SESUG Paper 216-2019

A Table 1 Macro that Produces Publication-Ready Results: %Table1nDone

Martha Wetzel, Emory University

ABSTRACT

Academic papers in many fields include a table summarizing the demographic characteristics of the sample and/or treatment groups—the ubiquitous Table 1. These tables often require an overall summary, as well as a between-group comparison. Furthermore, these tables summarize mixed data types, including both continuous and categorical data, requiring different statistical tests and SAS® procedures. Without automation, analysts can spend hours per project calculating and arranging the results into the correct format, often having to redo the entire table when an investigator realizes there is a data error. The %Table1nDone macro was created to reduce analyst time spent on Table 1s.

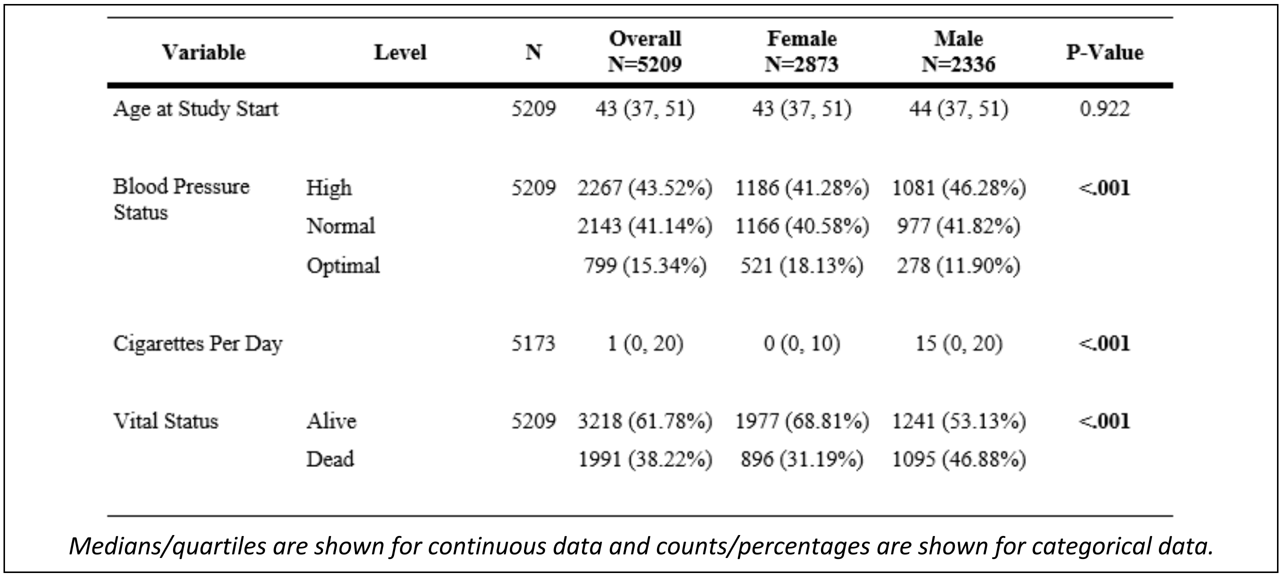
This paper presents a new Table 1 macro that calculates summary statics overall and by group, performs the corresponding statistical testing as required, and produces an RTF file containing the final summary (i.e., the Table 1). The %Table1nDone macro expands on existing Table 1 macros by 1) streamlining variable input via an Excel file, 2) creating a table with both overall and by-group summary data, 3) producing an RTF table in the format expected by many journals, and 4) saving permanent data sets of key information for analyst review.

Logic is built into the macro to select the appropriate statistical test based on user-supplied factors such as variable type (e.g., categorical, continuous) and data factors (e.g., number of comparison groups, cell-size counts, distributional assumptions). Publication-ready output contains results formatted as “N (%)”, “mean (standard deviation)”, or “median (quartile 1, quartile 3)”, depending on the type of data. In addition, the macro produces a report for the analyst to review for unexpected values in the data. This paper provides an overview of the macro’s capabilities, a description of the use and required parameters, an explanation of the statistical tests included, examples of output, and links to the macro code.

INTRODUCTION

Academic papers so often begin with a table containing descriptive statistics about the population studied that such tables are commonly referred to as “Table 1s.” Univariate analyses comparing the sample characteristics between key groups can also be included in the Table 1. The purpose of this macro is to streamline and standardize the production of these tables. To that end, this macro outputs statistics from univariate analysis in a publication-ready RTF table, an example of which is shown in Output 1.

Output 1. Sample Output from %Table1nDone



Variable	Level	N	Overall N=5209	Female N=2873	Male N=2336	P-Value
Age at Study Start		5209	43 (37, 51)	43 (37, 51)	44 (37, 51)	0.922
Blood Pressure Status	High	5209	2267 (43.52%)	1186 (41.28%)	1081 (46.28%)	**<.001**
	Normal		2143 (41.14%)	1166 (40.58%)	977 (41.82%)	
	Optimal		799 (15.34%)	521 (18.13%)	278 (11.90%)	
Cigarettes Per Day		5173	1 (0, 20)	0 (0, 10)	15 (0, 20)	**<.001**
Vital Status	Alive	5209	3218 (61.78%)	1977 (68.81%)	1241 (53.13%)	**<.001**
	Dead		1991 (38.22%)	896 (31.19%)	1095 (46.88%)	

Medians/quartiles are shown for continuous data and counts/percentages are shown for categorical data.

This paper outlines the process of using the macro, provides links to the macro, and documents the statistical methods the macro employs. The intended audience for this paper consists of SAS users who are familiar with the use and interpretation of the statistical tests described in the Statistical Methods section. This macro is not a substitute for statistical proficiency; rather, it is intended speed the workflow of SAS users who spend substantial time creating this type of univariate analysis table.

Three pieces of output are created by %Table1nDone. The macro's main product is the publication-ready RTF table, an example of which is shown below as Output 1. This table can be customized to display an overall statistics column and/or columns showing the statistics by group. Additional options allow the inclusion of columns showing the response counts, the metrics used, and p-values for group comparisons. The second piece of output is a temporary report showing descriptive statistics for the continuous variables, including means, medians, and minimums/maximums. This table is created as a check for the analyst to make sure that the data do not contain, for example, any extreme outliers that had been previously unnoticed. Finally, the macro saves SAS data sets containing detailed information on additional statistics. In particular, for the continuous variables, the saved data set includes the results from the Shapiro-Wilk test for normality and F-tests for equal variance.

All examples in this paper draw from the sashelp.heart data set.

USE

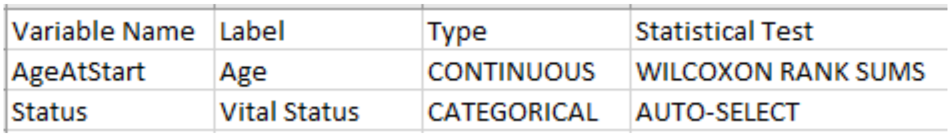
The steps for use of this macro are outlined below.

1. DATA PREPARATION

The input data should be at the individual level (e.g., if summarizing patient demographics, each line should represent one patient). Continuous variables must be numeric type and categorical variables must be character type. Variable labels can either be specified when running the macro or included in the SAS data.

2. VARIABLE LIST

The key to this macro is setting up an Excel driver. A driver template containing drop-down menus can be accessed at https://github.com/mpwetzel/SAS4Academia.The user inputs the list of requested variables into the Excel driver and designates whether each variable is continuous or categorical. Optionally, the user can specify custom variable labels. Furthermore, the user can specify the type of statistical testing performed by the macro. If the statistical testing type is not specified, the macro will use built-in logic to select the most appropriate test, as described in Statistical Methods of this paper. An example of a correctly completed driver is shown below in Figure 1.



Variable Name	Label	Type	Statistical Test
AgeAtStart	Age	CONTINUOUS	WILCOXON RANK SUMS
Status	Vital Status	CATEGORICAL	AUTO-SELECT

Figure 1. Excel Driver

It is recommended that the driver template be used for setting up the driver, in order to avoid issues with misspelling. If the driver template is unavailable, a driver can be created on a spreadsheet tab named "Variables." The columns in the driver should be as follows:

- Variable Name (Required): Add the names of all variables to be analyzed. All variables must be present in the input data set.
- Label (Optional): Enter the label to be used in the output for each variable. Note that this label will

overwrite existing variable labels. If left blank, the variable label in the SAS data set will be used.

- Type (Required): Designate variable as "CONTINUOUS" or "CATEGORICAL."
- Statistical Test (Optional): Select a statistical test. If omitted, the program will auto-select the tests.
    - Valid values for continuous variables are: AUTO-SELECT, TTEST, ANOVA, WILCOXON RANK SUMS, KRUSKAL WALLIS, and KOLMOGOROV-SMIRNOV.
    - Valid values for categorical variables: AUTO-SELECT, CHI SQUARE

3. MACRO CALL

The keyword parameters are defined below.

Required Parameters

- DATASET: Name of the input SAS data set
- DRIVER: File path and name for Excel driver file
- OUTPATH: File pathway where the RTF output should be saved
- FNAME: File name for the RTF output

Optional Parameters

- OVERALL: Y/N include the overall descriptive statistics. Default = Y
- BYCLASS: Y/N perform univariate analysis using the class variable specified in the CLASSVAR option. Default = Y
- CLASSVAR: Name of the grouping variable. This is required if the BYCLASS parameter is set to Y
- OUTLIB: Name of library to save final SAS data sets to. Default = Work
- ROUNDTO: Decimal place to round results to, in the format used by the round function. Default = 0.01
- DISPLAY_PVAL: Y/N display p-values in RTF report. Only valid with BYCLASS = Y. Default = Y
- DISPLAY_METRIC: Y/N display metric names in RTF report. Default = Y
- DISPLAY_N: Y/N display variable non-missing value counts in RTF report. Default = Y
- CLEARTEMP: Y/N delete temporary data sets associated with macro run. Default = Y

Example Call

The following code can be used to run the macro on the sashelp.heart data set:

```
%TABLE1NDONE(
      DATASET = sashelp.heart, /* Input data set */
      DRIVER = C:\Heart Analysis\Descriptives Driver_Heart.xlsx,
      OUTPATH= C:\Heart Analysis, /* File pathway for RTF output */
      FNAME= Table 1 Heart, /* File name for RTF output */
      OVERALL = Y, /* Include overall summary statistics */
      BYCLASS = Y, /* Include summary statistics by group */
      CLASSVAR = Sex, /* Class variable */
      DISPLAY_PVAL = Y, /* Display p-values in RTF report */
      DISPLAY_METRIC = Y, /* Display metric names in RTF report */
      DISPLAY_N = Y /* Display non-missing counts in RTF report */
      );
```

4. LOG AND OUTPUT REVIEW

Reviewing the log after program execution is essential. Custom warnings and error messages related to the Excel driver and statistical testing will print to the log.

Three types of output are created by this macro:

1. Publication-ready RTF output: RTF output is output to a specified destination. The style of the RTF output created by %Table1nDone is based on previously published tables styles (Molter 2007, Liu 2019).
2. Printed output to the results window: For continuous variables, tables are printed for user review to facilitate review of data.
3. Saved data sets: Up to four saved data sets are created. For continuous variables, the tables include full output from the UNIVARIATE procedure and the results from the F-test for equality of variance, in addition to the name of the statistical test used in the printed in the output.

STATISTICAL METHODS

This statistical testing logic of the macro is described below.

CATEGORICAL VARIABLES

For categorical variables, Chi square tests will be run. If the Statistical Test column in the Excel driver is set to AUTO-SELECT and a small-cell warning occurs, a Fisher's exact test will be run and used in the output. If the Statistical Test option is set to CHI SQUARE and a small-cell warning occurs, the report will contain the Chi square test and a warning will be printed to the log.

CONTINUOUS VARIABLES

This macro is capable of running the following tests on continuous variables: T-tests, ANOVA, Wilcoxon rank sums, Kruskal-Wallis, and Kolmogorov–Smirnov tests.

The input driver allows the user to specify a test or have the macro automatically select a test. The macro logic identifies a most appropriate test based on normality (per Shapiro-Wilk test), the number of groups, and equality of variance (based on a folded F-test or Levene's test), as shown in Table 1. If a user-selected test is not the same as the macro-identified "most appropriate" test, the results of the user-selected test will be printed in the RTF table and a warning will print to the log.

If the user selects an invalid option (e.g., T-test when there are three groups), the program will default to the auto-selected option and a warning will print to the log.

Table 1: Statistical Tests

Distribution	Number Groups	Variance	Test
Normal	2	Equal	Pooled T-test
Normal	≥3	Equal	ANOVA
Normal	2	Unequal	Satterthwaite T-test
Normal	≥3	Unequal	Out of macro scope
Non-Normal	2	Equal	Wilcoxon rank sums
Non-Normal	≥3	Equal	Kruskal-Wallis
Non-Normal	2	Unequal	Kolmogorov–Smirnov test
Non-Normal	≥3	Unequal	Out of macro scope

Some data may have a distribution not suitable for any of the statistical tests performed by %Table1nDone. When the AUTO-SELECT option is chosen, the macro will not print a p-value in the RTF output if normality and variance testing indicate the need for a statistical test outside of the macro's capabilities. Instead, a note will be printed to the log recommending a different form of analysis.

CODE

The full program for %Table1nDone and a template for the Excel driver can be found on the author's GitHub page, https://github.com/mpwetzel/SAS4Academia.

CONCLUSION

Many SAS programmers spend hours for each project creating formatted tables to display results from univariate analyses. The %Table1nDone macro is intended to streamline production of univariate analysis tables without turning the analysis into a black box that removes the analyst from the process. To this end, the macro creates a summary report and detailed SAS data sets for the user, in addition to a publication-ready RTF table. This paper prepares SAS users to implement the macro by elucidating the statistical methods and the process for using the macro.

REFERENCES

Molter, Michael J. 2007. "Tips and Tricks for Creating the Reports Your Clients Need to See." Proceedings of Northeast SAS User Group Conference, Baltimore, MD. Available at: https://www.lexjansen.com/nesug/nesug07/bb/bb12.pdf.

Liu Y, Nickleach DC, Zhang C et al. Carrying out streamlined routine data analyses with reports for observational studies: introduction to a series of generic SAS® macros [version 2; peer review: 2 approved]. F1000Research 2019, 7:1955.

ACKNOWLEDGMENTS

Many thanks for the support, feedback, and user acceptance testing performed by the Pediatric Biostatistics Core at Emory University, comprised of Courtney McCracken, Janet Figueroa, Michael Kelleman, Scott Gillespie, Erika Bettermann, Amanda Thomas, and Ishaan Dave. The author would also like to thank Ray Berens for code feedback.

CONTACT INFORMATION

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

Martha Wetzel
Emory University
Martha.wetzel@emory.edu
https://github.com/mpwetzel/SAS4Academia