

Survey Estimates and Variance Estimation Using the SURVEYMEANS Procedure

Hossein N. Yarandi, Ph.D.* College of Nursing and Biostatistics Unit, University of Florida

Shawn M. Kneipp, Ph.D., ARNP, College of Nursing, University of Florida

(* - principal contact)

National surveys, conducted by the various governmental organizations, are usually characterized with a complex sample design. The characteristics of a complex sample design may include clustering, stratification, disproportionate sampling, and multiple stages of sample selection. Analyzing data from a complex sample survey cannot be done accurately using the standard SAS Procedures such as MEANS, UNIVARIATE, REG, or GLM. When using these procedures, it is assumed that the data are selected from a simple random sample.

A simple random sample requires that one have access to all the subjects in the population and the probability of selecting each subject in the population is the same for each and every subject. In addition, the chance of selecting a subject is independent even if other subjects are chosen. In contrast, when the data are collected from a survey with a complex sample design, the variance of the variables measured in the survey will be underestimated when using the standard statistical methods for nonsurvey data resulting in lower confidence intervals and higher Type I error by rejecting a true null hypothesis (Cohen 1997). A number of alternatives, called replication methods, have been developed for estimating variances and weighting procedures for the types of complex sample designs usually encountered in practice (Judkins and Wright 1990; Flyer, Morganstein, and Rust 1989).

In SAS Version 8 and beyond, three procedures are added to SAS/STAT software (SAS/STAT, 2000). These procedures are: SURVEYMEANS, SURVEYREG, and SURVEYSELECT. They are designed to specifically analyze complex survey data. The purpose of this article is to demonstrate the difference in variance estimation using the SURVEYMEANS procedures for a variable, such as AGE, from the Medical Expenditure Panel Survey. The data will be analyzed without the sampling weights, with sampling weights, and stratified sampling.

Medical Expenditure Panel Survey

The Medical Expenditure Panel Survey (MEPS) is cosponsored by the Agency for Healthcare Research and Quality (AHRQ) and the National Center for Health Statistics (NCHS). MEPS provides nationally representative estimates of health care use, its expenditure, sources of payment, and insurance coverage for the U.S. civilian non-institutionalized population (MEPS HC-003, 1996). MEPS also includes a national representative survey of nursing homes and their residents. MEPS is comprised of four component surveys: the household component, the medical provider component, the insurance component, and the nursing home component. In addition, it allows for

health services research intended to guide health care policy.

Designed to capture the changing dynamics of health care delivery and its insurance system, MEPS began in 1996. An important feature of the MEPS is that it uses a panel survey design in which minority and low-income households are oversampled. Medical expenditure data are collected at both the individual and household levels through a prescreening telephone interview, a mailed questionnaire, and a telephone followup interview for nonrespondents. The data reflect demographic characteristics, income, employment, health conditions, health status, access and use of medical care services, charges and payments, satisfaction with care, and health insurance coverage (MEPS HC-003, 1996).

Replication Methods for Analyzing Data from Complex Survey

Replication methods involve the selection of subsamples from the whole sample, then the estimated variances among the subsamples are used to estimate the variance of the full sample (Judkins and Wright 1990; and Wolter 1985). Another approach is the Taylor Expansion Method that can be used in estimating variances for the types of complex sample designs and weighting procedures usually encountered in practice. There are various replication approaches for creating subsamples from the full sample but the two widely used methods are Balanced Repeated Replication and the Jackknife Replication. Balanced repeated replication is generally used with multistage stratified sample designs while the Jackknife replication is preferred when there is no explicit stratification. The jackknife replication method can, however, be used when systematic sampling has been employed (Cohen, Burt, and Jones 1986; and Flyer 1987). Frequently, in complex survey data sets the full sample weights are provided. In addition, to employ a replication method, two additional replicate weights called the primary sampling unit (PSU) and the variance strata are usually required.

MEPS uses multistage sampling: At the first stage of sampling, a small number of primary sampling units (PSUs) are selected. Most of these PSUs consist of individual counties, but sometimes they include two or more adjacent counties. A random samples of PSUs are selected in the first stage of sampling. During the second and additional stages of sampling, only individuals living within these selected PSUs are selected.

To decrease the variability of parameter estimators based on data from the completed survey, sampling theory suggests that it is better to sample large-population PSUs with a higher probability than small-population PSUs (Korn and Graubard, 1999). Recall that, one of the design considerations of MEPS is to provide reliable estimates the minority groups. To help accomplish this, PSUs with larger proportions of these minorities were included in the sample with higher probabilities. Additionally, from a meteorological perspective, one can classify PSUs into a small number of groups ("strata") and then sample a small number of PSUs from each stratum. This is known as stratified sampling, and is frequently used to decrease the variability of estimated quantities. Sample survey methodology allows one to sample units from the strata with differing probabilities provided by the survey designers. The data from each individual are associated with a sample weight, which is essentially the number of people in the target population that he or she represents. In calculating these weights, consideration is given to the differential probabilities with which individuals are sampled.

SAS Procedures

There are three SAS procedures designed to manage survey data. They include SURVEYMEANS Procedure, SURVEYREG Procedure, and SURVEYSELECT Procedure (SAS/STAT, 2000).

The SURVEYMEANS Procedure takes into account the sample design used to select the survey sample and provides descriptive statistics of survey population such as means, standard error, and confidence limits. The sample design can be a complex survey sample design with stratification, clustering, and unequal weighting. The Taylor expansion method is used to estimate the variance. It computes sampling errors of estimators by obtaining a linear approximation for the estimators (Wolter 1985). The estimate of the variance is the variation among PSU when there are clusters. When a stratified design is used, the procedure pools stratum variance estimates to compute the overall variance estimate.

The SURVEYREG can handle complex survey sample designs by performing regression analysis for survey data. The procedure fits linear models for survey data and computes regression coefficients and their variance-covariance matrix. The procedure also provides the predicted values for the sample survey data, computes the regression coefficient estimators by generalized least squares estimation using element-wise regression, and calculates the significance tests for the model effects and for any specified estimable linear functions of the model parameters. The procedure assumes that the regression coefficients are the same across strata and the PSUs.

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample or a sample according to a complex multistage sample design that includes stratification, clustering, and unequal probabilities of sample selection.

An Example

The variable AGE is chosen of a subsample of women with the age range of 18-46 years from the MEPS 1996 full-year data set. Procedure SURVEYMEANS is used to compute the mean, standard error, and 95% confidence limit for the variable age. Figure 1 shows the results of the analysis without using the sampling weights.

Quite often in complex surveys, respondents have unequal weights, which reflect unequal selection probabilities and adjustments for nonresponse and stratification. In such surveys, the appropriate sampling weights must be used to obtain valid estimates for the study population. Figure 2 illustrates the mean, standard error and 95% confidence interval for the variable AGE with unequal sampling weights. These statistics are different from the estimates shown in Figure 1 where the summary statistics are computed without the sampling weights. In Figure 2, the standard error is higher using the sampling weights. Finally, Figure 3 displays the analysis using the stratified information provided in the MEPS data set. Compared to the results in Figure 2, the estimates of the mean are the same. However, the standard error is larger when the number of strata are specified. When the design is stratified, the SURVEYMEANS procedure pools stratum variance estimates to compute the overall variance estimate (Figure 3) and the degrees of freedom equals the number of PSUs minus the number of strata in the sample design.

In summary, analyst using complex survey data, such as MEPS, to address their research questions must use

the procedures designed to deal with complex survey data. Otherwise, they risk the possibility of obtaining incorrect results. In particular, using standard SAS procedures for survey data that have a large variability results are biased estimates and will not accurately represent the survey population.

The SURVEYMEANS Procedure					
Data Summary					
Number of Observations			4490		
Statistics					
Variable	N	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean
age	4490	32.166370	0.117263	31.936477	32.396262

Figure 1. Data Summary measures of variable age without the sampling weights

The SURVEYMEANS Procedure					
Data Summary					
Number of Observations		4490			
Sum of Weights		56838646			
Statistics					
Variable	N	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean
age	4490	31.967844	0.217365	31.541822	32.393934

Figure 2. Data Summary measures of variable age with the weights

The SURVEYMEANS Procedure					
Data Summary					
Number of Strata		42			
Number of Observations		4490			
Sum of Weights		56838646			
Statistics					
Variable	N	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean
age	4490	31.967844	0.443845	31.068181	32.8664207

Figure 3. Data Summary measures of variable age with the weights and the stratified information

Author Contact Information

Hossein N. Yarandi, Ph.D.*
Associate Professor
College of Nursing and Biostatistics Unit
University of Florida
Campus Box 100187
Gainesville, FL 32610-0187
Telephone: (353) 846-0658
Fax: (352) 846-1624
E-mail: yarandi@ufl.edu

Shawn M. Kneipp, Ph.D., ARNP
Assistant Professor
College of Nursing
University of Florida
Campus Box 100187
Gainesville, FL 32610-0187
Telephone: (353) 392-9207
Fax: (352) 846-1624
E-mail: skneipp@nursing.ufl.edu

* Correspondence should be directed.

References

- Cohen, S. (1997). An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. The American Statistician, 51, 293-299.
- Cohen, S., Burt, V., and Jones, G. (1986). Efficiencies for variance estimation for complex survey data. The American Statistician, 40, 157-164.
- Flyer, P. (1987). Finite population correction for replication estimates of variance. Proceedings of the Section on Survey Research Methods of the American Statistical Association, 732-736.
- Flyer, P., Morganstein, D., and Rust, K. (1989). Complex survey variance estimation and contingency tables analysis using replication. Proceedings of the Section on Survey Research Methods of the American Statistical Association, 110-119.
- Judkins, D. and Wright, D. (1990). National health interview survey: report on variance estimation. Rockville, MD: Westat.
- Korn, E. & Graubard, B. (1999). Analysis of Health Surveys, New York: John Wiley & Sons.
- MEPS HC-003 (1998). Population characteristics and utilization data for 1996, Agency for Health Care Policy and Research, Rockville, MD.
- SAS Institute Inc. (2000). SAS/STAT User's Guide, Cary, NC: SAS Institute.
- Wolter, K. (1985). Introduction to variance estimation, New York: Springer-Verlag.