of the difference is

$$\text{Var} \left[ \hat{D} \right] = \sum_{h_2=1}^{H_2} 1 - \frac{\pi_{h_2}}{\pi_{h_2}} N_{\bullet h_2} S_{\bullet h_2}^2 + \sum_{h_1=1}^{H_1} 1 - \frac{\pi_{h_1}}{\pi_{h_1}} N_{h_1 \bullet} S_{h_1 \bullet}^2 - 2 \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} \frac{\Delta_{h_1 h_2} N_{h_1 h_2} S_{h_1 h_2}}{\pi_{h_1} \pi_{h_2}} \left( y_{2i} - \bar{y}_{\bullet h_2} \right) \left( y_{1i} - \bar{y}_{h_1 \bullet} \right),$$

(4.3)

where $S_{h_1 h_2} = \frac{1}{\left( N_{h_1 h_2} - 1 \right)} \sum_{i \in U_{h_1 h_2}} \left( y_{2i} - \bar{y}_{\bullet h_2} \right) \left( y_{1i} - \bar{y}_{h_1 \bullet} \right)$.

The variance in (4.3) can be expressed in a more standard form by converting some of the summations into stratum variances and covariances and approximating the sampling rates using the actual sample and population sizes achieved in each stratum. One approach is to substitute actual marginal sampling rates for terms like $1 - \pi_h$ and replacing $1/\pi_h$ by a stratum population size divided by the actual stratum sample size gives. Also, noting that $\frac{\Delta_{h_1 h_2}}{\pi_{h_1} \pi_{h_2}} = \frac{1}{\pi_{h_1} \pi_{h_2}} \left( 1 - \text{max} \left( \pi_{h_1}, \pi_{h_2} \right) \right)$ and replacing $\pi_{h_1}$ with $\frac{n_{h_1 \bullet}}{N_{h_1 \bullet}}$ and $\pi_{h_2}$ with $\frac{n_{h_2 \bullet}}{N_{h_2 \bullet}}$, we can obtain a covariance similar to the one in (4.3) that accounts for achieved sample sizes. This gives the following alternative variance of (3.6):

$$\text{Var} \left[ \hat{D} \mid n_1, n_2 \right] = \sum_{h_2=1}^{H_2} \left( 1 - \frac{n_{h_2 \bullet}}{N_{h_2 \bullet}} \right) N_{h_2 \bullet} S_{h_2 \bullet}^2 + \sum_{h_1=1}^{H_1} \left( 1 - \frac{n_{h_1 \bullet}}{N_{h_1 \bullet}} \right) N_{h_1 \bullet} S_{h_1 \bullet}^2 - 2 \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} \frac{\left[ \text{max} \left( \frac{n_{h_1 \bullet}}{N_{h_1 \bullet}}, \frac{n_{h_2 \bullet}}{N_{h_2 \bullet}} \right) \right]}{\text{max} \left( \frac{n_{h_1 \bullet}}{N_{h_1 \bullet}}, \frac{n_{h_2 \bullet}}{N_{h_2 \bullet}} \right)} N_{h_1 h_2} S_{h_1 h_2} \left[ \hat{y}_{\bullet h_2} - \bar{y}_{h_1 \bullet} \right] \left[ \bar{y}_{\bullet h_2} - \bar{y}_{h_1 \bullet} \right].$$

(4.4)

For the population ratio in (3.7), the first-order Taylor series approximation (as in Nordberg (2000) and Wood (2008)) is $\hat{R} - R = \left[ \hat{T}(t_2 \mid n_2) - R \hat{T}(t_1 \mid n_1) \right] / T(t_1 \mid n_1)$, which leads to the following approximate variance:

$$\text{Var} \left( \hat{R} \right) = \frac{\text{Var} \left( \hat{T}(t_2 \mid n_2) \right) + \hat{R}^2 \text{Var} \left( \hat{T}(t_1 \mid n_1) \right) - 2 \hat{R} \text{Cov} \left( \hat{T}(t_1), \hat{T}(t_2 \mid n_2) \right)}{\left( \hat{T}(t_1 \mid n_1) \right)^2},$$

(4.5)

where we have already derived the separate variance and covariance terms in (4.3) and (4.4). The relative difference has the same variance approximation, since we can write $RD = \left[ T(t_2) - T(t_1) \right] / T(t_1) = R - 1$. Thus, the variance is equivalent to (4.5).
5. Variance Estimators

Assuming that the counts \( N_{h_1}, N_{h_2}, \) and \( N_{h_2} \) are known, \( \bar{y}_{h_1} = \frac{1}{n_{h_1}} \sum_{i \in U_{h_1}} \delta_i(t_1) y_{1i} \) and \( \bar{y}_{h_2} = \frac{1}{n_{h_2}} \sum_{i \in U_{h_2}} \delta_i(t_2) y_{2i} \) are conditionally unbiased estimators of the strata population means. Since conditionally (and approximately unconditionally) unbiased estimators for the strata variances \( S_{h_1}^2 \) and \( S_{h_2}^2 \) are

\[
S_{h_1}^2 = \frac{1}{(n_{h_1} - 1)} \sum_{i \in U_{h_1}} \delta_i(t_1) (y_{1i} - \bar{y}_{h_1})^2
\]

\[
S_{h_2}^2 = \frac{1}{(n_{h_2} - 1)} \sum_{i \in U_{h_2}} \delta_i(t_2) (y_{2i} - \bar{y}_{h_2})^2
\]

the within-year variance estimators are the standard stratified simple random sampling variance estimators:

\[
\text{var}\left[ \hat{T}(t_1 | n_1) \right] = \sum_{h_1=1}^{H_1} \left( 1 - \frac{n_{h_1}}{N_{h_1}} \right) \frac{N_{h_1}^2}{N_{h_1} S_{h_1}^2}
\] (5.1)

\[
\text{var}\left[ \hat{T}(t_2 | n_2) \right] = \sum_{h_2=1}^{H_2} \left( 1 - \frac{n_{h_2}}{N_{h_2}} \right) \frac{N_{h_2}^2}{n_{h_2} S_{h_2}^2}
\] (5.2)

These are conditionally unbiased for (4.1) and (4.2). Using sample-based estimates for each (4.3) component, we have the approximate estimator of \( \text{Var} \left[ \hat{D} \right] \):

\[
\text{Var} \left[ \hat{D} \right] = \sum_{h_1=1}^{H_1} \frac{1 - \pi_{h_1} N_{h_1} S_{h_1}^2}{\pi_{h_1}^2} + \sum_{h_1=1}^{H_1} \frac{1 - \pi_{h_1} N_{h_1} S_{h_1}^2}{\pi_{h_1}^2} + \sum_{h_1=1}^{H_1} \frac{1 - \pi_{h_1} N_{h_1} S_{h_1}^2}{\pi_{h_1}^2} N_{h_1} S_{h_1}^2
\]

\[
-2 \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} \frac{\Delta_{h_1 h_2} N_{h_1} c_{h_1 h_2}}{\pi_{h_1} \pi_{h_2}}
\] (5.3)

where \( c_{h_1 h_2} = \frac{1}{(n_{h_1 h_2} - 1)} \sum_{i \in s_{h_1 h_2}} (y_{2i} - \bar{y}_{h_2})(y_{1i} - \bar{y}_{h_1}) \) is the unweighted covariance between the variable \( y \)-values for units in both years’ samples. An alternative variance estimator based on (4.4) is

\[
\text{var} \left[ \hat{D} | n_1, n_2 \right] = \sum_{h_2=1}^{H_2} \left( 1 - \frac{n_{h_2}}{N_{h_2}} \right) \frac{N_{h_2}^2}{n_{h_2} S_{h_2}^2} + \sum_{h_1=1}^{H_1} \left( 1 - \frac{n_{h_1}}{N_{h_1}} \right) \frac{N_{h_1}^2}{n_{h_1} S_{h_1}^2}
\]

\[
-2 \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} \left\lfloor 1 - \max \left( \frac{n_{h_1}}{N_{h_1}}, \frac{n_{h_2}}{N_{h_2}} \right) \right\rfloor \frac{N_{h_1} c_{h_1 h_2}}{\max \left( \frac{n_{h_1}}{N_{h_1}}, \frac{n_{h_2}}{N_{h_2}} \right)}
\] (5.4)
For the ratio and relative difference estimators, the variance is also found by plugging in the sample-based estimates into expression (4.5):

\[
\text{var}(\hat{R}) = \text{var}(\hat{RD}) \\
\approx \text{var}(\hat{T}(t_2|n_2)) + \hat{R}^2 \text{var}(\hat{T}(t_1|n_1)) - 2\hat{R}\text{cov}(\hat{T}(t_1|n_1),\hat{T}(t_2|n_2)),
\]

(5.5)

where the associated variance and covariance estimators are given in (5.3) and (5.4).

### 6. Domain-Level Modifications

How to estimate the variance of between-year change in domain-level totals depends on the type of domain. If the domain is a “planned domain,” i.e., if the domain categories are similar to the stratification categories of variables used in the sample design, then the stratification variable \( h_1 \) and \( h_2 \) in the (5.3)-(5.5) formulas needs to be redefined as the intersection of the domain indicator and the design stratum identifier. Thus, the number of population and sample (stratum x domain) jumpers needs to be tabulated from the frame file (or estimated from the sample) and used in all calculations, to account for units that shift across domains between the two years. If the domains are “analysis domains”, i.e., ones not related to the strata, then the stratification identifiers are unchanged.

For both types of domains, the variable of interest within each year needs to be modified (recoded in the sample dataset), for each domain \( d \), as follows:

\[
z_{1di} = \begin{cases} 
y_{1i} & \text{if } i \in d \text{ in year 1} \\
0 & \text{otherwise}
\end{cases} \quad \text{and} \quad z_{2di} = \begin{cases} 
y_{2i} & \text{if } i \in d \text{ in year 2} \\
0 & \text{otherwise}
\end{cases}
\]

Multiplying the domain indicator by each variable value each year will create the above variables to use in place of \( y_{1i} \) and \( y_{2i} \) in the formulas (i.e., the stratum-level population and sample counts, including the strata jumpers, remain the same). This corresponds to the same approach used in SUDAAN’s domain estimation (with the “subpop” statement, Shah et al., 1993).

For significance tests, Henry et al. (2008) used \( Z = 1.96 \) as the critical value for the national-level estimates. For domain-level estimates, if the number of units within a particular domain is less than 60, we suggest using \( t_{DF} \), with the following rule-of-thumb for the degrees-of-freedom involving the number of domain units minus the total number of strata for each year: \( DF = \min(n_{d1} - H_1, n_{d2} - H_2) \). However, given that our smallest domain size exceeded 7,000 units, there were negligible differences between the above \( t_{DF} \) vs. \( Z = 1.96 \) in our application. From this, the corresponding \( t_{DF} \) results are omitted.

### 7. Results

To account for SOI’s sample including prior year returns in each sample, we matched the most recent tax return within each year together in both the population and sample. This led to ignoring a few cases where a taxpayer filed more than one return in a single year even though these returns were used in estimating single-year totals. There were 154,772 returns that overlapped in the 2004 and 2005 samples; matching on the most recent tax periods resulted in 143,707 of these returns being used to estimate the covariance. Thus, doing this led to a slight underestimation of the covariances, but the impact of this was much less than ignoring the covariance term completely.

We consider eight variables of interest whose differences between SOI’s Tax Years 2004 and 2005 Individual samples were published (IRS 2006), by nineteen analysis domains formed using categories of the taxpayers’ size of Tax Year 2005 Adjusted Gross Income. Table 2 shows the point estimate of the

2647
relative differences in the yearly totals, relative to TY 2004, by these domains. SOI’s sample is designed to
oversample returns with larger income; the domains cover ranges of both small and large income; generally
the sampling variances are larger within the smaller income domains. The relative differences are widely
ranging, from 0.2% (for Taxable Interest Income in the $25-30,000 category) to 1,970.0% (for Alternative
Minimum Tax in the $30-40,000 category).

Figures 1 through 4 show the confidence intervals (CIs) for the Table 2 relative differences when
estimating or ignoring the covariance term in (5.4). Each plot shows the CIs for two variables; extremely
large values of CI endpoints for one relative difference were truncated in Figures 2 (Total Income Tax in
the Negative or No Adjusted Gross Income category), 3 (Alternative Minimum Tax in the $30-40,000
category and Net Capital Gains (less loss) in the $5-10,000 category) for display purposes. Also, in Figure
3, some domains for the Alternative Minimum Tax totals were collapsed due to disclosure for the single-
year totals (IRS 2006).

In all figures, ignoring the covariance lead to excessively large variance estimates, since the benefit of
the large sample overlap is ignored. Generally this lead to wider CIs, but it depended on both the variable
and domain of interest. For each variable, ignoring the covariance in some domains resulted in a CI wider
to the extent that it covered zero when ignoring the covariance but did not when estimating the covariance.
However, the excessively large relative differences shown in Table 1 have corresponding large sampling
variances; their confidence intervals cover zero regardless of estimating or ignoring the covariance. The
interpretation of these results is that small relative differences (in the range of 1.9-4%) can be significantly
different from zero, while extremely large relative differences may not be, depending on the magnitude of
sampling error. These illustrations are useful in gauging the statistical significance of between-year change.

8. Conclusions

We extend theory developed in Henry et al. (2008) to estimate the variance of the between-year differences
in domain-level totals. The large overlap of units between samples resulted in a large covariance term in
both the conditional variance estimator, even at the domain-level. Our estimators allow us to gauge whether
both small and extremely large differences between SOI’s 2004 and 2005 Individual tax return samples
were statistically significantly different from zero. This is useful information for economists and other data-
users interpreting the SOI sample results and making inferences about year-to-year change.

Despite large computing resources needed to match the two year’s population files, it was not difficult
to compute the variance estimates once the \( n_{hk} \), \( N_{hk} \) and \( c_{hk} \) quantities were produced. The ratio and
relative difference estimators of between-year change were also not difficult to produce, nor were the
domain-level extensions. All variance and covariance formulas were easily programmable in SUDAAN
and SAS, respectively.

REFERENCES


Change in Totals from Two Stratified Bernoulli Samples,” *2008 Proceedings of the Section on Survey

46.


---

**Table 1. Partition of Universe at Two Times**

<table>
<thead>
<tr>
<th>Time $t_2$ Stratum Membership</th>
<th>Time $t_1$ Stratum Membership</th>
<th>Stratum universe at time $t_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (deaths in $t_1$)</td>
<td>1</td>
<td>$H_2$</td>
</tr>
<tr>
<td>1</td>
<td>$U_{10}$</td>
<td></td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td></td>
</tr>
<tr>
<td>$H_1$</td>
<td>$U_{H_1,0}$</td>
<td>$U_{H_1,1}$</td>
</tr>
<tr>
<td>Stratum universe at $t_2$</td>
<td>--</td>
<td>$U_{\bullet 1} = \bigcup_{h=0}^{H_1} U_{h1}$</td>
</tr>
</tbody>
</table>

---

2649
Table 2. Between-Year Relative Difference (%) Estimates

<table>
<thead>
<tr>
<th>Domain</th>
<th>Adjusted Gross Income</th>
<th>Taxable Income</th>
<th>Total Income Tax</th>
<th>Business or profession net income (less loss)</th>
<th>Alternative minimum tax</th>
<th>Net capital gain (less loss)</th>
<th>Charitable contributions</th>
<th>Charitable contributions other than cash</th>
</tr>
</thead>
<tbody>
<tr>
<td>No adjusted gross income</td>
<td>-1.0</td>
<td>-</td>
<td>55.0</td>
<td>2.2</td>
<td>65.5</td>
<td>16.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>under $5,000</td>
<td>-2.9</td>
<td>31.3</td>
<td>32.2</td>
<td>-7.8 **</td>
<td>-64.0</td>
<td>-3.9</td>
<td>26.2</td>
<td></td>
</tr>
<tr>
<td>$5,000 under $10,000</td>
<td>-0.3</td>
<td>-7.3</td>
<td>-5.5</td>
<td>3.0 **</td>
<td>-1,316.8</td>
<td>-9.4</td>
<td>7.6</td>
<td></td>
</tr>
<tr>
<td>$10,000 under $15,000</td>
<td>0.1</td>
<td>-2.8</td>
<td>-2.9</td>
<td>1.6</td>
<td>-39.0</td>
<td>117.4</td>
<td>-4.5</td>
<td>-17.5</td>
</tr>
<tr>
<td>$15,000 under $20,000</td>
<td>-1.4</td>
<td>-4.3</td>
<td>-4.8</td>
<td>-1.3</td>
<td>-70.8</td>
<td>1.1</td>
<td>-5.2</td>
<td>20.9</td>
</tr>
<tr>
<td>$20,000 under $25,000</td>
<td>0.7</td>
<td>-3.0</td>
<td>-2.9</td>
<td>-0.8</td>
<td>195.3</td>
<td>9.5</td>
<td>-5.8</td>
<td>-11.0</td>
</tr>
<tr>
<td>$25,000 under $30,000</td>
<td>2.7</td>
<td>0.2</td>
<td>0.3</td>
<td>7.2</td>
<td>215.8</td>
<td>24.5</td>
<td>-3.2</td>
<td>-8.3</td>
</tr>
<tr>
<td>$30,000 under $40,000</td>
<td>0.3</td>
<td>-2.2</td>
<td>-3.3</td>
<td>7.7</td>
<td>1,970.0</td>
<td>39.9</td>
<td>-5.2</td>
<td>-10.3</td>
</tr>
<tr>
<td>$40,000 under $50,000</td>
<td>0.5</td>
<td>-1.8</td>
<td>-3.1</td>
<td>-5.7</td>
<td>57.4</td>
<td>88.9</td>
<td>-3.1</td>
<td>-3.9</td>
</tr>
<tr>
<td>$50,000 under $75,000</td>
<td>1.6</td>
<td>0.1</td>
<td>0.5</td>
<td>1.7</td>
<td>10.5</td>
<td>27.0</td>
<td>-0.4</td>
<td>-1.5</td>
</tr>
<tr>
<td>$75,000 under $100,000</td>
<td>3.2</td>
<td>1.9</td>
<td>0.5</td>
<td>7.5</td>
<td>18.8</td>
<td>25.0</td>
<td>3.9</td>
<td>-7.7</td>
</tr>
<tr>
<td>$100,000 under $200,000</td>
<td>11.0</td>
<td>9.5</td>
<td>8.1</td>
<td>13.5</td>
<td>29.3</td>
<td>30.8</td>
<td>8.3</td>
<td>3.8</td>
</tr>
<tr>
<td>$200,000 under $500,000</td>
<td>16.6</td>
<td>16.2</td>
<td>14.5</td>
<td>13.2</td>
<td>29.4</td>
<td>42.1</td>
<td>13.7</td>
<td>15.6</td>
</tr>
<tr>
<td>$500,000 under $1,000,000</td>
<td>21.1</td>
<td>20.5</td>
<td>18.7</td>
<td>16.0</td>
<td>44.6</td>
<td>36.5</td>
<td>19.5</td>
<td>-6.9</td>
</tr>
<tr>
<td>$1,000,000 under $1,500,000</td>
<td>23.3</td>
<td>22.8</td>
<td>22.5</td>
<td>33.9 **</td>
<td>27.5</td>
<td>25.1</td>
<td>0.7</td>
<td>-0.7</td>
</tr>
<tr>
<td>$1,500,000 under $2,000,000</td>
<td>25.4</td>
<td>25.5</td>
<td>23.2</td>
<td>35.1 **</td>
<td>40.3</td>
<td>20.6</td>
<td>47.1</td>
<td></td>
</tr>
<tr>
<td>$2,000,000 under $5,000,000</td>
<td>28.9</td>
<td>28.6</td>
<td>26.9</td>
<td>27.5 **</td>
<td>35.3</td>
<td>33.7</td>
<td>18.8</td>
<td></td>
</tr>
<tr>
<td>$5,000,000 under $10,000,000</td>
<td>35.7</td>
<td>35.7</td>
<td>34.1</td>
<td>62.2 **</td>
<td>44.0</td>
<td>34.9</td>
<td>13.0</td>
<td></td>
</tr>
<tr>
<td>$10,000,000 +</td>
<td>46.4</td>
<td>46.7</td>
<td>44.4</td>
<td>64.5 **</td>
<td>52.2</td>
<td>36.0</td>
<td>38.6</td>
<td></td>
</tr>
<tr>
<td>Total (all returns)</td>
<td>9.3</td>
<td>10.0</td>
<td>12.4</td>
<td>9.1</td>
<td>33.7</td>
<td>40.6</td>
<td>10.8</td>
<td>10.8</td>
</tr>
</tbody>
</table>

Notes: ** indicates suppressed estimate (collapsed with preceding domain(s)); - indicates Not Applicable.
Figure 1. Confidence Intervals for Between-Year Relative Difference Estimates When Ignoring vs. Estimating the Covariance, by Adjusted Gross Income (AGI) Category: Adjusted Gross Income and Taxable Interest Income

<table>
<thead>
<tr>
<th>AGI Category</th>
<th>Adjusted Gross Income</th>
<th>Taxable Interest Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>No AGI / Negative AGI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under $5,000</td>
<td>Sig.</td>
<td>Non-Sig.</td>
</tr>
<tr>
<td>$5,000 under $10,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$10,000 under $15,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$15,000 under $20,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$20,000 under $25,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$25,000 under $30,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$30,000 under $40,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$40,000 under $50,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$50,000 under $75,000</td>
<td>Sig.</td>
<td>Non-Sig.</td>
</tr>
<tr>
<td>$75,000 under $100,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$100,000 under $200,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$200,000 under $500,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$500,000 under $1,000,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1,000,000 under $1,500,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1,500,000 under $2,000,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$2,000,000 under $5,000,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$5,000,000 under $10,000,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$10,000,000+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (All Returns)</td>
<td>Sig.</td>
<td>Non-Sig.</td>
</tr>
</tbody>
</table>

LEGEND:
* = relative difference point estimate
— = CI when ignoring covariance
— = CI when estimating covariance
| | = truncated reldiff. and CI endpoints
D = Suppressed estimate (collapsed with above domains)
NA = Not applicable (AGI category doesn't have variable amount)
Sig. = 95% CI covers zero when ignoring the covariance, but doesn't when estimating it
Figure 2. Confidence Intervals for Between-Year Relative Difference Estimates When Ignoring vs. Estimating the Covariance, by AGI Category, Total Income Tax and Business or Profession Net Income (Less Loss)

<table>
<thead>
<tr>
<th>AGI Category</th>
<th>Total Income Tax</th>
<th>Business or Profession Net Income (Less Loss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No AGI / Negative AGI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under $5,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$5,000 under $10,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$10,000 under $15,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$15,000 under $20,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$20,000 under $25,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$25,000 under $30,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$30,000 under $40,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$40,000 under $50,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$50,000 under $75,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$75,000 under $100,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$100,000 under $200,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$200,000 under $500,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$500,000 under $1,000,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1,000,000 under $1,500,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1,500,000 under $2,000,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$2,000,000 under $5,000,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$5,000,000 under $10,000,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$10,000,000+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (All Returns)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Relative Difference (%)
Figure 3. Confidence Intervals for Between-Year Relative Difference Estimates When Ignoring vs. Estimating the Covariance, by AGI Category, Alternative Minimum Tax and Net Capital Gain (Less Loss)

AGI Category | Alternative Minimum Tax | Net Capital Gain (Less Loss)
---|---|---
No AGI / Negative AGI
Under $5,000
$5,000 under $10,000
$10,000 under $15,000
$15,000 under $20,000
$20,000 under $25,000
$25,000 under $30,000
$30,000 under $40,000
$40,000 under $50,000
$50,000 under $75,000
$75,000 under $100,000
$100,000 under $200,000
$200,000 under $500,000
$500,000 under $1,000,000
$1,000,000 under $1,500,000
$1,500,000 under $2,000,000
$2,000,000 under $5,000,000
$5,000,000 under $10,000,000
$10,000,000+
Total (All Returns)

Relative Difference (%)
Figure 4. Confidence Intervals for Between-Year Relative Difference Estimates When Ignoring vs. Estimating the Covariance, by AGI Category, Charitable Contributions and Charitable Contributions Other Than Cash

AGI Category

<table>
<thead>
<tr>
<th>Charitable Contributions</th>
<th>Charitable Contributions Other Than Cash</th>
</tr>
</thead>
<tbody>
<tr>
<td>No AGI / Negative AGI</td>
<td>NA</td>
</tr>
<tr>
<td>Under $5,000</td>
<td>NA</td>
</tr>
<tr>
<td>$5,000 under $10,000</td>
<td>NA</td>
</tr>
<tr>
<td>$10,000 under $15,000</td>
<td>NA</td>
</tr>
<tr>
<td>$15,000 under $20,000</td>
<td>NA</td>
</tr>
<tr>
<td>$20,000 under $25,000</td>
<td>NA</td>
</tr>
<tr>
<td>$25,000 under $30,000</td>
<td>NA</td>
</tr>
<tr>
<td>$30,000 under $40,000</td>
<td>NA</td>
</tr>
<tr>
<td>$40,000 under $50,000</td>
<td>NA</td>
</tr>
<tr>
<td>$50,000 under $75,000</td>
<td>NA</td>
</tr>
<tr>
<td>$75,000 under $100,000</td>
<td>NA</td>
</tr>
<tr>
<td>$100,000 under $200,000</td>
<td>NA</td>
</tr>
<tr>
<td>$200,000 under $500,000</td>
<td>NA</td>
</tr>
<tr>
<td>$500,000 under $1,000,000</td>
<td>NA</td>
</tr>
<tr>
<td>$1,000,000 under $1,500,000</td>
<td>NA</td>
</tr>
<tr>
<td>$1,500,000 under $2,000,000</td>
<td>NA</td>
</tr>
<tr>
<td>$2,000,000 under $5,000,000</td>
<td>NA</td>
</tr>
<tr>
<td>$5,000,000 under $10,000,000</td>
<td>NA</td>
</tr>
<tr>
<td>$10,000,000+</td>
<td>NA</td>
</tr>
<tr>
<td>Total (All Returns)</td>
<td>NA</td>
</tr>
</tbody>
</table>

Legend:
* = relative difference point estimate
— = CI when ignoring covariance
— = CI when estimating covariance
| | = truncated reldiff. and CI endpoints
D = Suppressed estimate (collapsed with above domains)
NA = Not applicable (AGI category doesn’t have variable amount)
Sig. Non-sig. = 95% CI covers zero when ignoring the covariance, but doesn’t when estimating it

Relative Difference (%)