ABSTRACT

The Jonckheere-Terpstra test is a nonparametric test for ordered alternatives that is not currently available in the SAS™ system. This paper reviews the use of the Jonckheere-Terpstra test and presents an algorithm that can be used to calculate the Jonckheere-Terpstra test statistic and its asymptotic p-value. The algorithm is written to be run in VMS™ but could be easily generalized to run in a macro environment on other operating systems. The advantages of this algorithm over previously presented methods for calculating the Jonckheere-Terpstra statistic using the SAS™ system are discussed.

KEY WORDS

Jonckheere, Terpstra, Nonparametrics, Ordered Alternatives

INTRODUCTION

In laboratory experiments investigating the teratogenic potential of chemical compounds, we are frequently interested in assessing the significance of observed dose-response trends in variables such as mortality and anomalies. In many cases the form of the dose-response curve will be unknown but can be assumed to be a monotonic increasing (decreasing) function of dose (Wilson, 1973). In this case the k-sample rank test proposed by JONCKHEERE (1954) and TERPSTRA (1952) can be employed. This procedure tests the hypothesis that the k populations are identical against the alternative that the populations are stochastically ordered in a specified manner.

In this paper, the Jonckheere-Terpstra test will be reviewed. An algorithm to perform the Jonckheere-Terpstra test will be presented. To illustrate the use of this algorithm, an example will be presented.

JONCKHEERE-TERPSTRA TEST

Suppose we have random samples of sizes \( n_1, n_2, \ldots, n_k \) from each of \( k \) populations and let \( F_i \) denote the distribution function of the \( i \)th population \( (i=1, \ldots , k) \). The Jonckheere-Terpstra test is used to test the following hypotheses:

\[
H_0: F_1 = F_2 = \ldots = F_k \\
H_A: F_1 \leq F_2 \leq \ldots \leq F_k
\]

where at least one of the inequalities is strict. Denote the \( j \)th observation from the \( i \)th sample as \( X_{ij} \). To compute the test statistic, first compute the \( k(k-1)/2 \) Mann-Whitney counts \( U_{uv} \), \( u < v \) as follows:

\[
U_{uv} = \sum_{i=1}^{n_i} \sum_{j'=1}^{n_{i'}} \Phi(x_{iu}, x_{i'j'})
\]

where

\[
\Phi(a, b) = \begin{cases} 
1 & \text{if } a < b \\
0.5 & \text{if } a = b \\
0 & \text{if } a > b
\end{cases}
\]

Then the test statistic is defined as follows:

\[
J = \sum_{u<v} U_{uv} = \sum_{u=1}^{k-1} \sum_{v=u+1}^{k} U_{uv}
\]

For small samples, this statistic should be compared to tables to find the level of significance (see Lehman, 1975 for tables). For larger samples, define

\[
J^* = \frac{J - E(J)}{\text{VAR}(J)^{1/2}}
\]

Under the null hypothesis, this can be rewritten (Lin, 1970) as follows:

\[
J^* = \frac{\sum_{i=1}^{H} (n_i - 1) d_i}{\left( \sum_{i=1}^{H} n_i(n_i-1)/2 \right)^{1/2}}
\]

where there are \( H \) distinct observations and the \( h \)th observation was observed \( d_h \) times and we define

\[
N = \sum_{i=1}^{H} n_i
\]

Then when \( H_0 \) is true, the statistic \( J^* \) has an asymptotic \( N(0,1) \) distribution. Therefore, at an \( \alpha \) level of significance:

\[
\text{reject } H_0 \text{ if } J^* > Z(\alpha) \\
\text{accept } H_0 \text{ if } J^* < Z(\alpha)
\]

The JONCKHEERE-TERPSTRA test could likewise be used to test for a decreasing trend by reversing the inequalities in the function \( \Phi(a, b) \).

COMPUTER ALGORITHM

Various means of executing the Jonckheere-Terpstra test using the SAS™ system have been suggested (for example Morris, 1989 and Hoop, 1985). The program used in this paper has the advantage of producing a data set which contains the test statistic, p-value and the direction of the test result that can be incorporated into reporting programs used to generate final reports. Another advantage of this program is that it analyzes any number of variables with a single execution of the procedure.
This program was written using the SAS™ system version 5.16 on a VAX™ 785 using the VMS™ operating system.

**Input To Algorithm**

The data must be in a data set named TRANS and this data set must have the variables SUBJ, GROUP as well as the variables to be analyzed. All variables must be numeric and no other variables can be in the data set. The variable SUBJ is an identifier for the subject number and the variable GROUP is a numeric identifier for the treatment group. The group numbers must be ordered in the hypothesized order of the test.

**Example Input Data:**

```plaintext
<table>
<thead>
<tr>
<th>Observation</th>
<th>Subject</th>
<th>Group</th>
<th>Test1</th>
<th>Test2</th>
<th>Test3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0.93</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0.87</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

The input data must contain at least two nonempty groups. There must be at least three observations and two unique observations in the data set. If these conditions are not met the program outputs missing values for the test statistic, p-value and direction flag.

**Program Execution**

A utility program which calculates the Jonckheere-Terpstra test statistic for each variable in the input data set is presented in Appendix 1. The one sided test in the direction suggested by the data is performed and the p-value is reported with a direction flag denoting which one sided test was done. If the p-values from the tests for increasing and decreasing trends are equal, the p-value is reported with a flag indicating that no trend was detected. If the variance of \( J \) is equal to zero, and \( J \) and the expected value of \( J \) are equal, the program sets the test statistic to be equal to zero. If the p-value is less than 0.0005, the program sets the p-value to be equal to 0.0005.

Every observation must have nonmissing values for SUBJ and GROUP, however missing values are acceptable for the analysis variables, calculations take missing values into account. Calculations are performed in the data sets TRANS and TABLE and the macro variables NGRPS, NTOTAL and Ni for i=1 to the number of groups.

**Program Output**

The program outputs a data set named TRANS which contains one observation per variable analyzed. The following variables are output:

- **NAME.** = the name of the variable analyzed
- **JONC** = test statistic \( J \)
- **PJONC** = the one sided p-value using the normal distribution approximation for large samples

**EXAMPLE**

Suppose we have the following data:

```plaintext
<table>
<thead>
<tr>
<th>GROUP</th>
<th>SUBJ</th>
<th>RESULT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.30</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.37</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.40</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.60</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.60</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.36</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.45</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.55</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.60</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.65</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.67</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.28</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.30</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.41</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.42</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.29</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.29</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>0.35</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0.45</td>
</tr>
</tbody>
</table>
```

If we create a data set called TRANS containing this data in the variables GROUP, SUBJ and RESULT, then we can call the Jonckheere-Terpstra program to perform the test. Our resulting data set looks like this:

```plaintext
<table>
<thead>
<tr>
<th>OBS</th>
<th>JONC</th>
<th>PJONC</th>
<th>JONC.D</th>
<th>NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2.112</td>
<td>0.0173775</td>
<td>-</td>
<td>RESULT</td>
</tr>
</tbody>
</table>
```

Thus our trend result is decreasing with a p-value of 0.017.

SAS is a registered trademark of SAS Institute Inc., Cary, NC, USA.
VAX and VMS are registered trademarks of Digital Equipment Corp., Maynard, MA, USA.
REFERENCES


Terpstra, T.J., 1952: The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. Indagationes Mathematicae 14, 327-333.

PROC SORT DATA=TRANS; BY GROUP; RUN;

DATA WEL_; SET TRANS END-GAP; BY GROUP; IF FIRST.GROUP THEN COUNT1=1; IF NOT THEN CALL SYMPUT(CHAR(4),LEFT(COUNT1)); RUN;

PROC MEANS DATA=TRANS HFORMAT=1.16 HLABEL=('NO VALUE', 'VALUE'); VAR GROUP; OUTPUT OUT=TABLE N=1; RUN;

DATA TABLE_; SET TABLE_; LENGTH C $1; RETAIN N 0; C="" 1 LEFT(C); CALL SYMPUT(C,LEFT(N));
IF N EQ 0 THEN CALL SYMPUT(CHAR(1),LEFT(N)); ELSE CALL SYMPUT(CHAR(2),LEFT(N)); RUN;

* DELETE EXTRA VARIABLES FROM THE INPUT DATA SET. *
DATASETS GROUPS=S;

* TRANSPOSE DATA AND CALCULATE TEST STATISTIC *
PROC TRANPOSE DATA=TRANS OUT=TRANS;
RUN;

DATA TRANS_; SET TRANS;
NEXT _NAME_;
RUN;

DATA TRANS;
* ARRAYS; COL = ARRAY OF VALUES TO BE ANALYZED GROUP = ARRAY OF GROUP IDENTIFIERS VAL = ARRAY OF GROUP VALUES TAB = ARRAY OF THE NUMBER OF OCCURRENCES OF EACH VALUE IN THE ARRAY VAL = ARRAY OF THE NUMBER OF OCCURRENCES OF EACH VALUE IN EACH GROUP VARIABLES: N = THE TOTAL NUMBER OF OBSERVATIONS NGROUP = THE NUMBER OF GROUPS NSAMPLE = THE NUMBER OF OBSERVATIONS FROM EACH GROUP NSAMPLE(foo) = THE NUMBER OF OBSERVATIONS FOR EACH VALUE foo = VALUE EXPECTED VALUE OF THE JOHNSON'S STATISTIC JONC = NORMALIZED JOHNSON'S STATISTIC JONC_D = TRANS DIRECTION *
SET TRANS_; ARRAY COL[*] COL-ARRAYTOTAL; ARRAY VAL[*] VAL-ARRAYTOTAL; ARRAY TAB[*] TAB-ARRAYTOTAL; ARRAY NSAMPLE[*] NSAMPLE-ARRAYTOTAL;

* FILL IN THE ARRAY WITH THE NUMBER OF OBSERVATIONS IN EACH TREATMENT GROUP; LENGTH _ID_ 3; _ID_ = 1 THEN TO _ID_ TO _ID_; PRINT "NO VALUE", _ID_, PRINT "VALUE", _ID_; RETAIN _ID_; RUN;

* THE SET IS NOW MADE DIFFERENT VALUES WHERE ARE, AND DETERMINE THE NUMBER OF OBSERVATIONS AT EACH VALUE;