Accounting for group classification error in variance estimates using the American Community Survey [*]

Matthew W. Brault
U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Prepared for the 2014 ACS Data Users Conference
Washington, DC, May 29-30, 2014
Updated: July 3, 2014[†]
SEHSD Working Paper Number 2014-18

An advantage of the American Community Survey is that its large sample size gives researchers the ability to categorize groups of individuals based on some secondary characteristic. For instance, one could classify people as living in 'high' or 'low' poverty areas based on the poverty rate of the census tract in which they live. Because this assignment is based on survey data, it is not without sampling error. When talking collectively of groups that share the characteristic, the concept is more nebulous than a simple aggregation and measures of uncertainty should include the error associated with making the group classification. Often the variance from this classification error can be many times that of the sampling variance in the final estimate. Using poverty areas as an example, I show how this error can be calculated and incorporated into the variance of estimates using the 2006-2010 American Community Survey. Based on my recalculations, the standard errors with the classification error were often 10 to 50 times as large as those without. Under a total error variance framework, both the Census Bureau and data users should be aware of this source of uncertainty and, where necessary, incorporate it into their estimates.

Introduction

A major advantage to using the American Community Survey (ACS) is the large sample of cases available to produce statistics for population subgroups defined by geographic, household, or individual characteristics. The ACS is the only sources of social and economic data for geographies down to the census tract level, but also one of the few data collection efforts capable of describing relatively rare populations such as specific race, ethnic and ancestry groups, and detailed industries and occupations.

Having consistently defined measures across these groups makes classifying them along those lines possible. Foreign-born ethnic groups could be classified by median year of entry to define "recent" versus "past" entrants. Languages could be classified by the degree of linguistic isolation, industries by the percentage of workers with employer-provided health insurance, or counties by… just about any number of characteristics. Defining the subgroups by another sample characteristic, however, introduces an additional source of error that must be accounted for in estimates of variance.

[*] This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily that of the U.S. Census Bureau.
[†] This update reflects comments from conversations with data users outside of the U.S. Census Bureau.

General Case

Consider the following. A survey contains some characteristic (census tracts, occupations, etc.) for which there are *K* categories. We wish to classify those categories into a schema based on a secondary characteristic θ (poverty rate, median earnings, etc.). Since θ is obtained from sample data, it is subject to sampling error and so the estimate for each category k is θ_k with standard error σ_k. The *K* categories are classified by making comparisons of each θ_k against some critical value τ. If θ_k is greater than τ then some classification indicator (y_k) is set to 1, otherwise it is set to 0. Our final statistic is then a measure of what percentage of the population belong to category $K = k$ with a θ greater than τ.

Often researchers will treat the classification of *y* as truth and assume there is no error associated with the classification, or that the error is minimal and thus ignorable. This paper will show that this assumption is naïve and that the magnitude of the error is significant. Ignoring it greatly overestimates the certainty of the final estimate. Using a total error variance framework, the uncertainty associated with the classification is treated as a form of measurement error and incorporated as such.

Calculating this type of measurement error is relatively easy in the ACS using the successive differences replication (SDR) method for calculating variance (U.S. Census Bureau, 20009). Consider for each category $K = k$, the statistic θ_k can be described as the function $m(z_i, W_i)$ where z_i is a vector of survey characteristics, W_i is the survey weight for the survey respondent, and i denotes a member of group $K = k$. For a simple example, z_i is a single variable and θ_k is the weighted mean value of z, such that $m(z_i, W_i) = \sum(W_i \cdot z_i) / \sum W_i$. Using replicate weights $W_{i,r}$ where $r = 0 \ldots 80$, ($W_{i,0}$=Production Weight), then replicate estimates $\theta_{k,r}$ can be determined by $m(z_i, W_{i,r})$. Each replicate estimate can then be compared to the critical value τ to create replicate classification indicators $y_{k,r}$.

Under the naïve assumption, $y_{k,0} = y_k$ is the only indicator used because the estimate for that group is treated as truth. The variance of the final estimate ($\tilde{Y}_0 = Y_0 = Y$) is the weighted mean value of y_k across all *k* categories such that:

$$\widetilde{Y_r} = \frac{\sum_i (y_k W_{i,r})}{\sum_i W_{i,r}}$$

$$\widetilde{var}(Y) = \frac{\sum_R (Y_0 - \widetilde{Y_r})^2}{R(1-\varepsilon)^2}$$

Taking the uncertainty of the classification into account, we allow the indicators to vary along with the replicate weights. Using the separately calculated indicators ($y_{k,r}$), the variance of the final estimate becomes:

$$Y_r = \frac{\sum_i (y_{k,r} W_{i,r})}{\sum_i W_{i,r}}$$

$$var(Y) = \frac{\sum_R (Y_0 - Y_r)^2}{R(1-\varepsilon)^2}$$

An error associated with classification can be described as the difference ($\delta_{k,r}$) between each replicate indicator and the indicator from the production weight ($\delta_{k,r} = y_{k,0} - y_{k,r}$). By definition, the deviation is zero for the production weight as the naïve and sophisticated estimates are the same in this scenario. However, the classification error is the variance of this deviation, as calculated by the SDR method, because the uncertainty comes from sampling error:

$$\Delta_r = \frac{\sum_i (\delta_{k,r} W_{i,r})}{\sum_i W_{i,r}}$$

$$var(\Delta) = \frac{\sum_R (\Delta_0 - \Delta_r)^2}{R(1-\varepsilon)^2}$$

It can also be shown that the weighted mean of this difference across groups is equal to the difference between the replicate estimates from the naïve and sophisticated calculations.

$$\Delta_r = \frac{\sum_i (\delta_{k,r} W_{i,r})}{\sum_i W_{i,r}} = \frac{\sum_i \left((y_{k,0} - y_{k,r}) W_{i,r} \right)}{\sum_i W_{i,r}} = \frac{\sum_i (y_{k,0} W_{i,r})}{\sum_i W_{i,r}} - \frac{\sum_i (y_{k,r} W_{i,r})}{\sum_i W_{i,r}} = \widetilde{Y_r} - Y_r$$

Decomposing the sophisticated variance calculation, we can see that the sophisticated calculation is the sum of the naïve calculation of variance for the estimate, error variance associated with classification, and the sampling covariance between the two (as they are derived from the same sample).

$$\begin{aligned} var(Y) = \frac{\sum_R (Y_0 - Y_r)^2}{R(1-\varepsilon)^2} &= \frac{\sum_R \left(Y_0 - \widetilde{Y_r} + \widetilde{Y_r} - Y_r \right)^2}{R(1-\varepsilon)^2} \\ &= \frac{\sum_R \left[\left(Y_0 - \widetilde{Y_r} \right)^2 + \left(\widetilde{Y_r} - Y_r \right)^2 + 2 \left(Y_0 - \widetilde{Y_r} \right) \left(\widetilde{Y_r} - Y_r \right) \right]}{R(1-\varepsilon)^2} \\ &= \frac{\sum_R \left(Y_0 - \widetilde{Y_r} \right)^2}{R(1-\varepsilon)^2} + \frac{\sum_R \left(\widetilde{Y_r} - Y_r \right)^2}{R(1-\varepsilon)^2} + 2 \frac{\sum_R \left(Y_0 - \widetilde{Y_r} \right) \left(\widetilde{Y_r} - Y_r \right)}{R(1-\varepsilon)^2} \\ &= \widetilde{var}(Y) + var(\Delta) + 2cov(Y, \Delta) \end{aligned}$$

In the latter half of this paper, I will show that the error variance associated with classification is quite large in comparison to the naïve variance, making it non-ignorable, using the Census Bureau brief *Areas With Concentrated Poverty: 2006-2010* (Bishaw, 2011) as an example.

It is also important to note that the method described thus far requires variation of *z* within group *k*. If *z* does not vary within the group, then the directly calculated standard error of $\theta_k$would be zero. The classification would be near certainty. This issue can occur when a characteristic is rare and thus may have none or few sampled individuals in the group. For instance, a 0 percent or 100 percent estimate has no directly calculated variance.

In the case of a directly calculated zero variance, I utilize a procedure for determining standard errors that uses the denominator of the estimate and state's average weight. This method is used by the Census Bureau to create standard errors and margins of error for

zero percent estimates released on American Factfinder (U.S. Census Bureau, 2009). From that standard error, I simulate R replicate estimates assuming a normal distribution $\mathcal{N}(0, \gamma\sigma_k^2)$ where $\gamma = (1-\varepsilon)^2$ and ε is the fay coefficient, commonly set at 0.5, and taking the absolute value of the random draw. This factor γ corrects for difference between the standard and SDR methods for calculations of variance.

Lastly, for comparison purposes, I use only the tract estimate and standard error to simulate sampling distributions for all areas, using the method similar to that used for the 0 and 100 percent estimates. This approach ignores the replicate estimates that were used to calculate the standard error and essentially regenerates synthetic replicate estimates using random draws from a normal distribution. This method, however, ignores any covariance between the indicators and derived estimates. However, this may prove to be a useful method when ACS estimates are combined with data from other surveys. The number of simulated estimates can be fit to the parameters of the survey to which the indicators are being attached. For example, if attaching to the Current Population Survey Annual Social and Economic Supplement, which uses 160 replicate weights to calculate variance, 160 simulated estimates could be drawn.

Example using Poverty Areas

In *Areas With Concentrated Poverty: 2006-2010*, census tracts were categorized into four categories: (I) tracts with poverty rates below the national estimate (about 13.8 percent), (II) tracts with poverty rates at or above the national estimate but below 20.0 percent, (III) tracts with poverty rates at or above 20.0 percent but below 40.0 percent, and (IV) tracts with poverty rates at or above 40.0 percent (Bishaw, 2011). Tracts in category III and category IV were considered "poverty areas". From Table 1 in the brief showed that 61.4 (± 0.1) percent of the U.S. population for whom poverty status was determined were living in category I areas, 16.0 (± 0.1) percent in category II areas, 19.1 (± 0.1) percent in category III areas, and 3.5 (± 0.1) percent in category IV areas. For the population in poverty, 30.6 (± 0.1) percent lived in category I areas, 19.2 (± 0.1) percent lived in category II areas, 37.8 (± 0.1) percent lived in category III areas, and 12.4 (± 0.1) percent lived in category IV areas. Each of the margins of error listed does not take into account the error associated with categorizing a tract's poverty rate.

Like Bishaw, I calculated poverty rates for 72,254 census tracts in the United States. Puerto Rico was excluded from this analysis because it was not included in Bishaw. Also, like Bishaw, my calculations showed 42,383 tracts were in category I, 11,574 tracts were in category II, 14,823 tracts were in category III, and 3,474 tracts were in category IV. But this does not tell the entire picture.

Of the 72,254 census tracts, the median tract size was 3,883 people while tracts ranged from 3 people to 29,369 people in the poverty universe. From the distribution of tract sizes, 98 percent of tracts had populations between 805 (1st percentile) and 9,373 (99^{th} percentile). The size of the tract is important as the precision of estimates tends to decrease with size. Also, of the 72,254 tracts, 517 had poverty rates of 0 percent and another 18 had rates of 100 percent. These areas had standard errors ranging from 0.28 percent to 95.74 percent, as calculated using the ACS production method. Figure 1 shows the poverty rates and their margins of error (90 percent confidence interval) for 100 randomly selected tracts. Many estimates have error bounds that stretch across the poverty area category boundaries. While this figure only shows a select few tracts, it is representative.

The coefficient of variation (CV) is a measure of precision for non-zero estimates. Figure 2 shows the distribution of CVs for tracts with non-zero poverty rates. About half of tracts have a CV that is less than 0.30. At this level, a tract with a poverty rate of 27 percent, for example, would have a standard error of 8.1 percent, and a 90 percent confidence interval that stretches from 13.7 percent to 40.3 percent. This demonstrates that a tract with a median CV could still have error bounds that stretch across all 4 poverty area categories.

Having calculated the tract poverty rates using each of the replicate weights, replicate indicators were determined, and sophisticated variances calculated. Table 1 shows those results. The sophisticated variance on the percentage of the population living in category I areas was 0.0589 yielding a standard error of 0.2426. This standard error was almost 20 times as large as the one calculated using the naïve method, shown in Figure 3.[1] The sophisticated variance on the percentage in category II areas was 0.2231 yielding a standard error of 0.4724. This was 53 times as large as the naïve standard error. The ratio of the standard errors for categories III and IV were 14.1 and 53.4, again showing large differences in the uncertainty attached to the percentages. Restating the earlier statistics with the new variances, 61.4 (± 0.4) percent of the U.S. population for whom poverty status was determined were living in category I areas, 16.0 (± 0.8) percent were in category II areas, 19.1 (± 0.3) percent were in category III areas, and 3.5 (± 0.4) percent were in category IV areas.

By state, the standard errors followed similar patterns, shown in table 2. Sophisticated standard errors were roughly 10 to 50 times as large as the naïve standard errors. The smallest ratio was for the percentage of people in category IV areas in Utah at 7.1 times as large. The largest ratio was for the percentage of people in category II areas in Rhode Island at 55.0 times as large. The median sophisticated CV across all states and the four categories was 0.101, compared with a median naïve CV of 0.004.

So far, the standard errors were generated from indicators derived directly from replicate estimates. To perform this method, however, requires the restricted access data held by the Census Bureau. To overcome this problem, one could produce a set of simulated replicate estimates based on the ACS estimate and the standard error, taken from FactFinder. Table 3 shows the break down of measurement error variance and covariance that sum with the naïve variance to produce the total (sophisticated) variance. Compared to the covariance estimates in table 2, the simulated indicators produce covariances that are smaller in magnitude than the replicate weight-based covariances. This makes sense because the replicate weight-based indicators preserve the fact that both the indicator and the final estimate derive from the same sample. The simulated indicators ignore this relationship and thus covariance should be closer to zero. Overall, covariance makes up a very small percentage of the total variance (less than 3 percent) so this component may be assumed away under necessary circumstances.

Table 4 shows the standard errors on the U.S. and state estimates using simulated replicates and compares them to the replicate weight-based standard errors. The ratios were between 0.72 and 1.30 across all categories and states and the average was 1.016.

[1] The margins of error shown in Bishaw (2011) are rounded to the nearest tenth of a percentage point. For margins of error that would round to zero, a minimum value of 0.1 percent is used. The standard errors used for comparisons in this paper are unrounded. Hence, margins of error may appear to be less than 20 times as large.

This means a typical standard error using the simulation was only 1.6 percent higher than its non-simulated standard error. Figure 4 shows the degree of correlation between the two (r=0.990). The estimates tend to cluster around the 1:1 line, with few extreme outliers. The ratio was higher (or lower) when the typical standard error was small, thus a small difference could produce a large ratio.

Discussion

An advantage of the American Community Survey is that it may be used to determine the characteristics of numerous population subgroups such as those defined by geographic boundaries, occupations, or place of birth. Furthermore, it is possible to then classify subgroups into groups defined by those characteristics. When these estimates are used in subsequent tabulations or models, it is important to take into account the error with making classifications. As the population subgroups get smaller and thus have smaller sample sizes, it becomes more likely that the estimates on which the classification is based will have larger variances. This uncertainty must be reflected in the margins of errors that accompany final estimates based on the classification.

As the example of poverty areas shows, this source of error cannot be ignored. Assigning census tracts to poverty categories resulted in estimates of uncertainty that were 10 to 50 times as large as the margins of error when the classification error was unaccounted. Ignoring this source of error can greatly overstate the level of confidence in the final estimate

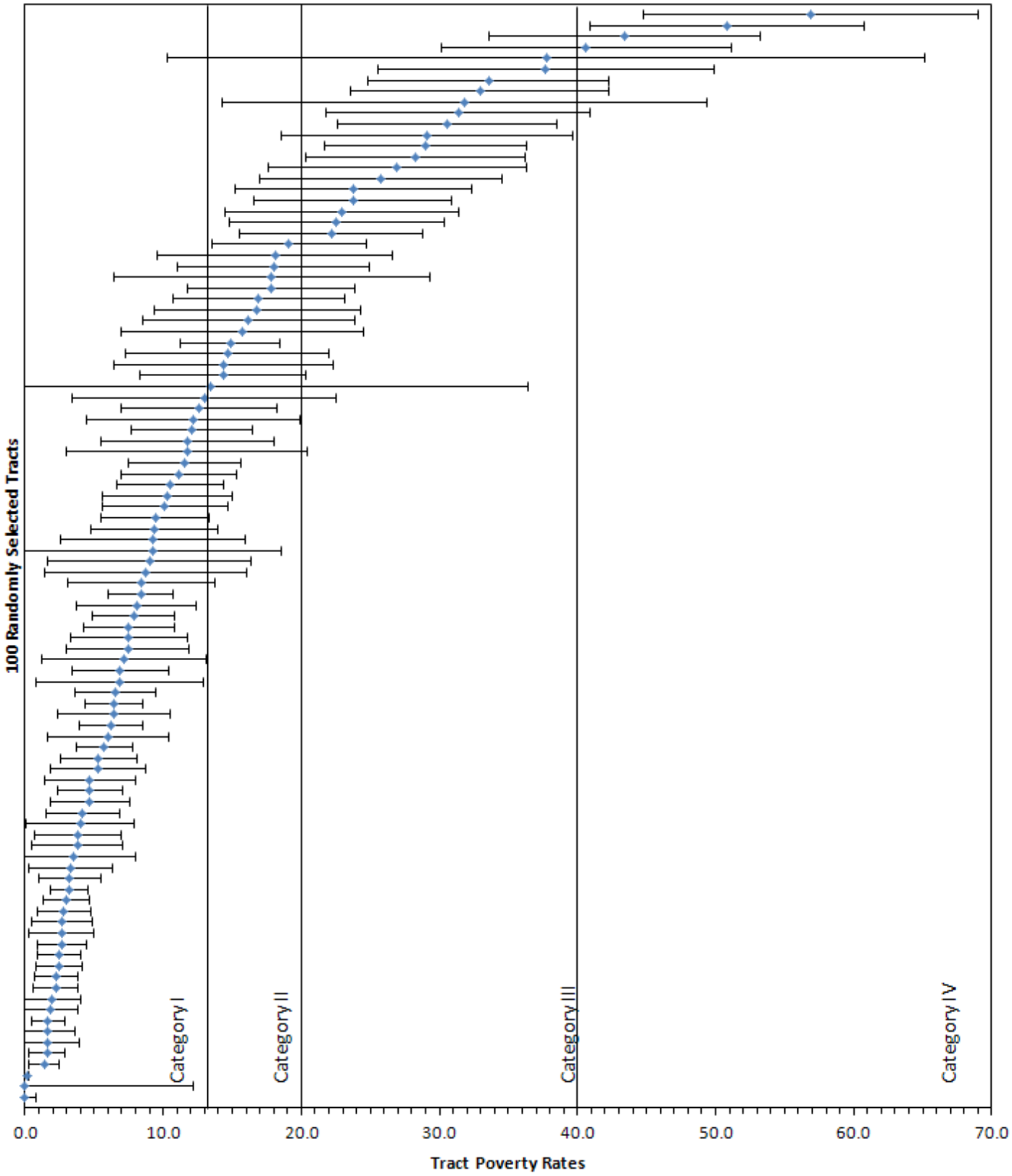
Ultimately, methods such as the one described in this paper should provide greater utility with using ACS for research. It is easy to see why some data users are hesitant to use small domain estimates given the sizes of the margins of error that accompany them. By aggregating the small domain estimates together, and properly accounting for that error, they can be confident in their uncertainty, while tackling the desired research question.

References

Bishaw, Alemayehu. *Areas with Concentrated Poverty: 2006-2010*, ACSBR/10-17, U.S. Census Bureau, Washington, DC, 2011. Available at <www.census.gov/prod/2011pubs/acsbr10-17.pdf>.

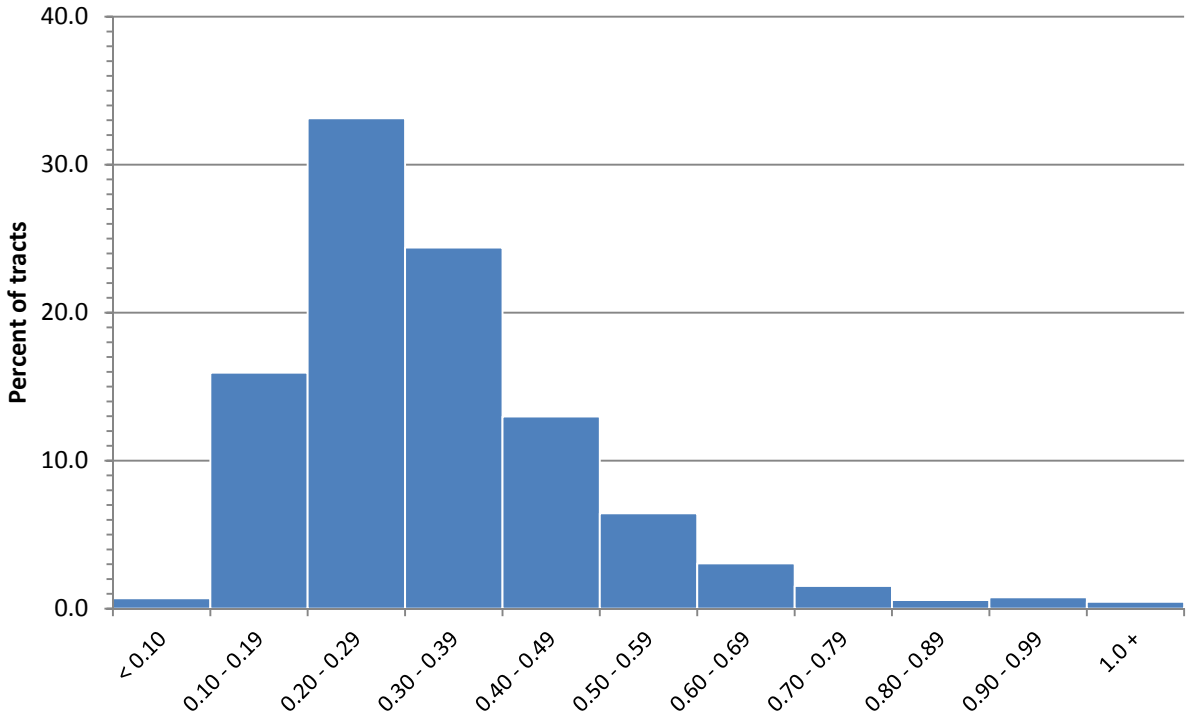
U.S. Census Bureau, *Design and Methodology*, American Community Survey, U.S. Government Printing Office, Washington, DC, 2009. Available at <www.census.gov/acs/www/Downloads/survey_methodology /acs_design_methodology.pdf>.

Figure 1. Poverty rates for 100 randomly selected tracts



Source: U.S. Census Bureau, American Community Survey, 2006-2010

Figure 2. Distribution of tract coefficients of variation



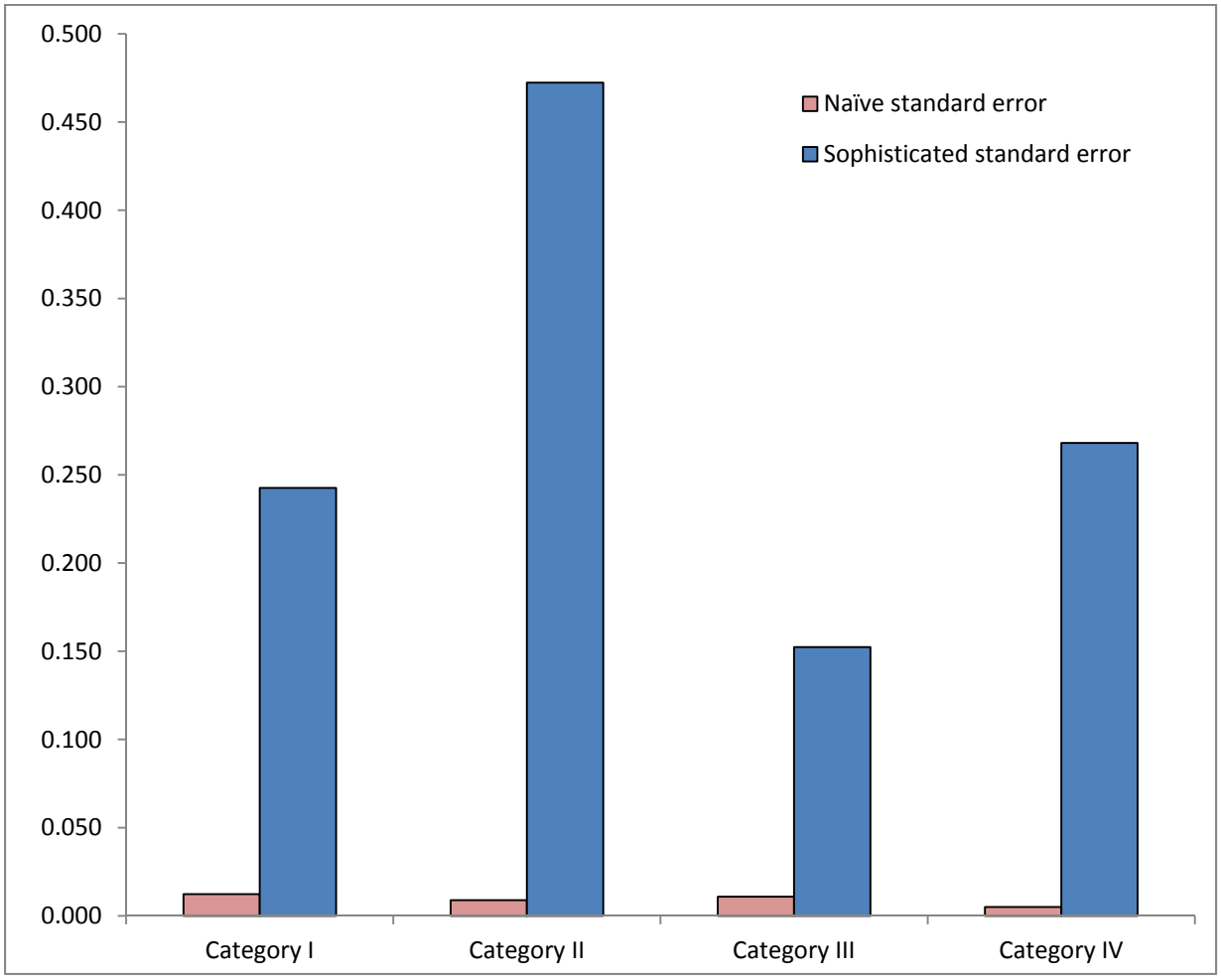
Source: U.S. Census Bureau, American Community Survey, 2006-2010

Table 1. Variance associated with the percentage in poverty categories – Replicate estimates

	Category I	Category II	Category III	Category IV
Naïve variance ($\widetilde{var}(Y)$)	0.000153	0.000079	0.000118	0.000025
Naïve standard error	0.012355	0.008887	0.010850	0.005017
Sophisticated variance ($var(Y)$)	0.058874	0.223126	0.023230	0.071863
Sophisticated standard error	0.242639	0.472362	0.152413	0.268073
Measurement error variance ($var(\Delta)$)	0.062034	0.225113	0.024111	0.071109
Covariance ($cov(Y, \Delta)$)	-0.001656	-0.001033	-0.000499	0.000364
Ratio of standard errors	19.64	53.15	14.05	53.43
Percentage of sophisticated variance				
Naïve variance	0.26	0.04	0.51	0.04
Measurement error variance	105.37	100.89	103.79	98.95
Covariance	-2.81	-0.46	-2.15	0.51

Source: U.S. Census Bureau, American Community Survey, 2006-2010

Figure 3. Naïve and Sophisticated Standard Errors



Source: U.S. Census Bureau, American Community Survey, 2006-2010

Table 2. State Estimates and Standard Errors

State	Estimate				Naïve S.E.				Sophisticated S.E.				SE Ratio			
	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV
Alabama	45.30	22.67	27.32	4.71	0.09	0.08	0.09	0.04	1.83	2.40	1.31	0.76	20.4	29.4	15.3	17.4
Alaska	76.87	14.94	8.19	-	0.22	0.21	0.12	-	3.97	4.89	1.94	0.04	17.7	22.8	15.9	-
Arizona	57.43	14.49	22.91	5.17	0.08	0.06	0.08	0.05	1.02	1.29	0.93	0.71	12.4	20.7	12.1	15.0
Arkansas	39.39	24.19	32.04	4.38	0.09	0.10	0.12	0.06	2.19	3.52	4.00	0.78	24.9	35.3	34.6	13.2
California	61.12	15.73	20.55	2.60	0.04	0.03	0.03	0.01	0.59	0.79	0.45	0.40	16.5	30.4	14.0	27.4
Colorado	65.28	14.51	18.52	1.69	0.08	0.07	0.08	0.03	1.87	2.23	1.02	0.74	22.1	31.4	13.1	23.8
Connecticut	78.88	9.57	8.84	2.71	0.07	0.07	0.07	0.05	1.06	1.60	1.44	0.37	15.5	23.4	21.7	7.9
Delaware	73.19	17.37	7.89	1.55	0.17	0.16	0.12	0.06	3.20	4.45	2.49	0.77	18.7	28.7	20.2	11.8
District of Columbia	47.79	17.48	26.54	8.19	0.24	0.22	0.27	0.17	2.95	4.79	3.83	2.05	12.3	21.5	14.4	11.9
Florida	61.22	18.37	17.81	2.61	0.05	0.04	0.04	0.02	0.80	1.19	0.68	0.34	15.9	28.3	15.5	14.0
Georgia	52.31	18.18	25.97	3.55	0.06	0.06	0.06	0.03	0.96	1.31	1.06	0.36	16.4	21.8	17.9	14.0
Hawaii	79.20	9.99	9.84	0.97	0.18	0.14	0.12	0.04	2.26	2.76	1.72	0.58	12.2	19.5	14.2	13.6
Idaho	58.98	26.89	12.55	1.59	0.13	0.12	0.10	0.04	2.90	3.97	2.88	0.71	21.7	32.7	28.1	20.0
Illinois	67.41	13.69	15.50	3.40	0.05	0.03	0.04	0.03	0.74	1.17	0.70	0.54	15.9	34.4	16.0	19.8
Indiana	63.52	16.32	17.28	2.87	0.05	0.05	0.05	0.03	0.99	1.24	0.95	0.53	20.9	24.7	18.5	15.2
Iowa	71.83	14.98	11.86	1.33	0.07	0.06	0.06	0.03	1.62	1.72	1.57	0.54	22.6	30.1	26.2	19.7
Kansas	67.98	15.33	14.57	2.13	0.08	0.07	0.08	0.04	2.07	2.00	1.94	0.47	25.5	28.1	25.3	10.7
Kentucky	42.93	20.40	32.24	4.43	0.07	0.08	0.09	0.05	1.35	1.69	1.43	0.55	18.3	22.4	15.4	11.6
Louisiana	45.44	19.02	28.92	6.62	0.09	0.08	0.09	0.06	1.48	2.02	1.35	0.64	15.7	25.3	15.5	11.2
Maine	65.02	20.13	13.87	0.98	0.10	0.08	0.10	0.03	2.67	3.03	1.79	0.50	26.7	36.4	18.5	14.3
Maryland	82.34	8.65	8.05	0.96	0.06	0.05	0.05	0.02	1.94	1.71	0.71	0.24	30.2	35.9	14.9	13.8
Massachusetts	75.15	9.18	13.56	2.12	0.06	0.04	0.05	0.03	0.88	1.00	1.22	0.50	15.8	24.0	24.4	17.4
Michigan	61.26	14.68	18.32	5.74	0.04	0.03	0.05	0.03	0.81	1.13	0.79	0.42	18.2	37.5	15.6	12.9
Minnesota	77.20	10.98	9.16	2.66	0.05	0.04	0.05	0.03	0.87	1.17	0.99	0.29	17.1	27.4	21.8	9.3
Mississippi	29.07	25.22	36.61	9.10	0.11	0.11	0.12	0.08	4.06	4.52	2.27	0.99	37.5	39.6	18.4	12.2
Missouri	58.55	19.82	18.72	2.90	0.07	0.06	0.06	0.03	1.18	2.09	1.49	0.39	16.9	34.3	24.6	13.1
Montana	54.70	24.59	19.64	1.07	0.15	0.16	0.11	0.06	3.80	4.79	2.56	1.30	25.1	30.7	22.7	21.2
Nebraska	70.06	15.95	12.41	1.57	0.09	0.09	0.09	0.04	1.72	2.22	1.94	0.72	18.3	25.6	22.7	18.6
Nevada	69.07	14.11	15.18	1.64	0.13	0.11	0.12	0.05	1.97	2.01	1.92	0.55	15.4	18.7	16.6	12.0
New Hampshire	86.40	8.62	4.48	0.50	0.10	0.08	0.07	0.03	2.50	2.32	1.49	0.34	24.0	29.6	20.4	10.7
New Jersey	79.96	7.91	10.39	1.74	0.04	0.04	0.04	0.02	1.18	1.47	0.81	0.38	26.3	41.1	22.2	15.9
New Mexico	39.76	22.29	33.62	4.33	0.13	0.13	0.15	0.07	2.61	3.98	2.11	1.32	20.4	30.0	14.1	18.3
New York	62.16	13.07	19.92	4.85	0.04	0.03	0.04	0.02	0.50	0.63	0.54	0.52	12.2	20.9	13.7	21.1
North Carolina	51.01	22.29	23.44	3.26	0.06	0.05	0.06	0.03	1.39	2.39	1.21	0.36	23.3	45.3	20.3	12.8

State	Estimate				Naïve S.E.				Sophisticated S.E.				SE Ratio			
	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV
North Dakota	72.07	13.93	12.01	2.00	0.14	0.14	0.12	0.05	3.24	3.37	1.73	0.78	23.3	24.9	14.5	15.2
Ohio	63.10	14.46	17.44	4.99	0.04	0.04	0.04	0.03	0.91	1.15	0.70	0.36	22.8	32.7	17.0	11.7
Oklahoma	46.18	23.75	27.34	2.73	0.09	0.08	0.09	0.05	2.70	2.53	1.62	0.54	28.7	30.0	17.1	11.6
Oregon	56.04	24.19	18.42	1.35	0.09	0.08	0.09	0.03	2.31	2.44	1.74	0.47	24.7	30.9	20.1	17.0
Pennsylvania	70.55	11.35	14.05	4.06	0.03	0.03	0.04	0.03	0.67	0.95	0.79	0.32	19.6	28.4	22.4	11.1
Rhode Island	72.50	5.79	19.27	2.44	0.13	0.10	0.12	0.06	3.95	5.27	2.51	0.87	30.1	55.0	21.7	14.6
South Carolina	47.41	21.23	27.71	3.64	0.09	0.09	0.10	0.04	1.87	1.85	1.33	0.52	19.9	20.0	13.4	12.6
South Dakota	63.96	18.66	11.89	5.49	0.15	0.13	0.16	0.09	2.99	4.49	3.87	1.83	20.5	35.9	23.8	21.0
Tennessee	47.70	21.95	26.05	4.30	0.06	0.06	0.07	0.05	1.49	2.32	1.21	0.54	23.2	36.0	16.3	11.7
Texas	50.18	17.22	26.59	6.01	0.04	0.04	0.04	0.03	0.70	0.81	0.76	0.47	16.1	22.1	17.4	17.6
Utah	76.49	11.09	10.13	2.30	0.08	0.07	0.07	0.04	2.09	1.96	1.17	0.28	24.6	26.7	16.4	7.1
Vermont	74.58	15.85	8.50	1.07	0.11	0.11	0.11	0.06	2.73	3.15	1.75	0.50	23.9	27.8	16.4	8.7
Virginia	74.35	13.69	10.28	1.68	0.05	0.04	0.04	0.02	0.98	1.62	0.99	0.22	18.7	38.5	25.9	9.6
Washington	67.58	15.96	14.85	1.61	0.05	0.06	0.06	0.02	1.06	1.33	0.79	0.59	19.3	23.9	12.8	27.7
West Virginia	36.96	32.00	28.62	2.42	0.13	0.11	0.12	0.05	2.95	4.00	2.34	0.60	23.6	35.0	19.4	12.6
Wisconsin	73.66	13.33	9.70	3.31	0.06	0.04	0.06	0.03	1.04	1.31	0.82	0.34	18.3	30.4	14.6	9.8
Wyoming	79.16	13.47	7.24	0.13	0.22	0.19	0.11	0.02	3.35	3.70	2.70	0.73	15.1	19.1	24.8	30.2

Source: U.S. Census Bureau, American Community Survey, 2006-2010

Table 3. Variance associated with the percentage in poverty categories – Simulated Replicate estimates

	Category I	**Category II**	**Category III**	**Category IV**
Sophisticated variance ($var(Y)$)	0.021983	0.183061	0.029690	0.092355
Sophisticated standard error	0.148266	0.427856	0.172308	0.303900
Measurement error variance ($var(\Delta)$)	0.021481	0.184285	0.029275	0.091577
Covariance ($cov(Y, \Delta)$)	0.000175	-0.000652	0.000149	0.000377
Ratio of standard errors	12.00	48.15	15.88	60.57
Percentage of sophisticated variance				
Naïve variance	0.69	0.04	0.40	0.03
Measurement error variance	97.72	100.67	98.60	99.16
Covariance	0.79	-0.36	0.50	0.41

Source: U.S. Census Bureau, American Community Survey, 2006-2010

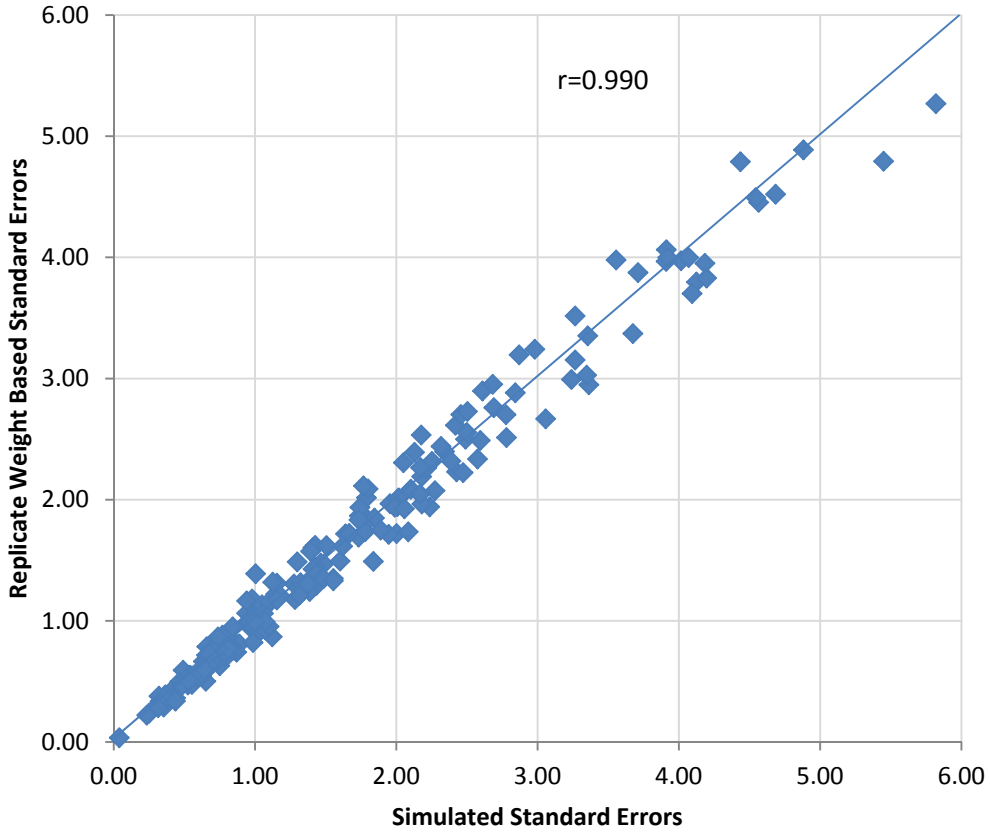
Table 4. Ratio of Simulated Standard Errors to Replicate Weight-based Standard Errors

State	Simulated Standard Errors				Replicate Weight-based Standard Errors				Ratio			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
United States	0.15	0.43	0.17	0.30	0.24	0.47	0.15	0.27	0.61	0.91	1.13	1.13
Alabama	1.77	2.34	1.32	0.85	1.83	2.40	1.31	0.76	0.97	0.98	1.01	1.13
Alaska	4.02	4.88	1.74	0.04	3.97	4.89	1.94	0.04	1.01	1.00	0.90	1.11
Arizona	1.04	1.43	1.07	0.80	1.02	1.29	0.93	0.71	1.01	1.11	1.16	1.13
Arkansas	2.18	3.27	3.92	0.82	2.19	3.52	4.00	0.78	0.99	0.93	0.98	1.06
California	0.49	0.79	0.47	0.40	0.59	0.79	0.45	0.40	0.83	1.01	1.03	1.01
Colorado	1.77	2.43	1.05	0.79	1.87	2.23	1.02	0.74	0.95	1.09	1.03	1.07
Connecticut	0.94	1.41	1.43	0.41	1.06	1.60	1.44	0.37	0.89	0.88	0.99	1.12
Delaware	2.87	4.57	2.59	0.81	3.20	4.45	2.49	0.77	0.90	1.02	1.04	1.06
District of Columbia	3.36	4.44	4.20	2.17	2.95	4.79	3.83	2.05	1.14	0.93	1.10	1.06
Florida	0.70	1.15	0.74	0.33	0.80	1.19	0.68	0.34	0.88	0.97	1.09	0.98
Georgia	0.96	1.15	1.06	0.37	0.96	1.31	1.06	0.36	1.00	0.88	1.00	1.02
Hawaii	2.22	2.69	1.67	0.64	2.26	2.76	1.72	0.58	0.98	0.97	0.97	1.11
Idaho	2.61	3.91	2.84	0.67	2.90	3.97	2.88	0.71	0.90	0.99	0.99	0.95
Illinois	0.87	0.94	0.72	0.62	0.74	1.17	0.70	0.54	1.17	0.81	1.04	1.14
Indiana	1.06	1.39	1.10	0.57	0.99	1.24	0.95	0.53	1.07	1.12	1.15	1.08
Iowa	1.43	1.64	1.39	0.60	1.62	1.72	1.57	0.54	0.88	0.95	0.89	1.10
Kansas	2.27	2.01	2.01	0.53	2.07	2.00	1.94	0.47	1.10	1.00	1.03	1.11
Kentucky	1.55	1.73	1.41	0.54	1.35	1.69	1.43	0.55	1.15	1.02	0.99	0.97
Louisiana	1.47	1.79	1.49	0.67	1.48	2.02	1.35	0.64	0.99	0.89	1.10	1.06
Maine	3.06	3.35	1.77	0.56	2.67	3.03	1.79	0.50	1.15	1.11	0.99	1.11
Maryland	2.24	1.95	0.79	0.25	1.94	1.71	0.71	0.24	1.15	1.14	1.11	1.05
Massachusetts	0.77	0.95	1.15	0.47	0.88	1.00	1.22	0.50	0.87	0.95	0.94	0.94
Michigan	0.87	1.05	0.75	0.41	0.81	1.13	0.79	0.42	1.07	0.93	0.95	0.97
Minnesota	1.12	1.12	1.07	0.35	0.87	1.17	0.99	0.29	1.29	0.96	1.08	1.22
Mississippi	3.91	4.68	2.17	1.00	4.06	4.52	2.26	0.99	0.96	1.04	0.96	1.01
Missouri	0.98	1.80	1.60	0.37	1.18	2.09	1.49	0.39	0.83	0.86	1.07	0.94
Montana	4.12	5.45	2.50	1.37	3.80	4.79	2.56	1.30	1.09	1.14	0.98	1.05
Nebraska	2.00	2.47	1.99	0.66	1.72	2.22	1.94	0.72	1.16	1.11	1.03	0.92
Nevada	1.96	2.02	2.06	0.54	1.97	2.02	1.92	0.55	0.99	1.00	1.07	0.98
New Hampshire	2.49	2.39	1.84	0.34	2.50	2.32	1.49	0.34	1.00	1.03	1.23	0.98
New Jersey	1.28	1.49	0.77	0.32	1.18	1.47	0.81	0.38	1.09	1.01	0.95	0.85
New Mexico	2.42	3.56	1.77	1.13	2.61	3.98	2.11	1.32	0.93	0.89	0.84	0.85
New York	0.65	0.75	0.55	0.53	0.50	0.63	0.54	0.52	1.30	1.19	1.02	1.01
North Carolina	1.00	2.13	1.31	0.44	1.39	2.39	1.21	0.36	0.72	0.89	1.09	1.21
North Dakota	2.98	3.67	2.08	0.74	3.24	3.37	1.73	0.78	0.92	1.09	1.20	0.95
Ohio	1.03	1.10	0.72	0.35	0.91	1.15	0.70	0.36	1.13	0.95	1.03	0.99
Oklahoma	2.46	2.18	1.62	0.57	2.70	2.53	1.62	0.54	0.91	0.86	1.00	1.06
Oregon	2.05	2.32	1.78	0.55	2.31	2.44	1.74	0.47	0.89	0.95	1.03	1.17
Pennsylvania	0.64	0.84	0.66	0.33	0.67	0.95	0.79	0.32	0.95	0.89	0.84	1.02
Rhode Island	4.18	5.82	2.78	0.74	3.95	5.27	2.51	0.87	1.06	1.10	1.11	0.85
South Carolina	1.74	1.85	1.55	0.52	1.87	1.85	1.33	0.52	0.93	1.00	1.17	0.99
South Dakota	3.24	4.54	3.71	1.74	2.99	4.49	3.87	1.83	1.08	1.01	0.96	0.95
Tennessee	1.30	2.25	1.18	0.59	1.49	2.32	1.21	0.54	0.87	0.97	0.97	1.10

State	Simulated Standard Errors				Replicate Based Standard Errors				Ratio			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
Texas	0.70	0.89	0.69	0.48	0.70	0.81	0.76	0.47	1.01	1.10	0.91	1.03
Utah	2.10	2.18	1.15	0.31	2.09	1.96	1.17	0.28	1.01	1.11	0.98	1.11
Vermont	2.50	3.27	1.89	0.53	2.73	3.15	1.75	0.50	0.92	1.04	1.08	1.06
Virginia	0.96	1.51	0.96	0.23	0.98	1.62	0.99	0.22	0.98	0.93	0.98	1.06
Washington	1.03	1.42	0.84	0.60	1.06	1.33	0.79	0.59	0.97	1.07	1.05	1.02
West Virginia	2.68	4.07	2.57	0.64	2.95	4.00	2.34	0.60	0.91	1.02	1.10	1.07
Wisconsin	1.01	1.28	0.99	0.44	1.04	1.31	0.82	0.34	0.97	0.98	1.20	1.30
Wyoming	3.36	4.09	2.78	0.68	3.35	3.70	2.70	0.73	1.00	1.11	1.03	0.93

Source: U.S. Census Bureau, American Community Survey, 2006-2010

Figure 4. Correlation between Simulated and Replicate-based Standard Errors -- States



Appendix A – SAS code for tabulation with sophisticated variances

```
libname acs '/PATH/';

/*geoid can be replaced with the level of geography or group to be classified*/

data extract;
  set acs.psam_pus (where=(povpip in(0:501)) keep=pwgtp: povpip geoid);

  if povpip<100 then pov=1;
  else pov=0;

run;
proc means noprint completetypes data=extract;
  class geoid pov;
  types geoid geoid*pov;
  var pwgtp pwgtp1-pwgtp80;
  output out=geo_sums sum=cnt0-cnt80;
run;
proc sort data=geo_sums; by geoid; run;
data geo_categories (keep=geoid pcat:);
  set geo_sums;
  by geoid;

  array cnt [0:80] cnt0-cnt80;
  array tot [0:80];
  array pct [0:80];
  array pcat [0:80] pcat0-pcat80;

  retain tot pcat;
  if first.geoid and pov=. then do i=0 to 80;
    tot[i]=cnt[i];
      pct[i]=.;
      pcat[i]=.;
  end;
  if pov=1 then do i=0 to 80;
    if cnt[i]=. then cnt[i]=0;
    pct[i]=100*cnt[i]/tot[i];

    if pct[i]<13.8 then pcat[i]=1;
    else if 13.8<=pct[i]<20.0 then pcat[i]=2;
    else if 20.0<=pct[i]<40.0 then pcat[i]=3;
    else if pct[i]>=40.0 then pcat[i]=4;
  end;
  if last.geoid;
run;
proc sort data=extract; by geoid; run;
proc sort data=geo_categories; by geoid; run;
data pov_area_dataset;
  merge extract (in=a) geo_categories;
  by geoid; if a;
run;
proc computab data=pov_area_dataset notranspose out=OutputTable;
  cols col1-col9;
  rows row1-row84;
  rows row2-row81 /noprint;
```

```
  array wgts [0:80] pwgtp pwgtp1-pwgtp80;
  array pcat [0:80] pcat0-pcat80;

  do i=0 to 80;
    table[(i+1),(pcat[i]+1)]=table[(i+1),(pcat[i]+1)]+wgts[i];
  end;

colsum: if _row_ in(1:81) then do;
          col1=sum(of col2-col5);
          if col1>0 then do;
            col6=100*col2/col1;
            col7=100*col3/col1;
            col8=100*col4/col1;
            col9=100*col5/col1;
          end;
        end;

        /*VARIANCE*/
rowsum: row82=(uss(of row2-row81)/20)-(row1*sum(of row2-row81)/10)+(4*row1**2);
        row83=sqrt(row82); /*SE*/
        row84=row83*1.645; /*MOE*/

run;
```