Paper 172-31

The Big Four: Analyzing Complex Sample Survey Data Using SAS®, SPSS®, STATA®, and SUDAAN®

Arlene B. Siller and Linda Tompkins

U.S. Department of Health and Human Services, Centers for Disease Control and Prevention,
National Center for Health Statistics

Abstract

The National Center for Health Statistics (NCHS) conducts surveys with probability-based complex sample designs to produce estimates of health conditions, vital statistics, medical establishment characteristics, disparities in health status, and use of health care by race and ethnicity, socioeconomic status, and other population traits in the United States. Researchers usually present descriptive statistics such as means and 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 four statistical packages (SAS, SPSS, STATA, and SUDAAN) using data from two NCHS surveys.

Introduction

The complex design of sample surveys dictates that data analysis procedures be able to account for multiple stages of sampling, stratification, and clustering. NCHS conducts such surveys to produce estimates of health conditions and access to medical care for the U.S. population. These usually include descriptive statistics such as means and totals and their standard errors. The accompanying standard errors must be computed using procedures that take into account the complex nature of the sample design. In recent years, statistical software designers have increased their efforts to incorporate such procedures. This paper compares estimates produced with four well-known statistical packages using data from the National Ambulatory Medical Care Survey (NAMCS) and the National Survey of Family Growth (NSFG), presents the results of the comparison, and provides appropriate code for each package.

National Ambulatory Medical Care Survey (NAMCS)

- Administered to a nationally representative sample of visits to nonfederally employed office-based physicians who are primarily engaged in direct patient care
- Conducted periodically since 1973 and continuously since 1989
- Collected via patient visit encounter forms completed by the physician
- Provides statistics on the demographic characteristics of patients and services provided, including information on diagnostic procedures, patient management, and planned future treatment

National Survey of Family Growth (NSFG)

- Administered to a nationally representative sample of men and women aged 15–44 years in the civilian noninstitutionalized population of the United States
- Conducted periodically between 1973 and 2002
- Collected via personal interviews conducted in the home
- Provides data on marriage, divorce, contraception, and infertility

Complexity of sample designs

The multistage area probability designs of NAMCS and NSFG include clustering, stratification, and the assignment of unequal probabilities of selection to sample units. The complexity of these sample designs causes a departure from the assumption that independent sample points have equal probabilities of selection. Specialized statistical software is required 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, negating the validity of resulting confidence intervals or statistical significance tests. All software systems compared in this paper use the Taylor series linearization method for estimating population characteristics from complex sample survey data.

Highlights of survey software

The procedures listed below are designed to analyze data derived from a complex sample survey for each of the four packages: SAS, SPSS, STATA, and SUDAAN. The procedures or modules handle the following survey-design features: stratification, multiple stages of cluster sampling, probability sampling weights, and poststratification. In addition, each package has more extensive features for managing and processing data and for performing statistical procedures not explored here.

SAS Version 9

- **PROC SURVEYSELECT** allows selection of probability-based samples while employing clustering, stratification, and unequal probabilities of selection.
- **PROC SURVEYFREQ** produces one- to *n*-way frequency and crosstabulation tables and associated tests of independence.
- **PROC SURVEYMEANS** computes estimates of the survey population means and totals and associated standard errors.
- **PROC SURVEYLOGISTIC** investigates the relationship between discrete responses and a set of explanatory variables.
- **PROC SURVEYREG** estimates regression coefficients by generalized least squares, using element-wise regression, assuming that the regression coefficients are the same across strata and primary sampling units.

SPSS Version 13

- **CSPLAN module** specifies design information for sample selection or analysis and must be created for use by all other modules.
- **CSSELECT module** chooses units according to a sample design specified by CSPLAN.
- **CSDESCRIPTIVES module** estimates sums, means, and ratios with standard errors and design effects for whole populations or subpopulations.
- **CSTABULATE module** produces one- or two-way tabulations with standard errors, design effects, coefficients of variation, odds ratios, relative risks, and tests of independence.

- **CSGLM module** produces linear regression models including analysis of variance and covariance models.
- **CSLOGISTIC module** produces binary and multinomial logistic regression models with linear predictor specification options similar to CSGLM.

STATA Version 9

- **SVYSET** sets variables for data.
- **SVY:TABULATE** produces two-way tabulations.
- **SVY:MEAN** computes estimates of survey population means and totals and associated standard errors.
- **SVY:REGRESS** computes general linear regression models.
- **SVY:LOGIT** produces logistic regression models.
- Other STATA procedures for the analysis of complex sample data (all with the **SVY:** prefix) include GNBREG, HECKMAN, HECKPROB, INTREG, LVREG, MLOGIT, NBREG, OLOGIT, OPROBIT, POISSON, PROPORTION, RATIO, and TOTAL.

SUDAAN Version 9

- **CROSSTAB** computes frequencies, percent distributions, odds ratios, relative risks, and their standard errors (or confidence intervals) for tabulations.
- **DESCRIPT** produces estimates of means, totals, proportions, percentages, geometric means, quantiles, and their standard errors.
- **REGRESS** fits linear regression models to continuous outcomes and performs hypothesis tests for model parameters.
- **LOGISTIC** produces logistic regression models for binary data and computes hypothesis tests for model parameters and estimates odds ratios and their 95% confidence intervals for each model parameter.
- Other SUDAAN procedures for analysis of complex sample data include SURVIVAL, RATIO, and MULTILOG.

Conclusion

Analysis of complex sample survey data must take into account characteristics of the sample design, including stages of sample selection, clustering, stratification, and unequal probabilities of selection. The packages examined here—SAS, SPSS, STATA, and SUDAAN—produced identical results using the Taylor series linearization method. Therefore, other factors such as cost, availability of point-and-click operation, overall data management capabilities, and alternative methods of variance estimation will affect the user's choice of software.

Comparison results

Several discrete and continuous variables were selected from NAMCS and NSFG. The computed percentages and means and their associated standard errors (shown in tables 1–4) are identical across packages. **[NOTE: These data are presented for illustrative purposes only and should not be used to make inferences about the U.S. population.]**

Table 1. Percentages of selected demographic characteristics with standard errors: National Ambulatory Medical Care Survey, 2002

Discrete variable	*SAS*	*SPSS*	*STATA*	*SUDAAN*
	Percent (SE)	*Percent (SE)*	*Percent (SE)*	*Percent (SE)*
Sex				
Male	40.552 (0.589)	40.552 (0.589)	40.552 (0.589)	40.552 (0.589)
Female	59.448 (0.589)	59.448 (0.589)	59.448 (0.589)	59.448 (0.589)
Sex and race				
Male:				
White	86.675 (1.030)	86.675 (1.030)	86.675 (1.030)	86.675 (1.030)
Black	9.029 (0.786)	9.029 (0.786)	9.029 (0.786)	9.029 (0.786)
Other	4.296 (0.683)	4.296 (0.683)	4.296 (0.683)	4.296 (0.683)
Female:				
White	85.675 (1.191)	85.675 (1.191)	85.675 (1.191)	85.675 (1.191)
Black	10.749 (1.079)	10.749 (1.079)	10.749 (1.079)	10.749 (1.079)
Other	3.577 (0.507)	3.577 (0.507)	3.577 (0.507)	3.577 (0.507)
MSA:[1]				
MSA	85.978 (3.236)	85.978 (3.236)	85.978 (3.236)	85.978 (3.236)
Non-MSA	14.022 (3.236)	14.022 (3.236)	14.022 (3.236)	14.022 (3.236)
MSA[1] and race				
MSA:				
White	85.647 (1.183)	85.647 (1.183)	85.647 (1.183)	85.647 (1.183)
Black	10.135 (0.989)	10.135 (0.989)	10.135 (0.989)	10.135 (0.989)
Other	4.217 (0.665)	4.217 (0.665)	4.217 (0.665)	4.217 (0.665)
Non-MSA:				
White	88.735 (2.910)	88.735 (2.910)	88.735 (2.910)	88.735 (2.910)
Black	9.536 (3.007)	9.536 (3.007)	9.536 (3.007)	9.536 (3.007)
Other	1.730 (0.374)	1.730 (0.374)	1.730 (0.374)	1.730 (0.374)

[1] MSA is metropolitan statistical area.

NOTE: SE is standard error.

Table 2. Means of selected continuous characteristics with standard errors: National Ambulatory Medical Care Survey, 2002

Continuous variable	*SAS Mean (SE)*	*SPSS Mean (SE)*	*STATA Mean (SE)*	*SUDAAN Mean (SE)*
Age	43.947 (0.516)	43.947 (0.516)	43.947 (0.516)	43.947 (0.516)
Male	42.259 (0.596)	42.259 (0.596)	42.259 (0.596)	42.259 (0.596)
Female	45.090 (0.550)	45.090 (0.550)	45.090 (0.550)	45.090 (0.550)
Time (in minutes) with doctor	17.517 (0.328)	17.517 (0.328)	17.517 (0.328)	17.517 (0.328)
Male	17.636 (0.346)	17.636 (0.346)	17.636 (0.346)	17.636 (0.346)
Female	17.436 (0.354)	17.436 (0.354)	17.436 (0.354)	17.436 (0.354)

NOTE: SE is standard error.

Table 3. Percentages of selected demographic characteristics with standard errors: National Survey of Family Growth—Cycle 6, 2002

Discrete variable	*SAS Percent (SE)*	*SPSS Percent (SE)*	*STATA Percent (SE)*	*SUDAAN Percent (SE)*
Sex				
Male	49.831 (0.768)	49.831 (0.768)	49.831 (0.768)	49.831 (0.768)
Female	50.169 (0.768)	50.169 (0.768)	50.169 (0.768)	50.169 (0.768)
Sex and race				
Male:				
White	76.283 (1.103)	76.283 (1.103)	76.283 (1.103)	76.283 (1.103)
Black	13.474 (0.785)	13.474 (0.785)	13.474 (0.785)	13.474 (0.785)
Other	10.243 (0.625)	10.243 (0.625)	10.243 (0.625)	10.243 (0.625)
Female:				
White	76.531 (0.909)	76.531 (0.909)	76.531 (0.909)	76.531 (0.909)
Black	15.073 (0.753)	15.073 (0.753)	15.073 (0.753)	15.073 (0.753)
Other	8.396 (0.478)	8.396 (0.478)	8.396 (0.478)	8.396 (0.478)
Marital status				
Married	44.116 (0.830)	44.116 (0.830)	44.116 (0.830)	44.116 (0.830)
Widowed	0.271 (0.043)	0.271 (0.043)	0.271 (0.043)	0.271 (0.043)
Divorced	7.696 (0.326)	7.696 (0.326)	7.696 (0.326)	7.696 (0.326)
Separated	2.372 (0.134)	2.372 (0.134)	2.372 (0.134)	2.372 (0.134)
Never married	45.545 (0.847)	45.545 (0.847)	45.545 (0.847)	45.545 (0.847)

NOTE: SE is standard error.

Table 4. Means of selected continuous characteristics with standard errors: National Survey of Family Growth—Cycle 6, 2002

Continuous variable	*SAS Mean (SE)*	*SPSS Mean (SE)*	*STATA Mean (SE)*	*SUDAAN Mean (SE)*
Age	29.900 (0.153)	29.900 (0.153)	29.900 (0.153)	29.900 (0.153)
Male	29.827 (0.231)	29.827 (0.231)	29.827 (0.231)	29.827 (0.231)
Female	29.973 (0.172)	29.973 (0.172)	29.973 (0.172)	29.973 (0.172)
Number of babies born alive (female respondents)	1.291 (0.032)	1.291 (0.032)	1.291 (0.032)	1.291 (0.032)
Number of lifetime female sex partners (male respondents)	4.185 (0.060)	4.185 (0.060)	4.185 (0.060)	4.185 (0.060)

NOTE: SE is standard error.

Sample code for percentages and mean

Sample code examples do not include data set creation or any variable transformation necessary for analysis.

SAS®

```
proc surveyfreq;
stratum strata variable;
cluster cluster variable;
weight sample weight;
table  var1  var2*var3;
run;

proc surveymeans;
stratum strata variable;
cluster cluster variable;
weight sample weight;
var var4;
run;
```

NOTE:

- PROC SURVEYMEANS with a CLASS statement can also be used for discrete variables.

SPSS®

```
Cstabulate
/plan file =”disk location of CSPLAN file”
/tables variables = var1  var2
/cells desired cell statistics
/statistics desired estimates
/missing scope = table

csdescriptives
/plan file =”disk location of CSPLAN file”
/summary variables = var4
/subpop table = var2 desired domain
/mean
/statistics desired estimates
/missing scope = analysis
```

NOTES:

- CSPLAN file must be created before analysis. This file identifies the design strata, cluster, and sample weight information.
- Point-and-click method is used and code is captured from background.

STATA®

```
use “disk location of data file”
svyset [pweight=sample weight], strata(strata variable)  psu(cluster variable)
svy:tab  var1  var2,  desired cell statistics

use “disk location of data file”
svyset [pweight=sample weight], strata(strata variable)  psu(cluster variable)
svy:mean  var4,  desired cell statistics

use “disk location of data file”
svyset [pweight=sample weight], strata(strata variable)  psu(cluster variable)
svy:mean  var4, over(var5)  desired cell statistics
```

NOTES:

- Be careful to differentiate between brackets [ ] and parentheses ( ).
- Use **over** for domain analyses.

SUDAAN®

```
proc sort;
by strata variable;
cluster PSU variable;

proc crosstab;
nest   strata variable    cluster variable;
subgroup  var1  var2  var3;
levels  number of levels for each discrete variable;
weight sample weight;
table  var1  var2*var3;
run;

proc descript;
nest   strata variable    cluster variable;
weight sample weight;
table  var4;
run;
```

NOTE:

- Data set must be sorted by **NEST** variables.

Information resources

Ambulatory Health Care Data.
http://www.cdc.gov/nchs/about/major/ahcd/ahcd1.htm.

National Survey of Family Growth.
http://www.cdc.gov/nchs/nsfg.htm.

Cochran WG. Sampling techniques, 3d ed. New York:
John Wiley & Sons. 1977.

Wolter KM. Introduction to variance estimation. New York:
Springer-Verlag. 1985.

Acknowledgments

SAS is a Registered Trademark of the SAS Institute Inc.,
Cary, North Carolina.

SPSS is a Registered Trademark of SPSS Inc., Chicago, Illinois.

STATA is a Registered Trademark of STATACorp LP,
College Station, Texas.

SUDAAN is a Registered Trademark of the Research Triangle Institute,
Research Triangle Park, North Carolina.

Contact information

Arlene B. Siller
Centers for Disease Control and Prevention
National Center for Health Statistics
3311 Toledo Road, Room 3317
Hyattsville, MD 20782
(301)458-4498
Fax: (301)458-4032
Email: ASILLER@CDC.GOV

Linda Tompkins
Centers for Disease Control and Prevention
National Center for Health Statistics
3311 Toledo Road, Room 3115
Hyattsville, MD 20782
(301)458-4533
Fax: (301)458-4031
Email: LIT3@CDC.GOV