RELATIVE COMPUTATIONAL EFFICIENCY OF THE LINEARIZED AND BALANCED REPEATED REPLICATION PROCEDURES FOR COMPUTING SAMPLING VARIANCES

Kurt Maurer, Gretchen Jones and Earl Bryant

National Center for Health Statistics

INTRODUCTION

Two widely used generalized techniques for computing sampling errors of estimates derived from complex surveys are the Balanced Repeated Replication Procedure (BRR) and the Linearized Procedure (LIN) based on Taylor series approximations. Several research studies have compared the two methods from the standpoint of variance, bias, and mean square error of estimated sampling variances. 1,2,3,4/ Each of these studies has shown that both procedures produce reliable estimates of sampling variances when the estimates are based on a reasonably large number of degrees of freedom.

For linear-type estimators such as means and population totals, these empirical studies have shown that the bias in variance estimates derived from either BRR or Linearized estimators is trivial, even when the number of degrees of freedom or the number of strata is small. And even for certain non-linear estimators such as simple or partial correlation coefficients, and regression coefficients, bias is a relatively insignificant component of the mean square error for BRR as well as LIN. Variance is the main problem, but fortunately the magnitude of the variance, as well as bias can be controlled by increasing the number of strata or the number of degrees of freedom in the estimator. On the basis of empirical evidence cited, the estimators seem to be consistent, i.e., the mean square error decreases as the number of degrees of freedom increases. For BRR, this means that the mean square error of the variance decreases with an increase in the number of replicates since there is a direct relationship between the number of replicates required for orthogonality and the number of strata.5/ (For a two PSU-per-stratum sample design, the number of half-sample replicates required to achieve orthogonality is at most three more than the number of strata.)

A criticism often heard of BRR relates to computational efficiency, especially in situations where a large number of replicates is required. The popular opinion seems to be that the Linearization method in such situations is preferred since data would need to be passed through the computer only once, while for BRR, the data would need to be read many times to compute half-sample estimates. The purpose of this paper is to compare the two methods with respect to the amount of computer time required to generate variances. The factors investigated that affect computing time include number of strata, number of replicates, size of the data file, and the size of table for which variances are required.

The data for this study were collected in two health surveys of the U.S. population, the National Health Interview Survey (HIS) of 1976 and the National Health and Nutrition Examination Survey (HANES) of 1971-1975. Both surveys were based on rather complex sample designs with a first stage selection of PSU's (counties or groups of contiguous counties) and a second-stage selection of segments or clusters of housing units (HU's).

The HIS design was based on 376 PSU's; one PSU was selected per stratum with probability proportional to the 1970 population. Within PSU's, compact segments of approximately four HU's were selected at a rate such that the over-all probability of selection was a constant. Then all civilian non-institutionalized persons who lived in sample housing units were interviewed. Approximately 40,000 households and 120,000 people are interviewed each year.

The HANES design is similar to the HIS design in that a self-weighting sample of housing units was selected in two stages. The designs differ in that for HANES a sample of persons was selected from sample housing units with differential probabilities depending on a person's age and sex. They also differed in the number of PSU's, the way the PSU's were selected, and the number of sample persons. The initial design of the HANES sample called for a total of 65 PSU's and about 28,000 sample persons. Fifteen of the PSU's were self-representing while the remaining 50 resulted from selecting two PSU's per stratum without replacement. To be able to produce early estimates of the nutritional status of the American people, the first year or two of HANES (round 1) concentrated on examining a representative sub-sample of the population. The sub-sample consisted of a random selection of one of the two initially selected PSU's from each of the 25 strata and a selection of 10 of the initially selected self-representing PSU's.

After examinations for the 65 PSU design had been completed, a decision was made to extend the survey to an additional 35 PSU's and to increase the sample of adults 25-74 years of age. In this supplemental survey, only health data (in contrast to nutrition data) were collected. The sample of additional PSU's consisted of the 10 "self-representing" PSU's that were in the round 1 sample and 25 PSU's selected independently from the 25 non-self-representing strata. Thus estimates of certain health parameters for the adult population were based on data collected in 90 distinct PSU's.

The estimators for the two surveys were also similar. In addition to weights based on reciprocals of selection probabilities, the data were adjusted for non-response and were post-stratified to known population totals provided by the U.S. Bureau of the Census, according to age, race, and sex.

For the HIS, the 376 PSU's were collapsed into 149 strata. Variances were computed using the BRR method based upon 152 balanced half-sample replicates and were computed using the

LIN method based upon the 149 strata with 2 PSU's per stratum.

For the HANES, estimates have been derived from 35, 65 and 90 PSU designs. Variances for these designs were computed using the BRR method based upon 20, 40 and 81 replicates respectively. All were half-sample replicates, except for the 90 PSU design which were third-sample replicates, since 3 PSU's were selected from each stratum. For the LIN method, the PSU pairings were the same as for BRR in non-self-representing strata. However, for self-representing strata, variance computations were based on first stage units or clusters of about 8 housing units each, the number of clusters per stratum ranging from 92 to 235 depending on the size of a stratum.

In this study, comparisons of computer times required to compute variances were made based upon the number of replicates, the number of records being processed, and the size of table. The following table summarizes the research plan:

Size of Table	Number of Replicates	Number of Records
3 x 3	152	6900
3 x 3	81	6900
3 x 3	40	6900
3 x 3	20	6900
3 x 3	40	20,000
3 x 3	40	15,000
3 x 3	40	10,000
3 x 3	40	5,000
10 x 11	40	6900
10 x 11	40	6900
10 x 11	40	6900
10 x 11	40	20,000
26 x 14	40	20,000
10 x 9 (6 Tables)	40	20,000

Variances for each of the above situations have been computed using two BRR programs that have been developed by Gretchen Jones of the National Center for Health Statistics and a linearized program, called the "STDERR" program, that was developed by B.V. Shah of the Research Triangle Institute 6/ and is to be included as part of SAS 7/ in the near future. All of the tables for this paper were run on an IBM 370 model 158 computer with a 360 operating system.

The HANES BRR program was specifically developed and written for use on the HANES and previous health examination surveys. Even so, it is a highly generalized program with extensive recoding and labeling features. It is capable of generating tables with two nested row variables, two nested column variables and multiple planes all in one pass of the input data file. More than one table can be produced per job but one pass of the input data file is required for each table. The HANES BRR program requires 8 disk drives, 2 tape drives, a terminal for job entry and 1 printer. The program uses about 250K bytes of core storage.

The Hanes BRR program was written in PL/I and operates in the following manner. Records containing demographic data, medical data, the vector of half-sample weights, plus the full sample weight are read into the first major program from a direct access device. A detail output record is created from each input record containing the cell code plus the vector of half-sample weights and the full sample weight. The weights on these records are accumulated for each cell and totals and subtotals are calculated. These half-sample and total sample estimates are read into the third major program, which calculates the requested statistics and prints out the desired tables.

The HIS BRR program, also written in PL/I, was specifically developed and written for use on the HIS, but unlike the HANES BRR program it has limited labeling features and no recoding features. The HIS BRR program requires 5 disk drives, 3 tape drives, a terminal for job entry and 1 printer. This program also uses about 250K bytes of core storage. Each time a new set of variance items is desired, a program must be written which reads the weighted data from a direct access device and accumulates these weights for each data item catagorized by PSU-age-race-sex. The output of this program contains the item identification, the PSU-age-race-sex code and the accumulated weight. The main program reads these records, the half-sample indicator matrix, and the post-stratification factor matrix. The half-sample estimates are accumulated by using the PSU code to access the appropriate row of the half-sample indicator matrix and multiplying the weight by the appropriate post-stratification factors. The statistics are then calculated and printed.

At this time the SAS LIN program interfaces with but is not a part of SAS. Because of this configuration, it takes advantage of the extensive recoding features of SAS. The input data are read from a SAS file contained on a direct access device. The program is written in FORTRAN and uses FORTRAN to produce its output. The version of the program available to us at the time did not produce neatly formatted crosstabulations, but this capability has been included in a more recent version of the program.

FINDINGS

It was hypothesized that four factors would affect the amount of time required to produce variance estimates: the number of strata, the number of PSU's the size of the table and the number of records used as input.

The number of strata or PSU's appears to have little effect on CPU time for the SAS LIN program (Tables 1 and 2) but does affect the CPU time for the BRR programs. The amount of CPU time required for the HANES BRR program based on 81 replicates was about twice the time required for 20 replicates. In Table 1 it can be seen that the HANES BRR program took more than twice as much CPU time to produce tables as did the SAS LIN program. However, in Table 2 the HANES BRR program took as little as one-third the CPU time required by the SAS LIN program. The difference between Tables 1 and 2 is that the first contains CPU times based upon 9 cells per table and the second was based upon 110 cells per table.

Table 1. Computer running times for the BRR and LIN procedures according to the number of replications, strata, and PSU's, based on 9 cells per table and 6900 input records.

Survey and Number of Replications	Procedure		Number of (Pseudo) Strata		Number of PSU's Per (Pseudo) Stratum
	BRR	LIN	Self Representing	Non-Self Representing	
HIS	(CPU time in seconds)				
152	54.09	23.43	35	114	2
HANES					
81	61.73	28.27	15	25	3*
40	44.87	27.01	15	25	2*
20	26.91	28.10	4	15	2*

*"PSU's" in HANES self-representing strata for the Linearization method consisted of clusters of about 8 housing units. The number of clusters per stratum ranged from about 92 to 235.

Table 2. Computer running times for the BRR and LIN procedures according to the number of replications, strata, and PSU's based on 110 cells per table and 6900 input records.

Survey and Number of Replications	Procedure		Number of (Pseudo) Strata		Number of PSU's per (Pseudo) Stratum
	BRR	LIN	Self Representing	Non-Self Representing	
HANES	(CPU time in seconds)				
81	65.90	85.55	15	25	3*
40	45.45	85.56	15	25	2*
20	29.68	87.09	4	15	2*

*"PSU's" in HANES self-representing strata for the Linearization method consisted of clusters of about 8 housing units. The number of clusters per stratum ranged from about 92 to 235.

The effect of the number of cells in each table on the CPU time became even more apparent as a significant factor in Table 3. The CPU time for the HANES BRR program was unaffected by the size of the table but the time required for the SAS LIN program was dramatically affected, increasing from 74 seconds for a small table to 639 seconds for a rather large table. From Table 3 it can be seen that for small tables the HANES BRR program took about 40 percent more time than the SAS LIN program but for large tables it took as little as 17 percent of the time required for the SAS LIN program. For an intermediate size table of 110 cells, a typical size table at NCHS, the HANES BRR program took only 45 percent as much time as did the SAS LIN program. It was thought that there might be some increased savings when several tables were run at the same time, but that does not appear to be the case. In fact the average amount of time required per table of 90 cells each, when 6 tables were run at the same time, was higher than it was for 1 table of 110 cells. This was true for both the HANES BRR and SAS LIN programs. Also, when the 6 tables of 90 cells each were run at the same time, the SAS LIN program took more than twice the amount of time required by the HANES BRR program.

The last comparison made for this paper, based on 40 pseudo-strata and 9 cells per table, shows the effect of the number of records read as input to the two programs (Table 4). As expected, the CPU time increased for both programs with an increase in the number of input records, but the increase was larger for the HANES BRR program. All of the CPU times were greater for the HANES BRR program because the times in Table 4 were based upon small tables. If they had been based upon somewhat larger tables they would have been less for the HANES BRR program, as indicated in Table 3.

The time-consuming aspect of the HANES BRR program was in the input-output (I/O) operation. The input records were long, as they contained the vector of half-sample weights. Thus, the larger the number of replicates, the longer the input records, and consequently the more time required to produce variances. On the other hand, once the half-sample estimates were accumulated, the time required for computing variances was trivial. Thus, after the I/O was finished, the size of the table had little effect on computer time. The HIS BRR program does not require as much I/O time as the HANES BRR program because

the half-sample weights are not read in, but are calculated in the program. However, for each variance, accessing the half-sample indicator matrix for each accumulation by PSU-age-race-sex is not a trivial process, and accounted for the increase in calculation time for each cell in the table. Similarly for the SAS LIN program, the I/O time was small, but the calculation of each variance was time-consuming, because the Taylorized deviation algorithm was derived for each cell of the table. It appears that for small tables it is more efficient to use the SAS LIN or HIS BRR programs, and for larger tables, it is more efficient to use the HANES BRR program.

Table 3. Computer running times for the BRR and LIN procedures according to the number of tables and number of cells per table, based on 40 pseudo-strata and 20,000 input records, HANES.

Number of Tables	Number of Cells per Table	Procedure	
		HANES BRR	SAS LIN
		(CPU time in seconds)	
6	90	758.07	1694.20
1	364	105.93	638.81
1	110	109.03	241.23
1	9	105.90	74.40

Table 4. Computer running times for the BRR and LIN procedures according to the number of input records, based on 9 cells per table and 40 pseudo-strata, HANES.

Number of Records	Procedure	
	HANES BRR	SAS LIN
	(CPU time in seconds)	
20,000	105.90	74.40
15,000	79.47	55.72
10,000	56.16	38.79
5,000	31.79	20.13

CONCLUSIONS

It was our original intention to make a comparison of the computational efficiency of the Balanced Repeated Replication and the Linearized procedures for estimating variances. However, through the course of our investigation we found the task to be a difficult one. We do know that the two BRR programs and the SAS LIN program could be written more efficiently and perhaps that will be a beneficial outcome of this study. The SAS LIN program has its only apparent inefficiency in the production of large tables. The reason for this inefficiency is that the program re-derives the Taylorized deviation algorithm for each cell of a table. The HANES BRR program could be made more efficient by using extensive assembler routines to handle its input-output needs. Since the input-output needs of the HANES BRR program are great, a significant reduction in time could be realized, thereby making the production of small tables more efficient.

It is our conclusion that neither the BRR procedure nor the LIN procedure, per se, is inherently more efficient than the other. Rather, the amount of time required to produce variance estimates is dependent upon how the procedures are programmed. And the way in which the procedures are programmed is dependent upon the researcher's needs.

REFERENCES

1/ Woodruff, Ralph S. and Causey, Beverley D., "Computerized Method for Approximating the Variance of a Complicated Estimate." Journal of the American Statistical Association, June 1976, Vol. 71, No. 354.

2/ Bean, Judy A., "Distribution and Properties of Variance Estimators for Complex Multi-Stage Probability Samples - An Empirical Distribution," Vital and Health Statistics, Series 2, Number 65, National Center for Health Statistics, March 1975.

3/ Frankel, Martin, Inference from Sample Surveys, Ann Arbor, Michigan, University of Michigan, 1971.

4/ Baird, James T., Jr., Relative Stability of Selected Correlation and Regression Statistics in Complex Sampling. Unpublished Doctoral Dissertation, American University, 1976.

5/ McCarthy, Philip J., "Replication: An Approach to the Analysis of Data from Complex Surveys," Vital and Health Statistics, Series 2, Number 14, National Center for Health Statistics, April 1966.

6/ Shah, B. V., "STDERR: Standard Errors Program for Sample Survey Data (Preliminary SAS Version)", Research Triangle Institute, August 1976.

7/ Barr, Anthony J., et. al., "A User's Guide to SAS 76", SAS Institute Inc., 1976.