Understanding Variance Estimator Bias in Stratified Two-Stage Sampling

Khoa Dong[1], Tim Trudell[1], Yang Cheng[1], Eric Slud[1,2]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233
[2]University of Maryland, 4176 Campus Drive, College Park, MD 20742

Abstract
The Current Population Survey (CPS) is one of the oldest surveys in the United States and the source of numerous high-profile economic statistics, including the national unemployment rate. Balanced repeated replication (BRR) is a commonly used variance estimation method at the U.S. Census Bureau when there is no design-based variance estimator available. Due to its sampling design, CPS requires collapsing of non-self-representing (NSR) strata to make pseudo-strata in order to implement BRR. These pseudo-strata should ideally contain exactly two perfectly matched primary sampling units (PSUs). This paper examines properties of BRR estimator when estimating variance of response rate of eligible housing units (HUs) in CPS NSR strata. In addition, we will present a bias study of the BRR variance estimator using simulations based on CPS data.

Key Words: Current Population Survey, Variance Estimator, Bias, Balanced Repeated Replication, PSU Matching

1. Introduction

The Current Population Survey (CPS) is a household survey sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics. CPS is the source of many high-profile economic statistics, including the national unemployment rate, and provides data on a wide range of issues relating to employment and earnings. CPS is also a source of information for the study of survey methodology (Technical Paper 66).

CPS monthly sample has about 72,000 households. This sample is designed to produce national and state estimates of labor force characteristics of the civilian noninstitutionalized population 16 years of age or older (CNP 16+). The first stage of sampling involves dividing the United States into primary sampling units (PSUs) – which consist of a single county or group of counties. We define and select PSUs every ten years. There are two types of PSU: self-representing (SR) and non-self-representing (NSR). NSR PSUs are then grouped into strata on the basis of independent information that is obtained from the decennial census or other sources. SR PSUs are always in sample (selected with probability one). For each NSR stratum, one PSU is selected with probability proportional by size. We define the size as CNP 16+ population from Census 2010 data. In 2010 design, 852 PSUs were selected in the first stage including 506 SR and 346 NSR.

After selecting PSUs, we do systematic sampling of clusters of four households within those sampled PSUs. We currently use BRR method for total variance estimation for NSR strata.

Disclaimer: Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

2. Problem Description

Although household response rate is not the primary outcome variable of interest in CPS, it is used as an illustrative variable for purposes of survey design and analysis in this paper. Suppose we want to estimate the monthly response rate, p, and the variance of its point estimator, $V(\hat{p})$, for eligible housing units (HUs) in CPS NSR strata from March 2017 to March 2018. We estimate $V(\hat{p})$ using BRR method as implemented in CPS. The sample is at household level: one record for each sampled HU in each month. The response y_i has binary outcome: 1 for response and 0 for nonresponse. Here we define household response rate as ratio of number of interviewed HUs over number of interviewed plus type A non-interviewed HUs. These type A households are ones that the field representative has determined to be eligible for a CPS interview but for which no usable data were collected. The plots in Figures 1 and 2 below respectively show estimated response rates $\hat{p}, \hat{V}_{BRR}(\hat{p})$ together with $\hat{V}_{BRR}(1-\hat{p})$ in CPS monthly NSR data over the period March 2017 – March 2018.

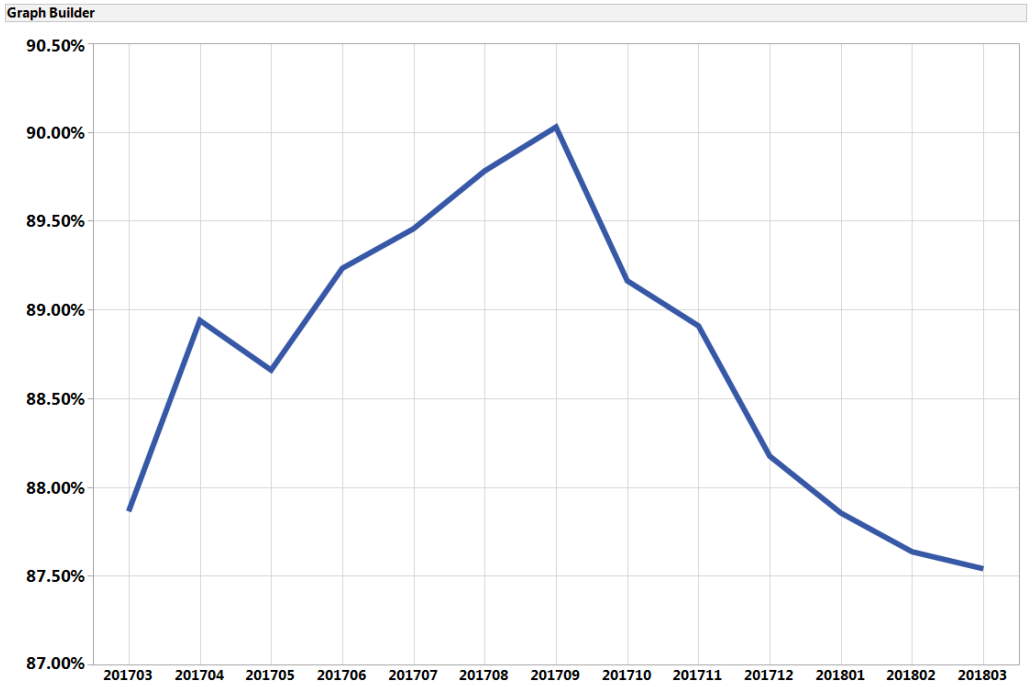

Figure 1: Response rate of eligible HUs in CPS NSR strata March 2017 – March 2018

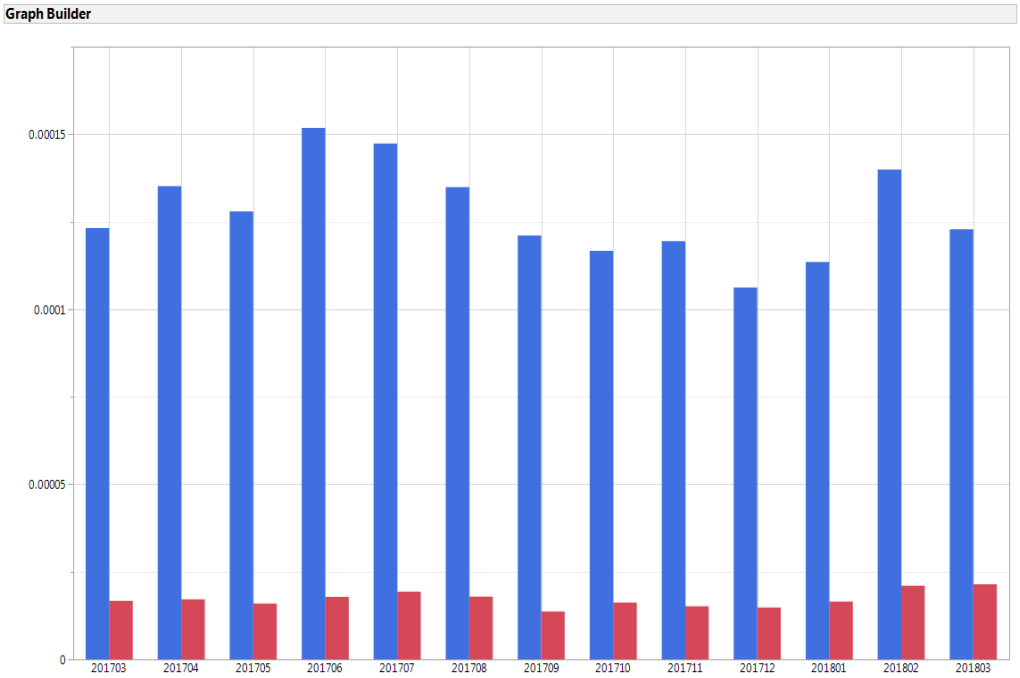

Figure 2: BRR estimated variance of response rate (blue) and nonresponse rate (red) in eligible HUs in CPS NSR strata March 2017 – March 2018

Estimated response rates $\hat{p}$ over this time period are about 87.5% to 90.0%. We expect to see $\hat{V}_{BRR}(\hat{p}) \approx \hat{V}_{BRR}(1-\hat{p})$. However, this symmetric pattern does not show up in Figure 2. Estimated variance of response rate is much higher than that of nonresponse rate. This suggests that our chosen variance estimator introduces bias in some way.

3. Balanced Repeated Replication with Pseudo-Strata

Because only one PSU is selected per NSR stratum and we do systematic sampling within sampled PSUs, there is no direct unbiased design-based variance estimator. CPS currently uses balanced repeated replication (BRR) variance estimation method for NSR strata, which was introduced in McCarthy (1966, 1969). BRR originated from half-sample replication which first emerged at the U.S. Census Bureau in the early 1960s to estimate variances of unadjusted and seasonally adjusted estimates derived from CPS (Wolter p. 110).

Suppose we want to estimate a population total Y and the variance of its point estimator $V(\hat{Y})$. The population is divided into L strata; h indexes the strata. Let $\hat{Y}_h$ be an estimate of stratum h total Y_h. An unbiased estimator of Y is:

$$\hat{Y} = \sum_{h=1}^{L} \hat{Y}_h$$

Fay-method BRR variance estimator has the form (Fay 1984):

$$\hat{V}_{BRR}(\hat{Y}) = \frac{4}{R}\sum_{r=1}^{R}(\hat{Y}_r - \hat{Y})^2$$

where $\hat{Y}_r$ = the r-th replicate whole-population estimate of Y; $\hat{Y}$ = the full sample estimate of Y; R = number of replicates (CPS uses $R = 160$ replicates).

BRR requires selecting two PSUs per stratum with replacement. Since CPS selects only one PSU per stratum, we need to collapse PSUs to make pseudo-strata. These pseudo-strata should ideally contain exactly two perfectly matched PSUs. Consider the simple case when L is even, and we estimate the variance of $\hat{Y}$ by combining the L strata into G groups of two strata each ($L = 2G$). Rewrite $\hat{Y}$ as:

$$\hat{Y} = \sum_{g=1}^{G} (\hat{Y}_{g1} + \hat{Y}_{g2})$$

where $\hat{Y}_{g1}$ and $\hat{Y}_{g2}$ are estimated totals from first and second strata in group g. Hence, the true variance is:

$$V(\hat{Y}) = \sum_{g=1}^{G} [V(\hat{Y}_{g1}) + V(\hat{Y}_{g2})] = \sum_{g=1}^{G} (\sigma_{g1}^2 + \sigma_{g2}^2)$$

The r-th replicate estimate of Y:

$$\hat{Y}_r = \sum_{g=1}^{G} [(1 + 0.5\delta_{gr})\hat{Y}_{g1} + (1 - 0.5\delta_{gr})\hat{Y}_{g2}]$$

δ_{gr} = r-th entry of an appropriately indexed row depending on g of a fixed Hadamard matrix; $\delta_{gr} = 1$ means the first PSU in g-th group is selected, and $\delta_{gr} = -1$ means the second PSU in g-th group is selected. For a given fixed Hadamard matrix, we have $\sum_{r=1}^{R} \delta_{kr} = 0$ ($\forall\, k \neq 1$) and $\sum_{r=1}^{R} \delta_{hr}\delta_{kr} = 0$ ($\forall\, h \neq k$). Note that $\delta_{gr}^2 = 1$ since $\delta_{gr} = \pm 1$.

Next, we will expand $\hat{V}_{BRR}(\hat{Y})$ and show that it is an unbiased estimator for $V(\hat{Y})$ only when the pair of PSUs in each group are perfectly matched.

$$(\hat{Y}_r - \hat{Y})^2 = \sum_{g=1}^{G} \frac{1}{4}\delta_{gr}^2 (\hat{Y}_{g1} - \hat{Y}_{g2})^2 + \sum_{g=1}^{G}\sum_{k\neq g}^{G} \frac{1}{4}\delta_{gr}\delta_{kr}(\hat{Y}_{g1} - \hat{Y}_{g2})(\hat{Y}_{k1} - \hat{Y}_{k2})$$

$$\frac{4}{R}\sum_{r=1}^{R}(\hat{Y}_r - \hat{Y})^2 = \frac{4}{R}\sum_{r=1}^{R}\sum_{g=1}^{G} \frac{1}{4}(\hat{Y}_{g1} - \hat{Y}_{g2})^2 + \frac{4}{R}\sum_{g=1}^{G}\sum_{k\neq g}^{G} \frac{1}{4}(\hat{Y}_{g1} - \hat{Y}_{g2})(\hat{Y}_{k1} - \hat{Y}_{k2})\sum_{r=1}^{R}\delta_{gr}\delta_{kr}$$

Therefore,

$$\hat{V}_{BRR}(\hat{Y}) = \sum_{g=1}^{G} (\hat{Y}_{g1} - \hat{Y}_{g2})^2 = \sum_{g=1}^{G} (\hat{Y}_{g1}^2 + \hat{Y}_{g2}^2 - 2\hat{Y}_{g1}\hat{Y}_{g2})$$

and,

$$E\left\{\sum_{g=1}^{G} (\hat{Y}_{g1}^2 + \hat{Y}_{g2}^2 - 2\hat{Y}_{g1}\hat{Y}_{g2})\right\} = \sum_{g=1}^{G} [V(\hat{Y}_{g1}) + \mu_{g1}^2 + V(\hat{Y}_{g2}) + \mu_{g2}^2 - 2\mu_{g1}\mu_{g2}]$$

$$= \sum_{g=1}^{G} (\sigma_{g1}^2 + \sigma_{g2}^2) + \sum_{g=1}^{G} (\mu_{g1} - \mu_{g2})^2$$

$$= V(\hat{Y}) + Bias^2 \quad \text{where } \sigma_{gh}^2 = \mathrm{V}(\hat{Y}_{gh}) \text{ and } \mu_{gh} = E(\hat{Y}_{gh}) \tag{1}$$

The bias squared term $\sum_{g=1}^{G}(\mu_{g1} - \mu_{g2})^2$ is nonnegative and would add to the expectation of the variance estimate. Again, the bias squared term would be zero if the pair of PSUs in each group were perfectly matched. In CPS, NSR pseudo-strata were formed by combining NSR PSUs into groups of two (or three for states with an odd number of NSR strata). The objective function used to match the NSR PSUs is based on a set of covariates related to civilian labor force (CLF) statistics: unemployment level, CLF level, and children aged 0-17 at or below 200% poverty level.

4. Simulation Design

We ran simulations in which the design variance can be approximated to assess performance of the BRR variance estimator for various response rates $p = 0.03, 0.06, \ldots, 0.99$. We used one month of CPS data (March 2018) with pseudo-strata information. The sample is at household level with sample size $n \approx 15{,}000$. For each household i, we generate response y_i as a Bernoulli trial outcome with success probability p. In many circumstances with independently identically distributed response $y_i's$ and large sample, the pure design variance is very close to the pure model variance.

We ran 5,000 simulations for each p. For each simulation, we compute:

- Total number of eligible households: $\hat{N} = \sum_{i=1}^{n} w_i$ where w_i is base weight of household i.
- Full sample estimated response count: $\hat{Y} = \sum_{i=1}^{n} w_i y_i$
- Replicate r estimated response count: $\hat{Y}_r = \sum_{i=1}^{n} w_i y_i f_{ir}$ where f_{ir} is a function of δ_{gr} and f_{ir} is either 1.5 or 0.5.
- Design variance of $\hat{Y}$: $V_D(\hat{Y}) \approx V_M(\hat{Y}) = V(\sum_{i=1}^{n} w_i y_i) = \sum_{i=1}^{n} w_i^2 V(y_i) = \sum_{i=1}^{n} w_i^2 \, p(1-p)$
- Design variance of $\hat{p} = \frac{\hat{Y}}{\hat{N}}$: $V_D(\hat{p}) = \left(\frac{1}{\hat{N}}\right)^2 V_D(\hat{Y})$. We are treating $\hat{p}$ as a scaled total with known N instead of a ratio estimator here.
- BRR variance of $\hat{Y}$: $\hat{V}_{BRR}(\hat{Y}) = \frac{4}{160}\sum_{r=1}^{160}(\hat{Y}_r - \hat{Y})^2$
- BRR variance of response rate $\hat{p}$: $\hat{V}_{BRR}(\hat{p}) = \left(\frac{1}{\hat{N}}\right)^2 \hat{V}_{BRR}(\hat{Y})$

5. Simulation Results and Discussion

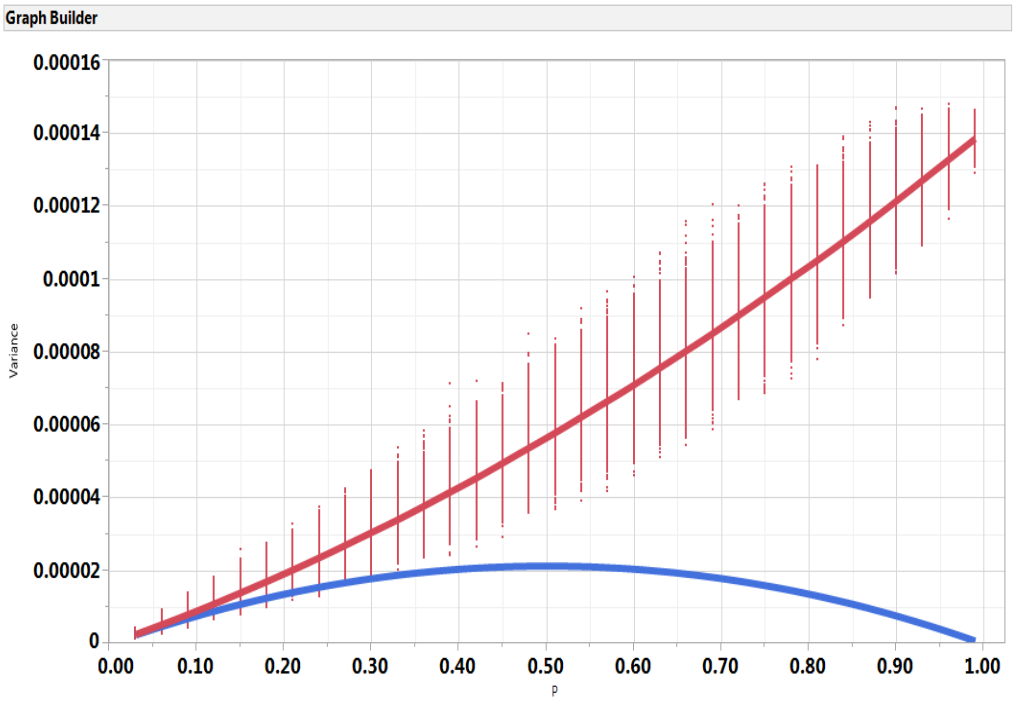

Figure 3: BRR estimated variance (red) versus design variance (blue) for various p

In Figure 3, $\hat{V}_{BRR}(\hat{p})$ computed from each simulation is represented by one red dot. The red curve smoothly connects the average value of those red dots for each p. The blue curve denotes an approximate theoretical variance curve, which is very close to the design variance $V_D(\hat{p})$. There is significant bias in BRR variance estimator as p increases. Even after accounting for variance in simulations, the red curve is pointing in a different direction from the blue curve. We conjecture that the bias is due to the pair of PSUs in each group not being matched well with respect to response rate. To confirm this observation, we estimate using external data the bias term $\sum_{g=1}^{G}(\mu_{g1}-\mu_{g2})^2$ in (1) and subtract it from $\hat{V}_{BRR}(\hat{p})$. The result should be very close to the design variance $V_D(\hat{p})$. To compute μ_{g1}, we have $E(\hat{Y}_{g1}) = E_D[E_M(\hat{Y}_{g1}|s_{g1})] = E_D\left[E_M\left(\sum_{i\in s_{g1}} w_i\, y_i|s_{g1}\right)\right] = E_D\left[E_M\left(\sum_{i\in U_{g1}} w_i\, y_i I_i(i\in s_{g1})|s_{g1}\right)\right] = E_D\left[\sum_{i\in U_{g1}} w_i\, I_i(i\in s_{g1})E_M(y_i)\right] = E_D\left[\sum_{i\in U_{g1}} w_i\, I_i(i\in s_{g1})p\right] = p\sum_{i\in U_{g1}} w_i\, E_D[I_i(i\in s_{g1})] = p\sum_{i\in U_{g1}} w_i\,\pi_i = p\left(\sum_{i\in U_{g1}} 1\right) = p\times N_{g1} = \mu_{g1}$ where N_{g1} is the total number of households in stratum $g1$.

Since we do not have current information on total number of households for strata to exactly calculate $\sum_{g=1}^{G}(\mu_{g1}-\mu_{g2})^2$, we use 2010-design data available in the American Housing Survey (AHS). In addition, there are some slight differences in HU universe between CPS and AHS. We compare CPS-estimated HUs to AHS-estimated HUs in 2010 and apply an adjustment factor of 0.9 to N_{g1} to account for these differences.

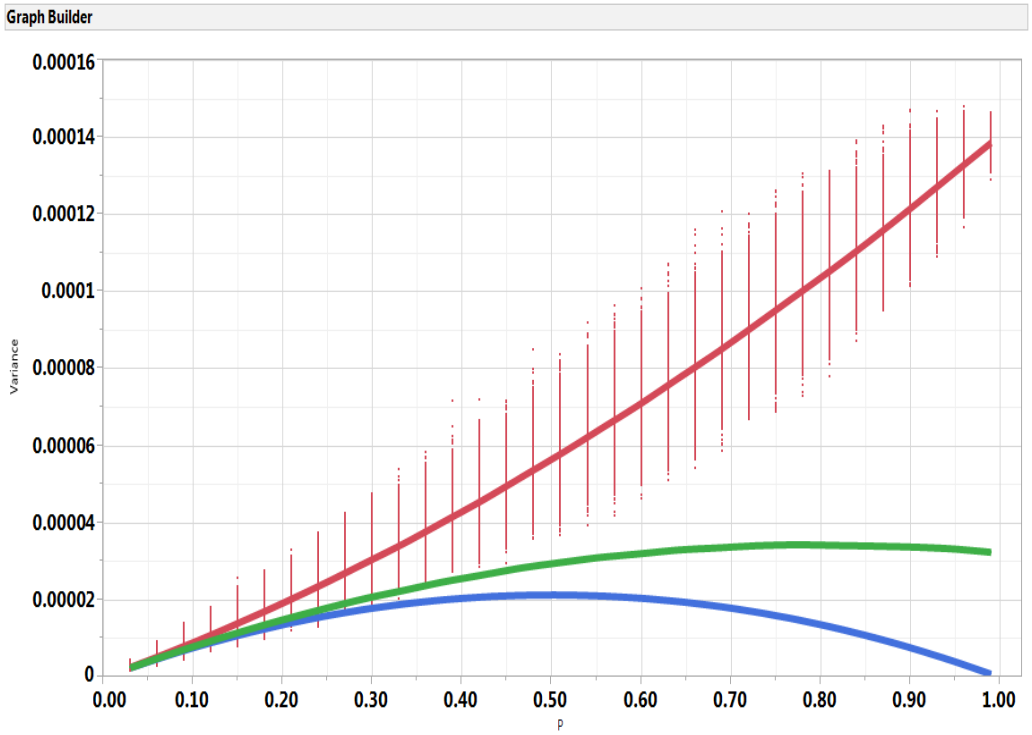

Figure 4: $\hat{V}_{BRR}(\hat{p})$ (red), $V_D(\hat{p})$ (blue), and bias-adjusted $\hat{V}_{BRR,adj}(\hat{p})$ (green)

After adjusting for bias, BRR estimated variance is significantly closer to the design variance. However, there is still a gap between $\hat{V}_{BRR,adj}(\hat{p})$ and $V_D(\hat{p})$ as p gets close to 1 due to: (a) AHS data on total number of HUs in strata is not up-to-date (2010-design information); (b) in CPS the NSR strata were collapsed based on a set of covariates related to civilian labor force statistics but not HU response rate. Although we do not have available information to account for all bias in BRR variance estimator, Figure 4 really confirms our conjecture that this bias issue is wholly due to poor matching of the NSR PSUs.

One of CPS main outputs is the national unemployment rate, and it has been historically below 10.0%. Hence, the bias issue in CPS BRR variance estimator is relatively ignorable. However, when we are interested in another outcome such as labor force participation rate, we should be aware of possible systematic bias in our BRR variance estimator. One way of accounting for bias would be to use recent American Community Survey estimates or modeled estimates based on other external but current data on employment to compute $\sum_{g=1}^{G}(\mu_{g1} - \mu_{g2})^2$ and subtract it from BRR estimated variance.

Acknowledgements

We would like to thank John Jones and Jeffrey Corteville of Demographic Statistical Methods Division and Dr. Yves Thibaudeau of Center for Statistical Research and Methodology at U.S. Census Bureau for useful suggestions and comments.

References

Cheng, Yang (2012) "Overview of Current Population Survey Methodology". *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 3965−3979.

Fay, R.E. (1984) ''Some Properties of Estimates of Variance Based on Replication Methods,'' *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 495−500.

Judkins, David (1990). "Fay's method for variance estimation." *Journal of Official Statistics*, Vol 6, No. 3, 1990

McCarthy, P.J. (1966). "Replication: An Approach to the Analysis of Data from Complex Surveys." *Vital and Health Statistics* Series 2 No. 14

McCarthy, P.J. (1969). "Pseudo-Replication: Half Samples." *Review of the International Statistical Institute*, Vol. 37, No. 3, pp. 239-264

Wolter, K.M. (2008). Introduction to Variance Estimation, New York: Spring-Verlag.

U.S. Census Bureau (2006). Current Population Survey Design and Methodology Technical Paper 66.