# Alternative Variance Estimates in the Current Population Survey and American Community Survey

Michel Boudreaux, Michael Davern, Peter Graven

## Long Abstract

### Introduction

The Current Population Survey's Annual and Social Economic Supplement (CPS) and American Community Survey (ACS) are widely used demographic datasets produced by the U.S. Census Bureau (CB). They are used by researchers for a variety of tasks and by governments to determine appropriations and to simulate alternative policy proposals. Both surveys are based on multi-stage area probability samples. As such, variance estimates calculated assuming that the data were collected as part of a simple random sample (SRS) (the default setting in many statistical packages) produce biased standard errors and invalid statistical inferences.

Census releases a set of replicate weights for the CPS and ACS files. The replicate weights have been released in the ACS since its inception and were added to the CPS in 2009 due to concerns over the validity of previous methods. This was a welcome addition, however, most software packages lack a canned routine for SDR and the method's implementation otherwise require high computing resources. These hurdles make it difficult for unsophisticated analysts to make accurate statistical inferences. The CB also releases a set of variance parameters called generalized variance parameters (or design factors in the ACS). While using these parameters can generate standard errors that are computationally easy, their use in the CPS was shown by Davern et al. (2006)[1] to produce biased results. Since Davern's finding the Census Bureau has adjusted the generalized variance parameters in the CPS, but it remains unknown if this adjustment sufficiently corrected the problem. Furthermore, there has been no demonstration of the relative bias of the ACS design factors.

This paper, using the replicate weights as a gold standard, compares alternative variance estimates in the public use CPS and ACS. We describe the performance of the simple random sample assumption and the variance parameters relative to the SDR estimate. We also describe and evaluate a Taylor series estimate that uses publically available geographic variables as a proxy for true sampling information. In the CPS we define the strata as unique metro areas or the remaining state balance. Households form clusters. This Taylor series approach was first described by Davern and Colleagues (2006) and was found to produce relatively unbiased standard errors compared to a Taylor series estimate using true sampling information obtained on the internal Census file. Here we extend their method for use in the ACS. We define strata as unique Public Use Micro-data Areas (PUMA) and each household or group quarters person is its own cluster.

### Method

[1] Davern, M. et al. "Unstable Infernces? An Examination of Complex Survey Sample Design Adjustments Using the Current Population Survey for Health Services Research" Inquiry. 43 (Fall): 283-297.

In both surveys, we explore 4 variance techniques: SDR, Taylor-series linearization using pseudo-sampling variables, variance parameters, and the simple random sample method. We examine three socio-economic estimates (% with health insurance, % in poverty, mean personal income). Each estimate was computed for 10 selected states and for the U.S. as a whole. The variables were chosen because they express different interclass correlations, an important determinant of the true standard error. Each variance method was implemented in Stata, version 10. SDR was implemented by modifying Stata's jackknife command[2] , although Stata 11 now has a SDR routine, the only canned routine we know of.

Our primary measure of interest is the ratio of the alternative variance estimate to the replicate weight method. Secondly, we were interested in the design effect (the ratio of the alternative variance estimate to the SRS estimate) of each variable under the SDR and Taylor series methods. Since the practical value of each method is not only a function of bias, but also of production cost, we report the CPU time for the replicate weight and the Taylor series method.

## Results

For the proportion with health insurance, initial results from the CPS show that the average ratio of the Taylor series method to the SDR is .94 across all 50 states and the District of Columbia. The GVP method averaged .83, a substantial improvement from previous studies. For poverty the ratio for Taylor Series was .91 and .97 for the GVP. The average ratio of the Taylor series method to the SDR method for personal income was .85, a surprisingly modest improvement over the SRS method (.80).

The Taylor Series method ran approximately 5 times faster than the SDR method on average (across the variables considered) when estimating the national mean and 6.5 times faster when estimating each individual state. The gains in computing efficiency appear to outweigh relatively minor losses in bias.

## Conclusion

The corrected variance parameters and alternative Taylor series produce similar results to SDR for all the variables we considered in the majority of geographic domains considered. Given the high computational costs of SDR, these alternatives are suitable for analysts using public use data. SRS variance estimates are significantly biased downward and generalized variance parameters or design factors perform better than assuming SRS but still tend to be further from gold standard and can be difficult to implement as statistical package routines are not designed to accommodate them.

[2] http://cps.ipums.org/cps/repwt.shtml