Analyzing Complex Sample Survey Data: A New Beginning

Linda Tompkins and Arlene B. Siller
CDC/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 8's PROC SURVEYMEANS, and an alternative program, SUDAAN.

Highlights of the National Health Interview Survey (NHIS)

- Nationally representative sample
- Continuously conducted since 1957
- Personal interviews are conducted in the home
- Civilian, noninstitutionalized population
- Topics cover a broad range of health issues

Sample design

- Multistage area probability design, including clustering**,** stratification, and the assignment of unequal probabilities of selection
- 358 primary sampling units (PSUs) drawn from approximately 1,900 geographically defined PSUs covering the 50 States and the District of Columbia
- Households selected from Census -defined tracts and block groups

Comparison of software systems

The complex nature of the NHIS design causes 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. Otherwise, the standard errors produced, as for a simple random sample, would generally underestimate the true population value. These erroneous estimates negate the validity of resulting confidence intervals or tests of statistical significance. Both software systems use Taylor series linearization for estimating population characteristics from complex sample survey data.

SUDAAN (Research Triangle Institute) version 7.5

- Analytical tools include four descriptive statistics procedures: CROSSTAB, RATIO, DESCRIPT, and RECORDS, as well as four modeling procedures: REGRESS, LOGISTIC, MULTILOG, and SURVIVAL
- Procedure statements and syntax are similar to those in SAS®
- Procedures are available in a stand-alone package, as well as procedures that are "callable" from within the SAS program. (Note: Although both packages are available on many computing platforms, for example, mainframe, PC, and Unix, RTI plans no further mainframe development after Version 7.0)
- Callable version has been developed for PC SAS Version 8

SAS (SAS Institute) version 8

- Now includes three experimental procedures designed to analyze data derived from a complex sample survey:
  1) PROC SURVEYSELECT provides methods for selecting probability-based samples, while simultaneously employing clustering, stratification, and unequal probabilities of selection
  2) PROC SURVEYMEANS computes estimates of the survey population means, totals, and the associated standard errors
  3) PROC SURVEYREG performs regression analysis for survey data
- Version 7 experimental procedures became production products with Version 8 of the SAS System

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 are 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 indicated on a LEVELS statement. Estimated percents produced by PROC MEANS are identical to those produced by PROCs SURVEYMEANS and CROSSTAB. However, the standard errors are smaller, will create narrower confidence intervals, and will impact any tests of statistical significance. Results from PROCs SURVEYMEANS and CROSSTAB are comparable.

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 required 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 study.

Table 1: Comparison of SAS and SUDAAN percents and standard errors

*Demographic characteristics*** *(percent of respondents)*	*SAS Version 6.12 PROC MEANS*		*SAS Version 8 PROC SURVEYMEANS*		*SUDAAN Version 7.5 PROC CROSSTAB*	
Elderly respondents (aged 65 years and over)	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)

** NOTE: 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. 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 characteristics (means)	*SAS Version 6.12 PROC MEANS*		*SAS Version 8 PROC SURVEYMEANS*		*SUDAAN Version 7.5 PROC CROSSTAB*	
Age of respondent	44.5486	(.0917)	44.5486	(.1505)	44.5486	(.1505)
Body mass index	26.2164	(.0282)	26.2164	(.0376)	26.2164	(.0376)
Frequency of vigorous activity (times per week)	1.4085	(.0137)	1.4085	(.0187)	1.4085	(.0187)

Two strictly hypothetical simple linear regressions were run in both packages for illustrative purposes only:
1) Hypertension = age sex race smoking body mass index
2) Body mass index = age sex race vigorous exercise
The regression estimates and their standard errors are shown below.

Table 3: Comparison of SAS and SUDAAN regression coefficients and standard errors

Independent and dependent variables		*SAS Version 6.12 PROC REG*		*SAS Version 8 PROC SURVEYREG*		*SUDAAN Version 7.5 PROC REGRESS*	
		Beta	*SE Beta*	*Beta*	*SE Beta*	*Beta*	*SE Beta*
1) Hypertension intercept		-0.548	0.016	-0.552	0.015	-0.552	0.015
Age		0.009	0.000	0.009	0.000	0.009	0.000
Sex	**Male**	-0.021	0.004	-0.021	0.005	-0.021	0.005
	Female	0.000	0.000	0.000	0.000	0.000	0.000
Race	**White**	0.005	0.009	0.005	0.008	0.005	0.008
	Black	0.087	0.011	0.087	0.010	0.087	0.010
	Other	0.000	0.000	0.000	0.000	0.000	0.000
Smoking	**Yes**	0.002	-0.005	0.002	0.005	0.002	0.005
	No	0.000	0.000	0.000	0.000	0.000	0.000
Body mass index		0.014	0.000	0.014	0.001	0.014	0.001
2) Body mass index intercept		23.661	0.141	23.661	0.170	23.661	0.170
Age		0.019	0.002	0.019	.0002	0.019	.0002
Sex	**Male**	1.130	0.056	1.130	0.066	1.130	0.066
	Female	0.000	0.000	0.000	0.000	0.000	0.000
Race	**White**	1.180	0.125	1.180	0.146	1.180	0.146
	Black	2.857	0.147	2.857	0.176	2.857	0.176
	Other	0.000	0.000	0.000	0.000	0.000	0.000
Vigorous exercise		-0.123	0.011	-0.123	0.013	-0.123	0.013

Sample Code

```
/* member(v6freqs.sas)    */
libname  in1 'c:\nchssug\sas';
libname library 'c:\nchssug\sas\formats';
run;
data new;
set in1.adults;
run;

   /* Establish format library */
proc format library=library;
run;
```

SAS Callable SUDAAN version 7.5

```
 /* Run frequencies and means in SUDAAN */
/* Must first sort dataset by NEST
  variables    */
proc sort data=new;
by stratum psu;
run;
  /* Discrete variables */
proc crosstab design=wr;
nest stratum psu;
subgroup sex origin racerec r_maritl
region ;
levels 2 3 3 9 4;
tables sex origin racerec r_maritl region;
weight fwgt;
setenv colwidth=16 decwidth=5;
title 'PROC CROSSTAB – Statistics for
Discrete Variables';
run;
  /* Continuous variables */
proc descript design=wr ;
nest stratum psu ;
var age_p bmi vig ;
weight fwgt ;
setenv colwidth=16 decwidth=5;
title 'PROC DESCRIPT – Statistics for
Continuous variables';
run;
```

SAS Version 6.12

```
/* Run frequencies and means in Version
   6.12 */
proc freq;
tables sex origin racerec r_maritl region;
format sex sexf. origin yesnof. racerec
racerf. region regionf.;
weight fwgt;
title 'PROC FREQ – Frequency Tables for
Discrete Variables';
title2 'Version 6.12';
run;
proc means n mean sum stderr std;
var age_p bmi vig;
weight fwgt;
title 'PROC MEANS – Statistics for
Continuous Variables';
title2 'Version 6.12';
run;
```

SAS Version 8

```
/* Run frequencies and means in version 8
   */
 /*  Discrete or continuous variables */
proc surveymeans data=new;
class newage origin racerec newmar region;
strata stratum;
cluster psu;
weight fwgt;
format sex sexf. origin yesnof. racerec
racerf. r_maritl mstatf. region regionf.;
title 'PROC SURVEYMEANS Example';
title2 'Version 8 Weighted';
run;
```


For more information

CONTACT:
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

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