COMPARISON OF TWO VARIANCE-ESTIMATION METHODS FOR A STANDARDIZED ECONOMIC PROCESSING SYSTEM

Antoinette Tremblay and Richard S. Sigman, U.S. Bureau of the Census
Antoinette Tremblay, U.S. Bureau of the Census, Washington, D.C. 20233

Key Words: VPLX, SUDAAN, Jackknife

1.0 Introduction

The United States Census Bureau is re-engineering its post-data-collection processing systems for economic surveys. The goal of this effort is to replace 15 separate systems, which are used for the post-data-collection processing of 113 current surveys, with a single processing system, the Standardized Economic Processing System (StEPS). StEPS will consist of general-purpose modules performing specific procedures applicable to a large number of surveys. One of the StEPS modules will calculate variances of estimates from surveys using single-stage stratified or probability-proportional-to-size sampling. This paper describes a study that we conducted to evaluate incorporating into the StEPS variance module two already developed programs: VPLX (Variances from comPLeX sample surveys) program developed by Fay (1995), and SUDAAN (SUrvey DAta ANalysis for multi-stage sample designs) developed by Research Triangle Institute (1992). VPLX calculates variances using the jackknife or random groups, whereas SUDAAN uses a Taylor approximation. In this paper we characterize the statistical properties of standard errors calculated by VPLX and SUDAAN over a large number of samples selected from a population of generated data resembling data from the Bureau's 1994 Farm and Ranch Irrigation Survey (FRIS).

Because of time constraints, this initial paper only addresses variances of estimates from surveys using single-stage stratified sampling. Also, a third potential StEPS variance estimation package, WesVarPC developed by WESTAT, is not discussed in this paper. Thus, the goal of the research presented in this paper is to recommend the use of either SUDAAN or VPLX to calculate variances for the StEPS surveys that use single-stage stratified sampling.

1.1 StEPS

The StEPS variance-estimation module will consist of several submodules, each calculating variances appropriate for a particular sample design. Among the sample-based surveys that will use StEPS, 33 now calculate variances. Six of these surveys use Poisson sampling, and a submodule written in SAS will calculate their variances. Six of the surveys use multi-stage sampling, and another submodule will calculate the variance through the use of VPLX. Lastly, 21 surveys use single-stage stratified or PPS sampling. Two approaches for these surveys are proposed: VPLX or SUDAAN.

Initially, StEPS will run on ALPHA workstations under UNIX. VPLX is available on this platform, but SUDAAN is not. Therefore, if SUDAAN is utilized by StEPS, it is assumed that Research Triangle Institute will be successful in the development of a SAS-callable SUDAAN that will run on a SUN workstation. StEPS would then use SUDAAN via SAS/CONNECT. All of the VPLX and SUDAAN analyses that we describe in this paper, however, were run on a Digital Equipment Corporation VAX computer.

1.2 The Variance Estimation Programs

VPLX uses replicate weights for variance estimation and it can be used for very simple and very complex sample designs. VPLX has four options: random groups, balanced-half-sample replicates, ordinary or stratified jackknife, and user-specified replicate weights. SUDAAN calculates variances for linear estimates (and for linear approximations to nonlinear estimates) associated with standard sample designs. It uses a Taylor approximation to linearize nonlinear estimators and then uses the appropriate s-squared variance formula. Two of the six designs handled by SUDAAN are single-stage stratified designs and single-stage PPS designs.

Simulation studies have indicated that the variances calculated by the methods used in SUDAAN have smaller mean squared errors than variances calculated via VPLX (Wolter, 1985). However, the VPLX documentation contends that some of the more recent research is not as negative about the VPLX methods.

2.0 Farm and Ranch Irrigation Survey[1]

The 1994 FRIS was conducted to supplement the basic irrigation data collected from all farm and ranch operators during the 1992 Census of Agriculture. Information collected in the FRIS included acres irrigated by category of land use, acres and yields of irrigated and nonirrigated crops, quantity of water applied, number of irrigation wells and pumps, and expenditures for maintenance and repair of irrigation equipment and facilities.

The FRIS universe included all irrigated farms identified in the 1992 Census of Agriculture, excluding farms in Hawaii and Alaska, horticultural specialty farms, and abnormal farms. The universe included some operations erroneously identified as irrigating in the 1992

[1] Some of the material in this section is taken from Census Bureau publication AC92-RS-1.

Census either due to reporting or Census processing errors. The universe consisted of 246,427 farms.

The number of farms included in the FRIS sample with certainty was 1,175. The remainder of the universe was then stratified on the basis of geography (first-level stratifier) and 1992 Census-reported total irrigated acres (second-level stratifier). An independent systematic sample was selected within each stratum, yielding an additional 18,823 sampled farms.

The number of responding sampled farms was 12,735. Sampling weights were adjusted for nonresponse by strata, and separate-ratio estimation with Census-reported irrigated acres as the auxiliary variable was used to produce estimates for each of the 27 leading irrigating states, the 21 remaining combined states, and at the national level. Also, estimates were made for each of 18 Water Resources Areas (WRAs), which are geographical areas corresponding to major drainage basins. The state estimates and the combined-state estimate were first-level stratum estimates. In the eastern United States the WRA estimates were also first-level stratum estimates, whereas in the midwestern and western United States the WRA estimates were domain estimates. Variances were calculated for the estimates of selected items using the method of random groups.

3.0 Study Population

Because FRIS data was available only for respondents, we developed a study population of simulated FRIS data from which to draw Monte Carlo samples. This study population consisted of the 37,593 farms in the FRIS sampling frame in the six states of Arizona, Colorado, Nevada, New Mexico, Utah and Wyoming. These six states contained all of WRA 14 (Upper Colorado) and also all of WRA 15 (Lower Colorado). Table 1 shows FRIS frame and sample counts, plus the number of FRIS responding farms, for the six study states.

We matched to each farm in the study population its Census-reported irrigated acres and either reported values (for certainty farms) or simulated values (for non-certainty farms) for two FRIS variables: (1) acre-feet of water applied from all sources, and (2) maintenance and repair expenses for irrigation. We used the following prediction equation to produce simulated FRIS data:

$$y_i = \begin{cases} g_i & x_i < x_0 \\ \beta(x_i - x_0) + g_i & x_i \geq x_0 \end{cases}$$

where

y_i = FRIS data for farm i,

x_i = Census-reported irrigated acres for farm i,

and g_i is a gamma random variable such that

$$E(g_i|x_i) = \begin{cases} \alpha+\beta x_i & x_i < x_0 \\ \alpha+\beta x_0 & x_i \geq x_0 \end{cases}$$

and

$$Var(g_i|x_i) = x_i^{2\gamma}\sigma^2 .$$

Then

$$E(y_i|x_i) = \alpha+\beta x_i$$

and

$$Var(y_i|x_i) = x_i^{2\gamma}\sigma^2 .$$

The analysis of economic data reported in Steel and Fay (1995) suggested to us the use of gamma-distributed errors, plus this permitted us to compare our empirical results for ratio estimation to analytical results in Krewski and Rao (1981).

We estimated the parameters x_o, α, β, γ, and σ^2 from the matched FRIS and Census data available for the FRIS respondents in selected model strata (combinations of FRIS strata) in four of the six study states. When predicting FRIS data for a state or model stratum that lacked estimated parameters, we used the parameters estimated in a similar state and/or model stratum. We estimated α, β, γ, and σ^2 by assuming the regression model

$$y_i = \alpha+\beta x_i+x_i^{\gamma}e_i$$

and using iterative weighted least squares to estimate α, β, and γ. If α was not significantly different from zero, we set it to zero. The estimate for σ^2 was the residual mean sum of squares for the regression model transformed by dividing each term by x_i^{γ}. We estimated x_0 by setting it equal to the largest value of x_i such that there existed a matched y_i equal to zero.

Tables 2 and 3 list the estimated model parameters. Since we deleted a small number of outliers, the sample sizes for the same state and model stratum, but for different FRIS variables, can be slightly different. Tables 4 and 5 compare various quantities calculated with actual FRIS data with the same quantities calculated with simulated data.

4.0 Monte Carlo Results

We independently selected 5,000 stratified-simple-random samples from the study population, using the same sampling rates that were used in the FRIS. Our Monte Carlo sampling was with replacement, but in each Monte Carlo sample we selected farms without replacement. We simulated nonresponse by using a missing-at-random model in which the probabilities of

nonresponse matched the FRIS nonresponse rates by stratum. In each Monte Carlo sample we adjusted the sampling weights based on that sample's counts of responding farms.

In 4,000 of the 5,000 samples, we used SUDAAN to calculate simple-weighted estimates and associated standard errors for individual states and the six states combined--these were stratum estimates--and for WRAs 14 and 15, which were domain estimates. For 2,000 of these 4,000 samples, we used SUDAAN to calculate separate-ratio estimates and associated standard errors for the six states combined and WRAs 14 and 15. For these same 2,000 samples, we used VPLX to also calculate simple-weighted and separate-ratio estimates, plus the associated standard errors (via stratified jackknife), at the same levels as for SUDAAN.

By calculating the standard deviation of the estimates over Monte Carlo samples, we were able to estimate the true standard error of the calculated estimates. This permitted us to estimate the biases of the estimated standard errors calculated by VPLX and SUDAAN. Tables 6 and 7 contain estimates of the relative biases, coefficients of variation, and relative root-mean-square errors of the estimated standard errors calculated by VPLX and SUDAAN. (The denominator of all of these quantities is the estimate of the true standard error of the calculated estimates.) The appendix lists the formulas we used to calculate the standard errors that are enclosed in parentheses in Tables 6 and 7. (The less stable estimate of the variance for the separate ratio estimates in WRA 14 for estimated Acre Feet is due, we believe, to small sample size and low correlation between estimated Acre Feet and Census-reported total acres irrigated.)

5.0 Conclusions, Recommendation and Further Research

Based on the results, we make the following conclusions:

- For simple-weighted estimation, the biases and variances of the standard errors calculated by VPLX and SUDAAN are nearly identical.
- For simple-weighted estimation, we found statistically-significant negative biases in some of the standard errors calculated by VPLX and SUDAAN. These may have been a result of the weight adjustment for nonresponse.
- For separate-ratio estimation, the standard deviations calculated by SUDAAN can have significant negative bias for domain estimates. This occurred less often for VPLX and when it did, the absolute bias of VPLX was less than that for SUDAAN.
- For separate-ratio estimation, the variances of standard errors calculated by VPLX are larger than those calculated by SUDAAN.
- For separate-ratio estimation, the larger absolute bias of SUDAAN and the larger variance of VPLX tend to balance each other when one considers the root mean square errors of calculated standard errors. SUDAAN tends to have slightly smaller relative root-mean-square errors, however, in most situations. When the relative root-mean-square of SUDAAN was substantially less than that for VPLX, SUDAAN had a much larger negative bias.

As stated in the introduction, the goal of this research was to recommend the use of either SUDAAN or VPLX to calculate variances for the StEPS surveys that use single-stage stratified sampling. The choice between the two programs is not obvious in terms of the studied statistical properties, plus survey size is not a constraint for either program. Thus, administrative considerations such as cost and available support play a more critical role. Compared to SUDAAN, VPLX is more flexible: basically, anything that can be set up in a formula can be done in VPLX. Also, development versions of VPLX correctly take into account the effect of imputed data on variance estimation. VPLX is 'license-free', and consulting is more readily available since its developer/maintainer is resident at the Census Bureau. Thus, we are recommending VPLX for StEPS method of variance estimation.

Even though we have made this recommendation, there are additional areas of needed research:

- Compare the random-groups and stratified jackknife methods in VPLX, because several StEPS surveys currently use the random-groups method;
- Compare VPLX to WesVarPC, because WesVarPC has an interactive user-interface and VPLX does not; and
- Analyze the applicability of the variance estimation programs to surveys using probability-proportional-to-size sampling.

6.0 Acknowledgements

The authors are grateful to Bob Fay and George Train for assisting us in the use of VPLX and to Bob Smith and Carolyn Swan for sharing their knowledge on the sample design, estimation methods, and available data sets for the 1994 Farm and Ranch Irrigation Survey.

7.0 References

Fay, R.E. (1995), "VPLX: Variance Estimation for Complex Samples, Program Documentation," unpublished Bureau of the Census report.

Kendall, M. and A. Stuart (1977), *The Advanced Theory of Statistics*, Volume 1, Fourth Edition, New York, NY: McMillan.

Krewski D. and J.N.K. Rao (1981), "Inference from Stratified Samples: Properties of the Linearization,

Jackknife and Balanced Repeated Replication Methods," *The Annals of Statistics,* Volume 9, pp. 1010-1019.

Research Triangle Institute (1992), *SUDAAN User's Manual*, Release 6.0. SAS Institute Inc. (1993), *SAS/INSIGHT User's Guide*, Version 6, Second Edition, Cary, NC: SAS Institute Inc.

Steel P. and R.E. Fay (1995), "Variance Estimation for Finite Populations with Imputed Data," *Proceedings of the Survey Research Section*, American Statistical Association, Alexandria, VA, pp. 374-379.

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

APPENDIX

Lemma 1. If X_1 and X_2 are random variables such that $Var(X_1)=O(n^{-1})$ and $Var(X_2)= O(n^{-1})$, then an approximate upper bound for the standard deviation of X_1/X_2 is

$SD(X_1/X_2) < SD(X_1)/|\theta_2| + |\theta_1/\theta_2|C_2$,

where $SD(X_1)$ is the standard deviation of X_1, θ_1 is the mean of X_1, θ_2 is the mean of X_2 and C_2 is the coefficient of variation of X_2.

Proof: Follows from the Taylor-approximation to $Var(X_1/X_2)$:

$$Var(X_1/X_2) \approx Var(X_1)/\theta_2^2 + \theta_1^2 Var(X_2)/\theta_2^4 - 2\theta_1 Cov(X_1,X_2)/\theta_2^3$$

Lemma 2. If a sample $X_1, X_2, ..., X_n$, is selected from a normal population and S is the sample standard deviation of $X_1, X_2, ..., X_n$, then the coefficient of variation of S is approximately $1/\sqrt{2n}$.

Proof: Follows from result on pages 250 and 258 of Kendall and Stuart (1977), which states that for a normal parent distribution, the large-sample standard deviation of S is $\sigma^2/2n$.

Proposition 1: If A and B are random variables such that $Var(A)=O(n^{-1})$ and B is the sample standard deviation of a sample $X_1, X_2, ..., X_m$, selected from a normal population, then an approximate upper bound for the standard deviation of A/B is

$$SD(A/B) < SD(A)/|B^*| + |A^*/B^*|/\sqrt{2m},$$

and an approximate upper bound for the standard deviation of (A-B)/B is

$$SD[(A-B)/B] < SD(A)/|B^*| + |1+(A^*-B^*)/B^*|/\sqrt{2m},$$

where A^* and B^* are the mean of A and B, respectively.

Proof: Follows from lemmas 1 and 2.

Proposition 2. If $X_1, X_2, ..., X_n$ are selected from a normal distribution with mean μ and variance σ^2, an approximate standard error for the sample coefficient of variation is $V[(1+2V^2)/2n]^{1/2}$, where $V=\sigma/\mu$.

Proof: See page 258 of Kendall and Stuart (1977), and specifically Example 10.5 on page 248.

Table 1. FRIS farm counts for study population.

Level	Number of Farms			
	FRIS Sampling Frame		FRIS Non-certainty Sample	FRIS Non-certainty Respondents
	All Farms	Non-certainty Farms		
All 6	37,593	37,089	2,760	1,796
AZ	3,141	2,945	395	225
CO	12,645	12,572	590	412
NV	1,791	1,668	403	233
NM	5,988	5,959	491	325
UT	9,609	9,575	415	271
WY	4,419	4,370	466	230
WRA 14	8,191	8,159	373	248
WRA 15	3,765	3,566	440	249

Table 2. Estimated model parameters for Acre Feet.

State	Model Stratum	n	x_0	α	β	γ	σ^2
CO	large	188	1060	0	1.6	0.75	34.2
CO	medium	180	0	0	2.0	1.10	0.8
NM	medium	149	244	0	2.9	1.05	2.1
UT	medium	147	0	0	2.0	0.75	39.0
WY	large	183	1800	372	1.3	0.75	48.5

Table 3. Estimated model parameters for Maintenance Expenses.

State	Model Stratum	n	x_0	α	β	γ	σ^2
CO	large	188	1900	0	6.5	1.1	8.2
CO	medium	180	0	0	5.9	1.0	53.3
NM	medium	147	0	0	8.7	0.9	486.8
UT	medium	149	0	0	5.5	1.0	82.2
WY	large	182	2100	0	2.2	0.7	572.1

Table 4. Comparisons between actual and simulated data for Acre Feet.

Model	Model stratum	Unweighted Stats				Means		
		Skewness		Correlation with x_i		Weighted Sample		Population simulated
		actual	simulated	actual	simulated	actual	simuated	
CO	large	2.07	1.51	0.60	0.63	1399	1265	1349
CO	medium	2.44	2.03	0.47	0.49	334	373	373
NM	medium	1.78	3.96	0.60	0.60	411	513	490
UT	medium	1.44	1.85	0.57	0.48	348	321	313
WY	large	1.74	2.64	0.39	0.50	1600	1506	1631

Table 5. Comparisons between actual and simulated data for Maintenance Expenses.

Model	Model stratum	Unweighted Statistics				Means		
		Skewness		Correlation with x_i		Weighted Sample		Population simulated
		actual	simu-lated	actual	simu-lated	actual	simu-lated	
CO	large	2.32	1.71	0.45	0.46	5079	5446	5960
CO	medium	2.44	4.25	0.29	0.31	1076	1246	1098
NM	medium	3.07	3.79	0.37	0.19	1636	1505	1463
UT	medium	6.33	5.40	0.31	0.24	856	832	885
WY	large	2.61	2.18	0.44	0.31	2593	2098	2138

Table 6. Estimates of relative biases, coefficients of variation (CVs), and relative root-mean-square errors (RMSEs) of the SUDAAN and VPLX estimated standard errors for estimated Acre Feet.

Simple Wtd.	Relative Bias (%)			CV (%)			Relative RMSE (%)		
	S (s.e.)	V (s.e.)	|S|-|V| (s.e.)	S	V	S-V	S	V	S-V
All 6	0.96 (1.30)	0.96 (1.30)	0.005 (0.001)	7.64	7.64	-0.001	7.70	7.70	-0.000
AZ	-1.13 (1.47)	-1.12 (1.47)	0.004 (0.001)	16.13	16.13	-0.003	16.17	16.17	-0.003
CO	-1.07 (1.39)	-1.08 (1.39)	-0.008 (0.001)	12.30	12.30	-0.002	12.35	12.35	-0.002
NV	-1.22 (1.37)	-1.22 (1.37)	0.006 (0.001)	12.01	12.01	0.002	12.07	12.07	0.002
NM	-1.07 (1.38)	-1.06 (1.38)	0.001 (0.001)	12.26	12.26	-0.001	12.30	12.30	-0.001
UT	0.10 (1.42)	0.10 (1.42)	0.001 (0.000)	13.30	13.30	-0.000	13.30	13.30	-0.000
WY	1.15 (1.36)	1.15 (1.36)	0.007 (0.000)	10.16	10.16	-0.000	10.22	10.22	0.001
State Avg.	-0.54 (0.57)	-0.54 (0.57)	-0.001 (0.000)	12.69	12.69	-0.001	12.73	12.74	-0.001
WRA 14	-0.90 (1.40)	-0.93 (1.40)	-0.031 (0.004)	12.90	12.90	0.001	12.93	12.93	-0.001
WRA 15	-2.00 (1.46)	-2.00 (1.46)	0.003 (0.001)	16.28	16.28	-0.003	16.40	16.40	-0.002
Separate Ratio									
All 6	-0.84 (1.78)	1.37 (1.91)	-0.534 (0.145)*	9.35 (0.15)	13.84 (0.22)	-4.491	9.39	13.91	-4.521
WRA 14	-40.17 (1.21)	-15.05 (3.04)	25.126 (2.037)	11.63 (0.19)	75.81 (1.76)	-64.174	41.82	77.29	-35.463
WRA 15	-7.06 (1.85)	-1.59 (1.97)	5.481 (0.259)	17.05 (0.28)	18.37 (0.30)	-1.312	18.46	18.43	0.025

S=SUDAAN V=VPLX

*s.e. of the difference of the signed relative biases.

Table 7. Estimates of relative biases, coefficients of variation (CVs), and relative root-mean-square errors (RMSEs) of the SUDAAN and VPLX estimated standard errors for estimated Maintenance Expenses.

Simple Wtd.	Relative Bias (%)			CV (%)			Relative RMSE (%)		
	S (s.e.)	V (s.e.)	|S|-|V| (s.e.)	S	V	S-V	S	V	S-V
All 6	-1.25 (1.31)	-1.26 (1.31)	-0.013 (0.002)	9.21	9.21	0.000	9.30	9.30	-0.001
AZ	-3.46 (1.58)	-3.46 (1.58)	0.003 (0.001)	22.58	22.58	-0.001	22.84	22.84	-0.000
CO	-0.04 (1.41)	-0.06 (1.41)	-0.024 (0.003)	12.95	12.95	-0.000	12.95	12.95	-0.001
NV	-4.13 (1.72)	-4.13 (1.72)	0.006 (0.001)	29.21	29.21	-0.003	29.50	29.50	-0.002
NM	-2.19 (1.48)	-2.19 (1.48)	0.001 (0.001)	17.57	17.57	0.000	17.70	17.70	0.001
UT	-5.75 (1.53)	-5.75 (1.53)	-0.000 (0.000)	21.57	21.57	0.000	22.32	22.32	0.000
WY	-1.55 (1.55)	-1.56 (1.55)	-0.007 (0.000)	20.22	20.22	-0.000	20.28	20.28	-0.001
State Avg.	-2.85 (0.63)	-2.86 (0.63)	-0.003 (0.000)	20.68	20.68	-0.001	20.93	20.93	-0.000
WRA 14	-0.14 (1.53)	-0.21 (1.53)	-0.071 (0.012)	18.50	18.50	-0.000	18.50	18.50	-0.001
WRA 15	-3.05 (1.56)	-3.04 (1.56)	0.004 (0.001)	21.42	21.42	-0.001	21.63	21.63	-0.000
Separate Ratio									
All 6	-2.00 (1.75)	-1.36 (1.76)	0.643 (0.018)	8.81 (0.14)	8.96 (0.14)	-0.151	9.03	9.06	-0.029
WRA 14	-4.68 (1.95)	1.40 (2.08)	3.281 (0.133)*	19.54 (0.32)	21.36 (0.35)	-1.819	20.10	21.41	-1.313
WRA 15	-6.53 (1.95)	-2.31 (2.03)	4.222 (0.215)	20.80 (0.34)	21.75 (0.36)	-0.949	21.80	21.87	-0.070

S=SUDAAN V=VPLX

*s.e. of the difference of the signed relative biases.

This paper reports the general results of research undertaken by Censu Bureau Staff. The views are attributable to the authors and do not necessarily reflect those of the Census Bureau.