UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

8/27/10

2009 AMERICAN COMMUNITY SURVEY VARIANCE MEMORANDUM SERIES #ACS09-V-3

MEMORANDUM FOR Daniel H. Weinberg
Assistant Director for ACS and Decennial Census

From: David C. Whitford ***signed 8/27/10***
Chief, Decennial Statistical Studies Division

Prepared By: Michael Starsinic
ACS Variance Estimation and Statistical Support Branch
Decennial Statistical Studies Division

Alfredo Navarro
Assistant Division Chief, ACS Statistical Design
Decennial Statistical Studies Division

Subject: Applying a Finite Population Correction Factor to ACS Variance Estimates

Background

The American Community Survey (ACS) has not in the past, and does not currently, use a finite population correction (FPC) factor in its variance estimation methodology. However, the Census 2000 long form variance estimation methodology did include such a factor. One-year ACS samples are not large enough for an FPC to have much impact on variances. However, with 5-year ACS estimates, up to 50 percent of housing units in certain blocks may have been in sample over the 5-year period. Applying an FPC factor will enable us to get a more accurate estimate of the variance, particularly for small areas. Research began on assessing the feasibility and the magnitude of the effect of incorporating an FPC factor into ACS variance estimates.

Methodology

The long form used the same successive difference variance methodology that the ACS employs, so it was expected that the long form FPC method could be adapted to the ACS variances.

The replicate factors, $f_{i,j}$ (i = replicate #, j = sample unit #), in a typical successive differences variance estimator are defined as follows:

$$f_{i,j} = 1 + 2^{-1.5} a_{i1,j} - 2^{-1.5} a_{i2,j} = \begin{cases} 1 \\ 1 + 1/\sqrt{2} \\ 1 - 1/\sqrt{2} \end{cases}$$

where $a_{i1,j} = \pm 1$ and $a_{i2,j} = \pm 1$ are the appropriate cells from a Hadamard matrix.

The long form applied an FPC factor directly to the replicate factors:

$$f^*_{i,j} = 1 + \left(2^{-1.5} a_{i1,j} - 2^{-1.5} a_{i2,j} \right) \times \sqrt{1 - n_{eff} / N}$$

where n_{eff} is defined as the observed number of long form sample respondents, and N is defined as the uncorrected census count. The FPC is typically applied as a multiplicative factor "outside" the variance formula. However, under certain simplifying assumptions, the variance using the replicate factors after applying the FPC factor is equal to the original variance multiplied by the FPC factor. This method allows a direct application of the FPC to each housing unit's or person's set of replicate weights, and a seamless incorporation into American Community Survey Office (ACSO's) current tabulation methodology, rather than having to keep track of multiplicative factors when tabulating across areas of different sampling rates. For further information on the theory behind the FPC adjustment, see the Appendix to this memorandum.

Replicate factors are assigned to sampled housing units at the very beginning of the weighting process, and each of the sets of replicate base weights are processed through the weighting system as the production weights are. It is expected that the variance improvement in the variance estimate will carry though the weighting, and will be seen when the final weights are used.

The FPC factor could be applied at any geographic level. Since the ACS sample is controlled at the small area level (mainly census tracts and governmental units), a low level of geography was desirable. At higher levels, the high sampling rates in specific blocks would likely be masked by the lower rates in surrounding blocks. For that reason, we decided to apply the factors at the tract level.

The FPC factor is applied to housing units only. Group quarters persons do not have an FPC factor applied to their replicate factors.

ACS Application

Because of the ACS computer-assisted personal interviewing (CAPI) sub-sampling, a single FPC factor was not deemed suitable. To develop an ACS application we considered basic features of the ACS sampling and weighting methodology, mainly the sampling of mail-computer assisted telephone interview (CATI) non-respondents. There are at least two models of non-response in sample surveys. One is the stochastic model in which every element in the population has a distinct probability of responding (or not) if selected into the survey. A second and older model is the fixed-response model. This model views the *population* as being composed of two strata – those that would respond if selected and those that would not. The fixed-response model basically assumes that for a fixed set of survey conditions (e.g., survey budget, saliency of the

survey content, planned non-response follow-up methods, etc), the survey's target population can be viewed as being composed of those people who will respond to the survey if invited and those that will not, and that this effect is *fixed* unless the survey conditions are changed.

One advantage of the fixed-response model is its rather simple form for stating the composition of total variance into the mail-CATI respondents and the CAPI respondents. In the fixed-response model using simple random sampling, people in both "strata" (respondents and non-respondents) will be selected into the sample, with the non-respondents becoming the sample's non-respondents.

Under the fixed response model, a two-tiered FPC was proposed.

$$F_1 = \frac{n_{mail} + n_{CATI}}{N \times R}, \text{ and}$$

$$F_2 = \frac{n_{CAPI}}{N \times (1 - R)}, \text{ where } R = \frac{w_{mail} + w_{CATI}}{w_{samp}}$$

N is the unweighted sample universe count, and n_{mail}, n_{CATI}, and n_{CAPI} are the unweighted counts of respondents by mode. R is the weighted proportion (using the unbiased sampling weights) of those who respond by mail or CATI.

For mail and CATI respondents,

$$\text{FPC Factor} = \sqrt{1 - F_1}$$

and for CAPI respondents,

$$\text{FPC Factor} = \sqrt{1 - F_2}$$

Simulation and Results

Unfortunately, no five-year ACS test files were available that contained both a full set of replicate weights and a full suite of person and housing unit characteristics. Instead, the FPC factors were calculated and applied to 2006-2008 three-year ACS production data. We would expect some improvement in variances from applying an FPC factor to three-year data, but not as much as with five-year data. Puerto Rico was excluded from the simulation because the source dataset for N (sample universe counts) did not include Puerto Rico. In production, Puerto Rico would be included.

Once new replicate weights were computed, variances both with and without the FPC factor were calculated for all data profile estimates (except medians) for states, all counties, all places, and tracts in three states (Maryland, New Mexico, and Pennsylvania).

For nonzero estimates, we calculated the ratio of the standard error of the estimate with the FPC factor to the standard error of the estimate without the FPC factor.

Table 1: Distribution of Ratios of SEs, 2006-2008 FPC Simulation

	Q1	Median	Q3
State	97.70%	98.17%	98.58%
County	96.51%	97.68%	98.41%
Place	95.92%	97.53%	98.48%
Tract	96.81%	97.79%	98.56%

Across all geographic areas within a summary level and all computed profile estimates, the median ratio is showing about a 1.8 to 2.5 percent improvement in the standard errors, due to the FPC factor. Some geographic areas experienced larger improvements, from 5 to 10 percent.

Table 2 below looks at the rounded margin of error (MOE) for nonzero estimates, which is how the MOE would be published on American FactFinder. If a standard error improves by such a small amount that the published MOE would not change, then it really has not improved from the public's perspective.

Table 2: Comparison of Rounded MOEs by Summary Level and Estimate Type, 2006-2008 FPC Simulation

Geo	Est Type	# Est	MOE Up	No Change	MOE Down
State	HHld Count	2,193	0.5%	0.1%	99.4%
	HU Count	6,782	2.3%	0.3%	97.4%
	Pop Count	12,739	0.4%	3.0%	96.6%
	Proportion	1,479	0.1%	86.5%	13.4%
	Ratio	714	0.0%	33.8%	66.2%
	total	23,907	0.9%	8.1%	91.0%
County	HHld Count	134,006	0.7%	1.4%	97.9%
	HU Count	395,844	2.4%	2.8%	94.8%
	Pop Count	691,822	0.4%	7.7%	91.9%
	Proportion	86,436	0.1%	29.5%	70.4%
	Ratio	43,837	0.2%	21.9%	77.9%
	total	1,351,945	1.0%	7.5%	91.5%
Place	HHld Count	924,608	0.8%	14.1%	85.1%
	HU Count	2,450,741	1.1%	14.9%	84.1%
	Pop Count	4,260,022	0.5%	9.8%	89.7%
	Proportion	463,536	1.2%	14.5%	84.3%
	Ratio	336,255	4.1%	11.3%	84.7%
	total	8,435,162	0.9%	12.1%	87.1%
Tract	HHld Count	185,983	0.2%	3.9%	95.9%
	HU Count	487,045	0.4%	5.7%	93.9%
	Pop Count	880,502	0.1%	4.2%	95.7%
	Proportion	92,400	0.1%	9.9%	90.1%
	Ratio	64,356	1.0%	11.4%	87.6%
	total	1,710,286	0.2%	5.2%	94.6%

Across all summary levels, and for nearly all estimate types, the percentage of estimates where the MOE decreases is quite large, and the percentage where the MOE increases is very small. Proportions and ratios at the state and county levels show a large percentage did not change. This is because all nonzero MOEs that would round to zero (e.g. 0.049 percent) are rounded up to the smallest publishable value (e.g. 0.1 percent). So, small MOEs that got smaller through the use of the FPC factor might not have their published MOE change.

Recommendation

The empirical simulation suggests a small but noticeable improvement in standard errors from applying the FPC factor to the 3-year estimates data set. We anticipate a larger improvement of the standard errors for the 5-year 2005-2009 estimates. Therefore, we recommend applying the FPC to the 2005-2009 production 5-year estimates and the 2007-2009 production 3-year estimates.

attachments

cc:
Susan Schechter (ACSO)
Alfredo Navarro (DSSD)
Karen King
Mark Asiala
Steve Hefter
Michael Starsinic
Sirius Fuller
Keith Albright

Appendix: Some Additional Theory Behind the ACS FPC Adjustment

The ACS variances are calculated using the successive differences replication methodology, which was developed by Bob Fay to appropriately handle systematically selected samples such as the ACS, the Current Population Survey (CPS), and the 2000 long form. Eighty replicate weights are created by first applying sets of replicate factors to each observation's initial base weight, and then reprocessing the weighting methodology independently on each set of replicate initial weights. The variance of an estimate is calculated from the sum of the squared differences between the production estimate and the 80 replicate estimates, created from the 80 sets of replicate weights.

The replicate factors, $f_{i,j}$ (i = replicate #, j = sample unit #), in a typical successive differences variance estimator are defined as follows:

$$f_{i,j} = 1 + 2^{-1.5} a_{i1,j} - 2^{-1.5} a_{i2,j} = \begin{cases} 1 \\ 1 + 1/\sqrt{2} \\ 1 - 1/\sqrt{2} \end{cases}$$

where $a_{i1,j} = \pm 1$ and $a_{i2,j} = \pm 1$ are the appropriate cells from a Hadamard matrix.

Instead, we are proposing applying the FPC adjustment directly to the replicate factors:

$$f^*_{i,j} = 1 + \left(2^{-1.5} a_{i1,j} - 2^{-1.5} a_{i2,j} \right) \times FPC, \text{ where } FPC = \sqrt{1 - \text{sampling fraction}}$$

Let's follow through with the algebra associated with this FPC adjustment using the simplified assumption that there are no further weighting adjustments after the initial weights are assigned. First, define

$$g_{i,j} = 2^{-1.5} a_{i1,j} - 2^{-1.5} a_{i2,j}$$

So, $f_{i,j} = 1 + g_{i,j}$.

Let $w_{0,j}$ be the production weight for the j^{th} sample unit. Then the replicate weight $w_{i,j}$ is

$$w_{i,j} = w_{0,j} * f_{i,j} = w_{0,j} * (1 + g_{i,j})$$

Now let's define the estimate and replicate estimates as

$$x_0 = \sum_{j=1}^{n} w_{0,j} \quad \textit{and} \quad x_i = \sum_{j=1}^{n} w_{i,j}$$

The variance of x_0 is then

$$Var(x_0) = \frac{4}{80}\sum_{i=1}^{80}\left(x_i - x_0\right)^2 = \frac{4}{80}\sum_{i=1}^{80}\left(\sum_{j=1}^{n} w_{i,j} - \sum_{j=1}^{n} w_{0,j}\right)^2 = \frac{4}{80}\sum_{i=1}^{80}\left(\sum_{j=1}^{n}\left(w_{i,j} - w_{0,j}\right)\right)^2$$

$$= \frac{4}{80}\sum_{i=1}^{80}\left(\sum_{j=1}^{n}\left(w_{0,j}(1+g_{i,j}) - w_{0,j}\right)\right)^2$$

$$= \frac{4}{80}\sum_{i=1}^{80}\left(\sum_{j=1}^{n} w_{0,j} g_{i,j}\right)^2$$

The FPC-adjusted replicate factor is

$$f_{i,j}^{*} = 1 + \left(2^{-1.5} a_{i1,j} - 2^{-1.5} a_{i2,j}\right) \times FPC = 1 + g_{i,j} \times FPC$$

The variance using the FPC-adjusted replicate factors is

$$Var^{*}(x_0) = \frac{4}{80}\sum_{i=1}^{80}\left(x_i^{*} - x_0\right)^2 = \frac{4}{80}\sum_{i=1}^{80}\left(\sum_{j=1}^{n} w_{i,j}^{*} - \sum_{j=1}^{n} w_{0,j}\right)^2 = \frac{4}{80}\sum_{i=1}^{80}\left(\sum_{j=1}^{n}\left(w_{i,j}^{*} - w_{0,j}\right)\right)^2$$

$$= \frac{4}{80}\sum_{i=1}^{80}\left(\sum_{j=1}^{n}\left(w_{0,j}\left(+ g_{i,j} \times FPC\right) - w_{0,j}\right)\right)^2$$

$$= \frac{4}{80}\sum_{i=1}^{80}\left(\sum_{j=1}^{n} w_{0,j} g_{i,j} \times FPC\right)^2$$

$$= FPC^2 \times \frac{4}{80}\sum_{i=1}^{80}\left(\sum_{j=1}^{n} w_{0,j} g_{i,j}\right)^2$$

$$= FPC^2 \times Var(x_0)$$

$$SE^{*}(x_0) = FPC \times SE(x_0)$$

So, in this simplified example, applying the FPC to the replicate factors yields exactly the original standard error multiplied by the FPC.

Why is the FPC adjustment made to the replicate factor and not on the "outside" of the variance calculation? For simplicity, this example only included one FPC adjustment. The ACS FPC's are defined at the tract level. For an estimate that included persons or housing units crossing multiple tracts, the FPCs would have to be adjusted depending on what tracts the observations comprising estimate were included in. By applying the FPC to the replicate factor, that step has already been taken care of, and the FPC adjustment does not need to be recalculated for each estimate tabulated.