Using SAS for Nonparametric Statistics

Gerry Hobbs,
Departments of Statistics and of Community Medicine
West Virginia University

Definition

A nonparametric procedure is a statistical procedure that has certain desirable properties that hold under relatively mild assumptions regarding the underlying population(s) from which the data are obtained – Hollander and Wolfe. The definition is not only vague but it is also extremely inclusive. “Statistical procedure” can refer to a test or an estimator. There are a number of new methodologies that can be seen as nonparametric. This discussion will be limited to traditional rank based analyses. In this presentation we will exploit the idea that if the data are replaced by ranks then we can often apply the usual computations to the ranks and arrive at a test that is equivalent to an established nonparametric procedure.

First we will review what SAS does directly. Suppose we have a standard ANOVA-like situation with a set of treatments named GROUP and response named VALUE. The “PROC” below is a bare-bones invocation of the relevant procedure and some selected sample output follows.

PROC NPAR1WAY;
CLASS GROUP;
VAR VALUE;

Analysis of Variance for Variable value
Classified by Variable group

<table>
<thead>
<tr>
<th>group</th>
<th>N</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>3</td>
<td>24.0</td>
</tr>
<tr>
<td>y</td>
<td>7</td>
<td>18.0</td>
</tr>
</tbody>
</table>

Source               DF  Sum of Squares  Mean Square  F Value    Pr > F
Among                1    75.60   75.60     13.7455    0.0060
Within               8    44.00   5.50

Wilcoxon Scores (Rank Sums) for Variable value
Classified by Variable group

<table>
<thead>
<tr>
<th>group</th>
<th>N</th>
<th>Sum of Scores</th>
<th>Expected Under H0</th>
<th>Std Dev</th>
<th>Mean Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>3</td>
<td>27.0</td>
<td>16.50</td>
<td>4.374167</td>
<td>9.0</td>
</tr>
<tr>
<td>y</td>
<td>7</td>
<td>28.0</td>
<td>38.50</td>
<td>4.374167</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Kruskal-Wallis Test

Chi-Square      5.7622
DF             1
Pr > Chi-Square 0.0164
The first part of the output is from a standard 1-way ANOVA. The second part corresponds to what is
known as a Kruskal-Wallis (KW) test. In a different guise it is also called the Wilcoxon Rank Sum Test
(WRST). The p-value for the Chi-square approximation is quite a bit larger than the p-value derived from
ANOVA, 0.006. That certainly doesn’t happen all the time but neither is it unusual, at least in those
situations where the data are reasonably well behaved.

Because the KW test is rank based, the actual distributions of the test statistics are discrete while the
approximations are all based on continuous probability models, usually normal or Chi-Square. SAS
provides an exact solution. Simply add EXACT WILCOXON; to the PROC to get additional output as
follows.

```
Exact Test
One-Sided Pr >= S           0.0083
Two-Sided Pr >= |S - Mean|   0.0167
```

PROC NPARIWAY essentially replaces the data by ranks and then evaluates them in a general purpose
linear rank statistic (see Hajek & Sidak for details). We can transform the ranks to produce other
nonparametric linear rank tests. SAS has facilities to do several such tests but we will illustrate just one.
The Van der Waerden (VW) test is a normal scores test in that we replace the data with functions of the
ranks that are related to the normal distribution function. Placing the VW option in the PROC statement
produces the following.

```
Van der Waerden Scores (Normal) for Variable value
     Classified by Variable group

          group   N   Sum of Scores  Expected Under H0  Std Dev Under H0    Mean Score
           x     3    2.848221       0.0     1.201694       0.949407
           y     7   -2.848221      0.0     1.201694     -0.406889

Van der Waerden One-Way Analysis

          Chi-Square  5.6177
          DF            1
          Pr > Chi-Square  0.0178
```

The EXACT statement provides a mechanism by which exact p-values may be gotten in cases where large
sample approximation (LSA) p-values may be suspect. For larger sized samples even the very efficient
algorithm used by SAS may take a long time to get exact results. Alternatively, we can use the MC option
on the EXACT statement to do a simulation in order to estimate the p-value. By default SAS uses 10,000
samples to estimate the value. Part of the NPARIWAY display associated with the MC option is shown
below.

```
Monte Carlo Estimates for the Exact Test

          One-Sided Pr >= S
          Estimate          0.0084
         99% Lower Conf Limit  0.0060
         99% Upper Conf Limit  0.0108
```
Two-Sided Pr >= |S - Mean|  
Estimate                    0.0166  
99% Lower Conf Limit        0.0133  
99% Upper Conf Limit        0.0199

There is good agreement between the Monte Carlo estimates and the exact p-values. Confidence intervals for the true p-values are also part of the display.

The rank transform approach to nonparametric methods has been summarized as ranking the data and applying the usual methods to the ranks themselves or, in slight extension, functions of the ranks that are called scores. Several longstanding nonparametric tests are either exactly equivalent to rank transform tests or are nearly equivalent to them. PROC RANK can be used to get the ranks and PROC GLM can be used to do the usual computations.

PROC RANK;  
VAR VALUE;  
PROC GLM;  
CLASS GROUP;  
MODEL VALUE = GROUP;

Part of the display from GLM is shown next.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum Sq</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>52.5000</td>
<td>52.5000</td>
<td>14.00</td>
<td>0.0057</td>
</tr>
<tr>
<td>Error</td>
<td>8</td>
<td>30.0000</td>
<td>3.7500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>9</td>
<td>82.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The rank transform approach to this problem would utilize F=14 as the test statistic and/or the p-value, .0057, and declare the means to be different. It may not be obvious why this presentation should be considered equivalent to the KW test. In order to demonstrate that we can note that the KW statistic can be derived from the output above as the sum of squares on the model line, SSM = 52.5 divided by the mean square for Total (not printed). That denominator is 82.5/9 = 9.17. The ratio, SSM/MST, nearly matches the value reported by PROC NPAR1WAY. The difference relates to a continuity correction SAS uses in NPAR1WAY. The equivalence of the two tests stems from the fact that the F value is related to the KW Chi-Square (H) value by the equation $F = \frac{H}{(K-1)/(N-1-H)/(N-k)}$. That relationship is easy to show and is monotonic. F and H increase and decrease together. The test statistics are therefore equivalent. The only difference between the procedures is that we happen to use the Chi-Square distribution in one case and the F distribution in the other, in order determine significance. A similar association exists between Student’s t-test applied to ranks and the WRST discussed earlier in this presentation.

While the above discussion is interesting (to some) it is not clear what practical advantage there is in using PROC GLM to do what amounts to the same thing as the KW test when PROC NPAR1WAY will give those results directly. One advantage is that multiple comparisons are available in PROC GLM but not in PROC NPAR1WAY. While PROC MULTTEST can be used to produce certain distribution free multiple comparison tests, randomization tests for instance, we will focus here on a test that is analogous to a common parametric procedure. Conover\(^8\) describes a method wherein the ranks of the observations can be subjected to the least significant difference (LSD) computations used on the raw data. Simply adding the statement

`MEANS GROUP / LSD;`

to the previous PROC GLM step is all that is needed to cause SAS to carry out the required computations. Excluding the titles and disclaimers about type I error rate, the new output is displayed below.
Alpha                            0.05
Error Degrees of Freedom            8
Error Mean Square                3.75
Critical Value of t           2.30600
Least Significant Difference   3.0815
Harmonic Mean of Cell Sizes       4.2

NOTE: Cell sizes are not equal.
Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>t Grouping</th>
<th>Mean</th>
<th>N</th>
<th>group</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4.000</td>
<td>7</td>
<td>y</td>
</tr>
<tr>
<td>B</td>
<td>9.000</td>
<td>3</td>
<td>x</td>
</tr>
</tbody>
</table>

Of course we only have 2 groups so this step is redundant. This procedure has an interesting and unusual characteristic in that the larger the overall test statistic, H, the smaller the least significant difference “yardstick” is for declaring pairwise comparisons to be significant. Among SAS implemented procedures, the Waller-Duncan procedure has this same characteristic. In addition to the procedure discussed above, “Any of the popular multiple comparisons techniques, including Scheffe’s, Tukey’s, Duncan’s and Fisher’s methods, as well as many other may be applied …. with good results”9. All of the listed “popular methods” are available in SAS and all the user needs to do in order to implement them is to change the option on the MEANS statement according to his/her wishes.

**Ordered Inference**

In 2-sample problems it is common to test directional alternative hypotheses, often called 1-sided alternatives. One-sided tests are more powerful than the corresponding 2-sided test in those cases where the directional alternative is true. Directional alternatives are useful in multi-sample problems. For example, the alternative hypothesis $H_a: \mu_1 \geq \mu_2 \geq \mu_3$ may be of interest in a situation where the three treatments correspond to different dosages and we anticipate that the response (or toxicity) will increase along with dose. Parametric tests for this situation have been proposed but are not in widespread use. ANOVA results often lead to contrasts aimed at linear effects in an attempt to “prove” similar ideas. Such attempts are not wholly satisfactory since “linear” and “ordered” don’t mean the same thing. The Jonckheere-Terpstra (JT) test is a rank based test that was developed for ordered alternative hypotheses and that is available in SAS. For purposes of example consider the fictitious data set below. The response values correspond to STRENGTH.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>18</td>
<td>20</td>
<td>24</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>19</td>
<td>17</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>15</td>
<td>19</td>
<td>26</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>18</td>
<td>21</td>
<td>25</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>21</td>
<td>16</td>
<td>23</td>
<td>23</td>
</tr>
</tbody>
</table>

While groups A, B and C seem to have similar values, groups D and E are larger. The pattern is consistent with the ordered alternative idea but the association is hardly linear. While it seems odd to use a categorical procedure to analyze a continuous response the JT test is implemented is PROC FREQ. The SAS code leads to some output, part of which is displayed.

PROC FREQ;
STATISTIC (JT) 218.0000
Z                      3.7270
One-sided Pr > Z       <.0001
Two-sided Pr > |Z|      0.0002

The normal approximation Z-statistic and the one-sided p-value are the appropriate values to report in this situation. Exact methods are available by adding the statement EXACT JT; to the code above. The NOPRINT option (not shown here) on the TABLES statement may be used to suppress the substantial contingency table output. Even for small problems the computations may be prohibitive for many machines. As an alternative, Monte Carlo methods are available in PROC FREQ. They may be used to estimate the exact p-values in lieu of exact results. The statement EXACT JT / MC; can be used to accomplish that task. The resulting output is displayed below.

Monte Carlo Estimates for the Exact Test

One-sided Pr >= JT
Estimate                   1.00E-04
99% Lower Conf Limit       0.0000
99% Upper Conf Limit       3.576E-04

Two-sided Pr >= |JT|
Estimate                   1.000E-04
99% Lower Conf Limit       0.0000
99% Upper Conf Limit       3.576E-04

Again, the 1-sided results at the top of the display are the ones appropriate to this example. If multiple comparison methods are used after the JT test then the analyses described the last section are available. If 1-sided pairwise comparisons are needed, it is appropriate to divide the p-values by two if the direction of the observed differences is consistent with the alternative hypothesis.

Analyses involving Blocking Factors

In the previous example we now imagine that the rows correspond to a blocking factor (BLOCK) and agree to ignore the previously assumed order in the treatment responses. Having done that the data are in form appropriate for the analyses often applied to randomized complete block designs. The idea is to investigate treatment differences while controlling for a nuisance factor in cases where we don’t think there is any likelihood of interaction between the two. The Friedman test is the standard nonparametric method applied to data of this type. The Friedman test is also available as part of PROC FREQ. This test involves ranking the data within blocks only. Subsequent calculations are performed on those ranks. Friedman’s test may be implemented by the following SAS code.

PROC FREQ;
TABLES BLOCK*GROUP*STRENGTH / CMH2 SCORE=RANK NOPRINT;

The NOPRINT option was used solely to suppress the substantial contingency table output. Output from the procedure is shown below.

Summary Statistics for group by strength
Controlling for block

Cochran-Mantel-Haenszel Statistics (Based on Rank Scores)
<table>
<thead>
<tr>
<th>Statistic Alternative Hypothesis</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Nonzero Correlation</td>
<td>1</td>
<td>12.4163</td>
<td>0.0004</td>
</tr>
<tr>
<td>2 Row Mean Scores Differ</td>
<td>4</td>
<td>15.5102</td>
<td>0.0038</td>
</tr>
</tbody>
</table>

The second line labeled “Row Mean Scores Differ” contains the interesting information. The test statistic, 15.5102, is referred to a Chi-Square distribution with four degrees of freedom. That results in a p-value of .0038. Therefore, the hypothesis of no difference is rejected. While PROC FREQ provides a mechanism for getting the Friedman test for the overall equality of means it does not provide a mechanism for comparing pairs of means subsequent to that analysis.

PROC GLM can be used to perform a rank analog of the ANOVA appropriate to RCB designs. With minimal effort, that same analysis can be used to derive the Friedman test and to do pairwise comparisons between means. The SAS code

```
PROC RANK;
BY BLOCK;
VAR STRENGTH;
PROC GLM;
CLASS BLOCK GROUP;
MODEL STRENGTH = BLOCK GROUP;
```

produces the following output. The Type III results are identical to the Type I results and are, therefore, not displayed. Some other output has been edited out.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>8</td>
<td>38.00000000</td>
<td>4.75000000</td>
<td>6.91</td>
<td>0.0005</td>
</tr>
<tr>
<td>Error</td>
<td>16</td>
<td>11.00000000</td>
<td>0.68750000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>24</td>
<td>49.00000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>block</td>
<td>4</td>
<td>0.00000000</td>
<td>0.00000000</td>
<td>0.00</td>
<td>1.0000</td>
</tr>
<tr>
<td>group</td>
<td>4</td>
<td>38.00000000</td>
<td>9.50000000</td>
<td>13.82</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

The F-statistic and associated p-value displayed provide a the “rank transform” solution to the analysis of this data. The p-value for GROUPS, <.0001, strongly suggests groups differences. Further, the Friedman statistic itself can be gotten with PROC PENCIL as follows. First, we calculate a denominator. That value is gotten by taking the corrected total degrees of freedom and subtracting the block degrees of freedom from it (24 - 4 = 20). That number is divided into the corrected total sum of squares (49/20 = 2.45). The result of that calculation is divided into the sum of squares for GROUP (also, MODEL), 38/2.45 = 15.5102. That value matches the test statistic given by PROC FREQ. Again, there is a simple monotone transformation the rank transform statistic and the traditional Friedman value so the approaches are equivalent. The difference is reduce to an argument about the accuracy of two LSA’s.

Beyond simple understanding, the advantage of this approach is that by adding a MEANS statement, with an appropriate option, any of the common multiple comparison methods used for parametric analyses can be used. That such tests are valid and asymptotically distribution free has been established'.
Beyond Normality

Standard t-tests and ANOVA methods are frequently thought to be sensitive to the assumption of normality. That really isn’t true. Central limit results of various types are such that the standard parametric approaches are reasonably robust to departures from the assumption of normally distributed errors. That being said, it should be added that linear rank tests may be more powerful than their parametric counterparts under certain distributional assumptions. What can be disastrous for standard parametric procedures is the existence of outliers. That is to say that while they are robust, they are not generally resistant. To illustrate the idea we will return to the original example of this presentation. In that example we had three X’s, seven Y’s and the Chi-Square approximation to the p-value (labeled Kruskal-Wallis Test) was seen to be .0164 while the ANOVA p-value was reported as being .0060. If we change the second of the X values from 24 to 224 we substantially increase the difference between the two sample means (from 6 to more than 70!). Surprisingly, the ANOVA or T-test based p-value increases to a value > .10 as is shown below.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Sq</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUP</td>
<td>1</td>
<td>11088.933</td>
<td>11088.9</td>
<td>3.3212</td>
<td>0.1059</td>
</tr>
<tr>
<td>Error</td>
<td>8</td>
<td>26710.667</td>
<td>3338.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>9</td>
<td>37799.600</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Here the WSR test statistic and its p-value remain unchanged. While the preceding situation involves an extreme outlier it is not unusual for questionable observations to be present in a carefully planned study (and “all over the place” in observational data). More frequently than we would like to admit the results of our analyses depend on whether we throw out outliers or leave them in. That puts the analyst in an ethically compromising situation. As we see from the preceding example, that conundrum can sometimes be avoided by the use of rank tests.

References

4. Arbuthnott, J., “An argument for divine providence taken from the constant regularity observed in the births of both sexes. Philosophical Transactions Vol. 27. pp. 186-190, 1710