# Variance Estimation for Estimators of Between-Year Change in Totals from Two Stratified Bernoulli Samples

*Kimberly Henry and Valerie Testa, Internal Revenue Service*
*and Richard Valliant, University of Michigan*

## ▶ Introduction and Universe of Tax Returns for Two Years

This paper contains the theoretical background necessary 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. The underlying theory here is modified from the theory in Valliant and Casady (1998). Related work for similar sample designs can be found in Berger (2004), Nordberg (2000), and Wood (2008).

We consider two estimators of the finite population total: the Horvitz-Thompson (HT) and post-stratification (PS) estimators. SOI uses the PS estimator to estimate yearly totals, but both are considered here for comparison purposes. 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. Both estimators are affected by the location of sample units within strata in both years, so we define the following:

- $U_{h_1 0}$ = returns in stratum $h_1$ that file only at time $t_1$ (deaths after time $t_1$ and before time $t_2$)
- $U_{0h_2}$ = returns in stratum $h_2$ that file only at time $t_2$ (births after time $t_1$ and before time $t_2$)
- $U_{h_1h_2}$ = returns in stratum $h_1$ 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 on the following page. If the set of strata is the same in the two years, then $H_1 = H_2$ and returns on the diagonal of Table 1 remain in the same stratum between the two years, while the "stratum jumpers" (returns that shift to different strata between the years) lie on the off-diagonal. 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 $h_1$ of Table 1, respectively:

$$U_{h_1\bullet} = \bigcup_{h_2=0}^{H_2} U_{h_1h_2} \text{ and } U_{\bullet h_2} = \bigcup_{h_1=0}^{H_1} U_{h_1h_2} .$$

## ▶ Sample Design

The stratified Bernoulli design is used by most of SOI's cross-sectional studies. 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

$$r_i < \pi_h, \tag{2.1}$$

where $\pi_h$ is the pre-assigned sampling rate for stratum $h$ that tax return $i$ belongs to. Stratification for SOI's Form 1040 sample used various criteria, including size of total gross positive/negative income indexed for inflation and an indicator of the return's "useful-

This paper was originally presented at the 2008 Joint Statistical Meetings held in Denver, Colorado, on August 3–7, 2008.

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

| | Time $t_2$ Stratum Membership | | | | |
|---|---|---|---|---|---|
| Time $t_1$ Stratum Membership | 0 (deaths in $t_1$) | 1 | ⋯ | $H_2$ | Stratum universe at time $t_1$ |
| 0 (births in $t_2$) | -- | $U_{01}$ | ⋯ | $U_{0H_2}$ | -- |
| 1 | $U_{10}$ | $U_{11}$ | ⋯ | $U_{1H_2}$ | $U_{1\bullet} = \bigcup_{h_2=0}^{H_2} U_{1h_2}$ |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ |
| $H_1$ | $U_{H_1 0}$ | $U_{H_1 1}$ | ⋯ | $U_{H_1 H_2}$ | $U_{H_1\bullet} = \bigcup_{h_2=0}^{H_2} U_{H_1 h_2}$ |
| Stratum universe at time $t_2$ | -- | $U_{\bullet 1} = \bigcup_{h_1=0}^{H_1} U_{h_1 1}$ | ⋯ | $U_{\bullet H_2} = \bigcup_{h_1=0}^{H_1} U_{h_1 H_2}$ | |

ness" for tax policy modeling purposes, to create 208 strata (see Testa and Scali 2005 for details). Every week of the IRS processing year, all 1040, 1040A, and 1040EZ returns in the frame population were assigned to a stratum as the data were transcribed by IRS. Due to this weekly schedule of available frame data, SOI uses stratified Bernoulli sampling to select its samples. Similar procedures are used by other tax agencies (e.g., Revenue Canada, Cooney 2008).

SOI's Individual sample consisted of two parts within each stratum. First, a 0.05 percent stratified Bernoulli sample of approximately 65,000 returns was selected, called the Continuous Work History Sample (CWHS, Weber 2004). A Bernoulli sample was also selected independently from each stratum, with rates ranging from 0.01 to 100 percent. The full sample thus consists of the CWHS plus all additional returns selected via the Bernoulli sample. For Tax Year 2005, the sampling rates were increased and ranged from 0.05 to 100 percent. Table 2 contains the sample and estimated population sizes (using the sample weights, both after the sample was selected and those numbers realized after SOI's data capture and cleaning.

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.

**Table 2. Realized Sample Sizes and Estimated Population Sizes, Tax Years 2004 and 2005**

| | Tax Year 2004 | Tax Year 2004 |
|---|---|---|
| *Sample Sizes* | | |
| Originally selected | 200,778 | 292,966 |
| After data cleaning | 200,295 | 292,837 |
| *Estimated Population Sizes* | | |
| Originally selected | 133,189,982 | 134,494,440 |
| After data cleaning | 132,226,042 | 134,372,678 |

- *Deaths*: units are not present in the population file, so they are not in the sample.

- *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 are four conditions related to whether or not a stratum jumper is selected for the $t_2$ sample, which are shown in Table 3. They depend on the size of the two years' sampling fractions (relative to each other) and if the unit was in the year $t_1$ sample.

The Table 3 conditions are further explained, where $h_1$ denotes the year $t_1$ stratum and $h_2$ for year $t_2$. These probabilities are needed for subsequent variance calculations.

**Table 3. Sample Inclusion of Stratum Jumpers in Year**

| | *Was unit $i$ in the sample in year $t_1$ ?* | |
|---|---|---|
| *Sampling rate relationship* | **Yes** | **No** |
| $\pi_{h_1} \leq \pi_{h_2}$ | Yes | Maybe |
| $\pi_{h_1} > \pi_{h_2}$ | Maybe | No |

## Returns selected in year $t_1$:

- If $\pi_{h_1} \leq \pi_{h_2}$, then unit $i$ will automatically be included in the year $t_2$ sample with conditional probability of selection 1 and unconditional probability of selection $\pi_{h_2}$, since $r_i < \pi_{h_1} \leq \pi_{h_2}$.
- If unit $i$ switches to a stratum with a smaller sampling rate, then it will be in the year $t_2$ sample if $r_i < \pi_{h_2}$ with conditional probability $\pi_{h_2}/\pi_{h_1}$ (and not in the sample with probability $1-(\pi_{h_2}/\pi_{h_1})$.

## Returns not selected in year $t_1$:

- If $\pi_{h_1} \leq \pi_{h_2}$, then a unit will only be in the year $t_2$ sample if $\pi_{h_1} < r_i \leq \pi_{h_2}$ with conditional probability $(\pi_{h_2}-\pi_{h_1})/(1-\pi_{h_1})$ (and not in with probability $1-(\pi_{h_2}-\pi_{h_1})/(1-\pi_{h_1})$).
- If unit $i$ switches to a stratum with a smaller sampling rate, then it will not be sampled at time $t_2$ since $\pi_{h_2} < \pi_{h_1} \leq r_i$ does not meet condition (2.1).

This sample selection method also ensures a large overlap between two years, since a sampled unit is selected in both years if $r_i \leq \pi_{h_1} \leq \pi_{h_2}$. There are a small number of 1040 tax returns where $r_i$ changes, even though the composition of the return itself remains the same (which can occur, for example, on a married joint tax return that switches the primary and secondary SSN's, since $r_i$ is calculated using the primary SSN). The overlapping of units across different year's samples also 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. For example, in the 1040 sample, "marriages" occur when two tax returns that previously filed separately file as a joint-married return the next year and "divorces" when a joint married tax return becomes two separate entities (which includes both legally divorced taxpayers and married taxpayers who choose to file their returns separately). We use the following rules in the covariance estimation, which generally lead to underestimating it:

- *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 that are in both years are tracked and considered the same unit in both years.

Sample design changes also result in sampling rate changes between years. For example, the 1040 samples included returns selected for the CWHS part of the sample using certain endings of the SSN. For Tax Year 2004, only half of the possible SSN endings were edited and used to produce the 2004 cross-sectional estimates and a defacto 100 percent increase in sampling rates. However, all selected CWHS returns were edited for the 2005 sample, resulting in approximately 60,000 additional returns used to produce the 1040 cross-sectional sample estimates. Also, the Congressionally-mandated five-year Foreign Income study was selected in the 2006 sample, resulting in sampling rate increases to include approximately 15,000 more returns in the associated strata. Our estimators can account for such changes.

# ▶ Estimators for Totals and Their Change

## Notation and Probabilities of Selection

A Bernoulli sample is selected within each stratum as described in Section 3, where $\pi_h$, the stratum sampling

rate in a given year, is also the probability of selection for all units in stratum $h$. The following random variables denote sample inclusion of unit $i$ at times $t_1$ and $t_2$:

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

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

$$E\left[\delta_i(t_1)\right] = \pi_{h_1}, \quad Var\left[\delta_i(t_1)\right] = \pi_{h_1}(1-\pi_{h_1})$$
$$E\left[\delta_i(t_2)\right] = \pi_{h_2}, \quad Var\left[\delta_i(t_2)\right] = \pi_{h_2}(1-\pi_{h_2}).$$

To compute $E[\delta_i(t_2)\delta_i(t_1)] - E[\delta_i(t_2)]E[\delta_i(t_1)]$ for the covariance, note that $\delta_i(t_2)\delta_i(t_1) = 1$ only when a unit is in the sample for both time periods. Thus, 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:

$$Cov\left[\delta_i(t_2), \delta_i(t_1)\right] = \min(\pi_{h_1}, \pi_{h_2}) - \pi_{h_1}\pi_{h_2}$$
$$= \Delta_{h_1 h_2}$$

## Finite Population Totals of Interest

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

$$T(t_1) = \sum_{h_1=1}^{H_1} \sum_{i \in U_{h_1 \bullet}} y_{1i} \tag{3.1}$$

$$T(t_2) = \sum_{h_2=1}^{H_2} \sum_{i \in U_{\bullet h_2}} y_{2i} \tag{3.2}$$

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

## The Horvitz-Thompson Estimator

The Horvitz-Thompson (HT) estimators for $T(t_1)$, $T(t_2)$ are

$$\hat{T}_\pi(t_1) = \sum_{h_1=1}^{H_1} \sum_{i \in U_{h_1 \bullet}} \frac{\delta_i(t_1) y_{1i}}{\pi_{h_1}} = \sum_{h_1=1}^{H_1} \hat{T}_{h_1 \bullet} \tag{3.3}$$

$$\hat{T}_\pi(t_2) = \sum_{h_2=1}^{H_2} \sum_{i \in U_{\bullet h_2}} \frac{\delta_i(t_2) y_{2i}}{\pi_{h_2}} = \sum_{h_2=1}^{H_2} \hat{T}_{\bullet h_2}, \tag{3.4}$$

where $1/\pi_{h_1}$ and $1/\pi_{h_2}$ are the base weights for unit $i$ at times $t_1$ and $t_2$ and

$$\hat{T}_{h_1 \bullet} = \frac{1}{\pi_{h_1}} \sum_{i \in U_{h_1 \bullet}} \delta_i(t_1) y_{1i},$$
$$\hat{T}_{\bullet h_2} = \frac{1}{\pi_{h_2}} \sum_{i \in U_{\bullet h_2}} \delta_i(t_2) y_{2i}$$

are HT estimators for each year's stratum totals. The estimators $\hat{T}_\pi(t_1)$ and $\hat{T}_\pi(t_2)$ are unconditionally (i.e., over all possible samples for each year) unbiased for $T(t_1)$ and $T(t_2)$, respectively.

## The Poststratified (Conditional HT) Estimator

Even though the HT estimator is unconditionally unbiased for the population total, it can have a high variance since the sample size is a random variable under Bernoulli sampling. Instead, 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 $n_{h_1 \bullet} = \sum_{i \in U_{h_1 \bullet}} \delta_i(t_1)$. Assuming that $n_{h_1 \bullet} > 0$, it can be shown that conditional on $\{n_{1\bullet}, n_{2\bullet}, \cdots, n_{H_1 \bullet}\}$, the sample design at time $t_1$ is a stratified simple random sample with stratum sample

sizes $n_{1\bullet}, n_{2\bullet}, \cdots, n_{H_2\bullet}$. Thus, for $N_{h_1\bullet}$ denoting the number of population units in stratum $h_1$ at time $t_1$, the (conditional) HT estimator for $T(t_1)$ is

$$\hat{T}_{\pi c}(t_1) = \sum_{h_1=1}^{H_1} \frac{N_{h_1\bullet}}{n_{h_1\bullet}} \sum_{i \in U_{h_1\bullet}} \delta_i(t_1) y_{1i} \,. \tag{3.5}$$

Similarly, if $n_{\bullet h_2} = \sum_{i \in U_{\bullet h_2}} \delta_i(t_2) > 0$, then conditional on $\{n_{\bullet 1}, n_{\bullet 2}, \cdots, n_{\bullet H_2}\}$, the sample design at time $t_2$ is a stratified simple random sample with stratum sample sizes $n_{\bullet 1}, n_{\bullet 2}, \cdots, n_{\bullet H_2}$. For $N_{\bullet h_2}$ being the number of population elements in stratum $h$ at time $t_2$, the (conditional) HT estimator for $T(t_2)$ is

$$\hat{T}_{\pi c}(t_2) = \sum_{h_2=1}^{H_2} \frac{N_{\bullet h_2}}{n_{\bullet h_2}} \sum_{i \in U_{\bullet h_2}} \delta_i(t_2) y_{2i} \,. \tag{3.6}$$

SOI uses $\hat{T}_{\pi c}(t_1)$ and $\hat{T}_{\pi c}(t_2)$ to estimate time-specific totals, which are conditionally unbiased for their corresponding population totals. They are special forms of the PS estimator, where the poststrata are the same as the design strata. The PS estimators (which are not unconditionally unbiased for the population totals, since they involve a ratio) for times $t_1$ and $t_2$ are

$$\hat{T}_{PS}(t_1) = \sum_{h_1=1}^{H_1} \frac{N_{h_1\bullet}}{\hat{N}_{h_1\bullet}} \hat{T}_{h_1\bullet} \tag{3.7}$$

$$\hat{T}_{PS}(t_2) = \sum_{h_2=1}^{H_2} \frac{N_{\bullet h_2}}{\hat{N}_{\bullet h_2}} \hat{T}_{\bullet h_2} \,, \tag{3.8}$$

where $\hat{N}_{h_1\bullet} = \frac{1}{\pi_{h_1}} \sum_{i \in U_{h_1\bullet}} \delta_i(t_1)$ is the estimated number of stratum $h_1$ population units for year 1. For year 2, $\hat{N}_{\bullet h_2}$ is similarly defined. It can be shown, by definition of $\hat{T}_{h_1\bullet}$ and $\hat{T}_{\bullet h_2}$, that $\hat{T}_{PS}(t_1)$ reduces to $\hat{T}_{\pi c}(t_1)$ and $\hat{T}_{PS}(t_2)$ reduces to $\hat{T}_{\pi c}(t_2)$.

## Estimators of Change

The change in level between two time points is denoted by

$$D = T(t_2) - T(t_1) \,. \tag{3.9}$$

The conditional and unconditional estimators of time-specific totals lead to two estimators of this difference. Based on the unconditional HT estimators, we have

$$\hat{D}_{\pi} = \hat{T}_{\pi}(t_2) - \hat{T}_{\pi}(t_1) \,, \tag{3.10}$$

which is unconditionally unbiased for the change in level between time $t_1$ and $t_2$. By breaking $\hat{T}_{\pi}(t_1)$ into the sum of births for time $t_2$ and units in both year's sample summed over the year 1 strata, and $\hat{T}_{\pi}(t_2)$ into the sum of the deaths for time $t_1$ and the units in both samples over the year 2 strata, expression (3.10) can be rewritten as:

$$\begin{aligned} \hat{D}_{\pi} = & \sum_{h_2=1}^{H_2} \sum_{i \in U_{0h_2}} \frac{\delta_i(t_2) y_{2i}}{\pi_{h_2}} - \sum_{h_1=1}^{H_1} \sum_{i \in U_{h_1 0}} \frac{\delta_i(t_1) y_{1i}}{\pi_{h_1}} \\ & + \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} \sum_{i \in U_{h_1 h_2}} \left[ \frac{\delta_i(t_2) y_{2i}}{\pi_{h_2}} - \frac{\delta_i(t_1) y_{1i}}{\pi_{h_1}} \right] \end{aligned} \tag{3.11}$$

Similarly, based on the conditional HT estimators,

$$\hat{D}_{PS} = \hat{T}_{PS}(t_2) - \hat{T}_{PS}(t_1) \tag{3.12}$$

which is conditionally unbiased for the change in level. Similarly, this estimator can be rewritten as:

$$\begin{aligned} \hat{D}_{PS} = & \sum_{h_2=1}^{H_2} \sum_{i \in U_{0h_2}} \frac{N_{\bullet h_2} \delta_i(t_2) y_{2i}}{n_{\bullet h_2}} - \sum_{h_1=1}^{H_1} \sum_{i \in U_{h_1 0}} \frac{N_{h_1\bullet} \delta_i(t_1) y_{1i}}{n_{h_1\bullet}} \\ & + \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} \sum_{i \in U_{h_1 h_2}} \left[ \frac{N_{\bullet h_2} \delta_i(t_2) y_{2i}}{n_{\bullet h_2}} - \frac{N_{h_1\bullet} \delta_i(t_1) y_{1i}}{n_{h_1\bullet}} \right] \end{aligned} \tag{3.13}$$

## ▶ Theoretical Variances

## Single Year Variances

### Unconditional HT Estimators

Using standard Poisson sampling results (Result 3.2.1 from Sarndal *et al.*, (1992)), the variances of the unconditional HT estimators for times $t_1$ and $t_2$ are

$$Var\left[\hat{T}_{\pi}(t_1)\right] = \sum_{h_1=1}^{H_1} \frac{\left(1-\pi_{h_1}\right)}{\pi_{h_1}} \sum_{i \in U_{h_1 \bullet}} y_{1i}^2 \qquad (4.1)$$

$$Var\left[\hat{T}_{\pi}(t_2)\right] = \sum_{h_2=1}^{H_2} \frac{\left(1-\pi_{h_2}\right)}{\pi_{h_2}} \sum_{i \in U_{\bullet h_2}} y_{2i}^2 \qquad (4.2)$$

**Unconditional Variance of PS Estimators**

Using $N_h\left(\frac{1}{\pi_h}-1\right) \approx \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)$, Sarndal et al. (expression 3.2.7, p. 65) approximated the unconditional variance of the PS estimators. Their ap-proximation, adjusted for our notation, is

$$Var\left[\hat{T}_{PS}(t_1)\right] \approx \sum_{h_1=1}^{H_1} N_{h_1 \bullet}\left(\frac{1}{\pi_{h_1}}-1\right) S_{h_1 \bullet}^2 \qquad (4.3)$$

$$Var\left[\hat{T}_{PS}(t_2)\right] \approx \sum_{h_2=1}^{H_2} N_{\bullet h_2}\left(\frac{1}{\pi_{h_2}}-1\right) S_{\bullet h_2}^2 , \qquad (4.4)$$

where

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

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

are the population stratum variances for $t_1$ and $t_2$ and the population strata means are

$$\bar{Y}_{h_1 \bullet} = \frac{1}{N_{h_1 \bullet}} \sum_{i \in U_{h_1 \bullet}} y_{1i}$$

$$\bar{Y}_{\bullet h_2} = \frac{1}{N_{\bullet h_2}} \sum_{i \in U_{\bullet h_2}} y_{2i} .$$

**Conditional Variances of PS Estimators**

Holt and Smith (1979) observed that conditioning on an achieved post stratum sample size is inferentially more appropriate than averaging over all possible sample sizes, as in (4.1) and (4.2). The theoretical conditional variances of the PS estimators for both years are simply the variances of a total under stratified simple random sampling:

$$Var\left[\hat{T}_{PS}(t_1)\right] = \sum_{h_1=1}^{H_1} \frac{N_{h_1 \bullet}^2}{n_{h_1 \bullet}}\left(1-\frac{n_{h_1 \bullet}}{N_{h_1 \bullet}}\right) S_{h_1 \bullet}^2 \qquad (4.5)$$

$$Var\left[\hat{T}_{PS}(t_2)\right] = \sum_{h_2=1}^{H_2} \frac{N_{\bullet h_2}^2}{n_{\bullet h_2}}\left(1-\frac{n_{\bullet h_2}}{N_{\bullet h_2}}\right) S_{\bullet h_2}^2 \qquad (4.6)$$

The variances in (4.5) and (4.6) are preferable since they reflect the precision for the sample sizes actually obtained. In comparing expressions (4.1) to (4.5) and (4.2) to (4.6), in almost all practical situations, the conditional variances of the PS estimators are substantially smaller than the unconditional variances of the HT estimators. To see this, in general, we can write (expanding Expression 3.2.5 in Sarndal et al. (1992) to stratified sampling):

$$\begin{aligned} Var_{STBE}\left[\hat{T}_{\pi}\right] &= Var_{STSI}\left[\hat{T}_{\pi}\right] + \sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right) S_h^2 \left[1-\frac{1}{N_h}+\frac{1}{CV_{yh}^2}\right] \\ &\approx Var_{STSI}\left[\hat{T}_{\pi}\right] + \sum_{h=1}^{H} N_h\left(\frac{N_h}{n_h}-1\right) \bar{Y}_h^2 \left[1+\frac{CV_{yh}^2}{N_h}\right] \end{aligned} \qquad (4.7)$$

Thus, the theoretical variance of a total under stratified Bernoulli sampling is equal to the stratified simple random sampling variance plus an additional factor (of the same magnitude) that depends on the stratum population mean of the study variable and the population coefficient of variation:

$$CV_{yh} = \frac{S_h}{\bar{Y}_h}$$

This means that we should expect differences in the variance of each HT total, whose size depends on the strata means of the underlying variable of interest.

## Variance of the Difference

**Unconditional Variance of the HT Estimators**

For the unconditional HT estimators, it can be shown that the unconditional variance of the difference is the variance from the year 1, births plus the variance from the year 2 deaths, plus the variance from returns in the same or different strata in both years. In formula form, this is:

$$Var\left[\hat{D}_\pi\right]=\sum_{h_2=1}^{H_2}\frac{1-\pi_{h_2}}{\pi_{h_2}}\sum_{i\in U_{\bullet h_2}}y_{2i}^2+\sum_{h_1=1}^{H_1}\frac{1-\pi_{h_1}}{\pi_{h_1}}\sum_{i\in U_{h_1\bullet}}y_{1i}^2 +\sum_{h_1=1}^{H_1}\sum_{h_2=1}^{H_2}\sum_{i\in U_{h_1h_2}}\left[\frac{\left(1-\pi_{h_2}\right)y_{2i}^2}{\pi_{h_2}}+\frac{\left(1-\pi_{h_1}\right)y_{1i}^2}{\pi_{h_1}}-2\frac{\Delta_{h_1h_2}y_{2i}y_{1i}}{\pi_{h_1}\pi_{h_2}}\right] \quad (4.8)$$

**Unconditional Variance of the PS Estimators (Linear Approximation)**

Using linear approximations to the PS estimators, the unconditional variance of the difference is

$$Var\left[\hat{D}_{PS}\right]\approx\sum_{h_2=1}^{H_2}\frac{1-\pi_{h_2}}{\pi_{h_2}}N_{\bullet h_2}S_{\bullet h_2}^2+\sum_{h_1=1}^{H_1}\frac{1-\pi_{h_1}}{\pi_{h_1}}N_{h_1\bullet}S_{h_1\bullet}^2 -2\sum_{h_1=1}^{H_1}\sum_{h_2=1}^{H_2}\left[\frac{\Delta_{h_1h_2}N_{h_1h_2}S_{h_1h_2}}{\pi_{h_1}\pi_{h_2}}\right] \quad (4.9)$$

where $S_{h_1h_2}=\frac{1}{N_{h_1h_2}-1}\sum_{i\in U_{h_1h_2}}\left(y_{2i}-\bar{Y}_{\bullet h_2}\right)\left(y_{1i}-\bar{Y}_{h_1\bullet}\right)$

**Unconditional Variance of PS Estimators (Substitution Form)**

The variance in (4.8) 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 substituting marginal actual sampling rates for terms like and replace by a stratum population size divided by the actual stratum sample size. Using these, we get

$$Var\left[\hat{D}_{PS}\right]\approx\sum_{h_2=1}^{H_2}\left(1-\frac{n_{\bullet h_2}}{N_{\bullet h_2}}\right)\frac{N_{\bullet h_2}^2}{n_{\bullet h_2}}S_{\bullet h_2}^2+\sum_{h_1=1}^{H_1}\left(1-\frac{n_{h_1\bullet}}{N_{h_1\bullet}}\right)\frac{N_{h_1\bullet}^2}{n_{h_1\bullet}}S_{h_1\bullet}^2 -2\sum_{h_1=1}^{H_1}\sum_{h_2=1}^{H_2}\left[\frac{\Delta_{h_1h_2}N_{h_1h_2}S_{h_1h_2}}{\pi_{h_1}\pi_{h_2}}\right] \quad (4.10)$$

# ▶ Variance Estimators

## Single-Year Variance Estimators

We assume that the terms $N_{h_1\bullet}$ and $N_{\bullet h_2}$ are known for all $h_1$,$h_2$ and consider both cases of the $N_{h_1h_2}$ being known and unknown. Assuming that the $N_{h_1h_2}$ are unknown, these are conditionally unbiased estimators of the strata population means:

$$\bar{y}_{h_1\bullet}=\frac{1}{n_{h_1\bullet}}\sum_{i\in U_{h_1\bullet}}\delta_i(t_1)y_{1i}$$

$$\bar{y}_{\bullet h_2}=\frac{1}{n_{\bullet h_2}}\sum_{i\in U_{\bullet h_2}}\delta_i(t_2)y_{2i}\ .$$

Since conditionally (and approximately unconditionally) unbiased estimators for the strata variances $S_{h_1\bullet}^2$ and $S_{h_1\bullet}^2$ are

$$s_{h_1\bullet}^2=\frac{1}{n_{h_1\bullet}-1}\sum_{i\in U_{h_1\bullet}}\delta_i(t_1)\left(y_{1i}-\bar{y}_{h_1\bullet}\right)^2$$

$$s_{\bullet h_2}^2=\frac{1}{n_{\bullet h_2}-1}\sum_{i\in U_{\bullet h_2}}\delta_i(t_2)\left(y_{2i}-\bar{y}_{\bullet h_2}\right)^2,$$

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

$$var\left[\hat{T}_{PS}(t_1)\right]=\sum_{h_1=1}^{H_1}\frac{N_{h_1\bullet}^2}{n_{h_1\bullet}}\left(1-\frac{n_{h_1\bullet}}{N_{h_1\bullet}}\right)s_{h_1\bullet}^2 \quad (5.1)$$

$$var\left[\hat{T}_{PS}(t_2)\right]=\sum_{h_2=1}^{H_2}\frac{N_{\bullet h_2}^2}{n_{\bullet h_2}}\left(1-\frac{n_{\bullet h_2}}{N_{\bullet h_2}}\right)s_{\bullet h_2}^2 \quad (5.2)$$

These are conditionally unbiased for (4.5) and (4.6).

## Variance Estimators of the Differences

**The Unconditional Difference**

Using sample-based estimates for each (4.8) component, we get the following estimate of $Var\left[\hat{D}_\pi\right]$:

$$var\left[\hat{D}_\pi\right]=\sum_{h_2=1}^{H_2}\frac{1-\pi_{h_2}}{\pi_{h_2}^2}\sum_{i\in s_{\bullet h_2}}y_{2i}^2+\sum_{h_1=1}^{H_1}\frac{1-\pi_{h_1}}{\pi_{h_1}^2}\sum_{i\in s_{h_1\bullet}}y_{1i}^2 -2\sum_{h_1=1}^{H_1}\sum_{h_2=1}^{H_2}\sum_{i\in s_{h_1h_2}}\frac{\Delta_{h_1h_2}y_{2i}y_{1i}}{\min\left(\pi_{h_1},\pi_{h_2}\right)\pi_{h_1}\pi_{h_2}} \quad (5.3)$$

## The Conditional Difference

Using sample-based estimates for each (4.10) component, we have the approximate estimator of $Var\left[\hat{D}_{PS}\right]$:

$$var\left[\hat{D}_{PS}\right] \approx \sum_{h_2=1}^{H_2}\left(1-\frac{n_{\bullet h_2}}{N_{\bullet h_2}}\right)\frac{N_{\bullet h_2}^2}{n_{\bullet h_2}}s_{\bullet h_2}^2 + \sum_{h_1=1}^{H_1}\left(1-\frac{n_{h_1\bullet}}{N_{h_1\bullet}}\right)\frac{N_{h_1\bullet}^2}{n_{h_1\bullet}}s_{h_1\bullet}^2 - 2\sum_{h_1=1}^{H_1}\sum_{h_2=1}^{H_2}\frac{\Delta_{h_1h_2}\hat{N}_{h_1h_2}c_{h_1h_2}}{\pi_{h_1}\pi_{h_2}} \quad (5.4)$$

where $\hat{N}_{h_1h_2} = \frac{n_{h_1h_2}}{\min\left(\pi_{h_1},\pi_{h_2}\right)}$, $\pi_{h_2} = \frac{n_{\bullet h_2}}{N_{\bullet h_2}}$, $\pi_{h_1} = \frac{n_{h_1\bullet}}{N_{h_1\bullet}}$ and the *unweighted* covariance between the *y*-values for units in both years' samples is $c_{h_1h_2} = \frac{1}{n_{h_1h_2}-1}\sum_{i\in s_{h_1h_2}}\left(y_{2i}-\bar{y}_{\bullet h_2}\right)\left(y_{1i}-\bar{y}_{h_1\bullet}\right)$. Note that the (5.4) covariance term is an estimator of the unconditional covariance, but it does partially account for achieved sample sizes in using $\frac{n_{h_1\bullet}}{N_{h_1\bullet}}$ and $\frac{n_{\bullet h_2}}{N_{\bullet h_2}}$ in place of $\pi_{h_1}$ and $\pi_{h_2}$.

## ▶ Data Elements Required For Variance Estimators

The following information is needed to evaluate (5.3) and (5.4):

- The strata occupied by each sample unit at times $t_1$ and $t_2$ and the data values for every unit in the sample each time, $y_{1i}$ and $y_{2i}$.
- $N_{h_1\bullet}$, the population size of stratum $h_1$ at time $t_1$ $N_{\bullet h_2}$ and the population size of stratum $h_2$ at time $t_2$.
- $\pi_{h_1}$ the probability of selection for a unit in stratum $h_1$ at time $t_1$, and $\pi_{h_2}$, the selection probability for a unit in stratum $h_2$ at time $t_2$.
- $n_{h_1h_2}$,$N_{h_1h_2}$ and $c_{h_1h_2}$, the number of sample and population units and unweighted covariance in each of the $(h_1h_2)$ cells.

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 totals. Doing so led to a slight underestimation of the covariances, but the impact of this was much less than ignoring the covariance term.

## ▶ Results

Table 4 shows the ratios of the unconditional to conditional estimates for each year's estimated total, the difference between them, and the associated variance estimates for eight variables estimated from SOI's Tax Year 2004 and 2005 Individual samples. While the point estimates in Table 4 are essentially identical, the unconditional variances are much larger (as much as six to ten times larger, in Adjusted Gross Income and Taxable Income) than the corresponding conditional variance estimates when the means are larger for each year, as expected from (4.7).

Table 5 contains the ratio of the variance estimates of the unconditional and conditional differences in the 2004 and 2005 totals for our eight variables, when estimating and ignoring the covariance term in (5.3) and (5.4). We also include results when estimating the Nh1h2 or using the known counts in (5.4). Ignoring the covariance leads to excessively large variance estimates because the benefit of having a large sample overlap is ignored. For example, ignoring the covariance would result in an estimated variance of the difference in HT estimators that was 68% too large and 57% to 59% too large for the variance of the difference in the PS estimators. While estimating the Nh1h2 did not lead to much different covariance estimates for these national-level variables, the HT estimates had higher decreases in the variance of the difference due to the estimated covariances for all variables except net capital gain (less loss), where the percentages were close. We also calculated the t-statistics to test whether the

**Table 4. Single-Year and Between-Year Difference Estimates, by Tax Year and Variable of Interest**

| | Single-Year Estimates | | | | | | Between-Year Difference Estimates | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Estimated Means (in $'s)* | | *HT Total / PS Total* | | *HT Variance / PS Variance* | | *HT Total / PS Total* | *HT Variance / PS Variance** | |
| *Variable* | *2004* | *2005* | *2004* | *2005* | *2004* | *2005* | *(2005-2004)* | *(5.4) v1* | *(5.4) v2* |
| Adjusted Gross Income | 51,342 | 55,238 | 0.994 | 1.001 | 10.638 | 9.739 | 1.074 | 9.609 | 9.737 |
| Taxable Income | 35,320 | 38,231 | 0.994 | 1.001 | 6.642 | 6.157 | 1.066 | 6.347 | 6.448 |
| Total Income Tax | 6,292 | 6,957 | 0.995 | 1.000 | 3.891 | 3.827 | 1.041 | 4.322 | 4.419 |
| Business or profession net income (less loss) | 1,870 | 2,007 | 1.004 | 1.007 | 1.278 | 1.306 | 1.042 | 1.801 | 1.887 |
| Alternative minimum tax | 99 | 130 | 1.002 | 1.001 | 1.194 | 1.218 | 0.999 | 1.416 | 1.426 |
| Net capital gain (less loss) | 3,568 | 4,934 | 1.001 | 1.001 | 1.175 | 1.225 | 1.001 | 1.535 | 1.568 |
| Charitable contributions | 1,252 | 1,365 | 0.996 | 1.001 | 1.272 | 1.270 | 1.051 | 1.934 | 2.066 |
| Charitable contributions other than cash | 328 | 358 | 0.997 | 1.001 | 1.015 | 1.008 | 1.037 | 1.087 | 1.098 |

* v1 is expression (5.4) using the estimated $\hat{N}_{h_1h_2}$ counts; v2 is (5.4) using the known $N_{h_1h_2}$ counts.

**Table 5. Ratios of Variances of Differences when Ignoring the Covariance to Variances When Estimating It**

| *Variable* | ***HT Difference*** | ***PS Difference (Using Estimated $\hat{N}_{h_1h_2}$)*** | ***PS Difference (Using Known $N_{h_1h_2}$)*** |
|---|---|---|---|
| Adjusted Gross Income | 1.682 | 1.574 | 1.595 |
| Taxable Income | 1.660 | 1.636 | 1.662 |
| Total Income Tax | 1.648 | 1.844 | 1.885 |
| Business or profession net income (less loss) | 1.537 | 2.145 | 2.248 |
| Alternative minimum tax | 1.242 | 1.456 | 1.466 |
| Net capital gain (less loss) | 1.099 | 1.403 | 1.433 |
| Charitable contributions | 1.546 | 2.352 | 2.512 |
| Charitable contributions other than cash | 1.077 | 1.158 | 1.169 |

population differences were zero. While the t-statistics had different values depending on which variance estimator was used, all were highly significant, and were thus omitted.

## ▶ Conclusions

The large overlap of units between SOI's 2004 and 2005 Individual tax return samples resulted in a large covariance term in both the conditional and unconditional variances. In comparing the (5.3) and (5.4) estimates, using the unconditional (5.3) formulas produced larger estimates for the separate year variances. This is due to the unconditional formulas incorporating the extra variability due to random sample sizes. The most interesting result was that these variances are much larger than the associated PS estimates of the single year totals, despite the fact the HT and PS point estimates of totals are almost identical. The SOI data support expression (4.7), that is, the mean of the variable of interest affects the size of the variance estimate. Since valid inferences are obtained using the conditional variance estimates, in summary, the best estimation strategy is using expression (5.4) to estimate the variance of the difference in PS estimators with the known $N_{\bullet h_1}$, $N_{\bullet h_2}$, and $N_{h_1}N_{h_2}$ counts.

Despite large computing resources needed to match the two year's population files, it was not difficult to compute the estimates once the $n_{h_1h_2}$, $N_{h_1h_2}$ and $c_{h_1h_2}$ quantities were produced. Notably, after the two population files were merged, only the $N_{h_1h_2}$ counts were needed.

A more complicated estimator of between-year change, such as a relative change in the two years' estimated totals, would require a more sophisticated approach (such as the Taylor series approximation used in Berger 2004 and Nordberg 2000). Our estimators would also need to be slightly modified for domain estimation. Such extensions are future consideration topics.

## ▶ References

Berger, Y.G. (2004), "Variance Estimation for Measures of Change in Probability Sampling," *The Canadian Journal of Statistics*, 32, 451–467.

Brewer, K.R.W., L.J. Early, and S.F. Joyce (1972), "Selecting Several Samples from a Single Population," *Australian Journal of Statistics*, 14, 231–239.

Cooney, S. (2008), Personal communication.

Holt, D. and T.M.F. Smith (1979), "Post Stratification," *Journal of the Royal Statistical Society A*, 142, 33–46.

Nordberg, L. (2000), "On Variance Estimation for Measures of Change When Samples are Coordinated by the Use of Permanent Random Numbers," *Journal of Official Statistics*, 14, No. 4, 363–368.

Sarndal, Swensson, and Wretman (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Testa, V. and J. Scali (2006), *Statistics of Income–2004 Individual Income Tax Returns, IRS, Publication 1304*, 23–27.

Valliant, R. and R. Casady (1998), "The Variance of An Estimator of Change From the Sample of Partnership Tax Returns," *Statistical Issues Related to SOI Family Cross-Sectional and Longitudinal Studies*, Contract No. TIRN0-96-D-0030 Task 5, Washington DC.

Weber, M. (2004), "The Statistics of Income 1979–2002 Continuous Work History Sample Individual Tax Return Panel," http://www.irs.gov/pub/irs-soi/04webasa.pdf.

Wood, J. (2008), "On the Covariance Between Related Horvitz-Thompson Estimators," *Journal of Official Statistics*, 24, No. 1, 53–78.