

SESUG Paper 003-2021

Statistical Test Selector (STS) For Researchers

Dr. Kannan Deivasigamani, Centene - TSS
Dr. Douglas Lunsford, Tutor With The Best, LLC

Abstract

This SESUG paper demonstrates how a SAS® macro can be used to programmatically select the suitable statistical test for a given scenario. Gitanjali, Manikandan, and Raveendran (2014) have presented in their book written for post graduates to help with research methodology in a medical setup to help choosing the apt statistical test for their research (pg. 90-91). While not all statistical tests are included to make it comprehensive, the book contains required information on most of the commonly used statistical tests and the criteria used in selecting those as appropriate. The book by Gitanjali, Manikandan, and Raveendran (2014) is the seed to this technical SAS paper that is anticipated to come-in as an easy and quick reference macro mostly by budding researchers, students and others involved in statistical testing using research data via SAS. As SAS is a commonly used statistical software, it made more sense to the author to develop a macro to accept different parameters required in deciding the appropriate research for a specific design and setup.

The SAS macro, as it is designed to be used as a Statistical-Test-Selector, will hence-forth be referenced by the acronym STS. As one can imagine, this macro can be used as part of general-purpose non-domain specific research. The process could be utilized in the beginning of most research studies where a statistical test selection needs to be made. This decision is suggested by this simple SAS Macro call. This would enable SAS statisticians proceed forward without re-inventing the wheel for the commonly used statistical tests to be able to select one from the many available. As this macro is not exhaustive, there may be some situations where other criteria need to be used that are not listed here resulting in statistical tests that are not accounted for in this macro. Some uncommon or infrequent techniques including but not limited to Survival analysis, Monte-Carlo and others are not in the scope of this tool in its current version.

INTRODUCTION

Statistics while interesting, could be daunting sometimes given its complexity especially if one is a beginner or a student or a budding statistician or even an experienced one if one has to go through the refined combing to identify the appropriate test for the current study. If one is interested in saving time by running a quick one-line macro this tool may come in handy. If that first step can be done by a simple one-line call to a macro with a handful of parameters, statisticians and SAS statistical programmers would find the easy and time savings from this macro. That is exactly the intention of this attempt in devising this macro.

Shields and Twycross (2021) have detailed several tests and their application of when they are used which they have cited from the work by Hicks (1996), Polgar, and Thomas (2000). There are several criteria that can drive the type of test chosen in a specific instance. Gibson (2014) stated that selecting the right statistical method in a research study is necessarily a part of the planning process in the research.

STATISTICAL TESTS

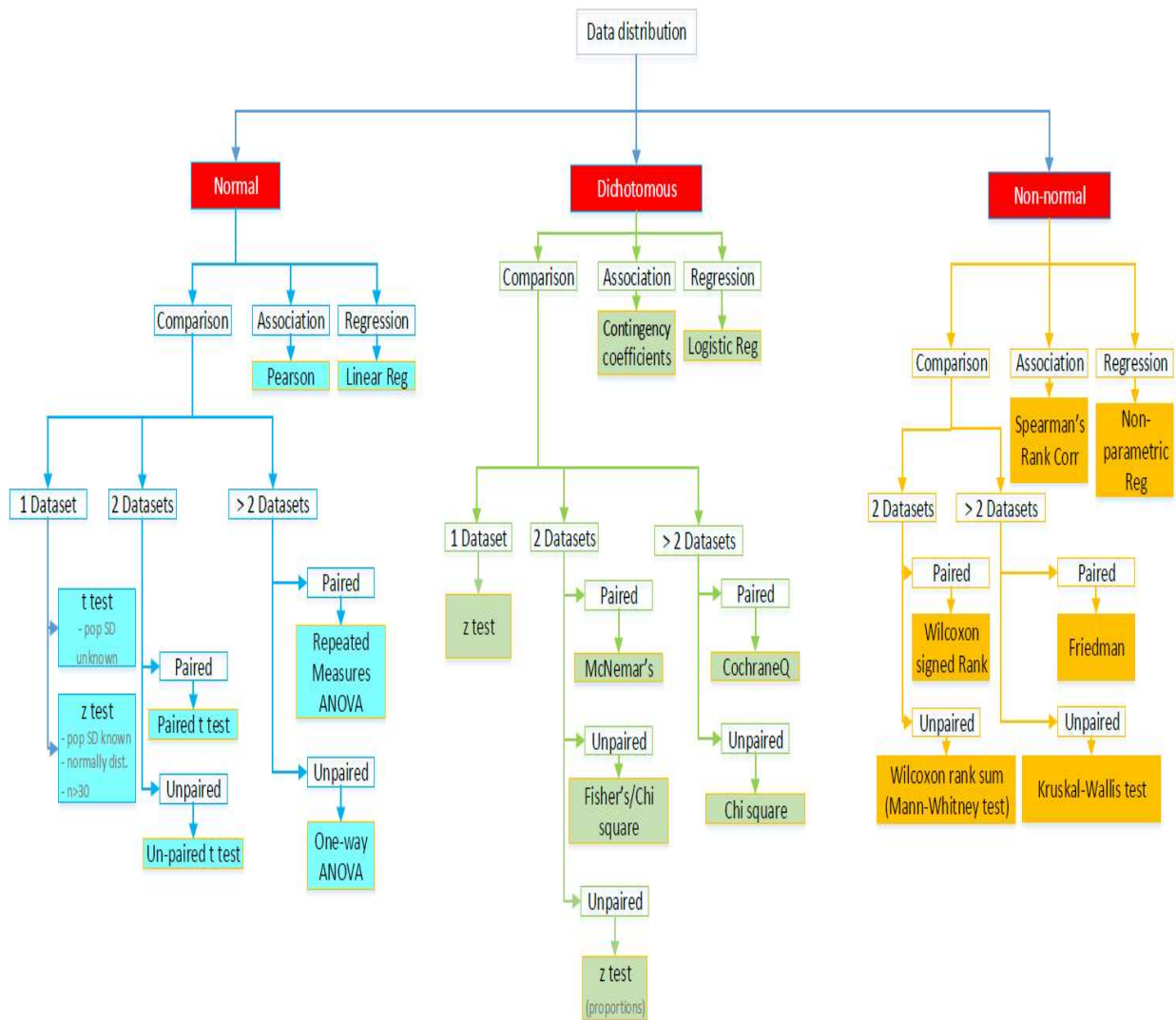
Flores-Ruiz, Miranda-Novales, and Villasís-Keever (2017) have stated that research design, number of measurements, and the scale of measurement of the variables determine the type of statistical test a research will require. Murray and Knowles (2014) presented a variation of the statistical test selection process using 4 aspects stated below. They further built a visual decision tree based on the possible deviations based on the above 4 aspects of the research conducted.

- (i) Number of variables measured
- (ii) The scale of measurement
- (iii) Difference or relationship and
- (iv) Independent or paired.

The task of selecting the right statistical test can be a daunting experience for a novice researcher and this SAS macro is expected to eliminate the guess work of the researcher for the most part. The macro list is not exhaustive but includes most of the commonly used statistical tests into consideration that are both Parametric and non-Parametric. The decision tree currently identifies t-test, z-test, paired t-test, unpaired t-test, One-way ANOVA, repeated measures ANOVA, Pearson's correlation, Spearman's Correlation, Linear Regression, Non-parametric Regression, Wilcoxon signed rank, Mann-Whitney test, Friedman test, Kruskal-Wallis test, Contingency coefficients, Logistic regression, McNemar's test, Fisher's test, CochranQ test, and Chi Squared test. While there are many other tests that are less common and less frequently used, it might be a better enhancement for the future and are not within the current scope of this paper. Addition of more statistical tests to STS might also increase the number of input parameters required to make the decision tree work in order to account for the additional statistical tests. In order to help a reader, grasp the hierarchical decision tree structure easily, a pictorial representation of the various statistical tests and the criteria that lead to those is shown in the figure below. Suner and Dycle (2014) stated the issues due to lack of a clear method to choose the statistical tests and its impact. They went ahead portraying a visual flowchart limiting to Anova and Friedman's test as common ones to their domain of interest (biological and medical studies).

"Lack of an algorithm presenting the most common statistical tests used in biomedical research in a single flowchart causes several problems such as shifting users among the algorithms, poor decision support in test selection and lack of satisfaction of potential users "(Suner and Dycle, 2014)

A PICTORIAL REPRESENTATION OF THE PROCESS



MACRO OPERATION

STA has been designed to be called with a simple invocation of a one-line macro call within a SAS program and then would consider the values provided by the user to arrive at the decision that would be printed in the output log. There is a series of branching logic that the execution flows through and depending on the values and decision tree, the waterfall methodology moves along the hierarchical structure till it gets to the final test. The different macro variables expect values that are mostly “Y” or “N” except for the last 2 macro inputs parameters that expect numerical values. Such as 2 or 3 or other whole number values. Such a call will be easy for the user to provide and makes the STS macro easy to use.

A typical call would look like the following:

```
Options mlogic mprint symbolgen errors=1 obs=max;

%sts(normal=Y,
      non_normal=Y,
      dichotomous=Y,
      Comparison=Y,
      association_of_2vars=Y,
      regression_of_2vars=Y,
      paired=Y,
      number_of_datasets=2,
      number_of_variables=2);
```

As a typical SAS programmer would have realized that the macro uses non-positional parameters and therefore also serves increasing clarity. The related options above the invocation of the macro aids in viewing the logic of the macro and with error identification information that would become part of the SAS log upon execution.

SOURCE CODE

```
1  /* macro "sts" stands for "Statistical Test Selector" */
2  %macro sts(normal=,
3      non_normal=,
4      dichotomous=,
5      comparison=,
6      association_of_2vars=,
7      regression_of_2vars=,
8      paired=,
9      number_of_datasets=,
10     number_of_variables=);
11
12     %if &normal=Y
13     %then
14         %do;
15             %if &comparison=Y
16             %then
17                 %do;
18                     %if number_of_datasets=2
19                     %then
20                         %do;
21                             %if &paired=Y
22                             %then
23                                 %let selected_test='Paired t test';
24                             %else
25                                 %if &paired=N
26                                 %then
27                                     %let selected_test='Unpaired t test';
28                             %end;
29                         %else
30                         %do;
31                             %if &paired=Y
32                             %then
33                                 %let selected_test='Repeated Measures ANOVA';
34                             %else
35                                 %if &paired=N
36                                 %then
37                                     %let selected_test='One Way ANOVA';
38                             %end;
39                         %end;
40                     %else
41                     %if &association_of_2vars=Y
42                     %then
43                         %let selected_test='Pearson correlation';
44                     %else
45                     %if &regression_of_2vars=Y
46                     %then
47                         %let selected_test='Linear regression';
48                     %end;
49                 %else
50                 %if non_normal=Y
51                 %then
52                     %do;
53                         %if &comparison=Y
54                         %then
55                             %do;
```

```

56:         %if number_of_datasets=2
57:         %then
58:             %do;
59:                 %if %paired=Y
60:                 %then
61:                     %let selected_test='Wilcoxon signed rank test';
62:                 %else
63:                 %if %paired=N
64:                 %then
65:                     %let selected_test='Wilcoxon rank sum / Mann-Whitney';
66:                 %end;
67:             %else
68:             %do;
69:                 %if %paired=Y
70:                 %then
71:                     %let selected_test='Friedman test';
72:                 %else
73:                 %if %paired=N
74:                 %then
75:                     %let selected_test='Kruskal-Wallis test';
76:                 %end;
77:             %end;
78:         %else
79:         %if %association_of_2vars=Y
80:         %then
81:             %let selected_test="Spearman's Rank Correlation";
82:         %else
83:         %if %regression_of_2vars=Y
84:         %then
85:             %let selected_test='Nonparametric regression';
86:         %end;
87:     %else
88:     %if dichotomous=Y
89:     %then
90:         %do;
91:             %if %comparison=Y
92:             %then
93:                 %do;
94:                     %if number_of_datasets=2
95:                     %then
96:                         %do;
97:                             %if %paired=Y
98:                             %then
99:                                 %let selected_test="McNemar's test";
100:                             %else
101:                             %if %paired=N
102:                             %then
103:                                 %let selected_test="Fisher's test (or) Chi square test";
104:                             %end;
105:                         %else
106:                         %do;
107:                             %if %paired=Y
108:                             %then
109:                                 %let selected_test="Cochrane Q";
110:                             %else

```

```

111         %if %paired=N
112         %then
113             %let selected_test='Chi square test';
114         %end;
115     %end;
116     %else
117     %if %association_of_2vars=Y
118     %then
119         %let selected_test="Contingency coefficients";
120     %else
121     %if %regression_of_2vars=Y
122     %then
123         %let selected_test='Logistic regression';
124     %end;
125     %PUT '*****';
126     %PUT " SELECTED TEST=%SELECTED_TEST. ";
127     %PUT '*****';
128 %mend sts;
129
130 %sts( normal=Y,
131       non_normal=Y,
132       dichotomous=Y,
133       comparison=Y,
134       association_of_2vars=Y,
135       regression_of_2vars=Y,
136       paired=Y,
137       number_of_datasets=2,
138       number_of_variables=2);

```

The raw source code shown above can be placed within a SAS program as part of in-line code or even be included in an auto-call library in one's organizational setting in which case, all users can call the macro with just one-line call without having to have a copy of the macro source code as part of their SAS program. It is up to the users' personal preference and also a preference to rename if required.

Whether the STS macro code is in-line or within an auto-call library as part of the SAS environment, the actual invocation is just one line and the execution of the call will produce an output SAS Log with the selected statistical test name written as part of the message within the log as shown below.

PRACTICAL APPLICATION (Log)

The macro expects different parameters as inputs

```
159 150      %sts( normal=Y,
160 MLOGIC(STS): Beginning execution.
161 151          non_normal=Y,
162 152          dichotomous=Y,
163 153          comparison=Y,
164 154          association_of_2vars=Y,
165 155          regression_of_2vars=Y,
166 156          paired=Y,
167 157          number_of_datasets=2,
168 158          number_of_variables=2);
169 MLOGIC(STS): Parameter NORMAL has value Y
170 MLOGIC(STS): Parameter NON_NORMAL has value Y
171 MLOGIC(STS): Parameter DICHOTOMOUS has value Y
172 MLOGIC(STS): Parameter COMPARISON has value Y
173 MLOGIC(STS): Parameter ASSOCIATION_OF_2VARS has value Y
174 MLOGIC(STS): Parameter REGRESSION_OF_2VARS has value Y
175 MLOGIC(STS): Parameter PAIRED has value Y
176 MLOGIC(STS): Parameter NUMBER_OF_DATASETS has value 2
177 MLOGIC(STS): Parameter NUMBER_OF_VARIABLES has value 2
178 SYMBOLGEN: Macro variable NORMAL resolves to Y
179 MLOGIC(STS): %IF condition &normal=Y is TRUE
180 SYMBOLGEN: Macro variable COMPARISON resolves to Y
181 4 The SAS System
182
183 MLOGIC(STS): %IF condition &comparison=Y is TRUE
184 MLOGIC(STS): %IF condition number_of_datasets=2 is FALSE
185 SYMBOLGEN: Macro variable PAIRED resolves to Y
186 MLOGIC(STS): %IF condition &paired=Y is TRUE
187 MLOGIC(STS): %LET (variable name is SELECTED_TEST)
188 MLOGIC(STS): %PUT '*****'
189 '*****'
190 MLOGIC(STS): %PUT " SELECTED TEST=&SELECTED_TEST. "
191 SYMBOLGEN: Macro variable SELECTED TEST resolves to 'Repeated Measures ANOVA'
192 " SELECTED TEST='Repeated Measures ANOVA' "
193 MLOGIC(STS): %PUT '*****'
194 '*****'
195 MLOGIC(STS): Ending execution.
196 159
```

LIMITATION

The macro, in its current design accepts various input parameters that are essential in determining the statistical test that would be appropriate based on various factors in the design. However, the macro does not currently perform the right test and provide the critical values or inferences as outcome of the statistical procedure that can be readily used by the Statistician or the data scientist to translate it into the form of a business decision or operational action item. The other possible enhancement that could be helpful to others is if the list is made close to be exhaustive to include other statistical tests along with their decision trees incorporated into STS.

CONCLUSION

The macro can be further enhanced for performing the limitation described above. The next progressive step for the development of this macro would be to enable implementation of the right test automatically that can be readily used by the data scientist. This can save time and help a quick turn-around of data analysis impacting a business in a timely fashion.

REFERENCES

- Flores-Ruiz, E., Miranda-Novales, M. G., & Villasís-Keever, M. Á. (2017). [The research protocol VI: How to choose the appropriate statistical test. Inferential statistics]. *Revista Alergia Mexico (Tecamachalco, Puebla, Mexico : 1993)*, 64(3), 364–370. <https://doi.org/10.29262/ram.v64i3.304>
- Gibson, D. (2014-10-16). Planning, choosing, and using statistics. In *Methods in Comparative Plant Population Ecology*, Second Edition. : Oxford University Press. Retrieved 12 Aug. 2021, from <https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199671465.001.0001/acprof-9780199671465-chapter-7>.
- Gitanjali, Manikandan, Raveendran (2014). *A Practical Approach to PG Dissertation*. Second Edition. Pharma Med Press, Hyderabad, India.
- SUNER, A., KARAKÜLAH, G., & DÝCLE, O. (2014). Towards a Web-Based Decision Support Tool for Selecting Appropriate Statistical Test in Medical and Biological Sciences. *Studies in Health Technology & Informatics*, 205, 48–52.
- Twycross, Alison, Shields, Linda. "Statistics made simple: Part 4 choosing the right statistical test: this is one of a series of short papers on aspects of research by Alison Twycross and Linda Shields." *Paediatric Nursing*, vol. 16, no. 8, 2004, p. 24. Gale OneFile: Health and Medicine, link.gale.com/apps/doc/A123331705/HRCA?u=uphoenix&sid=bookmark-HRCA&xid=b9dbc972. Accessed 12 Aug. 2021.

RECOMMENDED READING

- Base SAS® Procedures Guide
- SAS® Macro Language: Reference Guide®
- SAS® Macro Programming

CONTACT INFORMATION

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

Dr G. Douglas Lunsford
Tutor With The Best, LLC
gLunsfor@gmail.com
<https://www.linkedin.com/in/g-douglas-lunsford-4a72609/>

Dr. Kannan Deivasigamani
Centene Corporation
TextDrK@gmail.com
<https://www.linkedin.com/in/kannandeivasigamani>