VARIANCE ESTIMATION IN THE CPS OVERLAP TEST

Robin Fisher, Jenny Thompson, Bureau of the Census
Edwin Robison, Michael Welch, Bureau of Labor Statistics
Michael Welch, Bureau of Labor Statistics, Washington, DC 20212

Key words: complex sample design; generalized variance function; design effect

1.0 Introduction

The modernization program for the Current Population Survey (CPS) involves the development of a redesigned labor force questionnaire and the use of a completely automated data collection environment. These changes will be introduced into the CPS and thus become the basis for the official U.S. labor force statistics in January 1994.

A major part of this redesign program is the field testing of the new questionnaire and technology during the period July 1992 through December 1993. From this test, referred to as the CATI/CAPI[1] Overlap (CCO) survey, BLS and Census will be able to examine differences between national labor force characteristics that result from the current CPS questionnaire and procedures and those that result from the new methods. A secondary objective of the overlap survey is to enable analysis of effects on the data due to interview modes.

Comparisons of estimates between the CCO and CPS will be made with t-tests and chi-square tests modified for the complex sample designs. The basic challenge of analysis planning was to develop reasonable and efficient methods for variance estimation in a short time frame. The resulting methods, based on modifications to generalized functions of known forms, are presented in this report.

Section 2 reviews the CPS and CCO sample designs and weighting procedures. Variance estimation in the present CPS and the general approach taken for CCO are discussed in section 3, and the estimation formulas are presented in section 4. In the last section, we provide some recommendations and cautionary notes for future applications.

2.0 Survey Designs and Procedures

Current Population Survey. The CPS is collected each month from a multi-stage probability sample of approximately 59,000 occupied, households. Each decade, the entire CPS sample is selected by first dividing the United States into primary sampling units (PSU's) which are counties or groups of contiguous counties. These PSU's are grouped in strata within state boundaries, and, within a stratum, one PSU is selected with probability proportional to population size. Within PSU's, households are grouped by residential characteristics, sorted geographically, and systematically selected in four-unit clusters.

The monthly sample consists of eight rotation groups; each group enters the sample for four consecutive months, leaves the sample for the next eight months, and returns to the sample for four more consecutive months. Thus, 75% of the sampled households are common from month to month, and 50% are common from year to year, and labor force statistics are correlated between time periods. Such correlations need to be taken into consideration in the computation of changes and averages over time.

The CPS estimation process involves several stages of weighting. In the basic weighting procedure, persons are weighted by the sampling interval or inverse of their selection probability; these weights may be further adjusted for subsampling in the field. The basic weights are then adjusted for household nonresponse.

The next step involves a two-stage ratio adjustment. The first stage corrects race distributions in sampled PSU's to agree with state totals. The second stage applies iterative raking procedures to adjust estimated state civilian noninstitutional populations over age 16 (CNP) to

[1]CATI/CAPI refers to computer assisted telephone/personal interviewing

control totals and to adjust national CNP estimates to age-sex-race and Hispanic origin control totals. Weights resulting at this stage are termed first-and-second-stage-combined (FSC) weights.

Finally, a composite estimate is made which consists of a weighted average of two components. The first component is the two-stage ratio (FSC) estimate based on the entire sample from the current month; and the second is the composite estimate for the previous month plus an estimate of month-to-month change based on the six rotation groups common to both months.

Further information on CPS sample design and estimation procedures can be found in Bureau of Labor Statistics (1992), and Robison (1992) provides a recent technical summary.

CATI/CAPI Overlap Survey. The CPS is designed to meet both national and state reliability requirements; however, the CCO survey is a national-based design only and has a sample size of approximately 12,000 occupied households per month. The CCO PSU's are stratified within region (as opposed to states for CPS) and this difference results in a slightly larger proportion of between-PSU variance for CCO.

For the most part, the CCO estimation procedures closely follow those of CPS. However, the first stage factors are calculated by region, not state; and the FSC estimates, though controlled to the same national demographic categories as in CPS, are not adjusted to state totals.

Split Panels. To examine possible mode effects, split panel comparisons are being made within each surveys to compare mode of interview and across surveys to compare questionnaires.

CPS interviews for months-in-sample 1 and 5 are conducted by personal visit, and for months-in-sample 2-4 and 6-8, by telephone. The current questionnaire has been automated, and a small part of the total sample is also interviewed from a centralized CATI facility for months-in-sample 2-4 and 6-8. To enable CPS comparisons between CATI and field, PSU's are selected from large metropolitan areas, and within these PSU's, households are randomly assigned to a test group (eligible for CATI) or a control group (field). Comparing data between these groups, excluding personal visit interviews, provides an indication of the combined effects of centralized and automated interviewing.

In the CCO, sample households within selected PSU's are also randomly assigned to CATI and field panels. The field interviews (personal and telephone) are conducted with laptop computers (CAPI), and this split panel comparison measures a centralization or CATI effect given the new questionnaire.

The effect of the new, fully automated CPS questionnaire, given a CATI environment, is the third mode effect being tested. This comparison uses estimates for months-in-sample 2-4 and 6-8, constructed from CPS-CATI and CCO-CATI panels in PSU's common to both surveys.

In comparing national level estimates between CPS and CCO, the second stage (FSC) estimates are being used. The mode effects estimates, however, are not nationally representative since the CATI PSU's in both surveys are not randomly assigned. Thus, baseweighted estimates are used for the mode comparisons, and these are further adjusted to account for probability of panel assignment.

3.0 Variance Estimation

Although the estimation methods for CPS do not produce unbiased estimates, biases due to ratio adjustments and sources of nonsampling errors are believed to be small enough so that sample-based standard errors can be used to construct useful confidence intervals. Because it would be too costly to develop standard errors for all CPS estimates, generalized variance function (GVF) techniques (see, for example, Wolter, 1985, Chapter 5) are used to calculate sets of standard errors for various types of labor force characteristics.

The generalized variance curves are based on estimates of variances determined through balanced repeated replication (BRR) methods as described in Chapter 3 of Wolter's text. Due to the cost and computer resources involved in the replication process, this method is used to produce

variances for relatively few characteristics, and the Census Bureau last applied the methods to the 1987 data.

The 1987 replicate weight file includes information relevant to estimating the variance for over 600 different characteristics, such as the number of employed males aged 16-19, or the number of unemployed, nonwhite teenagers. From these characteristics or items, about 20 distinct GVF's are produced. These procedures are described in Rothhaas (1993). Decennial census and other information are then used to adjust the 1987 estimates to the desired point in time.

For the overlap survey, methods for variance estimation had to be developed for national comparisons and mode effects studies. The most practical approach for national estimates was to use CPS GVF's with suitable adjustments to account for differences in the CCO design. Each data item for a GVF received its own adjustment, creating, in effect, several hundred GVF's for use in the CCO. Maximum use was made of the 1987 CPS replicate samples and the 1990 decennial census data to estimate the various adjustment factors. For the mode effects, second stage control totals could not be used, and variability of sample size had to be accounted for.

The GVF approach had several advantages over other methods considered, such as resampling. The GVF estimates were known to be stable over time; the curves and methods for updating them were already developed, and analysts were experienced in their use.

4.0 Methodology

In this section, we describe the variance estimation formulas used for estimates from the CPS and the overlap test. At the national level, different adjustment factors are used to modify GVF parameters, predetermined for CPS estimates, to account for differences in the CCO sample design and to account for differences between second stage and composite estimates; a third adjustment factor is developed to account for correlations among estimates when computing time averages for estimates at any level. These adjustment factors are described in section 4.1. For the split panel variance estimates, it is necessary to start with baseweights and make special adjustments to account for the probability of panel assignment. These techniques are discussed in section 4.2.

4.1 Special Variance Adjustments

Converting CPS Variances to CCO. Let $\hat{X}_s$ be an estimator for characteristic X from survey s. From the sample design, we can partition total variance as

$$\mathrm{Var}(\hat{X}_s) = \mathrm{Var}_w(\hat{X}_s) + \mathrm{Var}_b(\hat{X}_s),$$

where the two components denote within-PSU and between-PSU variances respectively. Suppose we know the proportion of within-PSU variance, $p_{w,s}$. Then, we have

$$\mathrm{Var}_w(\hat{X}_s) = p_{w,s}\mathrm{Var}(\hat{X}_s).$$

Applying this to each survey gives

$$\frac{\mathrm{Var}_w(\hat{X}_{cco})}{\mathrm{Var}_w(\hat{X}_{cps})} = \frac{p_{w,cco}\mathrm{Var}(\hat{X}_{cco})}{p_{w,cps}\mathrm{Var}(\hat{X}_{cps})}.$$

Assume the ratio of within-PSU variances is equal to the ratio of sampling intervals. Let $f_{si} = SI_{cco}/SI_{cps}$; we then have

$$\mathrm{Var}(\hat{X}_{cco}) = f_{si}\left(\frac{p_{w,cps}}{p_{w,cco}}\right)\mathrm{Var}(\hat{X}_{cps}).$$

The proportion of within-PSU variance, $p_{w,s}$, is assumed to be stable over time; it is estimated for each survey by using both CPS replicate variances and decennial census data. The replicate variances provide estimates of total variance. Between-PSU variances are calculated for a limited number of characteristics from decennial census data.

Since each characteristic of interest, X_s may not have an exact counterpart in the census computations or in the replicate variances, we let X_s^* be a proxy for X_s. For example, total unemployment may serve as a proxy for male unemployment for between-PSU variance calculations. Denoting replicate and census (between-PSU) variance estimates by additional

subscripting, we can compute the within-PSU variance proportions as follows.

$$\hat{p}_{w,cps} = \frac{Var(X^*_{cps})_{rep} - Var_b(X^*_{cps})_{cens}}{Var(X^*_{cps})_{rep}}$$

$$= \frac{V^*_{w,cps}}{Var(X^*_{cps})_{rep}}, \quad \text{and}$$

$$\hat{p}_{w,cco} = \frac{f_{si} V^*_{w,cps}}{f_{si} V^*_{w,cps} + Var_b(X^*_{cco})_{cens}}.$$

The estimated variance for a CPS composite estimate, $\hat{X}$, is given by the generalized variance function $a\hat{X}^2 + b\hat{X}$, where the predetermined a and b parameters depend on the characteristic. We can then write

$$\hat{Var}(\hat{X}_{cco}) = f_{cco}(a\hat{X}^2_{cco} + b\hat{X}_{cco}), \qquad (4.1)$$

$$\text{where, } f_{cco} = f_{si}\left(\frac{\hat{p}_{w,cps}}{\hat{p}_{w,cco}}\right).$$

Converting Composite Estimates to FSC. Since the generalized variance parameters are determined from replicate variances based on composite estimates, (4.1) can be re-expressed as

$$\hat{Var}(\hat{X}_{cco})_{comp} = f_{cco}(a\hat{X}^2_{cco} + b\hat{X}_{cco}).$$

Next, we need to apply a variance inflation factor to adjust this estimator for use with second stage (FSC) estimates. Again, making use of replicate variances, define

$$f_{fsc} = \frac{Var(X^*_{cps})_{fsc,rep}}{Var(X^*_{cps})_{comp,rep}}.$$

Assume this ratio is stable over time. Then

$$\hat{Var}(\hat{X}_{cps})_{fsc} = f_{fsc}(a\hat{X}^2_{cps} + b\hat{X}_{cps})$$

and

$$\hat{Var}(\hat{X}_{cco})_{fsc} = f_{fsc} f_{cco}(a\hat{X}^2_{cco} + b\hat{X}_{cco}) \qquad (4.2)$$

Converting Monthly Estimates to T-month Averages. An additional conversion factor is required to account for variance reduction when computing variances of averages of monthly levels. The sample overlap in both CPS and CCO induces significant, positive correlations between estimates. Assuming CPS and CCO have the same autocorrelation structure, we can use correlations computed from the replicate samples. Define

$$f_{T,fsc} = \frac{1}{T^2}\sum_h \sum_l Corr(X^*_h, X^*_l)_{rep}, \quad h,l = 1,..,T.$$

Let the estimated average of the characteristic over T months be

$$\hat{\bar{X}}_{T,s} = \frac{1}{T}\sum_h \hat{X}_{h,s}.$$

The factor, $f_{T,fsc}$, can be applied to either survey. To apply to CCO, for example, combine (4.1) and (4.2) to obtain

$$\hat{Var}(\hat{\bar{X}}_{T,cco})_{fsc} = f_{cco} f_{fsc} f_{T,fsc}(a\hat{\bar{X}}^2_{T,cco} + b\hat{\bar{X}}_{T,cco})$$

$$= a'\hat{\bar{X}}^2_{T,cco} + b'\hat{\bar{X}}_{T,cco}. \qquad (4.3)$$

4.2 Split Panel Estimates

The mode effects comparisons address differences in estimates of characteristics between surveys given a common mode of interview, and they address differences between modes of interview given a common survey. In both cases, we can let s denote either the test or control panel. The estimates of characteristics are compared within selected PSU's and are computed by using the baseweights and an additional adjustment accounting for probability of panel selection. (See Thompson, 1993.) Since there is no second stage adjustment, the variances cannot be expected to approach zero for characteristic levels near the population value, and we need to consider an alternative approach to estimating variances.

Consider the following decomposition of total variance.

$$Var(\hat{X}) = Var(E(\hat{X}\,|\,\hat{N})) + E(Var(\hat{X}\,|\,\hat{N}))$$

$$= Var(Wnp) + E(W^2 np(1-p)DEF_{bas})$$

$$= W^2 p^2 Var(n) + DEF_{bas}\, W^2 p(1-p)E(n). \qquad (4.4)$$

In this expression, $\hat{N}$ is an estimator of N (CNP), and $\hat{N}$ = Wn, where W is a constant baseweight; p = X/N is the proportion in the population having the characteristic of interest; and DEF_{bas} is the design effect for the baseweighted characteristic. Equation (4.4) partitions the total variance into two components: the first measures the contribution to variance due to estimating population size, and the second measures the contribution given a fixed population. This latter component is analogous to the variances estimated by GVF's for the national estimates. Now let W = SI, the known sampling interval. If we assume $(SI)E(n) = E(\hat{N}) = N$, then we can express (4.4) as

$$\begin{aligned}\mathrm{Var}(\hat{X}) &= (SI)^2 p^2 \mathrm{Var}(n) + (SI)X(1-p)DEF_{bas} \\ &= p^2 \mathrm{Var}(\hat{N}) + (SI)X\left(1 - \frac{X}{N}\right)DEF_{bas} \\ &= \left(CV^2(\hat{N}) - \frac{(SI)DEF_{bas}}{N}\right)X^2 + (SI)DEF_{bas}X.\end{aligned}$$

We can estimate the parameters in this model using the replicate samples. Again, let asterisks denote replicate estimates, and let

$$\hat{\mathrm{Var}}(\hat{X}) = a^*\hat{X}^2 + b^*\hat{X}, \qquad (4.5)$$

where

$$a^* = \hat{CV}^2(N^*_{cps}) - \frac{(SI_{cps})DEF^*_{bas}}{N^*_{cps}},$$

$$b^* = (SI_{cps})DEF^*_{bas},$$

and

$$DEF^*_{bas} = \frac{\mathrm{Var}(X^*_{cps})_{rep} - p^{*2}\mathrm{Var}(N^*_{cps})_{rep}}{(SI_{cps})X^*_{cps}(1-p^*)}.$$

We assume the components of the a* and b* parameters are stable over time and that the design effect is the same for CPS and CCO. To apply (4.5) to CCO, one need only change the sampling interval in both parameters.

Equation (4.5) provides an estimator of a variance of a baseweighted CPS characteristic using the full sample. We need to modify this estimator to account for panel size and probability of panel assignment. Without loss of generality, assume panel s is from the CPS sample. Let $\hat{X}_{js}$ be the baseweighted estimate in PSU j, panel s, and let $\hat{N}_{js}$ be the corresponding baseweighted estimate of CNP. Define $P_j(s)$ as the probability that the sampled household in PSU j is assigned to panel s and let

$$\hat{X}_s = \sum_j \frac{\hat{X}_{js}}{P_j(s)}$$

and

$$\hat{N}_s = \sum_j N_{js}.$$

Then, if estimates are uncorrelated across PSU's,

$$\begin{aligned}\mathrm{Var}(\hat{X}_s) &= \mathrm{Var}\left(\sum_j \frac{\hat{X}_{js}}{P_j(s)}\right) \\ &= \sum_j \frac{\mathrm{Var}(\hat{X}_{js})}{P_j^2(s)}.\end{aligned}$$

If we further assume

$$\mathrm{Var}(\hat{X}_{js}) = \frac{N_{js}}{N_s}\mathrm{Var}\left(\sum_h \hat{X}_{hs}\right),$$

then it follows that

$$\mathrm{Var}(\hat{X}_s) = \mathrm{Var}\left(\sum_h \hat{X}_{hs}\right)\frac{1}{N_s}\sum_j \frac{N_{js}}{P_j^2(s)}. \qquad (4.6)$$

We now make the assumption that the latter variance term in (4.6) can be modeled by the relationship expressed in (4.5). That is, let

$$\hat{\mathrm{Var}}\left(\sum_h \hat{X}_{hs}\right) = a^*_s\hat{X}_s^2 + b^*\hat{X}_s, \qquad (4.7)$$

where

$$a^*_s = \hat{CV}^2(\hat{N}_s) - \frac{(SI_{cps})DEF^*_{bas}}{\hat{N}_s}.$$

However, we do not have an estimate of $CV(\hat{N}_s)$. If we assume

$$CV(\hat{N}_s) = \frac{N^*_{cps}}{N_s} CV(N^*_{cps}),$$

then apply (4.5) through (4.7), we have

$$\hat{Var}(\hat{X}_s) = \left(\frac{N^*_{cps}}{\hat{N}_s} a^* X_s^2 + b^* \hat{X}_s \right) \frac{1}{\hat{N}_s} \sum_j \frac{\hat{N}_{js}}{P_j^2(s)} \quad . \tag{4.8}$$

Both a* and b* parameters from (4.5) can now be applied. Note that the a* parameter in (4.8) is modified by an adjustment for the baseweighted CNP in the panel, and both parameters are modified by a factor accounting for probability of panel selection.

5.0 Conclusions

CPS modernization includes a redesigned questionnaire and computer-assisted interviewing for implementation in January 1994. To analyze the effects of the new questionnaire and data collection technology on labor force estimates, an overlap sample is being conducted during the period July 1992 through December 1993. For national level comparisons, variance estimates are based on the CPS GVF's, adjusted to account for second stage estimation and differences in sample designs. Variances for interview mode effect estimates are modeled at the PSU level to account for random sample size and split panel assignment.

While generalized variances are stable and easy to apply, there are several limitations on the methods used. The overlap GVF's, each based on a single characteristic from the original CPS GVF curve, are useful only in a restricted range of the characteristic; occasionally, an item needs to be subdivided, say by age or race, and the GVF's become less reliable. Some CCO data items are new and are difficult to link to one of the given curves. Additionally, the various adjustments are based on assumptions or models that need empirical validation.

The new questionnaire entails some modifications to traditional data items. Additional risk in applying CPS GVF's is taken when the variance behavior of new items do not resemble that of the past. Automation of the interviewing process and other operational changes could also change the behavior of variances. For example, if response variability on certain characteristics is reduced as intended, the component of nonsampling error for these characteristics, which currently contributes to estimated variances, could be lower for the new procedures.

While the GVF methods are being applied to analyze differences between old and new questionnaire results during the overlap period, efforts to evaluate and improve the reliability of CPS variance estimation procedures continue. After the 1994 implementation of the new CPS survey and processing system, BLS and Census are planning for increased use of replication methods. However, GVF's, based on new replicate variances, will continue to provide error measures for many CPS estimates.

6.0 References

Bureau of Labor Statisics (1992), *BLS Handbook of Methods* (Bulletin 2414), Washington, DC: U.S. Dept. of Labor.

Robison, E. (1992), "The Currrent Population Survey - Technical Summary of Design and Methodology," internal memorandum, Office of Employment and Unemployment Statistics, Bureau of Labor Statistics, Washington, DC: U.S. Dept. of Labor.

Rothhaas, R. (1993), "CPS New and Improved Standard Error Parameters for Labor force Characteristics," internal memorandum, Demographic Statistical Methods Division, Bureau of the Census, Washington, DC: U.S. Dept. of Commerce.

Thompson, J. (1993), "CCO Weighting and Estimation: Weighting for Mode Effects Analysis," internal memorandum, Demographic Statistical Methods Division, Bureau of the Census, Washington, DC: U.S. Dept. of Commerce.

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