COMPARISON OF PROCEDURES TO ACCOUNT FOR CERTAINTY PRIMARY SAMPLING UNITS

Frank Potter, Donsig Jang, Esther Friedman, Nuria Diaz-Tena, Bidisha Ghosh
Mathematica Policy Research, Inc. Princeton, New Jersey 08543-2393

KEY WORDS: Variance Estimation, Certainty PSUs, Balanced Repeated Replication

The purpose of this paper is to describe and assess some methods for accounting for certainty primary sampling units (PSUs) when using a pseudo-replication procedure (specifically balanced repeated replication (BRR) procedure) for variance estimation[1]. We compare these alternatives to the variance estimation procedure using the explicit variance estimation equations with Taylor series linearization of nonlinear estimators. The context for this study is a complex national survey of children in which some PSUs are selected with certainty and others are not, and children are selected within PSU using stratified random sampling. The paper summarizes our work to investigate possible methods for accounting for the certainty PSUs when using the BRR procedure.

Each certainty PSU can be regarded as a separate sampling stratum for variance estimation of survey statistics. Consequently, the classical multi-stage stratified design based estimation method or very straightforward for linear statistics and when using the Taylor series linearization method for nonlinear statistics, respectively (see for details Wolter 1985 and more recently LaVange et al. 1996). Popular statistical software, such as SAS and STATA, contain procedures for survey data analysis based on this procedure. SUDAAN (Research Triangle Institute 2001) offers the user the option to use this procedure as well as options for two pseudo-replication procedures (BRR and the jackknife). The spectrum of pseudo-replication methods includes methods that predate high-speed computers (such as BRR and jackknife) to methods that existence is largely dependent on the availability of high-speed computers (such as bootstrapping). An exposition on these variance estimation methods was given by Wolter (1985) and more recent literature on pseudo-replication variance estimation methods includes papers by Rust and Rao (1996) and Rao and Shao (1996).

[1]Pseudo-replications are random subsamples of a specific survey sample, as opposed to true replications of the sampling design, which entails the selection of multiple independent samples using the same sampling design.

1. VARIANCE ESTIMATION

The sampling variance of an estimate derived from survey data for a statistic (such as a total, a mean or proportion, or a regression coefficient) is a measure of the random variation among estimates of the same statistic computed over repeated implementation of the same sample design with the same sample size on the same population. Variance estimation depends on the population characteristics, the form of the statistic, and the nature of the sampling design. The two general forms of statistics are linear combinations of the survey data (e.g., a total) and nonlinear combinations of the survey data. Nonlinear combinations include the ratio of two estimates (e.g., a mean or a proportion in which both the numerator and the denominator are estimated) and more complex combinations such as regression coefficients. For linear estimates with simple sample designs (such as a stratified or unstratified simple random sample) or complex designs (such as stratified multi-stage designs), explicit formulae for the design-based variance estimator are available to estimate the sampling variance. For the nonlinear estimates with simple or complex sample designs, explicit equations are not generally available for variance estimation and various approximations or computational algorithms are used to provide an essentially unbiased estimate of the sampling variance that to a greater or lesser degree accounts for the survey design.

Variance estimators for complex sample designs take on two primary forms: (a) the procedures based on the Taylor series linearization of the nonlinear estimator using explicit sampling variance equations of linear statistics and (b) the procedures based on forming pseudo-replications of the sample. Within the class of pseudo-replication procedures, the BRR procedure, the jackknife procedure, and the bootstrap procedure are most widely used and discussed (Shao and Tu 1995). The discussion here will be limited to the Taylor series linearization procedure and the BRR procedures because these two procedures were actually considered for variance estimation of the survey example presented below.

1.1 Taylor Series Linearization Procedure

The Taylor series linearization procedure is based on classical statistical methods in which a nonlinear statistic can be approximated by a linear combination of the components within the statistic. The accuracy of the approximation is dependent on the sample size and the complexity of the statistic. For most

commonly used nonlinear statistics (such as ratios, means, proportions, and regression coefficients), the linearized form has been developed and has good statistical properties under the large sample approximation. Once a linearized form of an estimate is developed, the explicit equations for linear estimates can be used to estimate the sampling variance. Because the explicit equations can be used, the sampling variance can be estimated using many of the features of the sampling design (e.g., stratification, multiple stages of selection, unequal selection rates within strata, and finite population corrections for equal and unequal probability samples). This is the primary variance estimation procedure used in SUDAAN, SAS, Stata, and other software packages to accommodate many simple and complex sampling designs. SAS and Stata impose some limitations on the designs supported, whereas SUDAAN accommodates most complex designs including stratified multi-stage design using with-replacement or without-replacement probability proportional to size sampling. To be able to calculate the variance using survey analysis software, sample design information (such as stratum, analysis weight, sampling stage, etc.) is needed for each sample unit.

When certainty PSUs are in a sample, the certainty PSUs do not contribute to the sampling variance at the PSU level (that is there is no between-PSU contribution) and the variance between units within each certainty PSU can be the variance component contributed by the certainty PSU. This situation can be handled by treating the PSUs as strata. This results in one or more pseudo-strata defined by the certainty units.

1.2 Balanced Repeated Replication Procedure

The balanced repeated replication (BRR) procedure is designed for use with stratified multi-stage sample designs in which two primary sampling units are selected with replacement in each stratum. The BRR procedure was developed at the Census Bureau (and formalized by McCarthy (1966)) for the estimation of sampling variances before the availability of sophisticated high-speed computers for large national surveys. Two PSUs, or sometimes only one PSU, are selected from each stratum and a half sample is then constructed at each replicate (pseudo replicate), and the sampling variance is estimated by computing the variation among the survey estimates calculated for each half-sample. To minimize the number of replicates without the loss of statistical efficiency, the process for forming the half-samples is constrained to ensure a "balance" among the half-samples (Wolter 1985). For some estimates (especially ratio) for small subpopulations, the BRR procedure can be unreliable due to undefined denominator values obtained from some of replicates. To avoid this and guarantee to have non-missing value for all replicates, a modified BRR procedure (called "Fay's method") is proposed and commonly used (Judkins 1990).

An advantage of replication is its ease of use at the analysis stage. The BRR approach does not require the development of a linearized form of the estimator, so sampling variances can be computed for some forms of complex nonlinear estimates or non-smooth estimators that either cannot be or have not been incorporated in software using the Taylor series linearization procedure. Instead, the same estimation procedure is used on all replicates and the full sample, and the actual variance computation is readily computed. The procedure can be applied to most statistics as well as for subgroups. Software for replication methods requires either replicate weights or sample design information, including the sample weight and stratification information. WesVarPC (Brick et al. 1996) is a popular software program that has the ability to compute sampling errors using replication methods. SUDAAN also has the capacity to compute sampling variances using the BRR method.

Another advantage is that the procedure accounts for sampling variance associated with the adjustments used in weighting the data. By developing weighting adjustments for each replicate, the full effect of the adjustments, such as for nonresponse and poststratification, can be accounted for in the calculation of sampling variances. On the contrary, it requires an initial expenditure of effort in forming the replicates, computing a separate set of weights for each replicate, and applying all the nonresponse and poststratification adjustments independently to each replicate.

However, for sampling designs using simple stratified random samples, without-replacement sample selection with high sampling rates, or certainty selection of primary sampling units, the BRR procedure is not directly appropriate and adaptations are required to produce unbiased (or nearly unbiased) sampling variance estimates (Wolter 1985, Rao and Shao 1996 and Rao and Shao 1999). This paper is to consider a couple of design approximation options suited for BRR application using the data from the National Survey of SSI Children and Families (NSCF) in which a handful of certainty PSUs were selected.

2. THE SURVEY

The National Survey of SSI Children and Families (NSCF) collected data on children with disabilities and their families who received or applied for Supplemental Security Income (SSI). The survey, sponsored by the Office of Research, Evaluation, and Statistics of the Social Security Administration (SSA), had two major objectives:

1. To provide information on the characteristics, experiences, and needs of the current cross-section of SSI child recipients and their families

2. To evaluate the effects of the Personal Responsibility and Work Opportunity Act of 1996 (P.L. 104-193; otherwise known as the Welfare Reform Act) on SSI children

The NSCF, which was administered in 2001-2002, was the first national survey of SSI children in more than 20 years. The survey was intended to fill a critical data need by providing current information on the health and well-being of SSI children and their families.

Mathematica Policy Research, Inc. (MPR) developed the sampling design for the NSCF and questionnaire under a separate contract with SSA (Potter and Mitchell 2000). A complex multivariate allocation algorithm was used to minimize survey costs and sample size subject to precision constraints for more than 150 survey estimates. The NSCF used a two-stage sample design. SSA administrative records were used as a list frame from which 75 primary sampling units (PSUs) were formed; after PSUs were selected, the list was used as the sampling frame for the selection of individual children within PSUs. The final sample size was 11,971 cases.

The NSCF sample design involves stratification and unequal probabilities of selection. Variance estimates calculated from NSCF data must incorporate the sample design features in order to obtain the correct estimate. For the NSCF, 75 PSU selections were made using a probability minimal replacement sequential selection procedure (Chromy 1979) and a composite size measure (Folsom et el 1987) accounting for 8 strata of children in each PSU. One PSU had a sufficiently large composite measure value to receive two selections and 5 other PSUs had composite measure values sufficiently large to classify these as certainty PSUs. That is for NSCF, the 75 selections resulted in 6 certainty PSUs (one with a double selection and 5 with single selections) and 68 noncertainty PSUs selected. Within each PSU, children were stratified into 8 sampling strata and we selected a simple random sample in each stratum. Details of the sampling design and weighting procedures for the NSCF are in Potter and Diaz-Tena (2003).

3. VARIANCE ESTIMATION IN THE NSCF

3.1 Design-Based Taylor Series Linearization Method

For noncertainty PSUs, the 68 PSUs were paired to form 34 pseudo-strata[2]. For the 6 certainty PSUs, we formed 48 strata of children using the 8 within-PSU sampling strata. Although the sample was selected using a probability minimal replacement sequential sampling algorithm, we assumed a with-replacement design to simplify the variance estimation procedure.

3.2 Balanced Repeated Pseudo-Replication or BRR Procedure

For the PSU selection, no explicit stratification was used (although we did impose implicit stratification) so a single sample was selected. For the number of degrees of freedom for a complex survey, the rule of thumb is the number of PSUs less the number of strata. By this logic and ignoring the certainty selections, the number of degrees of freedom is 74. We choose to develop 72 pseudo-replicates for the BRR method to allow consistency between the two variance estimation approaches.

The six certainty PSUs posed a problem that has received some discussion in the survey sampling literature (Wolter 1985, Rao and Shao 1996 Rao and Shao 1999). We considered two alternatives (1) pairing certainty PSUs to form pseudo-strata following a collapsed strata approach described below and (2) forming pseudo-PSUs by randomly assigning sampled cases to 2 pseudo-PSUs in each certainty PSU.

a. BRR Option 1: 38 Pseudo Strata:

1 pseudo-stratum for the double-hit PSU +
3 pseudo-strata for the single-hit PSUs + 34 pseudo-strata for the 68 noncertainty PSUs

For the noncertainty PSUs, 34 strata were constructed with two non-certainty PSUs within each stratum. The certainty PSU with double hit is then regarded as a stratum in which units were evenly divided into two groups. For the remaining 5 certainty PSUs, the largest PSU is treated as the double hit PSU in that sampled children were evenly divided into two groups and the other four are pair-

[2]Because the sample was selected using a probability minimal replacement sequential selection procedure (Chromy 1979), pseudo-strata for noncertainty PSUs were formed by pairing adjacent sampling zones. Using this process, we take into account the implicit stratification of the PSUs.

wise matched. The original sample design is approximated by a two PSU per stratum design with 38 strata.

b. BRR Option 2: 40 Pseudo Strata:

1 pseudo-stratum for the double-hit PSU + 5 pseudo-strata for the single-hit certainty PSUs + 34 pseudo-strata for the 68 noncertainty PSUs

For the noncertainty PSUs, 34 strata were constructed with two non-certainty PSUs within each stratum. Each of 6 certainty PSUs is regarded as separate strata and sampled children were randomly partitioned into two groups. The original sample design is approximated by a design with 40 strata and two PSUs per stratum.

Other approaches found in the literature were deemed infeasible for the current survey. One extension of the method of splitting the sample in certainty PSUs uses repeatedly grouped balanced half samples (Rao and Shao 1996). In this method, the units in the certainty PSUs would be independently repeatedly grouped into two pseudo-PSUs for, say, T times. For example, if 20 sample members were in a certainty PSU, then two pseudo-PSUs would be developed each with 10 observations and this would be repeated 15 times. Replicate weights would be prepared and used to make 15 estimates of the sampling variance. These 15 estimates would be averaged to develop the final sampling variance estimate. This procedure exhibited good properties in the relative bias and coverage rates for 95 percent confidence intervals in a simulation study for a value as low as T=15. However, for a national survey of this scope, the use of this procedure is not practical.

Another ad hoc method would divide the sample within each certainty PSU into, say, m pairs of observation assuming the sample size is evenly divisible by 2 (Wolter 1985). That is if 20 sample members were in a certainty PSU, ten pseudo-PSUs would be developed each with 2 observations. The statistical properties of this procedure were investigated in a simulation study and this procedure also exhibited good results in the relative bias and coverage rates for 95 percent confidence intervals (Rao and Shao 1996). Once again, this procedure is not practical in the current study because of the number of pseudo-PSUs and the number of pseudo-strata that would be generated.

In some survey designs with certainty PSUs, the post hoc development of pseudo-PSUs can be avoided by forming secondary sampling units within these PSUs at the initial sample selection phase. Using these SSUs (defined by zip code or other factors), a sample of SSUs can be selected. If an even number of SSUs were selected in each certainty PSU, the sample design would conform more closely to a BRR model. There would pseudo-strata of noncertainty PSUs and pseudo-strata of SSUs within the certainty PSUs.

4. ANALYSIS

For the analysis, we considered frame variables that would be related to analysis variables of interest to SSA such as gender, race, diagnostic categorizations and a payment related variable. We also used the full sample (11,971 children) so that replicate-specific nonresponse adjustments were not required for each set of the BRR pseudo-replicates. All weights were post-stratified to the same frame totals instead. In addition, all identification information was removed from the file before any analysis to avoid any confidentiality concerns. SUDAAN was used to compute all sampling variances (based on both Taylor series linearization method and BRR) to avoid differences caused by different software.

5. METHODS FOR COMPARISONS

To avoid confounding the assessment of the BRR procedures by the Taylor series linearization process, we computed the sampling variance estimates for totals: for gender, race (4 categories), diagnostic categories and a payment variable on the frame file. That is, these estimates were simple linear estimates. In addition, for the payment variable, the mean and median was also computed to get assessment of differences caused by nonlinear estimators. These estimates were computed for the full population and for the population in each of 8 sampling strata. The sampling strata represented between 3 and 42 percent of the national total (see Potter and Mitchell 2000 for a detailed explanation of the sample design and stratification).

We computed a relative difference score for each standard error computed with the reference standard error defined by the explicit variance estimation equation with Taylor series linearization for the mean and median statistics, that is

Relative Difference (Percent) = 100 * [SE(BRRi) – SE(Design-based)] / SE(Design-based)

where SE(BRRi) is the standard error based on the BRR option (i = 38 strata or 40 strata) and SE(Design-based) is the standard error using the design-based equations (for nonlinear estimates using the Taylor series linearization).

6. RESULTS

In Table 1, we show in the first set of rows of, the mean, median, and standard deviation of the relative differences in estimates of the totals overall and by gender. The next set of rows, we show similar information for total population estimates for the four race categories (Hispanic, Black, Unknown or Not

Given, and White) and in the third set of rows for the two disability diagnostic categories. Because these estimates are totals, they are simple linear estimators and, therefore, this is a comparison of the BRR options and the explicit variance estimation equations. No clear pattern is shown for these relative differences. For the last set of rows (payment in dollars), we computed the total and two nonlinear statistics (the mean and median payments). For the total of payments, the mean, median and standard deviation of the relative differences between standard errors for the BRR options and explicit equations are consistent with the other linear statistics, and this is the same for the mean payment, which includes the Taylor series linearization for the mean. For the median payment statistic, the mean and median of the relative differences show more skewness and standard deviation of the relative differences is substantially larger than that for the relative differences of any other statistic. This may reflect the limitations of the Taylor series linearization procedure in working with percentile statistics like the median. Nevertheless, the pattern was the same for both BRR options.

Overall, there was no clear pattern between the estimated standard errors computed using the pairing 4 certainty units into 2 collapsed pseudo-strata or dividing the sample in these certainty PSUs into 2 pseudo-PSUs.

We subsequently looked at the percentage of the sampling variance associated with the certainty PSUs and determined that less than 5.0 percent of the sampling variance for the variables used in the analysis was attributable to the certainty PSUs. Because the contribution of certainty PSUs to the total sample variance is so minimal, the choice of design option for BRR method may not matter much in this application. We choose to use the 40 pseudo-strata BRR design because it more closely followed the literature. However, we were concerned that randomly dividing the 5 single-hit certainty PSUs into 2 pseudo-PSUs would result in small sample sizes for some domains in these pseudo-PSUs.

7. DISCUSSION

In complex sample surveys, PSUs are generally selected with probability proportional to size and without replacement. Therefore, certainty PSUs are often encountered. For the Taylor series linearization with the explicit estimation equations, we can rely on classical sampling theory for variance estimates. For the BRR procedure, the underlying assumption is that two PSUs are selected with replacement in each stratum. We conducted a literature search to identify procedures used by others to fit a sample design into the BRR model. The literature shows that, when the sample in a certainty PSU randomly split into two pseudo-PSUs, variance estimates are upwardly biased (Rao and Shao 1996 and Rao and Shao 1999). Moreover, it may be difficult to split some certainty PSUs into two pseudo-PSUs if the sample within the PSU is highly stratified and with small sample counts.

We explored a series of approaches and identified two strategies to study more fully. One approach combined the use of the split sample approach with a collapsed stratum component and the other approach used only the split sample approach. We found little to distinguish the results in comparison to the use of explicit estimation equations with Taylor series linearization for nonlinear statistics. We choose to use the 40 pseudo-strata BRR design because it followed more closely the literature available. However, we are concerned that randomly dividing the 5 single-hit certainty PSUs into 2 pseudo-PSUs would result in small sample sizes for some domains in these pseudo-PSUs.

We suggest that further research be conducted to develop and evaluate other methods to handle certainty PSUs or to handle stratified simple random samples when the use of the BRR variance estimation procedure is desired.

REFERENCES

Brick J.M., P. Broene, P. James and J. Severynse (1996), "A User's Guide to WesVarPC." Rockville, MD: Westat, Inc.

Chromy, J.R. (1979) "Sequential Sample Selection Methods." Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 401-406.

Folsom, R.E., F.J. Potter, and S.R. Williams. (1987) "Notes on a Composite Size Measure for Self-Weighting Samples in Multiple Domains." Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 792-796.

Judkins, D.R. (1990) "Fay's Method for Variance Estimation." Journal of Official Statistics, Vol. 6, No. 3, pp. 223-239.

LaVange L., S.C. Stearns, J.E. Lafata, G.G. Koch and B.V. Shah (1996) "Innovative Strategies using SUDAAN for Analysis of Health Surveys with Complex Survey Samples." Statistical Methods in Medical Research, Vol. 5, pp.311-329.

McCarthy, P.J. (1966) "Replication: An Approach to the Analysis of Data from Complex Surveys." Vital and Health Statistics, ser. 2, no. 14, Washington D.C.: National Center for Health Statistics.

Potter, F.J. and N. Diaz-Tena (2003) "Weighting, Nonresponse Adjustments and Imputation: National Survey of SSI Children and Families." Prepared for the Social Security Administration by Mathematica Policy Research, Inc. under contract 0660-00-60021 (Task 910)

Potter F.J. and S. Mitchell (2000). "Report on Sampling Design and Estimated Survey Costs: Evaluation of the Effects of the 1996 Welfare Reform Legislation on Children with Disabilities: Survey Design and OMB Clearance Package." prepared for the Social Security Administration by Mathematica Policy Research, Inc.

Rao J.N.K and J. Shao (1996), "On Balanced Half-Sample Variance Estimation in Stratified Random Sampling." Journal of the American Statistical Association. Vol. 91 pp. 343-348.

Rao J.N.K and J. Shao (1999), "Modified Balanced Repeated Replication for Complex Survey Data." Biometrika. Vol. 86 pp. 403-415.

Research Triangle Institute (2001). "SUDAAN User's Manual, Release 8.0." Research Triangle Park, NC: Research Triangle Institute

Rust, K.F. and Rao, J.N.K. (1996). "Variance Estimation For Complex Surveys Using Replication Techniques. Statistical Methods in Medical Research, Vol. 5, pp. 283-310.

Shao J. and D. Tu. (1995) "The Jackknife and Bootstrap". New York: Springer-Verlag.

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

TABLE 1

RELATIVE DIFFERENCES BETWEEN BRR STANDARD ERRORS AND EXPLICIT ESTIMATION EQUATIONS

	Balanced Repeated Replication Strata							
	38 Strata	40 Strata	38 Strata	40 Strata	38 Strata	40 Strata	38 Strata	40 Strata
	Overall		Male		Female			
Mean	4.53%	5.07%	2.73%	4.42%	-2.28%	-2.35%		
Median	2.87%	3.31%	0.59%	2.87%	-1.78%	-2.33%		
Standard Deviation	0.047	0.039	0.057	0.054	0.023	0.011		
Race	Hispanic		Black		Unknown/Not Given		White/Other	
Mean	4.32%	-0.65%	1.16%	0.27%	-0.89%	0.04%	0.24%	0.24%
Median	4.34%	-0.81%	1.59%	-0.12%	-1.99%	0.23%	0.07%	0.25%
Standard Deviation	0.050	0.025	0.016	0.009	0.019	0.021	0.004	0.004
Diagnosis	Diagnosis 1		Diagnosis 2					
Mean	0.78%	1.74%	0.12%	1.20%				
Median	-0.55%	0.50%	0.41%	0.06%				
Standard Deviation	0.042	0.047	0.036	0.038				
Payment (Dollars)	Total Dollars		Mean Dollars		Median Dollars			
Mean	1.92%	1.80%	0.42%	0.85%	-0.35%	0.80%		
Median	0.71%	0.42%	0.16%	0.41%	-4.12%	-6.32%		
Standard Deviation	0.044	0.035	0.013	0.016	0.761	0.803		

Relative differences (in percent) between BRR standard error and standard error based on Taylor series linearization for nonlinear estimators.

Note: 38 Strata: 1 pseudo-stratum for the double-hit PSU, 3 pseudo-strata for the single-hit PSUs, 34 pseudo-strata for the 68 noncertainty PSUs.

40 Strata: 1 pseudo-stratum for the double-hit PSU, 5 pseudo-strata for the single-hit certainty PSUs, 34 pseudo-strata for the 68 noncertainty PSUs.