AN APPLICATION OF BALANCED REPEATED REPLICATION TO THE ESTIMATION OF VARIANCE COMPONENTS

Judy A. Bean, University of Iowa
George A. Schnack, National Center for Health Statistics

1. INTRODUCTION

For purposes of designing surveys, survey statisticians need to know the components of variation inherent in the stages of a sampling plan or be able to estimate them from previous surveys. This paper presents the results of applying the balanced repeated replication (BRR) technique to data collected in the Health Interview Survey in order to estimate the variance components of four statistics.

In recent years, the BRR method has been adopted for estimating variances of estimates from complex probability surveys but has not been employed for estimating variance components. In 1975 Casady (3) showed for the first time that the BRR method can be adapted to estimate the variance components of a linear estimator from a two-stage stratified design. When sampling without replacement at both stages, the BRR estimators of total variance and within variation are biased. The between variability is estimated by subtracting the within estimate from the estimate of total variance. Bean (2) has derived another version of the BRR technique that yields unbiased estimates of the within component for the same sample design. However, no one has investigated the use of the method for survey designs that are more complicated than a simple two-stage stratified one.

2. METHODOLOGY

2.1 The BRR Estimators

Before describing the methodology of this study, the BRR estimators of variance components for a simple design will be featured.

Let us consider a finite population of N primary units classified into L strata each containing N_i units (i = 1, 2, ..., L) with $N = \sum_{i=1}^{L} N_i$.

Each primary unit consists of M_{ij} elements. Denote by X_{ijk} the measurement of interest on the k^{th} element in the j^{th} primary unit of the i^{th} stratum and by

$$X_{ij.} = \sum_{k=1}^{M_{ij}} X_{ijk}, \quad X_{i..} = \sum_{j=1}^{N_i} X_{ij.},$$

and $X_{...} = \sum_{i=1}^{L} X_{i..}$ the population primary unit total, the population stratum total and the population total. A random sample of n_i units is drawn without replacement from the i^{th} stratum; within each selected primary unit, m_{ij} elements are selected randomly without replacement. The n_i's and m_{ij}'s are assumed to be even numbers. The customary unbiased estimator of the population total, X..., is

$$X' = \sum_{i=1}^{L} \frac{N_i}{n_i} \sum_{j=1}^{n_i} \frac{M_{ij}}{m_{ij}} \sum_{k=1}^{m_{ij}} X_{ijk} \tag{1}$$

and the variance of X' is:

$$\sigma^2_{X'} = \sum_{i=1}^{L} N_i^2 \; (1 - f_i) \; n_i^{-1} \; S_i^2$$

$$+ \sum_{i=1}^{L} \sum_{j=1}^{N_i} N_i \; M_{ij}^2 \; (1 \; f_{ij}) \; n_i^{-1} \; m_{ij}^{-1} \; S_{ij}^2 \tag{2}$$

where f_i = the first stage sampling fraction, f_{ij} = the second stage sampling fraction,

$$S_{ij}^2 = \sum_{k=1}^{M_{ij}} (X_{ijk} - \bar{X}_{ij.})^2 \; (M_{ij} - 1)^{-1}, \text{ and}$$

$$S_i^2 = \sum_{j=1}^{N_i} (X_{ij.} - \bar{X}_{i..})^2 \; (N_i - 1)^{-1} \; .$$

The first term on the right-hand side of the equation (2) is the variability between primary units. The second term is the variation among the elements within the primary units.

To obtain an estimate of the total variance for X' by the BRR procedure, the n_i sampled primary units are randomly split into two groups, each of size $n_i/2$. Next, using an orthogonal matrix (for more details see McCarthy (7)), A half-samples are created by randomly selecting one of the two groups of the primary units from each of the L strata. Utilizing only the data from each half-sample, A estimates of the population parameter are made. The BRR estimate of $\sigma^2_{x'}$, is

$$\hat{\sigma}^2_{x'} = \sum_{\alpha=1}^{A} (X'_\alpha - X')^2 \; A^{-1} \tag{3}$$

To estimate the within component, denoted as σ^2_w, each of the primary units is considered to be a pseudostratum. Here, the m_{ij} sampled elements are randomly placed in one of two equal sized groups. A half-sample, thus, consists of choosing one of the two groups of elements from each of the n_i primary units. The data from a half-sample is subjected to the same estimation procedure as the data from the total sample, creating another estimate of X... By means of a second orthogonal pattern, B estimates of X' are produced. Then an estimate of the within component is:

$$\hat{\sigma}^2_w = \sum_{\beta=1}^{B} (X'_\beta - X')^2 \; B^{-1} \tag{4}$$

2.2 Sample Data

The data for the study were those collected in the 1973 Health Interview Survey (HIS) of 120,493 civilian noninstitutional individuals. A description of the survey has been published by the National Center for Health Statistics (9) but the sample design and estimation procedure used will be outlined to illustrate its intricacies.

The sampling plan of HIS is to select one primary unit which is either a county or group of counties of the United States from each of 376 strata with probability proportional to size. Some of the strata contain only one primary unit. The second stage units chosen are clusters of approximately 4 households. For each selected household, information concerning a person's perception of his/her health is gathered for each person residing at the household.

After these data are subjected to an extensive editing procedure, estimates of morbidity are produced using a complex estimation equation. The equation includes unequal weighting caused by unequal probabilities of selection, nonresponse adjustment and two ratio adjustments.

To recapitulate, features of the design are unequal probabilities of selection, stratification, clustering and strata containing only one unit. For estimation purposes, an adjustment for nonresponse and two ratio adjustments are performed.

2.3 Study Design

An underlying assumption of the BRR method is that at least two units are chosen from each stratum; however, for surveys not fulfilling this requirement, the practice is to pair primary units based on characteristics of the strata they represent. The sampled primary units in HIS from strata consisting of more than one unit were collapsed to form pseudostrata; strata consisting of one primary unit each were grouped together in a particular fashion to form an additional set of pseudostrata which will be called self-representing (SR) pseudostrata. A distinction is made between the two groups because the variation in the SR pseudostrata only reflects the within variability, not the between variation. This is taken into account when estimating the variance of estimates.

Even after the 376 strata are collapsed into pairs, there are still 160 pseudostrata which means a 160 x 160 orthogonal matrix is needed to estimate the variance of an estimate. The number of half-samples required for the BRR method equals the first multiple of 4 large as or larger than the number of pseudostrata. Since each primary unit is assumed to be a pseudostratum for estimating the within component, the size requirement for an orthogonal matrix here is greater than 160 x 160. Because the main objective of the investigation was to simply demonstrate that the BRR method can be applied, the decision was made to use only data from the South region. The reason for choosing this geographic location was that the South was the largest; it consists of data for 38,053 persons.

For clarification the steps involved in the preparation of the sample data for use by the BRR method are reviewed.

A. Estimation of total variance:

Here the SR primary units were grouped to form 10 pseudostrata; the remaining units were paired into an additional 61 strata. Within each stratum there must be two primary units. For the 10 SR pseudostrata, the clusters of households within each pseudostratum were randomly partitioned into two groups. The other 62 pseudostrata consisted of two primary units each. Thus, with a total of 71 pseudostrata, the size requirement for the orthogonal matrix is 72 x 71.

B. Estimator of within variance:

To use the BRR method here, the assumption of two units selected from each stratum must be met. First, each of the sampled primary units in the 61 non-SR pseudostrata was considered to be a pseudostratum resulting in 122 pseudostrata. Secondly, within each of these primary units the clusters of 4 households were randomly allocated into one of two groups. The partitioning of the 10 SR pseudostrata for step A was retained for this step. Because a 132 x 132 orthogonal matrix does not exist, a 136 x 132 matrix was employed.

2.5 Variance Estimators

For each statistic produced, its variability was estimated in two ways using the BRR method; these two versions are described by McCarthy (7). The variance estimators are:

$$\hat{\sigma}^2_{\theta''} = \sum_{\alpha=1}^{72} (\theta'_\alpha - \theta'')^2/72 \qquad (5)$$

and $${}_c\hat{\sigma}^2_{\theta''} = \sum_{\alpha=1}^{72} (\theta^*_\alpha - \theta'')^2/72 \qquad (6)$$

where θ'' = the final nonresponse ratio adjusted estimate, θ'_α = the nonresponse ratio adjusted estimate secured from the α^{th} half-sample, and θ^*_α = the nonresponse ratio adjusted estimate secured from the α^{th} complement half-sample (the primary units not in the α^{th} half-sample).

A comment on the estimates produced from the half-samples is necessary. As mentioned earlier, three sets of adjustment factors are applied in order to take advantage of ratio estimation, poststratification and imputation for nonresponse. Therefore, the correct method for estimation is the calculation of these adjustment factors for each particular half-sample. This is straight forward but requires considerable work. Studies by Simmons and Baird (10) and Kish and Frankel (4,5) indicate that the adjustment factors based on the parent sample can be applied without the estimates being seriously biased. Contrarily, the results of investigations by Bean (1) and Lemeshow (6) conclude that the adjustment of factors should be computed for each specific half-sample. Due to costs and time for this feasibility study, the adjustment factors for the entire sample were applied to estimate within and total variance.

There were 132 pseudostrata (10 SR pseudostrata and 122 others) and no known 132 x 132 orthogonal matrix; thus, an orthogonal matrix, 136 x 132, was utilized in computing the BRR estimate of the within component of variation. The estimators are:

$$\hat{\sigma}^2_w = \sum_{\beta=1}^{136} (\hat{\theta}_\beta - \theta'')^2/136 \qquad (7)$$

and $${}_c\hat{\sigma}^2_w = \sum_{\beta=1}^{136} (\tilde{\theta}_\beta - \theta'')^2/136 \qquad (8)$$

where $\hat{\theta}_\beta$ = the nonresponse ratio adjusted estimate produced from the β^{th} half-sample, and $\tilde{\theta}_\beta$ = the nonresponse ratio adjusted estimate produced from the β^{th} complement half-sample.

3. EMPIRICAL RESULTS

As stated previously using the Health Interview Survey data for the South region, the BRR technique was applied to produce estimates of total variance and within variation. McCarthy (8) shows that if the average of the half-sample estimates of the parameter is essentially the same as the total sample estimate of the parameter, the differential bias of the average and the estimator θ" will be close to zero. This is important since the BRR estimate of variance will reflect that differential bias. A relationship between the variance of the mean of the half-sample estimates and the variance of θ" is derived. From this McCarthy infers that when this differential bias is small the estimate of the variance of θ" is "good".

For the data presented in this paper, Table 1 gives the mean of the half-sample estimates, the mean of the complement half-sample estimates and the total sample estimates. The means are close to the value of θ"; thus, the inference is that the BRR estimate of variance is "good". Besides this evidence, Bean (1) has demonstrated that the BRR method yields a satisfactory estimate of variance of a ratio estimator.

In calculating an estimate of within variability, half-sample estimates of the population parameters are computed. The mean of these half-sample estimates and the mean of their complement half-sample estimates are presented in Table 2 along with total sample estimates. These three estimates are almost identical, meaning the differential bias here is near zero. One may wish to argue that if this bias is close to zero the estimate of variance which in this situation is an estimate of the within component is a "good" estimate; however, such an argument is based on the fact that the type of relationship found for the total variance estimate must hold for the within component estimate. To date, there is no derivation of the relationship of the variances here so the results are to be interpreted cautiously. The conclusion is that the differential bias between the mean of the half-samples estimates/complement half-sample estimates and θ" is negligible so the within component estimate is not inflated by the bias.

The estimates of variance using the BRR method are shown in Table 3. For example, 72.69% of the population living in the South saw a doctor last year. This estimate has a variance of 15.0×10^{-8}; the variance is partitioned into 3.0×10^{-8} from sampling the primary units and 11.9×10^{-8} from sampling within the primary units. For three variables, the within estimate is less than the total; both BRR methods give similar values. For the variable dental visits, the within estimate is larger than the estimate for total. Thus, the between estimate is negative which causes some embarrassment. Presently, no answer to the question of what can be done when this event happens is available. To assume the component is zero implies the primary units do not vary among themselves which is unlikely.

Perhaps a more meaningful statistic is displayed in Table 4. The numbers in the table give the percent contribution of each component. The within component contributes approximately 79% of the variability for the three variables number of restricted activity days, number of bed disability days and proportion of population seeing a physician. The variables represent aggregate estimators and a PQ type.

4. DISCUSSION

The results presented here are encouraging but a considerable amount of research remains. The reason for encouragement is that earlier studies performed by statisticians at the National Center for Health Statistics suggest that, for a typical statistic in HIS, the between PSU contribution to variance is in the range of 10% to 20%. For this study the between PSU component is about 20%. One concern about the findings is that the components are too similar. Later work, not given here, indicates that with a different pairing of the PSU's, more realistic component values are obtained. Therefore, an investigation of the effect of varying the pairing scheme may be necessary. We are presently preparing to do additional computations using other statistics for the full 376-PSU sample design in order to assess the problem.

One of the criticisms made of the BRR technique is that the estimates of variance components can not be computed using this method. However, with the work of Casady (3) and this feasibility study this criticism is no longer valid. The purpose of the investigation, to demonstrate that the BRR technique can be utilized to produce estimates of variance, has been accomplished. Whether or not these are the "best" estimates of the components cannot be answered. The limited evidence presented indicates that the estimates are reasonable. Not only are investigations comparing different methods for estimating variance components needed but further theoretical work must be done in order to estimate the variance within the strata. The estimates of stratum variances are the crucial values in designing other surveys.

REFERENCES

1. Bean, Judy A., (1975), "Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples," Data Evaluation and Methods Research, PHS Publication No. 75-1339, Series 2, No. 65, Washington, D.C.: Government Printing Office.

2. Bean, Judy A., (1977), "Unbiased BRR Estimator of Within Component for Two Stage Stratified Design." University of Iowa (Internal Memorandum).

3. Casady, Robert J., (1975), "The Estimation of Variance Components Using Balanced Repeated Replication," Proceedings of the Social Statistics Section of the American Statistical Association, 352-357.

4. Kish, Leslie and Frankel, Martin R. (1968), "Balanced Repeated Replication for Analytical Statistics," Proceedings of the Social Statistics Section of the American Statistical Association, 2-10.

5. ________________ (1970), "Balanced Repeated Replication for Standard Errors", Journal of the American Statistical Association, 65, 1071-1094.

6. Lemeshow, Stanley (1976), "The Use of Unique Statistical Weights for Estimating Variances with the Balanced Half-Sample Technique," Proceedings of the Social Statistics Section of the American Statistical Association, 507-512.

7. McCarthy, Philip J. (1966), "Replication: An Approach to the Analysis of Data from Sample Surveys," Vital and Health Statistics, PHS Publication No. 1000, Series 2, No. 14, Washington, D.C.: Government Printing Office.

8. McCarthy, Philip J. (1969), "Pseudoreplication: Further Evaluation and Application of the Balanced Half-Sample Technique," Vital and Health Statistics, PHS Publication No. 1000, Series 2, No. 31, Washington, D.C: Government Printing Office.

9. National Health Survey (1958), "The Statistical Design of the Health Household-Interview," Health Statistics, PHS Publication No. 584-A2, Washington, D.C.: Government Printing Office.

10. Simmons, Walt R. and Baird, James T., Jr. (1968), "Pseudoreplication in the NCHS Health Examination Survey," Proceedings of the Social Statistics Section of the American Statistical Association, 19-30.

11. Tepping, Benjamin J. (1968), "The Estimation of Variance in Complex Surveys, "Proceedings of the Social Statistics Section of the American Statistical Association, 11-18.

Table 1. Comparison of the Estimate for the Total Sample with the Averages of the Half-Sample Estimates Used in Estimating Total Variance[a]

		Averages	
Variable	θ''	Half-Sample	Complement Half-Sample
Number of restricted activity days	1,197.57 x 10^6	1,201.56 x 10^6	1,193.56 x 10^6
Number of bed disability days	479.18 x 10^6	480.25 x 10^6	478.09 x 10^6
Number of dental visits	81.24 x 10^6	80.68 x 10^6	81.81 x 10^6
Proportion of population seeing a physician	72.69 x 10^{-2}	72.73 x 10^{-2}	72.76 x 10^{-2}

[a] See the text for a description of these estimates.

Table 2. Comparison of the Estimate for the Total Sample with the Averages of the Half-Sample Estimates Used in Estimating Within Variation[a]

		Averages	
Variable	θ''	Half-Sample	Complement Half-Sample
Number of restricted activity days	1,197.57 x 10^6	1,196.29 x 10^6	1,198.82 x 10^6
Number of bed disability days	479.18 x 10^6	477.48 x 10^6	480.86 x 10^6
Number of dental visits	81.24 x 10^6	81.24 x 10^6	81.24 x 10^6
Proportion of population seeing a physician	72.69 x 10^{-2}	72.67 x 10^{-2}	72.71 x 10^{-2}

[a] See the text for a description of these estimates.

Table 3. Balanced Repeated Replication Estimates of Total Variance and Components for Four Variables[a]

	Variance Estimates					
	Half-Sample			Complement		
Variable	Total	Between	Within	Total	Between	Within
Number of restricted activity days	$1,063.8 \times 10^{12}$	228.7×10^{12}	835.1×10^{12}	$1,063.7 \times 10^{12}$	228.6×10^{12}	835.1×10^{12}
Number of bed disability days	252.1×10^{12}	53.7×10^{12}	198.4×10^{12}	252.1×10^{12}	53.8×10^{12}	198.3×10^{12}
Number of dental visits	652.4×10^{10}		703.6×10^{10}	652.5×10^{10}		703.6×10^{10}
Proportion of population seeing a physician	150.0×10^{-7}	30.4×10^{-7}	119.6×10^{-7}	150.9×10^{-7}	31.8×10^{-7}	119.1×10^{-7}

[a] See the text for a description of these estimates. A blank indicates the estimate of variance was negative.

Table 4. The Percent Contribution of Each Component to the Total Variance

Variable	Contribution	
	Between	Within
Number of restricted days	21.5%	78.5%
Number of bed disability days	21.3%	78.7%
Number of dental visits[a]		
Proportion of population seeing a physician	20.3%	79.7%

[a] A blank indicates the percentage was either negative or over a hundred.