ON THE PERFORMANCE OF JACKKNIFE VARIANCE ESTIMATION FOR SYSTEMATIC SAMPLES WITH SMALL NUMBERS OF PRIMARY SAMPLING UNITS

John Burke and Keith Rust,* Westat Inc.
Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

1. Introduction

Many surveys of human populations are conducted using a sample design that involves the use of stratified multistage sampling. Often a relatively small sample of Primary Sampling Units (PSUs) is selected, using explicit and/or implicit stratification. Then, within each PSU, further sampling takes place to select the ultimate sample of units to be surveyed.

This type of design can give rise to a difficulty when making inferences about population subgroups. It may frequently happen that there is a particular subgroup of interest which contributes a substantial number of ultimate units to the sample, but which is drawn from only a few of the PSUs. The fact that the sample size is substantial means that, absent very extreme design effects, the subsample representing the population subgroup in question is very likely to give estimates for many parameters of interest that are sufficiently reliable to be useful. We are not discussing here the problem of rare a subpopulation with a sparse sample.

The fact that the sample is derived from only a few PSUs, however, makes reliable inference difficult. This is because, using any of the techniques available to give approximately unbiased estimates of sampling variance, the resulting precision of the sampling error estimates derived will be low. This means that, even though the estimate of the parameter of interest may have small or moderate variability, with a coefficient of variation (relative standard error) of less than ten percent for example, the unreliability of the estimated sampling variance makes it difficult to construct confidence intervals with the stated levels of coverage.

This is because direct variance estimators must, explicitly or implicitly, estimate the between PSU component of variance. The precision of this is limited by the number of PSUs from which the subpopulation is drawn. The true number of degrees of freedom for variance estimation might well be considerably less than the number of PSUs for three reasons. First, the variance estimator used may be designed to reflect the impact of PSU stratification. This will decrease the bias of the variance estimates, but also decreases the number of degrees of freedom. The use of paired PSUs for use in variance estimation is an example of this that is used very frequently in practice. With such an approach the maximum number of degrees of freedom achievable is reduced by almost fifty percent. Second, the sampling distribution within some PSUs (at least) may be affected by outliers in the population, which reduces the precision of variance estimates (that is, the number of degrees of freedom). Third, whereas the full sample may consist of PSUs of about equal size, when considering subpopulations, the distribution across PSUs (each containing some members of the subpopulation) may vary considerably. This can drastically reduce the number of degrees of freedom available.

In this paper we examine this phenomenon more closely for a particular survey for which, after the fact, there was considerable interest in subpopulation mean estimates derived from just a few PSUs. Using a simulated population, we examine the true levels of confidence interval coverage obtained using standard large sample procedures. We particularly wished to find out if there was evidence, for the kind of population involved, of a "breakdown point" in the sample size of PSUs, above which confidence interval coverage had acceptable rates, but below which it was unsatisfactory. In fact, as will be seen, we found that the problem of having a small number of degrees of freedom was somewhat overshadowed by the unpredictable relationship between sample size and sampling error when using systematic samples of small size. We did not find evidence of a particular breakdown point for sample size, although it is clear that, with these kind of data, it is unwise to estimate a confidence interval directly from a sample of two PSUs.

In Section 2, we describe the sample design for the private school population for the 1994 Trial State Assessment of Educational Progress. Section 3 describes the simulated population that we constructed. In Section 4, we describe the procedures used to obtain true sampling errors and true confidence interval coverage for samples ranging from two to thirty PSUs. Section 5 includes a summary of findings and our conclusions.

2. The 1994 Trial State Assessment of Educational Progress - Private School Component

In 1994, 45 states and jurisdictions participated in the National Assessment of Educational Progress (NAEP) state assessment of reading at grade 4. The NAEP program is administered by the National Center for Education Statistics (NCES). The NAEP state program included private schools for the first time in 1994.

*The authors wish to thank Bruce Spencer, who provided the suggestion that the performance of the jackknife variance estimator for the NAEP private school state samples should and could be investigated. Thanks are due also to Bruce Kaplan and his colleagues at Educational Testing Service, who provided us with the reading assessment data used in this study. The views expressed are those of the authors, and do not represent official positions of Westat or the National Center for Education Statistics.

The public school sample for each state consisted of about 100 schools, selected with probability proportional to grade 4 enrollment, using explicit and implicit stratification. The private school sample size varied across states, roughly in proportion to the proportion of grade 4 students enrolled in private schools. A few states with relatively low private school enrollment had a sample of just six schools. Some states had over twenty percent of students enrolled in private schools, and had sample sizes as high as thirty schools.

The sampling frame used was the file of private schools obtained from Quality Education Data, Inc. In each state the population was stratified into Catholic diocesan schools and other (Catholic diocesan schools account for about sixty percent of grade 4 private school enrollment nationally). The second major stratification variable was Metropolitan Statistical Area (MSA) status (MSA/non-MSA). After sorting the schools by these two variables, this list of schools was then sorted, within each of the four cells so created, by a variable which gave the median household income (1989) of the ZIP Code area in which the school was located. A systematic sample of schools was drawn in each state, with school selection probabilities proportional to a monotone function of estimated grade four enrollment. For full details of the sample design see Chapter 3 of Carlson and Allen (1995).

Within each selected school a systematic equal probability sample of thirty grade 4 students was selected. If the school had fewer than thirty students, all students were included in the sample.

The original intention of the sample design for private schools was to ensure that, when the public and private school samples were combined, reliable estimates of mean student reading proficiency, overall and for various demographic subgroups, would be obtained for each state. Subsequent to the assessment NCES decided to publish mean reading proficiency and other statistics for private school students for all states where prescribed minimal school and student response rates were achieved. While all participating states obtained the required student response rates, in many cases the school response rate was not adequate, and no private school results are to be reported for such states.

Sampling errors for private school estimates were prepared using the jackknife approach (see Wolter, 1985). In using these jackknife variance estimators, the expectation was that the number of degrees of freedom available would be close to (although somewhat less than) the number of replicates created. Rust (1986, 1984) discusses the likely degrees of freedom to be obtained from a range of variations of the jackknife procedure, including this approach.

3. An Artificial Study Population

To study the behavior of this method of variance estimation, when used with systematic samples of populations such as those encountered in sampling private schools for state NAEP, we created a population with 105 observations as follows. We used the observed mean reading proficiency from 105 schools that participated in the 1994 NAEP state assessment. We then treated these 105 schools as if they were the entire population of a single fictitious state. The schools were sorted systematically using the characteristics used to stratify the private school samples in each state (Catholic/non-Catholic, MSA/non-MSA, median household income of ZIP Code).

An examination of the distribution of school means, when sorted in sampling order, revealed that the systematic sort does little or nothing to explain the variation between school means. This is perhaps not especially surprising, since the school means are based on samples of thirty students. This illustrates that the school level variables used to sort the sample systematically do not explain a great deal of the combined school and student variance, which is not surprising. However, with student samples of size thirty in each school one might have expected to see a little more evidence of gains from the stratification of schools.

4. Sampling Variances and Variance Estimates

Having obtained the population described in Section 3, we then proceeded to evaluate the confidence interval coverage of standard error estimates obtained via the jackknife procedure, for all even numbered sample sizes from 2 to 30. In doing this, we evaluated the true sampling variance, the true distribution of the variance estimator, and the true confidence interval coverage for each sample size. The results given are exact, and are not from simulations.

However, in doing this, we have ignored the within school variance, treating the estimate of the school mean for each of the 105 schools as a fixed quantity. This means that results are an evaluation for a single stage systematic sample drawn with probability proportional to size. As such, we speculate that the sampling distribution, variance estimator and confidence interval coverage are likely to be less well-behaved than for the actual two-stage sampling used for NAEP. In particular, the sampling distribution is likely to resemble less closely a normal distribution than is the case in practice.

Although, as noted, the use of systematic sampling is likely to have given little reduction in sampling variance, the use of sampling with probability proportional to size is appropriate for this single stage sample. That is because the population mean that is being estimated is the student mean reading proficiency, rather than the mean of the school means. Thus schools contribute to this population mean in proportion to their enrollment size, so that selecting schools with probability proportional to enrollment constitutes an efficient design.

For each even numbered sample size from 2 to 30, we evaluated the true sampling variance We then considered the performance of jackknife variance estimates, and confidence intervals generated using them, again considering the distribution of all possible samples. We considered two methods of jackknife variance estimation. With the first we replicated closely the procedure used for the private school samples for the 1994 NAEP state assessment. We formed n replicates, where n is the total sample size. For samples of size 10 and fewer we formed the replicates by deleting each sample school in turn, and calculating the replicate estimate of mean reading proficiency as the mean of the remaining (n-1) schools. Letting $\hat{x}_t$ denote the mean associated with the t th replicate, we obtained the variance estimator (for $n \leq 10$) of

$$\mathrm{var}_{NAEP}\left(\hat{x}\right) = \frac{n-1}{n}\sum_{t=1}^{n}\left(\hat{x}_t - \hat{x}\right)^2 .$$

This is the standard jackknife variance estimator, unbiased for linear estimators when using a simple random sample (with replacement) of PSUs. Under favorable conditions this variance estimator has (n-1) degrees of freedom.

For samples of 12 or more schools, the replication scheme reflected the primary stratification into Catholic and non-Catholic schools. All such samples contained at least 2 schools of each kind. Let n_1 denote the number of Catholic schools, and n_2 the number of non-Catholic. The first n_1 replicates were formed by deleting each Catholic school from the sample in turn. In each case the remaining Catholic schools were reweighted by a factor of $n_1/\left(n_1 - 1\right)$, while the non-Catholic schools retained their full-sample weight. For the remaining n_2 replicates, each non-Catholic school was removed in turn, the remaining non-Catholic schools were reweighted by a factor of $n_2/\left(n_2 - 1\right)$ and the Catholic schools were given their full-sample weight.

Under this replication scheme, the appropriate variance estimator, when $n \geq 12$, is

$$\mathrm{var}_{NAEP}\left(\hat{x}\right) = \frac{\left(n_1 - 1\right)}{n_1}\sum_{t=1}^{n_1}\left(\hat{x}_t - \hat{x}\right)^2 + \frac{\left(n_2 - 1\right)}{n_2}\sum_{t=n_1+1}^{n_1+n_2}\left(\hat{x}_t - \hat{x}\right)^2 .$$

This is the standard jackknife variance estimator for a stratified (Catholic/non-Catholic) simple random sample with replacement. Under favorable conditions this variance estimator has $\left(n_1 + n_2 - 2\right)$ degrees of freedom.

The second form of jackknife variance estimator involved pairing the sampled schools, with adjacent schools in the systematic sort being paired. A replicate was formed by deleting one school from the sample, doubling the weight of its complementary pair member, and recalculating the mean score. The procedure was carried out n times, by dropping each school in turn. Letting $\hat{x}_t^*$ denote the replicate estimate when school t is omitted, a variance estimator that is approximately unbiased for linear estimators is given by

$$var_{PAIR}\left(\hat{x}\right) = \frac{1}{2}\sum_{t=1}^{n}\left(\hat{x}_t^* - \hat{x}\right)^2 .$$

This variance estimator should have less positive bias than var_{NAEP}, since it captures most of the sampling variance reduction (if any) that results from the systematic sampling procedure, which var_{NAEP} fails to do. On the other hand, var_{PAIR} is less precise than var_{NAEP}, having a likely maximum of n/2 degrees of freedom, rather than the (n-1) or (n-2) degrees of freedom for var_{NAEP}.

In using these two variance estimators to estimate confidence intervals, we considered two methods for creating confidence intervals in each case. This was achieved by using two different choices in each case for the coefficient, applied to the estimated standard error in forming two-sided 95 percent confidence intervals. That is, we varied the value of t in the expression

$$\hat{x} \pm t\sqrt{var\left(\hat{x}\right)} .$$

One choice was to use the value, t=1.96. This is often used in practice, but is most appropriate when $\mathrm{var}\left(\hat{x}\right)$ is very precisely estimated (that is, has many degrees of freedom). The alternative was to use the 97.5th percentile of a t distribution, using $\left(n-1\right)$ degrees of freedom for choice of the t distribution when using var_{NAEP} for $n \leq 10$, using $\left(n-2\right)$ degrees of freedom for var_{NAEP} for $n \geq 12$, and using n/2 degrees of freedom with var_{PAIR}.

The results of these different approaches are shown in the following tables. Table 1 shows the results when using var_{NAEP}. Column 2 shows the true mean squared error (MSE) of the mean.

The remaining columns in the table relate to the performance of the variance estimator, var_{NAEP}. Column 3 shows its mean across all possible samples, and Column 4 shows its bias. Column 5 shows the sampling variance of var_{NAEP}, and Column 6 its coefficient of variation (cv). Note that the degrees of freedom (df) of a variance estimator are related to this quantity via the expression

$$df = 2/cv\left(var\left(\hat{x}\right)\right) .$$

This shows that var_{NAEP} has 0.90 degrees of freedom with a sample of size 2 (compared with an expected 1 degree of freedom), and 5.88 degrees of freedom with a sample of size 30 (compared with an expected 28 degrees of freedom). Thus all sample sizes give rise to very few degrees of freedom, with this population and sample design.

Column 7 shows the mean square error (MSE) of var_{NAEP}, combining the bias from Column 4 with the variance from Column 5. It can be seen that the variance predominates with smaller sample sizes, while the squared bias predominates for samples of size 28 and 30.

The statistics of prime interest for our analysis are given in Columns 8 and 9. Column 8 shows the true coverage probability, over all possible samples, of the 95 percent confidence interval, given by

$$x \pm 1.96 \sqrt{var_{NAEP}(\hat{x})} \, .$$

The figures in Column 8 show that the true level of confidence interval coverage is somewhat erratic. Even having used t=1.96 when the variance estimators have relatively few degrees of freedom, for samples of size 22, 24, 28, and 30 the confidence intervals always include the true mean. The coverage is 76 percent for samples of size 2.

In using a t coefficient of 1.96 to form 95 percent confidence intervals, it is implicitly assumed that the variance estimator in question has many degrees of freedom (say 30 or more). Since that assumption is clearly inappropriate, we repeated the process of creating confidence intervals, but using a t coefficient that reflected the number of replicates formed. That is, as described above, we used the 97.5th percentile point from a t distribution with m degrees of freedom, where m=(n-1) for n≤10, and m=(n-2) for n≥12. This process of course creates wider confidence intervals than using 1.96, especially for small sample sizes.

The results are given in Column 9, which shows the effect of the alternative approach to forming confidence intervals. Substantial improvement in achieved coverage is seen for the smaller sample sizes. In fact, the major problem evident now is that the coverage rates are generally over 95 percent, and achieve 100 percent for samples of size 6, 22, 24, 28, and 30. In these cases the substantial positive relative bias of the variance estimator is leading to this phenomenon.

It is noticeable that poor coverage is achieved for samples of size 26 (86.6 percent). In fact, most of the confidence intervals that fail to include the mean only just fail to do so. Thus the exact coverage rate gives a somewhat overly pessimistic summary of confidence interval performance in this case.

A striking feature of the results in the table is seen by examining the second column. This shows that the true level of sampling error does not decrease monotonely with increasing sample size. This lack of monotonicity is not just observed among the smaller sample sizes. Although n=30 gives rise to the smallest mean square error (2.6851), the second lowest is n=28, next is n=22, then n=20, n=24, and n=26. This behavior contrasts, of course, with simple random samples, drawn with or without replacement, where the sampling error decreases monotonely with increasing n. Column 6 shows also that the precision of the variance estimator is also not monotone with sample size.

Given that the major drawback to the NAEP jackknife variance estimator with this population is the substantial positive bias demonstrated for some sample sizes, it is of interest to consider the performance of the paired jackknife variance estimator var_{PAIR}. This estimator attempts to reduce the bias present in var_{NAEP} from failing to reflect all of the gains from systematic sampling. This is expected to come at the expense of decreased precision of variance estimation. The results are shown in Table 2.

Comparing Column 4 of Table 2 (showing the bias of the paired jackknife) with Column 4 of Table 1 (showing the bias of the NAEP jackknife), we see evidence of reduced bias for larger sample sizes (20 to 30), but not for smaller sample sizes. Looking at Column 6, we see that the coefficient of variation of the estimator is indeed higher for var_{PAIR} for small sample sizes (for n=2 the two estimators are the same). This difference disappears for larger sample sizes, and indeed for n=28 and n=30, the paired approach actually gives rise to more precise estimates of variance despite having, nominally, considerably fewer degrees of freedom.

To see the impact of these differences in bias and precision on confidence interval coverage, we compare the second last column (Column 8) of Table 2 with that of Table 1. This comparison is for the case when t=1.96 is used as a coefficient. This shows that the two approaches give very similar coverage for samples of size 14 and greater, while for smaller sample sizes on balance var_{NAEP} is preferable to var_{PAIR}. This suggests that the decision to use var_{NAEP} with the NAEP private school samples, rather than using var_{PAIR}, was a good one, albeit not importantly so. The NAEP sample sizes varied from 6 to 30 or so across states, with a mean of around 14.

We turn now to a consideration of the use of a different coefficient for forming confidence intervals using the paired jackknife approach. For a coefficient we used the 97.5th percentile point from a t distribution with n/2 degrees of freedom. The result is shown in the last Column 8 of Table 2. Comparison with column in Table 2 shows that this approach improved coverage generally, bringing it closer to the desired 95 percent in most cases, and never causing a noticeable deterioration. In comparison with the last Column 8 in Table 1 we see that the paired jackknife gave closer to 95 percent coverage for samples of size 4, 12, and 24, but performed significantly worse than the NAEP jackknife for sample sizes of 10. This constitutes the most serious of the few points of discrepancy, since both jackknife estimators gave rise to under coverage, more serious for the paired jackknife.

5. Summary

The aim of this exercise was to examine the performance of the jackknife variance estimator, and large sample confidence intervals created using it, for small systematic samples from a population of a particular type. Our first finding was that in fact the true levels of sampling error for samples of this type are rather unpredictable. In particular, the relationship of sampling error to sample size is far from monotone, even for sample sizes between 20 and 30.

Given this situation, it seems that the jackknife estimator used in NAEP performs quite well, especially when used with a *t* coefficient that reflects the limit on the degrees of freedom available, rather than using 1.96. The exception is with samples of size 2 where, even though the confidence interval coverage may approach the stated level, the extreme width of the confidence intervals, and the variability in confidence interval width, render these of very little use.

Modifying the jackknife procedure, with the intention of decreasing the bias at the expense of greater variance, through the use of the paired jackknife procedure, generally had the intended result. The impact of this on confidence interval coverage was mixed, and not extensive. On balance we would argue that the NAEP jackknife procedure gives better results than the paired jackknife, but the difference is not great. We did also consider the effect of the addition of a finite population correction factor to the paired variance estimator, and found that it made little difference to the results.

We are encouraged by these results that valid inferences can be made from samples in the range of 6 to 30, using the jackknife procedure, with populations of schools and students within them, for estimating mean student proficiency. While confidence interval coverage sometimes strayed from the stated levels, it was generally in a conservative direction. The stability of the confidence intervals formed seems acceptable for all but the smallest sample sizes of 2 (and possibly 4). We recommend the use of the jackknife procedure for a design such as this. However, we caution that such designs do give rise to very few degrees of freedom for variance estimation, and the true level of sampling error for a given sample size probably cannot be predicted in advance with much certainty.

References

Carlson, J. and Allen, N. (eds). (1995, to appear). The Technical Report of the 1994 NAEP Trial State Assessment in Reading. National Center for Education Statistics: Washington, DC.

Rust, K.F. (1984). Techniques for Estimating Variances for Sample Surveys. Ph.D. Thesis, University of Michigan: Ann Arbor, MI.

Rust, K. (1986). Efficient Replicated Variance Estimation. Proceedings of the Survey Research Methods Section. American Statistical Association: Alexandria, VA.

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

Table 1. Performance of jackknife variance estimator using NAEP replication scheme (var_{NAEP})

Sample size (1)	MSE of estimated mean (2)	Expected variance (3)	Bias of estimated variance (4)	True variance of estimated variance (5)	cv of estimated variance (6)	MSE of estimated variance (7)	*Coverage rate of 95% confidence interval (8)	**Coverage rate of 95% confidence interval (9)
2	152.32	138.03	-13.96	93,736.92	2.22	93,931.74	0.7552	0.9484
4	52.07	71.05	18.98	6,171.13	1.11	6,531.47	0.9122	0.9813
6	26.82	46.42	19.61	1,912.25	0.94	2,296.77	0.9971	1.0000
8	30.97	32.26	1.29	500.53	0.69	502.20	0.9501	0.9538
10	27.25	24.96	-2.27	292.23	0.68	297.37	0.9053	0.9168
12	14.11	23.02	8.91	311.92	0.77	391.27	0.9492	0.9813
14	15.87	18.76	2.90	173.95	0.70	182.34	0.9802	0.9905
16	18.39	16.36	-2.03	57.86	0.46	61.98	0.9099	0.9099
18	10.95	13.27	2.33	44.39	0.50	49.82	0.9570	0.9739
20	7.19	12.92	5.74	31.16	0.43	64.07	0.9656	0.9656
22	4.46	10.40	5.94	23.53	0.47	58.77	1.0000	1.0000
24	8.34	10.95	2.61	40.55	0.58	47.39	1.0000	1.0000
26	9.28	9.77	0.49	13.96	0.38	14.19	0.8657	0.8657
28	3.25	9.12	5.86	12.26	0.38	46.63	1.0000	1.0000
30	2.69	8.56	5.87	8.47	0.34	42.97	1.0000	1.0000

* based on *t* coefficient of 1.96

** based on *t* coefficient from appropriate *t* distribution

Table 2. Performance of jackknife variance estimator using paired replication scheme (var_{PAIR})

Sample size (1)	MSE of estimated mean (2)	Expected variance (3)	Bias of estimated variance (4)	True variance of estimated variance (5)	cv of estimated variance (6)	MSE of estimated variance (7)	*Coverage rate of 95% confidence interval (8)	**Coverage rate of 95% confidence interval (9)
2	152.32	138.03	-13.96	93,736.92	2.22	93,931.74	0.7552	0.9484
4	52.07	74.02	21.95	12,928.24	1.54	13,410.12	0.8439	0.9656
6	26.82	47.18	20.37	2,969.01	1.15	3,383.89	0.9532	0.9991
8	30.97	30.94	-0.03	633.66	0.81	633.66	0.8809	0.9363
10	27.25	21.84	-5.38	424.14	0.94	453.11	0.7131	0.8608
12	14.11	19.41	5.30	383.95	1.01	412.01	0.9128	0.9492
14	15.87	18.47	2.60	169.33	0.70	176.11	0.9811	0.9894
16	18.39	15.65	-2.74	93.16	0.62	100.68	0.9099	0.9099
18	10.95	11.55	0.61	50.32	0.61	50.69	0.9570	0.9739
20	7.19	12.70	5.52	38.07	0.49	68.53	0.9527	0.9656
22	4.46	11.16	6.70	30.76	0.50	75.61	1.0000	1.0000
24	8.34	10.28	1.94	54.03	0.71	57.81	0.9615	0.9615
26	9.28	9.11	-0.17	13.65	0.41	13.68	0.8617	0.8657
28	3.25	7.19	3.93	5.98	0.34	21.46	0.9914	1.0000
30	2.69	7.62	4.94	3.97	0.26	28.36	1.0000	1.0000

* based on *t* coefficient of 1.96

** based on *t* coefficient from appropriate *t* distribution