Statistical Significance of Ranking Paradoxes

Anna E. Bargagliotti and Raymond N. Greenwell[1]

February 28, 2009

[1]Anna E. Bargagliotti is an Assistant Professor in the Department of Mathematical Sciences at the University of Memphis, Memphis, TN 38152 (e-mail: abargag@yahoo.com); and Raymond N. Greenwell is a Professor in the Department of Mathematics at Hofstra University, Hempstead, NY 11549-1030 (matrng@hofstra.edu).

Abstract

When nonparametric statistical tests are used to rank-order a list of alternatives, Simpson-like paradoxes arise, in which the individual parts give rise to a common decision, but the aggregate of those parts gives rise to a different decision. Haunsperger (Haunsperger 2003) and Bargagliotti (Bargagliotti, submitted manuscript, 2008) have shown that the Kruskal-Wallis (Kruskal and Wallis 1952), Mann-Whitney (Mann and Whitney 1947), and Bhapkar's V (Bhapkar 1961) nonparametric statistical tests are subject to these types of paradoxes. We further investigate these ranking paradoxes by showing that when they occur, the differences in ranking are not statistically significant.

Keywords Kruskal-Wallis, nonparametric, ranking, Mann-Whitney, statistical test

1 Introduction

Nonparametric statistical test based on ranks can be used to test for differences among alternatives (Lehman 1975). Each test, defined by a test statistic, utilizes a unique nonparametric procedure that analyzes the ranked data and provides an overall ranking of the alternatives. For example, the Kruskal-Wallis test (Kruskal and Wallis 1952) defines a test statistic in terms of the rank-sums for each alternative. This rank-sum procedure yields a rank-ordering of the alternatives (i.e. the alternative with the largest rank sum is ranked first, the alternative with the second largest ranked sum is ranked second, etc.). Depending on which procedure is used to analyze the ranks, different rank-orderings of the alternatives may occur (Bargagliotti and Saari, submitted manuscript, 2008). Particularly interesting types of inconsistencies are Simpson-like paradoxes, in which the individual data sets give rise to one overall ranking, but the aggregate of the data sets gives rise to a different ranking (Haunsperger and Saari 1991; Haunsperger 1992; Haunsperger 1996; Haunsperger 2003; Bargagliotti, submitted manuscript, 2008). Haunsperger (2003) has shown this paradox exists when the Kruskal-Wallis nonparametric statistical procedure on n samples is used to rank-order a list of alternatives. Bargagliotti (submitted manuscript, 2008) has shown that these paradoxes also occur when the Mann-Whitney (Mann and Whitney 1947) and Bhapkar's V (Bhapkar 1961) procedures are used.

In this paper, we further investigate these ranking paradoxes by exploring the nonparametric statistical test outcomes and determining whether these paradoxes persist at the test level. Focusing on the Kruskal-Wallis test, the Mann-Whitney test, and Bhapkar's V test, we determine whether the test statistics are statistically significant for data sets whose structures lead to the Simpson-like paradoxes explored in the literature.

Section 2 of the paper illustrates paradoxes occurring at the procedure level in the case where the two identical ranked data sets are aggregated and analyzed using the Kruskal-Wallis, Mann-Whitney, and V procedures. Section 3 of the paper explores the statistical significance of the three tests for the data structures introduced in Section 2. Section 4 explores a necessary condition data structures must meet in order to not be subject to the ranking paradoxes of Section 2. Section 5 explores a sufficient condition for generating such a paradox. In each of these sections, we compute the statistical significance for each test statistic for the general forms of data that meet these conditions. Section 6, the last section of the paper, discusses the implications of these results.

2 Replication and Ranking Paradox

It may be the case that data is collected on two or more occasions. In this setting, one could analyze the data separately or consider incorporating it into one data set before doing the analyses. The Simpson-like paradox described in Bargagliotti (submitted manuscript, 2008) and Haunsperger (2003) occurs when the analysis of each of the separate data sets leads to one rank-ordering of the alternatives but the analysis of the complete data leads to another rank-ordering. Of course this paradox depends on the procedure used to analyze each of the data sets. The following simple example is directly taken from Haunsperger (2003) to illustrate the paradox occurring with the Kruskal-Wallis procedure.

Example 1. Consider the following two sets of data, which contain favorability scores for three candidates C_1, C_2, and C_3:

C_1	C_2	C_3
5.89	5.81	5.80
5.98	5.90	5.99

and

C_1	C_2	C_3
5.69	5.63	5.62
5.74	5.71	6.00

By replacing the lowest entry with the number 1, the second lowest entry with the number 2, etc., the favorability scores are transformed into rankings. Each of the two sets give rise to the exact same matrix of ranks:

C_1	C_2	C_3
3	2	1
5	4	6

The Kruskal-Wallis procedure computes the rank-sums for each alternative C_1, C_2, and C_3 to be 8, 6, and 7 respectively. Using these data sets, this means that the intrinsic ranking produced by the Kruskal-Wallis procedure is $C_1 \succ C_3 \succ C_2$. When the two raw data sets are combined as

C_1	C_2	C_3
5.89	5.81	5.80
5.98	5.90	5.99
5.69	5.63	5.62
5.74	5.71	6.00

the corresponding re-ranking is

C_1	C_2	C_3
8	7	6
10	9	11
3	2	1
5	4	12

The rank-sums are then 26, 22, and 30 respectively. With these data, the Kruskal-Wallis procedure yields the overall ranking of $C_3 \succ C_1 \succ C_2$.

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

C_1	C_2
5.89	5.81
5.98	5.90

the Mann-Whitney procedure ranks these data and obtains:

C_1	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:

$\underline{C_1}$	$\underline{C_2}$
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 (submitted manuscript, 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 (submitted manuscript, 2008), however, examples do exist that cause these same paradoxes to occur using the V test.

3 Statistical Significance

In cases such as Example 1, Haunsperger (2003) defines the 2×3 matrix of ranks from above as not *consistent under replication* (i.e., one does not obtain the same overall ranking when replicating and aggregating the data set).

The main purpose of the nonparametric tests is to test whether the observations for each alternative are sampled from the same population or whether the populations differ (Mann and Whitney 1947; Kruskal and Wallis 1952; Bhapkar 1961). In this section, we take the ranking analysis one step further and explore the test outcomes for the data parts and the whole to see whether the Simpson-like paradox persists. To illustrate, we compute the Kruskal-Wallis, Mann-Whitney, and V test statistics for Example 1 above.

The Kruskal-Wallis test statistic is

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

where N is the number of data, m is the number of columns, n_k is the number of data in column k, $\overline{r}_k$ is the average rank of the data in column k, and $\overline{r}$ is the average rank of the data. For simplicity, we will restrict ourselves to the case in which $n_k = n$ does not vary with k, so that $N = mn$. These are the only cases that are considered in the ranking paradox literature.

An approximation when N is not too small is

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

In Example 1, this leads to $KW = 0.286$, $p = 0.867$ for the original matrix of ranks, and $KW = 0.615$, $p = 0.735$ for the combined matrix of ranks. For N small, however, the chi-square approximation is not highly accurate. The values of p for the exact distribution, using the tables in Alexander and Quade (1968), are 0.917 and 0.770. These values do not change our conclusion, nor do they do so elsewhere in this paper, so for simplicity, we will continue to use the chi-square approximation. In other words, there is no statistically significant difference in rankings between the three choices and thus the paradox does not persist beyond the rankings.

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 μ=$n^2/2$ and $\sigma^2 = n^2(2n+1)/12$ where n is the number of observations per alternative. We thus compute $Z = (U-\mu)/\sigma$.

For the data in Example 1, we compare only C_1 and C_2. The number of tuples that C_1 and C_2 win are 3 and 1 respectively. The other two possible pairwise comparisons have equally matched number of tuples won for each alternative. In the full ranked data case, every pairwise comparison has one of the two alternatives winning 10 tuples and the other winning 6. Without loss of generality, we thus only compute the C_1 versus C_2 pair of alternatives. This leads to $Z = -0.775$ with $p = 0.439$ for two-sided test on the original matrix of ranks, and $Z = -0.577$ with $p = 0.564$ for the combined matrix of ranks. Because the number of observations is small ($n^2 < 20$) in both data sets, the normal approximation may not be accurate. For small cases, the p-value can be found by directly computing the distribution of U. 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 hypothesis of uniformity fails to be rejected for both the part and the whole data set.

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

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

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.

By calculating the test statistics for the Kruskal-Wallis, Bhapkar's V, and Mann-Whitney tests for general data forms, the subsequent sections show that although certain data structures are subject to paradoxes at the ranking level, they are not subject to the same paradoxes at the test significance level. If the paradox persisted, this would mean that the parts data sets would lead to one conclusion about the null hypothesis, but when aggregated together, the statistical conclusion would differ.

4 Conditions for Consistent Outcomes

Haunsperger proves many surprising results about matrices of ranks. In Haunsperger (2003) she defines a matrix to be *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 m candidates. Haunsperger observes, "Indeed, being row-ordered is such a strong property that the reader would be justified in complaining that it is unrealistic. Not only is this claim correct, but the reader might be even more shocked to learn that the *only* matrices of ranks that are consistent under replication are row-ordered" (Haunsperger, 2003, p. 265). This is demonstrated by the following theorem.

Theorem 1 *(Haunsperger 2003) An $n \times m$ matrix of ranks is consistent under replication if and only if it can be row-ordered.*

We extend this analysis by investigating how low the statistical significance associated with a row-ordered matrix might be. High significance (that is, small p-values) can be achieved by putting the lowest ranks in the first column, the next highest in the next column, and so forth. To achieve low significance, consider the matrix of ranks

$\underline{C_1}$	$\underline{C_2}$	$\underline{C_3}$	$\ldots$	$\underline{C_m}$
1	2	3	$\ldots$	m
$m+1$	$m+2$	$m+3$	$\ldots$	$2m$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$(n-1)m+1$	$(n-1)m+2$	$(n-1)m+3$	$\ldots$	nm

For data of this form, it is straightforward to show that

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

and

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

where $u_j = \sum_{i=1}^{n} i^{j-1}(i-1)^{m-j}/n^m$. When $j = m$ and $i = 1$, this formula contains the expression 0^0. In this formula and elsewhere in this paper, 0^0 is defined to equal 1.

Using the chi-square approximation, we calculate the Kruskal-Wallis and V test statistics and their associated p-values for various values of m and n, as displayed in Tables 1 and 2. Notice that none of the values in the tables approach statistical significance for either test statistic, except for values when $n = 1$ in Table 2. But for these values, the chi-square approximation is particularly poor. In fact, when $n = 1$, the V statistic can only take on one value, and so the statistical significance of that value is meaningless.

Table 1: 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

Table 2: 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

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:

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

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.

Summarizing these ideas leads us to the following general theorem.

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.*

This illustrates that although a matrix being row-ordered may seem unrealistically strict, in fact it is always possible to find row ordered data that yields no significant difference in rankings between the candidates. From that perspective, being row-ordered is not strict at all. For there to be a significant difference between the candidates, even stricter requirements on the matrix are needed.

5 Statistical Significance of Replicated Data

As a consequence of Theorem 1, a matrix with unanimous rankings across all the rows but one is not consistent under replication. One might suspect that for a matrix that is almost unanimous in ranking, a large number of replications would be needed to make the matrix inconsistent. Haunsperger dispels this suspicion for the Kruskal-Wallis procedure with the following theorem.

Theorem 3 *For any $n \geq 1$ and $m \geq 2$, let r_0 be the $n \times m$ matrix of ranks*

$\underline{C_1}$	$\underline{C_2}$	$\underline{C_3}$	$\ldots$	$\underline{C_m}$
1	2	3	$\ldots$	m
$m+1$	$m+2$	$m+3$	$\ldots$	$2m$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$(n-1)m+1$	$(n-1)m+2$	$(n-1)m+3$	$\ldots$	nm

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

Example 2. The proof of the theorem is by construction similar to the following:

$$r_0 = \begin{array}{ccc} \underline{C_1} & \underline{C_2} & \underline{C_3} \\ 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ \underline{10} & \underline{11} & \underline{12} \\ 22 & 26 & 30 \end{array}$$

$$r = \begin{array}{ccc} \underline{C_1} & \underline{C_2} & \underline{C_3} \\ 2 & 1 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ \underline{10} & \underline{11} & \underline{12} \\ 23 & 25 & 30 \end{array}$$

Notice that $C_3 \succ C_2 \succ C_1$ by Kruskal-Wallis ranking in either case. Now consider the following aggregate of two matrices, each with the ranking r.

$\underline{C_1}$	$\underline{C_2}$	$\underline{C_3}$
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

Observe that the ranking has now changed to $C_3 \succ C_1 \succ C_2$.

The V test and the Mann-Whitney procedure also exhibit the same sensitivity to non row-ordered data sets. For the data matrices r and r_0 in Example 2, the V test and the pairwise comparison yields the ranking $C_3 \succ C_2 \succ C_1$ while both procedures output $C_3 \succ C_1 \succ C_2$ for the combined matrix.

Exploring the significance for these data matrices, $KW = 0.615$ 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 $p = 0.200$.

These results show there is no statistically significant difference between the three candidates in either the original matrix, the matrix with two ranks switched, or the combined matrix of ranks in Example 2 using the Kruskal-Wallis test, V test, or the Mann-Whitney test.

To generalize the previous example, consider the matrix of ranks

$\underline{C_1}$	$\underline{C_2}$	$\underline{C_3}$	$\ldots$	$\underline{C_m}$
2	1	3	$\ldots$	m
$m+1$	$m+2$	$m+3$	$\ldots$	$2m$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$(n-1)m+1$	$(n-1)m+2$	$(n-1)m+3$	$\ldots$	nm

For this matrix,

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

Using the chi-square approximation, we calculated p-values for various values of m and n, as displayed in Table 3.

Table 3: 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 in the previous section, i.e.

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

where $u_j = \sum_{i=1}^{n} i^{j-1}(i-1)^{m-j}/n^m$. This is because interchanging the first row entries for candidates 1 and 2 does not affect the number of m-tuples either candidate wins.

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, $Z = ((-n+2)/n)\sqrt{3/(2n+1)}$. Using the normal distribution, the p-values are displayed for various values of n in Table 4. 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 at the beginning of Sec. 4. Therefore, as before, none of the values approach statistical significance for the Kruskal-Wallis, V test, and the Mann-Whitney.

Table 4: 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

Now consider the combined matrix of ranks

$\underline{C_1}$	$\underline{C_2}$	$\ldots$	$\underline{C_m}$
$mn+2$	1	$\ldots$	$mn+m$
$mn+m+1$	$mn+m+2$	$\ldots$	$mn+2m$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$mn+(n-1)m+1$	$mn+(n-1)m+2$	$\ldots$	$2nm$
3	2	$\ldots$	$m+1$
$m+2$	$m+3$	$\ldots$	$2m+1$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$(n-1)m+2$	$(n-1)m+3$	$\ldots$	$nm+1$

For this matrix,

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

Using the chi-square approximation, we calculated p-values for various values of m and n, as displayed in Table 5.

Table 5: Kruskal-Wallis p-values

$m\backslash n$	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

Although the theorem includes the $n = 1$ case, this construction makes no sense when $n = 1$. When $n = 2$, there is no paradox, because one row has the ranks in one order, and the other row in a different order. Among the cases with $n \geq 3$, the smallest p value is 0.653. (Using the tables in Alexander and Quade (1968), the exact calculation gives a corresponding p-value of 0.683.) Thus, there is no statistically significant difference between the ranking of the candidates.

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

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

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

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

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

The associated p-values for different values of m and n are displayed in Table 6.

Table 6: 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:

$\underline{C_1}$	$\underline{C_2}$	$\underline{C_2}$	$\underline{C_j}$	$\underline{C_i}$	$\underline{C_j}$
3	1	1	3	1	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 $Z = -\sqrt{3/(2n+1)}$. Table 7 illustrates the values of Z for various size data along with their respective p-values.

Table 7: Mann-Whitney Test Statistics and p-values

n	Z_{C_1,C_2}	p-value
2	-0.577	0.282
3	-0.320	0.374
4	-0.210	0.417
5	-0.151	0.440

n	Z_{C_2,C_j}	p-value
2	-1.155	0.124
3	-0.961	0.168
4	-0.840	0.200
5	-0.756	0.225

n	Z_{C_i,C_j}	p-value
2	-0.577	0.282
3	-0.480	0.315
4	-0.420	0.337
5	-0.378	0.352

As before, the values are not statistically significant. On the other hand, 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.

6 Implications and Conclusions

Through our analyses in this paper, we have shown that data matrices that lead to ranking paradoxes when replicated and aggregated do not lead to statistical paradoxes. Because of this, from a statistical perspective, although the ranking paradoxes shown in the literature are surprising, we hope that this analysis makes them less disturbing. We have shown that in the cases in which the paradoxes arise, the difference between

the ranking of the candidates is not statistically significant. From another perspective, when there is a statistically significant difference between the ranking of the candidates, these paradoxes do not occur. In order for the null hypothesis of uniformity to be rejected, data structures must be more restricted. This means that for the test statistics to reject the null hypothesis, we must have a structural condition on the data set that is stronger than the row-ordered condition for the replication paradox to occur. The significance of what we've shown in Section 4 is that being row-ordered is not as strict of a condition as one might think. In another words, being row-ordered is a fairly weak condition (i.e., it doesn't mean there is a significant difference between the candidates), and yet that's sufficient to avoid paradoxes of inconsistency under replication. In the data sets that are not quite row-ordered in Section 5 and do produce paradoxes, there is still no significant difference between the candidates. This shows that the disagreement in the ranking of the aggregate data with the ranking given by the rows is due to the fact that the process used to aggregate the data weaken a candidate enough for the ranking to shift. From a statistical perspective, however, there is no significant difference between the candidates in any case, and the weakening of one candidate does not change this.

The ranking paradoxes and their lack of statistical significance should be of broad interest to statisticians. Both the Kruskal-Wallis and the Mann-Whitney tests are considered standard approaches to determining significant differences among alternatives in the nonparametric case. Although the V procedure is less widely used, it still provides an interesting comparison to the more popular tests. These results raise important questions about which of these three tests is most appropriate to use for a given data set. Furthermore, the analysis in this article only applies to the paradoxes introduced by Haunsperger using her constructions. It is possible that different constructions might lead to these paradoxes or to entirely new paradoxes. Further research is also necessary to determine whether paradoxes such as those that have arisen at the procedural level can arise at the statistical level. It is possible that different data structures might lead to such paradoxes. Finally, there are many other paradoxes, such as those in apportionment methods or voting methods with preference schedules, that could be analyzed in the same manner as we have done in this paper. We hope that this article provokes further investigation into these questions.

References

Alexander, D. A., and Quade, D. (1968), "On the Kruskal-Wallis Three Sample H-Statistic," *University of North Carolina School of Public Health, Institute of Statistics Mimeo Series No. 602.*

Bargagliotti, A. E., "Aggregating and decision making using ranked data: problems and solution," to appear.

Bargagliotti, A. E., and Saari, D. G., "Symmetry of nonparametric statistical tests on three samples," to appear.

Bhapkar, V. P. (1961), "A nonparametric test for the problem of several samples," *The Annals of Mathematical Statistics,* 32, 1108-117.

Haunsperger, D. B. and Saari, D. G. (1991), "The lack of consistency for statistical decision procedures," *The American Statistician,* 45, 252-255.

Haunsperger, D. B. (1992), "Dictionary of paradoxes for statistical tests on k samples," *Journal of the American Statistical Association,* 87, 249-255.

Haunsperger, D. B. (1996), "Paradoxes in Nonparametric Tests," *The Canadian Journal of Statistics,* 24, 95-104.

Haunsperger, D. B. (2003), "Aggregated statistical ranks are arbitrary," *Social Choice and Welfare,* 20, 261-272.

Kruskal, W. H., and Wallis, W. A. (1952), "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association,* 47, 583-621.

Lehman, E. L. (1975), *Nonparametrics: Statistical methods based on ranks,* Holden-Day, Inc.

Mann, H. B., and Whitney, D. R. (1947), "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics,* 18, 50-60.