Approaches to the Nonparametric Tests for Samples Location Comparison in SAS® Software

C.J. Tian and S. Shiue
Janssen Research Foundation

Abstract

The p-values of the nonparametric tests for multiple samples location comparison can be obtained by using one of the three procedures: NPAR1WAY, FREQ, or IML in SAS®. It is observed that the results from these approaches are different.

The main purpose of this paper is to explain the differences of the three approaches. Moreover, the comparison of the three SAS® approaches with the statistical software StatXact® is also included.

I. Statistical Background

As one of the basic statistical techniques, the nonparametric testing methods of multiple samples location comparison are often used in the applied fields. Especially in the pharmaceutical industries, the technique is widely used to perform analysis on ranked data. These tests, such as the Kruskal-Wallis test, Wilcoxon rank sum test, median test, are frequently appeared in the statistical reports of clinical trials for comparison among treatment groups.

Let \( x_{ij}, \quad i=1,2, \ldots, n_j, \quad j=1, \ldots, k \) be \( k \) samples from a continuous distribution differing only (or at most) in their so-called location parameters \( \mu_j, \quad j=1,2, \ldots, k \). The problem considered here is to test the following hypotheses:

- Null hypothesis: \( H_0: \mu_1 = \mu_2 = \cdots = \mu_k \)
- Alternative hypothesis: \( H_A: \mu_i \neq \mu_j \) for some \( i \neq j \).

A typical rank procedure for the nonparametric analysis is to rank all observations jointly by assigning the rank 1 to the smallest value, the rank 2 to the second smallest one, ... and the midrank to each of tied observations in the tied case. Let \( r_{ij}, \quad i=1,2, \ldots, n_j, \quad j=1, \ldots, k \) be \( k \) ranked samples. The statistical inference of nonparametric analysis is merely based on the statistics of ranked samples. A number of different statistics have been proposed for the problem of multiple samples location comparison in the literature, some of them are currently included in SAS®.

II. Approaches in SAS®

2.1. Approach based on PROC NPAR1WAY

The procedure NPAR1WAY is probably the first choice in SAS® to perform the nonparametric analysis of multiple samples location comparison. It provides the following four popular tests:

- Kruskal-Wallis test,
- Median test,
- Van der Waarden test,
- and Savage test.

As is well known, the SAS® procedure NPAR1WAY performs one-way analysis of variance on ranked data. It implies such pre-assumption that all \( k \) samples must be independent. The property of independence contributes the limitations on the analysis implemented in PROC NPAR1WAY. Apparently, it is not possible to considered any kind of 'interaction' or 'association' with other variables through this approach. Besides, only asymptotic p-values are generated from the procedure. Also, a continuity correction is performed in the calculation of p-values.

2.2. Approach based on PROC IML

To provide the exact p-values of the above four tests in the PROC NPAR1WAY, a program was developed by J. L. Barry(1993)\(^\text{9}\), which is based on SAS® procedure IML. The approach is restricted to the case of two independent samples.

2.3. Approach based on PROC FREQ

The FREQ procedure is designed to produce one-way to k-way frequency and crosstabulation tables and to perform some statistical analysis related to these tables. When the option SCORE=RANK is used in FREQ procedure, the analysis is actually performed on the ranked samples, so that the results would be equivalent to the nonparametric methods. For example, in the case when there is only one stratum, the ANOVA statistic is a Kruskal-Wallis test. In other words, the item 'Row Mean Scores Differ' in Cochran-Mentel-Haenszel Statistics, which can be found in the output of PROC FREQ, presents the results from the Kruskal-Wallis test. If there are multiple strata, the results then turn to present the stratified Kruskal-Wallis test.

The advantage of the approach is that the 'association', a type of dependence among samples, can be concerned in the procedure FREQ.

Note that the p-values generated from the approach are also asymptotic results. However, the p-values of the approach are different from those generated by procedure NPAR1WAY, because the continuity correction is not performed in the procedure FREQ.

2.4. Approach by Invoking StatXact® in SAS®

The statistical software StatXact® is a special tool for the categorical data analysis, it is getting popular recently. Some nonparametric tests are included in the software, for example, the Kruskal-Wallis test. The software can be used in different ways. If ranked samples are imported to StatXact®, then the corresponding results may be used for the purpose of nonparametric inference. In such case, the results of the item 'Linear-by Linear Association' in StatXact® are equivalent to the item 'Row Mean Scores Differ' of CMH test in SAS®.

An interface between StatXact® and SAS® is now available\(^\text{10}\). By using the interface we can invoke StatXact® in a SAS® system: import data from a SAS® dataset; access StatXact® and store the output results into a SAS® dataset.

Both types of exact and asymptotic p-values can be obtained.
through the approach. Note that the asymptotic approximation herein is not treated by any kind of continuity correction.

2.5. Comparison of the Four Approaches

Comprehensively, two major types of data are considered: unstratified (independent samples) or stratified (dependent samples); three types of p-values: exact p-values, asymptotic p-values without continuity correction as well as the asymptotic p-values with continuity correction, are generated through all four approaches. However, no one covers all functions. The differences of the above approaches are briefly summarized in the following table.

Table 1. Comparison Summary

<table>
<thead>
<tr>
<th>Approach</th>
<th>Main Tool</th>
<th>Stratified</th>
<th>P-Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>PROC NPAR1WAY</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>B</td>
<td>PROC IML</td>
<td>X</td>
<td>X*</td>
</tr>
<tr>
<td>C</td>
<td>PROC FREQ</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>D</td>
<td>Stata**</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

* with continuity correction.
** only for two samples comparison.

III. Wilcoxon Rank Sum Test

For convenience, we use the popular Wilcoxon rank sum test as an example to explain the differences in details. The Wilcoxon rank sum test is a special case of the Kruskal-Wallis test for two samples. Without loss of generality, we always assume $n_1 \leq n_2$. The statistic for the testing is the so-called rank sum, which is defined as

$$ T_1 = \sum_{i=1}^{2} T_i, $$

and two related statistics are

$$ E = T_1 - (1/2) n_1 (n_1 + n_2 + 1), $$
$$ S = \left[ (1/12) n_1 n_2 (n_1 + n_2 + 1) \right]^{1/2}. $$

The distribution of the statistic $T_1$ can be determined by the permutation method or asymptotic approach. The permutation method leads to the exact p-values. In the asymptotic approach, there are two different algorithms

(2) \[ Z = E / S, \]

or

(3) \[ Z^* = (E - \gamma) / S, \]

where the continuity correction $\gamma$ is defined as -0.5 when $E$ is negative, as 0.5 when $E$ is positive, or as 0 when $E = 0$.

From the central limit theorem, under certain conditions either one of $Z$ and $Z^*$ is considered as the standard normal variate.

Among the four approaches in Table 1, the asymptotic p-values of approach A are calculated by using the formula (3) and the other asymptotic p-values from approaches C and D are produced by using the formula (6).

A traditional way to get exact p-value is to use a numerical table of p-values, which is usually attached to some common reference books\(^{5,6}\). In the table, the one-side p-values, as a function of the variables $n_1$, $n_2$, and $T$,

\[ P(n_1, n_2, T) = \text{Prob} (T_1 \leq T), \]

are listed for all cases $3 \leq n_1, n_2 \leq 10$.

All p-values mentioned in the rest of the section are one side p-values.

In order to have some intuitive impression on the differences among the three types of p-values, we generated all p-values through the approaches A, C, and D for $4 \leq n_1, n_2 \leq 20$. Some interesting observations are received from the comparison. The asymptotic p-values are providing a better fitting to the exact p-values than the asymptotic p-values with continuity correction. All of the differences $|P_1 - P_2|$, $|P_1 - P_3|$, and $|P_2 - P_3|$ are smaller than 0.02 when sample sizes $n_1$, $n_2$ are larger than 14, where $P_1$, $P_2$, and $P_3$ represent the exact p-values, asymptotic p-values, and asymptotic p-values with continuity correction respectively.

IV. A Numerical Example

A numerical example in [5, p.191] is used here for the demonstration.

The two samples are

Group 1 = 6.5, 9.1, 8.4, 3.7
and Group 2 = 14.5, 11.4, 16.4, 11.9, 10.3

The sample sizes $n_1$, $n_2$ are 4 and 5, and the rank sum $T_1$ is easily obtained as 10.

Suppose that the data have been imported into a SAS\(^*\) dataset, aa, with two variables: group and value. The following SAS\(^*\) statements are for producing p-values of the Wilcoxon rank sum test through the four different approaches.

Approach A.

```sas
proc nparlway data=aa wilcoxon;
```

Approach B.

```sas
%permtest('wilcoxon',4,5,10,pval);
```
Approach C.

```sas
proc freq data=aa; table group*value/cmg
   scores=rank,
   output out=out CMHRMS;
proc print data=out;
title 'Results from proc FREQ ';
```

Approach D.

Suppose that the system files of the software StatXact are installed in directory c:\sxturbo, and a SAS® dataset bb is stored in a directory c:\user, which contains two variables: group and rank (not the original value).

```sas
options nowait;
libname out xport 'c:\user\sas2sxt.xpt';
libname db 'c:\user';
data Qut.bbiset db.bb;
data _null_;
file 'c:\user\sas2sxt.cmd';
put 'RE c:\user\sas2sxt.log';
put 'IN SAS c:\user\sas2sxt.xpt ro=group co=rank';
put 'WX EX';
put 'XL c:\user\sas2sxt.out';
put 'QU';
run;
X 'c:\sxturbo\sxt c:\user\sas2sxt.cmd';
The SAS® program is for using the interface between SAS® and StatXact® and the output from the program is actually generated by the software StatXact®, which is stored as an ASCII file 'c:\user\sas2sxt.out'. It is easy to convert the file into a SAS® dataset file by using simple SAS® statement.

All p-values from the four different approaches are listed in the Table 2.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Main Tool</th>
<th>P-Values (2-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>PROC NPAR1WAY</td>
<td>0.0200</td>
</tr>
<tr>
<td>B</td>
<td>PROC IML</td>
<td>0.0159</td>
</tr>
<tr>
<td>C</td>
<td>PROC FREQ</td>
<td>0.0143</td>
</tr>
<tr>
<td>D</td>
<td>StatXact®</td>
<td>0.0143</td>
</tr>
</tbody>
</table>

The above numerical results are produced under the following computer environment:

1. VAX/VMS with network DECNET
2. SAS® Version 6.08
3. P C -- COMPAQ XL560
4. StatXact® Turbo Version 2.14a

V. Conclusion Remarks

A few more words are necessary for the approach D. There is not any mainframe version of StatXact® available currently, so that the approach is only implemented on PCs. Also, when accessing StatXact from a SAS® system, the active screen is held by the system command SXT of the software StatXact®. It means that invoking the software StatXact in a SAS® batch program is actually impossible under the DOS environment. From the practical viewpoint, the total time, which are consumed for importing SAS® dataset, invoking StatXact and exporting results to SAS® dataset, would be counted while comparing with other SAS® procedures.

To choose a proper approach in practice is a real complexity. It depends on a number of factors, such as the experiment design, the ranges of sample sizes, the accuracy requirement as well as computer environment, and so on.

A new version of PROC NPAR1WAY is under developing, which will include the exact p-values, as informed by the SAS® Institute. We are expecting a better solution to the samples location comparison in the upcoming new version of SAS®.

References