PRESERVING DEGREES OF FREEDOM IN A MULTI-MODE, MULTI SITE SURVEY

Ralph DiGaetano, J. Michael Brick, and Ismael Flores-Cervantes, Westat Inc.
Ralph DiGaetano, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Variance Estimation; Degrees of Freedom, Replication

1. Replication Methodology

The two major methods for estimating variances from a complex survey are replication and Taylor series estimation. Wolter (1985) is a useful reference on the theory and applications of these methods. For these methods, procedures have been developed to account for the sample design employed in a complex survey. Factors such as the selection of sample clusters (PSUs, blocks, households) in multistage sampling and the use of differential sampling rates to permit the oversampling of a targeted subpopulation can be appropriately reflected in estimates of sampling error.

Both methods for estimating variances involve the identification of variance estimation strata and PSUs. For replication methods this is followed by the generation of sample weights that replicate the full sample weight. These replicate weights are based on subsamples (replicates) where one or more PSUs within a stratum are randomly excluded and the remaining PSUs within the stratum are reweighted to account for this random exclusion. The variance of an estimate of a parameter is then obtained by comparing estimates of the same parameter generated using each replicate weight to the estimate based on the full sample weight. A function of the sums of the squared deviation of each replicate estimate from the corresponding full sample estimate provides an estimate of sampling error.

One advantage of the replication approach is operational efficiency. Once replicate weights are constructed, it is a very straightforward matter to compute estimates of sampling errors. No special care is needed for subgroups of interest, and no knowledge of the sample design is required. If, in the process of doing data analysis, an estimator not previously considered becomes of interest, replication methodology can be easily used to develop an appropriate estimate of variance.

A second advantage is that all components of the design and estimation strategy can be reflected in the estimates of variability. Variances are affected by both the nonresponse and poststratification adjustments. Replicate weights can be developed that reflect all such aspects of weighting. Currently existing software for using the Taylor Series method for variance estimation can only reflect the very last step in poststratification. Nonresponse adjustments and any other calibration method (including the matching of multiple demographic distributions via raking) will not be reflected in the variance estimates.

2. Number of Replicate Weights and Degrees of Freedom

A disadvantage of the replication method is that costs in computer resources may be incurred if a substantial number of replicates are used. Thus, a problem that is sometimes encountered in the development of replicates is how to limit the number of replicates to a reasonable number (say, roughly 100) while still preserving a sufficient number of degrees of freedom to help ensure the stability of variance estimates for parameter estimates of interest.

An approach developed for the National Survey of American Families (NSAF), where the jackknife replication methodology was employed, is described and evaluated in this paper. The jackknife method as employed for the NSAF is based on having two PSUs sampled per stratum and creating one replicate weight for each of those variable estimation strata. The approach focuses on strategies for combining variance strata to reduce the number of replicates while maintaining an adequate number of degrees of freedom for variance estimation purposes. The findings from this evaluation, in addition to other analyses, led to the development of a new approach for construction of the replicate weights to enhance the numbers of degrees of freedom even further. The new approach is mentioned in the last section of the paper.

3. Sample Design of the National Survey of American Families

The target population for the NSAF is the population under 65 years of age (in several specific areas and nationwide) with particular focus on households with children and the low income population. To help maximize the coverage of the target population while limiting costs, the sample design used two modes of data collection. A Random Digit Dial (RDD) Telephone Survey covered those households with telephones while an area probability sample covered those households without phones. It was important to cover nontelephone households since they contain a disproportionately high number of poor persons in the country.

For the NSAF one factor contributing to the design effect for national estimates is the creation of 15 separate geographical sampling strata or sites: 13 states, one county within a targeted state, and the remainder of the country. The oversampling required to produce separate estimates for the study areas increased the design effect for corresponding national estimates.

Households without children and those above 200 percent of poverty were subsampled for the RDD component of the survey but not for the area sample. Within sampled households children were sampled and the adult most knowledgeable about the sampled child was selected with certainty. A sample of adults not responsible for the children in the household was also selected. More details on the sample design for the NSAF are summarized by Waksberg et al (1997).

4. Applying Replication Principles in the NSAF

The choice of the number of replicates is based on the desire to obtain an adequate number of degrees of freedom to ensure stable estimates of variance while not having so many as to make the cost of computing variance estimates unnecessarily high. See Rust (1986) for a discussion of the relationship of the degrees of freedom and the stability of the sample estimate. Generally, one targets at least 30 degrees of freedom where possible in order to obtain relatively stable variance estimates. The number of replicates is an upper bound on the number of degrees of freedom associated with estimates of variance. Thus, the number of replicates should be at least 30 but is generally targeted higher because other factors can reduce the contribution of a replicate to the total number of degrees of freedom.

For a complex survey, factors affecting the number of degrees of freedom associated with a particular variance estimate include: the variability and distribution of sampled units within and across the variance estimation strata; the sampling rates and sample sizes within strata; and the number of replicates ultimately generated. Often, the variability within strata can be regarded as a constant and the units within strata have a kurtosis approximately equal to 3 (3 is the value for a normal distribution). In such cases, 40 to 50 replicates will provide a sufficiently stable variance estimator for most analytic purposes, permitting the assumption of an underlying normal distribution for confidence intervals and hypothesis tests instead of a t-distribution with a specified (and estimated) number of degrees of freedom. To the extent that there may be some variation between strata for some subpopulations of interest, more replicates may be desirable. For the approach described here 82 replicate weights were constructed for national estimation purposes with a view to obtaining at least 30 degrees of freedom for variance estimates for the various target populations.

The number of degrees of freedom varies by the target populations of interest. Populations of special interest for analytic purposes for the NSAF include the general population and the poor, both nationally and for each site separately. Children too were of interest. For estimates for the general population, most of the degrees of freedom would be expected to come from the RDD component of the survey, since roughly 95 percent of all households have telephones. Thus, 60 variance estimation strata were established for the RDD component. For national estimates for the poor, many of whom do not have telephones, a sufficient number of degrees of freedom should be available for analytic purposes from the 82 replicates constructed. These 82 replicates are based on combinations of many more variance estimation strata from both the RDD and area components, discussed below.

4.1 Initial Variance Strata

"Combining" strata in the context of jackknife replication can perhaps best be explained by example. Suppose we have two variance strata A and B, each with two variance estimation PSUs. We will call the PSUs within stratum A A1 and A2 and similarly for stratum B. If we combine the strata A and B for variance estimation purposes, we treat them as a single stratum "AB" with two PSUs: A1 and B1 are treated as a single PSU AB1, and A2 and B2 as a single PSU AB2. For the sample replicate weight associated with stratum AB, the sample weights for persons within A1 and B1 are either randomly selected to be doubled (in which case those of persons within A2 and B2 are set to zero) or set to zero (in which case those of A2 and B2 are doubled). "Combining" strata in jackknife replication is analogous to establishing partially balanced replicates with Balanced Repeated Replication (BRR).

Combining strata does not introduce bias into the variance estimates but does reduce the number of replicates, thus reducing the number of degrees of freedom and the stability of the variance estimates. Our strategy for combining strata attempts to reduce substantially the number of replicates while still preserving enough degrees of freedom to achieve the analytic goals described earlier.

Determining the initial number of variance estimation strata, and thus potential number of replicates from which to work, was fairly straightforward. The noncertainty (nonself-representing or NSR), area strata were readily identified based on the sample design for the selection of PSUs. There were generally two PSUs per stratum (after collapsing since generally one PSU per stratum was selected). The exception was Texas, since, with 7 PSUs, three variance estimation strata were created where one stratum has 3 PSUs.

For each certainty stratum (self-representing or SR PSU), initial variance strata were formed by pairing area sample segments (blocks or groups of blocks) that were sampled consecutively in the systematic sample of segments selected from the PSU. Within a self-representing PSU, segment selection represented the first stage of sample selection. Thus, the potential number of variance estimation strata that could be created was generally half the number of segments contained within the PSU. However, each resulting variance stratum makes a relatively small contribution to the total number of degrees of freedom compared to the noncertainty strata from the same site. Thus, more combining was undertaken within the certainty strata.

For the nation as a whole 208 "preliminary" variance strata were created from the area sample: 42 associated with noncertainty sample strata and 166 for the self-representing PSUs. Note that the 166 preliminary strata already represented some combinations of variance strata.

For the RDD component of the study a large number of variance strata were possible since each pair of adjacent, sampled phone numbers could be regarded as a stratum. However, such a large number of strata are unnecessary to achieve stable variance estimates. A total of 60 variance estimation strata were created for telephone numbers within each of the 15 sites. This was accomplished as follows.

First, the sampled telephone numbers were arranged in the same sort order used in sample selection. Then, adjacent sampled telephone numbers were paired to establish "initial" variance estimation strata (the first two sampled phone numbers represented the first "initial" stratum, the third and fourth sampled phone numbers, the second "initial" stratum, etc.) Finally, each pair was sequentially assigned to one of 60 "final" RDD variance estimation strata (the first pair to variance estimation stratum 1, the second to stratum 2,..., the sixtieth pair to stratum 60, the sixty first pair to stratum 1, etc.).

Note that, using this approach for combining strata, each site has 60 replicate weights for the RDD component on which to base variance estimates. For estimates for the nation as a whole, we combined the 60 strata from all 15 sites, providing 60 replicate weights for national estimates as well. Consequently, estimates for the general population within the various geographical entities of interest should have relatively stable variance estimates.

4.2 Combining the Area and Telephone Sample Replicates

In all, there were 268 variance strata available for variance estimation at the national level: 208 from the various site area samples and 60 from the RDD sample (after combining all RDD strata from all sites). This is far more than is necessary to achieve an adequate level of precision for survey estimates. The number of area sample replicates for each site was determined to help maximize the stability of variance estimates for site level estimation. However, the resulting number of area sample replicates obtained for national or regional estimates does not proportionately increase the precision for estimated sampling errors at the national or site levels. Thus, the combining of some strata was undertaken to achieve operational efficiency while still permitting the establishment of stable variance estimates at the national, regional, and site level.

The strategy for combining variance estimation strata was based on analytic concerns. Estimates of analytic interest include those at the national, region, and study area level, for the general population, the poor, and children.

Important factors considered in combining strata are

- Most of the variance for national estimates for the general population is from with the RDD sample since it represents roughly 95 percent of the population--the 60 replicates available from the RDD sample should provide a substantial number of degrees of freedom;
- The certainty strata from the area sample (nontelephone household sample) represent roughly 20 percent of the total measure of size associated with the segments eligible for sample selection--thus, most of the variability associated with nontelephone households comes from the noncertainty strata;
- The percentage of the total measure of size associated with certainty strata for the study areas ranged from 13 percent (Mississippi) to 70 percent (Massachusetts and New Jersey) with a median of about 42 percent--thus, the contribution of certainty strata to the variance of study area level estimates for nontelephone households varies considerably;
- For cost reasons, the number of noncertainty sample PSUs was small. Thus, strata associated with noncertainty area samples potentially could have a dramatic effect on the number of degrees of freedom obtained, particularly for an individual

site, if the proportion of the population found in non-telephone households is relatively high for a given site; and

- Important contributions to the degrees of freedom for estimates for the poor will be obtained from both the RDD and the area samples--roughly, 30 percent of the population under 200 percent of the poverty threshold lives in households without telephones. Thus, by maintaining a relatively large number of degrees of freedom from the noncertainty area sample replicates, in tandem with the replicates available from a combination of RDD and area sample certainty replicates, an adequate number of degrees of freedom should help preserve the stability of variance estimates at all levels of estimation.

Taking these factors into account, the general strategy for combining strata was as follows. Most of the degrees of freedom from the area sample can be expected to come from the non-self-representing PSUs. Thus, for each of the study areas separately, combine only variance strata from the certainty PSUs with the RDD variance strata. This should help preserve most of the degrees of freedom (albeit relatively few) available from the area sample for national estimates. In so doing, most of the degrees of freedom for study area level estimates for the general population should be preserved as well.

Before combining the 166 variance strata from the area certainty PSUs with the 60 variance strata constructed for the RDD survey, the 166 variance strata were first reduced to 60 by combining. This was done so that no variance strata from the same site or region were combined and only rarely (3 of 63 variance strata) involved variance strata from the five sites accounting for 75 percent of the total measure of size associated with the certainty strata. By avoiding or limiting such overlap, the stability of national, regional, and site estimates for nontelephone households is only slightly less than that which would have been available without such combining.

Once the 60 new variance strata for the area sample certainty strata were formed, the resulting strata for each study area were combined with the 60 RDD variance strata for that study area. By then combining across all 15 sets of 60 strata, 60 strata were established for national estimation. For national estimates for the general population from the area and RDD samples combined, this has little impact on the stability of the variance estimates since the number of degrees of freedom associated with the area certainty strata is small. The combining of variance strata reduces the number of degrees of freedom associated with estimates for the poor. However, the estimates for the poor include the non-certainty areas of the nontelephone survey from which additional degrees of freedom are obtained. Thus, the overall number of degrees of freedom for estimates for the poor should be sufficient to ensure relatively stable variance estimates.

The decision of how many strata to construct was based on the expected number of degrees of freedom that were already available from the other strata for the various populations of interest. For the poor population the degrees of freedom for variance estimates are somewhat reduced over what is potentially available due to the combination of the RDD and certainty strata into 60 strata. The addition of roughly 20 additional strata from the noncertainty strata helps ensure stable variance estimates in this case. Thus, the 42 strata available from the various site samples were combined into 22 final strata for variance estimation purposes. Again, no strata from the same site are combined. The only combinations within census regions were Alabama and Mississippi with the Balance of U.S. strata.

Table 1 provides a visual illustration of how the strata from the area sample and RDD sample were combined for three states in two different regions. Note that overlap of strata was permitted for California and Alabama but not Colorado and California, since the latter two are in the same region. For those 3 states, 68 strata (and thus replicates) would be obtained overall. For each state separately we obtained 64 (Alabama), 62 (California, and 62 (Colorado). Nationally, we obtained 82 variance strata. For each site, 60 plus the ultimate number of area non-certainty strata established for that site represents the total number of strata obtained.

5. Evaluating the Degrees of Freedom

To evaluate this approach, we looked at estimates for the state of California. The estimates were for all children and poor children in the state.

We used the following equation to estimate the number of degrees of freedom (*DF*) associated with the variance of an estimate of a parameter θ:

$$DF = \frac{2[v(\theta)]^2}{V(v(\theta))};$$

θ = The estimate of interest;
$v(\theta)$ = The variance of θ ; and
$V(v(\theta))$ = The variance of $v(\theta)$.

The value for V(v(θ)) was estimated as follows (Rust, 1986):

$$V(v(\theta)) = \sum_{g=1}^{G}\left\{\left[\sum_{h\in g}^{H_g}\frac{W_h{}^4\sigma_h^4(\beta_h-3)}{n_h^3}\right]+2\left(\sum_{h\in g}^{H_g}\frac{W_h{}^2\sigma_h^2}{n_h}\right)^2\right\};$$

W_h: the proportion of persons in the population;

σ_h^2: the variance of the PSU estimate;

n_h: the number of sampled PSUs;

β_h: the Kurtosis of the PSU estimates;

h: indexes original variance strata;

g: indexes final "combined" variance strata;

H_g: the number of original strata in combined stratum g; and

G: the final number of combined strata.

We assumed β_h was equal to 3 (see Kish, 1965 for a justification of this).

The estimate of $V(v(\theta))$ appeared to be most sensitive to the value of W_h assigned. The proportion of persons in nontelephone households in a PSU or segment (the parameter W_h for the area sample) was estimated in two different ways: from survey data and from the Public Use Microdata System (PUMS) representing a 5 percent sample of households from the 1990 Census. The PUMS data were far more stable and considered much more accurate than those available from the survey, where data from roughly 10 nontelephone households per non-certainty PSU were collected. The estimated proportion of W_h from the survey for non-telephone households in non-certainty strata in California was about .044, from the PUMS about .023 or approximately half as much. (Corresponding figures for the certainty areas of California was about .045 from the survey and .022 from PUMS.)

Table 2 summarizes the evaluation. It gives estimates of DF if there was: no overlap at all; the overlap obtained using the above methodology described; and full overlap of the RDD and area strata. If the proposed approach was effective, a reduction in the number of replicates would not result in a corresponding proportionate reduction in degrees of freedom.

Using the PUMS estimates of W_h, a reduction in the number of replicates of about 17.5% (from 75 to 62) resulted in a corresponding reduction in degrees of freedom of about 14% (from 60.22 to 51.79) for California children as a whole and of about 11.75% (from 35.09 to 30.97) for poor children. On the other hand if we had chosen to eliminate only two more replicates by combining all area strata with the RDD strata (going from 62 down to 60), we would have lost a disproportionately large number of degrees of freedom. Again using the PUMS estimates of W_h, a reduction in the number of replicates of about 3.25% (from 62 to 60) resulted in a corresponding reduction in degrees of freedom of about 12% (from 51.79 to 45.64) for California children as a whole and of about 20% (from 30.97 to 24.73) for poor children. For poor children the estimated number of degrees of freedom would have fallen below the targeted 30. Similar results were obtained using the W_h estimated using survey data, but the estimated number of degrees of freedom are substantially lower using survey based estimates of W_h. If telephone households truly represented 91 percent of California households the number of degrees obtained for the survey would be low. However, PUMS and other data suggest that the percentage is likely to be at least 95.5 percent in 1990 and even higher when the NSAF was carried out in 1997. Thus, the actual number of degrees of freedom are likely to be at least as large as the number associated with the PUMS estimates.

To summarize the finding of the evaluation considering the state of California, the proposed approach of combining area sample strata with the RDD strata while retaining most of the degrees of freedom associated with the non-certainty strata seems to have accomplished its purpose. The number of replicate weights was kept relatively low (under 100) while the number of degrees of freedom retained provide relatively stable estimates of variance for analytic purposes.

6. An Alternate Approach

After considering the results of this and other evaluations of the variance estimates from the NSAF, several modifications were made in the estimation strategy. One of the most important results of this evaluation was the quantification of the small number of degrees of freedom that were available for the low income population of children. As the evaluation showed, this was associated with the contribution of the non-certainty strata to the variance estimates. One non-certainty PSU was selected from each stratum, and the number of strata ranged from 2 to 10 in the study areas. The pairing of PSUs to permit variance estimation resulted in a low number of degrees of freedom for these strata. Notice that this is true irrespective of the method of variance estimation. Despite the fact that the maximum number of degrees of freedom were retained for these strata within study area (since there was no combining of non-certainty strata within study area), the number of degrees of freedom is small.

An alternative strategy was developed to obtain additional degrees of freedom by treating the non-certainty PSUs as certainty PSUs for variance estimation purposes. In this approach variance estimation strata consist of pairs of segments (second stage sampling units) instead of PSUs. In so doing, the between PSU component of variance is ignored for these strata, thus incurring some potential for bias by understating the variance. We examined the gains obtained by a reduction in the variance of the variance estimate achieved by the increased degrees of freedom, and it outweighed the loss due to bias. The result was a net reduction in the mean square error of the variance estimate. In fact, any bias would be very small because most of the variance was due to segment-level variability rather than between-PSU variability. The approach of pairing the non-certainty PSUs into variance estimation strata described in this paper actually overstates the variance to some extent since a between-stratum component is introduced into the variance estimate that is not part of the sample design. Thus, we believe the alternative strategy is superior for the NSAF.

7. References

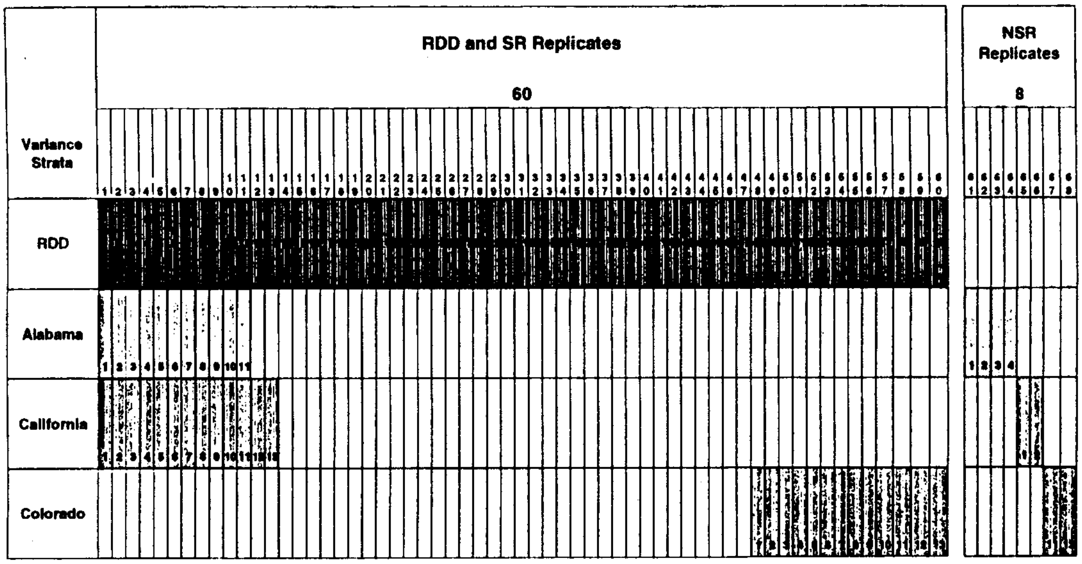
Kish, Leslie (1965). *Survey Sampling*, John Wiley & Sons, Inc., p.289-291.

Rust, Keith (1986). Efficient replicated variance estimation. Proceedings of Survey Research Methods Section of the American Statistical Association, 81-87.

Waksberg, J.; Brick, J.M.; Shapiro, G.; Flores-Cervantes, I.; Bell, B. (1997). Dual-Frame RDD and area sample for household survey with particular focus on low-income population. Proceedings of Survey Research Methods Section of the American Statistical Association, 713-718.

Wolter, Kirk M., (1985). Introduction to Variance Estimation, Springer-Verlag, New York, Inc.

Table 1. Illustrative replicate distribution after combining across variance strata (reflecting the three states only)



Resulting Replicate Count:
60+4+2+2=68

Table 2. Results for estimates of children in California

	Number of replicates	Survey		PUMS	
		CA Total DF	CA Poor DF	CA Total DF	CA Poor DR
Combined RDD and Area Samples					
3 No Overlap	75	19.67	10.81	60.22	35.09
3 Certainty Overlap	62	18.45	10.45	51.79	30.97
3 With Full Overlap	60	16.92	9.96	45.64	24.73