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

Why Subset?

- May not need to analyze all the observations collected for analysis
- Allows analyst to use a smaller, condensed dataset that contains only the necessary observations
- Less code to write

National Immunization Survey (NIS)-Teen

- Center for Disease Control (CDC) began collecting the NIS-Teen survey in 2008
- Immunization related data collected from parents and providers
- Examples and results shown are from the NIS-Teen 2010 data
  - Outcome of interest is being up-to-date (UTD) for the Tetanus (TET) vaccine for females only
  - Parent collected information used
  - Weight variable is the provider weight variable (PROVWT)
  - Clusters are represented by the unique identifier, SEQNUMT, which is assigned to each teen that participated in the survey
  - Estimation area or stratum are identified by the ESTIAPT variable
  - Domain variable is sex

Methods

- Three SAS® procedure examples, SURVEYFREQ, FREQ and GENMOD are displayed
- Domain variables can be used in the survey procedures available in SAS® for subset analysis
- Subset weighted analysis can be performed outside of the survey procedures using standardized weights or specifying the survey parameters when available
- Individual standardized weights (ISW) (Program 6: SAS® Code)

ISW Formula:

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

Step 1: Calculate the sum of PROVWT

Step 2: Divide each respondent's PROVWT value by the sum of PROVWT

Step 3: Multiply by the total sample size of respondents with a PROVWT value

Results

Results from SAS® programs 1 to 5

SAS® Procedure	Program	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

SAS® Programming

Program 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;
```

Program 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;
```

Program 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;
```

Program 4: 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;
```

Program 5: 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;
```

Program 6: Calculating the ISW

```
*Step 1;
proc means data=nis10 sum n;
var provwt;
output out = provwt_2 sum =
sum_provwt n = n_provwt;
run;
*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;
```

Discussion

- Researchers and analysts need to be aware that various methods produce variable estimates and standard errors
- SAS® survey procedures use all survey parameters, which can be difficult if the focus is on analyzing multiple subsets within a dataset
- The SAS® manual advises against using the WHERE statement to perform weighted analysis of subsets of survey data
- The examples shown analyze a subset of survey data while calculating similar or more conservative SE
- When a SAS® procedure does not permit specifying the stratum or cluster variables, the ISW is an alternative method to performing the analysis
  - Using the original weights without the capability of specifying the stratum and cluster variables yields very small standard errors

Conclusions

- Multiple options are available for calculating weighted subset analysis
  - PROC SURVEYFREQ
  - PROC FREQ
  - PROC GENMOD
- Estimates across the various methods remain consistent, but the standard errors vary depending on the method used
- Stratum and cluster variables for a survey can be applied in PROC GENMOD using the REPEATED statement and SUBJECT option
- Ability to utilize the capabilities of GENMOD , such as GEE, allows researchers and analysts the opportunity to discover potential new findings previously unavailable for survey data

Caution

- The methods presented here apply to the NIS-Teen survey, but may produce different results on other complex survey designs
- In order to determine if these methods apply to a complex design survey other than the NIS-Teen, tests similar to the various examples shown could be used

Acknowledgements

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

Contact Us

Jessica J. Hale jessica-j-hale@ouhsc.edu

David M. Thompson david-thompson@ouhsc.edu

Paul M. Darden paul-darden@ouhsc.edu