Discussion of: Variance Estimation for Complex Surveys

Keith Rust
Westat, 1600 Research Boulevard, Rockville, MD 20850

Key Words: Replicated variance estimation, jackknife, two-phase sampling, multiple imputation

1. Introduction

It is a pleasure to have participated in this session, and to have the opportunity to review these four interesting and insightful papers. I will be begin by making some general remarks about the topic, and then discuss the individual papers.

2. Why is Research on Variance Estimation Still Needed and Ongoing?

It seems that there is a session or more on the topic of variance estimation for complex surveys at every Joint Statistical Meetings. It is worth speculating as to why this might be, given that this topic has been well-researched over about the past 50 years. I think we can identify at least four reasons for continued interest in this topic.

2.1 Advances in Imputation for Missing Data

Methods of imputation for missing survey data continue to be a very active area of research. Each method of imputation has its own implications for variance estimation, and practical methods of variance estimation, to be implemented on a routine basis, have not yet been fully developed. Indeed this is the topic of paper in this session by Baskin and Thompson, discussed below.

2.2 Advances in Computing

Continued advances in computing power have affected variance estimation in two ways. First, more complex estimators of model parameters and other statistics of interest are feasible, using iterative procedures with large amounts of data and many iterations. Methods are needed to estimate the sampling variance of the resulting estimators. Second, more intensive methods of variance estimation are feasible. Thus routine application of bootstrap procedures, with hundreds or thousands of replicates, is not possible.

2.3 Interactive Online Analysis Tools

Many statistical agencies and other organizations are providing analytic tools on the internet which allow users to generate their own estimates. These must be accompanied by appropriate methods for calculating sampling errors, which must be generated in real time. Thus robust procedures must be implemented that will be effective for any analysis that a user might choose.

2.4 Increased Standards for Reporting Sampling Error

It seems that there is continues to be steady progression in the extent to which statistical and administrative agencies and scientific journals stipulate and enforce standards regarding the reporting of sampling errors for statistical analyses in general, and the analysis of survey data in particular. This in turn applies pressure on the field to develop appropriate and practical methods for estimating sampling errors for the analysis of survey data, in a wide variety of applications.

The four papers in this session, between them, provide a fairly comprehensive cross-section of the areas where research into the generation of sampling errors from survey data is currently active – for model-assisted survey estimation, for imputation, and for two-phase sampling. The other area that seems to be most active is for small area estimation – so look for at least one paper on that topic at next year's meetings!

3. Shao and Wang: Variances for Model-Assisted Regression Estimators in Stratified Surveys

Shao and Wang discuss the asymptotic bias and variance properties of a variety of regression estimators, in the context of stratified sampling. Studying the asymptotic properties of estimators under stratified sampling always presents a difficulty, since the asymptotic setting is rather ambiguous. Should sample and population sizes grow within a fixed number of strata, or should the number of strata grow and the sample and population sizes remain bounded within each stratum? Or is there some sensible combination of these?

The authors consider both of these asymptotic scenarios separately. This leads to the conclusion that different regression estimators are superior under different asymptotic conditions. When the sample size increases within each of a fixed number of strata, it is clear that stratum-by-stratum regression estimators are to be preferred. In hindsight at least this is intuitively fairly obvious, as the inefficiency of stratum-by-stratum regression decreases to nothing as the sample size within each stratum increases, whereas the effects of model misspecification on cross-stratum regression estimators does not decrease with increased sample size. However, when stratum sample sizes are small and the number of strata increases, the best choice among the regression estimators considered depends upon the how much the true regression models vary across strata: if the models are similar across strata, then cross-stratum regression estimation is to be preferred, but as the regression models vary more widely across strata, at some point the effectiveness of cross-stratum regression estimators breaks down and stratum-by-stratum estimators are again preferable.

This would seem to leave the practitioner in a quandary as to which regression estimator to choose, but in fact suggests that a compromise approach should be used. The sample design should be stratified as deeply as is reasonably possible, so as to take advantage of the variance reduction gains of stratification. Then when using regression estimators the best approach is likely to be to group strata together into major strata, and then use a separate regression for each major stratum, common across strata within the major stratum.

If this enlightening area of research, as the most effective method of regression estimation, is to be pursued by the authors or others, I suggest two extensions. The first is

consider the role of ratio estimation – that is, regression estimation with an intercept of zero – since it is known that if the true regression model contains no intercept then ratio estimation can be more efficient than regression estimation. The second, related, idea is to consider heteroscedastic regression models for the population, as initially discussed by Brewer (1963).

Turning to the question of variance estimation, perhaps not surprisingly given the results for the regression estimators themselves, Shao and Wang find that different variance estimators are suggested by different asymptotic assumptions. This makes it difficult to decide which approach is most suitable in any given application. The authors conducted a simulation study. They investigated the performance of the variance estimators suggested when the sample size is large within each stratum using samples of 200 per stratum, and investigated methods that the asymptotic results suggested would be good for small sample sizes per stratum using a design of 600 strata with four units selected per stratum. The simulation confirmed the findings of the asymptotic results as to the desirable properties of each of these variance estimators in the context where its use was suggested by the asymptotic results. However, it would be very useful to extend this in the following ways. First, we would like to see how each of the variance estimators performed when applied under conditions where its use was not suggested by the asymptotic results. That is, how robust are the methods to the choice of asymptotic assumptions. Second, it would be useful to compare the methods under a design that is between these two extremes and likely to be of the kind encountered in practice; for example, a design with 100 strata and 20 units selected per stratum.

I found this paper very interesting and thought provoking, and Dr. Wang's presentation of it was extremely clear. I congratulate the authors on a very informative contribution to the areas of regression estimation and associated variance estimation for sample survey data.

4. Baskin and Thompson: Variances for Imputed Data

In their paper Baskin and Thompson follow up on previous research, where they observed unexpected behavior of the Rao-Shao variance estimator for imputed data, when used with proportions. This was in contrast to the case of continuous data, where the results were as expected.

The application is to the Medical Expenditure Panel Survey (MEPS), which uses a 'nearest neighbor hot-deck' method of imputation. In determining the properties of an imputation variance estimator, it is important to be clear what exactly the method of imputation is. The term 'hot deck' implies a stochastic imputation procedure, whereas as 'nearest neighbor' implies a deterministic method. It seems that the method used for MEPS is a blend of these. This is important, as the Rao-Shao method is appropriate for hot deck imputation. Rao and Shao (1992) showed that their method is consistent when hot deck imputation is used, whereas the standard jackknife underestimates the variance, while the Burns (1990) method overestimates. However, Chen and Shao (2001) showed that the Rao-Shao method has an upwards bias when applied to nearest neighbor imputation. With nearest neighbor imputation, the Rao-Shao method is essentially the same as the Burns method.

Chen and Shao (2001) proposed a modified Rao-Shao jackknife variance estimator for use with nearest neighbor imputation. Using y to denote the original data, z to denote the

transformed data, * to indicate imputed values, *i* to denote replicate, *l* to denote stratum, and *v* to denote imputation class:

$$z_{lij}^{*} = \begin{cases} y_{lj} & if\ y_{lj}\ not\ imputed \\ y_{lj}^{*} + g_i^{(vl)}(\bar{y}_{liv} - \bar{y}_{lv}) & if\ y_{lj}\ imputed \end{cases}$$

$where\ 0 < g_i^{(vl)} < 1$ is a function of the distances among neighbors in class *v*. If $g = 1$ in all cases this reduces to the standard Rao-Shao jackknife. While this approach is presumably not directly applicable in the MEPS case, since it does not use standard hot deck imputation, this does suggest an analytical approach that might lead to a suitable variance estimator in the MEPS case.

The question remains as why the unadjusted Rao-Shao method works well empirically in the MEPS application for continuous variables, but not for discrete ones. Presumably the answer lies in the nature of the particular imputation approach used in MEPS, or else the distances among neighbors differs for these two types of variables in the MEPS case (so that the g factor in the above expression tends to be close to 1 for continuous variables, but substantially less than one for at least some discrete variables.

I encourage the authors to continue their research. In many applications the effect of imputations on inference is ignored, and it is commendable that MEPS is endeavoring to ensure that the effect of imputation variance is included in the measures of uncertainty provided.

5. White and Opsomer: Variances for Two-Phase Sampling

As in the case of Baskin and Thompson's paper, the work of White and Opsomer is motivated by a real application – in this case variance estimation for the National Survey of College Graduates (NSCG), which has a two-phase design. The literature on replicated variance estimation for two-phase designs is restricted to cases with straightforward second phase sampling procedures; either simple random sampling or Poisson sampling. The second phase design for the NSCG is much more complicated, and in particular, systematic sampling is used, which tends to have very different properties than either simple random sampling or Poisson sampling.

White and Opsomer show empirically that when second phase frame variables are considered, using the reweighted expansion estimator, the resulting replicate variance estimators have a large relative bias for those phase two frame variables that are used to create the sort order.

To address this, the authors attempt an ingenious approach. The two-phase jackknife approach is able to reflect the complexity of *estimation* well, but is not able to reflect the complexity of the second phase *design*. Thus White and Opsomer modified the estimation procedure to poststratify the estimates using the sort variables. In their empirical study, this method proved highly effective in reducing the relative bias of replicate variance estimators for second-phase frame variables.

This raises several interesting questions, and I hope that the authors will pursue at least some of them:

1) How much did the poststratification actually reduce sampling variance, in addition to reducing the bias of variance estimation? That is, to what extent did this modification to the estimation actually improve the estimates themselves, in addition to improving the variance estimates?

2) The research was conducted using the Phase 2 frame variables? What happens to the survey variables themselves? It would be very informative to conduct a simulation to see what happens when the method is applied to analysis variables that are correlated with the frame variables in various ways, rather than just considering the frame variables themselves?

3) Does this technique of augmenting the estimation procedure so as to reduce the marginal impact of the design features on the sampling variance, thus resulting in variance estimators with low bias, suggest a more general approach that might be used in other cases where complex designs lead to difficulties in estimating variances well. Can one utilize an approach of having redundancy across design and estimation (such as using systematic selection and poststratification using the same auxiliary data), so that the variance estimator reflects the estimation gains, even if it cannot directly reflect the design gains? This idea is somewhat analogous to the advice sometimes given to analysts of survey data: Include the design variables in your analysis model and then you do not have to explicitly account for the design features when making inferences.

6. Reist and Larsen: Calibration and Multiple Imputation

Multiple Imputation (MI) is a well-known and powerful technique for reflecting the variance due to imputation, and it is increasingly commonly applied in survey application. However, it is well-documented that MI is biased, and possibly negatively, when the imputation process is not 'proper'. There are two main sources of improper multiple imputation in survey applications. The first is if analysis variables are omitted from consideration in the imputation model. The other is when the survey weights are informative but are not used in the imputation model.

Kim *et al* (2006) showed that the bias in variance estimation due to ignoring survey weights during the MI process is a function of the true imputation variance: if there is little imputation variance, there is little bias in estimating it. Reist and Larsen attempt to take advantage of this relationship. They employ calibration as a means to reduce the imputation variance, with the aim of reducing the ensuing bias in MI. The simulation set-up employed by the authors is very thorough, and provides a good model for others planning to do research into imputation variance estimation.

The approach used here does raise one question for analysis: If the calibration is effective in reducing the imputation variance to a sufficiently low level that the bias of multiple imputation is modest, then why not just ignore the effect of imputation variance and dispense with multiple imputation in this case?

The approach used here has interesting parallels with the approach used by White and Opsomer in their approach to variance estimation for a complex two-phase sample. In each case, the investigators have investigated the approach of modifying the parameter estimator, utilizing auxiliary information, as a means of handling difficulties in estimating variances. However, in this case it is clear that, if the approach is successful in

reducing of variance estimation via multiple imputation, it is also reducing the variance of the parameter estimators, since to be effective it must reduce the imputation variance itself.

One wonders though if this use of calibration to reduce the imputation variance might involve a trade-off. Would sampling variance, and thus perhaps total variance, be reduced more by calibrating using variables known to be correlated with the survey variables of interest, but which are not design variables?

With regard specifically to the results of the empirical research that Reist and Larsen conducted, it seems that it would be useful to carry out similar analyses, but using *x* and *y* variables that are not as highly correlated as those in the research presented. Nevertheless, the results of their work look promising.

7. Conclusion

Variance estimation for complex surveys continues to be an active area of research, with real issues that need to be addressed. As is evidenced by these four papers, current ongoing research is making progress towards solutions. It seems that we can look forward to presentations on these topics at the Joint Statistical Meetings for the foreseeable future. Once again I think the presenters and their co-authors for a very stimulating session.

References

Baskin, R.M., and Thompson, M.S. (2012). Anomolies Under Jackknife Variance Estimation Incorporating Rao-Shao Adjustment in the Medical Expenditure Panel Survey – Insurance Component. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Brewer, K.R.W. (1963). *Ratio Estimation and Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process*. Australian Journal of Statistics, 5(3), pp.93-105.

Burns, R.M. (1990). Multiple and Replicate Item Imputation in a Complex Sample Survey. *Proceedings of the Sixth Annual Research Conference*, U.S. Bureau of the Census, pp.655-665.

Chen, J., and Shao, J. (2001). Jackknife Variance Estimation for Nearest-Neighbor Imputation. *Journal of the American Statistical Association*, 96, No. 453, pp.260-269.

Kim, J.K., Brick, J.M., Fuller, W.A., and Kalton, G. (2006). On the bias of the Multiple-Imputation Variance Estimator in survey sampling. *Journal of Royal Statistical Society*, B 68, pp.509-521.

Rao, J.N.K., and Shao, J. (1992). Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation. *Biometrika*, 79(4), pp.811-822.

Reist, B.M., and Larsen, M.D. (2012). Post-Imputation Calibration Under Rubin's Multiple Imputation Variance Estimator. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Shao, J., and Wang, S. (2012). Asymptotic Variance Estimation and Comparison of Model-Assisted Regression Estimators in Sample Surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

White, M., and Opsomer, J.D. (2012). Replicate Variance Estimation in a Two-Phase Sample Design Setting – Simulation Study with 2010 National Survey of College Graduates Data. *Proceedings of the Survey Research Methods Section*, American Statistical Association.