COMPARISON OF VARIANCE ESTIMATION METHODS FOR COMPLEX SAMPLE DESIGNS
UNDER EXTREME CONDITIONS
Edward L. Spitznagel, Washington University

ABSTRACT
Modern survey sampling designs frequently involve complex strategies far removed from simple random sampling. Three popular methods for approximating the variances of estimators — Taylor series linearization, balanced repeated replications, and jackknife repeated replications — are currently in common use. All three methods have been studied through simulation and found to give approximately the same results. However, conditions involving weighting and cluster sizes have been fairly tame in comparison with extremes sometimes encountered in practice. This paper examines these methods under extreme conditions, for the purpose of estimating rates that themselves are extreme, in the one to five percent range.

The results, generally, are that all three methods remain nearly identical in terms of bias and standard error, under all single extremes and combinations thereof used to challenge them. They are found, however, to be inferior in terms of standard error to the jackknife.

INTRODUCTION
The history behind this paper lies in a large survey of mental health conditions known as the Epidemiological Catchment Area program. A description of this program is given in the acknowledgments section of the paper. Sites from five different regions of the country were involved in data collection. Although all five sites used distinct sampling plans, all plans are the typical multi-stage clustered designs found in most survey work: Groups of adjacent houses, called primary sampling units (PSUs), were selected at random, then one adult respondent was chosen at random from each household. Using groups of adjacent houses dramatically reduces interviewer travel, but it tends to lower the efficiency of estimates because of intraclass correlation. Weighting to compensate for sampling one individual per household and also for certain deliberate oversampling built into the design can be expected to further reduce the efficiency of estimates. Finally, the estimates of greatest interest are rates of specific disorders, such as schizophrenia and major depression, which tend to be low and therefore difficult to estimate efficiently under any conditions.

For initial results, the SAS procedure SURVEYREG [1], which uses the Taylor series linearization method to account for clustering and weighting, was used to estimate variances. This procedure also furnishes estimates of the design effects (DEFF's), computed as the ratio of the variance estimate to the binomial variance estimate for the same rate and size of sample. The design effects exhibited rather large and erratic behavior, even within sites, and it was this behavior that led to the investigations reported here.

Specific questions to be asked and answered in this paper are: 1) Is the Taylor series method less reliable than jackknife repeated replication or balanced repeated replication? 2) Which of several extreme conditions are associated with lowered performance of variance estimation techniques? 3) Are there variance estimation techniques that are better than Taylor series, jackknife repeated replication, and balanced repeated replication?

METHODS
Five different variance estimation methods were simulated: Taylor series linearization, jackknife repeated replication, balanced repeated replication, the ordinary jackknife (applied to the clusters), and Efron's bootstrap with 128 replications.

The sampling situation simulated is similar to an artificial design that all five sites are capable of generating: 120 clusters organized into 60 strata of 2 units each. The sample size was taken to be either 5000 or to be variable, averaging to 3000, which is close to the number in the base household sample at each site.

For the initial investigations, SAS was used to generate the data, and SAS MATRIX procedure code was used to analyze it. In this mode, it was easy to experiment with changes, and up to 100 trials could be run in reasonable time — sufficient to delineate trends, but not enough for precise comparisons. For final results, data was generated using IMSL random number generators, and FORTRAN code was used to analyze it. Unless otherwise stated, results given in this paper are based on 1000 of these trials per set of conditions.
The "extreme" conditions simulated are described below. These are fairly representative of conditions at the St. Louis ECA site, but they might or might not be representative of other sites.

Rates of occurrence simulated were 5% and 1%. Most rates of psychiatric disorder in the general population fall within this range. For example, six-month prevalence for alcohol abuse/dependence is approximately 5%, for major depression is approximately 3%, and schizophrenia is approximately 1% [2]. Positive cases were generated by cutting a standard normal deviate at 1.645 (for 5%) and at 2.326 (for 1%). This threshold model allowed the addition of another normal deviate as a cluster effect.

Varying cluster sizes were simulated to range uniformly from 2 to 48, for an expected mean size of 25 and expected total sample size of 3000. (Actual range at the St. Louis site was 2 to 76, with a slightly longer tail to the right.) The effect of these varying cluster sizes was then compared with the effect of cluster sizes fixed equal to 25.

Cluster effects were simulated by adding to the base normal deviate a cluster-associated normal deviate of mean zero and standard deviation 0.25. This had the effect of substantially shifting prevalence from one cluster to another. For example, then, in the 5% prevalence case, roughly 1/4 of all observations had prevalence shifted up to 7% or higher, and another 1/4 had prevalence shifted down to 3% or lower. (Cutpoints themselves were shifted to maintain net prevalences approximately 5% and 1%.)

Weights were simulated in the range from 9 to 600, with a long tail to the right, to match the pattern in the St. Louis site. The actual extremes in the St. Louis data range from a low value of 3.4 to a high of 650. (The weights at each site represent the combined effect of household size, compensation for over- or under-sampling by design, and a post-stratification weight. The weights were deliberately chosen to sum to the size of the population sampled, rather than back to the sample size, to warn users of the data that it is important to use appropriate methods for estimating variances.)

It should be pointed out that this paper is concerned solely with the behavior of variance estimates for rates. This is a "first things first" report, and it is hoped at some future time to report on the behavior of the various methods in more sophisticated work involving rates, such as logistic regression.

RESULTS

In the following discussion, the term variance refers to the variance of the rate estimate. The term standard error refers in turn to the standard deviation of the estimate of this variance. (In all cases, the rate estimate is simply a weighted mean; what we are interested in is estimating the variability (variance) of this weighted mean.) The abbreviations JRR and BRR will be used for the jackknife repeated replication and balanced repeated replication methods, respectively. The term jackknife will be used to refer to the ordinary jackknife applied to the 120 clusters.

The first and most important result is that over all extremes tested, the Taylor series, JRE, and BRR yielded practically identical estimates of variance, and hence are comparable in terms of both bias and standard deviation. (This equivalence is known to break down for statistics more complex than rates, such as correlation coefficients [3, pp 322-324].) Therefore, in the results given below, only one of the three, JRR, will be represented, the values for the Taylor series and BRR methods being virtually identical.

Simple random sampling: the base for comparison. When there are no differences in cluster sizes, no cluster effects, and all weights are equal, the design reduces to simple random sampling. We use this case as a yardstick against which to measure other designs. Figure 1 compares the means of the three variance estimates and their standard errors with the mean and standard error of the estimate based on the binomial model. As can be seen, there is practically no bias, and so the important comparison is in terms of standard error. In terms of standard error, JRE and the bootstrap were nearly the same, with the ordinary jackknife somewhat more stable.

All three methods have standard errors substantially larger than the estimate of variance based on the binomial model. This leads to a dilemma: If we use variance estimation techniques based on clusters because we suspect the existence of a small cluster effect, we wind up paying a large price in terms of the stability of our variance estimates. This is the same flavor as the age-old problem of whether it is better to seek minimum variance unbiased estimators or to seek minimum mean-square error estimators that may well possess small amounts of bias.

The effect of varying cluster sizes. As Figure 2 shows, the variance estimates remain unbiased in the face of wildly differing cluster sizes, while standard errors of the estimates increase slightly. Only the case of 5% prevalence is shown, the 1% case looking much the same.
Figure 1a. Equal cluster sizes, no cluster effects, and equal weights ("simple random sampling"), prevalence 5%.

Figure 1b. Equal cluster sizes, no cluster effects, and equal weights ("simple random sampling"), prevalence 1%.

Figure 2. Unequal cluster sizes, varying from 2 to 48, prevalence 5%.

Figure 3. Cluster effects ("intracluster correlation"), prevalence 5%.
The effect of varying cluster rates. As Figure 3 shows, the mean values of the variance estimates increased due to the intracluster correlation, but the estimates remained unbiased, as we would hope. There is once again a modest increase in standard errors as compared to Figure 1. Only the case of 5% prevalence (actually 4.5%, due to an error in setting the cutpoint) is shown.

The effect of varying weights. Figure 4 shows that rather large differences have resulted from varying weights. First, a modest amount of bias is now evident in the case of 5% prevalence; all three methods have over-estimated the variance by about 10%. Second, in the case of 1% prevalence, standard errors have become large relative to the sizes of the estimates. This means that the estimates of variance are much less stable, and one can expect computed design effects to vary widely from this cause alone.

All extremes combined. When all extremes were applied together, no synergistic effects were found. As one might expect from the condition of varying weights being the single major influence, the results were practically identical to those displayed in Figure 4, and are therefore not displayed separately.

As far as the St. Louis data set is concerned, the message is clear. The highly variable weights are the overwhelming source of possible instabilities in variance estimates. They have the same effect as outlier values in regression. One of the causes of the few unusually high weights was a change in sampling design midway through the survey. The design had called for 2:1 oversample of black respondents. A check on yield halfway in the survey indicated that this goal was not being met, primarily because some regions being sampled that had been believed to be predominately black were found to be racially balanced.

The sampling plan was revised so that in the second half of the study the sampling rate for these areas was drastically reduced, and since weights are determined as reciprocals of sampling rates, the weights for certain individuals skyrocketed. It would appear that after certain temporal elements are checked out, weights could be redistributed among respondents of these areas to make variation in weights more reasonable.

An interesting by-product of this study is the finding that the ordinary jackknife tends to yield more stable estimates of
variance than the three specialized survey sampling methods do. The explanation is twofold: First, no simulations of stratification have been reported here, so the specialized methods have lost the advantage they have under conditions of stratification. Second, a method like JRR, which trades units within strata, does not make as many independent comparisons as does the ordinary jackknife, which ignores stratification. While one should be duly cautious about ignoring stratification in a survey sample, it should be possible to define relatively few strata each containing many units and to use a tool like the jackknife within the strata.

Although the bootstrap did not perform as well as the jackknife in this study, one should not completely dismiss this wonderful tool as a possible aid in analyzing survey data. The bootstrap replications were limited to 128 in this study because of constraints on computer time. However, one small test was run with 512 bootstrap replications, in which it was found that the bootstrap estimates decreased in standard error but still had somewhat larger standard error than the jackknife estimates. That is, there is more information out there to be harvested via the bootstrap, though the number of replications may become prohibitive.

Although the extreme of low rates was studied in this simulation, the extreme of low sample size was not. Other investigators have examined the effect of this condition on the Taylor series and bootstrap methods and found the bootstrap to perform better [4] in terms of bias.

In this study, Taylor series, JRR, and BRR were found to give nearly identical estimates and therefore to be equivalent in terms of both bias and standard error. However, this study has been concerned only with rate estimates. The methods can be expected to show more differences when applied to other estimation problems. They have, indeed, been found to behave differently in several other studies [3, pp 322-324]. For the estimation of subgroup rates, such as the rate of major depression in white males above age 65, logistic regression could be a much better tool than simply subsetting the data met down to the small number of respondents meeting the conditions. With this in mind, it would be very helpful to understand the behavior of the variance estimation methods when applied to logistic regression.

ACKNOWLEDGMENTS

This research was supported by the Epidemiological Catchment Area Program (ECA). The ECA is a series of five epidemiologic research studies performed by independent research teams in collaboration with staff of the Division of Biometry and Epidemiology (DBE) of the National Institute of Mental Health (NIMH). The NIMH Principal Collaborators are Darrel A. Regier, Ben Z. Locke, and William W. Eaton; the NIMH Project Officer is Carl A. Taube. The Principal Investigators and Co-Investigators from the five sites are: Yale University, U01 MH 34224--Jerome K. Myers, Myrna M. Weissman, and Gary Tischler; Johns Hopkins University, U01 MH 33870--Morton Kramer, Ernst Grunenberg, and Sam Shapiro; Washington University, St. Louis, U01 MH 33808--Lee B. Robins and John R. Reizer; Duke University, U01 MH 33846--Dan Blasser and Linda George; University of California, Los Angeles, U01 MH 35865--Richard Rough, Marvin Kamar, Javier Escobar, and M. Audrey Burnam. This work acknowledges support of this program.

REFERENCES


Address:
Edward L. Spitznagel
Department of Mathematics
Washington University
St. Louis, MO 63130