

Statistical Significance of Ranking Paradoxes

Anna E. Bargagliotti and Raymond N. Greenwell

University of Memphis and Hofstra University

Goal: To investigate Simpson-like paradoxes described by Haunsperger (2003), in which the individual parts give rise to a common decision, but the aggregate of those parts give rise to a different decision. Haunsperger's paradoxes use the Kruskal-Wallis method of summing ranks. In previous research we investigated the significance of the Kruskal-Wallis statistic for the differences in ranking when these paradoxes occur. Bargagliotti (2008) has shown that these paradoxes also arise using the Mann-Whitney and Bhapkar's V test, so we now investigate significance of these paradoxes.

Example (Haunsperger):

<u>C_1</u>	<u>C_2</u>	<u>C_3</u>
5.89	5.81	5.80
5.98	5.90	5.99

<u>C_1</u>	<u>C_2</u>	<u>C_3</u>
5.69	5.63	5.62
5.74	5.71	6.00

<u>C_1</u>	<u>C_2</u>	<u>C_3</u>
3	2	1
5	4	6
8	6	7

$C_1 \succ C_3 \succ C_2$ by Kruskal-Wallis ranking.

<u>C_1</u>	<u>C_2</u>	<u>C_3</u>
5.89	5.81	5.80
5.98	5.90	5.99
5.69	5.63	5.62
5.74	5.71	6.00

<u>C_1</u>	<u>C_2</u>	<u>C_3</u>
8	7	6
10	9	11
3	2	1
<u>5</u>	<u>4</u>	<u>12</u>
26	22	30

$$C_3 \succ C_1 \succ C_2$$

The 2×3 matrix of ranks from before, along with the statistical procedure used, is not *consistent under replication*.

$$KW = \frac{12}{N(N+1)} \sum_{k=1}^m n_k (\bar{r}_k - \bar{r})^2,$$

where N is the number of data, m is the number of columns, n_k is the number of data in column k , \bar{r}_k is the average rank of the data in column k , and \bar{r} is the average rank of the data. Here $n_k = n$ and $N = mn$.

An approximation when N is not too small is

$$p = P \left(\chi_{m-1}^2 \geq KS \right).$$

In the previous example, $KW = 0.286$, $p = 0.867$ for original matrix of ranks, and $KW = 0.615$, $p = 0.735$ for combined matrix of ranks.

The Mann-Whitney procedure considers pairs of alternatives and ranks the data. For example, comparing C_1 and C_2

$\underline{C_1}$	$\underline{C_2}$
5.89	5.81
5.98	5.90

the Mann-Whitney procedure ranks these data and obtains:

$\underline{C_1}$	$\underline{C_2}$
2	1
4	3

Unlike the Kruskal-Wallis procedure, the Mann-Whitney procedure analyzes the ranks by tallying the number of times an observation for C_1 is larger than an observation for C_2 and vice versa. Because there are 2 observations per alternative, there are 2×2 possible tuples to consider. They are (2,1), (2,3), (4,1), and (4,3). From these comparisons, we see that C_1 beats C_2 three times leaving C_2 to beat C_1 only one time.

Repeating this process for all pairwise comparisons, C_1 and C_3 each beat each other two times and C_2 and C_3 each beat each other two times. This leads to a non-strict cyclical overall ranking of the alternatives, i.e. $C_1 \succ C_2$, $C_2 \sim C_3$, $C_1 \sim C_3$.

When considering the whole data set and comparing C_1 to C_2 , the observations are ranked as follows:

C_1	C_2
6	5
8	7
2	1
4	3

Analyzing these data, the Mann-Whitney procedure has C_1 being greater than C_2 ten times and C_2 being greater than C_1 six times. Repeating this procedure for all pairwise comparisons, one obtains the overall ranking $C_3 \succ C_1 \succ C_2$. Again, as with the Kruskal-Wallis procedure, the paradox (procedure results from the parts not matching the procedure results from the whole) persists using the Mann-Whitney procedure.

In addition to illustrating the Simpson-like paradox, Example 1 shows how two different procedures, Kruskal-Wallis and Mann-Whitney, may lead to different overall rankings. This type of inconsistency is due to symmetric structures in the data sets. These structures and inconsistencies have been completely characterized in Bargagliotti and Saari (2008) by building on ideas in Haunsperger (1992).

Using the V test to analyze the ranks in Example 1, the procedure considers all possible 3-tuples (a_i, b_j, c_k) where a_i is an observation for C_1 , b_j is an observation for C_2 , and c_k is an observation for C_3 , and counts the number of 3-tuples for which each alternative has the largest entry. For the 2×3 data set in Example 1, alternative C_1 has the largest entry in 3 of the 8 possible 3-tuples, C_2 has the largest entry in one, and C_3 has the largest entry in 4. Therefore, for this data matrix, the V procedure outputs $C_3 \succ C_1 \succ C_2$ as the overall ranking of the alternatives.

For the full ranked data matrix in the example, there are a total of 4^3 possible 3-tuples. Alternative C_1 has the largest entry in 17, C_2 has the largest entry in 11, and C_3 has the largest entry in 36. This analysis of rank procedure thus outputs $C_3 \succ C_1 \succ C_2$ overall ranking of the alternatives for this matrix. The V procedure does not lead to the same inconsistencies for these particular data matrices as do the Kruskal-Wallis and Mann-Whitney procedures. As shown in Bargagliotti (2008), however, examples do exist that cause these same paradoxes to occur using the V test.

The Mann-Whitney statistic is

$$U = \min(U_{C_i}, U_{C_j})$$

where U_{C_i} = the number of times an entry of C_i beats an entry of C_j , and U_{C_j} = the number of times an entry of C_j beats an entry of C_i . For N large enough, this statistic is normally distributed with $\mu = n^2/2$ and $\sigma^2 = n^2(2n + 1)/12$ where n is the number of observations per alternative. We thus compute $Z = (U - \mu)/\sigma$.

In the case of only two observations per alternative, the p -value is 0.667. In the case of the aggregated data matrix with four observations per alternative, the p -value is directly tabulated in Mann and Whitney (1947) as 0.343, which doubles for the two-tailed test to 0.686. In both cases, it is not statistically significant.

The V test statistic is

$$V = N(2m-1) \left[\sum_{j=1}^m p_j \left(u_j - \frac{1}{m} \right)^2 - \left(\sum_{j=1}^m p_j \left(u_j - \frac{1}{m} \right) \right)^2 \right]$$

where m = number of alternatives, N = number of total observations, p_j = (number of observations for alternative j)/ N , v_j = number of m -tuples j wins, and $u_j = v_j$ /(number of m -tuples).

As with the Kruskal-Wallis test, an approximation when N is not too small is

$$p = P \left(\chi_{m-1}^2 \geq V \right).$$

In Example 1, with $u_1 = 3/8$, $u_2 = 1/8$, and $u_3 = 4/8$ for the original matrix of ranks, this leads to $V = 0.729$ and $p = 0.695$, and $V = 1.52$, $p = 0.467$ for the combined matrix of ranks. Therefore, just as with the Kruskal-Wallis and Mann-Whitney test, the alternatives are not considered statistically different.

A matrix of ranks is called *row-ordered* if the observations of each candidate can be put in an order so that every row of the matrix gives rise to the same ranking of the candidates.

Theorem 1. (Haunsperger) An $n \times m$ matrix of ranks is consistent under replication if and only if it can be row-ordered.

What is the lowest statistical significance associated with a row-ordered matrix?

Consider the matrix of ranks

$$\begin{array}{ccccc}
 1 & 2 & 3 & \dots & m \\
 m + 1 & m + 2 & m + 3 & \dots & 2m \\
 \dots & \dots & \dots & \dots & \dots \\
 (n - 1)m + 1 & (n - 1)m + 2 & (n - 1)m + 3 & \dots & nm
 \end{array}$$

$$KW = \frac{m^2 - 1}{mn + 1}$$

$$V = nm(2m-1) \left[\sum_{j=1}^m \frac{1}{m} \left(u_j - \frac{1}{m} \right)^2 - \left(\sum_{j=1}^m \frac{1}{m} \left(u_j - \frac{1}{m} \right) \right)^2 \right]$$

where $u_j = \sum_{i=1}^n i^{j-1} (i - 1)^{m-j} / n^m$.

Kruskal-Wallis p -values

$m \backslash n$	1	2	3	4
2	0.317	0.439	0.513	0.564
3	0.368	0.565	0.670	0.735
4	0.392	0.644	0.764	0.830
5	0.406	0.702	0.827	0.887

V test p -values

$m \backslash n$	1	2	3	4
2	0.221	0.386	0.823	0.540
3	0.189	0.508	0.649	0.727
4	0.154	0.556	0.721	0.804
5	0.126	0.562	0.750	0.840

When $n = 1$, the V statistic can only take on one value, and so the statistical significance of that value is meaningless.

Considering the matrix of ranks above, a pairwise comparison of C_i and C_j for $i \neq j$ can also be made. Letting $i < j$ and reordering the entries for C_i and C_j , we obtain the following ranked data:

$$\begin{array}{cc} \underline{C_i} & \underline{C_j} \\ 1 & 2 \\ 3 & 4 \\ \dots & \dots \\ 2n - 1 & 2n \end{array}$$

The Mann-Whitney test statistic for these data is

$$U = \min(U_{C_i}, U_{C_j}) = n(n - 1)/2,$$

where U_{C_i} = the number of times an entry of C_i beats an entry of C_j , and U_{C_j} = the number of times an entry of C_j beats an entry of C_i .

This statistic is normally distributed with $\mu = n^2/2$ and $\sigma = n^2(2n + 1)/12$ and thus converting to the standard normal distribution $Z = -\sqrt{3/(2n + 1)}$. This Z value has a maximum magnitude of -0.577 when $n = 1$, and thus will never be less than -1.96 , the critical value for 0.05 significance level using the normal distribution. It therefore will never be statistically significant.

Theorem 2. For all m and for all n , a set of row-ordered data exists that leads the Kruskal-Wallis, V , and Mann-Whitney test to conclude there is not enough evidence to show the observations were sampled from different distributions

Theorem 4. (Haunsperger) For any $n \geq 1$ and $m \geq 2$, let r_0 be the $n \times m$ matrix of ranks

$$\begin{array}{cccccc}
 1 & 2 & 3 & \dots & m \\
 m + 1 & m + 2 & m + 3 & \dots & 2m \\
 \dots & \dots & \dots & \dots & \dots \\
 (n - 1)m + 1 & (n - 1)m + 2 & (n - 1)m + 3 & \dots & nm
 \end{array}$$

Let r be the matrix of ranks made from r_0 by switching 2 adjacent entries x_{ij} and $x_{i(j+1)}$ for some $1 \leq i \leq n$, $1 \leq j \leq m - 1$. Only two data sets with matrix of ranks r are needed to have their aggregate ranking other than

$$C_m \succ C_{m-1} \succ \dots \succ C_1.$$

Example:

	$\underline{C_1}$	$\underline{C_2}$	$\underline{C_3}$
$r_0 =$	1	2	3
	4	5	6
	7	8	9
	<u>10</u>	<u>11</u>	<u>12</u>
	22	26	30

	$\underline{C_1}$	$\underline{C_2}$	$\underline{C_3}$
$r =$	2	1	3
	4	5	6
	7	8	9
	<u>10</u>	<u>11</u>	<u>12</u>
	23	25	30

$C_3 \succ C_2 \succ C_1$ by Kruskal-Wallis ranking in either case.

$\underline{C_1}$	$\underline{C_2}$	$\underline{C_3}$
14	1	15
16	17	18
19	20	21
22	23	24
3	2	4
5	6	7
8	9	10
$\underline{11}$	$\underline{12}$	$\underline{13}$
98	90	112

$$C_3 \succ C_1 \succ C_2$$

$KW = 0.615$ with $p = 0.735$ for matrix r_0 ,
 $KW = 0.500$ with $p = 0.779$ for matrix r , and
 $KW = 0.620$ with $p = 0.733$ for the combined matrix of ranks.

Using the V test, $V = 0.638$ with $p = 0.727$ for r_0 ,
 $V = 0.638$ with $p = 0.727$ for r , and
 $V = 0.430$ with $p = 0.807$ for the combined matrix.

Using the Mann-Whitney test, for any pair (C_i, C_j) where $i < j$ in matrix r_0 , $Z = -0.577$ with $p = 0.282$.

In the matrix r , comparing C_1 with C_2 gives $Z = -0.289$ and $p = 0.386$.

Comparing C_1 or C_2 with C_3 gives $Z = -0.577$ with $p = 0.282$.

For the combined matrix,

comparing (C_1, C_2) gives $Z = -0.210$ and $p = 0.417$,

comparing (C_1, C_3) gives $Z = -0.420$ and $p = 0.337$, and

comparing (C_2, C_3) gives $Z = -0.840$ and $p = 0.200$.

Consider the matrix of ranks

$$\begin{array}{ccccc}
 2 & 1 & 3 & \dots & m \\
 m + 1 & m + 2 & m + 3 & \dots & 2m \\
 \dots & \dots & \dots & \dots & \dots \\
 (n - 1)m + 1 & (n - 1)m + 2 & (n - 1)m + 3 & \dots & nm
 \end{array}$$

$$KW = \frac{m^2 - 1}{nm + 1} + \frac{24(1 - n)}{mn^2(nm + 1)}$$

Kruskal-Wallis p values

$m \backslash n$	1	2	3	4
2	0.317	1.000	0.827	0.773
3	0.368	0.651	0.733	0.779
4	0.392	0.682	0.789	0.846
5	0.406	0.722	0.838	0.894

The V test statistic for this matrix is identical to that of the row-ordered matrix considered earlier.

When comparing alternative C_1 pairwise with any other alternative C_j using the generalized matrix of ranks, the following re-rankings of these data are made:

$\underline{C_1}$	$\underline{C_j}$
2	1
3	4
...	...
$2n - 1$	$2n$

For this generalized matrix, $Z = ((-n + 2)/n)\sqrt{3/(2n + 1)}$.

Mann-Whitney outcomes

n	Z	p -value
2	0.000	0.500
3	-0.218	0.414
4	-0.289	0.387
5	-0.313	0.377
6	-0.320	0.374
7	-0.319	0.375

As n approaches infinity, Z approaches 0 and thus the p -value approaches 0.5. All other pairwise comparisons are equivalent to those made from the row-ordered generalized matrix considered previously.

Consider the matrix of ranks

$$\begin{array}{cccc}
 mn + 2 & 1 & \dots & mn + m \\
 mn + m + 1 & mn + m + 2 & \dots & mn + 2m \\
 \dots & \dots & \dots & \dots \\
 mn + (n - 1)m + 1 & mn + (n - 1)m + 2 & \dots & 2nm \\
 3 & 2 & \dots & m + 1 \\
 m + 2 & m + 3 & \dots & 2m + 1 \\
 (n - 1)m + 2 & (n - 1)m + 3 & \dots & nm + 1
 \end{array}$$

$$KW = \frac{1}{m(2mn + 1)} \left(m^3 + 9m^2 - 22m + \frac{12m}{n} - \frac{24}{n} + \frac{24}{n^2} \right)$$

p values

$m \setminus n$	1	2	3	4
2	0.121	0.564	0.749	0.834
3	0.156	0.500	0.653	0.733
4	0.198	0.566	0.728	0.810
5	0.244	0.642	0.801	0.875

$n = 1$ case makes no sense.

No paradox with $n = 2$.

Among $n \geq 3$, the smallest p value is 0.653.

Using the V statistic to test for differences among alternatives, we obtain the following complicated formula:

$$V = nm(2m - 1) \left[\frac{1}{m}F(n, m)^2 + \frac{1}{m}G(n, m)^2 + \frac{1}{m} \sum_{j=3}^m M(n, m)^2 - \left(\frac{1}{m}F(n, m) + \frac{1}{m}G(n, m) + \frac{1}{m} \sum_{j=3}^m M(n, m) \right)^2 \right].$$

In the above equation, the expression for C_1 is

$$F(n, m) = \frac{\sum_{i=1}^n (i+1)i^{m-2} + \sum_{i=n+1}^{2n-1} i^{m-1}}{(2n)^m} - \frac{1}{m},$$

the expression for alternative C_2 is

$$G(n, m) = \frac{\sum_{i=1}^{2n-1} (i+1)i^{m-2} - (n+1)n^{m-2}}{(2n)^m} - \frac{1}{m},$$

and the expression for C_j where $j > 2$ is

$$M(n, m) = \frac{\sum_{i=1}^n i^{j-2}(i+1)(i-1)^{m-j} + \sum_{i=n+1}^{2n} i^{j-1}(i-1)^{m-j}}{(2n)^m} - \frac{1}{m}.$$

V test p -values

$m \backslash n$	1	2	3	4
2	0.540	0.829	0.906	0.939
3	0.482	0.775	0.859	0.898
4	0.675	0.889	0.940	0.961
5	0.775	0.938	0.971	0.984

Three different pairwise comparisons arise from the general form of the combined matrix of ranks. The comparisons are (C_1, C_2) , (C_2, C_j) where $j \geq 3$, (C_i, C_j) where $i \neq 2$ and $j > 2$. The comparisons result in the following pairwise re-ranked data matrices:

$\frac{C_1}{3}$	$\frac{C_2}{1}$	$\frac{C_2}{1}$	$\frac{C_j}{3}$	$\frac{C_i}{1}$	$\frac{C_j}{2}$
4	2	2	5	3	4
6	5	6	7	5	6
...	...	8	9	7	8
$2n$	$2n - 1$
$2n + 2$	$2n + 1$	$n - 2$	$n + 1$
$2n + 3$	$2n + 4$	n	$n + 2$
$2n + 5$	$2n + 6$	$n + 3$	$n + 4$
...
$4n - 3$	$4n - 2$	$2n - 3$	$2n - 2$	$2n - 3$	$2n - 2$
$4n - 1$	$4n$	$2n - 1$	$2n$	$2n - 1$	$2n$

The (C_1, C_2) comparison yields $Z = -(2/n)\sqrt{3/(4n+1)}$,
 (C_2, C_j) where $j \geq 3$ has $Z = -2\sqrt{3/(2n+1)}$,
 (C_i, C_j) where $i \neq 2$ and $j > 2$ gives $Z = -\sqrt{3/(2n+1)}$.

Mann-Whitney Test Statistics and p -values

n	Z_{C_1, C_2}	p -value	n	Z_{C_2, C_j}	p -value	n	Z_{C_i, C_j}	p -value
2	-0.577	0.282	2	-1.155	0.124	2	-0.577	0.282
3	-0.320	0.374	3	-0.961	0.168	3	-0.480	0.315
4	-0.210	0.417	4	-0.840	0.200	4	-0.420	0.337
5	-0.151	0.440	5	-0.756	0.225	5	-0.378	0.352

The p -values for (C_2, C_j) where $j \geq 3$ are approaching significance, but the difference between these candidates was not in dispute. In Example 2, with $m = 3$ and $n = 4$, we found that $C_3 \succ C_2$ in r_0 , r , and the aggregate matrix.

Conclusion:

In the cases in which the paradoxes arise, the difference between the ranking of the candidates is not statistically significant.

Contrapositive:

When there is a statistically significant difference between the ranking of the candidates, these paradoxes do not occur (at least for the constructions in Haunsperger's 2003 article).