General Application of Youden’s Rank Sum Test for Outliers and Tables of One-Sided Percentage Points

T. A. Willke*

(February 18, 1964)

The rank sum test for outliers advanced by W. J. Youden provides a method for detecting if the measurement distribution of any one of a group of objects has a mean significantly different from the rest. This paper discusses a more general application of the rank sum procedure which permits a similar test on other parameters, such as the variance, with the same tables. Tables of the critical values of the extreme rank sum and the corresponding significance levels for one-sided tests are given in this paper to supplement similar tables for two-sided tests already published.

Introduction

W. J. Youden advanced in [1] a nonparametric rank sum procedure for detecting whether the distribution of measurements of any one of the rows of a two-way table of measurements had a mean that was noticeably different from the means of the other rows. In Youden’s paper approximate 5 percent points for a two-sided test were given. Because of the discrete distribution of the rank sum statistic exact 5 percent points ordinarily do not exist. This technique was further investigated in [2], and there the tables were extended to include approximate 3 percent and 1 percent points along with the actual values of the significance levels.

In all the writings and applications so far, this test has been used to detect an object (or laboratory) whose measurement distribution differs from the distribution of the other objects in its mean or median. However, the test, with the same tables, can often be used to detect an object whose distribution is different with respect to another parameter, for instance, the variance. The purpose of this note is to point out and illustrate the scope of possible applications of Youden’s rank sum test, and to give a table of critical regions and significance levels for one-sided tests.

1. The Ordinary and the General Application

The ordinary application of Youden’s test as discussed in [1] and [2] is as follows: Suppose there are I objects and J judges. Each judge independently ranks the I objects according to some property. Thus, each object receives a rank, an integer between 1 and I, from each judge, and the J ranks for each object are summed. Then if any object has a rank sum that is extremely high (or low) it can be concluded that the ranks were not assigned just by chance but that the objects tends to rank higher (or lower) than the rest.

In many cases the rankings are assigned on the basis of measurements. That is, there is a measurement for each object-judge combination, and, for a given judge, the object with the highest measurement gets a rank of 1, the object with the next highest gets a rank of 2, etc. (The ranks can be assigned in the reverse order if that is more convenient.) Then if an object gets an extreme rank sum it can be concluded that its measurements tend to be high or low accordingly. This application is actually a test on the means or medians of the measurement distributions for the different objects.

The general application of this technique provides a test on other parameters, and it can be described in the following way. Suppose we have a two-way classification (I×J) with Kij observations in cell (i, j). The range of the subscript i and j will be 1, 2, . . . , I, and 1, 2, . . . , J respectively throughout the paper. Let the rows correspond to objects being judged and the columns correspond to judges. Now fix attention on the jth column. Suppose the Kij measurements for row i are each distributed with a probability density function, f(xi; θj(i)), which depends on a row parameter θj that has the value θj(i) for the ith row. Note that each column may have its own distribution, fj, and its own set of row parameters θj(i).

We wish to test the hypothesis that all the row parameters for any given column are the same,

\[ H_0: \theta_j(1) = \theta_j(2) = \ldots = \theta_j(I), \quad j = 1, \ldots, J \]

against the alternative that one of the rows has parameters consistently higher or lower than the rest.

To use the rank sum test there must be an estimate \( \hat{\theta}_j(i) \) of the row parameter \( \theta_j(i) \) for each cell \((i, j)\). The distribution of these estimates must be continuous and must satisfy the following conditions.

a) For each \( j \), the joint distribution of \( \hat{\theta}_j(1), \hat{\theta}_j(2) \ldots \hat{\theta}_j(I) \) under \( H_0 \) is symmetric with respect to the row indexes.

b) The estimates in one column must be independent of the estimates in every other column, i.e., the estimates are mutually independent columnwise. This is the only between-column restriction.

* Figures in brackets indicate the literature references at the end of this paper.
* University of Maryland, College Park, Md.; part time at the National Bureau of Standards.
When the above requirements are satisfied, then the rank sum test can be used for $H_0$ as follows: Within each column compute the estimates $\hat{\theta}_j(i), \ldots, \hat{\theta}_j(I)$ of the row parameters and rank the rows accordingly. Then sum the ranks for each row. An extreme (large or small) value of the rank sum leads to rejection of $H_0$. The earlier papers, [1] and [2], treated the case $K_{ij} = 1$, in which the single observation in each cell is an estimate of the mean or median of the distribution in that cell. In the general application of the test, instead of ranking the rows (or objects) according to the measurements, we rank them according to estimates computed from the measurements. If (a) and (b) are fulfilled then every permutation of the ranks is equally likely and the original tables for the rank sum test apply. If one of the rows has parameters consistently higher (or lower) than the rest, then that row will tend to receive high (or low) ranks and hence an extreme rank sum. Of course, the efficacy of the test depends on the quality of the estimators, $\hat{\theta}_j(i)$.

The conditions (a) and (b) are sufficient for the tables in [1, 2], and in this paper to apply. It is interesting to note in what situation (a) is fulfilled.

For instance, (a) is implied by

(a') For each $j$, $\hat{\theta}_j(1), \ldots, \hat{\theta}_j(I)$ are mutually independent and identically distributed.

Thus, consider the common situation where (1) $K_{ij} = K_j$, that is, there is the same number of measurements in each cell in the same column, (2) the measurements are all independent between cells, and (3) $\hat{\theta}_j(i)$ is computed only from the measurements in cell $(i, j)$ and is computed in the same way for every cell. Then (a') holds and the test can be used. It is not necessary that there be the same number of measurements per cell from column to column. For example, suppose the test is on medians and $\hat{\theta}_j(i)$ is the sample median for each cell. Then it is permissible to have, say, three measurements per cell in the first column, one per cell in the second column, four per cell in the third, etc.

There are other situations where the $\hat{\theta}(i)$ are not computed from the measurements in cell $(i, j)$ only. For example, suppose $K_{ij} = 1$ and $\hat{\theta}(i)$ is the absolute deviation of the measurement in $(i, j)$ from the column mean. Then (a') does not hold because the $\hat{\theta}(i)$'s are not independent; however, (a) may still hold and then the test can be used. In this example, if the measurements in the cells are identically and independently distributed under $H_0$, then (a) is fulfilled because the joint distribution of the $\hat{\theta}(i)$ for each $j$, although not independent, is still symmetric with respect to the row indices.

**Example 1, Test on Means:** Consider the usual test on the means. If $K_{ij} = 1$, and $x_{ij}$ is the measurement in cell $(i, j)$, then $\hat{\theta}(i) = x_{ij}$ is the estimate of the $i$th row mean for column $j$. If $K_{ij} = K_j > 1$, then the cell average or the cell median would be used for $\hat{\theta}(i)$.

**Example 2, Test on Variance:** Suppose the measurements have the same distribution, row to row, except possibly for the variances $\sigma^2_{ij}$. We wish to test

$$H_0: \sigma^2_{ij} = \sigma^2_{ij} = \ldots = \sigma^2_{ij}, \quad \text{for } j = 1, \ldots, J$$

Here we need $K_{ij} = K_j \geq 2$, then we can take $\hat{\theta}(i) = \frac{x_{ij}}{\sqrt{K_j}}$, the sample within-cell variance. In this case the estimates are independent, hence (a') holds, and the rank sum test can be applied. A one-sided test might be appropriate here.

An alternative test for equality of variances could be based on the within-cell sample ranges, or any other dispersion measure.

**Example 3, Test on Accuracy:** Suppose $K_{ij} = 1$ and we wish to discover whether any row is substantially less accurate than the others. By this is meant that its means differ from the other row means, or it has larger variances, or any combination of these such that its measurements tend to differ from the measurements of the other rows.

The measurements $x_{ij}$ are assumed to have two parameter distributions with means $\mu_{ij}$ and variances $\sigma^2_{ij}$. We test

$$H_0: \mu_{ij} = \mu_{ij} = \ldots = \mu_{ij}$$

$$\sigma_{ij} = \sigma_{ij} = \ldots = \sigma_{ij}, \quad j = 1, \ldots, J$$

against the alternative described above. The rank sum test can be used by letting

$$\hat{\theta}(i) = |x_{ij} - x_j|$$

where $x_j = \frac{1}{I} \sum x_{ij}$, and then ranking the rows accordingly. If some row's measurements deviate consistently from the column averages, it will receive an extreme rank sum. Notice that the statistics $\hat{\theta}(i)$ satisfy condition (a), hence the regular rank sum tables apply. A one-sided test is appropriate here.

It is interesting to see just what the statistics $\hat{\theta}(i)$ do estimate. First, it clearly does not matter whether we rank by using $\hat{\theta}(i)$ or $\hat{\theta}(i)$ so we look at the latter. It turns out that

$$E[\hat{\theta}(i)] = E[(x_{ij} - x_j)^2]$$

$$= \frac{1}{F^2} \left[ (I - 1)\sigma^2_0 + \left\{ \sum (\mu_{ij} - \mu_{ij}) \right\}^2 + \sum \sigma^2_{ij} \right]$$

where $\Sigma$ means the sum over $i' \neq i$.

Consider the special case where all but one of the rows are the same. Let

$$\mu_{ij} = \mu_{ij} = \ldots = \mu_{ij} = \mu_j$$

$$\sigma^2_{ij} = \sigma^2_{ij} = \ldots = \sigma^2_{ij} = \sigma^2_j$$

and

$$\mu_{ij} = \mu_j + \epsilon_j$$

$$\sigma^2_j = \sigma^2_j + \beta_j$$

This application of Youden's test was first suggested by Robert B. Dean of the Borden Chemical Co. In a recent paper [3] T. W. Lashof proposed a "center outward" ranking procedure of a similar nature. He suggested that, say, for $n$ odd, the median measurement be awarded a "rank" of 1, the next larger and the next smaller both a "rank" of 2 and so on. No tables are available for Lashof's technique.
Thus

\[ E[\hat{\theta}_0(1)] - E[\hat{\theta}_0(2)] = \frac{1}{N^2} (\sigma_1^2 + \sigma_2^2) \]

Hence, if either the \( \beta_2 > 0 \) and they are large, or if the \( \alpha_j \)'s are large, then \( \theta_j(1) \) will tend to be larger than \( \theta_j(2), \ldots, \theta_j(N) \) and the first row will receive an extreme rank sum.

### Table 1: 

A one-sided rank sum test at nominal significance levels of 1, 3, and 5 percent

<table>
<thead>
<tr>
<th>( \alpha ) = P(\text{reject} &lt; R) = P(\text{max} &gt; I - R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I = 3</td>
</tr>
<tr>
<td>I = 4</td>
</tr>
<tr>
<td>I = 5</td>
</tr>
<tr>
<td>I = 6</td>
</tr>
<tr>
<td>I = 7</td>
</tr>
<tr>
<td>I = 8</td>
</tr>
<tr>
<td>I = 9</td>
</tr>
<tr>
<td>I = 10</td>
</tr>
<tr>
<td>I = 11</td>
</tr>
<tr>
<td>I = 12</td>
</tr>
<tr>
<td>I = 13</td>
</tr>
<tr>
<td>I = 14</td>
</tr>
</tbody>
</table>

#### No. of objects \( j \)

<table>
<thead>
<tr>
<th>No. of judges ( j )</th>
<th>( R )</th>
<th>( a )</th>
<th>( R )</th>
<th>( a )</th>
<th>( R )</th>
<th>( a )</th>
<th>( R )</th>
<th>( a )</th>
<th>( R )</th>
<th>( a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.0014</td>
<td>0.0016</td>
<td>0.0018</td>
<td>0.0020</td>
<td>0.0022</td>
<td>0.0024</td>
<td>0.0026</td>
<td>0.0028</td>
<td>0.0030</td>
<td>0.0032</td>
</tr>
<tr>
<td>4</td>
<td>0.0022</td>
<td>0.0024</td>
<td>0.0026</td>
<td>0.0028</td>
<td>0.0030</td>
<td>0.0032</td>
<td>0.0034</td>
<td>0.0036</td>
<td>0.0038</td>
<td>0.0040</td>
</tr>
<tr>
<td>5</td>
<td>0.0030</td>
<td>0.0032</td>
<td>0.0034</td>
<td>0.0036</td>
<td>0.0038</td>
<td>0.0040</td>
<td>0.0042</td>
<td>0.0044</td>
<td>0.0046</td>
<td>0.0048</td>
</tr>
</tbody>
</table>

See footnote at end of table.

57
To avoid duplication in the tables these values of \( R \) do not give \( \alpha \)'s as close as possible to the nominal significance levels listed. They are already given in the column immediately adjacent, hence the tabulated \( R \) gives another still higher (or lower) significance level.

### 2. Tables for a One-Sided Test

In both the ordinary application of the rank sum test for outliers and in some of the examples given in sec. 1, one-sided tests may be appropriate. Table 1 gives approximate 0.01, 0.03, and 0.05 one-sided probability points for the extreme rank sum. It is tabulated in much the same way as the table of two-sided probability points given in [2].

The table can be described as follows: Let \( I \) be the number of objects ranked and \( J \) be the number of judges. Then the smallest that a rank sum could be is \( J \) and the largest it could be is \( IJ \). Let \( r_i \) be the rank sum of the \( i \)th object, and let

\[
\begin{align*}
R_{\text{max}} &= \max r_i, \\
R_{\text{min}} &= \min r_i
\end{align*}
\]

Let \( R \) be an integer and \( \alpha \) be a probability such that

\[
\alpha = P(r_{\text{min}} \leq J + R) = P(r_{\text{max}} \geq IJ - R)
\]

under the null hypothesis that all rankings are equally likely.

Table 1 gives, for \( 3 \leq I \leq 15, 3 \leq J \leq 15 \), the values of \( R \) yielding significance levels, \( \alpha \), as close as possible to the nominal significance levels of 0.01, 0.03, 0.05. In small print next to each \( R \) is given the corresponding value of \( \alpha \).

For example, suppose \( I = 8 \) and \( J = 6 \). Then all the rank sums must be between 6 and 48. From table 1 we see that \( R = 4 \) yields \( \alpha = 0.0064 \). Thus, a one-sided test for large rank sums at a significance level 0.0064 can be made by rejecting the hypothesis when \( R_{\text{max}} \geq 48 - 4 = 44 \). That is, if any rank sum is 44 or more, it is rejected as too large. Similarly, the corresponding test for small rank sums would be to reject whenever \( R_{\text{min}} \leq 6 + 4 - 10 \).

From the same line of the table we see that a critical region of \( r_{\text{max}} \geq 42 \) has a significance level of \( \alpha = 0.028 \), and a critical region of \( r_{\text{max}} \geq 41 \) has \( \alpha = 0.052 \). Because of the discrete nature of the rank sum statistic, \( \alpha = 0.0064, 0.028 \), and 0.052 are the closest significance levels possible to 0.01, 0.03, and 0.05, respectively.

Table 1 can be thought of as an abbreviated table of the distribution function of the largest (or smallest) rank sum. In several places the values of \( R \) yielding \( \alpha \)'s closest to 0.03 and to 0.05 are the same. To avoid duplication and to offer a wider choice of significance levels in the table, the value of \( R \) belonging to both 0.03 and 0.05 is given in the 0.03 column and the next higher value of \( R \) is given in the 0.05 column even though it doesn’t “belong” there. The same procedure was followed whenever possible and when there was a duplication in the 0.01 and the 0.03 column. These places in the table are marked with an “*”.

### 3. Computation of the Table

The significance level \( \alpha \) for any given \( R \) cannot be computed exactly but bounds, \( \alpha_U \) and \( \alpha_L \), such that \( \alpha_L \leq \alpha \leq \alpha_U \) can be. Using eqs (2), (3), and (6) of [2] \( \alpha_U \) and \( \alpha_L \) were computed for \( I \leq 11 \). For the 0.01 (0.03) level, \( \alpha_U \) and \( \alpha_L \) always agree to (three) decimal places, hence the values in the table for \( \alpha \) are exact to the four (three) decimal places given.

For the 0.05 level \( \alpha_U \) and \( \alpha_L \) did not always agree when rounded to three places although

\[
\alpha_U - \alpha_L < 0.0005
\]

all the time. The rounded value for \( \alpha_L \) was chosen since \( \alpha_L \) is a better approximation to \( \alpha \).

For \( I \geq 12 \), \( \alpha_L \) was not computed, hence \( \alpha_U \), not \( \alpha \), is given in the table. From the computations performed for [2] for large \( I \) it can safely be assumed that \( \alpha_U \) agrees with \( \alpha \) to 3 decimal places in the 0.01 and 0.03 column and that the \( \alpha_L \)'s given in the 0.05 column are in error by at most 0.002. Hence \( \alpha_U \) can be used as a good approximation to \( \alpha \).

### 4. References


(Paper 68B2–116)