

Paper ST-141

Complex Survey Data Analysis in SAS® and the Link between PROC SURVEYFREQ and SURVEYMEANS

Taylor Lewis, University of Maryland, College Park, MD

ABSTRACT

This paper highlights two commonly used, closely related survey data analysis procedures for reporting and testing significance, SURVEYFREQ and SURVEYMEANS. They are counterparts to the more familiar PROC MEANS and FREQ, to be employed under complex survey designs, such as when stratification or clustering is involved. Issues and caveats related specifically to survey data analysis which may not be obvious to everyday programmers will be discussed. The situation where SURVEYMEANS and SURVEYFREQ can be substituted for each other is illustrated. SAS's survey analysis capabilities are compared with other statistical software packages, and the paper concludes with a look ahead to what is new in SAS's SURVEY procedures with SAS 9.2.

INTRODUCTION

PROC MEANS and PROC FREQ are two common SAS procedures for data exploration and statistical analysis of a continuous or categorical variable(s), respectively. You need to be aware, however, that the underlying assumptions these procedures maintain when calculating variances is that data come from a simple random sample (SRS) of an infinite population. We will see through a hypothetical situation of a fictitious company referred to herein as Alpha Corporation how a complex sample survey design can quickly nullify these assumptions. In such instances, you must call upon the SURVEY family of SAS procedures to correctly account for the survey design effects on standard errors—namely, this paper discusses amending SAS code with PROC SURVEYMEANS and PROC SURVEYFREQ..

THE FINITE POPULATION CORRECTION (FPC)

For sake of an example, consider Alpha Corporation's desire to determine the mean bonus amount employees received at the end of the 2007 work year along with a measure of overall job satisfaction. Year-end bonus is clearly continuous, but satisfaction is categorical, since a five-point response scale was provided ranging from "Very Satisfied" to "Very Dissatisfied." It is of interest to management whether bonus amount is associated with a higher satisfaction rating, corroborating or refuting a notion that higher pay is linked to an increase in job satisfaction. To test this, management surveys a random sample of 1,000 of the company's 20,000 employees.

This is the traditional SRS setup, with each employee having a $1,000 / 20,000 = .05$ inclusion probability. From the data retrieved, the mean $\bar{x}$ and standard error of the mean $se(\bar{x})$ of the year-end bonus is

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \text{and} \quad se(\bar{x}) = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

SAS will calculate this equivalently in PROC MEANS or PROC SURVEYMEANS. Assume the survey dataset with the 1,000 respondents is called ALPHA and the bonus variable named BONUS. Both procedures below produce the same mean and standard error:

```
proc means data=alpha mean stderr;
  var bonus;
run;
proc surveymeans data=alpha mean stderr;
  var bonus;
run;
```

But in many certain survey situations, when the proportion surveyed of the population is known and not negligible, there is technically an additional component that should be added the standard error equation known as the *finite*

population correction, or FPC. The FPC is defined as $\left(\frac{N-n}{N}\right) = \left(1 - \frac{n}{N}\right)$, where N = population size and n = sample size, and the standard error equation becomes

$$se(\bar{x}) = \sqrt{\frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \left(1 - \frac{n}{N}\right)}$$

Notice that the FPC has a range of [0, 1) and will continually decrease the standard error as a function of the sampling fraction. In the case of $n = N$, whereby you census the population, the FPC is 0 and thus the standard error is zero.

So with a sample size $n = 1,000$ from a population $N = 20,000$, the true standard error is slightly smaller when we incorporate the FPC. SAS offers two ways to define this correction in the PROC SURVEYMEANS statement. The first is by specifying TOTAL=<number>; the second by specifying RATE=<decimal>. For instance, you can add TOTAL=20000 or RATE=.05. Either method supplies SAS sufficient information to apply the correction.

STRATIFICATION AND WEIGHTING

A *complex survey design* is merely defined as one that is not SRS—that is, not every population element has known, equal selection probability. To demonstrate, we extend the Alpha Corporation's survey project to include a stratification scheme. The purpose of stratification is to presort sample units into non-overlapping groups, or strata, then perform an independent SRS within each. This technique is frequently used to ensure adequate sample sizes for certain domains of interest, often for preset precision goals. Moreover, when strata delineate homogeneous groupings of sample units, the overall estimates may be more precise. A complicating byproduct of this, however, is that sampling rates may vary across strata. When this occurs, the onus lies with the SAS programmer to calculate and provide the SURVEY procedure a weight variable—this will be illustrated shortly.

Assume the personnel database from which the Alpha Corporation's sample is drawn contains a variable indicating whether the employee is a non-supervisor, supervisor, or executive. Stratifying on this variable prior to sampling allows for sufficient sample sizes of estimates for each supervisory level and potentially decreases the variance of a companywide mean bonus estimate. Suppose the following sample plan is implemented:

Table 1. Alpha Corporation Stratified Sampling Scheme

Stratum	Supervisory Status	N	n	Rate	Weight $w_i = 1/\text{Rate}$
1	Non-Supervisor	17,500	2,500	0.143	7.0
2	Supervisor	2,000	1,000	0.500	2.0
3	Executive	500	200	0.400	2.5
Totals		20,000	3,700	0.185	

If data is collected in this fashion, you can no longer use the simplified SURVEYMEANS code shown previously; you must first inform SAS of the survey design information by issuing a STRATA statement that identifies the variable(s) in the analysis dataset indicating from which stratum the response belongs. If the supervisory status indicator variable is stored as SUPER in the ALPHA SAS dataset, the following amendment correctly accounts for the stratification scheme:

```
proc surveymeans data=alpha mean stderr;
  var bonus;
  strata super;
run;
```

The second task is to specify the FPC, which can no longer be supplied as a sole number or rate, but as three, one for each stratum. The standard error of a mean estimate encompassing data from multiple strata involves a pooled estimate of the individual stratum variances, so one number or rate would not suffice. Instead, you must supply a secondary dataset to the TOTAL= option in the PROC statement with the like-named STRATA variable and key variable _TOTAL_, which should contain the total population size, N , for each stratum.

Returning to Alpha Corporation's stratified sample scheme, a dataset named TOTALS was created with three observations and two variables: SUPER and _TOTAL_. The new SURVEYMEANS code is as follows:

```
proc surveymeans data=alpha total=totals mean stderr;
  var bonus;
  strata super;
run;
```

Comparably, a dataset with key variables SUPER and _RATE_ could be supplied to the RATE= option in the PROC statement.

The third task is to adjust for unequal sampling rates by creating a weight variable in the survey dataset equal to 1 divided by the sampling rate for each stratum. This particular weight is referred to in the survey literature as the *base weight*, or *sampling weight*, and can be thought of as a way to make the results more representative of the overall population when sampling rates vary. You can also conceptualize a weight to be the number of population elements the survey respondent speaks for: considering the non-supervisors, each of the 2,500 respondents represents himself or herself and 6 other non-supervisors, for a total weight of 7.

A naive programmer may expect SAS to calculate the base weights behind the scenes considering all pieces can be gathered from the survey dataset and TOTAL= secondary dataset, but SAS only issues the following note to the log when no WEIGHT statement is provided:

```
NOTE: You are using unequal sampling rates in a stratified design but did not
specify a WEIGHT statement. Unless you also specify a WEIGHT statement, the
analysis will assume equal weights for all observations.
```

The variable BASEWGT was defined in the ALPHA dataset as 7.0 for non-supervisors, 2.0 for supervisors, and 2.5 for executives, as per Table 1 above. At last, the following SAS code represents the correct way to analyze the bonus variable collected from Alpha Corporations stratified sample survey:

```
proc surveymeans data=alpha total=totals mean stderr;
  var bonus;
  strata super;
  weight basewgt;
run;
```

A few caveats to weighting: the base weight is appropriate in the absence of nonresponse and undercoverage, which we assume in this hypothetical scenario. A wide body of literature exists on methods to adjust base weights for nonresponse and undercoverage. For a summary, see Kalton and Kasprzyk (2003). Note also that if we sampled in equal proportion from each stratum, we would have a self-weighting design, in which assigning differential weights would not be required. This partially explains why no WEIGHT statement was required in the SRS design—all sample units maintained equal selection probabilities. In fact, omitting the WEIGHT statement is tantamount to assigning all units a weight of 1, 10, 100, or any equivalent number.

DOMAIN ANALYSIS

By default, SAS only calculates one overall estimate for each analysis variable provided in the VAR statement. To get subgroup analysis, such as year-end bonus broken out by supervisory status or gender, you must provide a DOMAIN statement with the variable(s) used to define survey data subgroups—you can also cross two or more variables with asterisks. SAS refers to this type of analysis as *domain*, or *subgroup*, analysis. In fact, an alias for the DOMAIN statement in PROC SURVEYMEANS is SUBGROUP.

Suppose management wanted to investigate mean bonuses by gender, supervisory status, and gender within each supervisory status. The following SAS code will output the appropriate analysis:

```
proc surveymeans data=alpha total=totals mean stderr;
  var bonus;
  strata super;
  weight basewgt;
  domain gender super super*gender;
run;
```

This is a crucial point to remember when analyzing survey data analysis SAS. If you are performing any kind of subgroup analysis, it is not recommended to use the BY or WHERE statements. In general, it is not recommended to subset the survey dataset in any way before running a SURVEY procedure. Statistically valid subpopulation analyses can only be guaranteed when the DOMAIN statement is used analyzing the entire dataset.

THE LINK BETWEEN SURVEYMEANS AND SURVEYFREQ

The CLASS statement operates quite differently in PROC MEANS than it does in PROC SURVEYMEANS. In the former, the CLASS statement allows the user to define subgroups for analysis, but it was already noted how the DOMAIN or SUBGROUP statement is the correct method of subgroup analysis in complex survey designs. In PROC SURVEYMEANS, the lone purpose of the CLASS statement is to override the default action for analyzing numeric variables. By default, any numeric variable listed in the VAR statement will be analyzed as continuous and any character variable analyzed as categorical. Continuously analyzed variables have an average value computed, whereas categorically analyzed variables have proportions of each level computed. Listing a numeric variable in both the VAR statement and CLASS statement ensures a categorical variable stored as numeric is analyzed appropriately.

In the Alpha Corporation's survey, a measure of overall employee satisfaction was collected and stored as SATISFACTION, a numeric variable with levels 1 through 5, where 1 is the value assigned for "Very Dissatisfied" and 5 the value assigned for "Very Satisfied." We do not want SAS to calculate a mean coded satisfaction level, but instead wish to estimate proportions of the company's employees who indicated the various degrees of satisfaction. The following SAS code achieves this:

```
proc surveymeans data=alpha total=totals mean stderr;
  class satisfaction;
  var satisfaction;
  strata super;
  weight basewgt;
run;
```

While a branch of PROC SURVEYMEANS can handle basic categorical data analysis, PROC SURVEYFREQ was developed specifically for it. Interestingly, the exact same SURVEYMEANS analysis can be obtained by submitting the following SURVEYFREQ code:

```
proc surveyfreq data=alpha total=totals;
  table satisfaction;
  strata super;
  weight basewgt;
run;
```

The only difference between the two procedures' outputs is that SURVEYMEANS reports proportions and SURVEYFREQ reports percents. Move the decimal point of all SURVEYMEANS output two digits to the right and all numbers will be equivalently reported as percents (or vice versa).

This paper has stressed how PROC MEANS and PROC FREQ do not calculate the correct standard errors of means for data emanating from complex survey designs, but there has been no mention of ill effects on means themselves, because there are none. In truth, PROC SURVEYMEANS and PROC SURVEYFREQ can be computationally intensive, especially with large datasets. If you are generating a report without standard errors, such as a report of only weighted means or frequencies, I would recommend staying with PROC MEANS or PROC FREQ for familiarity and efficiency. Both support a WEIGHT statement and would produce results identical to their SURVEY counterparts.

A BRIEF LOOK TO WHAT'S NEW IN SAS 9.2

Standard error equations for survey statistics become quickly more involved when moving from SRS to complex designs. As a testament to their complexity, this paper gives no specific formulas for standard errors beyond the SRS

accounting for the FPC. (We should be thankful software exist to do the work!) In reality, there are a number of approximations that can be employed, but SAS currently offers only the Taylor series linearization technique. Other commonly used methods are balanced repeated replication and the class of jackknife methods, available in some competing software packages such as STATA® and WesVar PC®. But with the release of Version 9.2, SAS introduces balanced repeated replication and jackknife methods to PROC SURVEYMEANS and PROC SURVEYFREQ, which is a great stride in the continuing evolution of survey analysis in SAS.

CONCLUSION

The goal of this paper was to provide an introduction to complex survey data analysis in SAS by highlighting some of the common syntax amendments required when moving from PROC MEANS and PROC FREQ to PROC SURVEYMEANS and PROC SURVEYFREQ. Three crucial considerations were presented by the fictitious Alpha Corporation's employee year-end bonus survey: 1) accounting for the FPC 2) assigning weights due to unequal sampling rates, and 3) domain analysis in lieu of BY or WHERE processing. Equation modifications were discussed, where pertinent and explicable to lay SAS users. Admittedly, there is ample room for a more thorough discussion of the numerous facets of survey data analysis that can crop up for SAS programmers—clustered sample designs and adjusting survey weights for nonresponse, to name a few. But at the very least, the moral is for SAS programmers to remain vigilant when analyzing complex survey data, because standard PROCs may no longer be valid and caveats exist.

REFERENCES

SAS Institute Inc. 2004. *SAS/STAT® 9.1 User's Guide*. Cary, NC: SAS Institute Inc.
SAS Institute Inc. 2008. "What's New in SAS/STAT 9.2."
<http://support.sas.com/documentation/cdl/en/whatsnew/61982/HTML/default/statugwhatsnew.htm>

RECOMMENDED READING

Kalton, G., and Flores-Cervantes, I. (2003). "Weighting Methods," *Journal of Official Statistics*, 19 (2), pp. 81 – 97.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name	Taylor Lewis
Enterprise	Joint Program in Survey Methodology, University of Maryland, College Park
Address	1218 LeFrak Hall
City, State ZIP	College Park, MD 20742
Work Phone:	301-314-7911
Fax:	301-314-7912
E-mail:	tlewis@survey.umd.edu

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 trademarks of their respective companies.