Austin Biometrics and Biostatistics

Open Access

Austin Publishing Group

Review Article

Survey Tables Binary: A SAS Macro for Publication Quality Tables of Complex Survey Data

Sunesara I[1]*, Lirette ST[1] and Griswold ME[1]
Department of Center of Biostatistics and Bioinformatics, University of Mississippi Medical Center, USA

***Corresponding author:** Sunesara I, Center of Biostatistics and Bioinformatics, University of Mississippi Medical Center, 2500 N State St, Jackson, MS, 39216, USA

Received: September 16, 2015; **Accepted:** December 08, 2015; **Published:** December 14, 2015

Abstract

Production of publication-quality tables can be time consuming and tedious. The repetitive copy/paste or the often inaccurate typing by hand is less than optimal solutions for a very common problem. Proc survey in SAS is a very powerful tool for complex multistage probability sampling designs, but digesting the output can be overwhelming. We present a SAS macro that gives the user concise publication quality tables for complex survey data which uses design variables such as stratification, clustering and sampling weights.

Keywords: Complex survey; Multi-stage sampling; Design variables; Population; SAS; Tables

Introduction

SAS proc survey procedures are available to handle complex Multi-Stage Probability Sampling Designs (MDPS), each producing a plethora of analytic output. Unlike other procedures in SAS and competing statistical packages, the survey procedures provide appropriate parameter estimates from a known probability sample by incorporating the necessary design weights. Generally the output produced is extremely valuable to the researcher but is not output in a concise, publishable format. Even when using ODs export functions of tables into output destinations such as html, pdf or rtf formats, the output often requires post transfer processing. Producing publication-quality tables by copying and pasting into formatted shells can be tedious, laborious, and prone to typing errors as well as needing further processing. In this paper we present a SAS macro which automates the production of publication ready tables for complex sampling survey data directly from SAS using the ODs capabilities. We illustrate the macro using a sample from the National Health and Nutritional Education Survey (NHANES) [1]. This study uses multi-stage sampling procedures, which introduces design variables for stratification and clustering, similar to the Medical Monitoring Project [2], and related sampling weights for analysis in order to infer back upon the population of interest from which the sampling frame was derived. In this work, we are most interested in estimates of population prevalence and, therefore, limit the macro mainly to producing proportions and their associated measures of variance and confidence.

Description of Example Datasets

For our example, a combined dataset (N=5871) of NHANES from years 2001 - 2006 is used for show-casing the macro. The dataset includes the subset of variables from NHANES shown in Table 1. Using this example data set; we wish to create (Tables 2 & 3) for demographic characteristics of our sample to illustrate the macro.

Features and options

Variance: For variance computation necessary to provide confidence intervals and errors, only Taylor series estimation [3] is currently available in the macro. The survey procedures in SAS do include resampling methods for variance estimation, such as, Balanced Repeated Replication (BRR) and Jackknife (JK); these additional methods are intended to be included in future releases and should be a straightforward addition.

Missingness: When requesting binary subgroup analysis, the default missingness structure for SAS survey procedures is Missing Completely at Random (MCAR) [4]. Therefore, the macro call assumes MCAR. The Not Missing Completely at Random (NOMCAR) option can be requested and is specified within the source code of the macro. The nomcar option is useful when one cannot assume data values are missing completely at random, and, thus, calculates the variance appropriately. This option applies only to Taylor series variance estimation [4]. However, as noted, this only applies to binary subgroup analysis (Table 2). For estimated means and percentages

Table1: Characteristics of participants

Characteristics	Levels	N(%)	95%CI
BMXBMI		23.08 (±0.12)*	[22.84,23.32]
RIDAGEYR		15.41 (±0.05)*	[15.3,15.52]
INDFMPIR		2.6 (±0.05)*	[2.49,2.71]
race	White	1603 (62.69%)	[58.37,67.01]
	Black	1965 (14.79%)	[11.93,17.64]
	Hispanic	2067 (16.95%)	[13.78,20.11]
	Others	236 (5.57%)	[4.3,6.85]
	Total	5871 (100%)	
RIDEXMON	Winter	3126 (42.27%)	[34.45,50.09]
	Summer	2745 (57.73%)	[49.91,65.55]
	Total	5871 (100%)	
RIAGENDR	Boys	2937 (50.64%)	[48.87,52.41]
	Girls	2934 (49.36%)	[47.59,51.13]
	Total	5871 (100%)	
vstatus	<48.1 nmol/L	2883 (31.2%)	[27.56,34.85]
	>=48.1 - 66.2 nmol/L	1690 (32.56%)	[30.13,34.99]
	>= 66.2 nmol/L	1298 (36.24%)	[32.85,39.63]
	Total	5871 (100%)	
RIDEXMON	Winter	3126 (42.27%)	[34.45,50.09]
	Summer	2745 (57.73%)	[49.91,65.55]
	Total	5871 (100%)	
bmigroup	< 85th percentile	5014 (87.21%)	[85.6,88.81]
	>= 85th percentile	857 (12.79%)	[11.19,14.4]
	Total	5871 (100%)	

*Estimated means and StdErr for Continous variables
Frequencies and Estimated percentages for categorical variables
‡ Estimates are considered unreliable. Data marked by an ‡ have a relative standard error of more than 30 %

Figure 1: Screenshot of Table 1 output for example dataset.

Citation: Sunesara I, Lirette ST and Griswold ME. Survey Tables Binary: A SAS Macro for Publication Quality Tables of Complex Survey Data. Austin Biom and Biostat. 2015; 2(4): 1028.

Table2: Characteristics of participants by Metabolic Syndrome

Characteristics	Levels	Total	Total(95%CI)	No	No(95%CI)	Yes	Yes(95%CI)	p-value
BMXBMI		23.08 (±0.12)*	[22.84,23.32]	22.58 (±0.11)*	[22.36,22.8]	31.73 (±0.6)*	[30.52,32.94]	<.0001
RIDAGEYR		15.41 (±0.05)*	[15.3,15.52]	15.39 (±0.05)*	[15.29,15.5]	15.68 (±0.18)*	[15.32,16.04]	0.0884
INDFMPIR		2.6 (±0.05)*	[2.49,2.71]	2.61 (±0.06)*	[2.5,2.72]	2.45 (±0.1)*	[2.24,2.66]	0.1035
race	White	1603 (62.69%)	[58.37,67.01]	1508 (58.99%)	[54.76,63.22]	95 (3.7%)	[2.74,4.66]	0.11
	Black	1965 (14.79%)	[11.93,17.64]	1889 (14.19%)	[11.44,16.94]	76 (0.6%)	[0.42,0.77]	
	Hispanic	2067 (16.95%)	[13.78,20.11]	1948 (15.98%)	[12.99,18.97]	119 (0.97%)	[0.67,1.27]	
	Others	236 (5.57%)	[4.3,6.85]	226 (5.38%)	[4.16,6.6]	10 (0.19%)‡	[0.03,0.36]	
	Total	5871 (100%)		5571 (94.54%)	[93.54,95.54]	300 (5.46%)	[4.46,6.46]	
RIDEXMON	Winter	3126 (42.27%)	[34.45,50.09]	2965 (40.22%)	[32.68,47.76]	161 (2.05%)	[1.47,2.62]	0.246
	Summer	2745 (57.73%)	[49.91,65.55]	2606 (54.32%)	[46.81,61.83]	139 (3.42%)	[2.5,4.33]	
	Total	5871 (100%)		5571 (94.54%)	[93.54,95.54]	300 (5.46%)	[4.46,6.46]	
RIAGENDR	Boys	2937 (50.64%)	[48.87,52.41]	2758 (47.1%)	[45.25,48.95]	179 (3.54%)	[2.73,4.34]	<.0001
	Girls	2934 (49.36%)	[47.59,51.13]	2813 (47.44%)	[45.7,49.17]	121 (1.92%)	[1.49,2.36]	
	Total	5871 (100%)		5571 (94.54%)	[93.54,95.54]	300 (5.46%)	[4.46,6.46]	
vstatus	<48.1 nmol/L	2883 (31.2%)	[27.56,34.85]	2701 (28.83%)	[25.45,32.21]	182 (2.38%)	[1.81,2.94]	<.0001
	>=48.1 - 66.2 nmol/L	1690 (32.56%)	[30.13,34.99]	1612 (30.76%)	[28.32,33.19]	78 (1.8%)	[1.31,2.29]	
	>= 66.2 nmol/L	1298 (36.24%)	[32.85,39.63]	1258 (34.96%)	[31.62,38.3]	40 (1.28%)	[0.75,1.81]	
	Total	5871 (100%)		5571 (94.54%)	[93.54,95.54]	300 (5.46%)	[4.46,6.46]	
RIDEXMON	Winter	3126 (42.27%)	[34.45,50.09]	2965 (40.22%)	[32.68,47.76]	161 (2.05%)	[1.47,2.62]	0.246
	Summer	2745 (57.73%)	[49.91,65.55]	2606 (54.32%)	[46.81,61.83]	139 (3.42%)	[2.5,4.33]	
	Total	5871 (100%)		5571 (94.54%)	[93.54,95.54]	300 (5.46%)	[4.46,6.46]	
bmigroup	< 85th percentile	5014 (87.21%)	[85.6,88.81]	4941 (85.56%)	[83.8,87.32]	73 (1.65%)	[1.04,2.25]	<.0001
	>= 85th percentile	857 (12.79%)	[11.19,14.4]	630 (8.98%)	[7.68,10.29]	227 (3.81%)	[3.12,4.5]	
	Total	5871 (100%)		5571 (94.54%)	[93.54,95.54]	300 (5.46%)	[4.46,6.46]	

*Estimated means and StdErr for Continous variables
Frequencies and Estimated percentages for categorical variables
‡ Estimates are considered unreliable. Data marked by an ‡ have a relative standard error of more than 30 %

Figure 2: Screenshot of Table 2 output for example dataset.

Table 1: Description of example dataset.

Variable Type	Variable Name	Variable Description	Variable Attribute
Popln* Characteristic	RIAGENDR	Gender, (Boys/Girls)	Categorical
Popln Characteristic	RIDAGEYR	Age at screening	Continuous
Popln Characteristic	BMIGROUP	Body Mass Index	Categorical
Popln Characteristic	RACE	Race	Categorical
Popln Characteristic	VSTATUS	Vitamin levels	Categorical
Subgroup	METSYN	Metabolic Syndrome	Categorical
Popln Characteristic	INDFMPIR	Family poverty index ratio	Continuous
Popln Characteristic	BMXBMI	Body Mass Index	Continuous
Design	SDMVSTRA	Sampling Stratum	Design
Design	SDMVPSU	Sampling Cluster	Design
Design	MEC6YR	Sampling Weight	Design

Footnote: Poplin*: Population

of overall participant characteristics (Table 1), a MCAR missingness structure is assumed.

Relative standard error: The Standard Error (STDErr) is primarily a measure of the sampling variability that occurs by chance when only a sample, rather than an entire universe, is surveyed [5,6]. Proper estimation of STDerr is important in providing appropriate estimates, p-values, and confidence intervals based on design weights. Relative Standard Error (RSE) is one of the criteria to check for reliability of estimates (mean or percent) [7]. RSE is obtaining by dividing the standard error by the estimate itself (RSE= STDErr / Estimate) [8]. The macro relies on understanding the order of computation, either row or column proportions as needed can be output. If the row option is specified in the macro, row proportions and STDErr will be calculated appropriately. Likewise, column proportions (the default) and STDErr can be calculated with the call option for clarity. The resulting RSE is then expressed as a percent, where 20% or 30% are commonly chosen as reliable estimates. For

Table 2: Table shell for overall participant's characteristics.

Characteristics	Levels	N (%) / MN (sd?)	95% CI
Body Mass Index			
Gender	Boys		
	Girls		
	Total		

this macro, the end user should specify 0.30 if they desire a cut point of 30% RSE. By default, the macro will calculate RSE at 20%. Unreliable estimates [7] based on RSE criteria only are marked by double dagger sign (‡) in the output generated by this macro at the specified RSE cut point.

Output: The macro creates a folder named "result" under the active directory that contains relevant output. If the folder similarly named is available all the output will be saved within it. Output file names consist of concatenation of (Tables 1 & 2), name of the data file, and suffix of current date and time.

Implementing the macro

Macro parameters: The macro call allows for several options as well as required fields as noted in Table 4.

To download the macro please uses the link (https://sites.google.com/site/imransunesara/macros-programs/sas-software).

Recommended steps to use the macro using example dataset.

Step 1) prepare the dataset: Apply formats to all categorical variables of interest. See appendix for details. Apply dummy coding (0=No, 1=Yes). Only necessary for (Table 2).

Step 2) Read in the Macro using %include statement.

Step 3) Plug in variables of interest.

% survey tables binary (strata = SDMVSTRA, cluster = SDMVPSU, weights = MEC6YR, data = Nhanes_01_06_metsys, categorical_vars = bmigroup RACE RIDEXMON RIAGENDR vstatus, continous_vars

Table 3: Table shell for binary (yes/no) subgroup (metabolic syndrome) with association statistics.

Characteristics	Levels	Total	Total (95%CI)	No	No (95%CI)	Yes	Yes (95%CI)	p-value
Body Mass Index								
Gender	Boys							
	Girls							
	Total							

Table 4: Macro parameters.

Parameter	Explanation	Mandatory/Optional
data	Dataset name only	Mandatory
groupvar	Binary Outcome or subgroup of interest (Should be coded as 0=No and 1=Yes) (Defines Columns to split)	Mandatory for Table 2
categorical_vars	Enter all categorical variables (e.g. Gender…) (Row Variables)	Mandatory
continous_vars	Enter all continuous variables (e.g. Age…) (Row Variables)	Mandatory
strata	Stratification variable	Mandatory
percent_kind	Row or Column percent (Default=column)	Mandatory
cluster	Cluster variable	Mandatory
weights	sampling weights	Mandatory
rse	Relative Standard Error (Default = 0.20) Input range 0.00 to 1.00 Recommended 0.20 or 0.30	Mandatory
title 1	Title for the Table of Overall Characteristics	optional
title 2	Title for the Table of Characteristics split by a binary variable	optional

= BMXBMI RIDAGEYR INDFMPIR, percent_kind = col, groupvar = metsyn, rse = 0.30, table1title = Characteristics of participants, table2 title = Characteristics of participants by Metabolic Syndrome);

Generated output: This macro uses ODs rtf and ODs markup (Excel xp tag set) [4]. Various outputs have been programmed into it, with and without grid lines (Figures 1 & 2) are screenshots of tables in the example data set.

Errors and limitations

Common errors and/or warning messages generated and displayed in the log file typically result from categorical variables (like race) having "zero" in one of the cells, due to which association statistics are not calculated. The final table produced will contain estimates, but the p-value will be excluded. Another possible error message could be "Lock is not available". The solution to this problem is to rerun the program. If error message persists, change the active directory to your project directory.

Conclusion

This macro helps in increasing productivity and reproducibility and also helps in preparing error free tables for summarizing data, reporting, and research publications.

Acknowledgement

Authors thank Dr. Warren May, Ph.D. for reviewing this manuscript. We would also like to thank the very supportive and informative SAS user community.

References

1. Zipf G, Chiappa M, Porter KS, Ostchega Y, Lewis BG, Dostal J. National health and nutrition examination survey: plan and operations, 1999-2010. Vital Health Stat 1. 2013; 1-37.
2. McNaghten AD, Wolfe MI, Onorato I, Nakashima AK, Valdiserri RO, Mokotoff E, et al. Improving the representativeness of behavioral and clinical surveillance for persons with HIV in the United States: the rationale for developing a population-based approach. PLoS One. 2007; 2: e550.
3. Rust K. Variance Estimation for Complex Estimators in Sample Surveys. Journal of Official Statistics. 1985; 1: 381-397.
4. SAS Institute Inc. SAS/STAT Software, Version 9.2. Cary, NC.
5. Schappert S, Burt C. Ambulatory Care Visits to Physician Offices, Hospital Outpatient Departments, and Emergency Departments: United States, 2001-2002. National Center for Health Statistics. Vital Health Stat. 2006; 1-66.
6. CDC. National Hospital Discharge Survey. 2014; 1979-1996.
7. Klein RJ, Proctor SE, Boudreault MA, Turczyn KM. Healthy People 2010 criteria for data suppression. Healthy People 2010 Stat Notes. 2002; 1-12.
8. Hing E, Cherry D, Woodell D. National Ambulatory Medical Care Survey: 2004 Summary. National Center for Health Statistics. Vital Health Stat. 2006; 1.

Citation: Sunesara I, Lirette ST and Griswold ME. Survey Tables Binary: A SAS Macro for Publication Quality Tables of Complex Survey Data. Austin Biom and Biostat. 2015; 2(4): 1028.