Calculating Subset Weighted Analysis Using PROC SURVEYFREQ and GENMOD

Jessica J. Hale, MS[1], David M. Thompson, PhD[2], and Paul M. Darden, MD[1]

1 Department of Pediatrics, School of Medicine; 2 Department of Biostatistics and Epidemiology, School of Public Health;
University of Oklahoma Health Sciences Center, Oklahoma City, OK, 73104

ABSTRACT

Stratum-specific weighted analysis is available in SAS procedures such as PROC SURVEYMEANS &
SURVEYLOGISTIC, which include the DOMAIN statement. However, other procedures that can model correlated
outcomes, including PROC GENMOD, do not. This presentation demonstrates a method of assigning individual
weights to each record in a dataset to perform weighted subset analysis on a correlated outcome without creating
domain variables or transferring analysis to a separate program.

This demonstration employs data from the National Immunization Survey Teen (NIS-Teen).The NIS-TEEN includes
provider & household weights and details the immunization status for several vaccines for each respondent. To
compare weighted estimates of vaccination rates, where respondents receive several vaccines and had adequate
provider data, a domain variable must be created or analysis moved into a different program. To calculate individual
weights, the sum of the provider weights of the respondents with a positive provider weight is calculated. Then, each
respondent's positive provider weight is divided by the sum of the positive provider weights & multiplied by the total
sample size of respondents with a positive provider weight.

The newly recalculated individual level provider weights can be used in the in SAS PROC SURVEYFREQ, allowing
for weighted variable by variable comparisons. It permits the use of PROC GENMOD, with multivariate data, to
perform weighted domain-specific analyses that account for correlations among the multiple vaccines offered to
individuals. Validation of the proper calculation of weighted results is proven by producing identical estimates as
analysis in STATA.

INTRODUCTION

After submission of the abstract, further research on the use of individual standardized weights (ISW) led to new
conclusions that this paper explains in detail. These findings apply to the NIS-Teen survey. In order to determine if
the methods presented here apply to a survey other than the NIS-Teen, similar tests of the various examples shown
could be used

The Centers for Disease Control (CDC) began the NIS-Teen survey in 2008 to measure progress toward the Healthy
People 2010 goal of having at least 90 percent of adolescents in the United States receiving routinely recommended
vaccinations. Data was collected in all 50 states and the District of Columbia in randomly selected households
through phone interviews. To ensure accuracy of the immunization data, health care providers of the teens surveyed
were mailed surveys which collected various data on immunizations. For each year the NIS-TEEN is collected, a
unique identifier, SEQNUMT, is assigned to each teen that participated in the survey. Estimation area is identified by
the ESTIAPT variable and allows for calculation of standard errors for state and national estimates of vaccination
rates. Two weight variables are calculated based on different criteria. There is a household weight, RDDWT, and a
provider weight, PROVWT. The RDDWT is used to calculate estimates based on teens that have completed
interviews. The PROVWT is calculated for those teens whose provider supplied a vaccine history. The main question
of interest that will be demonstrated using the NIS-Teen 2010 data throughout this paper is: what is the vaccination
rate of females for the Tetatnus (TET) vaccine? The outcome variable is R_TET_NOT_UTD with values of 0 for up-
to-date (UTD) and 1 for not-up-to-date (nUTD) and the domain variable that identifies the subset of interest is SEX,
which is coded 1 for males and 2 for females.

METHODS

Three SAS procedures, SURVEYFREQ, FREQ and GENMOD, are used to show how subset weighted analysis can
be performed and how they differ from one another. Table 1 compares the results from five examples

SAS Procedure	Output	Sample Subset	Survey Parameters Used in Analysis	Format of Parameter Estimate (Est) from Procedure	Est (percent)	Standard Error (percent)
SURVEYFREQ	1	All records	ESTIAPT, SEQNUMT, PROVWT	Percentage	78.4456	0.9346
	2	Females only	PROVWT	Percentage	78.4456	0.9374
FREQ	3	Females only	PROVWT	Proportion	78.4456	0.0149
	4	Females only	ISW	Proportion	78.4456	0.4881
GENMOD	5	Females only	ESTIAPT, SEQNUMT, PROVWT	Proportion	78.4456	0.9373

Table 1. Output results from examples 1 to 5

WEIGHTED ANALYSIS IN SURVEY PROCEDURES

EXAMPLE 1: PROC SURVEYFREQ – DOMAIN VARIABLE

The first example shows how PROC SURVEYFREQ calculates weighted subset analysis. Unlike PROC SURVEYMEANS, PROC SURVEYFREQ does not have a DOMAIN statement to identify subsets. Instead, PROC SURVEYFREQ incorporates a domain variable(s) in the TABLES statement which identifies the subset(s), or domain(s). In this example, the domain variable used is SEX; which is coded to 1 for males and 2 for females. Since the domain variable already exists in the dataset, a new domain variable to calculate our outcome of interest is not necessary. A portion of the procedure's output is provided.

SAS Code Example 1: Designating a domain variable in the TABLE statement for calculating subset specific analysis using PROC SURVEYFREQ

```
proc surveyfreq data=nis10;
tables sex*r_tet_not_utd/col row;
strata estiapt10;
cluster seqnumt;
weight provwt;
run;
```

```
          r_tet_not_utd 95% Confidence Limits      Row      Std Err of   95% Confidence Limits
  SEX                    for Percent            Percent   Row Percent      for Row Percent
_____

  MALE            0     40.3099     43.2319    81.5550       0.8094    79.9686       83.1415
                  1      8.5943     10.3000    18.4450       0.8094    16.8585       20.0314


               Total    49.7444     52.6917   100.000
  -------------------------------------------------------------------------------------------------
  FEMALE          0     36.8584     39.6762   ^78.4456      ^0.9346    76.6136       80.2776
                  1      9.5447     11.4846    21.5544       0.9346    19.7224       23.3864
               Total    47.3083     50.2556   100.000
  ^Values  referenced in Table 1
```

Output 1. Limited output from PROC SURVEYFREQ

EXAMPLE 2: PROC SURVEYFREQ – SUBSET ANALYSIS

The second example shows how PROC SURVEYFREQ calculates weighted subset analysis using the weight

variable while omitting the other survey parameters related to stratum or cluster. Using the WHERE statement, we specify the subset of interest and calculate weighted proportions for only this subset.

SAS Code Example 2 Designating a subset using the WHERE statement and calculating subset specific analysis using PROC SURVEYFREQ and only the weight variable, ignoring other survey parameters

```
proc surveyfreq data=nis10;
tables r_tet_not_utd/row cl;
weight provwt;
where sex = 2;
run;
```

r_tet_not_utd	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	95% Confidence Limits for Percent	
0	5511	5945067	118876	^78.4456	^0.9374	76.6081	80.2832
1	1364	1633515	79738	21.5544	0.9374	19.7168	23.3919
Total	6875	7578583	132906	100.000			

Frequency Missing = 2345

^Values referenced in Table 1

Output 2. Limited output from PROC SURVEYFREQ

Although the estimates calculated in output 1 and 2 are the same, the estimates' standard errors differ; the method using the WHERE statement and weight variable without the other design elements is more conservative. The weights provide sufficient information for PROC SURVEYFREQ to produce unbiased estimates; however, the STRATA and CLUSTER statements are important to accurately estimate sampling variability as well as utilizing the entire sample and specifying a domain rather than analyzing only a subset.

WEIGHTED ANALYSIS IN NON-SURVEY PROCEDURES

Because the standard survey procedures do not accommodate correlated outcomes, analysts may be interested in moving beyond their capabilities. However, because procedures like PROC FREQ do not allow for specification of survey parameters (including strata and clusters), the analyst must calculate and use individual standardized weights (ISW).

EXAMPLE 3: PROC FREQ – SUBSET ANALYSIS USING ORIGINAL WEIGHT

The third example shows how in PROC FREQ, using the original weight variable may calculate the same estimate as PROC SURVEYFREQ, but underestimates the population variance or standard error.

SAS Code Example 3: Designating a subset using the WHERE statement and calculating subset specific analysis using PROC FREQ and only the weight variable while ignoring other survey parameters.

```
proc freq data=nis10;
exact binomial;
tables r_tet_not_utd;
weight provwt;
where sex = 2;
run;
```

```
                                         Cumulative       Cumulative
           r_tet_not_utd    Frequency    Percent    Frequency     Percent
           ──────────────────────────────────────────────────────────────
                       0    5945067      ^78.45     5945067         78.45
                       1    1633515       21.55     7578583        100.00
                       Frequency Missing = 2446536.6537


                           Binomial Proportion for
                              r_tet_not_utd = 0
                      ──────────────────────────────────

                      Proportion (P)                0.7845
                      ASE                          ^0.0001
                      95% Lower Conf Limit          0.7842
                      95% Upper Conf Limit          0.7847
```

^Values referenced in Table 1

Output 3. Limited output from PROC FREQ

Table 1 shows how example 3 differs from examples 1 and 2 in producing estimates for the SE that are much smaller. The STRATA and CLUSTER statements provide important information that PROC FREQ does not duplicate.

ISW

Steps and Formula for Calculating Individual Weights

To calculate individual standardized weights, the original weight variable for each respondent must be divided by the sum of the weights and multiplied by the total sample size of respondents with weights. Through this process, we standardize the weights to each individual.

ISW Formula:

$$ISW = (w_{original}/\sum w_{original})*n$$

Step 1: Calculate the sum of the of the original weight variable

Step 2: Divide each respondent's original weight value by the sum of the original weight variable

Step 3: Multiply by the total sample size of respondents with an original weight value

SAS Code Example 4: Calculating the ISW

```
*calculating the sum of provider weight variable (sum_provwt) and the number of
positive provider weights (n_provwt);
*Step 1;
proc means data=nis10 sum n;
var provwt;
output out = provwt_2 sum = sum_provwt n = n_provwt;
run;
*Calculating the new standardized weights (ISW);
*the new Individual weight = (Weight / Sum weight) * number of positive provider
weights;
*Steps 2 and 3;
data nis_10_new;
if _n_ = 1 then set provwt_2;
set nis_10;
new_providerwt_sas = ((provwt/sum_provwt)*n_provwt);
run;
```

4

EXAMPLE 4: SUBSET ANALYSIS WITH PROC FREQ, USING ISW INSTEAD OF ORIGINAL WEIGHTS

The newly recalculated ISW along with a where statement identifying the subset of interest can be used in PROC FREQ, allowing for weighted variable by variable comparisons within a subset.

SAS Code Example 5: Using Individual Provider Weight in PROC FREQ to Calculate Subset Weighted Proportions

```
proc freq data=nis10_new;
exact binomial;
tables r_tet_not_utd;
weight new_provwt;
where sex = 2;
run;
```

```
                                         Cumulative    Cumulative
            r_tet_not_utd   Frequency   Percent   Frequency    Percent
          ─────────────────────────────────────────────────────────────────
                       0    5567.707     ^78.45    5567.707      78.45
                       1    1529.829      21.55    7097.535     100.00

                          Frequency Missing = 2291.2437868
                            Binomial Proportion for
                                r_tet_not_utd = 0
                          ──────────────────────────────────
                          Proportion (P)            0.7845
                          ASE                      ^0.0049
                          95% Lower Conf Limit      0.7749
                          95% Upper Conf Limit      0.7940
```

^Values referenced in Table 1

Output 4. Partial output from PROC FREQ

The bolded values in output 4 represent the percentages and standard errors that Table 1 compares to those estimated in examples 1 to 3 and 5.

EXAMPLE 5: PROC GENMOD – SUBSET ANALYSIS

PROC GENMOD can perform repeated measures and other analyses on correlated data. Example 5 determines whether PROC GENMOD will complete appropriately weighted subset analysis. In order to determine if appropriate weighted subset analysis can be completed in PROC GENMOD for repeated measures analysis, the weighted estimate and SE should be calculated in PROC GENMOD. Using GENMOD, the original weight and survey parameters can be incorporated into PROC GENMOD's REPEATED statement. The WEIGHT statement specifies the original weight, and the cluster and stratum variables are identified in the REPEATED statement and SUBJECT option by specifying the "cluster*stratum" variables.

SAS Code Example 6: Using PROC GENMOD to Calculate Subset Weighted Proportions

```
proc genmod data=nis10;
class seqnumt estiapt10;
model r_tet_not_utd = / dist=binomial link=identity;
weight provwt;
repeated subject =seqnumt(estiapt10);
where sex = 2;
run;
```

```
                        Analysis Of GEE Parameter Estimates
                        Empirical Standard Error Estimates


                              Standard    95% Confidence
                Parameter Estimate    Error       Limits          Z Pr > |Z|


                Intercept  ^0.7845  ^0.0094   0.7661   0.8028   83.69   <.0001

      ^Values referenced in Table 1
```

Output 5. Limited output from PROC GENMOD

DISCUSSION

Calculating subset weighted analysis has been demonstrated across three well known SAS PROCEDURES: SURVEYFREQ, FREQ and GENMOD. This paper compares five methods that produce identical estimates for weighted subset analysis, but with different standard errors. For the NIS-Teen data, the CLUSTER variable is SEQNUMT. There is one unique SEQNUMT assigned per participant, therefore all clusters are of equal size, with a single observation in each cluster. The STRATUM variable ESTIAPT has a mean size of 158.97 observations (or clusters) and a standard deviation of 18.05 observations for the subset sample of females with a provider weight. Again there are not clusters within the data based on the SEQNUMT variable and based on the standard deviation of the ESTIAPT variable of 18.05, not a large amount of variability within the STRATUM variable for our subset of interest; this may explain why the SE calculated with all the survey parameters included in PROC SURVEYFREQ versus the SE calculated with only the weight variable are very similar.

The SAS manual suggests that using a WHERE statement or creating a dataset with only the subset of interest and not including the survey parameters is an inappropriate method for calculating appropriate subset weighted estimates. For the NIS-Teen survey, this paper demonstrates that using a subset of the data and only the weight variable calculate the same parameter estimates and more conservative standard errors than when using the entire dataset and all survey parameters in PROC SURVEYFREQ. Analysts should be cautious in using this method for surveys other than the NIS-Teen. Extensive analytical tests should be run to ensure that ignoring the survey parameters and only analyzing a subset of data does not produce different parameter estimates or more liberal standard errors.

Calculation of ISW is necessary when attempting to perform weighted analysis for a subset in a non-survey procedure when the cluster and ESTIAPT variables cannot be specified. Using original weights yields standard errors that significantly underestimate the variability of the sample, but using an ISW calculates SE much closer to the SE calculated in the survey procedures. The main limitation using PROC FREQ to calculate weighted subset analysis is that the standard errors will be less conservative compared to other procedures that utilize all survey parameters. However, using an ISW versus the original weight variable in PROC FREQ produces more conservative standard errors which are closer to those calculated in procedures that allow for all survey parameters to be utilized. As example 3 demonstrates, the ISW offers the ability to use non-survey procedures that do not allow for specification of the survey parameters and decreases the amount of output needed to review. However, when all survey parameters can be specified in a non-survey procedure, such as PROC GENMOD, the original weight variables can be used instead of the ISW.

The ability to use the original weights to calculate weighted estimates using PROC GENMOD that produce similar variance estimates as PROC SURVEYFREQ has been shown by specifying the cluster and stratum variables in the REPEATED statement and SUBJECT option of GENMOD. Using these options yields the same parameter estimate and nearly the same exact SE as PROC SURVEYFREQ. This implies that if a non-survey procedure offers options where the cluster and stratum variables can be specified, appropriate subset weighted analysis can be performed using the original weights and other survey parameters. The ability to calculate appropriate weighted parameter and variance estimates for subsets from complex survey data opens new possibilities for researchers to utilize the analytic options available in GENMOD, such as repeated measures analysis and risk difference calculations.

CONCLUSION

Performing subset specific weighted analyses on survey data can be straightforward or complicated, depending on the number of subsets of interest and complexity of the analysis. This paper has demonstrated how the same

weighted estimates can be calculated using various methods, but how these methods differ from one other in accounting for the variance of the NIS-Teen survey. Also demonstrated is the possibility of running weighted subset analysis in PROC GENMOD. Having the option to run analysis PROC GENMOD on weighted survey data allows researchers and analysts the opportunity to analyze weighted correlated data leading to potential discovery of new significant findings in their data that are not currently available in survey procedures.

REFERENCES

SAS/STAT(R) 9.2 User's Guide, Second Edition. "Chapter 83: The SURVEYFREQ Procedure." September 2009. <http://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf> (November 19, 2013)

SAS/STAT(R) 9.2 User's Guide, Second Edition. "Chapter 85: The SURVEYMEANS Procedure." <http://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf> (November 19, 2013)

SAS/STAT(R) 9.2 User's Guide, Second Edition. "Chapter 35: The FREQ Procedure." <http://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf> (November 19, 2013)

SAS/STAT(R) 9.2 User's Guide, Second Edition. Chapter 37: The GENMOD Procedure. <http://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf> (November 19, 2013)

University of Toronto. "Normalized weights: is using them enough?" *National Longitudinal Survey of Children and Youth (NLSCY) Workshop 3*. October 2008. <http://www.utoronto.ca/rdc/files/3_NLSCY_Workshop__Nonresponse_and_Normalized_Weights_and_Pooling_Data _and_Full_Example.pdf> (November 19, 2013)

Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases, National Center for Health Statistics (U.S.). National Immunization Survey-Teen. A User's Guide for the 2010 Public-Use Data File2010:< http://www.cdc.gov/nchs/nis/data_files_teen.htm. Accessed 12/14/2011>.

Jain N, Singleton JA, Montgomery M, Skalland B. Determining accurate vaccination coverage rates for adolescents: the National Immunization Survey-Teen 2006. *Public Health Rep.* Sep-Oct 2009;124(5):642-651.

CONTACT INFORMATION

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

Jessica Hale
University of Oklahoma Children's Physicians, Department of Pediatrics
1200 N. Children's Avenue
Oklahoma City, Oklahoma 73104
Phone (405)271-4407
Email: Jessica-J-Hale@ouhsc.edu

Dr. David Thompson
University of Oklahoma Health Sciences Center, Dept. of Biostatistics and Epidemiology
801 Northeast 13th Street, Room 352
Post Office Box 26901
Phone (405)271-2229
Email: David-Thompson@ouhsc.edu

Dr. Paul Darden
University of Oklahoma Children's Physicians, Department of Pediatrics
1200 N. Children's Avenue
Oklahoma City, Oklahoma 73104
Phone (405)271-4407
Email: Paul-Darden@ouhsc.edu

ACKNOWLEDGEMENTS