

Estimating Design Effects for Means, Proportions and Totals from Complex Sample Survey Data Using SAS® Proc Surveymeans

Trent D. Buskirk, Ph.D.
Saint Louis University School of Public Health
Saint Louis, MO

Abstract

Recent advances in survey data analysis procedures continue to be made in the latest releases of SAS software. Despite these advances, functionality for computing design effects for complex sampling designs are not readily available in proc surveymeans. In this presentation we will use the 2010 BRFSS National survey data to illustrate how to compute design effects for means, totals and proportions computed from proc surveymeans by using proc univariate, proc means and proc freq.

Introduction

Data collected from complex sample surveys that involve multiple stages, stratification of populations and clustering of sampling units usually require the end user to incorporate sampling weights and these design variables into the analysis to avoid errors in inference, particularly in variance estimation. In most cases, sampling designs reflect strategies that are more than simple random samples from the population, in part because final sampling units are usually clustered within primary sampling units that may be stratified within the population of interest. More often than not, multistage designs are employed because a list, or sampling frame, of final sampling units (from which a simple random sample could be easily derived) is only available after a larger sampling unit is selected. For instance, in health settings it may not be possible to have a list of all nurse practitioners who are currently working within a hospital system, but once a particular hospital is selected for a sample, such a list of nurse practitioners currently on the payroll may be readily available.

The price paid in terms of efficiency for using a complex sample compared to a simple random sample (presumably without replacement) is called the DESIGN EFFECT. More formally, the design effect is the quotient of the sampling variance of an estimate of interest (i.e. means, totals or proportions) derived under the design based inference framework implied by the specific sampling design used and an estimate of the sampling variance for the same estimator derived under simple random sampling (without replacement) (SRSWOR) of the same number of final sampling units (Lohr, 2010). Effectively, the design effect quantifies the complex effects of stratification, clustering and weighting on the overall standard errors (Heeringa et al. 2010).

Another interpretation of the design effect is the increase/decrease in sample size needed under the current design to collect the same (statistical) information that could have been obtained if a simple random sample had been conducted of the same size. Design effects larger than 1 would then imply that the current sampling design is less efficient than an SRSWOR design since more final sampling units were needed in order to produce an estimate with the same variance. Similarly, design effects that are less than 1 indicate that the sampling design is more efficient than a SRSWOR design because fewer final sampling units were required to achieve the same level of precision in the estimator that one would have obtained under a SRSWOR design-based framework. Design effects are also the basis of computing effective sample sizes as described in Heeringa et al. (2010).

In general, under stratified sampling designs, for example, the design effect can be lower than one if the final sampling units within strata are more homogenous as compared to across strata. Clustered designs generally result in design effects that exceed one unless elements within a cluster are more heterogeneous or whenever clusters tend to be homogeneous with respect to the outcome of interest.

Design effects as well as point estimates, confidence intervals and other statistics under complex sampling designs can now be derived using most modern statistical software packages including Stata, R, Sudaan, SPSS and SAS. Specifically in SAS an analyst can deploy Proc Survey Means for estimating population means and totals (and proportions), Proc Survey Freq for estimating proportions and other tests of association between categorical variables and Proc Survey Reg or Proc Survey Logistic for creating population based regression and logistic regression models, respectively. All of these survey procs require sampling weights and specifications of pertinent sampling design variables such as stratification and clustering variables. More details on these SAS Survey Procedures can be found throughout the literature including (Chen and Gorrell, 2008 and Berglund, 2009). Others have compared SAS analytic survey procedures to SUDAAN including Baisden and Hu (2003).

The design effects can be computed for many requested statistics within every survey analysis procedure, except proc survey means. While proc survey means can generate accurate point estimates and corresponding standard errors for means and totals, no design effects can be requested or computed directly with a single call to Proc SurveyMeans. In contrast, the deff and rowdeff and coldeff options in Proc SurveyFreq allow the analyst to request that SAS directly compute the design effect for overall proportion, row and column proportion estimates, respectively. Similarly, the deff option in Proc SurveyReg requests that SAS directly compute the estimated design effect for regression parameter estimates. In this paper, we will provide two methods for estimating design effects for estimates of population means (and proportions) and sums that would have been derived using Proc Survey Means. We will also illustrate both methods using two publically available data sets- the 2010 Behavioral Risk Factor Surveillance Survey (BRFSS) (<http://www.cdc.gov/brfss/>).

Design Effect Technical Details.

If the parameter of interest is defined as θ (i.e. such as a population mean, total or proportion) and is estimated using a design-based weighted estimator $\hat{\theta}$ then the design effect is formally defined as:

$$DE(\hat{\theta}) = \frac{V_D(\hat{\theta})}{V_{SRS}(\hat{\theta})}$$

and is estimated using variance estimators of $\hat{\theta}$ under the sampling design D as well as under simple random sampling (without replacement). The numerator variance in the design effect equation represents design effects of stratification and clustering, unequal weighting (i.e. differential selection probabilities) and over/under sampling from specific subpopulations of interest. We note that in many applications, with replacement sampling of primary sampling units (psu's) from across strata is generally assumed in order to operationalize variance estimation in the survey software. To this end, the sas survey procedures will generally approximate the true design variance by assuming with replacement sampling of the psu's and effectively ignore the finite population correction (fpc) factors in this computation. Under these circumstances, the variance computation of the estimate under simple random sampling should also ignore the fpc factor for computing the overall design effect. In the examples we provide in this paper we will assume with replacement sampling (e.g. small sampling fractions within

strata) and ignore the fpc factor in the computations of the variance under simple random sampling.

Analysts familiar with SUDAAN recognize that there are four possible design effects that can be requested for SUDAAN survey analytic procedures (e.g. DEFF1, DEFF2, DEFF3, and DEFF4). The distinction between such design effects center around assumptions about fixed or random subdomain sizes and whether or not sampling was done with replacement and unequal probabilities. DEFF4 is the default approach for estimating design effects for means, proportions and totals in SUDAAN. It essentially implements a model-based approach for estimating the variance of the estimate under simple random sampling. This specific design effect is what we will be computing in the approaches outlined in the next section. Specifically, the design effect of interest for us is the value of DEFF4 in SUDAAN for means, totals and proportions. We note that from our understanding, design effects computed using SAS survey procedures are also equal to SUDAAN's DEFF4. For more information about how SUDAAN differentiates between various types of design effects, consult the Sudaan 10 User's Manual (RTI, 2008).

Based on DEFF4 in SUDAAN, the reference formula for estimating the variance of a mean under simple random sampling (assuming no finite population correction factors) is:

$$\hat{V}_{SRS}(\bar{x}) = \frac{1}{n-1} \frac{\sum_{i=1}^n w_i [x_i - \hat{E}(x_i)]^2}{\sum_{i=1}^n w_i} \quad (1)$$

where $\hat{E}(x_i)$ is the weighted survey mean estimate for the outcome of interest. For totals, the estimated variance under SRS is $\hat{N}^2 \times \hat{V}_{SRS}(\bar{x})$ where $\hat{N} = \sum_{i=1}^n w_i$. We note that the value of "n" in formula (1) will generally depend on the number of observations that have valid values for the outcome of interest (or the number of respondents who fall in a particular subdomain, for subdomain estimates). Formula (1) serves as the basis for the methods we describe herein for computing the design effects for estimates derived from SAS Proc Surveymeans,

Computing Design Effects for Population Means, Proportions and Totals Using SAS Proc SurveyMeans

In this section we provide heuristic algorithms for two approaches that can be used to compute design effects for estimated means, proportions and totals derived from SAS Proc SurveyMeans. The first of these two methods relies on applications of both the survey means and the univariate procedures while the second approach relies on the surveyreg procedure. More details of these procedures are discussed in Berglund (2001) and An and Watts (1998). The first method can also be adapted to estimate design effects for means, totals and proportions related to SUBDOMAINS of interest. Detailed examples with corresponding SAS syntax are included in the next section.

Approach 1: Means and Re-Means

To estimate the design effect for means and population total estimates derived using Proc SurveyMeans, the user must currently compute the numerator and denominator of the design effect measure explicitly. To compute the numerator of the design effect (i.e. variance of the mean or total estimate under the complex sampling design), one should apply Proc SurveyMeans incorporating the proper stratification and clustering as well as the sampling

weights. To properly account for the missing data in the computation of the design-based variance estimate, the user should specify the NOMCAR (i.e. not missing completely at random) option in the proc surveymeans data statement. This option will provide a design based variance estimate that is consistent with what is computed in SUDAAN's proc descript. More information on this option can be found from the SAS technical support (<http://www.sas.com/support>) or in Berglund (2009).

We note that while proc surveymeans with NOMCAR option will make use of all observations in the computation of the design-based variance, only observations with a non-missing value for the outcome of interest will be used in deriving the point estimate. Similarly, only observations with a non-missing value on the outcome will be used in the estimate of the variance under simple random sampling. This approach is consistent with how SUDAAN computes the simple random sampling variance estimate. The denominator of the design effect (i.e. the variance of the estimator under simple random sampling) is then computed using sas proc univariate with a weight statement that specifies the design-based sampling weight as the weighting variable (and will apply only to observations with non-missing information on the outcome of interest).

The design effect is then estimated as the quotient of the variance under the surveymeans procedure divided by the variance derived under the weighted univariate procedure with sum of the weights as the variance divisor (i.e. vardef=WEIGHT option) divided by the sample size, less 1 (for means and proportions). We note that the variance estimate for totals under SRS is the variance derived from the weighted proc univariate with the default variance divisor (i.e. vardef=DF) multiplied by an estimate of the population total (i.e. weighted sum of all the weights).

Summarizing, the design effect can be computed using the Means and Re-Means approach using the following steps:

Step 1: Compute the variance (i.e. squared standard error) of the mean, proportion or total of interest using proc surveymeans accounting for the stratification, clustering and sampling weight variables;

```
proc surveymeans data=DATASET mean nomcar;
  strata  STRATUM_VAR;
  cluster CLUSTER_VAR;
  weight  SAMPLING_WEIGHT;
  var    VAR_OF_INTEREST;
  /*The output data set below will contain information that will be used to
  compute the NUMERATOR of the DESIGN EFFECT estimate of the MEAN*/
  ods output Statistics=SampDesignOut (rename=Stderr=DesignSE keep=Stderr
keep=mean);
run;
```

Step 2: Compute the variance of the mean, proportion or total of interest using proc univariate with the sampling weight variable used on the weight statement;

Step 2 NOTE: If computing a mean or proportion, use VARDEF=WGT option in the proc univariate data statement. If computing a total, use VARDEF=DF (default) option in proc univariate data statement.

```
/* Use Proc Univariate with a weight statement and Vardef Statement to
estimate the SE of the sample mean estimator under Simple Random Sampling
(here we assume NO FPC)*/
proc univariate data=DATASET vardef=WGT;
var VAR_OF_INTEREST;
weight SAMPLING_WEIGHT;
```

```
ods output moments=UNIVAR_Moments; run;
```

Step 3: Design effect is computed as the quotient of the result from step 1 divided by the result of step 2 divided itself by the overall sample size minus 1.

```
/*Transpose moments table to readily capture data on Sample Size and Standard
Error of the Mean under SRS */
proc transpose data=UNIVAR_Moments (keep=nvalue1)
prefix=Stat
/* The output data set below will contain information that will be used to
compute the DENOMINATOR of the DESIGN EFFECT estimate of the MEAN*/
Out=SRSOut (rename=stat1=n_size rename=stat3=SRSSE keep=stat1 keep=stat3);
run;

data DATASET_DesignEffect;
merge SampDesignOut SRSout;
DESIGN_Effect=(DesignSE/SRSSE)**2*(n_size-1); /*FOR MEANS and PROPORTIONS*/
DESIGN_Effect_Totals=(DesignSE/SRSSE)**2; /*For TOTALS*/
run;

proc print data=DATASET_DesignEffect;
var mean DesignSE Design_Effect; run;
```

Approach 2: Intercept-Only Regression

Under the traditional simple linear regression model, the regression intercept estimate is computed as $\beta_0 = \bar{y} - \beta_1 \bar{x}$. Thus, using a null, intercept-only regression model, properties of the sample mean estimator can be derived from corresponding properties of the regression intercept estimate, β_0 . This fact coupled with the fact that SAS Proc SurveyReg can provide estimates of design effects for each estimated regression parameters allows us to generate design effects for estimated means and proportions. This approach will not apply to estimates of population totals.

Executing these Proc Surveyreg specifications will generate an intercept estimate which should be equivalent to the mean or proportion (weighted). The output will also contain a corresponding design effect estimate. We note that under normal circumstances one would not attempt to apply a linear regression model for the purposes of modeling a binary outcome. We simply advocate it here because we are looking at interpreting the regression intercept estimate as an estimated proportion (i.e. point estimate) rather than using the model for other inferential purposes.

To properly account for missing data in the computation of the design-based variance of the intercept estimate the user should specify the NOMCAR option in the proc surveyreg data step. By using the NOMCAR option the surveyreg procedure will incorporate all observations (missing and non-missing for the outcome of interest) into the computation of **both** the design based and SRS-based variance estimates of the intercept. Specifically, the divisors in (1) are the sample size of the entire sample while the sums of weighted observations are only for those observations that have complete information for the outcome of interest. This difference in computation is inconsistent with approach 1 and with Sudaan computations. To correct for the potential underestimation of the SRS variance (or overestimation of the design effect) we need to apply a missing observation correction factor to the design effect computed from proc

surveyreg (with NOMCAR option) as follows: $DEcorr = \frac{n_{used}}{n_{total}}$ so the final design effect estimate is given by: $FinalDE = DE_{procSurveyReg} \times DEcorr$.

The Intercept-Only Regression approach for deriving design effect estimates for means and proportions can be implemented in SAS Proc Surveyreg with the following specifications and steps:

Step 1, Specification 1: Use the NOMCAR statement with Proc SurveyReg to produce the correct design-based variance for the estimates of means and proportions (in the presence of missing data for the outcome variable of interest);

Step 1, Specification 2: Specify the survey design and sampling weight variables appropriately in proc survey reg. If interest is given to estimating population proportions, convert your outcome variable to a binary variable where 1 indicates the condition of interest and 0 the other alternative. Also specify the nomcar option in the data statement.

Step 1, Specification 3: Specify the model statement using only the outcome variable and an “equals” sign with no covariates listed followed by the DEFF keyword as an option on the model statement.

```
proc surveyreg data=DATASET nomcar;
  strata STRATIFICATION_VARIABLE;
  cluster CLUSTER_VARIABLE;
  weight SAMPLING_WEIGHT;
  model VARIABLE_of_INTEREST = /deff;
  ods output datasummary = SurveyRegSum;
  ods output ParameterEstimates = MyParmEst
    (keep=designeffect keep=estimate keep=stderr);
run;
```

Step 2: Using the total number of cases and total number of cases used from Proc SurveyReg data summary table, compute the design effect correction factor $DEcorr = \frac{n_{used}}{n_{total}}$ and multiply this factor by the design effect estimate for the intercept estimate computed from step 1 by Proc SurveyReg. The resulting value is the correct design effect estimate and is consistent with the corresponding estimate computed using Approach 1.

```
proc transpose data=SurveyRegSum (keep=nvalue1)
  /*nvalue1 refers to the column of data that contains size of sample, number
  of used observations, etc.*/
  prefix=n
  /* The output data set below will contain information that will be used to
  compute the correction factor for the DESIGN EFFECT estimate of the MEAN*/
  out=SurveyRegSum (rename=n1=n_Total rename=n2=n_Used keep=n1 keep=n2);
run;

data revisedDesignEffect;
merge MyParmEst SurveyRegSum;
RevisedDE=(n_Used/n_Total)*DesignEffect;
rename Stderr=DesignSE;
run;
```

```
proc print data=RevisedDesignEffect;  
var estimate DesignSE RevisedDE;  
title 'Estimated Mean, Standard Error and Design Effect, revised from Proc  
SurveyReg Output';  
run;
```

Adapting These Approaches to Subdomain Estimation

The means and re-means approach can also be applied to estimating means, proportions and totals from proc surveymeans for specific subdomains of interest. For example, if you wanted to know the mean age of adults living in households with children, you could define a (sub)domain variable based on parental/guardian status and use such a variable with the domain statement in proc surveymeans to derive the appropriate estimate and corresponding standard error. Note that for subdomain estimation, the domain statement is used in conjunction with the full data set and no additional where or other subsetting is required (see Berglund, 2009 for example). When estimating the variance under simple random sampling using the means and re-means approach, the use of a where statement to subset the original data file to a subset containing only those observations in the domain of interest is indicated. The resulting standard error that is derived using the weighted proc univariate on the domain subset forms the basis for the denominator of the design effect. We note that while it should theoretically be possible to apply the regression intercept only approach using a domain statement as an additional specification, this approach has not yielded consistent estimates of the domain specific design effects with Approach 1 and SUDAAN. We suspect that these differences are directly related to how subdomain estimation is handled in the regression models but have not been able to confirm this based on the limited documentation for domain estimation and design effects in proc surveyreg. For these reasons, we do not recommend the Intercept Only Approach for estimating design effects for means/proportions related to subdomains of interest.

The Behavioral Risk Factor Surveillance Survey (BRFSS)

Briefly, the BRFSS is a disproportionately stratified random sample (in some states/territories, simple random sampling was used) of the adult U.S. household population. Generally, states represent levels of the stratification variable, but in many cases, subregions of states also formed another layer of stratification. The sample data collected from the BRFSS is weighted to account for the differential selection probabilities both within and across states and is post-stratified to population controls to adjust for selection and possibly undercoverage biases. The overall sampling design for the 2010 BRFSS contains no clustering, so only a strata statement is required in the specification of the survey design when using SAS survey procedures. Further information about the 2010 BRFSS including sampling design and weighting can be accessed in the Overview Documentation online at:

http://www.cdc.gov/brfss/technical_infodata/surveydata/2010.htm

In the examples used in this paper we will focus on the continuous variable age, a “parent or guardian variable” that was computed based on the number of children under the age of 18 living in the household and a lifetime diabetes variable (i.e. “ever been told by a doctor that you have diabetes?”). Specifically, in the next section we will compute the design effect for the estimated mean age of adults in households within the U.S. We will also compute the design effect for the estimated number of adults living in households in the state of Missouri who have ever been told they diabetes by a physician.

Example 1: Computing the Mean Age of Adults Living in U.S. Households using Means, Re-Means Approach

Based on the variable age, the BRFSS stratification variable _ststr and the sampling weight variable _finalwt, we will estimate design effect for the estimated mean age of adults living in households within the U.S. using the Means and Re-Means approach.

STEP 1:

```
/*Use Proc SurveyMeans to Compute the Standard Error of the Sample
  Mean Estimate under the Complex Sampling Design*/
proc surveymeans data=state.brfss2010 mean nomcar;
strata _ststr;
weight _finalwt;
var age;
title 'Average Adult Age BRFSS';
/*The output data set below will contain information that will be used to
compute the NUMERATOR of the DESIGN EFFECT estimate of the MEAN*/
ods output Statistics=BRFSS_SampDesignOut
  (rename=Stderr=DesignSE keep=Stderr keep=mean);
run;
```

STEP 2:

```
/* Use Proc Univariate with a weight statement and Vardef Statement to
estimate the SE of the sample mean estimator under Simple Random Sampling */
proc univariate data=DELIB.brfss2010 vardef=WGT;
var age;
weight _finalwt;
ods output moments=BRFSSMoments;
run;
```

```
/*Transpose moments table to readily capture data on Sample Size
and Standard Error of the Mean under SRS */
proc transpose data=BRFSSMoments (keep=nvalue1)
prefix=Stat
/* The output data set below will contain information that will be used to
compute the DENOMINATOR of the DESIGN EFFECT estimate of the MEAN*/
out=BRFSS_SRSOut (rename=stat1=n_size rename=stat3=SRSSE keep=stat1
keep=stat3); run;
```

STEP 3:

```
/*Now combine the two output data files and compute the Design Effects
directly using the output*/
data DELIB.BRFSSDesignEffect;
merge BRFSS_SampDesignOut BRFSS_SRSOut;
DESIGN_Effect=(DesignSE/SRSSE)**2*(n_size-1);
run;

proc print data=DELIB.BRFSSDesignEffect;
var mean DesignSE Design_Effect;
title 'Estimated Mean Age of U.S. Adults (BRFSS, 2010), Standard Error and
Design Effect';
run;
```

Executing the print statement above produces the following point estimate and design effect for the average age of adults living in households within the U.S. based on the 2010 BRFSS:

Estimated Mean Age of U.S. Adults (BRFSS, 2010), Standard Error and Design Effect

Obs	Mean	DesignSE	DESIGN_Effect
1	47.293424	0.054654	4.40141

Example 1(b): Computing the Mean Age of Adults Living in U.S. Households Using Intercept Only Regression Model Approach

Based on the variable age, the BRFSS stratification variable `_ststr` and the sampling weight variable `_finalwt`, we will estimate design effect for the estimated mean age of adults living in households within the U.S. using the Means and Re-Means approach. We note that there were a total of 451,075 observations in the analysis of which 4,220 were missing age data implying that $DE_{corr} = .990645$.

STEP 1:

```
proc surveyreg data=state.brfss2010 nomcar;
  strata _ststr;
  weight _finalwt;
model age = /deff;
ods output datasummary = SurveyRegSum;
ods output ParameterEstimates = MyParmEst
  (keep=designeffect keep=estimate keep=stderr);
title 'Average Age of U.S. Adults in Households (BRFSS, 2010) Using
REGRESSION Approach';
run;
```

The estimate and design effect (unadjusted) from this step are given below:

Estimated Regression Coefficients (with NOMCAR)					
Parameter	Estimate	Standard Error	t Value	Pr > t	Design Effect
Intercept	47.2934240	0.05465389	865.33	<.0001	4.44

STEP 2:

```
proc transpose data=SurveyRegSum (keep=nvalue1)
  prefix=n
/* The output data set below will contain information used to compute the
DENOMINATOR of the DESIGN EFFECT estimate of the MEAN*/
  out=SurveyRegSum
  (rename=n1=n_sizeTotal rename=n2=n_sizeUsed keep=n1 keep=n2);
run;

data revisedDesignEffect;
merge MyParmEst SurveyRegSum;
RevisedDE=(n_sizeUsed/n_sizeTotal)*DesignEffect;
rename Stderr=DesignSE;
run;
```

Applying the design effect correction to the results obtained from `proc surveyreg` yields the same design effect estimated above using approach 1 as shown in the following table:

Estimated Mean Age of U.S. Adults (BRFSS, 2010), Standard Error and Design Effect,
Revised from Proc SurveyReg Output

Obs	Estimate	DesignSE	RevisedDE
1	47.2934240	0.05465389	4.40141

Example 2: Computing the Total Number of Adults in Missouri Who Have Ever Been Told of Diabetes from a Physician (MO BRFSS, 2010)

Based on a recoding of the diabete2 variable for the Missouri portion of the 2010 BRFSS data file, the BRFSS stratification variable – _ststr and the sampling weight variable – _finalwt, we apply Approach 1 to estimate the total number of adults who have ever been told by a physician that they have diabetes. The hasesugar variable below is coded as 1 for those who have been told of diabetes and 0 for those adults in MO who have not.

STEP 1:

```
proc surveymeans data=state.MObrfss sum nomcar;
  strata _ststr;
  weight _finalwt;
  var hasesugarbin;
  ods output Statistics = SampDesignOUT (rename=StdDev=DesignSE keep=StdDev);
run;
```

STEP 2: Note – the vardef option for totals is set to the default value (df).

```
proc univariate data=state.MObrfss vardef=DF;
  var hasesugarbin;
  weight _finalwt;
  ods output moments=SRSOUT;
run;

proc transpose data=SRSOUT (keep=nvalue2)
  prefix=Stat
  /* The output data set below will contain information that will be used to
  compute the DENOMINATOR of the DESIGN EFFECT estimate of the TOTAL*/
  out=SRSOut (rename=stat1=Nhat rename=stat3=SRSvar rename=stat2=TotalEst
  keep=stat1 keep=stat2 keep=stat3);
run;
```

STEP 3:

```
data state.BRFSSMODesignEffectsFinal;
  merge SampDesignOut SRSout;
  DESIGN_Effect=(DesignSE**2)/(SRSVar*Nhat);
run;

proc print data=state.BRFSSMODesignEffectsFinal;
  var TotalEst DesignSE Design_Effect;
  title 'Estimated Total number of Adults living in Households in MO who have
  been told of Diabetes by a Physician (MO-BRFSS, 2010), Standard Error and
  Design Effect';
run;
```

Estimated Total number of Adults living in Households in MO who have been told of Diabetes by a Physician (MO-BRFSS, 2010), Standard Error and Design Effect

Obs	TotalEst	DesignSE	DESIGN_Effect
1	426646	22572	1.57640

Example 3: Estimating the Proportion of U.S. Adults Who Are Parents of Children Under 18 Who Have Some Type of Health Insurance Coverage

For this example, we make use of the BRFSS, 2010 stratification and weighting variables previously described. We also make use of the binary healthplan variable (e.g. 1 if any type of health insurance coverage, 0 otherwise), hlthplanbin as our outcome of interest and created a parent or guardian (POG) variable for adults who have at least one child living in the household. Those who do (i.e. POG=1) form our domain of interest.

STEP 1:

```
proc surveymeans data=state.brfss2010 mean nomcar;
  strata _ststr;
  weight _finalwt;
  domain POG;
  var hlthplanbin;
  title 'Percentage of Adults with Children who have Health Insurance';
  /*The output data set below will contain information that will be used to
  compute the NUMERATOR of the DESIGN EFFECT estimate of the MEAN for the POG=1
  Subdomain*/
  ods output Domain=BRFSS_SampDesignOutSubDom
    (rename=Stderr=DesignSE keep=Stderr keep=mean keep=POG);
run;
```

STEP 2:

```
proc univariate data=state.brfss2010 vardef=WGT;
  where POG=1;
  var hlthplanbin;
  weight _finalwt;
  ods output moments=BRFSSMomentsSubDom;
run;
quit;
ods html close;

/*Transpose moments table to readily capture data on Sample Size
and Standard Error of the Mean under SRS for the Subdomain POG=1*/
proc transpose data=BRFSSMomentsSubDom (keep=nvalue1)
  prefix=Stat
  /* The output data set below will contain information that will be used to
  compute the DENOMINATOR of the DESIGN EFFECT estimate of the MEAN for POG=1
  Subdomain*/
  out=BRFSS_SRSOutSubDom (rename=stat1=n_size rename=stat3=SRSSE keep=stat1
  keep=stat3);
run;
```

STEP 3:

```
/*Now combine the two output data files and compute the Design Effects
directly using the output*/
```

```

data state.BRFSSDesignEffectSubDom;
merge BRFSS_SampDesignOutSubDom (where=(POG=1)) BRFSS_SRSoutSubDom;
DESIGN_Effect=(DesignSE/SRSSE)**2*(n_size-1);
run;
quit;

ods html;
proc print data=state.BRFSSDesignEffectSubDom;
where POG=1;
var mean DesignSE Design_Effect;
title 'Estimated Percentage of U.S. Adults Who Are Parents/Guardians Who Have
Some Type of Health Insurance Coverage (BRFSS, 2010), Standard Error and
Design Effect';
run;

```

The resulting output for executing the syntax in steps 1-3 is given in the following table.

Estimated Percentage of U.S. Adults Who Are Parents/Guardians Who Have Some Type of Health Insurance Coverage (BRFSS, 2010), Standard Error and Design Effect

Obs	Mean	DesignSE	DESIGN_Effect
1	0.819415	0.002224	4.02567

Conclusion

While design effects are available for many statistics of interest in a majority of the analytic survey procs, they continue to be unavailable in proc surveymeans. By using either a null regression model or applying surveymeans and proc univariate in a sequential fashion, we have demonstrated how to derive correct design effects for means and/or proportions. Design effects for totals can also be computed using the means/re-means approach which also lends itself to easily computing design effects for means, totals and proportions for subdomains of interest.

References

- An, Anthony and Donna Watts. 1998 "New SAS Procedures for Analysis of Sample Survey Data." Proceedings of the Twenty-Third Annual SAS® Users Group International (SUGI) Conference, Nashville, TN.
- Baisden, K.L. and Hu, P. (2005) "The Enigma of Survey Data Analysis: Comparison of SAS ® Survey Procedures and SUDAAN Procedures," Paper 194-31, SUGI 31.
- Berglund, P. (2001) "Analysis of Complex Sample Survey Data Using the SURVEYMEANS and SURVEYREG Procedures and Macro Coding" SUGI 27.
- Berglund, P.A. (2009) "Getting the Most out of the SAS ® Survey Procedures: Repeated Replication Methods, Subpopulation Analysis, and Missing Data Options in SAS ® v9.2," Paper 246-2009, SUGI 2009.
- Lohr, S.L. (2010) Sampling Design and Analysis, 2nd Edition, Boston: Brooks/Cole.
- Research Triangle Institute (RTI) (2008). SUDAAN User's Manual, Release 10.0, Research Triangle Park, NC: Research Triangle Institute.

Acknowledgements

Special thanks to my family for their support during many long nights of SAS programming and writing. Also, thanks to the students of SLU-SPH BST-542, Spring 2011 for working through all the SAS Survey Procedures – an exercise which motivated this paper.

Contact Information

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

Trent D. Buskirk, Ph.D.
Department of Biostatistics, Saint Louis University's School of Public Health
3545 Lafayette Avenue, Salus Center 3rd Floor
Saint Louis, MO 63104
Phone: 314-698-7212
E-mail: trent.buskirk@gmail.com

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.