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

Kimberly Henry[1], Valerie Testa[1], and Richard Valliant[2]
[1]Statistics of Income, P.O. Box 2608, Washington DC 20013-2608
[2] University 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 that 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:

- $U_{h_1 0} =$ returns in stratum h_1 that file only at time t_1 (deaths after time t_1 and before time t_2)
- $U_{0 h_2} =$ returns in stratum h_2 that file only at time t_2 (births after time t_1 and before time t_2)
- $U_{h_1 h_2} =$ 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. 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:

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

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

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

where π_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 $r_i \le \pi_{h_1} \le \pi_{h_2}$. 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 π_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:

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

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

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

Since $\delta_i(t_2)\delta_i(t_1) = 1$ only when a unit is in the sample for both time periods, for $E\left[\delta_i(t_2)\delta_i(t_1)\right] - E\left[\delta_i(t_2)\right]E\left[\delta_i(t_1)\right]$, 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:

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

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

$$T(t_1) = \sum_{h_1=1}^{H_1} \sum_{i \in U_{h_1 \bullet}} y_{1i} \text{ and } T(t_2) = \sum_{h_2=1}^{H_2} \sum_{i \in U_{\bullet h_2}} y_{2i} , \tag{3.1}$$

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 $n_{h_1 \bullet} = \sum_{i \in U_{h_1 \bullet}} \delta_i(t_1)$. Assuming that $n_{h_1 \bullet} > 0$, it can be shown that conditional on $\mathbf{n}_1 = \{n_{1\bullet}, n_{2\bullet}, \cdots, n_{H_1 \bullet}\}$, 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 PS estimator for $T(t_1)$ is

$$\hat{T}(t_1 | \mathbf{n}_1) = \sum_{h_1=1}^{H_1} \frac{N_{h_1 \bullet}}{n_{h_1 \bullet}} \sum_{i \in U_{h_1 \bullet}} \delta_i(t_1) y_{1i} . \tag{3.2}$$

Similarly, for year 2, we assume that $n_{\bullet h_2} = \sum_{i \in U_{\bullet h_2}} \delta_i(t_2) > 0$ and conditional on $\mathbf{n}_2 = \{n_{\bullet 1}, n_{\bullet 2}, \cdots, n_{\bullet H_2}\}$, the time t_2 sample design is a stratified simple random sample. For $N_{\bullet h_2}$ being the number of population elements in stratum h_2 at time t_2, the PS estimator for $T(t_2)$ is

$$\hat{T}(t_2|\mathbf{n}_2) = \sum_{h_2=1}^{H_2} \frac{N_{\bullet h_2}}{n_{\bullet h_2}} \sum_{i \in U_{\bullet h_2}} \delta_i(t_2) y_{2i} \; . \tag{3.3}$$

SOI uses $\hat{T}(t_1|\mathbf{n}_1)$ and $\hat{T}(t_2|\mathbf{n}_2)$ 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

$$D = T(t_2) - T(t_1) \; . \tag{3.4}$$

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

$$\hat{D} = \hat{T}(t_2|\mathbf{n}_2) - \hat{T}(t_1|\mathbf{n}_1) \, , \tag{3.5}$$

which is conditionally unbiased for the change in level. By breaking $\hat{T}(t_1|\mathbf{n}_1)$ into the sum of deaths for time t_1 and units in both years' samples summed over the year 1 strata and $\hat{T}(t_2|\mathbf{n}_2)$ 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

$$\begin{aligned}\hat{D} = &\sum_{h_2=1}^{H_2} \sum_{i \in U_{0 h_2}} \frac{N_{\bullet h_2} \delta_i(t_2) y_{2i}}{n_{\bullet h_2}} - \sum_{h_1=1}^{H_1} \sum_{i \in U_{h_1 0}} \frac{N_{h_1 \bullet} \delta_i(t_1) y_{1i}}{n_{h_1 \bullet}} \\ &+ \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} \sum_{i \in U_{h_1 h_2}} \left[\frac{N_{\bullet h_2} \delta_i(t_2) y_{2i}}{n_{\bullet h_2}} - \frac{N_{h_1 \bullet} \delta_i(t_1) y_{1i}}{n_{h_1 \bullet}} \right]\end{aligned} \tag{3.6}$$

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

$$R = \frac{T(t_2)}{T(t_1)} \text{ and } RD = \frac{T(t_2) - T(t_1)}{T(t_1)} \, , \tag{3.7}$$

which are estimated by

$$\hat{R} = \frac{\hat{T}(t_2|\mathbf{n}_2)}{\hat{T}(t_1|\mathbf{n}_1)} \text{ and } \widehat{RD} = \frac{\hat{T}(t_2|\mathbf{n}_2) - \hat{T}(t_1|\mathbf{n}_1)}{\hat{T}(t_1|\mathbf{n}_1)} . \tag{3.8}$$

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:

$$Var\left[\hat{T}(t_1)\middle|\mathbf{n}_1\right] = \sum_{h_1=1}^{H_1} \frac{N_{h_1 \bullet}^2}{n_{h_1 \bullet}} \left(1 - \frac{N_{h_1 \bullet}}{n_{h_1 \bullet}}\right) S_{h_1 \bullet}^2 \tag{4.1}$$

$$Var\left[\hat{T}(t_2|\mathbf{n}_2)\right] = \sum_{h_2=1}^{H_2} \frac{N_{\bullet h_2}^2}{n_{\bullet h_2}} \left(1 - \frac{n_{\bullet h_2}}{N_{\bullet h_2}}\right) S_{\bullet h_2}^2 \tag{4.2}$$

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

$$Var\left[\hat{D}\right] \approx \sum_{h_2=1}^{H_2} \frac{1-\pi_{h_2}}{\pi_{h_2}} N_{\bullet h_2} S_{\bullet h_2}^2 + \sum_{h_1=1}^{H_1} \frac{1-\pi_{h_1}}{\pi_{h_1}} N_{h_1 \bullet} S_{h_1 \bullet}^2 \\ -2\sum_{h_1=1}^{H_1}\sum_{h_2=1}^{H_2} \frac{\Delta_{h_1 h_2} N_{h_1 h_2} S_{h_1 h_2}}{\pi_{h_1}\pi_{h_2}}, \tag{4.3}$$

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

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 $\dfrac{\Delta_{h_1 h_2}}{\pi_{h_1}\pi_{h_2}} = \dfrac{1-\max\left(\pi_{h_1}, \pi_{h_2}\right)}{\max\left(\pi_{h_1}, \pi_{h_2}\right)}$ and replacing π_{h_1} with $\dfrac{n_{h_1 \bullet}}{N_{h_1 \bullet}}$ and π_{h_2} with $\dfrac{n_{\bullet h_2}}{N_{\bullet h_2}}$, 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):

$$Var\left[\hat{D}\middle|\mathbf{n}_1, \mathbf{n}_2\right] \approx \sum_{h_2=1}^{H_2}\left(1-\frac{n_{\bullet h_2}}{N_{\bullet h_2}}\right)\frac{N_{\bullet h_2}^2}{n_{\bullet h_2}} S_{\bullet h_2}^2 + \sum_{h_1=1}^{H_1}\left(1-\frac{n_{h_1 \bullet}}{N_{h_1 \bullet}}\right)\frac{N_{h_1 \bullet}^2}{n_{h_1 \bullet}} S_{h_1 \bullet}^2 \\ -2\sum_{h_1=1}^{H_1}\sum_{h_2=1}^{H_2}\left[\frac{1-\max\left(\dfrac{n_{h_1 \bullet}}{N_{h_1 \bullet}}, \dfrac{n_{\bullet h_2}}{N_{\bullet h_2}}\right)}{\max\left(\dfrac{n_{h_1 \bullet}}{N_{h_1 \bullet}}, \dfrac{n_{\bullet h_2}}{N_{\bullet h_2}}\right)}\right] N_{h_1 h_2} S_{h_1 h_2}. \tag{4.4}$$

For the population ratio in (3.7), the first-order Taylor series approximation (as in Nordberg (2000) and Wood (2008)) is $\hat{R} - R \approx \left[\hat{T}(t_2|\mathbf{n}_2) - R\hat{T}(t_1|\mathbf{n}_1)\right] / T(t_1|\mathbf{n}_1)$, which leads to the following approximate variance:

$$Var\left(\hat{R}\right) \approx \frac{Var\left(\hat{T}(t_2|\mathbf{n}_2)\right) + \hat{R}^2 Var\left(\hat{T}(t_1|\mathbf{n}_1)\right) - 2\hat{R} Cov\left(\hat{T}(t_1), \hat{T}(t_2|\mathbf{n}_2)\right)}{\left(\hat{T}(t_1|\mathbf{n}_1)\right)^2}, \tag{4.5}$$

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 $N_{h_1\bullet}, N_{\bullet h_2}$, and $N_{h_1h_2}$ are known, $\bar{y}_{h_1\bullet} = \frac{1}{n_{h_1\bullet}}\sum_{i\in U_{h_1\bullet}}\delta_i(t_1)y_{1i}$ and $\bar{y}_{\bullet h_2} = \frac{1}{n_{\bullet h_2}}\sum_{i\in U_{\bullet h_2}}\delta_i(t_2)y_{2i}$ are conditionally unbiased estimators of the strata population means. Since conditionally (and approximately unconditionally) unbiased estimators for the strata variances $S^2_{h_1\bullet}$ and $S^2_{\bullet h_2}$ are

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

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

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

$$var\left[\hat{T}(t_1|\mathbf{n}_1)\right] = \sum_{h_1=1}^{H_1}\left(1-\frac{n_{h_1\bullet}}{N_{h_1\bullet}}\right)\frac{N^2_{h_1\bullet}}{n_{h_1\bullet}}s^2_{h_1\bullet} \tag{5.1}$$

$$var\left[\hat{T}(t_2|\mathbf{n}_2)\right] = \sum_{h_2=1}^{H_2}\left(1-\frac{n_{\bullet h_2}}{N_{\bullet h_2}}\right)\frac{N^2_{\bullet h_2}}{n_{\bullet h_2}}s^2_{\bullet h_2} . \tag{5.2}$$

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 $Var\left[\hat{D}\right]$:

$$\begin{aligned} Var\left[\hat{D}\right] \approx& \sum_{h_2=1}^{H_2}\frac{1-\pi_{h_2}}{\pi^2_{h_2}}N_{\bullet h_2}S^2_{\bullet h_2} + \sum_{h_1=1}^{H_1}\frac{1-\pi_{h_1}}{\pi^2_{h_1}}N_{h_1\bullet}S^2_{h_1\bullet} \\ &-2\sum_{h_1=1}^{H_1}\sum_{h_2=1}^{H_2}\frac{\Delta_{h_1h_2}N_{h_1h_2}c_{h_1h_2}}{\pi_{h_1}\pi_{h_2}} \end{aligned} \tag{5.3}$$

where $c_{h_1h_2} = \frac{1}{\left(n_{h_1h_2}-1\right)}\sum_{i\in s_{h_1h_2}}\left(y_{2i}-\bar{y}_{\bullet h_2}\right)\left(y_{1i}-\bar{y}_{h_1\bullet}\right)$ is the *unweighted* covariance between the variable y-values for units in both years' samples. An alternative variance estimator based on (4.4) is

$$\begin{aligned} var\left[\hat{D}|\mathbf{n}_1,\mathbf{n}_2\right] \approx& \sum_{h_2=1}^{H_2}\left(1-\frac{n_{\bullet h_2}}{N_{\bullet h_2}}\right)\frac{N^2_{\bullet h_2}}{n_{\bullet h_2}}s^2_{\bullet h_2} + \sum_{h_1=1}^{H_1}\left(1-\frac{n_{h_1\bullet}}{N_{h_1\bullet}}\right)\frac{N^2_{h_1\bullet}}{n_{h_1\bullet}}s^2_{h_1\bullet} \\ &-2\sum_{h_1=1}^{H_1}\sum_{h_2=1}^{H_2}\left[\frac{1-\max\left(\frac{n_{h1\bullet}}{N_{h_1\bullet}},\frac{n_{\bullet h_2}}{N_{\bullet h_2}}\right)}{\max\left(\frac{n_{h1\bullet}}{N_{h_1\bullet}},\frac{n_{\bullet h_2}}{N_{\bullet h_2}}\right)}\right]N_{h_1h_2}c_{h_1h_2} \end{aligned} \tag{5.4}$$

For the ratio and relative difference estimators, the variance is also found by plugging in the sample-based estimators into expression (4.5):

$$var\left(\hat{R}\right) = var\left(\widehat{RD}\right)$$

$$\approx \frac{var\left(\hat{T}(t_2|\mathbf{n}_2)\right) + \hat{R}^2 var\left(\hat{T}(t_1|\mathbf{n}_1)\right) - 2\hat{R}cov\left(\hat{T}(t_1|\mathbf{n}_1), \hat{T}(t_2|\mathbf{n}_2)\right)}{\left(\hat{T}(t_1|\mathbf{n}_1)\right)^2}, \qquad (5.5)$$

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:

$$z_{1di} = \begin{cases} y_{1i} & \text{if } i \in d \text{ in year1} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad z_{2di} = \begin{cases} y_{2i} & \text{if } i \in d \text{ in year 2} \\ 0 & \text{otherwise} \end{cases}.$$

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: $DF = \min\left(n_{d1} - H_1, n_{d2} - H_2\right)$. 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 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 $n_{h_1h_2}$, $N_{h_1h_2}$ and $c_{h_1h_2}$ 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

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

Henry, K., Testa, V., and Valliant, R. (2008). "Variance Estimation for Estimators of Between-Year Change in Totals from Two Stratified Bernoulli Samples," *2008 Proceedings of the Section on Survey Research Methods*, American Statistical Association.

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

Internal Revenue Service (2006), Statistics of Income–2004 Individual Income Tax Returns, IRS, Publication 1304. Internal Revenue Service, Statistics of Income Bulletin, Winter 2008, Appendix A, "SOI Sampling Methodology and Data Limitations," pp. 149-151.

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.

Rivest, L.P. (1999), "Stratum Jumpers: Can We Avoid Them?" *1999 Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Shah, B.V., Folsom, R. E., LaVange, L.M, Wheeless, S. C., Boyle, K.E., and Williams, R. L. "Statistical Methods and Mathematical Algorithms Used in SUDAAN," Research Triangle Institute, 1993.

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

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

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.

Table 1. Partition of Universe at Two Times

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

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

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

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

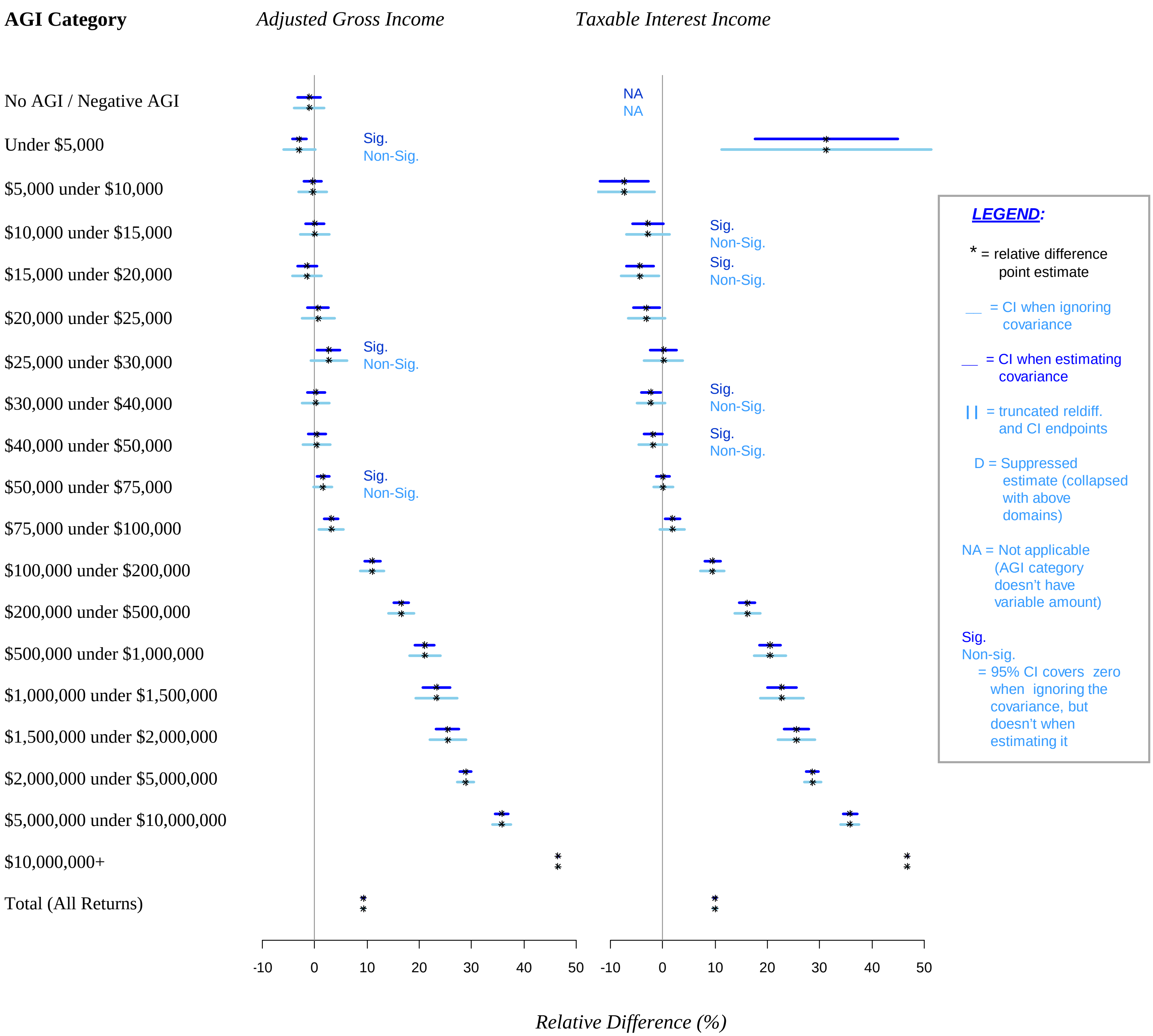

Figure 2. Confidence Intervals for Between-Year Relative Difference Estimates When Ignoring vs. Estimating the Covariance, by AGI Category, Total Income Tax and Business or Profession Net Income (Less Loss)

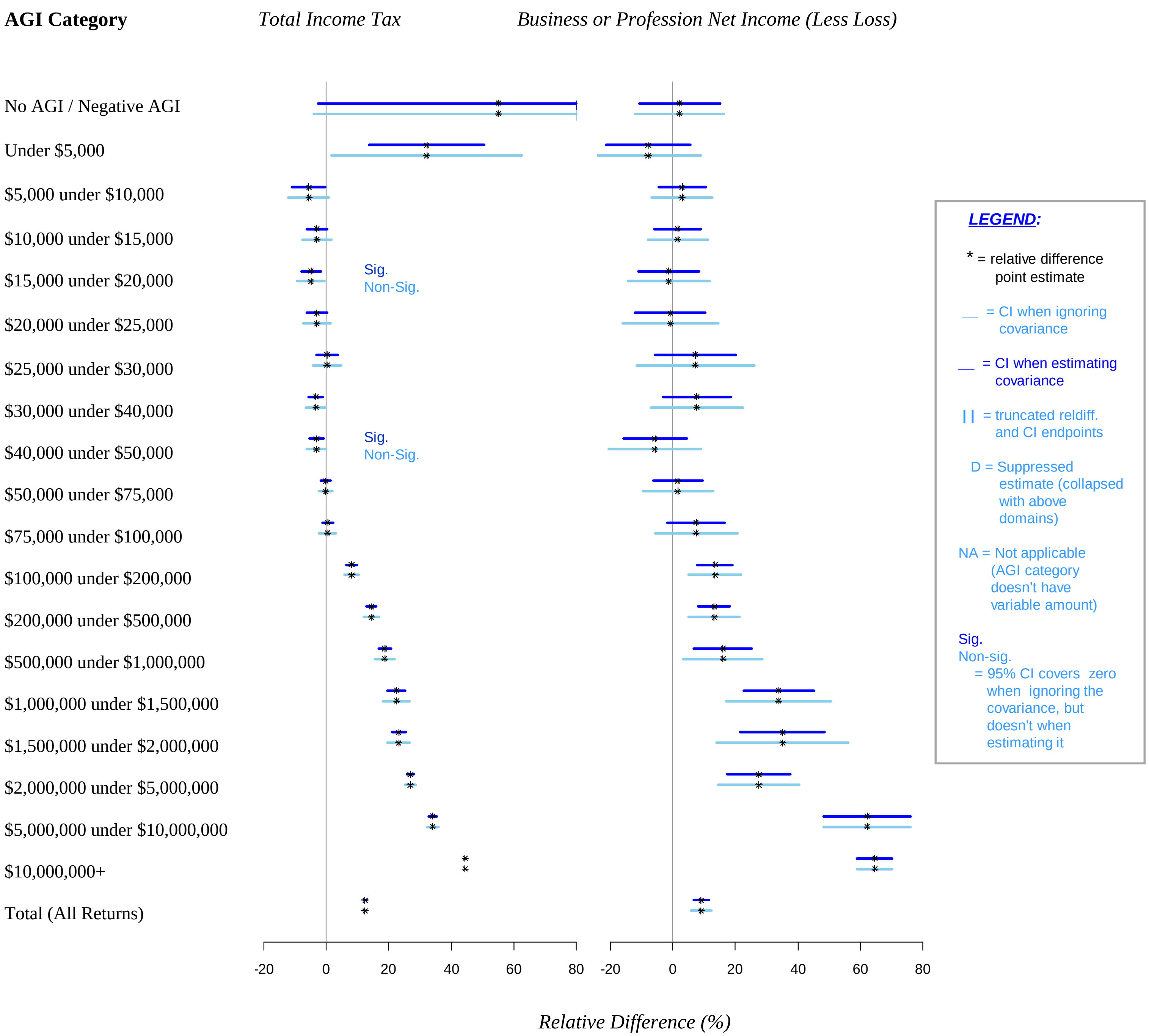

Figure 3. Confidence Intervals for Between-Year Relative Difference Estimates When Ignoring vs. Estimating the Covariance, by AGI Category, Alternative Minimum Tax and Net Capital Gain (Less Loss)

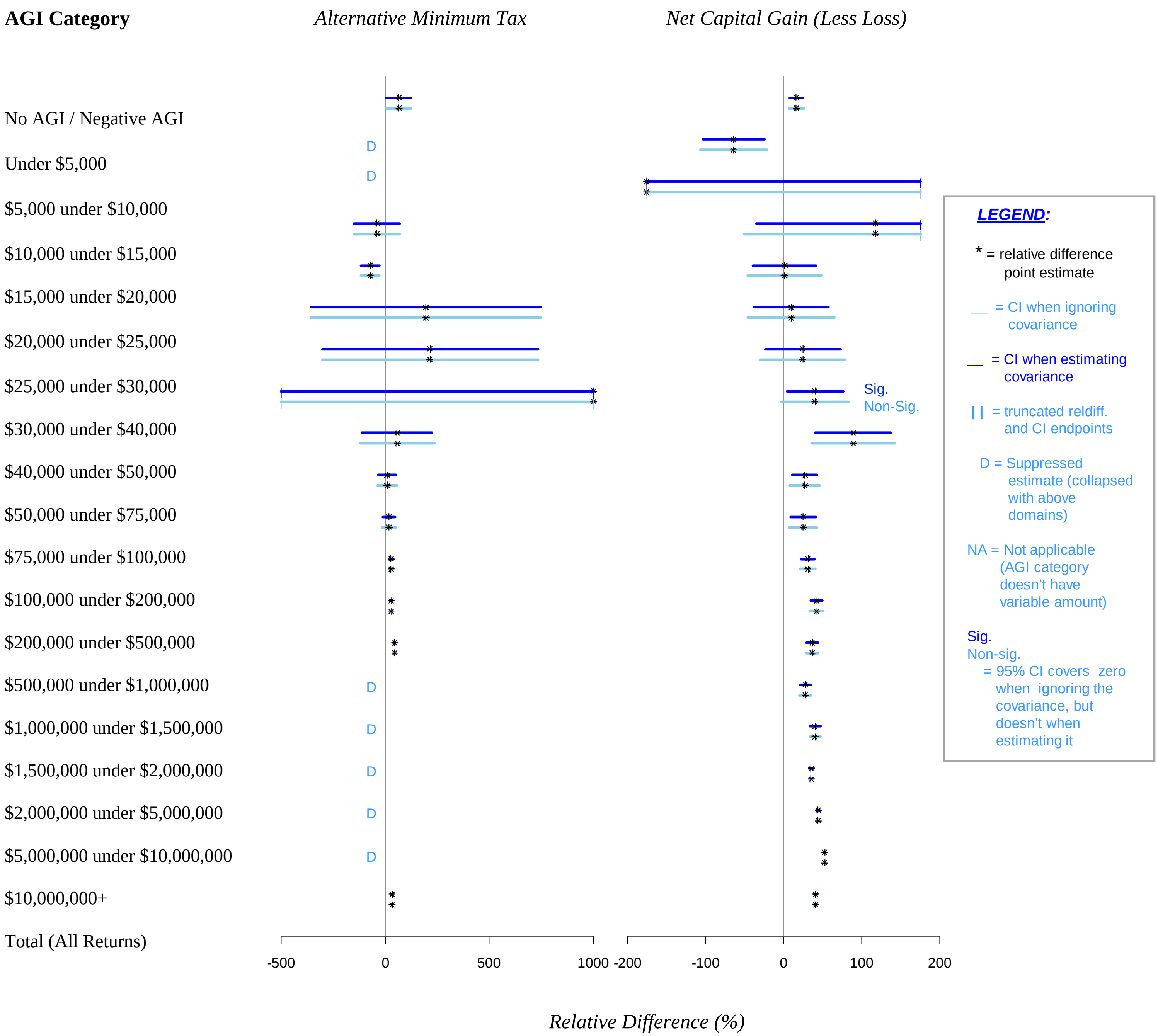

Figure 4. Confidence Intervals for Between-Year Relative Difference Estimates When Ignoring vs. Estimating the Covariance, by AGI Category, Charitable Contributions and Charitable Contributions Other Than Cash