VARIANCE ESTIMATES COMPARISON BY STATISTICAL SOFTWARE

Stanley S. Weng and Fan Zhang, Synectics for Management Decisions[1]
Michael P. Cohen, National Center for Education Statistics[1]
Stanley S. Weng, Synectics for Management Decisions, Inc., Arlington, VA 22201

Key Words: Complex survey, Variance estimation, Balanced repeated replication, Jackknife, Taylor linearization

This article reports a comparison analysis which, involving six statistical software routines in wide use for variance estimation for complex surveys, examined the variance estimates produced by those routines in a sample data setting from an NCES (National Center for Education Statistics) complex survey. This study helps identify reliable and capable statistical software for variance estimation, and, perhaps more meaningfully, is an effort to raise the standard of practice in the analysis of complex survey data.

1. Introduction

Most of the surveys of the NCES are large complex surveys. As well known, the sampling and weighting processes of complex surveys have much changed the methodology and algorithms of variance estimation.

Conventional statistical software packages such as SAS and SPSS can be only used to provide variance estimates for simple random samples. Naive use of such software for variance estimation on complex survey data, as often made in practice, may lead to underestimating the variances.

There are three methods widely used for variance estimation for complex surveys: the *balanced repeated replication* (BRR) method, the *jackknife* (JK) method, and the *Taylor series* method (Wolter, 1985). The first two methods are under the *replication* approach, and the third one under the *linearization* approach. A number of statistical software have been developed to perform these methods.

The BRR method has been used to estimate the sampling errors associated with estimates for all of the 1990-91 Schools and Staffing Survey (SASS) samples. In the BRR method, within each stratum, sampled schools are paired by the order they were selected. One school from each pair is placed into each replicate. Each replicate includes approximately half the total sample. The choice of when to place a school from a pair into a replicate is done in a balanced manner to reduce the variability of the variance estimates. See Kaufman and Huang (1993) for more detailed information on how SASS units are placed into balanced half-sample replicates. SASS uses 48 replicates for variance estimation, giving a reasonable degrees-of-freedom cushion for the validity of the z-test approximation when making inference. Each SASS public use file includes a set of 48 weighted replicates for BRR variance estimation.

The jackknife method has been used by the 1990 National Assessment of Educational Progress (NAEP) to estimate all sampling errors as presented in the various reports and provided good quality estimates of sampling variance for most statistics. A set of 56 jackknife replicate weights for students was developed, for the purpose, in the manner that models the design as one in which two PSUs were drawn with replacement per stratum (Johnson and Allen, 1992).

The Taylor series method has been used by the National Education Longitudinal Study of 1988 (NELS 88) and follow-ups to calculate standard errors as presented in various reports (Spencer et al., 1990, and Ingels et al., 1994).

However, NCES recently reported several occurrences where unexpected differences in variance estimates were produced by different statistical software routines. This resulted in a concern: if reliable results could be expected from available variance estimation routines, including their estimating approach, portability, and capabilities to accommodate the features of various complex surveys. This study was conducted to address the computational as well as methodological issue: whether different statistical programs, using different estimation methods and with different designs, produce significantly different results for complex surveys. The study compared the variance estimates, produced by six statistical software routines in wide use, from descriptive as well as regression analysis using the same data from an NCES complex survey.

We will present the analysis and results in Section 2, and make some discussions in Section 3 to serve the purpose of this study. The remainder of this section is a brief description of the six software routines selected for this study. They are:

SUDAAN (Shah, et al., 1992). Uses Taylor series approximations in conjunction with textbook-type variance formulas to calculate variance estimates.

PC CARP (Fuller, et al., 1986). Uses Taylor series method. It uses a general framework of linearization for the calculation of variance estimates, which could cover most sampling designs used in practice.

VPLX (Fay, 1995). Performs replication methods (BRR, JK, etc.). VPLX can create the jackknife replicate weights in general algorithm, and has the full

computational ability to handle hundreds of PSUs within a stratum without the need of grouping.

WESVAR (Westat, 1993a) and WESREG (Westat, 1993b). WESVAR handles basic survey estimates. WESREG handles regression analysis. Both programs perform either BRR or JK. The jackknife procedure of WESVAR and WESREG assumes a two-per-stratum sampling design.

REPTAB(Liebman, 1993). A SAS procedure, uses replication methods (BRR and JK).

STRATTAB (Ogden and Liebman, 1991). A SAS procedure using a Taylor series approximation.

2. Analysis and Results

2.1 Data

Data from the Teacher Survey of the 1990-91 School and Staffing Survey (SASS), as recommended by NCES statisticians, were used to apply the software routines for the variance estimates comparison. Below is a brief description of the SASS Teacher Survey.

The SASS Teacher Survey is a two-stage stratified probability sample. The school survey is the first stage of the sampling. Schools are selected within strata by a probability proportional to the number of teachers within the school. Within the first-stage school sample, a second-stage teacher sample is selected stratified by teacher experience level.

The SASS Teacher Survey sample design is very close to the standard two-stage sampling design, as the one adopted in the design by all statistical software for variance estimation for complex surveys: the stratified probability sampling with replacement at the first stage and simple random sampling at the second stage. For the analysis purpose of this study, because of the small sampling rates of schools within strata, it should not cause concern to treat the sample as from with-replacement sampling at the first stage. Stratum and PSU variables, as required for performing Taylor series and jackknife procedures, are well included in the data files, and the BRR replicate weights for teachers are also available in the files.

2.2 Analyses

Our analyses used the public school sample in the Teacher Survey. Variance estimates were produced for basic survey statistics, including means, percents/ proportions, and ratios, as well as regression coefficients, using the six selected software routines. Two analyses with different sets of variables were conducted for each kind of statistics (see Table 1, the first three columns).

Here is a list of some questions with abbreviated wording in the column Variables of Table 1:

Percent:

Master's degree--Do you have a master's degree?

Look forward to day--I usually look forward to each working day at this school.

Mean:

Salary--What is your academic base year salary for teaching in this school?

Ratio:

Schl hrs extra--School-related activities involving student interaction

Othr hrs extra--Other school-related activities

Regression (first):

Independent--Have you ever taken any undergraduate or graduate courses in the following subjects.

Before entering analysis, the data were necessarily shaped. For instance, the strata which contained only one PSU were appropriately collapsed. Missing values were also handled appropriately according to the design of the software routine applied. For those routines which do not have the capability of handling missing values, missing variables were handled in an external data step.

There are two versions of the jackknife procedure used in variance estimation for survey data: the *simple jackknife* (JK1) and the *stratified jackknife* (JK2). The simple jackknife is the basic algorithm of the jackknife procedure. The stratified jackknife is a generalization of the simple jackknife to stratified samples. For multi-stage stratified sample variance estimation, the simple jackknife is considered generally not able or not suitable to perform. Only the stratified jackknife was performed in this study.

The Taylor series procedure, as understood, is performed in conjunction with the selected sampling design. For SUDAAN, a number of standard designs as options are available. By the sampling design of SASS, the appropriate design option would be "without replacement (WOR)". However, under this design option, SUDAAN requires data on the number of PSUs for each stratum in the population. The variable, NUMSCH, in 1990-91 SASS public school file for this purpose, had some problems in its data. For instance, all PSUs in the same stratum should have the same values of NUMSCH, but this is not always the case. Therefore, our analysis used the design option "with replacement (WR)" which appeared suitable to the survey design and the data. In using PC CARP, the sampling design is identified by three components: the design variables entered into the analysis, the "Two-Stage" option, and the optional data of the sampling rates of strata. Since there were no sampling rates data available, our analysis also used the "with replacement" sampling design for PC CARP. STRATTAB was designed only for the standard sampling scheme assuming sampling with replacement at the first stage. Thus, the same design option was used for the three routines to perform the Taylor series procedure.

Some features of the software were noticed with the

conduction of the analyses.

(a) For WESVAR, if a given variable has a missing value for an observation, that observation is not used in the calculation of the total for that variable only. Effectively, this treats the missing value as zero in all computations. However, even to estimate the mean for a missing variable, this way of handling missing values will yield incorrect results. In fact, the WESVAR mean is treated as a ratio. Thus, the numerator will be calculated using only non-missing values, while the denominator will sum up to the weights for all observations. The same problem will occur when calculating ratio estimates. When the two variables involved have missing values in different cases, the resulting ratio estimate will be misleading. To avoid the problem, we handled the missing data in a SAS data step, prior to using WESVAR.

WESREG was supposed, as a regression procedure, to handle missing values in the usual way, as also stated in its manual: "Observations having missing values for the dependent variable or any of the independent variables are excluded from all estimates." However, our analysis showed that it is not the case with WESREG. There is no further information available to clarify how WESREG handles missing values. In our analysis, we then handled missing values in a SAS data step prior to using WESREG.

(b) The version of WESVAR used in our study does not have the ability to perform jackknifing from the stratum and PSU variables in the data. Moreover, the jackknife procedure in WESVAR is in a simplified form. As documented in the manual (Westat, 1993), the jackknife procedure is formulated only to the special case that there are two PSUs in each stratum. WESVAR cannot handle more than two PSUs in a stratum. When there are more than two PSUs in a stratum, even if the jackknife replicate weights are supplied, a simple use of WESVAR will give wrong results. In such a situation a grouping procedure must be conducted to make two (pseudo) PSUs in each stratum in order to meet WESVAR's jackknife frame.

Remark The new WesVarPC (Westat, 1995) can create the jackknife replicate weights from the design variables, however, still in the two-per-stratum form, remaining from the design of WESVAR.

2.3 Variance estimates

The estimates of the statistics and associated standard errors are presented in Table 1. The different variance estimation procedures were not involved in the estimation of the survey statistics. All the software routines produced identical estimates for all the statistics in the analysis. In Table 1, one column is used to present the estimates of the statistics, and the body of the table is for the variance estimates (standard errors) presented by the estimation method and the software used. For the analysis not available due to capability limitation of the software, an N/A indicator is put in the table. In the following, we examine the variance estimation results, mainly under same estimation method and also across the methods.

(1) BRR variance estimates

The three statistical software packages, VPLX, WESVAR, and REPTAB, provide BRR variance estimates for descriptive survey statistics. As generally designed for software performing BRR, replicate weights need to be supplied with the input data for all the three programs. With replicate weights supplied, only simple and standard calculations need to be conducted to obtain the BRR variance estimates. As Table 1 shows, for all the descriptive statistics (means, percents, and ratios), the three routines produced identical BRR variance estimates.

Among the six software packages, WESREG is the only one providing BRR variance estimates for regression coefficients. No comparison could be made for the BRR variance estimates for regression coefficients. However, some comparisons between the results by WESREG and by SUDAAN and PC CARP (both using Taylor series method) may be of interest, and are discussed later in this section.

(2) Jackknife variance estimates

The data set does not supply replicate weights for the jackknife procedure and thus WESVAR and REPTAB are not applicable. VPLX is the only software which provided jackknife variance estimates in this study. By using all PSUs in the jackknifing, VPLX reached great precision. The VPLX jackknife variance estimates appear the same, except for a minor difference for the mean of SALARY, as those produced by SUDAAN and PC CARP using the Taylor series method. This coincidence, as expected from the asymptotic property that the jackknife variance estimate tends to be close to the linearized variance estimate if both calculations employ the same PSUs and the statistic is smooth (Wolter, 1985), is an indication that VPLX has sound statistical design and is computationally reliable.

(3) Taylor series variance estimates

Three statistical software routines, SUDAAN, PC CARP, and STRATTAB, produced variance estimates using the Taylor series method.

Experience with large, complex sample surveys has shown that the Taylor linearization approximation often yields satisfactory results, except for highly skewed populations. Generally speaking, if the nonlinear estimator can be expressed as a smooth continuous function of population totals, the Taylor linear approximation would be valid (Wolter, 1985).

SUDAAN and PC CARP produced identical variance estimates for the descriptive survey statistics, means, percents, and ratios. For the first regression analysis, the variance estimates for the coefficients produced by the two programs appear slightly different. The differences may be due to different computational procedures handling complex functions of survey estimates. As for the second regression analysis which is simpler than the first one, the variance estimates by the two programs are similar.

The variance estimates produced by STRATTAB, only available for means and percents, appear to be of quite different magnitude (smaller) compared to those by SUDAAN and PC CARP. The differences could not be considered as within a reasonable range of error due to different computational procedures.

(4) Comparison across estimation methods

Though there seems no general comparison based on rigorous theoretical justification between the BRR, jackknife, and Taylor series methods for variance estimation - appraisal of their performance with different estimators and types of surveys has relied on empirical studies, an important property has been established that for nonlinear statistics that can be expressed as functions of estimated means of p variables - such as ratios, regression and correlation coefficients, the variance estimators from the linearization, the jackknife, and the BRR methods are asymptotically consistent (Krewski and Rao, 1981). This result is valid for any multistage design in which the primary sampling units (PSUs) are selected with replacement and in which independent subsamples are selected within those PSUs sampled more than once (Rao and Wu, 1988). The sample data used in our study can be considered under this situation, and are of large size. Meaningful information could be drawn.

For the descriptive survey statistics, the BRR variance estimates are very close to that produced by the Taylor series (using SUDAAN and PC CARP) and jackknife methods. And for most of them, the BRR variance estimate appears slightly lower.

For the first regression analysis, for six out of the eight regression coefficients, the BRR standard error (by WESREG) appears significantly different from, mostly higher than, the Taylor series estimate (by SUDAAN and PC CARP). The differences range from 14 percent to over 50 percent compared to the Taylor estimates. For the second regression, the BRR standard errors appear almost the same as the Taylor estimates. Since the first regression involves more regressors than the second one, the behavior of BRR method when performed to complex functions of survey estimates, such as regression coefficients, needs to be further explored. The comparison of jackknife estimates and Taylor series estimates was already made above between the results from VPLX and from SUDAAN and PC CARP.

3. Conclusion

For estimating variances, it would be expected that statistical software routines performing the same estimation method produce same results; while for routines using different methods it may not be expected that same results would be reached. Thus, identical variance estimates produced by two statistical routines performing different methods provide an indication of their reliable performance; while significant difference in the results produced by routines using the same method implies error existent with some of them.

This study thus helps identify reliable and capable statistical software for variance estimation for complex surveys. Reliable statistical software routines are available to perform all the three variance estimation methods.

This study may also be a motivation for further development of comprehensive statistical software for variance estimation of survey data.

To perform the BRR, the creation of the BRR replicate weights is an issue. All the statistical software routines for performing BRR, included in this study, require the replicate weights be supplied in the data input. This situation certainly limits the practice of calculating the BRR estimates. As already made available for general jackknifing, VPLX is going to make available a general algorithm for creating BRR replicate weights within the program. The new WesVarPC (Westat, 1995) has the capability of creating the BRR replicate weights. Such capability will expand the usability of the software and promote the use of the BRR method.

The progress of computing ability in recent years has been changing the consideration in designing statistical software for variance estimation for complex surveys. Computing cost seems no longer a big issue as it was years ago. The computing-intensive methods, like BRR and jackknife, can be performed in general versions as a usual matter. Unnecessary simplification in the estimation algorithm would merely limit the applicability of the software and reduce the power of the performance of the method.

Many NCES surveys use more complex sampling designs than the standard one as assumed for the BRR and the jackknife to apply. It seems necessary to make available the statistical software using more general algorithms for variance estimation, for example, the more general resampling procedure (Rao and Wu, 1988; and Kaufman, 1993a, 1993b, and 1995), and also the combination of linearization and replication methods, if the Taylor linearization does not bring the estimate to an appropriate form to which standard variance estimation formulas are applicable.

Table 1: Standard Errors Associated with Survey Estimates by Statistical Software

Analysis				Variance Estimation Method						
				BRR			JK2	Taylor series		
Data type	Survey statistics	Variables	Esti-mate	VPLX	WESVAR/ WESREG	REPTAB	VPLX	SUDAAN	PC CARP	STRAT TAB
Categorical	Percent(%)	Master's Degree								
		1: YES	46.980	.326	.326	.326	.393	.393	.393	.0259
		2: NO	53.020	.326	.326	.326	.393	.393	.393	.0259
		Look forward to day								
		1: ST AGREE	51.37	.341	.341	.341	.385	.385	.385	.019
		2: AGREE	40.39	.313	.313	.313	.363	.363	.363	.017
		3: DISAGREE	6.23	.163	.163	.163	.180	.180	.180	.014
		4: ST DISAGREE	2.01	.121	.121	.121	.107	.107	.107	.000
Continuous	Mean	Salary ($)	30,751	93.494	93.494	93.494	102.849	102.798	102.798	7.099
		AGE (=91-BIRTHYR)	42.576	.0751	.0751	.0751	.0732	.0732	.0732	.0028
	Ratio	Schl hrs extra/Hrs requ	.0886	.001	.001	.0011	.001	.0010	.0010	N/A
		Othr hrs extra/Hrs requ	.223	.0013	.0013	.0013	.0014	.0014	.0014	N/A
	Regression Coefficients	Independent:		N/A		N/A	N/A			N/A
		Math	72.152		155.231			188.72	194.46	
		Computer Science	232.656		397.865			258.61	258.10	
		Biology-Life Science	221.769		170.345			126.36	128.73	
		Chemistry	-27.725		379.888			355.94	359.63	
		Physics	323.148		369.148			323.39	319.11	
		Earth/Space science	339.624		345.140			309.12	310.74	
		Other nat science	435.344		369.957			264.33	266.24	
		AGE	451.914		44.107			48.32	44.26	
		Dependent: Salary								
		Independent:		N/A		N/A	N/A			N/A
		Look forward to day	-1.274		.0716			.0726	.073	
		BIRTHYR	-0.745		.0054			.0056	.006	
		Depen.: Years to retire								

N/A: Not available due to capability limitation of the software

NCES recently issued a note from the chief statistician regarding the technical approaches to performing analyses on NCES survey data (Ahmed, 1993) in the desire to perform more complex statistical analyses on NCES data taking account of the complex survey designs. In practice it is not unusual that analysis on complex survey data does not account for the complex design. As reported by a recent survey by the Census Bureau, for instance, many, if not most, journal articles in the social sciences do not account for the complex survey. More effort needs to be made to promote the survey data analysis practice, including the further development and employment of advanced statistical software for variance estimation and other analysis purposes. With today's computing ability and facilities, it is necessary and possible to raise, with our great effort, the standard of practice in the analysis of complex surveys.

[1] This paper reports the general results of research undertaken by staff members of Synectics for Management Decisions, Inc. and the National Center for Education Statistics (NCES). The views expressed are attributable to the authors and do not necessarily reflect those of Synectics or NCES.

References

Ahmed, S. W. (1993), "Technical Approaches to Performing Regression and Other Multivariate Techniques on NCES Survey Data - Where We Stand," A Note from the Chief Statistician. Washington, DC: National Center for Education Statistics.

Fay, R.E. (1989), "Theory and Application of Replicate Weighting for Variance Calculations," *Proceedings of the Survey Research Methods Section*, 212-217. Alexandria, VA: American Statistical Association,

Fay, R.E., (1995), *VPLX*. Washington DC: U.S. Bureau of the Census.

Fuller, W.A., Kennedy, W., Schnell, D., Sullivan, G., and Park, H.J. (1986), *PC CARP*. Ames, IA: Statistical Laboratory, Iowa State University.

Ingels, S.J., Scott, L.A., Rock, D.A., Pollack, M.J., Rasinski, K.A., and Wu, S.-C. (1994), *National Education Longitudinal Study of 1988, First Follow-up Final Technical Report, NCES Technical Report*. Washington, DC: National Center for Education Statistics.

Johnson, E.G. and Allen, N. (1992), *The NAEP 1990, NCES Technical Report*. Washington, DC: National Center for Education Statistics.

Kaufman, S. (1993a), "A Bootstrap Variance Estimator for the Schools and Staffing Survey," *ASA 1993 Proceedings of the Section on Survey Research Methods*.

Kaufman, S. (1993b), "Properties of the Schools and Staffing Survey's Bootstrap Variance," *ASA 1994 Proceedings of the Section on Survey Research Methods*.

Kaufman, S. and Huang, H. (1993), *1991 Schools and Staffing Survey: Sample Design and Estimation, NCES Technical Report*. Washington, DC: National Center for Education Statistics.

Kaufman, S. (1995), "Properties of the School and Staffing Survey's Bootstrap Variance Estimator," presented at the 1995 ASA Meeting.

Kish, L. and Frankel, M. (1974), "Inference from Complex Samples," *Journal of the Royal Statistical Society: Series B* (Methodological), 36: 2-37.

Krewski, D. and Rao, J.N.K. (1981), "Inference from Stratified Samples: Properties of Linearization, Jackknife and Balanced Repeated Replication Methods," *The Annals of Statistics*, 9, 1010-1019.

Liebman, E. (1993), *PC REPTAB* (with *PROC REPTAB*). Berkeley, CA: MPR Associates, Inc.

Ogden, C. and Liebman, E. (1991), *PC STRATTAB* (with *PROC STRATTAB*). Berkeley, CA: MPR Associates, Inc.

Rao, J.N.K. and Wu, C.F.J. (1988), "Resampling Inference with Complex Survey Data," *Journal of the American Statistical Association*, 83, 231-241.

Rust, K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys," *Journal of Official Statistics*, Vol 1, No. 4, Statistics, Sweden.

Sarndal, C.E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shah, B.V., Barnwelll, B.G., Hunt, P., and LaVange, S.C. (1992), *SUDAAN User's Guide* (software version 6.00, 1992). Research Triangle Park, NC: Research Triangle Institute.

Spencer, B.D., Frankel, M.R., Ingels, S.J., Tourangeau, R., and Owings, J.A. (1990), *National Education Longitudinal Study of 1988, Base Year Sample Design Report, NCES Technical Report*. Washington, DC: National Center for Education Statistics.

Westat,Inc. (1993a), *The WESVAR SAS Procedure, Version 1.2*. Rockville, MD.

Westat, Inc. (1993b), *The WESREG SAS Procedure*. Rockville, MD.

Westat, Inc. (1995), *A User's Guide for WesVarPC, Version 1.1*. Rockville, MD.

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