SESUG Paper 226-2019

A user-friendly and robust macro that produces a publication-ready Table 1

Daniel Brinton, PhD
Marshall Chew, MSHI
Medical University of South Carolina

ABSTRACT

The most common first deliverable of a statistical analysis project is colloquially known as the Table 1—which compares the baseline demographics and characteristics of two or more groups of subjects. Completing a Table 1 by hand can be time consuming and introduces opportunities for data entry error in an otherwise flawless analysis. Engineering out human error is a common goal of programmers, but auto-generating a Table 1 using portable code has proven to be a complex task. Typically, prior endeavors to create a Table 1 macro required lots of modification by the end user to work correctly. This Table 1 macro fills in the gaps by producing a publication-ready table 1 on any SAS® dataset, output as a Microsoft Excel spreadsheet, along with a SAS table of the contents therein.

Reported variables may be categorical, continuous, or binary. The macro reports the total sample size in each group. Categorical variables are reported by frequency and percent. Continuous variables are reported as mean ±SD, as well as median [IQR] when not normal. Moreover, testing for normality is accomplished, with results and normality test utilized reported in a comments column. Finally, comparison of differences between groups are statistically tested using appropriate statistical tests—with p-values reported, and the name of the statistical test used reported in the comments column.

INTRODUCTION

The Table 1 macro call function requests seven variable inputs at the onset of the macro execution. For the macro to function correctly the user must supply input for the macro variables data, out, class and at least one entry for the macro variables binary, categorical or continuous. Providing a file pathway for the macro variable spreadsheet ensures a Microsoft Excel spreadsheet will be auto generated and ready for publication excluding journal specific formatting.

THE MACRO VARIABLES

Below is the Macro call function for %TABLE1: The dataset is a modified version of sashelp.bweight to generate binary variables that were not present in the dataset. All listed variables are delimited by a space and punctuated by a comma. VARS_CAT, VARS_BIN and VARS_CONT refer to categorical, binary and continuous variables respectively:

```
%Table1(data=sashelp.bweight,
      out=Table1_bweight,
      VARS_BIN=Married,
      VARS_CAT=MomEdLevel,
      VARS_CONT=CigsPerDay MomWtGain Weight,
      CLASS=MomSmoke,
      SPREADSHEET="C:\Users\User\Desktop\Table1.xlsx");
```

VARS_BIN is for binary variables. These variables must utilize “1” either as a character or numeric category in the affirmative to function correctly. For example, the variable “Married” refers to 1=*Married* and 0=*Unmarried*. The affirmative 1 value and missing values are reported, and all others are discarded. Any binary variables that do not utilize zero-reference coding should be included in *VARS_CAT* instead.

VARS_CAT is for categorical variables. All categories and missing values are reported in the output table.

VARS_CONT is for continuous variables.

CLASS is the classification variable used to denote the comparison of two or more unpaired groups. If the user does not have any of the three types of variables, then the variable may be left blank and punctuated by a comma as seen below for VARS_CAT:

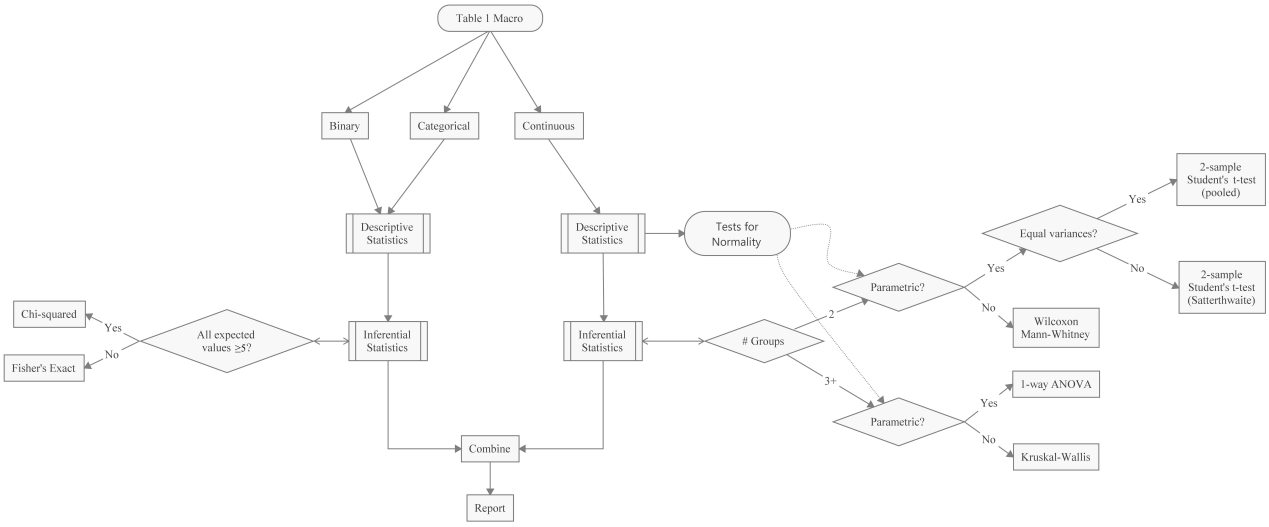
```
%Table1(data=sashelp.bweight,
      out=Table1_bweight,
      VARS_BIN=Married,
      VARS_CAT=,
      VARS_CONT=CigsPerDay MomWtGain Weight,
      CLASS=MomSmoke,
      SPREADSHEET="C:\Users\User\Desktop\Table1.xlsx");
```

VARS_CONT consist of continuous variables which are evaluated for normality and number of comparing groups. Once these criteria are determined, p-values for the corresponding Two sample Student's t-test, Wilcoxon-Mann Whitney test, One-Way ANOVA test or Kruskal-Wallis test are reported in the final table. An example of the final excel file output can be seen in Table 1.

THE MACRO PROCESS

Figure 1 shows how the macro proceeds in its operation. For the three types of variables that can be specified—binary, categorical, and continuous—descriptive statistics are first generated, followed by inferential statistics, then these are combined into one table, then this table is exported to a Microsoft Excel spreadsheet.

Figure 1. *Table 1 macro p*rocess diagram



DESCRIPTIVE STATISTICS

Binary variables report both the number and column percentage (e.g. *180 (18.3))* for the non-reference category (i.e. where *Male=1)*.

Categorical variables report both the number and column percentage for each of the categories represented within the categorical variable (e.g. Cardinal_Direction: North, South, East, West), as well as an additional category for missing, if necessary. This is helpful for evaluating the absence or presence of missing data, as well as its potential biasing implications on analyses. In the event one category is not represented within the data for one or more class groups, an emdash (—) is placed within the cell to show no values are present.

Continuous variables are reported as mean ± standard deviation (when the variable is normally distributed, or both mean ± standard deviation and median/IQR (when the variable is not normally distributed). Tests for normality are conducted and reported in the comments field. When there are less than or equal to 2000 observations, the Shapiro-Wilk test is conducted, otherwise the Kolmogorov-Smirnov test is conducted.

INFERENTIAL STATISTICS

Differences between CLASS groups are tested for both binary and categorical variables using the chi-squared test. In cases where one cell's expected value is less than five, a Fisher's exact test would be more appropriate. However, one limitation of this macro is that this condition is not tested, and is therefore incumbent upon the analyst to look for this warning in the log then conduct the appropriate test is required.

Differences between CLASS groups for continuous variables proceed as shown in Figure 1. In the case of two groups and parametric (normally-distributed data), a two sample Student's t-test is conducted, with either the pooled or Satterthwaite results reported depending on whether or not the variances are equal. In the case of two groups and non-parametric data, the Wilcoxon Mann-Whitney test is performed. In the case of three or more groups and parametric data, a one-way ANOVA is performed. Finally, in the case of three or more groups and non-parametric data, Kruskal-Wallis test is performed.

EXAMPLE: CALL OF MACRO

```
data temp_Cars;
      set sashelp.cars;
      Norm = rand('NORMAL',0,1);
      Drive4 = (DriveTrain='All');
      label       Norm = 'A random standard-normal number'
                  Drive4 = 'All-wheel drive?';
run;
```

Here a temporary dataset is created using the Cars dataset included with SAS. Since this dataset does not have a normally distributed continuous variable, one is created (*Norm*). Likewise, as a binary variable is not included in this datase, one is created by assigning all-wheel-drive cars to the *yes* condition (Drive4=1), and all front-wheel-drive and rearwheel drive cars to the *no* condition of this binary variable (Drive4=0).

The macro code may be called as follows:

```
%include "C:\Location\MACRO - Table 1.sas";
```

After the creation of this temporary dataset, this macro is called as follows:

```
%Table1(data=temp_Cars,
      out=Table1_cars,
      VARS_BIN=Drive4,
      VARS_CAT=Cylinders Type,
      VARS_CONT=EngineSize Horsepower MPG_City MPG_Highway MSRP Norm,
      CLASS=Origin,
      SPREADSHEET="C:\Users\User\Desktop\Table1 - Cars.xlsx");
```

The above macro call will utilize *temp_Cars* to create a table 1, outputting this both as a SAS table (*Table1_cars*) as well as a Microsoft Excel spreadsheet at the above-mentioned location. Comparisons between groups will be made using the variable *Origin* (i.e. Asia, Europe, USA). The example of the output of these results follows.

EXAMPLE: RESULTS

Table 1. Demographics and characteristics (Source: temp_Cars)

Characteristic		Asia	Europe	USA	p-value	Comments
n		158	123	147		The number of observations in each group (Origin)
A random standard-normal number		-0.0 ± 1.0	-0.1 ± 1.0	-0.2 ± 1.1	0.2699	Shapiro-Wilk, W=0.995, p=0.169 (Normal=1) One-Way ANOVA test
All-wheel drive?		34 (21.5)	36 (29.3)	22 (15.0)	0.0173	Chi-Squared test
Cylinders					<0.0001	Chi-Squared test
	10	—	—	2 (1.4)		
	12	—	3 (2.4)	—		
	3	1 (0.6)	—	—		
	4	74 (46.8)	25 (20.3)	37 (25.2)		
	5	—	7 (5.7)	—		
	6	69 (43.7)	54 (43.9)	67 (45.6)		
	8	12 (7.6)	34 (27.6)	41 (27.9)		
	Missing	2 (1.3)	—	—		
Engine Size (L)		2.8 ± 0.9	3.2 ± 1.0	3.6 ± 1.2		Shapiro-Wilk, W=0.959, p<0.001 (Normal=0)
		2.6	3.0	3.6	<0.0001	Kruskal-Wallis test
Horsepower		190.7 ± 59.4	251.9 ± 80.7	212.8 ± 63.7		Shapiro-Wilk, W=0.950, p<0.001 (Normal=0)
		187.5 [142.0 - 235.0]	225.0 [194.0 - 302.0]	200.0 [155.0 - 250.0]	<0.0001	Kruskal-Wallis test
MPG (City)		22.0 ± 6.7	18.7 ± 3.3	19.1 ± 4.0		Shapiro-Wilk, W=0.808, p<0.001 (Normal=0)
		20.5	19.0	18.0	<0.0001	Kruskal-Wallis test
MPG (Highway)		28.3 ± 6.8	26.0 ± 4.2	26.0 ± 5.4		Shapiro-Wilk, W=0.930, p<0.001 (Normal=0)
		27.0	26.0	26.0	0.0083	Kruskal-Wallis test
MSRP		24741.3 ± 11321.1	48349.8 ± 25318.6	28377.4 ± 11712.0		Shapiro-Wilk, W=0.778, p<0.001 (Normal=0)
		23032.5 [17200.0 - 28800.0]	40590.0 [33780.0 - 56595.0]	25520.0 [20310.0 - 33995.0]	<0.0001	Kruskal-Wallis test
Type					0.0001	Chi-Squared test
	Hybrid	3 (1.9)	—	—		
	SUV	25 (15.8)	10 (8.1)	25 (17.0)		
	Sedan	94 (59.5)	78 (63.4)	90 (61.2)		
	Sports	17 (10.8)	23 (18.7)	9 (6.1)		
	Truck	8 (5.1)	—	16 (10.9)		
	Wagon	11 (7.0)	12 (9.8)	7 (4.8)		

All values expressed as n(%), mean ±s.d., or median[Q1 - Q3]

LIMITATIONS

CLASS is our classifying variable used to denote the comparison of two or more unpaired groups. Paired group analysis is not a feature supported by this macro for statistical testing.

The table 1 macro will report the frequency values and Chi-squared test result, even if a Fisher's Exact test would be more appropriate (in cases where expected cell counts are < 5). In this instance, a warning is displayed (c.f. *Figure 2*). In this case, the user may choose to manually run a Fisher's Exact test and change the corresponding p-value.

Figure 2. Chi-Square Warning

```
WARNING: 57% of the cells have expected counts less than 5 for the table
      of Cylinders by Origin. Chi-Square may not be a valid test.
```

CODE AVAILABILITY

Full source code of this macro is available at the primary author's website, listed below in the contact information section.

CONCLUSION

This macro creates one of the first deliverables of a project—the Table 1, which typically is a comparison of the demographics and characteristics between two or more groups. This table is also handy for examining comparisons between groups where key information may be missing to ascertain the missing data mechanism to inform statistical analyses.

ACKNOWLEDGMENTS

This project was supported by the Health Resources and Services Administration (HRSA) of the U.S.Department of Health and Human Services (HHS) as part of the National Telehealth Center of Excellence Award (U66 RH31458-01-00). The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement, by HRSA, HHS or the U.S. Government.

CONTACT INFORMATION

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

Daniel Brinton, PhD
Assistant Professor
Medical University of South Carolina
brintond@musc.edu
http://people.musc.edu/~dlb24/
https://www.linkedin.com/in/danielbrinton/