



## Getting the Most out of the SAS® Survey Procedures: Repeated Replication Methods, Subpopulation Analysis, and Missing Data Options in SAS® v9.2

# Overview of Presentation Topics

- **Practical guidance on survey data analysis techniques using new features in the SAS® 9.2 Survey procedures**
  - **Repeated Replication methods for variance estimation**
    - Jackknife/Balanced Repeated Replication methods for variance estimation
  - **Subpopulation analyses**
    - Subpopulation analysis via the **DOMAIN** option
  - **Techniques for handling missing data**
    - **NOMCAR** (not missing completely at random) option and multiple imputation options for handling missing data

# Background Information on Complex Data

- Complex surveys are derived from sample designs that adjust for non-response and differing probabilities of selection.
- Complex samples differ from simple random samples as SRS designs assume independence of observations while complex samples do not.
- Most SAS® procedures assume a simple random sample and under-estimate variances when analyzing data from complex samples.
- Analysis of data derived from complex samples should include methods to properly account for the sample (**PROC SURVEYMEANS, SURVEYFREQ, SURVEYREG, SURVEYLOGISTIC**).

# Example Data and Complex Sample Variables

## ■ Data sets

- **National Comorbidity Survey Replication (NCS-R)**, with a focus on mental health.
- **National Health and Nutrition Examination Survey (NHANES)** 1999-2000 data set, with a focus on the medical examination data.

## ■ Complex design variables

- Both data sets include variables to incorporate the survey design into variance estimation computations.
- **Stratum** and **SECU** (Sampling Error Computing Unit) in the NCS-R data or **Replicate weights** in the NHANES data.
- **Probability Weights** to adjust for non-response and differing probabilities of selection.

# Variance Estimation Methods

## ■ Taylor Series Linearization Method

- Based on a method that derives a linear approximation of variance estimates that are in turn used to develop corrected standard errors and confidence intervals for statistics of interest.
- **SURVEYMEANS, SURVEYFREQ, SURVEYREG and SURVEYLOGISTIC** procedures all use the Taylor Series method as the default.
  - Very efficient computationally and appropriate for most statistics.
  - Cannot be used with data sets that provide only replicate weights.

*continued...*

# Variance Estimation Methods

## ■ Resampling Methods

- Follow a specified method of selecting observations defined as probability sub-samples or replicates, calculate the statistic of interest from each replicate and calculate variance estimates from replicate distributions.
- Methods include **Balanced Repeated Replication (BRR)** and **Jackknife Repeated Replication (JRR)**.
  - Preferred for some non-linear statistics and in situations where the coefficient of variation exceeds 0.2 from the Taylor Series linearization method.
  - Additional advantages of repeated replication methods are the simplicity of the process and applicability to a wide range of statistics.

*continued...*

# Repeated Replication Methods

## ■ Use of Repeated Replication Methods

- Some projects publish replicate weights rather than stratum and/or cluster variables, generally due to concerns about protecting respondent confidentiality.
- Providing only replicate weights limits the choice of appropriate variance estimation methods to repeated replication approaches.
- With the addition of Repeated Replication techniques in SAS® v9.2, the analyst can now use SAS® Survey procedures with the appropriate repeated replication method.

# Comparison of Variance Estimation Methods

- Comparison of Taylor Series Variance method with Jackknife and Balanced Repeated Replication methods
  - No use of “**varmethod=**” defaults to the Taylor Series method
  - Use “**varmethod=jackknife**” for the JRR method
  - Use “**varmethod=brr**” for BRR.

```
1. proc surveymeans data=ncsr mean stderr cv;  
2. proc surveymeans data=ncsr varmethod=jackknife;  
3. proc surveymeans data=ncsr varmethod=brr (printh);
```

- Example uses NCS-R data to perform a means analysis of **Major Depressive Disorder**.

*continued...*

# Prevalence of Major Depressive Episode

- Examination of Table 1.1 shows little difference between the three methods in the standard errors, not always the case.
- Note that the means are the same (as they should be) but the standard errors are slightly different.

Variable	Label	Mean	Std Error of Mean	Coeff of Variation
DSM_MDE	DSM-IV Major Depressive Episode(Lifetime) (Method=Taylor Series)	0.191711	<b>0.004877</b>	0.025438
	DSM-IV Major Depressive Episode(Lifetime) (Method= JRR)	0.191711	<b>0.004879</b>	NA
	DSM-IV Major Depressive Episode(Lifetime) (Method= BRR)	0.191711	<b>0.004952</b>	NA

# Analysis with Replicate Weights

- This example uses NHANES data to analyze diastolic blood pressure
- NHANES 1999-2000 data released with only replicate weights specified as JRR weights, therefore calls for the use of the **Jackknife Repeated Replication method**.

```
proc surveymeans data=nhanes9900 varmethod=jackknife;  
repweights wtmrep01-wtmrep52;  
weight wtmecl2yr; var bpxdi1; run;
```

Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
BPXDI1	Diastolic: Blood pres (1st rdg) mm Hg	6457	70.233553	0.518031	69.1940482	71.2730571

# Subpopulation Analysis

- Analysis of **subpopulations** is a common analytic practice yet confusion about how to correctly analyze subpopulations from survey data persists.
- With the addition of a **DOMAIN** statement in **PROC SURVEYREG/PROC SURVEYLOGISTIC** (new in SAS® 9.2), the analyst can correctly perform **DOMAIN** analyses in each of the main Survey analysis SAS® procedures (**SURVEYMEANS, SURVEYFREQ, SURVEYREG, and SURVEYLOGISTIC**).

# Domain Analysis and the “Subset” Approach

- A **DOMAIN** analysis uses all of the original sample through use of an indicator variable (ex. age 50+ or not, coded 1/0).
  - Correctly performs subpopulation analyses within each group of the indicator.
- Why is a subset approach (by/where/if) incorrect?
  - Problems with strata with “**singleton**” clusters due to the sub-setting step, interferes with variance estimates.
  - A “BY” group approach analyzes only a part of the original sample and **ignores the variability of the domain sample sizes** across the strata of the sample design. The result is **under-estimated variances**.
- The **DOMAIN** statement addresses these problems.

# Subpopulation Analysis Comparison

- Example: a comparison of prevalence of **Alcohol Abuse and Dependence** among those with **General Anxiety Disorder and Major Depressive Disorder and Drug Abuse (n=88)** with a **domain** and **where** statement (NCS-R data).
- Use of the **domain** versus a **where** statement changes the method while the rest of the code stays the same. This will produce very different sample sizes and standard errors.

```
proc surveymeans data=ncsr;  
strata str; cluster secu; weight finalp2wt;  
domain gad_mde_dra; *note: this is the domain statement approach;  
where gad_mde_dra=1; *note: this is the where statement approach;  
var dsm_ala dsm_ald; run;
```

*continued...*

# Subpopulation Analysis Comparison

Table 3.2 Domain Analysis				
gad_mde_dra	Variable	Label	Mean	Std Error of Mean
0	DSM_ALA	DSM-IV Alcohol Abuse(Lifetime)	0.125227	0.005485
	DSM_ALD	DSM-IV Alcohol Dependence(Lifetime)	0.049645	0.003283
1	DSM_ALA	DSM-IV Alcohol Abuse(Lifetime)	0.836008	<b>0.039781</b>
	DSM_ALD	DSM-IV Alcohol Dependence(Lifetime)	0.513451	<b>0.051268</b>

Table 3.4 Statistics			
Variable	Label	Mean	Std Error of Mean
DSM_ALA	DSM-IV Alcohol Abuse(Lifetime)	0.836008	<b>0.032639</b>
DSM_ALD	DSM-IV Alcohol Dependence(Lifetime)	0.513451	<b>0.036225</b>

Note: SE's are larger in the DOMAIN analysis, correctly including the entire sample.

# Missing Data and Survey Data Analysis

- Methods for handling missing data
  - **Exclude missing data**, default in most procedures.
  - **Use of the NOMCAR (“Not Missing Completely at Random”) option**, new in SAS® v9.2.
    - The **NOMCAR** feature includes and analyzes missing data cases as a separate **domain** rather than simply excluding these cases from the analysis.
  - **Impute missing data** prior to analysis (**PROC MI**), analyze imputed data sets using standard (including Survey) SAS® procedures, and then use tools for correctly analyzing imputed data sets (**PROC MIANALYZE**).

# Missing Data Example

- Data is from the NCS-R, variable of interest is **respondent income**.
- Missing data is simulated and is contained within a set of Stratum and SECU (Strata #35-42 and SECU=1), **857 cases are missing** (overall n=5692).
- Data is clearly **not missing at random**
  - Reason to assume excluding missing data would omit important information about original sample and additional respondent information in a multivariate analysis.

# Comparison of Methods for Handling Missing Data

- Methods of handling missing data on respondent income
  - **1 and 2. Exclusion of missing data in PROC MEANS and PROC SURVEYMEANS**
  - **3. Use of the NOMCAR (not missing completely at random) where missing data is analyzed as a separate domain in PROC SURVEYMEANS**
  - **4. Multiple imputation of missing data using PROC MI and subsequent analysis of imputed data sets using PROC SURVEYMEANS and PROC MIANALYZE.**
- Differing approaches illustrate how the method of handling missing data can affect **variances**.

## Exclude Missing Data

- Example: use of **PROC MEANS** and **PROC SURVEYMEANS** but both procedures **exclude missing data** by default.
- Given that this is **survey data**, the use of **PROC SURVEYMEANS** is appropriate for this analysis.

```
proc means data=ncsr nmiss sum mean std stderr min max;  
var inc_rsp; weight finalp2wt; run;  
  
proc surveymeans data=ncsr nmiss mean stderr ;  
strata str; cluster secu; weight finalp2wt;  
var inc_rsp; run;
```

## Use of NOMCAR Option

- Example: use of the **NOMCAR** option on the **PROC SURVEYMEANS** procedure statement.
- This option specifies that the **missing data be analyzed as a separate domain** rather than excluded from the analysis. (See full output in paper for more information)

```
proc surveymeans data=ncsr nomcar nmiss mean stderr;  
strata str; cluster secu;  
weight finalp2wt; var inc_rsp;  
run;
```

```
var inc_rsp;  
strata str;  
cluster secu;  
weight finalp2wt;  
run;
```

# Multiple Imputation of Missing Data

- Step 1: Multiple Imputation Step (**PROC MI**)

```
proc mi nimpute=5 data=ncsr out=outmi1;  
class sexf; monotone method=reg;  
var sexf age inc_rsp; run;
```

- Step 2: Analysis using standard SAS Procedure (**PROC SURVEYMEANS**)

```
ods output domain=outsummary;  
proc surveymeans data=outmi1;  
domain _imputation_; var inc_rsp;  
strata str; cluster secu; weight finalp2wt; run;  
Ods output close ;
```

*continued...*

# Multiple Imputation of Missing Data

- Step 3: **Use of PROC MIANALYZE** for analysis of output from Steps 1 and 2
  - 5 imputed data sets from **PROC MI** used in **SURVEYMEANS** analysis
  - “outsummary” data set saved from **PROC SURVEYMEANS** and used in **PROC MIANALYZE**
- This multi-step process imputes missing data, analyzes the imputed data sets using **PROC SURVEYMEANS** and then analyzes the SURVEYMEANS output and the variability introduced by imputation using **PROC MIANALYZE**.

```
proc mianalyze data=outsummary;  
model effects mean; stderr stderr;  
Run;
```

# Summary of Missing Data Results

- Table 4.13 illustrates how differing methods of handling missing data affect the standard errors.
- Use of the **NOMCAR** and **multiple imputation** methods produce the highest and likely most accurate standard errors due to inclusion of missing data as a separate domain or the imputation of missing data and analysis using **PROCS SURVEYMEANS** and **MIANALYZE**.

Table 4.13 Summary of Four Approaches

Variable (Method)	N Miss	Mean	Std Error of Mean
Inc_rsp (Proc Means with Missing Data Excluded)	857	24837	395.444
Inc_rsp (Proc SurveyMeans without NOMCAR)	857	24837	783.534
Inc_rsp (Proc SurveyMeans with NOMCAR)	857	24837	910.411
Inc_rsp (Imputed and Analyzed with Proc MI and MIANALYZE)	None	25039	829.964

# Conclusion

- This session has provided practical guidance on ways to optimize new features in the SAS® v9.2 Survey procedures
  - **Repeated Replication** methods for variance estimation.
  - Subpopulation analyses with use of the **DOMAIN** statement.
  - New options to handle missing data in survey data analyses such as the **NOMCAR** option and/or use of **multiple imputation techniques** in SAS ® v9.2.

# Questions?

- Comments/questions are welcome!
- Author contact: [pberg@umich.edu](mailto:pberg@umich.edu)