Paper 238-25

Analyzing Complex Sample Survey Data: A New Beginning

Linda Tompkins, National Center for Health Statistics, Hyattsville, MD
Arlene B. Siller, National Center for Health Statistics, Hyattsville, MD

ABSTRACT

The National Center for Health Statistics conducts surveys with probability-based complex sample designs to produce estimates of health conditions for the civilian, noninstitutionalized population of the United States. Researchers usually present descriptive statistics such as means, totals, and their standard errors. However, in order to make statistically valid population inferences from sample data, standard errors must be computed using procedures that take into account the complex nature of the sample design. This paper compares estimates produced with PROC MEANS, Version 7's PROC SURVEYMEANS, and an alternative program, SUDAAN. The paper will also include benchmarks for comparing the three procedures in a Windows environment, as well as insights into the efforts necessary to produce these statistical outcomes.

INTRODUCTION

Large-scale, household surveys are often conducted by government organizations and private institutions to estimate characteristics of interest for some underlying population based on those of individuals selected in the sample. Such surveys usually employ complex, probability-based designs intended to increase the precision of the estimates obtained. Methods used include the clustering of survey respondents, stratification, and the assignment of unequal probabilities of selection. However, increased precision comes at a cost -- that of a departure from the assumption of independent sample points having equal probabilities of selection. This departure creates a requirement for specialized statistical software to accurately compute estimates of population statistics and their standard errors. This paper will examine the estimates of selected health characteristics produced by SUDAAN (Research Triangle Institute), and by SAS® (SAS Institute), now experiencing a *new beginning* with its survey procedures.

BACKGROUND

The National Health Interview Survey has been conducted continuously since 1957 for the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC) by the U. S. Census Bureau. The NHIS serves as a major data collection program, enabling NCHS analysts and researchers worldwide to estimate the health characteristics for the noninstitutionalized, civilian population of the United States. Prior to 1995, the survey had consisted of a "core" questionnaire, containing a set of basic health and demographic items, and one or more "supplements" on current health topics. In 1997, the NHIS questionnaire was redesigned in an attempt to reduce respondent burden, the data collection budget, and workloads on NCHS and Census staff. It also had as a design goal to improve the reliability of NCHS statistics for racial, ethnic, economic, and geographic domains.

The design of the NHIS has traditionally been redesigned every ten years, following each decennial census. The current sampling plan follows a multistage area probability design that permits the representative sampling of approximately 50,000 households and 100,000 persons per year. The first stage consists of a sample of 358 primary sampling units (PSU's) drawn from approximately 1,900 geographically defined PSU's that cover the 50 States and the District of Columbia. The NCHS treats the NHIS as a two-stage sample for variance estimation purposes.

The redesigned NHIS for 1997 consists of three parts: (1) a Basic module, which will remain unchanged from year to year and consists of family, sample adult, and sample child components; (2) a Periodic module, which will collect more detailed data on specific topics from sample persons; and (3) a Topical module, which will collect data on an existing topic or on new emerging health issues, as the need arises. The survey was conducted using computer-assisted personal interviewing (CAPI) by Census Bureau interviewers. The year of surveying yielded data on 103,477 total persons residing in 39,832 households.

COMPARISON OF SOFTWARE SYSTEMS

Research documents and reports produced by analysts at NCHS typically present percentages, means, totals, and standard errors for selected health and demographic data collected in it's many data collection systems. Over recent years, SAS (SAS Institute, Inc) has been favored as the "software of choice". However, SAS products prior to Version 7 produced accurate point estimates, but produced inaccurate variances and standard errors, which were usually underestimated. The computational algorithms used failed to account for the complexity of the survey design (i.e., clustering, stratification, unequal probabilities of selection. The standard errors produced, as if it were a simple random sample, generally underestimated the true population value. These erroneous estimates negate the validity of resulting confidence intervals or tests of statistical significance. This left researchers in the position of having to perform "some" analysis in SAS or other software, but having to produce standard errors in a package with complex survey design capabilities.

Early prototypes of SUDAAN (Research Triangle Institute) were used for this purpose. As a result of the collaboration between RTI and several government agencies, these procedures have evolved into a concise software package capable of analyzing complex sample survey data. It's analytical tools include four descriptive statistics procedures: CROSSTAB, RATIO, DESCRIPT, and RECORDS, as well as four modeling procedures: REGRESS, LOGISTIC, MULTILOG, and SURVIVAL. In addition, SUDAAN procedure statements and syntax are similar to those in SAS. Moreover, SUDAAN procedures are available in a stand-alone package, as well as procedures which are "callable" from within the SAS program. (Note: Although both packages are available on many computing platforms, (e.g., mainframe, PC, Unix), RTI plans no further mainframe development after Version 7.0).

Responding to users' requests and needs, SAS Institute introduced three experimental procedures designed to analysis data derived from a complex sample survey in it's Version 7 release: (1) PROC SURVEYSELECT provides methods for selecting probability-based samples, while simultaneously employing clustering, stratification, and unequal probabilities of section; (2) PROC SURVEYMEANS computes estimates of the survey population means, totals, and the associated standard errors; and (3) PROC SURVEYREG performs regression analysis for survey data. These experimental procedures are scheduled to become production products with Version 8 of the SAS System.

The most commonly used procedures for estimating population characteristics from complex sample survey data are balanced repeated replication (BRR), the jackknife method, and Taylor series linearization. The latter technique will be used in examining estimates for selected demographic and health variables for this analysis.

RESULTS

(Note: results are incomplete at the time of publication, final results will be available at the Poster Session or contact the authors for a copy)

For this comparison, data from the 1997 NHIS public use data files was obtained. Because the files are large in size and PC processing of the data was desired, the sample adult component of the Basic NHIS module (data collected on one sample adult per responding household), was chosen. The file contains various demographic and health information collected on 31,116 adults. In addition, stratum and PSU information, used to specify the design structure to estimation procedure, and sampling weights, needed to produce unbiased estimates, are included. Although design information and weights are on the file, analysts are cautioned to inspect estimates produced for domains with small sample sizes, which may indicate the presence of unstable and unreliable variance estimates.

COMPARABILITY OF SURVEY STATISTICS **

Statistics are presented in Table 1 for five selected demographic variables with discrete categories. To obtain frequencies and standard errors in SAS Version 6.12, one category from each of the demographic variables was used to create dummy (0,1) variables to be input to PROC MEANS. PROC SURVEYMEANS can accept either discrete or continuous variables, provided that all categorical variables be named on a CLASS statement. SUDAAN requires that variables of this type be named on a SUBGROUP statement and that the number of values for each be named on a LABELS statement. Estimated percentages produced by PROC MEANS are identical to those produced by PROCs SURVEYMEANS and CROSSTAB. However, the standard errors are smaller, will create narrower confidences intervals, and impact any tests of statistical significance. Results from PROCs SURVEYMEANS and CROSSTAB are comparable.

Table 1: Comparison of SAS and SUDAAN Percentages and Standard Errors

Demographic Characteristic (% of Respondents)	SAS Version 6.12 PROC MEANS	SAS Version 7 PROC SURVEY-MEANS	SUDAAN Version 7.5 PROC CROSSTAB
Elderly Respondents (aged 65 yrs. +)	16.3909 (.1980)	16.3909 (.2677)	16.3909 (.2671)
Respondents of Hispanic Origin	9.8637 (.1569)	9.8637 (.2505)	9.8637 (.2506)
Black Respondents	11.1796 (.1658)	11.3440 (.2822)	11.3440 (.2822)
Married Respondents	58.6995 (.2591)	58.6995 (.3813)	58.6995 (.3813)
Live in the South	35.5433 (.2519)	35.5433 (.4486)	35.5433 (.4486)

**** PLEASE NOTE: DATA COLLECTED IN THE 1997 NHIS ARE PROVISIONAL AND ARE SUBJECT TO CHANGE. VARIABLES WERE SELECTED FOR THIS ANALYSIS FOR ILLUSTRATIVE PURPOSES ONLY. RESULTS SHOWN HERE SHOULD NOT BE USED TO MAKE INFERENCES ABOUT THE US POPULATION**.

Means and standard errors for selected health characteristics (continuous variables) are presented in Table 2 below. As was true for the demographic characteristics (discrete or dummy variables), the means across packages are identical, whereas the standard errors are reduced. Again, these underestimated standard errors impact the width of confidence intervals and tests of significance.

Table 2: Comparison of SAS and SUDAAN Means and Standard Errors

Demographic / Health Characteristic	SAS Version 6.12 PROC MEANS	SAS Version 7 PROC SURVEY-MEANS	SUDAAN Version 7.5 PROC DESCRIPT
Age of Respondent	44.5846 (.0917)	44.5846 (.1505)	44.5846 (.1505)
Body Mass Index	26.2164 (.0282)	26.2164 (.0376)	26.2164 (.0376)
Frequency of Vigorous Activity (times/week)	1.4085 (.0137)	1.4085 (.0187)	1.4085 (.0187)

Table 3: Comparison of SAS and SUDAAN Regression Coefficients and Standard Errors

Two *strictly hypothetical* simple linear regressions were run in both packages for illustrative purposes only:

1) heart disease = age sex race smoking body mass index

2) body mass index = age sex race vig

Independent and Dependent Variables	SAS Version 6.12 PROC REG	SAS Version 7 PROC SURVEYREG	SUDAAN Version 7.5 PROC REGRESS
heart disease intercept age sex race smoking body mass index			
body mass index age sex race exercise			

COMPARABILITY OF EFFICIENCY

All runs were made on a Dell 233 with 32 Mb RAM. The identical dataset containing 31,116 observations and 35 variables was input to each procedure. Also, the same number of variables was input at once, thereby making only one call to the procedure necessary. It must be noted, however, that SUDAAN procedures require that the input dataset be sorted by variables that describe the stages of the sample design (i.e., on the NEST statement). The "efficiency" of each procedure, as measured by "real time" recorded in the SAS log, is presented in Table 4 below:

Table 4: Comparison of SAS and SUDAAN Efficiency (Real Time)

Efficiency Measure	SAS Version 6.12 PROC MEANS	SAS Version 7 PROC SURVEYMEANS	SUDAAN Version 7.5 PROC DESCRIPT
SAS' log "real time" (In seconds)			

Efficiency Measure	SAS Version 6.12 PROC REG	SAS Version 7 PROC SURVEYREG	SUDAAN Version 7.5 PROC REGRESS
SAS' log "real time" (In seconds)			(Sort = 18.00 secs.)

CONCLUSIONS

Surveys are implemented to obtain data used to make inferences about the underlying population. The designs of these instruments have become increasing complex, often using multiple stages of sample selection, unequal probabilities of selection, clustering, and stratification. As care and effort are taken to structure these designs, great care and effort must be used in selecting the appropriate computer software to analyze the data collected. Therefore, the analyst must be aware of the many variance estimation techniques, as well as the available computer software necessary to compute accurate statistics from this complex sample survey data. If not, computed results may be erroneous and result in making incorrect inferences about the population being studied.

REFERENCES

An AB and Watts DL (1998), "New SAS Procedures for Analysis of Sample Survey Data.", *Proceedings of the Twenty Third Annual SAS Users Group International Conference*, 23,

Brogan DJ (1998), "Pitfalls of Using Standard Software Packages for Sample Survey Data." *Encyclopedia of Biostatistics*. P Armitage and T Colton, eds. John Wiley, New York.

Cochran WG (1977), *Sampling Techniques*. John Wiley & Sons, New York.

Cohen SB (1997), "An Evaluation of Alternative PC-Based Packages Developed for the Analysis of Complex Survey Data." *The American Statistician*, Vol. 51, No. 3.

Lepkowski JM and Bowles, J (1996), "Sampling Error Software for Personal Computers." *The Survey Statistician*, No. 35, pp. 10-17.

National Center for Health Statistics (1999). National Health Interview Survey: Research for the 1995-2004 Redesign. Vital Health Statistics 2 (126).

SAS Institue Inc. (1999), OnLine Documentation, Cary, NC. http://www.sas.com/rnd/app/da/new/dasurvey.html

SAS Institute Inc. (1989), *SAS/STAT User's Guide, Version 6, Fourth Edition, Vol. 2*, Cary, NC.

Shaw BV, Barnwell BG and Bieler GS, (1997). *SUDAAN User's Manual: Release 7.5*, Research Triangle Institute, Research Triangle Park, NC.

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

CONTACT INFORMATION:

Linda Tompkins
National Center for Health Statistics
6525 Belcrest Road
Room 915
Hyattsville, MD 20782
(301)458-4533
Fax: (301)458-4031
Email: LIT3@CDC.GOV

Arlene B. Siller
National Center for Health Statistics
6525 Belcrest Road
Room 952
Hyattsville, MD 20782
(301)458-4498
Fax: (301)458-4032
Email: ABS2@CDC.GOV

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.