Advice to Health Services Researchers: Be Cautious Using the "Where" Statement in SAS® Programs for Nationally Representative Complex Survey Data

Hemalkumar B. Mehta, Michael L. Johnson

Department of Clinical Sciences and Administration, College of Pharmacy, University of Houston, Houston, TX 77030

ABSTRACT:

Health services researchers often conduct research with nationally representative survey data where participants or patients are not sampled randomly but sampled using complex stratified multistage probability designs. Such datasets include cluster, strata and weight information which are essential for extrapolation of results to a national level. Several Proc Survey procedures are available in SAS® 9.2 which enables analysis of such data while preserving the complex sampling design and extrapolation of results. The first step researchers often perform is selection of a population of interest, i.e. selection of participants with certain inclusion criteria, from the main dataset. This can be accomplished in SAS® using the 'where' statement in data steps. However, use of the where statement for selecting a population of interest can defeat the purpose of the sampling design of such data and limits researcher's ability to generalize results. In the current paper using Medical Expenditure Panel Survey (MEPS) data, a nationally representative multistage probability survey, we show how to analyze such data while preserving sampling design and not using the where statement. The principles and techniques explained in this paper can be extended to any other disciplines where the researcher has to deal with complex survey data which involves cluster, strata and weight information in sampling design of the data.

Key words: Proc Survey procedures, Where statement, Multistage probability design, Health Services Research

INTRODUCTION:

Several nationally representative datasets pertaining to health services research use multi-stage probability sampling design for data collections (Table 1). Traditional SAS procedures, such as the PROC MEANS and the PROC REG procedure, compute statistics under the assumption that the sample is drawn from an infinite population by simple random sampling. These procedures generally do not correctly estimate the variance of an estimator if they are applied to a sample drawn by a complex sample design.[1]

Table 1: Selected nationally representative dataset which employs complex multi-stage probability survey design *

Data set	Website
Medical Expenditure Panel Survey (MEPS)	http://www.meps.ahrq.gov/mepsweb/
National Ambulatory Medical Care Survey (NAMCS)	http://www.cdc.gov/nchs/ahcd/about_ahcd.htm#NAMCS
National Hospital Ambulatory Medical Care Survey (NHAMCS)	http://www.cdc.gov/nchs/ahcd/about_ahcd.htm#NHAMCS
Behavioral Risk Factor Surveillance System (BRFSS)	http://www.cdc.gov/brfss/
National Health Nutrition and Examination Survey (NHANES)	http://www.cdc.gov/nchs/nhanes/about_nhanes.htm
Health Information National Trend Survey (HINTS)	http://hints.cancer.gov/
National Survey of Children's Health (NSCH)	http://childhealthdata.org/learn/NSCH
National Survey of Children with Special Health Care Needs (NS-CHSCN)	http://childhealthdata.org/learn/NS-CSHCN
National Health Interview Survey (NHIS)	http://www.cdc.gov/nchs/nhis.htm
National Hospital Discharge Survey (NHDS)	http://www.cdc.gov/nchs/nhds.htm
National Nursing Home Survey (NNHS)	http://www.cdc.gov/nchs/nnhs.htm
National Home and Hospice Care Survey (NHHCS)	http://www.cdc.gov/nchs/nhhcs.htm
National Home Health Aide Survey (NHHAS)	http://www.cdc.gov/nchs/nhhas.htm
National Survey of Residential Care Facility (NSRCF)	http://www.cdc.gov/nchs/nsrcf.htm

*This list is not comprehensive.

While working with multi-stage probability survey data, researchers must use survey procedures and should incorporate sample design to make statistically valid inferences for the population of interest. SAS provides the following survey procedures to analyze sample survey data with multi-stage probability sampling deign. These survey procedures give flexibility to analyze data collected using single or multi-stage sampling designs, with or without stratification, and with or without equal or unequal weighting.

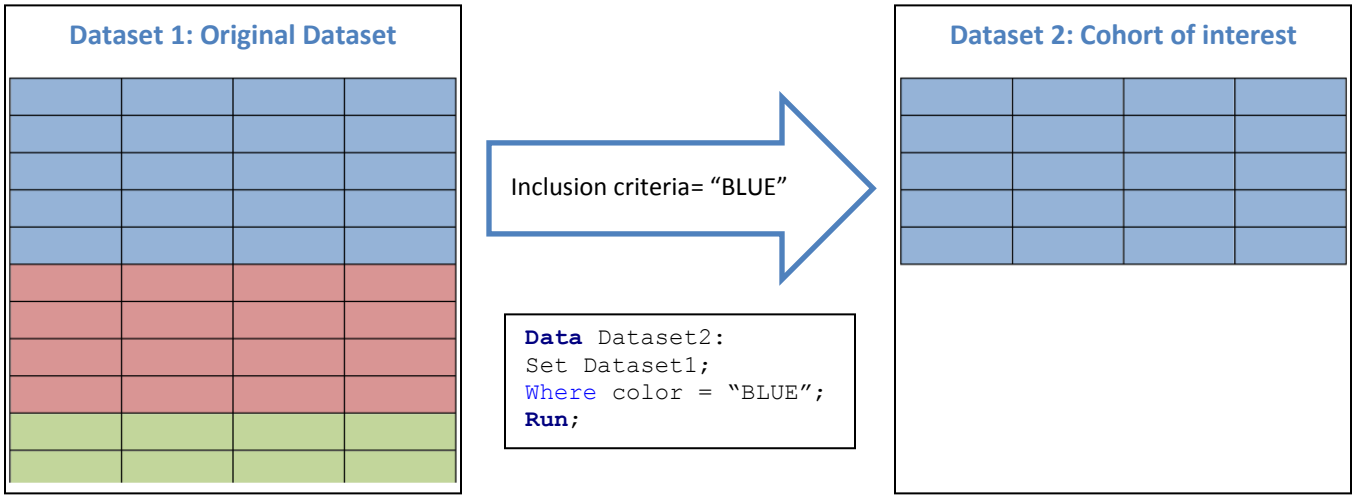
- Proc Surveymeans
- Proc Surveyfreq
- Proc Surveylogistic
- Proc Surveyreg
- Proc Surveyphreg

Generally speaking, health services researchers do not perform statistical analysis on the entire dataset. The first step researchers often perform is selection of a population of interest, i.e. selection of participants with certain inclusion criteria, from the original dataset. This can be accomplished in SAS using the 'Where' statement in data steps. Whether a researcher performs unweighted (PROC MEANS) or weighted statistical analyses (PROC SURVEYMENAS) on the subset of the original dataset, it is incorrect. In this paper we show that instead of performing statistical analysis on sub-setted dataset, researcher should use the entire dataset and perform weighted statistical analysis.

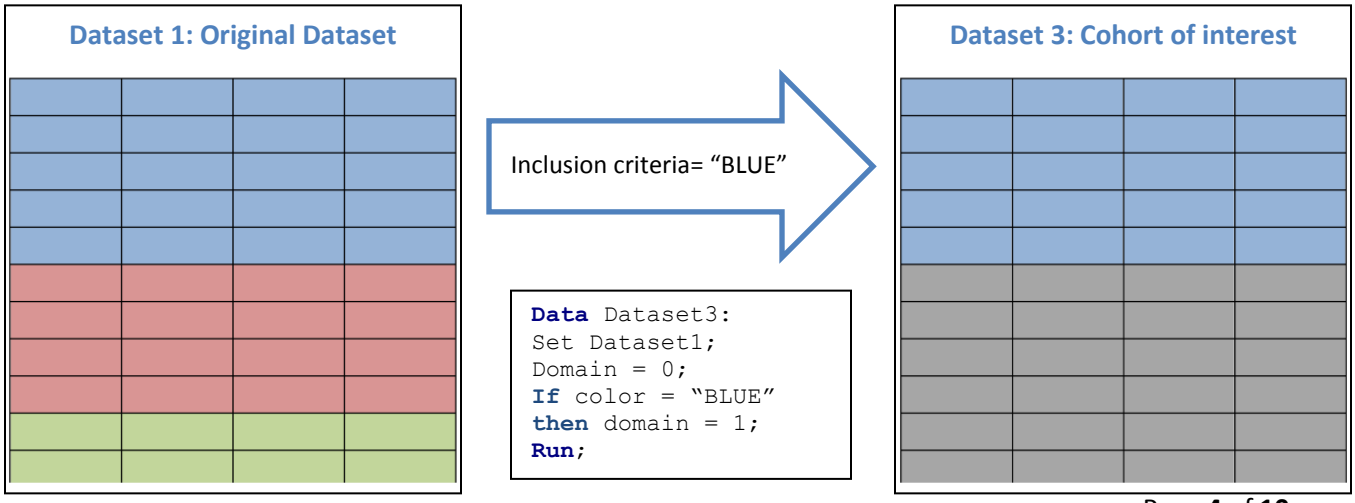
VISUAL DEMONSTRATION:

Dataset-1 is the original dataset containing all observations.

Approach – 1 (Incorrect): In this approach, the researcher will select observations that meet inclusion criteria. In the below example, inclusion criteria is 'BLUE' color; one can easily apply this inclusion criterion in SAS using 'Where' statement in the data steps and obtain dataset-2. Whether a researcher performs subsequent unweighted or weighted analysis on dataset-2, results will not be valid.



Approach – 2 (Correct): The correct approach is to make a dummy variable with value = 1 for the population of interest and value = 0 for rest of the population. In this case, new variable called "Domain" will be created in Dataset-3; for blue color, domain =1 and for rest of the observations domain = 0. One should perform weighted analysis on the Dataset – 3 including proper domain statement.



EMPIRICAL PROJECT:

Objective:

To find mean prescription medication expenditure among obese adults.

Research Methods:

Dataset:

The empirical project used Medical Expenditure Panel Survey (MEPS) – 2008 dataset. The MEPS is a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States. MEPS collects data on the specific health services that Americans use, how frequently they use them, the cost of these services, and how they are paid for, as well as data on the cost, scope, and breadth of health insurance held by and available to U.S. workers. The MEPS is not a simple random sample; it utilizes complex multistage probability design employing clustering, stratification and weighting which enables researchers to extrapolate results at the national level. The MEPS is considered as a reliable source of prescription drug use at national level because it employs several measures to make prescription data more accurate such as cross-checking of patient reported information at pharmacy store.

The full year consolidated data file was used for this project which is freely accessible through http://www.meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp.

Definitions/Inclusion criteria:

The study included obese adults with non-zero prescription drug expenditure with positive sampling weight. Adults were defined as Age ≥ 20 and obese patients were defined as Body mass Index (BMI) ≥30.

Statistical Analysis:

For the demonstration purpose, we performed both unweighted and weighted analysis to estimate mean prescription drug expenditure. Proc means and Proc Surveymeans were used for unweighted and weighted analysis, respectively.

(i) Approach – 1: In this case, population of interest was selected (adult obese with non-zero prescription drug expenditure and positive sampling weight). Unweighted and weighted analysis was performed on this cohort.

(ii) Approach –2: In this case, new variable called "Domain" was created; adult obese patients with non-zero prescription drug expenditure received a value of one and remaining patients received a value of zero. Weighted analysis was performed on this cohort including proper domain statement.

All statistical analyses were performed in the SAS® 9.2.

Results:

The original MEPS 2008 dataset consisted of 33,066 observations (meps08).

Approach –1: After applying inclusion criteria on the original dataset, the new dataset consisted of 4,539 observations (meps08_1).

SAS commands:

```
Data meps08_1;
Set meps08;
Where age08x >=18 and BMINDX53 >=30 and RXEXP08 > 0 and PERWT08F > 0;
Run;
```

Unweighted and weighted analysis was performed on meps08_1 dataset to estimate mean prescription expenditure.

SAS commands:

```
Title "Mean Expenditure using Proc means - Unweighted analysis
      (Approach -1)";
Proc means data = meps08_1 n mean stderr clm;
Var RXEXP08;
Where PERWT08F >0;
Run;
```

```
Title "Mean Expenditure using Proc Surveymeans - Weighted analysis
      (Approach -1)";
Proc surveymeans data = meps08_1  ;
Var RXEXP08;
Strata VARSTR;
Cluster  VARPSU ;
Weight PERWT08F;
Run;
```

Mean Expenditure using Proc means - Unweighted analysis (Approach -1)

The MEANS Procedure

Analysis Variable : RXEXP08 TOTAL RX-EXP 08				
N	Mean	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
4539	1742.67	43.9830704	1656.44	1828.89

Mean Expenditure using Proc Surveymeans - Weighted analysis (Approach -1)

The SURVEYMEANS Procedure

Data Summary	
Number of Strata	165
Number of Clusters	361
Number of Observations	4539
Sum of Weights	47456878.6

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
RXEXP08	TOTAL RX-EXP 08	4539	1847.654907	56.746470	1735.74285	1959.56696

Approach – 2: A new binary dummy variable called 'Domain' was created; patients meeting inclusion criteria received value of 1 for domain and others received value of 0 (meps08_2).

SAS commands:

```
Data meps08_2;
Set meps08;
If age08x >=18 and BMINDX53 >=30 and RXEXP08 > 0 and PERWT08F > 0 then
      DOMAIN = 1;
Else DOMAIN = 0;
Run;
```

Weighted analysis was performed on meps08_2 dataset to estimate mean prescription expenditure while incorporating domain statement.

```
Title "Mean Expenditure using Proc Surveymeans - Weighted analysis
      (Approach -2)";
Proc Surveymeans data = meps08_2;
Var RXEXP08;
```

```
Strata VARSTR;
Cluster VARPSU;
Weight PERWT08F;
Domain domain;
Run;
```

Mean Expenditure using Proc Surveymeans - Weighted analysis (Approach -2)

The SURVEYMEANS Procedure

Data Summary	
Number of Strata	165
Number of Clusters	370
Number of Observations	33066
Number of Observations Used	31262
Number of Obs with Nonpositive Weights	1804
Sum of Weights	304375942

Statistics						
Variable	**Label**	**N**	**Mean**	**Std Error of Mean**	**95% CL for Mean**	
RXEXP08	TOTAL RX-EXP 08	31262	812.929714	20.369219	772.769691	853.089738

Domain Analysis: DOMAIN							
DOMAIN	**Variable**	**Label**	**N**	**Mean**	**Std Error of Mean**	**95% CL for Mean**	
0	**RXEXP08**	TOTAL RX-EXP 08	26723	621.800153	19.589700	583.17703	660.42327
1	**RXEXP08**	TOTAL RX-EXP 08	4539	1847.654907	56.612641	1736.03723	1959.27259

Table 2: Comparison of results

	Approach -1		Approach-2
	Unweighted analysis	**Weighted analysis**	**Weighted analysis**
N	4539	4539	4539
Mean	1742.67	1847.654907	1847.654907
Std Error of Mean	43.9830704	56.746470	56.612641
95% CL for Mean	(1656.44, 1828.89)	(1735.74285, 1959.56696)	(1736.03723, 1959.27259)

Discussion:

Table 2 compares the results obtained from two approaches. Number of observation used to estimate mean prescription expenditures were same. In the first approach where unweighted analysis was performed, the mean value of prescription medication expenditure and standard error was different compared to weighted analysis of approach-2. The unweighted analysis assumes simple random sampling design of the data and does not take into account complex survey design; hence it gives incorrect estimate of mean value and underestimation of the standard error.[2, 3]

The estimated mean prescription expenditure is exactly same when weighted analysis was performed employing either approach 1 or 2. However, the standard error of mean is more when weighted analysis was performed on the subsetted dataset (approach-1). Also, the 95% confidence interval is wider.

The weighted analysis with proper domain statement (approach-2) estimated the correct mean prescription expenditure and standard error. The 95% confidence interval limit is also narrower.

The lesson to learn is that use of Proc Survey procedures may give an incorrect answer if the researcher is performing statistical analysis on the sub-setted dataset/population of interest. While working with complex survey data which utilizes multi-stage probability sampling design, one should perform weighted analysis using survey procedures on the entire dataset incorporating proper domain statement rather than performing analysis on the sub-setted dataset. In the current empirical project, we have shown differences in results using only one survey procedures (Proc Surveymeans). Future research can compare the results employing different survey procedures.

CONCLUSION:

Researcher should not select a population of interest using 'Where' statement or by any other means; performing weighted or unweighted statistical analysis on this subsetted population will give incorrect answer. The right approach is to keep all observations and creating a dummy variable (value=1 for population of interest and value=0 otherwise); performing appropriate weighted statistical analysis on this entire dataset with proper domain statement will give correct answer while preserving the complex survey design. It also enables researcher to extrapolate results at the national level.

REFERENCES:

1. Sample Survey Design and Analysis. Available at: http://support.sas.com/rnd/app/da/new/dasurvey.html
2. An, A., Watts D. New SAS Procedures for Analysis of Sample Survey Data. Available at: http://www.ats.ucla.edu/stat/sas/library/svy_survey.pdf
3. Chen, X., Gorrell, P. An Introduction to the SAS® Survey Analysis PROCs. NESUG 2008. Available at: http://www.nesug.org/proceedings/nesug08/sa/sa06.pdf

Contact Information:

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

Hemalkumar B. Mehta, MS
PhD Student (Pharmacy Administration),
Department of Clinical Sciences and Administration,
College of Pharmacy, University of Houston,
Texas Medical Center
1441Moursund Street,
Houston, TX 77030

Phone: 718-607-4967
E-mail: hbmehta3@uh.edu
Web: www.mehtahemal.com

Michael L. Johnson, PhD
Associate Professor,
Director of Graduate Studies,
Department of Clinical Sciences and Administration,
College of Pharmacy, University of Houston,
Texas Medical Center
1441Moursund Street,
Houston, TX 77030

Phone: 713-795-8353
E-mail: mikejohnson@uh.edu