Investigation of Variance Properties of Noise-Infused Estimates for the Survey of Business Owners (SBO)

Irene Brown[1], Marilyn Balogh[1], Anthony Caruso[1],
Beth Schlein[1], Katherine J. Thompson[1]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract
Beginning with the 2007 data release, the Survey of Business Owners (SBO) will employ random noise instead of cell suppression to perform disclosure avoidance processing. This paper reports the results of a simulation study conducted to assess the impact of noise infusion on the statistical properties of the SBO variance estimates. We present two alternative methods of estimating the additional variance component due to noise infusion, while using the SBO random group variance estimator to estimate the sampling variance component. We examine the coverage and bias properties of the alternative variance estimators for level estimates and percentage change estimates over repeated samples. As part of this analysis, we considered the impact of the prevalence of sensitive cells as well as the percentage of the total variance due to noise infusion. Our study showed that the effects on the variance estimates over repeated samples due to the addition of noise was negligible for the SBO estimates, due to the survey's large sampling variances.

Keywords: variance estimation, disclosure avoidance, noise-infusion, random group variance estimator

1. Introduction

The U.S. Census Bureau promises respondents confidentiality of data under Title 13 of the U.S. Code. For years, the Economic Directorate exclusively used cell suppression (with the p-percent rule) to achieve disclosure avoidance. The p-percent rule flags a cell as sensitive if p-percent of the top contributor's value is greater than the cell total minus the top *c* contributor's values. The primary disadvantage of cell suppressions is that a large percentage of data cells may be suppressed (unpublished). An alternative method, using noise to protect individual responses, outlined by Evans et al (1998) achieves the publication of more data cells without sacrificing the quality of the aggregate data while keeping respondents' data confidential. Beginning with the 2007 data release, the Survey of Business Owners (SBO) will employ random noise instead of cell suppression to perform disclosure avoidance processing.

The noise infusion methodology yields unbiased estimates, but increases the expected value of the variance estimates by adding an additional component to the total variance. This paper reports the results of research conducted to assess the impact of noise infusion on the statistical properties of the SBO variance estimates via a simulation study. We present two alternative methods of estimating this additional variance component caused

by noise infusion, while retaining the random group estimator employed by SBO to estimate the sampling variance component.

Using simulated data modeled from the SBO frame data, we examine the coverage and bias properties of the alternative noise infused variance estimators for level estimates and percentage change estimates over repeated samples. Because the random group variance estimator often requires a large number of samples to achieve unbiasedness, we examine the performance of the random group estimates over repeated samples without added noise as a baseline before examining the statistical properties of the two alternative complete variance estimators. In our analysis, we take the prevalence of sensitive cells as well as the percentage of the total variance due to noise infusion into consideration.

2. Background on SBO

The Survey of Business Owners and Self-Employed Persons (SBO) is part of the Economic Census, which the U.S. Census Bureau conducts, in years ending in "2" and "7." (www.census.gov/econ/sbo) The SBO supplies data users with estimates of total firm counts, receipts, payroll, and number of employees for businesses in the United States based on the race, gender, ethnicity, and veteran status of the majority business owners. The published data also include additional owner characteristics, such as age, education level, primary function of the business, type of business (inside or outside the home), type of customers and workers, and sources of financing for expansion, capital improvements and start-up costs. In addition to the totals listed above, for certain estimates, the SBO also provides estimates of percentage change from the prior period.

A new independent SBO sample is selected each data collection period using stratified systematic sampling. The sample frames divide the data into the following nine disjoint groups: American Indian, Asian, Black, Female, Hawaiian, Hispanic, Other, Publicly owned, and White Non-Hispanic. The sampling strata are defined by frame (one of the nine groups), state code, industry code (NAICS), and employer status. Companies that operate in multiple states are selected with certainty, along with companies whose payroll, receipts, or number of employees exceed stratum-specific size cutoffs. Otherwise, the companies are selected in each stratum systematically with a given probability after being sorted by the following: legal form of organization; likelihood of being male, female, or equally owned for sole proprietorship; and probability of belonging to the selected frame. Totals are Horvitz-Thompson estimates, using the inverse probability of selection as the sampling weight. A hot-deck imputation procedure adjusts for unit and item nonresponse. Variances are estimated via the random group methodology, with ten random groups. Race, gender, ethnicity, and veteran status data are not equally represented at all levels of estimates, leading to high estimates of variance in certain cases.

Prior to the 2007 data release, SBO used cell suppression for disclosure avoidance. Under the cell suppression method, the sensitivity of each cell depends on the distribution of the respondent values that are summed to form the cell value. Sensitive cells, defined by the p-percent rule, are suppressed, and additional cells (known as "complements") are suppressed to protect the sensitive cells. Given the multidimensional nature of published SBO data, complex linear programming methods are necessary to determine the placement of complementary suppressions. This complimentary cell suppression process requires significant programming resources and analyst review time, and publishable crosstabulations are limited by the capabilities of the linear programming software.

Approximately 27% of all published 2002 SBO cells were complementary suppressions. For the 2007 SBO, noise infusion is a promising alternative since far fewer cells will be suppressed and the limitations of complementary suppression do not apply.

3. Noise-infusion

The Evans-Zayatz-Slanta (EZS) noise method involves adding a predetermined amount of random "noise" to the micro data before tabulation. To do this, the establishment's data are multiplied by a factor that perturbs the data by a small percentage [Note: all establishments within the same company are perturbed in the same direction]. For example, if one were to perturb the reported data by about 10%, the noise adjustment factor would be close to 1.1 or .9. The "noise-infused" total for an establishment is obtained as

*Establishment value * [factor + (weight – 1)].*

The noisy establishment data are then summed up to get the cell total. The probabilistic model used to generate the noise adjustment factors must be symmetric about one. This achieves two objectives: (1) the expected value of noise being added to any cell is zero; and (2) in a given cell, there are equal expected amounts of establishments having positive amounts of noise and negative amounts of noise. These objectives accomplish the goal of minimizing the noise added to cells that are not at risk for disclosure. (Evans et al 1998)

The large sampling weights provide enough protection to avoid disclosure for the non-certainty units, so for SBO, the noise-factors will be added to **certainty units** only. SBO then calculates the noise-infused total as:

$$\hat{Y}_N = \sum_{i \in weight=1} (factor_i * y_i) + \sum_{i \in weight \neq 1} weight_i * y_i$$

All certainty companies (firm-level) in the SBO universe are randomly assigned a noise direction (either positive or negative). Each establishment is assigned a random noise factor ($1 \pm f$), where f is in $[a, b]$ and $1 \pm f$ is from a "split triangular" distribution. The factor (f) is random for each establishment, but the direction ($\pm$) is the same for all establishments in a company. Receipt, employment, and payroll values of each establishment are multiplied by the same noise factor. The SBO rounds estimates to the nearest 1,000 in each cell, which could potentially remove the effect of noise on small cases' values. To overcome this, based on research with 2002 SBO data, SBO increases or decreases each establishment data value by at least one unit. (Massell 2007)

The variance estimator for any **noise-infused** estimated total $\hat{Y}_N$ could be expressed as

$$\hat{v}_N(\hat{Y}_N) = \hat{v}_D(\hat{Y}) + \hat{v}_M(N)$$

where $\hat{v}_D(\hat{Y})$ is the sampling variance (obtained from **non-certainty** units only) of the original estimate $\hat{Y}$ and $\hat{v}_M(N)$ is the additional variance obtained by inducing noise. We estimate the first component using the method of random groups with ten assigned random groups to model SBO.

We evaluated two different estimators of the noise-infused variance component, $\hat{v}_M(N)$. The first estimator is derived by treating the noise infusion process as another stage of sampling. The first stage is the sample selection process and the second stage is the noise

infusion process. The variance of the noisy estimate is obtained by conditioning on both stages. Using the conditional variance identity

$$V(\hat{Y}_N) = V_1 E_2(\hat{Y}_N) + E_1 V_2(\hat{Y}_N),$$

the first component is approximated by the variance of the original (non-noisy) estimate ($V(\hat{Y})$). The estimate of the second component is

$$\hat{v}_M(N) = \sigma^2 * \sum_{i=1}^{n} y_i^2,$$

where σ^2 is the known variance of the probabilistic model used for noise assignment. For SBO, this is the variance of the split triangular distribution. Hereafter, we refer to this method as Variance Option 1.

The second estimator provided in Evans et al (1998) and derived under different assumptions is

$$\hat{v}_M(N) = (\hat{Y}_N - \hat{Y})^2.$$

Hereafter, we refer to this method as Variance Option 2. With Variance Option 2, the aggregate noise applied to each sample unit is viewed as a known bias added to the original estimate and the covariance between $\hat{Y}$ and $(\hat{Y}_N - \hat{Y})$ is assumed to be zero. Also, the variance of an estimate with little noise should be close to the original variance and the variance of an estimate with a lot of noise would be much larger.

There are several differences between Variance Option 1 and Variance Option 2. First, with the second variance estimator, there is a slight disclosure risk when $\hat{v}_D(\hat{Y}) = 0$, because the original estimate could be derived. In contrast, there is no similar disclosure risk with Variance Option 1. To derive Variance Option 1, no assumptions are made about the relationship between the added noise $(\hat{Y}_N - \hat{Y})$ and $\hat{Y}$. With Variance Option 2, it is impossible to develop an upper bound on the added variance due to noise. But with Variance Option 1, the added variance due to noise has an upper bound of

$$\hat{v}_M(N) <= \sigma^2 * n * y_i^2,$$

where y_i is the largest value for an item among all the establishments. A more detailed proof of each variance option is found in the appendix.

4. Simulation Study

4.1 Design

We modeled SBO populations from available 2002 and 2007 sampling frame data in selected states chosen by subject-matter experts. The five states modeled were New York (large number of firms with great diversity of race and gender and a small percentage of sensitive cells), Utah (small number of firms with little diversity and a large percentage of sensitive cells), Pennsylvania (large number of firms with moderate diversity and an average amount of sensitive cells), Georgia (mid-size number of firms with moderate diversity and an average amount of sensitive cells), and Missouri (mid-size number of firms with some diversity and an average amount of sensitive cells).

To obtain simulated (2002 and 2007) population frames, we modeled percentages of race, gender, ethnicity and publicly held ownership from the 2002 weighted response data

proportions. Receipt values were not simulated: the totals are computed from administrative data found on the sampling frame. SBO performs a hot-deck imputation procedure to correct for unit and item non-response. For this study, we ignored this imputation component, recognizing that our simulated variances will consequently be underestimates.

We selected 5,000 repeated stratified systematic samples from each population (2002 and 2007) using the SBO stratification, allocation, and sampling design. Within each sample, we independently applied noise to establishments with sample weight equal to one using the SBO methodology stated in Section III. For this study, we restricted the analysis to variances of total estimates of firm counts and receipts and their respective percent change comparisons. We added noise to firm counts to obtain a second set of estimates, that the sample is designed to obtain, for our analysis. It should be noted that this would not be done for the actual SBO publication.

We used the two variance estimation options presented in Section III to compute variances of the noise-infused estimates. To distinguish between variance effects due to noise infusion and variance effects due to sampling variance alone, we computed variances for the original estimates along with the noise-infused counterparts. Since the SBO is introducing noise in 2007, the first set of percent change comparisons compares current noise-infused estimates to prior original estimates. After 2007, SBO percent change estimates will include noise in both years. Thus, we consider three types of percent change estimates: original to original (baseline); original to noise-infused (mimicking the 2002/2007 production setting); and noise-infused to noise-infused (after 2007).

The following tables (Tables 1 and 2) give a description and subscript notation of estimates ($\widehat{\theta}_{Cis}$) and variance estimates ($\hat{v}_F(\widehat{\theta}_{Cis})$) used in the simulation study. The *C* subscript indexes the type of estimate, the *F* subscript indexes the type of variance estimate, *i* is the tabulation level and *s* is the sample.

Table 1: Values of "Type of Estimate" Indices Used in Simulation Study

C	Description
1	2002 Firm counts original
2	2002 Firm counts noise infused
3	2007 Firm counts original
4	2007 Firm counts noise infused
5	Firm counts change from original to original
6	Firm counts change from original to noise infused
7	Firm counts change from noise-infused to noise infused
8	2002 Total Receipts original
9	2002 Total Receipts noise infused
10	2007 Total Receipts original
11	2007 Total Receipts noise infused
12	Total Receipts change from original to original
13	Total Receipts change from original to noise infused
14	Total Receipts change from noise infused to noise infused

Table 2: Values of "Variance Estimator" Indices Used in Simulation Study

F subscript	Description
0	The sampling variance of the original estimate
1	The variance of noised infused estimate; Option 1
2	The variance of noised infused estimate; Option 2

In all 5,000 samples, we computed all fourteen types of estimates. We obtained empirical mean square errors (MSE) for each type of estimate as

$$MSE(\hat{\theta}_{Ci}) = \left(\frac{1}{5000}\right)\sum_{s=1}^{5000}(\hat{\theta}_{Cis} - \theta_{Ci})^2,$$

where θ_{Ci} is the corresponding population value for each estimate $\hat{\theta}_{Ci}$. In 1,000 of the 5,000 samples, we calculated all three types of variances per estimate. This yielded twenty-two variances in all per tabulation level to review.

4.2 Evaluation Criteria

We computed the following statistics to evaluate the properties of the variance estimators over repeated samples:

- Relative Bias - the ratio of a variance estimate over repeated samples to the empirical mean square error minus one, computed as

$$RB_F(\hat{\theta}_{Ci}) = \frac{1/1000\sum_{s=1}^{1000}\hat{v}_F(\hat{\theta}_{Cis})}{MSE(\hat{\theta}_{Ci})} - 1$$

- Coverage – the percentage of 90% confidence intervals (constructed using a t-statistic with 9 d.f.) that contain the true population estimate, calculated as

100% - (Lower error rate + Upper error rate)

Lower error rate: The percentage of estimates where the population total is **less** than the lower bound of the 90% confidence interval.

Upper error rate: The percentage of estimates where the population total is **greater** than the upper bound of the 90% confidence interval.

The "ideal" variance estimator will have relative bias near zero and coverage rates near 90-percent.

5. Results

5.1 State Characteristics

The five states were chosen to study the impact of the prevalence of sensitive cells. Table 3 presents summary statistics on the percentage of sensitive cells statewide by year for total receipts, determined independently within each sample. In this analysis, cell *i* is considered sensitive if it would be a primary suppression using the p-percent rule. In Table 3, we see Utah has the largest percentage of sensitive cells and New York has the smallest. (Tested at alpha=0.05)

Table 3: Percentages of Sensitive Cells In Simulated Populations

State:	2002 Total Receipts	2007 Total Receipts
New York	3.00%	4.80%
Utah	6.04%	12.60%
Pennsylvania	5.08%	8.18%
Georgia	5.16%	7.90%
Missouri	5.93%	11.53%

If the sampling variance is quite high, then it is likely that the additional variance component resulting from noise infusion may not have much effect on the statistical properties of the variance estimates. To examine this, we calculated the overall percentage of variance due to noise infusion for 2002 and 2007 total receipts by state using the 1,000 samples and all tabulation levels. Table 4 presents these percentages.

Table 4: Percentage of Variance due to Noise Infusion

	2002 Receipt Totals		2007 Receipt Totals	
State	Variance Option 1	Variance Option 2	Variance Option 1	Variance Option 2
New York	11.7%	12.2%	22.9%	21.2%
Utah	21.6%	21.2%	31.5%	29.9%
Pennsylvania	15.2%	16.2%	28.3%	26.5%
Georgia	15.4%	16.3%	27.2%	25.2%
Missouri	19.9%*	20.0%*	30.5%	28.5%

* Variance Option 1 is not significantly different from Variance Option 2

Not surprisingly, the variance estimates in Utah have the highest proportion total variance due to noise infusion, and the variance estimates in New York have the least. (Tested at alpha=0.05) There was no significant difference between Pennsylvania and Georgia for 2002 receipt totals. For Utah, Variance Option 1 (based on the known split triangular distribution combined with the squared total estimate) consistently yielded a larger percentage of variance due to noise than Variance Option 2 (the Evans et al estimator). The other states did not show this pattern. The percentage of variance due to noise infusion is on average small. Because of the inconsistency between the results for 2002 and 2007, we cannot say if either variance estimator option is consistently smaller than the other.

5.2 Relative Bias

To assess the bias properties of Variance Options 1 and 2, we performed a separate analysis of each variance estimator by type of estimate. This was a two-step process. First, we computed relative bias of each variance estimator for a given tabulation. Then, we used two-sided sign tests to test for overall unbiasedness, under the null hypothesis that the median value of the relative bias of a variance estimator (over all tabulation levels) is zero. We performed a total of twenty-two sign tests per state.

For each of the two variance estimator options, we performed eight tests:

- Totals – four tests (firms and receipts, 2002 and 2007 data)
- Percent change – four tests (firms and receipts, original/noise-infused and noise-infused/noise-infused)

For the original variance estimator (sampling variance only), we performed six tests:

- Totals – four tests (firms and receipts, 2002 and 2007 data)
- Percent change – two tests (firm and receipts, original/original).

The null hypothesis was rejected for the majority of the sign tests providing evidence that the variance estimates were biased.

Although the sign test provides evidence of the **existence** of bias, it does not examine the magnitude or direction of the bias and cannot be used to evaluate the impact of the bias on coverage rates. To examine the direction of the bias analytically, we examined histograms of relative bias for all variance estimators over all tabulation levels to obtain some anecdotal information on the direction of the bias. The graphical analysis provided indications of positively biased variance estimates for the majority of total estimates, and negatively biased variance estimates for change estimates. This pattern was the same across states and for all variance estimators.

Since both Options 1 and 2 yielded biased variance estimates, we used Wilcoxon signed rank tests to determine whether one variance estimator was systematically less biased than the other. We tested a one-sided null hypothesis that the relative bias of Variance Option 1 is greater than or equal to the relative bias of Variance Option 2, i.e.

$$H_0 : RB_1(\hat{\theta}_{Ci}) >= RB_2(\hat{\theta}_{Ci})$$
$$H_1 : RB_1(\hat{\theta}_{Ci}) < RB_2(\hat{\theta}_{Ci}) .$$

Approximately 10% of all tests were significant, as expected at an alpha level of 10%. We concluded that neither variance estimator option induced higher relative bias.

5.3 Coverage Rates

To analyze the statistical properties of the variance estimators on confidence interval coverage, we restricted the analysis to cells with unbiased estimates over repeated samples. This eliminated the confounding effects on coverage caused by using both a biased estimator and a biased variance estimator. We used normal tests (*z*-tests) to test for estimate bias. For estimates of total, we tested:

$$\mathrm{H}_0\text{: } \overline{\hat{\theta}}_{Ci} = \theta_{Ci}$$
$$\mathrm{H}_1\text{: } \overline{\hat{\theta}}_{Ci} \neq \theta_{Ci}$$

where $\overline{\hat{\theta}}_{Ci}$ is the average estimate of the characteristic *C* in cell *i* from the 5000 samples, and θ_{Ci} is the population value. The test statistic is

$$z = \frac{(\overline{\hat{\theta}}_{Ci} - \theta_{Ci})}{\sqrt{MSE(\overline{\hat{\theta}}_{Ci})/5000}} .$$

Under H_0, $z \sim N(0,\sigma^2)$. Reject H_0 if $|z| \geq z_{(.95)}$.

For change estimates, we tested

$$\mathrm{H}_0\text{: } \overline{\hat{\theta}^t_{Ci} - \hat{\theta}^{t-1}_{Ci}} = \theta^t_{Ci} - \theta^{t-1}_{Ci}$$
$$\mathrm{H}_1\text{: } \overline{\hat{\theta}^t_{Ci} - \hat{\theta}^{t-1}_{Ci}} \neq \theta^t_{Ci} - \theta^{t-1}_{Ci}$$

The test statistic is

$$z = \frac{((\overline{\hat{\theta}_{Ci}^{t} - \hat{\theta}_{Ci}^{t-1}}) - (\theta_{Ci}^{t} - \theta_{Ci}^{t-1}))}{\sqrt{MSE(\overline{\hat{\theta}_{Ci}^{t} - \hat{\theta}_{Ci}^{t-1}})/5000}}.$$

Under H_0, $z \sim N(0,\sigma^2)$. Reject H_0 if $|z| \geq z_{(.95)}$.

We constructed three sets of 90% confidence intervals for all unbiased estimates (one per variance estimator option). Then, we tested each coverage rate to determine whether it was different from the expected 90%, using the normal approximation of a binomial distribution with n=1000 (n = number of independent samples/trials). This gave us an overall proportion of non-nominal (different from 90%) coverage rates by state, characteristic, and variance option for each state and tabulation level. A majority of the coverage rates were non-nominal in all states.

We also examined whether the non-nominal confidence intervals were systematically too narrow (undercoverage) or too wide (overcoverage). All states showed the same pattern: systematic overcoverage for firm counts estimates of change and undercoverage for receipt totals. The pattern holds regardless of variance estimator.

To examine the "practical impact" of imperfect coverage (i.e., how far are the non-nominal confidence intervals from the optimal 90%), we created histograms of the coverage rates by state, estimate characteristic *C*, and variance option showing the median and the interquartile range (IQR). This histogram and univariate analysis is summarized in Table 5. Coverage is essentially the same for noise-infused and original estimates, except adding noise improves the coverage for receipt totals. Coverage rates are "close" to 90% for firm counts, which is expected given the sample is designed for firm counts.

Table 5: Summary of Histogram and Univariate Analysis of Coverage Rates by State

State	Variance Options	Total		Change	
		Firms	Receipts	Firms	Receipts
New York	Sampling Variance	Median ≈ 91 Narrow IQR	Median ≈ 85 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 90 Wide IQR
	Variance Option 1	Median ≈ 91 Narrow IQR	Median ≈ 88 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 90 Wide IQR
	Variance Option 2	Median ≈ 91 Narrow IQR	Median ≈ 88 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 90 Wide IQR
Utah	Sampling Variance	Median ≈ 90 Narrow IQR	Median ≈ 84 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 87 Wide IQR
	Variance Option 1	Median ≈ 90 Narrow IQR	Median ≈ 87 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 89 Wide IQR
	Variance Option 2	Median ≈ 90 Narrow IQR	Median ≈ 87 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 89 Wide IQR
Pennsylvania	Sampling Variance	Median ≈ 91 Narrow IQR	Median ≈ 84 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 88 Wide IQR
	Variance Option 1	Median ≈ 91 Narrow IQR	Median ≈ 88 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 90 Wide IQR
	Variance Option 2	Median ≈ 91 Narrow IQR	Median ≈ 88 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 90 Wide IQR

State	Variance Options	Total		Change	
		Firms	Receipts	Firms	Receipts
Georgia	Sampling Variance	Median ≈ 90 Narrow IQR	Median ≈ 87 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 90 Wide IQR
	Variance Option 1	Median ≈ 90 Narrow IQR	Median ≈ 90 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 90 Wide IQR
	Variance Option 2	Median ≈ 90 Narrow IQR	Median ≈ 90 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 90 Wide IQR
Missouri	Sampling Variance	Median ≈ 90 Narrow IQR	Median ≈ 84 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 87 Wide IQR
	Variance Option 1	Median ≈ 90 Narrow IQR	Median ≈ 88 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 89 Wide IQR
	Variance Option 2	Median ≈ 90 Narrow IQR	Median ≈ 88 Wide IQR	Median ≈ 92 Narrow IQR	Median ≈ 89 Wide IQR

6. Conclusion

For SBO, sampling variance accounted for the majority of the variance for a noise-infused estimate. The relative bias and coverage rates of unadjusted (original) estimates exhibit the same patterns and are essentially the same as the corresponding adjusted (noise-infused) estimates, except for the slight improvement in coverage rates for estimates of Receipt Totals. In the majority of the cases, although the coverage is non-nominal, the coverage rates are quite close to 90%. Based on these results, we concluded that the addition of noise did not impact the statistical properties of the SBO variance estimators. Due to our findings of biased variance estimates and non-nominal coverage rates, which arose from using the random group (sampling) variance estimator, SBO may wish to investigate alternate methods of calculating the sampling variance.

Our empirical research showed no advantage of one method of estimating variance for noise-infused estimates over the other. The results showed no statistical difference between the options and computing time was similar. Therefore, we present that there are two alternative variance estimators for noise-infused estimates: (1) Variance Option 1, which was derived by us; and (2) Variance Option 2 that is found in the noise infusion literature. Although, we do recommend using Variance Option 1 because it poses no disclosure risk. Moreover, an upper bound on the added variance due to noise can be calculated, and this information can be built into the sample design stage to obtain optimal allocations.

This study was carefully modeled to simulate the SBO. The data may or may not be representative of other programs. Moreover, we consider only one type of survey design (stratified systematic sampling), one type of sampling variance estimator (random groups), and one variation of the split-triangular probability distribution when assigning noise. Consequently, we do not recommend making inferences for other surveys based on these analyses. In particular, we caution against drawing conclusions about the impact of noise-infusion on variance estimates for other programs whose sampling variance component is not very large, as it is with SBO. We recommend a similar empirical analysis for other programs considering using these same methods of noise-infusion.

Acknowledgments

We acknowledge all the hard work, staff from the Company Statistics division at the U.S Census Bureau, did in working on this simulation study. Also, the staff from the Commodity Flow Survey for their added imputes. We also acknowledge Rita Petroni and Richard Moore Jr. for their review and comments on the paper.

References

Caldwell, C. (2007). "Application of Protective Noise for Disclosure Avoidance for the 2007 Economic Census Programs in Company Statistics Division." Internal memo, U.S. Census Bureau.

Caruso, A. (2007), "Using Protective Noise for Disclosure Avoidance with 2007 SBO Data." Internal memo, U.S. Census Bureau.

Casella, G. and Berger, R. L. (2001), Statistical Inference (Second Edition). Duxbury Press.

Conover, W. J. (1980), Practical Nonparametric Statistics (2nd ed.). New York: Wiley.

Evans, T., Zayatz, L., and Slanta, J. (1998). Using Noise for Disclosure Limitation of Establishment Tabular Data. *Journal of Official Statistics*, **4**, pp. 537-551.

Massell, P. and Funk, J. (2007), "Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding that Preserves Protection," *Proceedings of the Research Conference of the Federal Committee on Statistical Methodology*, Arlington, Virginia, November 5-7, 2007.

Schlein, B. "2002 SBO Estimation." Internal memo, U.S. Census Bureau.

Schlein, B. "Phase II Sample Selection for 2007 SBO." Internal memo, U.S. Census Bureau.

Wolter, K. M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Appendix

The Noise-Infused Estimate is

$$\hat{Y}_N = \sum [\, factor_i + (wgt_i - 1)] * y_i \,,$$

where y_i is the variable of interest in the survey,

wgt_i is the sampling weight,

$factor_i$ is a random variable from a split triangular distribution (see Figure 1) with $E(\, factor) = 1$ and $V(\, factor) = \sigma^2$.

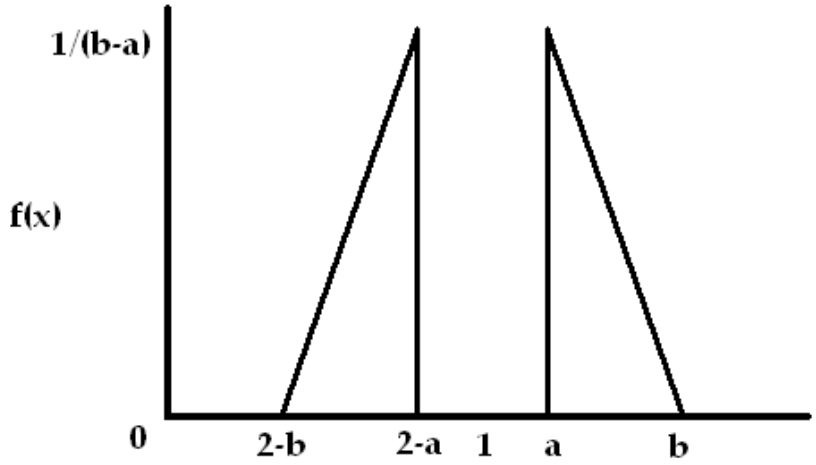

Figure 1: Split triangular distribution

Variance Option 1 (Derived by Brown et al):

This option treats the noise infusion process as another stage of sampling. The first stage is the sample selection process and the second stage is the noise infusion process.

The expected value of the noise-infused estimate is found using the conditional expectation identity:

$$E(\hat{Y}_N) = E_1E_2(\hat{Y}_N) = E\left(E\left(FACTOR\middle|Y\right)\right)$$
$$= E_1E_2\left(\sum[\,factor_i + (wgt_i - 1)] * y_i\right)$$
$$= E_1\left(\sum(E_2(\,factor_i) * y_i - y_i) + \sum(wgt_i * y_i)\right)$$
$$= E_1\left(\sum(y_i - y_i) + \sum(wgt_i * y_i)\right)$$
$$= E_1(\hat{Y}) = Y$$

The variance of the noise-infused estimate is found using the conditional variance identity:

$$V(\hat{Y}_N) = V_1E_2(\hat{Y}_N) + E_1V_2(\hat{Y}_N) = V(E(FACTOR|Y)) + E(V(FACTOR|Y))$$
$$= V_1(\hat{Y}) + E_1V_2\left(\sum[\,factor_i + (wgt_i - 1)] * y_i\right)$$
$$= V_1(\hat{Y}) + E_1V_2\left(\sum(\,factor_i * y_i)\right)$$
$$= V_1(\hat{Y}) + E_1\left(\sum(y_i^2 * V_2(\,factor_i))\right)$$
$$= V_1(\hat{Y}) + E_1\left(\sum y_i^2 * \sigma^2\right)$$
$$= V(\hat{Y}) + (\sigma^2 * n * E(Y^2))$$

$(\sum_{i=1}^{n} y_i^2)/n$ is an unbiased estimate of $E(Y^2)$, so $\sum_{i=1}^{n} y_i^2$ is an unbiased estimate of $n * E(Y^2)$.

Therefore,

$$\hat{v}_1(\hat{Y}_N) = \hat{v}_S(\hat{Y}) + \sigma^2 * \sum_{i=1}^{n} y_i^2 ,$$

where $\hat{v}_S(\hat{Y})$ is the estimated variance of the original estimate and σ^2 is the variance from the split triangular distribution.

Variance Option 2 (derived by Evans et al):

The noise-infused estimate is seen as:

$$\hat{Y}_N = \hat{Y} + \varepsilon ,$$

where $\hat{Y} = \sum wgt_i * y_i$ (Original estimate),

$\varepsilon = \sum (factor_i - 1) * y_i$ (The aggregated noise that is applied to sample units).

Based on the probability distribution chosen to get $factor_i$, ε has the following properties:

$E(\varepsilon) = 0$

$V(\varepsilon) = E(\varepsilon^2) - (E(\varepsilon))^2 = E(\varepsilon^2)$

and we assume $COV(\hat{Y}, \varepsilon) = 0$.

The expected value of the noise-infused estimate is then:

$$E(\hat{Y}_N) = E(\hat{Y} + \varepsilon) = E(\hat{Y}) + E(\varepsilon) = E(\hat{Y}) = Y$$

The Variance of the noise-infused estimate is then:

$$V(\hat{Y}_N) = V(\hat{Y} + \varepsilon) = V(\hat{Y}) + V(\varepsilon) + 2 * COV(\hat{Y}, \varepsilon)$$
$$= V(\hat{Y}) + E(\varepsilon^2)$$

An expected value of a function of estimates ($\varepsilon = \hat{Y}_N - \hat{Y}$) is approximately equal to the function of the estimates. (i.e. $E(\varepsilon^2) \approx (\hat{Y}_N - \hat{Y})^2$).

Therefore,

$$\hat{v}_2(\hat{Y}_N) = \hat{v}_S(\hat{Y}) + (\hat{Y}_N - \hat{Y})^2 ,$$

where $\hat{v}_S(\hat{Y})$ is the estimated variance of the original estimate, $\hat{Y}_N$ is the noise-infused estimate, and $\hat{Y}$ is the original estimate.