Statistical Significance of Ranking Paradoxes

Raymond N. Greenwell, Hofstra University

October 8, 2008

Abstract: Haunsperger (2003) has shown that when the Kruskal-Wallis nonparametric statistical test on n samples is 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. We further investigate these ranking paradoxes by showing that when they occur, the differences in ranking are not statistically significant.

In a series of articles [2, 3, 4, 5] Haunsperger has shown that when the Kruskal-Wallis nonparametric statistical test on n samples is 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. In this paper, we further investigate these ranking paradoxes by finding the statistical significance of the differences in ranking when these paradoxes occur.

Example 1 [5]: Consider the two sets of data

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

each of which gives rise to the exact same matrix of ranks

C_1	C_2	C_3
3	2	1
$\underline{5}$	$\underline{4}$	$\underline{6}$
8	6	7

and, hence, the same ordering $C_1 \succ C_3 \succ C_2$ by Kruskal-Wallis ranking. When the two data sets are combined

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

and reranked,

C_1	C_2	C_3
8	7	6
10	9	11
3	2	1
$\underline{5}$	$\underline{4}$	$\underline{12}$
26	22	30

the ranking has changed to $C_3 \succ C_1 \succ C_2$. In cases such as Example 1, Haunsperger says that the 2×3 matrix of ranks from before is not *consistent under replication.*

Let's take the analysis one step further and calculate the Kruskal-Wallis statistic for Example 1:

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

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 Haunsperger considers in her papers.

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 $KW = 0.2857$, $p = 0.867$ for the original matrix of ranks, and $KW = 0.6154$, $p = 0.735$ for the combined matrix of ranks. In other words, there is no statistically significant difference in rankings between the three choices. Since N is small in these cases, the chi-square approximation is not highly accurate. The values of p for the exact distribution, using the tables in [1], are 0.917 and 0.770. These values do not change our conclusion, nor do they do so elsewhere in this paper, so we will continue to use the chi-square approximation.

Haunsperger proves many surprising results about matrices of ranks. In [5] 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 of ranks that are consistent under replication are row-ordered.” This is demonstrated by the following theorem.

Theorem 1 [**5**]**.** 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 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

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

For this matrix, it is straightforward to show that

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

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

$m\backslash n$	1	2	3	4
2	0.3173	0.4386	0.5127	0.5637
3	0.3679	0.5647	0.6703	0.7351
4	0.3916	0.6444	0.7641	0.8297
5	0.4060	0.7024	0.8266	0.8874

Notice that none of the values in the table approach statistical significance. For the smaller values of m and n, the approximate values in the table are inaccurate. For example, for $m = 2$ and $n = 1$, the true p-value is 0.5. This illustrates that although a matrix being row-ordered may seem unrealistically strict, in fact it is possible for a matrix to be row ordered and yet for there to be no significant difference in ranking 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.

As a consequence of Theorem 1, a matrix with unanimous ranking 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 with the following theorem.

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

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

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

Example 2: The proof of Theorem 4 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_0.

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

Observe that the ranking has now change to $C_3 \succ C_1 \succ C_2$. But $KW = 0.615$, $p = 0.735$ for original matrix of ranks, $KW = 0.5$, $p = 0.779$ for matrix with ranks switched, and $KW = 0.62$, $p = 0.733$ for combined matrix of ranks. Thus 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.

Let us look at the situation in general. Consider the matrix of ranks

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

For this matrix,

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

This decreases with n, so has its maximum value at $n=1$, where

$$KW = m-1.$$

In this case, however, the statistic can only take on one value, and hence is not meaningful. In the next largest case, with $n=2$, the statistic has the value

$$KW = \frac{m^3-m-6}{m(2m+1)}.$$

The chi-square approximation yields a minimum p-value of 0.651 when $m = 3$. Using the tables in [1], the exact calculation for $m=3$, $n=2$ gives a p-value of 0.803.

Now consider the combined matrix of ranks

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

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 the following table.

$m\backslash n$	1	2	3	4
2	0.1213	0.5637	0.7488	0.8336
3	0.1561	0.5004	0.6525	0.7334
4	0.1979	0.5663	0.7276	0.8095
5	0.2438	0.6421	0.8010	0.8745

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 0.6525. (Using the tables in [1], the exact calculation gives a corresponding p-value of 0.683.) Thus, there is no statistically significant difference between the ranking of the candidates.

Although the paradoxes shown by Haunsperger 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. Conversely, when there is a statistically significant difference between the ranking of the candidates, these paradoxes do not occur.

The analysis in this article only applies to the the paradoxes of Haunsperger using her constructions. It is possible that different constructions might lead to these paradoxes or to entirely new paradoxes, in which case the analysis in this article does not apply. We hope that this article provokes further investigation into this area.

References

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

[2] Deanna B. Haunsperger and Donald G. Saari, The lack of consistency for statistical decision procedures, *The American Statistician* **45** (1991) 252-255.

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

[4] Deanna B. Haunsperger, Paradoxes in Nonparametric Tests, *The Canadian Journal of Statistics* **24** (1996) 95-104.

[5] Deanna B. Haunsperger, Aggregated statistical ranks are arbitrary, *Social Choice and Welfare* **20** (2003) 261-272.