Replicate Control Totals

David Judkins[1], Varma Nadimpalli[1], and Samson Adeshiyan[2]
Westat[1]
U.S. Bureau of Census[2, *]

Keywords: Variance Estimation, Raking, Poststratification

1. Introduction

One of the standard approaches to variance estimation in sample survey inference is to use replication methods such as the Jackknife or balanced repeated replications (BRR) (Wolter, 1985). These methods are particularly useful when complex estimation procedures such as poststratification, raking, and calibration (Särndal, Swensson, and Wretman, 1992) are employed since the procedures can often just be run on each replicate. The use of resampling methods for variance estimation is well established when poststratification involves known control totals (Lemeshow, 1979; Ernst and Williams, 1987; and Valliant, 2004). However, the problem is more difficult when sample-based control totals are used, where estimates from one survey are used as part of the estimation process for another survey.

Such techniques may be employed for two reasons. First, variances on estimates from a smaller survey can often be reduced through poststratification, raking or calibration to control totals from the larger survey. Second, bias might also be reduced if the larger survey has better coverage or higher response rates. When such techniques are employed, they often use control totals from the monthly Current Population Survey (CPS). A few solutions to this problem have been reported (Nadimpalli, Judkins and Chu [NJC], 2004, and Judkins, Mosher, and Botman, 1991), but none of these solutions have been totally satisfactory because they failed to take into account the covariances among CPS estimates of various population domains (i.e., the control totals). In this paper, we demonstrate that it is feasible to reflect the CPS covariances given cooperation between the Census Bureau and other statistical institutions.

The key step in the process was for the Census Bureau to release a special public use file (PUF) with replicate weights. Once this had been released, Westat staff were then able to estimate the variance-covariance matrix for CPS control totals. With this covariance matrix, it was then simple to generate pseudo replicated control totals for use in raking the replicate weights for the smaller survey.

In section 2, we present some background on the application. In section 3, we describe how the CPS replicate weights were generated. In section 4, we discuss how pseudo replicated control totals were generated. Results are given in section 5. We close with a brief discussion in section 6.

2. Application

The National Youth Anti-Drug Media Campaign was funded by Congress to reduce and prevent drug use among young people 9 to 18 years of age, by addressing youths directly as well as indirectly, and by encouraging their parents and other adults to take actions known to affect youth drug use (Hornik, et al., 2001). The primary tool for the evaluation is the National Survey of Parents and Youth (NSPY). The NSPY is a household-based survey with a sample of over 25,000 youths and 18,000 of their parents. The households were selected in stages using a stratified multistage probability sampling design. For the NSPY, parents were defined differently than in the CPS. NSPY parents included natural parents, adoptive parents, foster parents who lived in the same household as the sampled youths, as well as stepparents and other relatives serving as parents provided they lived with the child for at least six months. Up to two youths were selected per household. If they were siblings, then one of their parents was sampled. If they were not siblings, then one parent for each youth was selected.

For analysis purposes, separate sets of sampling weights were developed for youths, parents, and youth-parent dyads (e.g. see Hornik, et al., 2001, Appendix A), where a dyad was defined to be a unique youth-parent combination. The weighting of youths followed standard procedures and involved raking to control totals from demographic series which are generally treated as having zero sampling variance. The weighting of the parents was more complex. There are no demographic series for the population of parents. Moreover, since NSPY had a unique definition of parents, it was also impossible to use estimates of parents from the CPS as control totals. Instead, CPS estimates of the number of households in the U.S. with eligible youths were used as control totals in the preparation of NSPY household weights. Parent weights were then based on these household estimates, whereby adjustments were made for within-household subsampling of parents and for parental nonresponse.

The CPS control totals were estimated as the number of households with member youths aged 9 to 18, split by three dimensions. The dimensions were: 1) Potential father

[*] This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

figure age 28 or older present (2 levels), 2) Age mixture of children 9-18 (3 levels), and 3) Combined race and ethnicity of householder (4 levels).

NJC showed that the estimates derived from the monthly CPS data showed substantial month-to-month variation despite the use of composite estimation in CPS weighting. NJC demonstrated that stronger variance reduction can be achieved by taking advantage of the long CPS time series to smooth the monthly CPS estimates. Rather than smoothing just the nine margins of the table defined by the three raking dimensions, they smoothed the 24 interior cells and then re-estimated the margins by summing the smoothed interior cells of the table. These smoothed CPS control totals were used in the NSPY household raking step.

The focus of this paper is how best to estimate the variances on NSPY parent estimates given the use of smoothed sample-based control totals in the preparation of NSPY weights.

The procedure used to generate replicate baseweights for NSPY was described in Rizzo and Judkins (2004). This involved 100 replicate weights, 60 designed to measure between-PSU variance and 40 designed to measure within-PSU variance. The form of the Rizzo-Judkins estimator is

$$\hat{v}\left(\hat{\theta}\right) = \sum_{k=1}^{100} h_r \left(\hat{\theta}_r - \hat{\theta}_0\right)^2 ,$$

where h_r has a different value for the 60 replicate weights devoted to estimating between-PSU variance than for the 40 replicates devoted to estimating within-PSU variance.

To reflect the effects of using smoothed sample-based control totals in the estimation, NJC generated 100 sets of pseudo replicated control totals. This was done by first estimating the variances on the smoothed CPS estimates for the 24 interior cells of the raking table using the residuals from the smoothing model. Let this estimated variance for the k-th interior cell of the raking table be denoted by $\hat{\sigma}_k^2 = \text{var}\left(\tilde{S}_k\right)$, where $\tilde{S}_k$ is the smoothed estimate for the cell.

For each interior cell, 100 pseudo control totals were generated by

$$S_{kr} \sim N\left(\tilde{S}_k, \frac{\hat{\sigma}_k^2}{100 h_r}\right),$$

where $r = 1,\ldots,100$ indexes the replicates, and draws were independent across cells and replicates.

These 100 replicated interior cell series were then summed to obtain 100 pseudo control totals for each of the nine margins of the three-way table. The replicated margins were then used to rake the corresponding NSPY pre-raked household weights (household baseweights adjusted for household-level nonresponse). The same adjustments were then applied to each set of household weights for parent subsampling. Each replicate weight was also adjusted for nonresponse.

It is important to note here that the draws for each interior cell were independent across cells. This method had the advantage of producing nine margins that were consistent with each other, but it is clear that since the interior cells should be correlated with each other, rather than independent, that the variance across the 100 replicates for each margin was incorrect.

Although the NJC approach was reasonable given that no information on the CPS covariances of the interior cells was available at the time, we set out to improve upon this procedure.

3. CPS Replicate Weights for the 1990 Design

The CPS is a multistage sample design that selects primary sampling units (PSUs) composed of one or more adjacent counties at the first stage. PSUs with large populations are sampled with certainty and are referred to as self-representing (SR). Small PSUs, referred to as non-self-representing (NSR), are grouped into strata from which one PSU is selected at random. As a result, there are two components to the total sampling variance, one component measuring the variability between PSUs and the other measuring the variability within PSUs. However, the strategy for CPS total variance estimation is to consider variance estimation for SR and NSR PSUs separately.

3.1 NSR PSUs

Since the CPS sampling design calls for selecting only one NSR PSU per stratum, the sampling strata are collapsed into pairs or triplets. Then these pseudostrata are used to estimate the NSR PSU contribution to the total variance. The collapsed stratum estimator (Wolter, 1985) of total variance is given by:

$$v_{cs}\left(\hat{Y}\right) = \sum_{g=1}^{G} \frac{L_g}{L_g - 1} \sum_{h=1}^{L_g} \left(\hat{Y}_{gh} - P_{gh}\hat{Y}_g\right)^2$$

where

$h = 1,\ldots, L_g$ are the individual strata in a grouped strata,

P_{gh} is the proportion of the population in strata g that belongs to PSU h,

$\hat{Y}_{gh}$ are the estimated strata values of the characteristic of interest,

Y_{gh} are the true strata values of the characteristic of interest, and

$Y_g = \sum_h Y_{gh}$ are the true collapsed strata values.

3.2 SR PSUs

The second stage of the CPS sample design involves selecting segments of housing units (HUs) within each PSU. The segments, which comprises of approximately 4 HUs, are drawn via systematic sampling after the frame has been sorted by characteristics associated with labor force participation.

Wolter (1985) studied several variance estimators for systematic sampling. They were based on squared differences between neighboring samples cases. For the CPS SR PSUs, Fay and Train (1995) extended one of the estimators to handle addition of new sample units over time. Ignoring sample weights and finite population correction factors, the modified variance estimator is given by

$$v_{SD} = (1/2)\left[(y_n - y_1)^2 + \sum_{i=2}^{n} (y_i - y_{i-1})^2 \right]$$

where y_i, *1,..., n,* represents a systematic sample of segments from an ordered population.

3.3 Replicate Variance Estimator

Rather than using the functional forms as described above, CPS variance estimates are obtained from equivalent replicate variance estimators. They result from a combination of two replication methods: Fay's modified balanced half-sample method for NSR PSUs (Judkins, 1990) and successive difference replication for SR PSUs.

The general form of the replicate variance estimator is given by:

$$\hat{v}(\hat{Y}) = \sum_r c_r (\hat{Y}_r - \hat{Y}_0)^2$$

where (ignoring sample weights)

$\hat{Y}_0 = \sum_{i=1}^{n} y_i$ is the parent sample total

$\hat{Y}_r = \sum_{i=1}^{n} f_{ir} y_{ir}$ are replicate totals

f_{ir} are the replicate factors

$c_r = 4/k$ (k = number of replicates for CPS)

3.4 Replicate Factors for NSR PSUs

For each collapsed stratum, the variance estimator

$$v_g = \frac{L_g}{L_g - 1} \sum_{h=1}^{L_g} \left(\hat{Y}_{gh} - P_{gh} \hat{Y}_g \right)^2$$

can be expressed in quadratic form as $\mathbf{y}'\mathbf{C}\mathbf{y}$ where $\mathbf{y}$ is a vector of sample observations $\{y_i\}$, so that $\hat{Y} = \mathbf{1}'\mathbf{y}$ and $\mathbf{C}$ is a symmetric matrix.

Fay (1984) shows that the replicate variance estimator is equivalent to v_g when f_{ir} is given by

$$f_{ir} = 1 + 0.5 \sum_{j=1}^{m} a_{ir} \lambda_{(j)}^{0.5} e_{(j)}$$

where

$\lambda_{(j)}$ and $e_{(j)}$ are eigenvalues and eigenvectors of matrix C respectively,

a_{ir} are the elements of a Hadamard matrix $A=\{a_{ir}\}$ of order k, and

m is the number of eigenvalues (or eigenvectors)

So f_{ir}, which are approximately 1.5 or 0.5 for a pair of strata, provides k replicate factors for each NSR sample unit.

3.5 Replicate Factors for SR PSUs

Fay and Train (1995) shows that the replicate variance estimator is equivalent to v_{SD} when f_{ir} is given by

$$f_{ir} = 1 + (2)^{-3/2} a_{i+1,r} - (2)^{-3/2} a_{i+2,r} \text{ for } i < n, \text{ and}$$
$$f_{nr} = 1 + (2)^{-3/2} a_{n+1,r} - (2)^{-3/2} a_{2,r}$$

However in the successive difference variance estimator, the expression for f_{ir} is used for all i. So f_{ir}, which are approximately 1.7, 1.0 or 0.3, provides k replicate factors for each SR sample unit.

3.6 Weighting the Replicates

The final weights for CPS are obtained by beginning with the reciprocal of the probability of selection for each sample unit. This set of initial weights, also known as the baseweights, is then subjected to the following four successive adjustments: Noninterview adjustment; First-stage ratio adjustment due to sampling of one PSU per NSR strata; Second-stage ratio adjustment, which is an iterative proportional fitting technique, used to control CPS estimates of population to independent population estimates; and, the Composite ratio adjustment which employs estimates from previous months so as to reduce month-to-month variability of CPS estimates.

For each replicate, the replicate weights are obtained by first multiplying the replicate factors by the baseweight. They are then subjected to the same series of adjustments as described above, similar to the parent sample.

Once the CPS replicate weights had been prepared by Census Bureau staff, they were attached to a special PUF and released to Westat.

4. Generation of Pseudo Replicate Control Totals

Given the replicate weights on the special CPS PUF, the design based variance-covariance matrix for the 24 interior cells of the three-way table of interest was then estimated with SUDAAN (RTI, 2004). Of course, this variance-covariance matrix did not reflect the effect of the smoothing of the monthly CPS estimates that was performed. We knew how smoothing affected variances but had no information of the effect of smoothing on the covariances. Lacking better information to the contrary, we assumed that the smoothing preserved correlations. Specifically, let $\hat{C}_i$ be the regular CPS estimate the *i*-th interior cell of the raking table and $\tilde{S}_i$ be the smoothed estimate of the same cell. Then we estimated the covariance matrix for the vector of smoothed interior cells of the raking table as

$$\operatorname{cov}\left(\tilde{S}_i, \tilde{S}_j\right) = \operatorname{Corr}\left(\hat{C}_i, \hat{C}_j\right)\sqrt{\operatorname{var}\left(\tilde{S}_i\right)\operatorname{var}\left(\tilde{S}_j\right)}\,.$$

Let Σ be covariance matrix of smoothed CPS estimates estimated in this way, and let T be the Cholesky of Σ (upper triangular matrix T such that T'T = Σ). Let Z be a 24X100 matrix of iid standard normal variates. Let G be an intermediate matrix with row *k* and column *r* defined by

$$G_{kr} = \frac{Z_{kr}}{\sqrt{h_r} * \sqrt{100}}$$

Let E = T'G . Then the *r*-th pseudo replicate control total for the *k*-th interior cell of the raking table was taken as

$$\tilde{S}_{kr} = \tilde{S}_k + E_{kr}$$

As before, these were summed to obtain the nine margins of the raking table. As before, these were adjusted appropriately for parent subsampling and nonresponse.

5. Results

Variances were then calculated for a variety of NSPY statistics using three different sets of poststratified replicate weights. The first, called the Year 2000 procedure (Hornik, et al., 2001), treated the estimates derived from CPS as fixed totals without variance. The second, called the Year 2004 procedure, was the NJC procedures described in Section 2. The third, called the Year 2005 procedure, was the new procedure discussed in Section 4.

At the time we embarked on this research, we expected that the correlations between interior cells would be mostly negative since domain indicators for disjoint domains are negatively correlated on a simple random sample. This, in turn led us to expect that the Year 2005 procedure to produce variance estimates higher than the Year 2000 procedure but lower than the year 2004 procedure. However, when we examined Σ, we found strong negative and positive correlations.

A high positive correlation of +0.5 was observed between age mixtures within same race and father status. A high negative correlation of -0.6 was observed between race cells within child age mixture and father status. These strong correlations are probably due to residential segregation by race/ethnicity, family-friendly housing and strong clustering in the CPS design. The average absolute correlation of 24 cells with each other is 0.15. The average correlation is -0.006. Since this is still slightly negative, it still seemed reasonable to expect that the 2005 procedure would tend to produce variance estimates between those produced by the 2004 and 2005 procedures, but the expectation for a major change in variance estimates was clearly reduced.

We also noted that the average design effect on the CPS estimate of an interior cell of the raking table is 0.73, indicating some benefits for CPS household estimates from CPS poststratification of persons to demographic control totals. NJC previously noted that smoothing reduced CPS variances by 45 percent.

For the nine margins of the raking table, our expectations were borne out. The average design effect for these nine margins using the 2000 procedure was 0.0000. The

corresponding averaged design effects with the 2004 and 2005 procedures were 0.1531 and 0.0062, respectively.

However, our expectations were not borne out for other NSPY statistics. We calculated the average design effects on 40 statistics about parental substance abuse, parents' exposure to drug information, and parenting practices. We obtained an average design effect of 1.66 using the 2000 procedure, an average design of 1.63 using the 2004 procedure and an average design effect of 1.62 using the 2005 procedure. We also computed the correlations among the design effects. We obtained a correlation of 0.98 between the 2000 and both the 2004 and the 2005 procedures and obtained a correlation of 0.99 between the 2004 and the 2005 procedures.

6. Discussion

Why isn't the new method making more of a difference and why are the 2000 estimates the biggest when they should be the smallest? Possible explanations include: Variance on variance might be obscuring effects (just 100 degrees of freedom). Or perhaps the smoothed CPS variances are so small that treating them as zero is a reasonable approximation. Or perhaps, poststratification of household estimates by race and family structure has little effect on estimates of parental behaviors.

Although the choice of procedure for raking replicate weights clearly didn't matter much in this case, it is also clear that it must matter in other cases. Specifically, if the survey that serves as the source of the control totals has variances that are of a similar magnitude to those of the target survey, and if the variables involved in the poststratification are related to the substantive variables of interest in the target survey, then the 2000 procedure of treating the estimates from the source survey as having zero variance would clearly result in underestimates of target survey variances.

In summary, we have demonstrated that with the cooperation of the Census Bureau, it is possible to improve at least the theoretical properties of variance estimates for surveys that use CPS estimates in their own estimation procedures. Moreover, we have shown that it is not necessary to have the same number of replicate weights as the source survey.

Acknowledgement

The authors would like to thank Mr. Larry Cahoon, Mr. Harland Shoemaker, Jr. and Mr. Jeffrey S. Corteville of The U.S. Census Bureau for their guidance and help in the calculation of the replicate weights and PUF.

References

Ernst, L.R. and Williams, T. (1987). Some aspects of estimating variances by half-sample replication in the CPS. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 480-485.

Fay, R.E. (1984). Some properties of estimators of variance based on replication methods. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 495-500.

Fay, R.E. and Train, G.F. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties (Disc: 172-174). *Proceedings of the Section on Government Statistics of the American Statistical Association*, pp. 154-159.

Hornik, R., Maklan, D., Judkins, D., Cadell, D., Yanovitzky, I., Zador, P., Southwell, B., Mak, K., Das, B., Prado, A., Barmada, C., Jacobsohn, L., Morin, C., Steele, D., Baskin, R., and Zanutto, E. (2001). *Evaluation of the National Youth Anti-Drug Media Campaign: Second Semi-Annual Report of Findings - April 2001.* Rockville, Maryland: Westat.

Judkins, D. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6, 223-239.

Judkins, D.R., Mosher, W.D. and Botman, S. (1991). *National Survey of Family Growth: Design, Estimation, and Inference.* Vital and Health Statistics, September, 1991, Series 2, Data Evaluation and Methods Research; No. 109.

Lemeshow, S. (1979). The use of unique statistical weights for estimating variances with the balanced half-sample technique. *Journal of Statistical Planning and Inference,* 3, 315-323.

Nadimpalli V, Judkins, D.R., Chu A. (2004). Survey calibration to CPS household statistics. *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* pp. 4090-4094.

Rizzo, L., and Judkins, D.R. (2004). Replicate variance estimation for the National Survey of Parents and Youth. *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* pp. 4257-4263

RTI (2004). *Sudaan Language Manual, Release 9.0, First Edition.* Research Triangle Park, NC: RTI.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.

Valliant, R. (2004). The effect of multiple weighting steps on variance estimation. *Journal of Official Statistics*, 20, 1-18.

Westat (2002). *Wesvar 4.0 User's Guide.* Rockville, MD: Westat.

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