JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematics and Mathematical Physics
Vol. 68B, No. 2, April-June 1964

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][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 ($I \times J$) with K_{ij} observations in cell (i, j). The range of the subscripts i and j will be $1, 2, \ldots, I$, and $1, 2, \ldots, J$ 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 K_{ij} measurements for row i are each distributed with a probability density function, $f_j(x_i; \theta_j(i))$, which depends on a row parameter θ_j that has the value $\theta_j(i)$ for the ith row. Note that each column may have its own distribution, f_j, and its own set of row parameters $\theta_j(i)$.

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.

[1] 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 $\hat{\theta}_j(1), \ldots, \hat{\theta}_j(I)$ 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 $\hat{\theta}_j(i)$ are not computed from the measurements in cell (i, j) only. For example, suppose $K_{ij}=1$ and $\theta_j(i)$ is the absolute deviation of the measurement in (i, j) from the column mean. Then (a′) does not hold because the $\hat{\theta}_j(i)$'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 $\theta_j(i)$ 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 $\hat{\theta}_j(i)=x_{ij}$ is the estimate of the ith row mean for column j. If $K_{ij}=K_j>1$, then the cell average or the cell median would be used for $\hat{\theta}_j(i)$.

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

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

Here we need $K_{ij}=K_j\geq 2$, then we can take $\hat{\theta}_j(i)=s^2_{ij}$, 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.[2]

The measurements x_{ij} are assumed to have two parameter distributions with means μ_{ij} and variances σ^2_{ij}. We test

$$H_0: \mu_{1j}=\mu_{2j}=\ldots=\mu_{Ij} \qquad j=1,\ldots,J$$
$$\sigma_{1j}=\sigma_{2j}=\ldots=\sigma_{Ij},$$

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

$$\hat{\theta}_j(i)=|x_{ij}-x_{\cdot j}|$$

where $x_{\cdot j}=\frac{1}{I}\sum_{i=1}^{I} x_{ij}$, 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 $\hat{\theta}_j(i)$ 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 $\hat{\theta}_j(i)$ do estimate. First, it clearly does not matter whether we rank by using $\hat{\theta}_j(i)$ or $\hat{\theta}^2_j(i)$ so we look at the latter. It turns out that

$$E[\hat{\theta}^2_j(i)]=E[(x_{ij}-x_{\cdot j})^2]$$
$$=\frac{1}{I^2}\left\{(I-1)^2\sigma^2_{ij}+\left[\sum_{i'}(\mu_{ij}-\mu_{i'j})\right]^2+\sum_{i'}\sigma^2_{i'j}\right\}$$

where $\sum_{i'}$ means the sum over $i'\neq i$.

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

$$\mu_{2j}=\mu_{3j}=\ldots=\mu_{Ij}=\mu_j$$
$$\sigma^2_{2j}=\sigma^2_{3j}=\ldots\sigma^2_{Ij}=\sigma^2_j$$

and

$$\mu_{1j}=\mu_j+\alpha_j$$
$$\sigma^2_{1j}=\sigma^2_j+\beta_j$$

[2]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.

then

$$E[(x_{ij}-x_{.j})^2]=\left(\frac{I-1}{I}\right)\sigma_j^2+\left(\frac{I-1}{I}\right)^2(a_j^2+\beta_j)$$

and for $i \neq 1$,

$$E(x_{ij}-x_{.j})^2=\left(\frac{I-1}{I}\right)\sigma_j^2+\frac{1}{I^2}(\alpha_j^2+\beta_j).$$

Thus

$$E[\hat{\theta}_j^2(1)]-E[\hat{\theta}^2{}_j(2)]=\frac{I-2}{I}(\alpha_j^2+\beta_j)$$

Hence, if either the $\beta_j > 0$ and they are large, or if the α_j's are large, then $\hat{\theta}_j(1)$ will tend to be larger than $\hat{\theta}_j(2)$, . . . , $\hat{\theta}_j(I)$ and the first row will receive an extreme rank sum.

TABLE 1.—*Tables for a one-sided rank sum test at nominal significance levels of 1, 3, and 5 percent*

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

No. of objects I	$I=3$						$I=4$					
	0.01		0.03		0.05		0.01		0.03		0.05	
No. of judges J	R	α	R	α	R	α	R	α	R	α	R	α
3											0	0.062
4			0	0.037	0	0.037	0	0.0156	0	0.016	1	.078
5	0	0.0123	0	.012	1	.074	0	.0039	1	.023	2*	.082
6	0	.0041	1	.029	2*	.115	1	.0068	2	.027	3*	.082
7	1	.0110	1	.011	2	.049	2	.0088	3	.029	4*	.079
8	1	.0041	2	.021	3	.072	3	.0101	4	.030	5*	.074
9	2	.0084	3	.032	4*	.095	4	.0108	5	.029	6	.069
10	3	.0140	4	.045	5*	.119	5	.0110	6	.028	7	.063
11	3	.0060	4	.021	5	.059	6	.0110	7	.026	8	.058
12	4	.0094	5	.029	6*	.074	7	.0107	8	.025	9	.053
13	5	.0136	6	.038	7*	.090	8	.0103	9	.023	10	.048
14	5	.0062	6	.018	7	.047	9	.0098	10	.021	11	.043
15	6	.0088	7	.024	8	.057	10	.0092	12	.039	13*	.072

No. of objects I	$I=5$						$I=6$					
	0.01		0.03		0.05		0.01		0.03		0.05	
No. of judges J	R	α	R	α	R	α	R	α	R	α	R	α
3			0	0.040	1*	0.160			0	0.028	1*	0.111
4	0	0.0080	1	.040	2*	.120	0	0.0046	1	.023	2	.069
5	1	.0096	2	.034	3*	.090	1	.0046	3	.043	4*	.097
6	2	.0090	3	.027	4	.067	3	.0108	4	.027	5	.059
7	3	.0077	4	.021	5	.050	4	.0071	6	.037	7*	.072
8	4	.0063	6	.038	7*	.078	6	.0107	7	.023	8	.045
9	6	.0126	7	.028	8	.057	7	.0068	9	.028	10	.051
10	7	.0096	8	.021	9	.042	9	.0089	11	.032	12	.056
11	8	.0073	10	.031	11	.058	11	.0109	13	.036	14	.060
12	10	.0117	11	.023	12	.043	13	.0127	14	.022	15	.038
13	11	.0088	13	.032	14	.056	14	.0080	16	.024	17	.040
14	13	.0128	14	.024	15	.042	16	.0091	18	.026	19	.042
15	14	.0096	16	.031	17	.053	18	.0100	20	.027	21	.043

No. of objects I	$I=7$						$I=8$					
	0.01		0.03		0.05		0.01		0.03		0.05	
No. of judges J	R	α	R	α	R	α	R	α	R	α	R	α
3			0	0.020	1*	0.082	0	0.0156	0	0.016	1	0.062
4	1	0.0146	2	.044	3*	.102	1	.0098	2	.029	3	.068
5	2	.0087	3	.023	4	.052	3	.0137	4	.031	5	.061
6	4	.0125	5	.028	6	.055	4	.0064	6	.028	7	.052
7	5	.0067	7	.029	8	.054	7	.0131	8	.024	9	.043
8	7	.0078	9	.029	10	.052	9	.0116	11	.035	12	.058
9	9	.0083	11	.028	12	.048	11	.0099	13	.029	14	.046
10	11	.0085	13	.026	14	.044	13	.0083	16	.036	17	.056
11	13	.0084	15	.024	16	.040	16	.0114	18	.029	19	.044
12	15	.0080	18	.036	19	.055	18	.0092	21	.034	22	.050
13	17	.0076	20	.032	21	.049	21	.0115	23	.027	25	.056
14	20	.0115	22	.028	23	.043	23	.0092	26	.030	27	.044
15	22	.0105	24	.025	26	.055	26	.0110	29	.034	30	.048

No. of objects I	$I=9$						$I=10$					
	0.01		0.03		0.05		0.01		0.03		0.05	
No. of judges J	R	α	R	α	R	α	R	α	R	α	R	α
3	0	0.0123	0	0.012	1	0.049	0	0.0100	1	0.040	2*	0.100
4	1	.0069	2	.021	3	.048	2	.0150	3	.035	4*	.070
5	3	.0085	5	.038	6*	.070	4	.0126	5	.025	6	.046
6	5	.0078	7	.029	8	.051	6	.0092	8	.030	9	.050
7	8	.0121	10	.036	11	.059	9	.0114	11	.032	12	.050
8	10	.0091	12	.026	13	.042	11	.0076	14	.032	15	.048
9	13	.0114	15	.029	16	.045	14	.0081	17	.030	18	.044
10	15	.0082	18	.031	19	.047	17	.0082	20	.028	22	.058
11	18	.0094	21	.033	22	.047	20	.0081	23	.026	25	.051
12	21	.0103	24	.033	25	.047	24	.0114	27	.033	28	.045
13	24	.0109	27	.033	28	.046	27	.0105	30	.029	32	.054
14	27	.0113	30	.033	31	.045	30	.0096	33	.026	35	.047
15	30	.0115	33	.032	34	.044	33	.0087	37	.031	39	.054

No. of objects I	$I=11$						$I=12$					
	0.01		0.03		0.05		0.01		0.03		0.05	
No. of judges J	R	α	R	α	R	α	R	α_U	R	α_U	R	α_U
3	0	0.0083	1	0.033	2*	0.082	0	0.007	1	0.028	2	0.069
4	2	.0113	3	.026	4	.052	2	.009	3	.020	4	.040
5	4	.0086	6	.032	7	.054	5	.012	6	.022	8	.062
6	7	.0107	9	.031	10	.049	8	.012	10	.032	11	.049
7	10	.0110	12	.028	13	.043	11	.011	13	.026	15	.057
8	13	.0104	15	.025	17	.054	14	.009	17	.030	18	.043
9	16	.0094	19	.031	20	.044	18	.011	21	.032	22	.045
10	19	.0083	22	.026	24	.050	21	.008	25	.034	26	.045
11	23	.0105	26	.030	28	.055	25	.009	29	.033	30	.045
12	26	.0088	30	.033	31	.044	29	.010	33	.033	35	.057
13	30	.0102	33	.026	35	.047	33	.010	37	.032	39	.054
14	34	.0115	37	.028	39	.048	37	.010	41	.030	43	.051
15	37	.0093	41	.030	43	.050	41	.010	45	.029	47	.047

No. of objects I	$I=13$						$I=14$					
	0.01		0.03		0.05		0.01		0.03		0.05	
No. of judges J	R	α_U	R	α_U	R	α_U	R	α_U	R	α_U	R	α_U
3	0	0.006	1	0.023	2	0.059	0	0.005	1	0.020	2	0.051
4	2	.007	4	.032	5	.057	3	.013	4	.025	5	.046
5	5	.009	7	.028	8	.045	6	.012	8	.033	9	.052
6	8	.008	11	.033	12	.050	9	.009	12	.034	13	.050
7	12	.010	15	.035	16	.050	13	.010	16	.032	17	.046
8	16	.011	19	.035	20	.048	17	.010	20	.029	22	.054
9	19	.008	23	.033	24	.045	21	.009	25	.034	26	.046
10	23	.008	27	.031	29	.055	25	.008	29	.029	31	.050
11	28	.011	31	.028	33	.049	30	.010	34	.032	36	.053
12	32	.011	36	.033	38	.055	34	.009	39	.033	41	.054
13	36	.010	40	.029	42	.048	39	.010	43	.027	46	.055
14	40	.009	45	.033	47	.052	44	.011	48	.028	51	.054
15	45	.010	49	.029	51	.045	48	.009	53	.028	55	.044

See footnote at end of table.

TABLE 1.—*Tables for a one-sided rank sum test at nominal significance levels of 1, 3, and 5 percent*—Continued

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

No. of objects I	$I=15$					
	0.01		0.03		0.05	
No. of judges J	R	α_U	R	α_U	R	α_U
3	0	0.004	1	0.018	2	0.044
4	3	.010	5	.037	6	.062
5	6	.009	8	.025	10	.059
6	10	.010	12	.024	14	.051
7	14	.010	17	.030	19	.057
8	18	.009	22	.034	23	.045
9	23	.011	26	.027	28	.047
10	27	.009	31	.028	33	.046
11	32	.009	36	.027	38	.045
12	37	.010	41	.027	44	.053
13	42	.010	47	.032	49	.050
14	47	.010	52	.030	54	.046
15	52	.009	57	.028	59	.043

*In order to avoid duplication in the tables these values of R do not give α'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 ith object, and let

$$r_{max} = \max_{1 \leq i \leq I} r_i,$$

$$r_{min} = \min_{1 \leq i \leq I} r_i$$

Let R be an integer and α be a probability such that

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

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, α, 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 α.

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 $r_{max} \geq 48-4=44$. 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 $r_{min} \leq 6+4=10$.

From the same line of the table we see that a critical region of $r_{max} \geq 42$ has a significance level of $\alpha=0.028$, and a critical region of $r_{max} \geq 41$ 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 α'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 α for any given R cannot be computed easily, but bounds, α_U and α_L, such that $\alpha_L \leq \alpha \leq \alpha_U$ can be. Using eqs (2), (3), and (6) of [2] α_U and α_L were computed for $I \leq 11$. For the 0.01 (0.03) level, α_U and α_L always agree to four (three) decimal places, hence the values in the table for α are exact to the four (three) decimal places given.

For the 0.05 level α_U and α_L did not always agree when rounded to three places although

$$\alpha_U - \alpha_L < 0.0005$$

all the time. The rounded value for α_L was chosen since α_L is a better approximation to α.

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

4. References

[1] W. J. Youden, Ranking laboratories by round-robin tests. Materials Research and Standards **3,** 9–13 (1962).

[2] W. A. Thompson, Jr. and T. A. Willke, On an extreme rank sum test for outliers. Biometrika **50,** 375–383 (1963).

[3] T. W. Lashof, (1964) Ranking laboratories in round-robin tests. Materials Research and Standards (in press).

(Paper 68B2–116)