

THE QUANDARY OF SURVEY DATA: Comparison of SAS[®] Procedures and SUDAAN[®] Procedures

Katherine Baisden, SRI International, Menlo Park, California

ABSTRACT

Have you ever worked with survey data that are based on a stratified, clustered or complex sample design? Such designs impact the accuracy of variance estimates (e.g., standard error of the mean) and test statistics (e.g., chi-squares). Until recently, SAS programmers had to rely on other statistical software packages, such as SUDAAN[®], WesVar[®], and STATA[®], to produce accurate variance estimates and test statistics from complex sample designs. With the release of SAS V8.2 and the upcoming release of SAS V9.0, SAS has incorporated survey procedures (e.g., Proc Surveymeans, Proc Surveyfreq) to address this issue. This paper examines the differences between the Proc Surveymeans and Proc Surveyfreq procedures and the corresponding SAS[®]-callable SUDAAN v8.0 procedures. Using data examples, the paper highlights the differences in syntax and output. It will also discuss the available options and limitations of each package.

INTRODUCTION

To maximize the effort of survey data collection and to minimize the cost, researchers continue to develop increasingly complex sample designs. These designs include stratification, clustering, unequal probabilities of selection, and a multitude of the combinations of all these techniques. Simple random sample designs are a rarity in this day and age of survey research. These complex designs impact the accuracy of variance estimates and test statistics. The SAS programmer must expand beyond the traditional tools in his/her analytical handbag to deal with survey data today. Until recently, SAS programmers had to use additional software packages, such as SUDAAN, to produce correct variance estimates. Now, with SAS v8.2 (PROC SURVEYSELECT, PROC SURVEYMEANS AND PROC SURVEYREG), and with the pending release of SAS v9.0 experimental procedure SURVEYFREQ, some of the tools needed to deal with this type of survey data are available in SAS.

This paper compares and contrasts SAS and SAS-callable SUDAAN focusing on syntax and output for two of the most common procedures used in analysis; the crosstabulation/frequency and means procedure. This will be demonstrated using data from a study of teachers in the state of California. Schools were classified on three criteria: the percentage of emergency credentialed teachers in the school (EMERG: 1e 10%, 11-19%, 20%+), the size of the district (DISTSIZ: less than 5000, 5001-10,000, 10,000+) and the type of school (SCHL_LVL: elementary, middle and high school). Weights were developed for the data based on these stratification variables. Teachers were then selected from each of the strata. For analysis purposes, our statistician has classified this as a stratified sample with replacement.

The data examples will give you a highlight of the syntax (not all options can be included) for each of the procedures in SAS-callable SUDAAN and SAS.

GENERAL COMMENTS ABOUT SAS AND SAS-CALLABLE SUDAAN

Both SAS and SUDAAN procedures are based on the Taylor linear approximation method to calculate the variance estimates. However, SUDAAN does offer the option of using balanced repeated replicates (BRR) and jackknife weights. SUDAAN does not have the capability to calculate BRR or jackknife weights, but can use them if they are provided on the data set. The SAS-callable version of SUDAAN is designed to use within the framework of SAS. Within any SAS program you call into play SUDAAN and it uses the SAS dataset format. Thus, much of the syntax of a procedure is very similar. However there are some important differences to note.

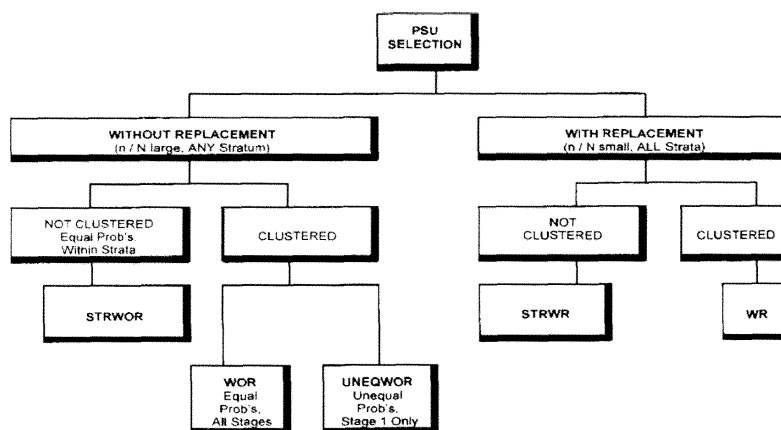
This paper will focus on PROC SURVEYMEANS and PROC SURVEYFREQ. Because PROC SURVEYFREQ is an experimental procedure at this time, there is no documentation available on it. The examples had to be restricted due to lack of access to documentation and all options of the procedure. Each procedure has many options and statistics available in each package, however due to space and limitations, this paper will highlight the most common.

DETERMINING THE IMPACT OF SAMPLE DESIGNS ON VARIANCE ESTIMATES

There is a way to measure the impact of complex sample designs (CSD) on variance estimates. A common measure is called the Design Effect (DE). The DE is a ratio. It takes the variance from the CSD and compares it to the variance that would have occurred under the assumption of simple random sampling (SRS). If the DE is close to 1.0 then one can assume the variances would have come out the same whether it was a CSD or a SRS design. Most of the time, the DE for a CSD is greater than one. The larger the DE, the more correlated are your respondents within clusters, leading to underestimated variances if analyzed with packages without the capabilities to go beyond the assumption of SRS.

$$DE = \text{variance of CDS} / \text{variance of SRS}$$

Before beginning any analysis it must be determined on what kind of sampling design the survey is based. SUDAAN offers you a choice of the following:



SAS and SUDAAN offer the following procedures:

SAS	SUDAAN	PURPOSE
	RECORDS	Print records from ASCII, SAS®, SPSS and SUDAAN
SURVEYFREQ	CROSSTAB	Produces weighted oneway and multiway frequencies
	RATIO	Produces ratio estimates and their standard errors for correlated data
SURVEYMEANS	DESCRIPT	Produces means, medians and quantiles and their standard errors
SURVEYREG	REGRESS	Fits linear models
	RLOGISTIC	Fits logistic regression models
	MULTILOG	Logistic model with categorical dependent variables
	SURVIVAL	Fits the discrete proportional hazards model
SURVEYSELECT		Helps you select a sample

The majority of my knowledge about these procedures comes from self-discovery and hands-on experience. Although both programs use very similar syntax, SUDAAN requires more detail. For example SUDAAN requests that, for every variable in the syntax you specify the number of levels in each variable (using the LEVELS statement). In addition, SUDAAN will not accept 0,1 coding schemes when dealing with categorical values; all

values for categorical variables must start with a 1. You do have the option to recode your variables on the fly within a SUDAAN procedure but it is another step that must be taken for successful completion of a procedure. Likewise, there is no default printing of output for SUDAAN. You must specify exactly what statistic you want printed and in what format. It is not as simple as requesting statistics on an OPTIONS statement within a SAS procedure. Unlike SAS, SUDAAN also does not provide the variable names in the output unless they are specified in the label of the variable.

SAS assumes that first-stage sampling is with replacement although reality bears witness that the vast majority of the time it is not. This can result in a slight overestimate of the variance, but this is very small.

PROC SURVEYMEANS IN SAS

```

Proc Surveymeans; Var T4B;
  Strata emerg distsize schl_lvl;
  Weight WGT1;
  Domain T40;
Run;
  
```

This analysis is requesting the overall mean of T4B (number of classes taught) and the mean for number of classes taught for each gender (T40). The stratification variables are EMERG, DISTSIZE, and SCHL_LVL as indicated on the STRATA statement. The DOMAIN statement indicates a breakdown of T4B by gender. Without specifying any statistic keywords, SAS provides the NOBS, MEAN, STDERR and CLM statistics by default. A /LIST option will provide basic information about (N, number of missing, strata variable levels) respondents in each stratum. (not presented here) Output is in Exhibit 1.

As in the PROC MEANS, when computing statistics for an analysis variable, SAS omits observations with missing values for that variable. In addition, it is important to note that in PROC SURVEYMEANS, if an observation has a missing value or non-positive value for the weight it will be excluded from the analysis. Observations are also excluded if there are missing values on the STRATA or CLUSTER statement, unless the MISSING option is used.

When the MISSING option is used with categorical data, the missing values are treated as a valid category.

As an experienced SAS programmer, you may want to sort the data set by T40 (gender) and use a BY statement. That method will produce a NOTE from SAS requesting a DOMAIN statement.

PROC DESCRIPT IN SUDAAN

(Overall Mean)

```
Proc Descript data=one filetype=SAS design=strwr;
Nest emerg distsize schl_lvl;
Weight WGT1;
Var T4B;

Setenv labwidth=28 colspce=1 colwidth=10
decwidth=4;
Print nsum="Sample Size" Wsum="Population size"
Mean semean="S.E."
Deffmean="Design effect" / style=nchs nsumfmt=f6.0
wsumfmt=f10.0
Deffmeanfmt=F6.2 semeanfmt=F7.4;
Rtitle "Mean of T4B";
Run;
```

(Mean by Gender)

```
Proc Descript data=one filetype=SAS design=strwr;
Nest emerg distsize schl_lvl;
Weight WGT;
Var T4B;
Subgroup T40;
Levels 2;
```

```
Setenv labwidth=28 colspce=1 colwidth=10
decwidth=4;
Print nsum="Sample Size" Wsum="Population size"
Mean semean="S.E."
Deffmean="Design Effect" / style=NCHS nsumfmt=f6.0
wsumfmt=f10.0 Deffmeanfmt=F6.2
Semeanfmt=F10.4;
Rtitle "Mean of T4B by T40";
Run;
```

This analysis is the same request as presented in the PROC SURVEYMEANS in the preceding section. The STWR design was used to correspond with the SAS assumptions. In SUDAAN the name of the procedure is DESCRIPT. You must specify the filetype and the design. The NEST statement is similar to the STRATA statement in SAS. The SUBGROUP statement corresponds to the DOMAIN statement in SAS; however, you must include a LEVELS statement indicating the number of levels for the variable. Design effects can be requested in SUDAAN, this is not true for the PROC SURVEYMEANS procedure. Output is in Exhibit 1.

The SETENV statement sets the output environment parameters, similar to the options statement in SAS. The PRINT statement is the place where you have to indicate

each statistic and a label for those statistics that you want in the output. The STYLE option is a particular way SUDAAN prints the output. NCHS style is printed according to the standards of the National Center for Health Statistics. Before you are done you must give a format for each statistic. If you have not given a large enough format, an "***" will appear in the output. You must then go back and change the format for that specific variable. The RTITLE statement is equivalent to the TITLE statement in SAS.

Unlike SAS, you have to execute the PROC DESCRIPT twice in order to get an overall mean of T4B and the separate means for T4B by gender (T40).

SUDAAN handles missing values very much like SAS. Observations that have missing values for weights and required sample design variables will be excluded from the analysis.

In both programs the point estimates and the standard errors are the same (within reasonable rounding error).

PROC SURVEYFREQ IN SAS

```
Proc Surveyfreq;
Strata emerg distsize schl_lvl;
Tables T40*T6 / chisq;
Weight WGT1;
TITLE 'Crosstab of T40 by T6';
Run;
```

This syntax is very similar to PROC FREQ in SAS. It is a crosstabulation of gender (T40) and T6 (Did respondent leave a teacher prep program for employment?). There is an addition of a STRATA statement indicating the stratification variables. When requesting a chi-square analysis with this procedure you will get a Rao-Scott chi-square test, which applies a design effect correction to the Pearson chi-square. Due to technical problems, I was unable to get PROC SURVEYFREQ to execute. SAS has diligently worked with me to try to resolve the problem, however at this time, I am unable to provide a data example. Hopefully, by the time of the presentation, I will be able to provide a data example of this procedure. Theoretically, the point estimates will not differ.

I look forward to the introduction of PROC SURVEYFREQ as a functional procedure because one-way and multiway frequencies are the mainstay of any analysis plan. Some of the options that should become available include the Rao-Scott chi-square, confidence limits for the percentages, Rao-Scott likelihood chi-square test and design effects.

PROC CROSSTAB IN SUDAAN

```
Proc Crosstab data=one filetype=SAS design=strwr;
Nest emerg distsize schl_lvl;
Weight WGT1;
Subgroup T40 T6;
Levels 2 2;
```

Tables T40*T6;

```
Setenv colwidth=9 decwidth=2 colspce=2;  
Print nsum wsum colper rowper totper  
  /wsumfmt=f9.0 nsumfmt=f9.0 cmhtest=all tests=all  
cmhfmt=f8.2  
Cmhdffmt=f8.0 cmhpvalfmt=f8.4 chisqfmt=f11.2;  
Rtitle "Crosstab of T40 by T6";  
Run;
```

The PROC CROSSTAB in SUDAAN follows the logic of the syntax presented in the PROC DESCRIPT. You must supply a DESIGN statement and a NEST statement. Besides specifying the crosstabulation in the TABLES statement, you must have a SUBGROUP statement and a corresponding LEVELS statement. SUDAAN will produce several types of chi-square tests including the Cochran-Mantel-Haenszel and the Pearson. Output is in Exhibit 2.

The crosstabulation output prints out the totals on the left, reversed from the traditional SAS output. One of the disadvantages of SUDAAN output is that it produces a single page for every table and every test statistic you request. It is not environmentally friendly.

LIMITATIONS OF EACH PACKAGE

One of the major limitations at this time in SAS is that the package does not offer the option of using balanced repeated replicates (BRR) or jackknife weights. Why is this so important? It is very common as a programmer/analyst to inherit data sets or do secondary dataset analysis. In many cases we do not have access to the actual formation of the sampling design. It is essential, especially in SUDAAN to be able to designate the sample design based on such information. With the use of balanced repeated replicates or jackknife weights, the syntax does not require any further information other than the supplied weights. This makes it much more useable for the end-user.

Although SUDAAN offers more options in terms of survey sampling designs and procedures, it is a cumbersome program to code. SUDAAN documentation is not the easiest to comprehend, especially if you are a novice. From an economic viewpoint, using SUDAAN is an additional expense in terms of licensing and training. SAS gives you the ease of coding and more print control of output, however at this time it is very limited in what it offers a user in terms of design and procedures.

CONCLUSION

Simple random sampling is like a rare gem in this day of social science research. We are dealing with increasingly more complex sample designs. These designs require the sophistication of SAS survey procedures or SUDAAN procedures. One must balance variety of choice with ease of coding. At this time SUDAAN is the most desirable package to use because of the variety of choice it offers in sample designs and the number of procedures available to

analyze the data. However, it is a program that is cumbersome to program, creating a more labor-intensive task than its counterpart in SAS. My conclusion is that SAS is moving in the right direction and I hope to see it incorporate the power of SUDAAN in terms of choice and number of procedures in the future. With the incorporation of these survey-based procedures in SAS, we look forward to greater ease in coding when dealing with complex sample designs. In the interim, if one has a design that fits the parameters of SAS design and statistical options offered now, welcome to automatic transmission. If your sample cannot meet the parameters of what SAS offers now, then you must contend with the manual transmission mode of SUDAAN. In the interim, we will have to switch back and forth between the two packages depending on our individual needs.

ACKNOWLEDGEMENTS

I wish to thank the Center for Education Policy at SRI International for their support in letting me learn and expand into the complex sample design programming. Special thanks goes to Andrea Lash for her mentoring and support. I also want to thank Hal Javitz for his technical assistance. A thanks goes to my fellow programmers, Peter Godard and Kathryn Valdes for their comments. A debt of gratitude to Betsy Davies-Mercier for editorial assistance. An overdue thanks to my husband, Rob Robbins for enduring late nights and lonely meals.

REFERENCES

- An, Anthony and Donna Watts (1998), "New SAS Procedures for Analysis of Sample Survey Data", Cary, NC, SUGI Proceedings 23rd, SAS Institute.
- Cassell, David L. and AnnMaria Rousey (2003), "Complex Sampling Designs Meet the Flaming Turkey of Glory", Design Pathways and Spirit Lake Consulting, SUGI Proceedings 28th, Seattle, WA.
- Research Triangle Institute (2001). SUDDAN User's Manual, Release 8.0, Research Triangle Park, NC: Research Triangle Institute.
- SAS[®] Institute, Inc., SAS/STAT User's Guide, Version 8, Volumes 1,2,3, Cary, NC: SAS Institute Inc., 1999. 3884PP.

CONTACT INFORMATION

Katherine Baisden
SRI International
333 Ravenswood Ave, BS381
Menlo Park, CA 94025
Phone: (650) 859-5944
Fax: (650) 859-3375
katherine.baisden@sri.com

EXHIBIT 1
The SURVEYMEANS Procedure

SURVEYMEANS T4B (# OF CLASSES) BY T40 (GENDER)

Number of Strata 27
Number of Observations 441
Sum of Weights 289707.2

Statistics

Variable	N	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean
T4b	423	2.852893	0.056202	2.742401	2.963385

Domain Analysis: T40

T40	Variable	N	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean
(1) FEMALE	T4b	319	2.373224	0.092859	2.190660	2.555788
(2) MALE	T4b	100	4.332771	0.226274	3.887909	4.777633

S U D A A N

Software for the Statistical Analysis of Correlated Data
Copyright Research Triangle Institute January 2003
Release 8.0.2

(OVERALL MEAN)

Number of observations read : 441 Weighted count : 289707
Denominator degrees of freedom : 414

Variance Estimation Method: Taylor Series (STRWR)

Mean of T4b
by: Variable, One.

Variable	Sample Size	Population size	Mean	S.E.	Design effect
T4b	423	276217	2.8529	0.0611	0.37

(MEAN BY GENDER)

Number of observations read : 441 Weighted count : 289707
Denominator degrees of freedom : 414

Variance Estimation Method: Taylor Series (STRWR)

Mean of T4b by T40 by: Variable, T40:GENDER.

Variable	T40:GENDER	Sample Size	Population size	Mean	S.E.	Design effect
T4b:TOTAL NUMBER OF CLASSES TAUGHT						
Total		419	274537	2.8625	0.0617	0.38
(1) FEMALE		319	205989	2.3732	0.0943	0.79
(2) MALE		100	68548	4.3328	0.2263	1.52

EXHIBIT 2
S U D A A N

Number of observations read : 441 Weighted count : 289707
Denominator degrees of freedom : 414

Variance Estimation Method: Taylor Series (STRWR)
Crosstab of T40 by T6
by: T40:GENDER, T6:LEAVE MA OR PREP PGM FOR FT PAID POSITION.

T40:GENDER		T6:LEAVE MA OR PREP PGM FOR FT PAID POSITION		
		Total	1 (YES)	2 (NO)
Total	Sample Size	386	20	366
	Weighted Size	251079	12686	238393
	Col Percent	100.00	100.00	100.00
	Row Percent	100.00	5.05	94.95
	Tot Percent	100.00	5.05	94.95
(1) FEMALE	Sample Size	294	14	280
	Weighted Size	189630	10622	179008
	Col Percent	75.53	83.73	75.09
	Row Percent	100.00	5.60	94.40
	Tot Percent	75.53	4.23	71.30
(2) MALE	Sample Size	92	6	86
	Weighted Size	61449	2063	59385
	Col Percent	24.47	16.27	24.91
	Row Percent	100.00	3.36	96.64
	Tot Percent	24.47	0.82	23.65

Pearson Chisq: 0.72
P-value of Pearson: .40
DF 1.00

Cochran-Mantel
Haenszel Chisq: 0.72
P-value of CMH 0.3955
DF 1.00