CURRENT POPULATION SURVEY VARIANCE PROPERTIES

Carol A. Gunlicks, Jeffrey S. Corteville, Khandaker Mansur, U.S. Bureau of the Census
Carol A. Gunlicks, U.S. Bureau of the Census, Washington, D.C. 20233 cgunlick@ccmail.census.gov

Key Words: Replication, Variances, Correlations

I. Introduction

This paper discusses the variance properties of the 1990 design Current Population Survey. Monthly variance estimates and other variance properties such as coefficients of variation, design effects, and month-to-month change variances are presented for major labor force characteristics at the national level. Quarterly, semiannual, and annual estimates of variance are also presented. This paper updates an earlier paper based on the 1970 design (Train, et al. 1978)

II. Description of the Current Population Survey

Overview

The Current Population Survey (CPS) is the nation's primary source of information on the labor force status of the population. Conducted monthly by the Bureau of the Census for the Bureau of Labor Statistics, the CPS is the source of one of the best known economic statistics - the nation's unemployment rate. The CPS is a two-stage sample design based on data from the 1990 decennial census and updated with new construction. The 2007 PSUs are grouped into strata within each state. Nationally, there are 428 PSUs in strata by themselves. These strata are self-representing (SR) and are generally the most populous PSUs in each state. The remaining PSUs are grouped by similar characteristics into non-self representing (NSR) strata. Then one PSU is selected from each stratum with the probability of selection proportional to the population of the NSR PSU.

Within-PSU Sampling

The CPS uses systematic sampling. Because the sample design is state based, the sampling ratio differs by state and sometimes within a state. It depends on state population size as well as both national and state reliability requirements.

Rotation of the Sample

Each monthly sample is divided into eight representative subsamples called rotation groups. A given rotation group is interviewed for a total of eight months, divided into two four-month periods. It is in sample four consecutive months, is not in sample the following eight months, and returns to the sample for another four consecutive months (4-8-4). In each monthly sample, one of the eight rotation groups is in the first month of enumeration, another rotation group is in the second, and so on. Under this system, 75 percent of the sample is common from month to month and 50 percent from year to year for the same month. This procedure provides a substantial amount of month-to-month and year-to-year overlap in the sample, thus providing better estimates of change and reducing discontinuities in the series of data without burdening any specific group of households with an unduly long period of inquiry.

Weighting and Estimation Procedures

Unbiased Estimate:

The estimation procedure involves weighting the data from each sample person by the inverse of the probability of the person being in the sample. This weight is called the baseweight. It gives a rough measure of the number of actual persons that the sample person represents. Since 1985, most sample persons within the same state have had the same probability of selection. Through the series of estimation steps outlined below, the selection probabilities are adjusted for noninterviews and survey undercoverage. Data from previous months are incorporated into the estimates through the composite estimation procedure. The unbiased estimate is the estimate before any of the following adjustments.

Noninterview Adjustment:

The weights for all interviewed households are adjusted to account for occupied sample households for which no information was obtained because of absence, impassable roads, refusals, or unavailability of the respondents for other reasons.

Ratio Adjustments:

The distribution of the population selected for the sample may differ somewhat, by chance, from that of the population as a whole in such characteristics as age, race, sex, and state of residence. Because these characteristics are closely correlated with labor force participation and other principal measurements made from the sample, the survey estimates can be substantially improved when weighted appropriately by the known distribution of these population characteristics. This is accomplished through two stages of ratio adjustment.

First-stage Ratio Adjustment:

The procedure corrects for differences that existed in each state at the time of the 1990 census between 1) the race distribution of the population in sample NSR PSUs and 2) the race distribution of all NSR PSUs.

Second-stage Ratio Adjustment:

This procedure substantially reduces the variability of estimates and corrects, to some extent, for undercoverage of persons and housing units in the CPS. The CPS sample-based estimates of population match independent population controls.

Composite Estimation Procedure:

The primary purpose of compositing is to reduce the variance of estimates of month-to-month change, by taking advantage of the overlap in sample from month to month. The (A=0.2, K=0.4) composite estimator was used.

For further information on the sample design and estimation, see *Employment and Earnings* published by the U.S. Department of Labor, Bureau of Labor Statistics.

III. Types of Errors included in Variance Estimates

All sample surveys are subject to a variety of errors which can occur at any stage of the survey process. These stages include frame creation, sample selection, data collection, data processing, and reporting and analyzing the results. Errors are usually classified as sampling errors or nonsampling errors. Each of these errors can result in either variance or bias.

Sampling error is easier to measure and control, since it results mainly from conducting the survey on a sample rather than the entire population of interest. The full extent of nonsampling error is unknown. Nonsampling errors in surveys can be attributed to many sources, e.g., the inability to obtain information about all persons in the sample; differences in the interpretation of questions; inability and unwillingness of respondents to provide correct information; inability to recall information; errors made in collecting and processing the data; errors made in estimating values for missing data; and failure to represent all sample households and all persons within sample households.

Nonsampling errors are also introduced through the 4-8-4 sample rotation scheme in the form of time-in-sample bias. In the 4-8-4 rotation scheme, the first and fifth interviews occur in person and the others typically occur over the telephone. It is well known that repeated interviewing of the same persons can frequently change response patterns. For some characteristics, the expected value of the estimates varies depending on the number of times a rotation group has been in sample (Bailar 1975). The time-in-sample bias also has an effect on both the ratio and composite estimators.

In general, the variances presented in this paper include components of sampling error and nonsampling error. Some examples are the variability in the nonresponse rate, response variance, interviewer variance, time-in-sample bias, and mode-of-interview effects. Unfortunately, the variance estimation procedure does not permit us to isolate many of these individual components.

IV. Description of Replication Methodology

The variance of an estimate based on any probability sample using any estimation procedure no matter how complex may be estimated by the method of replication. This method requires that the sample selection, the collection of data, and the estimation procedures be independently carried through (replicated) several times. The dispersion of the resulting replicate estimates can be used to measure the variance of the full sample. There are many different types of variance estimators based on replication including jackknife, balanced half sample, and successive difference.

Description of the Successive Difference Method

The general goal of the 1990 variance estimation methodology is to produce comparable variances and covariances for each month over the entire life of the design. Periodic maintenance reductions in the sample size and the continuous addition of new construction to the sample complicate the strategy needed to achieve this goal. However, research has shown that variance estimates are not adversely affected as long as the cumulative effect of the reductions is less than 20% of the original sample size (Kostanich 1996). Assigning all future new construction sample to replicates when they are originally defined provides the basis for consistency over time in the variance estimates.

The approach to estimating the 1990 design variances is called successive difference replication. The theoretical basis for the successive difference method was discussed by Wolter (1984) and extended by Fay (Fay and Train 1995) to produce the successive difference replication method used for the CPS. The following is a description of the application of this method. Successive USUs[1] (ultimate sampling units) formed from adjacent hit strings are paired in the order of their selection to capture the systematic nature of the CPS within-PSU sampling scheme. For each USU within a PSU, two pairs of neighboring USUs are defined based on the order of selection -- one with the USU selected before and one with the USU selected after it, e.g., (USU1, USU2), (USU2, USU3), (USU3, USU4), etc.

Assigning Replicate Factors

Total variance is composed of two types of variance, the variance due to sampling of housing units within PSUs (within-PSU variance) and the variance due to the selection of a subset of all NSR PSUs (between-PSU variance). Replicate factors are calculated using a 160 by 160 Hadamard orthogonal matrix[2]. Each column of the matrix forms a replicate and the cells of the assigned rows generate the replicate factors. To produce estimates of total variance, replicates are formed differently for SR and NSR sample. Note that between-PSU variance cannot be estimated directly using this methodology. Rather, it is

---

[1]An ultimate sampling unit is usually a group of four neighboring housing units.

[2]Rows 1 and 81 have been dropped from the matrix.

the difference between the estimates of total variance and within-PSU variance.

Total Variance

For total variance, estimating the NSR component of variance requires two PSUs per stratum. Because CPS only selects one PSU per stratum, we need to collapse NSR strata. NSR strata are collapsed within each state. One NSR PSU in each collapsed-stratum is randomly assigned a replicate factor of 1.5; the other is assigned a replicate factor of 0.5. These factors are assigned using one row of a Hadamard matrix for each collapsed-stratum. The factors are further adjusted to account for unequal sizes of the original strata within the collapsed-stratum (Wolter 1985). In most cases these collapsed-strata consist of a pair of strata except where an odd number of strata within a state forces a triplet to be formed. In this case two rows from the Hadamard matrix are assigned to the collapsed-stratum resulting in replicate factors of 0.5, 1.7, and 0.8 or 1.5, 0.3, and 1.2 for the three PSUs. All USUs in a collapsed-stratum are assigned the same row(s).

In SR sample, two rows of the Hadamard matrix are assigned to each USU creating replicate factors.

$$f_{ir} = 1 + (2)^{-\frac{3}{2}} a_{i+1,r} - (2)^{-\frac{3}{2}} a_{i+2,r}$$

where $a_{i,r}$ = a number in the Hadamard matrix (+1 or -1) for the i^{th} USU in the r^{th} replicate ($r = 1,\ldots,160$). This formula yields replicate factors of 1.7, 1.0, or 0.3.

The unbiased weights (baseweight x special weighting factor) are multiplied by the replicate factors to produce unbiased replicate weights. These unbiased replicate weights are further adjusted through the noninterview adjustment, the first-stage ratio adjustment, and the second-stage ratio adjustment just as the full sample is weighted. The compositing is also done using replication. A variance estimator for the characteristic of interest is a sum of squared differences between each replicate estimate $(\hat{Y}_r)$ and the full sample estimate $(\hat{Y}_0)$. The formula[3] is:

$$\mathrm{Var}(\hat{Y}_0) = \frac{4}{160} \sum_{r=1}^{160} (\hat{Y}_r - \hat{Y}_0)^2$$

Within-PSU Variance

The variance estimator for total variance with some modification can also be used for within-PSU variance. The replicate factors assigned to SR sample remain the same. For NSR sample, different row assignments are made for USUs to form pairs of USUs in the same manner that was used for the SR assignments. Thus, for within-PSU variance all USUs (both SR and NSR) have replicate factors of 1.7, 1.0, or 0.3.

We plan to make replicate weights available on the public use files for basic CPS and for supplements. These weights will differ slightly from the internal replicate weights described above in that there will be only 80 replicates rather than 160. These public use replicate weights also have been slightly modified to maintain confidentiality in small areas. Replicate weights will allow the user to calculate variance estimates for any available characteristic on their data file, including variance estimates of totals, percentages, ratios, means, or medians. Using these replicate variances for small areas, especially rural states, may produce unreliable variance estimates. More research is needed for these.

V. Variance Results

The sample design and the estimation procedures used in the CPS have changed many times during the history of the survey. Changes occur when reliability requirements are revised. To develop an efficient sample design, we estimate the components of variance attributable to each of the several stages of sampling and investigate the effect of the separate steps in the estimation procedure on the variance estimate.

Tables 1-4 show variance estimates computed, using replication methods, by stage of estimation. We also computed variance estimates by type (total, between- and within-PSU). Averages over fifteen months have been used to improve the reliability of the estimated monthly variances. Correlations for estimates separated by n months and design effects are also included. More tables are available from

Carol.A.Gunlicks@ccmail.census.gov

A. Variances on Monthly Estimates

Table 1 compares composited estimates of level and the associated variances to those of the FSC estimates (the estimate after both the first- and second-stage adjustments are applied). For example, the composited estimate of unemployed persons of is 7,320,000 and its standard error is 143,300. The CV for this characteristic is 1.96%. The FSC estimate for the same characteristic is 7,400,000 with standard error 146,800. The CV is 1.99%. CVs for the FSC and

---

[3]Usually balanced half-sample replication uses replicate factors of 2 and 0 with the formula, $\mathrm{Var}(\hat{Y}_0) = 1/k \sum_{r=1}^{k} (\hat{Y}_r - \hat{Y}_0)^2$, where k is the number of replicates. The factor of 4 in our variance estimator is the result of using replicate factors of 1.5 and 0.5. Note that the replicate factors 1.7, 1.0, and 0.3 for the self-representing portion of the sample were specifically constructed to yield "4" in this formula in order that the formula remains consistent between SR and NSR areas (Fay and Train 1995).

composite estimates are similar for most characteristics, indicating that compositing tends to have a small effect on the relationship between the level and standard errors of most estimates.

Although not shown in tables here, comparisons of the relative variance at a given stage of estimation to that of the unbiased estimate have been made. Recall that the unbiased estimate is the estimate before the noninterview adjustment. The relative variance (equal to the square of the coefficient of variation) of the unbiased estimate of unemployment level is 4.17×10^{-4}. The relative variance of the noninterview estimate for this characteristic would be 0.99 times as large. If the first stage adjustment is also included, the relative variance is raised to 1.07 times the size of the relative variance for the unbiased estimate of level. Including the second stage of estimation reduces the relative variance factor to 0.95. After the second-stage ratio adjustment, a composite estimator is used to improve estimates of month-to-month change by taking advantage of the 75 percent of the total sample that continues from the previous month. After compositing the relative variance factor for unemployment decreases to 0.93. The second-stage adjustment appears to greatly reduce the total variance, as intended. This is especially true for large characteristics such as employed - nonagriculture and civilian labor force. Without the second stage, the relative variances of these characteristics would be 5, 6, or even 9 to 10 times as large. For smaller groups such as unemployed and employed in agriculture, the second-stage adjustment does not have as dramatic an effect.

The application of the second-stage adjustment on the first-stage estimates of the civilian noninstitutional population (CNP) causes the estimates to converge to population controls and causes variances for most of these characteristices to be zero. The FSC relative variance factors for CNP are zero for CNP, because we are controlling to these estimates in the second-stage ratio adjustment. They are not zero for the teenage category, because the estimates don't completely converge to the control totals in these categories.

Within-PSU variance and total variance are computed as described previously in this paper. Between-PSU variance is estimated by subtracting the within-PSU variance from the total variance. Due to variation of the variance estimates, the estimate of between-PSU variance is sometimes negative.

For most characteristics the proportion of the total variance due to sampling housing units within PSUs (within-PSU variance) is larger than that due to sampling a subset of NSR PSUs (between-PSU variance). In fact, for many characteristics, the within-PSU component accounts for over 90 percent of the total variance. For civilian labor force and not in labor force characteristics almost all of the variance is due to sampling housing units within PSUs. For composited estimates of total and white employed in agriculture, the within-PSU component still accounts for 67 to 68 percent of the total variance, while the between-PSU component accounts for the remaining 32 to 33 percent.

Another way of looking at monthly estimates is through design effects. A design effect (deff) is the ratio of the variance from a complex sample design to the variance of a simple random sample design. Table 2 compares the design effects for the total variance for selected labor force characteristics. The deffs in this table were computed by solving $Var(\hat{X}) = [1/n][N^2PQ(deff)]$ for deff and replacing N/n in the formula with an estimate of the national sampling interval. Estimates of P and Q were obtained from the fifteen months of data.

For unemployed the deff is 1.35 for the uncomposited (FSC) estimate and 1.30 for the composited estimate. This means that the design of the CPS (including the sample selection and estimation) increases the variance by about 30 percent over the variance of an unbiased estimate based on a simple random sample of the same size. On the other hand, for the civilian labor force the design of the CPS decreases the variance by about 14 percent. Note that the deffs for composited estimates are generally lower than those for the FSC estimates indicating again the tendency of the compositing to reduce the variance of most estimates.

B. Variances on Month-to-Month Change

The variance estimates included in Table 3 are averages of 14 month-to-month change estimates (January 1996-March 1997). The table compares the efficiency of the FSC month-to-month change variance estimates with those of the composited estimates. The variance factor is the ratio of the variance of the composite estimate of month-to-month change to the variance of the FSC estimate of month-to-month change. As expected the variance factors for month-to-month change are all less than one, indicating the compositing reduces the variance of all month-to-month change estimates. Note that compositing reduces the variance more for characteristics that are found in large proportions in the sample, e.g., employed, than for those found in small proportions, e.g., unemployed.

C. Variances on 3-, 6-, 9-, and 12-month Averages

The average FSC and composite variances over periods up to one year show a decrease when compared to the monthly variance estimate. The averages computed for the various time periods are based on the periods of publication, e.g. the 3-month averages include January through March 1996 and 1997, April through June 1996, July through September 1996, and October through December 1996. If the samples were independent for each month, we would expect the ratio of the average variance of estimates over 3 months to be .33, but the month-to-month overlap in the sample causes the ratios to be larger (approximately .45 to .88 for the 26 characteristics we considered). As we would expect, the ratios decrease consistently as the sample is averaged over more months. Ratios over 12 months decrease to between .17 and .70.

D. Correlations

Table 4 provides the average correlations between composite estimates separated by n months (n=1,...,14). For each value of n, the correlation shown in the table is an average of the correlations computed for each of the 15/n pairs of months. The correlations are low for unemployed but higher for employed and civilian labor force. We don't expect much correlation for lags 4 through 8, because there is no overlap in sample during these periods. Some correlation in lags 4 through 8 is probably the result of geographic clustering in our within-PSU sample. Correlations are highest for the lowest lags, drop off as the lag increases and then increase slightly for the higher lags. This reflects the amount of overlap from month-to-month due to the rotation of the eight panels in the CPS sample.

VI. Future Plans

Designing and implementing a variance estimation system for a survey as complex as CPS was a very challenging task which took about 3 years. The system was designed with enough flexibility to accommodate most of the changes in sample design and estimation procedures that are likely to occur in the CPS during the next decade. The system has been used to estimate variances for estimates produced with the new composite weights, scheduled for release in December 1997. The system is currently being used to estimate variances for earnings and union membership characteristics, which are derived from questions asked of two rotation groups each month. With some modifications, the system also has been used to compute variances for the CPS Tobacco-Use supplement, the American Travel Survey, and the American Community Survey.

Also, as mentioned earlier, data users external to Census and the Bureau of Labor Statistics will eventually have access to replicate weights, so they can compute variance estimates for any type of estimate from variables on the public use data files.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authos and do not necessarily reflect those of the Census Bureau.

The authors thank Harland Shoemaker for his help in the planning and organization of this paper.

Supporting Materials:

Bailar, Barbara A. (1975) "The Effects of Rotation Group Bias on Estimates From Panel Surveys," *Journal of the American Statistical Association*, **349**, 23-30.

Fay, Robert E., (1984) "Some Properties of Estimates of Variance Based on Replication Methods," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, D.C., 495-500.

________ (1989) "Theory and Application of Replicate Weighting for Variance Calculations," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, 212-217.

Fay, Robert; Dippo, C.; and Morganstein, D. (1984), "Computing Variances from Complex Samples with Replicate Weights," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, D.C., 489-494.

Fay, Robert, and Train, G. (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," *Proceedings of the Section on Government Statistics*, American Statistical Association, Alexandria, VA, 154-159.

Fisher, Robin, and McGuinness R., "Correlations and Adjustment Factors for CPS (VAR80_1)," Bureau of the Census internal documentation memorandum, January 6, 1993.

Hanson, Robert H. (1978), *The Current Population Survey: Design and Methodology*, U.S. Department of Commerce, Bureau of the Census, Technical Paper 40, Washington, D.C.: U.S. Government Printing Office.

Kostanich, Donna, Memorandum from, "Proposal for Assigning Variance Codes for the 1990 CPS Design (VAR90-22)," Bureau of the Census internal documentation memorandum, June 17, 1996.

Lent, Janice (1991), "Variance Estimation for Current Population Survey Small Area Estimates," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, D.C., 11-20.

Train, George; Cahoon, L.; Makens, P. (1978), "The Current Population Survey Variances, Inter-Relationships, and Design Effects," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, D.C., 443-448.

U.S. Department of Commerce, Bureau of the Census, U.S. Department of Labor, Bureau of Labor Statistics, *Technical Paper 63: The Current Population Survey, Design and Methodology*, forthcoming.

U.S. Department of Labor, Bureau of Labor Statistics, *Employment and Earnings (1997)*, 43:1, Washington, D.C.: U.S. Government Printing Office.

Wolter, Kirk (1984), "An Investigation of some Estimators of Variance for Systematic Sampling," *Journal of the American Statistical Association*, 79, 781-790.

_______ (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag New York Inc.

Table 1. Reliability of Monthly Estimates: (January 1996 - March 1997 Monthly Average)

Civilian- Noninstitutional Population 16 years old and over	First and Second Stage Combined			Composited		
	Estimate (x 10^6)	Standard Error (x 10^4)	Coefficient of variation (%)	Estimate (x 10^6)	Standard Error (x 10^4)	Coefficient of variation (%)
Unemployed	7.40	14.68	1.99	7.32	14.33	1.96
Employed - Agriculture	3.38	16.01	4.76	3.36	15.17	4.53
Employed - Nonag.	123.53	35.05	0.28	123.43	32.55	0.26
Civilian Labor Force	134.30	31.62	0.24	134.11	29.37	0.22
Not in Labor Force	66.65	31.62	0.47	66.84	29.37	0.44

Table 2 . Design Effects after Various Stages of Weighting: (January 1996 - March 1997 Monthly Average)

Civilian- Noninstitutional Population 16 years old and over	Design Effect after Stage of Estimation				
	UB[1]	NI[2]	FS[3]	SS[4]	Comp.
Unemployed	1.18	1.25	1.34	1.35	1.30
Employed - Agriculture	3.07	3.20	3.52	3.44	3.10
Employed - Nonag.	4.43	4.52	5.99	1.15	0.99
Civilian Labor Force	5.42	5.50	7.77	1.00	0.86
Not in Labor Force	2.66	2.69	3.61	1.00	0.86

[1] Unbiased
[2] Noninterview
[3] First Stage
[4] Second Stage

Table 3. Effect of Compositing on Month-to-Month Change Variance Factors: (January 1996-March 1997 Monthly Average)

Civilian Noninstitutional Population 16 years old and over	Variance of Composited Estimate of Change[1] (x 10^9)	Variance Factor[2]
Unemployed	25.33	0.94
Employed - Agriculture	6.89	0.71
Employed - Nonag.	57.47	0.75
Civilian Labor Force	52.83	0.78
Not in Labor Force	52.83	0.78

[1] The arithmetic mean of the 14 variance estimates.
[2] The ratio of the variance of the composited estimate to the variance of the FSC estimate.

Table 4. Correlations Between Composited Estimates Separated by n Months: (January 1996 - March 1997 Monthly Average)

Civilian- Noninstitutional Population 16 years old and over	n													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Unemployed	0.38	0.22	0.13	0.06	0.03	0.04	0.02	0.04	0.10	0.11	0.06	0.03	0.07	0.07
Employed - Agriculture	0.85	0.76	0.68	0.60	0.59	0.59	0.58	0.58	0.60	0.63	0.65	0.67	0.67	0.58
Employed - Nonag.	0.73	0.57	0.44	0.35	0.30	0.29	0.27	0.24	0.24	0.27	0.30	0.29	0.21	0.17
Civilian Labor Force	0.69	0.52	0.38	0.28	0.21	0.17	0.14	0.16	0.16	0.19	0.23	0.23	0.20	0.11
Not in Labor Force	0.69	0.52	0.38	0.28	0.21	0.17	0.14	0.16	0.16	0.19	0.23	0.23	0.20	0.11