

Combinatorics and Statistical Issues Related to the Kruskal-Wallis Statistic

Anna E. Bargagliotti, Loyola Marymount University
Raymond N. Greenwell, Hofstra University

October 11, 2011

Department of Mathematics
Loyola Marymount University
Los Angeles, CA 90045
abargag@yahoo.com

Department of Mathematics
Hofstra University
Hempstead, NY 11549
matrng@hofstra.edu

Key words: Kruskal-Wallis, continuity correction, rank-sums, tournaments, row-ordered data, Young tableaux

ABSTRACT

We explore criteria that data must meet in order for the Kruskal-Wallis test to reject the null hypothesis by computing the number of unique ranked

data sets in the balanced case where each of the m alternatives has n observations. We show that the Kruskal-Wallis test tends to be conservative in rejecting the null hypothesis, and we offer a correction that improves its performance. We then compute the number of possible data sets producing unique rank-sums. The most commonly occurring data lead to an uncommonly small set of possible rank-sums. We extend prior findings about row- and column-ordered data structures.

1 Introduction

To determine whether the population distribution functions for m alternatives are significantly different, the Kruskal-Wallis test is commonly employed (Kruskal and Wallis 1952). The Kruskal-Wallis test is a nonparametric equivalent of the parametric one-way ANOVA. Due to its widespread use, it is imperative that the rejection probability of the test be calculated as accurately as possible. In addition, it is of interest to describe conditions that data sets must meet in order for the Kruskal-Wallis test to reject the null hypothesis.

In this paper, we consider the balanced data case where each of the m alternatives $A_1 \dots A_m$ has n observations to explore what criteria data must meet in order for the Kruskal-Wallis to reject the null. In order to answer this question, we consider the number of unique ranked data sets that exist for a fixed m and n as well as the accuracy with which the Kruskal-Wallis test rejects the null hypothesis.

Using simulations in R and in Mathematica, we note that the Kruskal-

Wallis test tends to be conservative in rejecting the null hypothesis. In order to account for the probability of rejection of a randomly selected data set being smaller than desired, we offer a correction to the statistic that improves its performance. Once the correction is made, we then investigate the criteria that a ranked data set must meet in order for the Kruskal-Wallis test to reject the null hypothesis. To do so, we use combinatorial efforts to compute the number of possible data sets producing unique rank-sums. In this manner, we note the conditions the rank sums must meet in order for the Kruskal-Wallis to reject.

Finally, our work builds on prior results by Bargagliotti and Greenwell (2011) and Haunsberger (2003) that point out that a data set being row-ordered is enough for it to produce consistent ranking outcomes but does not ensure consistency at the test rejection level. This paper further examines row-ordered data structures as well as column-ordered data structures and investigates the probability that the Kruskal-Wallis test rejects the null hypothesis when restricted to these data structures.

Section 2 of the paper explores the accuracy of the Kruskal-Wallis statistic and offers a correction. Sections 3 and 4 present conditions the rank-sums must meet in order for the Kruskal-Wallis to reject by first counting the number of possible data sets and then examining the number of distinct column rank-sums. Section 5 then examines how specific data structures discussed in the literature affect the outcome of the Kruskal-Wallis test. In each of these sections, we compute the statistical significance for the Kruskal-

Wallis for the general forms of data. Section 6, the last section of the paper, concludes and discusses the implications of these results.

2 Continuity Correction

In order to explore the type of data that leads to a Kruskal-Wallis test rejection, it is first important to understand the accuracy of the probability of rejection of the test. Ideally, the probability of rejecting the null hypothesis when the ranks are allocated at random should be α , but in practice, this is not necessarily true. Consider the balanced data set with n observations for each of the m alternatives:

$$\begin{array}{cccccc}
 \underline{A_1} & \underline{A_2} & \underline{A_3} & \dots & \underline{A_m} \\
 x_{11} & x_{21} & x_{31} & \dots & x_{m1} \\
 x_{12} & \dots & \dots & \dots & x_{m2} \\
 \dots & \dots & \dots & \dots & \dots \\
 x_{1n} & \dots & \dots & \dots & x_{mn}
 \end{array}$$

Except for the case when $m = 2$, the Kruskal-Wallis statistic tends to not reject the null hypothesis sufficiently often. To see this, we manipulate the Kruskal-Wallis statistic algebraically and then simulate the hypothesis test. The Kruskal-Wallis statistic is

$$KW = \frac{12}{N(N+1)} \sum_{i=1}^m n_i (\bar{r}_i - \bar{r})^2,$$

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. An approximation when N is not too small is

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

$m \backslash n$	2	3	4	5	6	7	8
2	0	0.09918	0.05646	0.05567	0.04210	0.05279	0.05129
3	0	0.01148	0.04011	0.04433	0.04184	0.04408	0.04476
4	0	0.02059	0.03317	0.03821	0.04005	0.04237	0.04281
5	0	0.01873	0.03101	0.03589	0.03833	0.04129	0.04236
6	0	0.01867	0.02870	0.03492	0.03731	0.03972	0.04196
7	0.00007	0.01810	0.02876	0.03448	0.03739	0.03985	0.04132
8	0.00038	0.01908	0.02810	0.03487	0.03827	0.03872	0.03949

Table 1: Fraction of times out of 100,000 H_0 rejected in simulation

In our case, with $n_k = n$, $N = mn$, $\bar{r}_i = R_i/n$, and $\bar{r} = (mn + 1)/2$, this leads to a rejection of the null hypothesis when the rank-sums meet the following condition:

$$\sum_{i=1}^m R_i^2 \geq \frac{mn^2(mn + 1)}{12} [\chi_{m-1}^2(\alpha) + 3(mn + 1)]. \quad (1)$$

The probability of rejecting the null hypothesis when the ranks are allocated at random is not exactly α because the Kruskal-Wallis statistic can only take on a finite number of discrete values. For example, in the case with $m = 3$ and $n = 2$, there are 15 unique sets of ranked data listed and analyzed in Bargagliotti and Saari (2011). Of these 15, none lead to the rejection of the null hypothesis at the $\alpha=0.05$ level. Furthermore, the exact distribution of the Kruskal-Wallis statistic is usually not used because it can be adequately approximated by the χ^2 statistic when the number of data is not too small. However, for small cases such as $m = 3$ and $m = 4$, the exact distribution of the Kruskal-Wallis statistic has been discussed in the literature (see Choi et al., 2003 and Iman et al., 2975).

In order to investigate the probability of the Kruskal-Wallis test rejecting

the null hypothesis given an arbitrary data set for a fixed m and n , we simulate the hypothesis test. Table 1 shows the fraction of times H_0 was rejected out of 100,000 trials using our simulation in the computer language R, randomly selecting combinations of rankings and using the χ^2 approximation with $\alpha = 0.05$. Looking at the table results, it can be seen that many of the values are not close to 0.05. This means that the probability with which the Kruskal-Wallis test is rejecting the null is not as accurate as may be desired.

To make the values in the above table closer to 0.05, one can use a continuity correction, as suggested for the two-sample Kolmogorov-Smirnov statistic in Kim (1969) and Greenwell and Finch (2004). In this case, rather than rejecting the null hypothesis when

$$P(\chi_{m-1}^2 \geq KW) < \alpha,$$

we instead reject the null hypothesis when

$$P(\chi_{m-1}^2 \geq KW) - e(m, n) < \alpha,$$

where $e(m, n)$ is a function designed to make the actual rejection probability as close to α as possible. It is interesting to note that a different continuity correction is preferred for $n = 2$ than for other values of n . By extending the above table for $n = 2$ for $m = 9$ to 20, and then doing a logarithmic regression, the best continuity correction for $n = 2$ is

$$e(m, 2) = \frac{0.053}{n^{0.044}}. \tag{2}$$

Doing a logarithmic regression of the above table for $m \geq 3$ and $n \geq 3$ gives the continuity correction

$$e(m, n) = \frac{0.056m^{0.53}}{n^{1.4}} \text{ for } m \geq 3, n \geq 3. \quad (3)$$

Applying the continuity correction (2) for $n = 2$ and the correction (3) for all other values of n leads to the Table 2, showing the fraction of times H_0 was rejected out of 100,000 trials.

The entries in Table 2 tend to be closer to 0.05 than those in Table 1. However, the continuity correction does not help when $m = 2$, for which the Kruskal-Wallis statistic tends to reject more than it should. For this case, there is not a clear pattern in terms of n . This, however, may be unavoidable. For example, when $m = 2$ and $n = 3$, there are only 10 different combinations of ranks, so the null hypothesis can only be rejected none of the time, or 10% of the time, or 20% of the time, etc. To get around this limitation one can use a randomized procedure, such as the one described in Greenwell and Finch (2004), in which for the same value of the statistic, sometimes the

$m \backslash n$	2	3	4	5	6	7	8
2	0	0.09918	0.05646	0.05567	0.06543	0.05279	0.05129
3	0	0.07303	0.05784	0.05105	0.05131	0.05022	0.05048
4	0.00938	0.04306	0.05279	0.05048	0.04965	0.05031	0.04930
5	0.03144	0.04723	0.05079	0.04993	0.04866	0.04999	0.04934
6	0.02987	0.04814	0.04927	0.05037	0.04901	0.04930	0.04958
7	0.02922	0.04810	0.05043	0.05032	0.04999	0.04996	0.04929
8	0.03040	0.05057	0.05060	0.05207	0.05146	0.04962	0.04819

Table 2: Fraction of times out of 100,000 H_0 rejected using continuity correction

null hypothesis is rejected and sometimes not. Despite this limitation for the case where $m = 2$, the continuity correction recommended here improves the overall performance of the statistic.

Tables in the Appendix show that the continuity correction also improves accuracy for $\alpha = 0.01$. and $\alpha = 0.10$.

3 Counting the Number of Possible Data Sets

Now that the rejection probability of the Kruskal-Wallis is being calculated as accurately as possible, we turn our attention to understanding the types of data structures that lead to a rejection. Our approach is to consider all possible data sets that can occur for a fixed m and n and then determine which data will lead to a Kruskal-Wallis rejection of the null hypothesis. To do so, we begin by counting the number of possible unique balanced ranked data sets that could occur for a fixed m and n . We then proceed by counting the number of possible unique rank-sum combinations that can occur. For example, for the case of $m = 3$ and $n = 2$, one possible ranked data set is

$$\begin{array}{ccc} \frac{A_1}{5} & \frac{A_2}{6} & \frac{A_3}{4} \\ \frac{3}{8} & \frac{1}{7} & \frac{2}{6} \end{array}$$

in which we have arranged the data so that in the column sums of ranks R_i , we have $R_1 \geq R_2 \geq R_3$. As mentioned above, all possible 15 different sets of ranked data for the case of $m = 3$ and $n = 2$ are listed and analyzed in Bargagliotti and Saari (2001). In general, let $r(m, n)$ be the number of such

cases. It is straightforward to prove the following theorem.

Theorem 1

$$r(m, n) = \frac{(mn)!}{m!(n!)^m}. \quad (4)$$

Table 3 illustrates the values of $r(m, n)$ for values of m and n between 2 and 8. As can easily be seen in the table, as m and n become large, $r(m, n)$ becomes extremely large.

It should be noted that each possible data set does not necessarily have a unique rank sum. For example, with $m = 3$ and $n = 2$, in addition to the data set listed above, the set of ranks

$\frac{A_1}{6}$	$\frac{A_2}{4}$	$\frac{A_3}{5}$
2	3	1

also leads to the column sums $R_1 = 8$, $R_2 = 7$, and $R_3 = 6$. Because the Kruskal-Wallis statistic is a function of the rank-sums of each alternative A_i , we investigate the distinct rank-sums that could occur for a fixed m and n . This will give us the number of possible unique values of the Kruskal-Wallis statistic for a fixed m and n , which is smaller than the number of allocations

$m \backslash n$	2	3	4	5	6	7	8
2	3	10	35	126	462	1716	6435
3	15	280	5775	126126	2858856	66512160	1577585295
4	105	15400	2627625	488864376	96197645544	$1.969 \cdot 10^{13}$	$4.148 \cdot 10^{15}$
5	945	1401400	2546168625	$5.195 \cdot 10^{12}$	$1.142 \cdot 10^{16}$	$2.648 \cdot 10^{19}$	$6.381 \cdot 10^{22}$
6	10395	190590400	$4.509 \cdot 10^{12}$	$1.234 \cdot 10^{17}$	$3.709 \cdot 10^{21}$	$1.191 \cdot 10^{26}$	$4.013 \cdot 10^{30}$
7	135135	36212176000	$1.319 \cdot 10^{16}$	$5.722 \cdot 10^{21}$	$2.779 \cdot 10^{27}$	$1.461 \cdot 10^{33}$	$8.143 \cdot 10^{38}$
8	2027025	$9.162 \cdot 10^{12}$	$5.929 \cdot 10^{19}$	$4.706 \cdot 10^{26}$	$4.263 \cdot 10^{33}$	$4.235 \cdot 10^{40}$	$4.505 \cdot 10^{47}$

Table 3: Number of Combinations of Rankings, $r(m, n)$

of ranks. Our computations extend the work done by Choi et al. (2003) and Iman et al. (1975). They proposed looking only at the distinct combinations of R_i in developing algorithms for computing the exact distribution of the Kruskal-Wallis statistic for small m values. As such, they recursively computed all possible rank-sum combinations for small values of m and their associated distribution.

To find all the unique possible combinations of R_i in the general case for a fixed m and n , note that the sum of all the ranks is $S = mn(mn + 1)/2$. Because switching the order of the alternatives has no affect on the number of possible data sets that can be created, then without loss of generality, we will restrict our discussion to data sets that have $R_i \geq R_j$ for $i < j$. With this restriction, then each R_i must be at least the total sum of all ranks not already allocated to previous R_i , divided by the total number of columns left. Also, each R_i cannot be so large that the remaining columns cannot be filled, even using all the smallest ranks. The smallest possible sum for the remaining $m - i$ columns is

$$1 + 2 + \cdots + n(m - i) = \frac{n(m - i)(n(m - i) + 1)}{2}.$$

We therefore have the following theorem.

Theorem 2

$$\left\lceil \frac{S - \sum_{k=1}^{i-1} R_k}{m - i + 1} \right\rceil \leq R_i \leq \min \left[R_{i-1}, S - \sum_{k=1}^{i-1} R_k - \frac{n(m - i)(n(m - i) + 1)}{2} \right], \quad (5)$$

where $\lceil \cdots \rceil$ indicates the ceiling function.

$m \setminus n$	2	3	4	5	6	7	8
2	3	5	9	13	19	25	33
3	13	50	145	338	685	1,250	2,113
4	76	630	3,173	11,466	33,268	82,400	181,521
5	521	9,285	81,441	455,741	1,893,961	6,365,425	18,276,481
6	3,996	151,652	2,315,021	20,044,960	119,298,580	543,960,010	*
7	32,923	2,658,131	70,592,121	945,549,475	*	*	*
8	286,202	49,061,128	*	*	*	*	*

(* indicates that the calculation took too much time to be completed.)

Table 4: Number of Combinations of Column Sums, $c(m, n)$

To illustrate the theorem with $m = 3$ and $n = 2$, the above inequality gives $7 \leq R_1 \leq 11$. For $R_1 = 7$, $7 \leq R_2 \leq 7$, for just the possibility $(7, 7, 7)$. (R_m is automatically determined by the previous values of R_i .) For $R_1 = 8$, $7 \leq R_2 \leq 8$, for the 2 possibilities $(8, 7, 6)$ and $(8, 8, 5)$. Similarly continue on to $R_1 = 11$, for which $5 \leq R_2 \leq 7$. Summing all of these possibilities leads to a total of 13. Denote this number by $c(3, 2)$. This is only slightly less than $r(m, n) = 15$. For larger m and n , however, the savings becomes enormous, as indicated by comparing Table 4, where we give the number of combinations of rank sums $c(m, n)$, with Table 3. Table 4 was created by running a computational program in R for all values of R_i satisfying inequalities (5).

Notice that the inequalities (5) are much tighter than those given by Choi et al. (2003). Due to this improvement, it would be interesting and desirable to prove the following conjecture.

Conjecture 1 *There exists an allocation of ranks leading to every set of*

rank column sums R_i satisfying inequalities (5).

Proof of this conjecture must specify a general way to generate a data set for a given set of rank sums that satisfy the inequalities (5). Although the conjecture appears to be quite simple, and all of our numerical explorations indicate that it is true, it is in fact difficult to prove. It is not clear that any existing combinatorial results can be applied to this scenario. In addition, all of our combinatorial efforts have fallen short of a proof that is valid in all cases. Therefore, we challenge the readers to find a proof for this conjecture.

3.1 Formula for $c(m, n)$

Ideally, it would be preferable to find a closed formula for $c(m, n)$ comparable to Equation (4). However, it has been observed in Choi et al. (2003) and MacMahon (1979) that it is unlikely that such a formula exists.

It is possible, however, to find a closed formula for row 1 of Table 4, that is, for $c(2, n)$. This sequence of numbers is sequence A099392 in the On-Line Encyclopedia of Integer Sequences (2011) (ignoring the initial two 1's in that sequence), for which the formula

$$a_n = \left\lceil \frac{(n+1)^2}{2} \right\rceil - n \tag{6}$$

is given. (We have changed the index in the formula cited from n to $n + 1$ to match our notation.) To see why this formula applies, notice that when $m = 2$, once R_1 is determined, the only remaining column sum, R_2 , is

automatically determined. For $m = 2$ and $i = 1$, formula (5) gives

$$\left\lceil \frac{n(2n+1)}{2} \right\rceil \leq R_1 \leq \frac{n(3n+1)}{2}. \quad (7)$$

The number of values of R_1 satisfying inequalities (7) is

$$\frac{n(3n+1)}{2} - \left\lceil \frac{n(2n+1)}{2} \right\rceil + 1 = \frac{3n^2 + 3n + 2}{2} - \left\lceil \frac{2n^2 + n}{2} \right\rceil - n.$$

By considering the two cases when n is even and odd, it is easily shown that this expression is the same as formula (6).

We therefore have the following theorem.

Theorem 3

$$c(2, n) = \left\lceil \frac{(n+1)^2}{2} \right\rceil - n$$

Although a general formula has not been found for $c(m, n)$, the values of $c(m, n)$ in Table 4 have been studied in the context of the number of possible point allocations in a round-robin tournament. The latter were studied in MacMahon (1979), where MacMahon demonstrated a recursive procedure for computing the point allocations when ties are possible. For example, notice that the $n = 2$ column in Table 4 (3, 13, 76, 521, ...) is the same as sequence A047730 in the On-Line Encyclopedia of Integer Sequences (2011), which gives the number of score sequences in tournaments with m players, when 4 points are awarded in each game. The $n = 3$ column (5, 50, 630, 9,285, ...) is the same as sequence A047736, which gives the number of score sequences when 9 points are awarded. Building on these prior findings, we have the following theorem.

Theorem 4 $c(m, n)$ is the number of score sequences in tournaments with m players, when n^2 points are awarded in each game.

To prove this theorem, we create a one-to-one correspondence between sets of rank sums and score sequences, which may be done by pairing the set of rank sums $\{R_i\}$ with the score sequence $\{S_i\} = \{R_i - n(n+1)/2\}$. This gives

$$\sum_{i=1}^m S_i = \frac{mn(mn+1)}{2} - m \cdot \frac{n(n+1)}{2} = \frac{m(m-1)}{2} \cdot n^2.$$

Alternatively, the total number of points given is

$$n^2(m-1) + n^2(m-2) + \cdots + n^2 + 0 = \frac{m(m-1)}{2} \cdot n^2,$$

which is the same sum.

MacMahon's recursive procedure for calculating the number of score sequences is still not efficient for large values of m and n . In light of this fact, computer computations like those generated in R to construct Table 4 remain key to counting the number of distinct possible ranks.

4 Rejection Probability for Combination of Ranks

If we consider each rank sum combination to be equally likely, then in order for the Kruskal-Wallis test to reject, the sum of the rank sums must meet the conditions described in equation (1). However, as noted in the previous section with the computation of $c(m, n)$ being smaller than $r(m, n)$ for all m

and n , in actuality, we know certain ranks combinations are more likely than others. Therefore, when trying to understand what type of ranked data will lead to a Kruskal-Wallis rejection, it may be helpful to better understand the distribution of the rank-sums.

Section 2 described the probability of rejecting the null hypothesis under the assumption that all distributions of ranks are equally likely, that is, all values of $r(m, n)$ are equally likely. In this section, we instead examine the probability of rejecting the null under the assumption that all possible column rank-sums are equally likely, that is, the distribution of values of $c(m, n)$ for a fixed m and n is uniform. As before, we do this using a simulation in R.

Table 5 shows the number of times H_0 is rejected at $\alpha = 0.05$ for various values of m and n . Table 6 shows the same information as a fraction of $c(m, n)$, thus giving the probability of rejecting H_0 under the assumption that the each of the rank-sums are equally likely. Table 6 indicates that the rejection probability increases and approaches 1 as m and n increase. This means that as more and more observations are obtained for a fixed n , then the Kruskal-Wallis is more likely to reject. In addition, the table illustrates that as more and more alternatives are considered, then the probability of rejecting also is tending to 1. Thus, when we consider the unique number of rank-sums, the Kruskal-Wallis test no longer rejects with probability equal to α . This illustrates that the distribution of possible rank-sums is overwhelmingly distributed around the rank-sums that do not lead to a rejection. In other words, the more commonly occurring rank-sums do not lead to a re-

$m \setminus n$	2	3	4	5	6	7	8
2	0	1	2	4	6	10	14
3	0	3	44	150	360	747	1,370
4	0	95	1,326	6,664	22,689	61,680	144,374
5	0	2,162	43,340	319,421	1,502,764	5,405,623	16,178,840
6	0	48,690	1,471,503	15,868,250	103,684,399	495,940,175	*
7	7	1,088,571	51,009,195	811,190,365	*	*	*
8	775	24,217,520	*	*	*	*	*

Table 5: Number of times H_0 rejected

$m \setminus n$	2	3	4	5	6	7	8
2	0	0.2000	0.2222	0.3077	0.3684	0.4000	0.4242
3	0	0.2000	0.3655	0.4645	0.5518	0.6128	0.6597
4	0.0132	0.2508	0.4869	0.6183	0.7051	0.7636	0.8056
5	0.0576	0.3640	0.6038	0.7360	0.8130	0.8610	0.8828
6	0.0686	0.4617	0.7023	0.8221	0.8844	0.9203	*
7	0.0969	0.5547	0.7817	0.8823	*	*	*
8	0.1293	0.6366	*	*	*	*	*

Table 6: Fraction of the time H_0 rejected with continuity correction

jection.

Although proving the limit approaches 1 explicitly is made difficult by the lack of a formula for $c(m, n)$, we can prove this result when $m = 2$ by using theorem 3. In this simple case, inequality (1) reduces to

$$R_1^2 + R_2^2 \geq \frac{n^2(2n+1)}{6} [\chi_1^2(\alpha) + 3(2n+1)]$$

under the constraint $R_1 + R_2 = n(2n+1)$. This can be simplified to

$$R_1 \geq \frac{n(2n+1)}{2} + n\sqrt{\frac{(2n+1)\chi_1^2(\alpha)}{12}}. \quad (8)$$

Combining (6) and (8) shows that the portion of the time that H_0 is rejected

is

$$\frac{n(3n + 1)/2 - \left\lfloor n(2n + 1)/2 + n\sqrt{(2n + 1)\chi_1^2(\alpha)/12} \right\rfloor}{\lceil (n + 1)^2/2 \rceil - n}.$$

The numerator is asymptotic to $n^2/2 - O(n^{3/2})$ and the denominator to $n^2/2$, so the fraction is asymptotic to 1.

Furthermore, there is a plausibility argument for the conjecture that the rejection probability approaches 1 for large m and n . Numerous combinations of ranks exist that lead to column rank sums that are close to each other and the null hypothesis is not rejected. On the other hand, for the more extreme case, in which the column rank sums are far apart, there are fewer combinations of ranks that lead to these column rank sums. In the most extreme case, when the n largest ranks are assigned to column 1, the next n ranks assigned to column 2, and so forth, there is only one combination of ranks that leads to this column rank sum. By only considering distinct rank sums, we are favoring these more extreme distributions of ranks. It is as if the extreme tail values in the normal distribution were to occur as frequently as values around the mean. In the limit, as the extreme tail values become further from the mean, the portion of rank sums that lead to rejection approaches 100%.

It turns out that the null hypothesis is always rejected (except for very small values of m and n) when the n largest ranks are assigned to column 1, the next n ranks assigned to column 2, and so forth. This leads to a

Kruskal-Wallis statistic of

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

Using the χ_{m-1}^2 approximation to the Kruskal-Wallis statistic, and the fact that the χ_{m-1}^2 statistic is approximately normal with mean $m - 1$ and standard deviation $\sqrt{2(m - 1)}$ for large m , the statistic above has a z -score of

$$\frac{m^2n^2 - m^2n + mn - n^2 - m + 1}{(mn + 1)\sqrt{2(m - 1)}}.$$

This is an increasing function of m and n and becomes large as soon as m and n become reasonably large, as Table 7 shows.

We therefore have the following result.

Theorem 5 *For $n > 2$, there exists a data set for all m such that the Kruskal-Wallis test will reject the null for $\alpha = 0.05$.*

5 Special Data Structures

Prior literature has discussed some very specific structures of ranked data. In particular, data could be row-ordered (Haunsperger (2003), Bargagliotti and

$m \backslash n$	2	3	4	5	6	7	8
2	0.990	2.020	3.064	4.114	5.167	6.223	7.279
3	1.286	2.600	3.923	5.250	6.579	7.909	9.240
4	1.497	3.015	4.539	6.065	7.593	9.122	10.652
5	1.671	3.359	5.051	6.745	8.440	10.135	11.831
6	1.824	3.662	5.502	7.345	9.188	11.031	12.875
7	1.963	3.936	5.913	7.890	9.869	11.847	13.826
8	2.091	4.191	6.293	8.396	10.500	12.603	14.708

Table 7: z -scores for extreme distribution of ranks

Greenwell (2011)), or both row- and column-ordered (Frame et al. (1954), Griffiths and Lord (2011)). In this section of the paper, we discuss each of these data structures separately. Our goal is to uncover whether these types of structures have affect on whether the Kruskal-Wallis test will reject the null hypothesis.

5.1 Row-Ordered Data

Haunsperger (2003) defines a matrix to be *row-ordered* if the observations of each alternative can be put in an order so that every row of the matrix gives rise to the same ranking of the m alternative. In Bargagliotti and Greenwell (2011), we showed that being row-ordered is not a sufficient condition for rejecting the null hypothesis. However, a row-ordered data set does create the largest possible Kruskal-Wallis statistic. This means that a particular type of row-ordered data minimizes the p -value obtainable for a fixed m and n .

Theorem 6 $\sum R_i^2$ is maximized when the largest ranks are in the first column, the next largest are in the next column, and so on.

Consider the data set of the form described in the theorem:

$$\begin{array}{ccccc}
 \underline{A_1} & \underline{A_2} & \underline{A_3} & \dots & \underline{A_m} \\
 x_{11} & x_{21} & x_{31} & \dots & x_{m1} \\
 x_{12} & \dots & \dots & \dots & x_{m2} \\
 \dots & \dots & \dots & \dots & \dots \\
 x_{1n} & x_{2n} & x_{3n} & \dots & x_{mn}
 \end{array}$$

where $x_{11} > x_{12} > \cdots > x_{1n} > x_{21} > \cdots > x_{2n} > x_{31} > \cdots > x_{mn}$. The theorem states that $R_1^2 + R_2^2 + \cdots + R_m^2$ will be maximized for data of this form. To see why this is true, suppose that we were to switch two of the entries in the data set, x_{ij} and x_{kp} , where $i < k$, so $x_{ij} - x_{kp} = q > 0$. This would lead the squares of the new column rank-sums to equal: $R_{i_{new}}^2 = (R_i - q)^2$ and $R_{k_{new}}^2 = (R_k + q)^2$. By distributing out these quantities, we see that $R_{i_{new}}^2 = R_i^2 - 2qR_i + q^2$ and $R_{k_{new}}^2 = R_k^2 + 2qR_k + q^2$. Because the quantity $2qR_i$ will always be larger than the quantity $2qR_k$, then it is straightforward to see that the original $R_i^2 + R_k^2$ obtained from the data described in the theorem is larger than $R_{i_{new}}^2 + R_{k_{new}}^2$.

This proves the theorem that the rank-sums are maximized when data have the row-ordered form described. The next theorem gives conditions on the rank-sums for the other extreme Kruskal-Wallis statistic and p -value for a fixed m and n .

Theorem 7 $\sum R_i^2$ is minimized when $R_1 = R_2 = \cdots = R_m$.

We see this by using the Lagrange multiplier method to find the minimum of the function $L = \sum R_i^2 + \lambda[nm(nm + 1)/2 - (R_1 + R_2 + \cdots + R_m)]$. Taking the partial derivatives with respect to R_1, R_2, \dots, R_m , and λ , setting them all equal to zero, and solving the system of equations, we find that the sum of the squares of the rank-sums is minimized when $R_1 = R_2 = \cdots = R_m = n(nm + 1)/2$. Data structures that lead to this extreme are not necessarily row-ordered. The following corollary describes such data sets for n even.

Corollary 1 For n even, $\sum R_i^2$ is minimized for a data set of the following form:

$$\begin{array}{ccccc}
 \frac{A_1}{m} & \dots & \frac{A_{m-2}}{3} & \frac{A_{m-1}}{2} & \frac{A_m}{1} \\
 m+1 & \dots & 2m-2 & 2m-1 & 2m \\
 \dots & \dots & \dots & \dots & \dots
 \end{array}$$

5.2 Young Tableaux

Another data structure discussed in the literature (Frame et al. (1954), Griffiths and Lord (2011)) is when both rows and columns are in decreasing order. This is an example of the combinatorial object known as a Young tableau. When each column has the same number of rows, this is a special case of a row-ordered matrix. It is well known that the number of such tableaux (or data sets in our application) is given by the hook length theorem. A hook is the set of elements consisting of an element in the tableau, plus all elements below and to the right of the element. The hook length of an element is the number of elements in the hook. The hook length theorem says that the number of ways to fill a Young tableau with N elements is $N!$ divided by the product of all hook lengths. We can apply the hook length theorem to count $y(m, n)$, the number of possible data sets that are row- and column-ordered. This leads to the following theorem.

Theorem 8

$$y(m, n) = \frac{(mn)! \prod_{k=0}^{n-1} k!}{\prod_{k=0}^{n-1} (m+k)!}$$

Table 8 lists values of $y(m, n)$. Notice that these numbers are much smaller than those in Table 1. Some, but not all, are smaller than those in Table 3. In particular, these numbers increase more slowly with m , but more rapidly with n , than those in Table 3. These numbers are symmetrical in m and n , something not obvious from the formula in the theorem above, but obvious from the definition of Young tableaux. Since some of these numbers are greater than $c(m, n)$, they clearly cannot all have distinct column sums.

Notice also that the numbers in column 1 or row 1 of Table 8 are the well-known Catalan numbers. Griffiths and Lord (2011) refer to the other numbers in Table 8 as generalized Catalan numbers.

Table 9 shows the probability of rejecting the null hypothesis of no difference between the columns under the assumption that all values of $y(m, n)$ are equally likely for fixed m and n . Because we only have a formula for the denominator of these probabilities, and not for the numerator, we used the Tableaux command in Combinatorica, an extension of Mathematica, which generates all Young tableaux. Then, we tested which of these Young tableaux

$m \setminus n$	2	3	4	5	6	7	8
2	2	5	14	42	132	429	1,430
3	5	42	462	6,006	87,516	1,385,670	23,371,634
4	14	462	24,024	1,662,804	140,229,804	$1.367 \cdot 10^{10}$	$1.490 \cdot 10^{12}$
5	42	6,006	1,662,804	701,149,020	$3.965 \cdot 10^{11}$	$2.786 \cdot 10^{14}$	$2.315 \cdot 10^{17}$
6	132	87,516	140,229,804	$3.965 \cdot 10^{11}$	$1.672 \cdot 10^{15}$	$9.490 \cdot 10^{18}$	$6.787 \cdot 10^{22}$
7	429	1,385,670	$1.367 \cdot 10^{10}$	$2.786 \cdot 10^{14}$	$9.490 \cdot 10^{18}$	$4.751 \cdot 10^{23}$	$3.210 \cdot 10^{28}$
8	1,430	23,371,634	$1.490 \cdot 10^{12}$	$2.315 \cdot 10^{17}$	$6.787 \cdot 10^{22}$	$3.210 \cdot 10^{28}$	$2.208 \cdot 10^{34}$

Table 8: Hook length theorem calculations

$m \backslash n$	2	3	4	5	6	7	8
2	0.0000	0.2000	0.1429	0.1667	0.2121	0.1935	0.1979
3	0.0000	0.2381	0.3139	0.3443	0.4034	0.4459	*
4	0.0714	0.3918	0.5177	0.5976	*	*	*
5	0.2619	0.5746	0.7098	*	*	*	*
6	0.4091	0.7245	*	*	*	*	*
7	0.5245	0.8397	*	*	*	*	*
8	0.6287	*	*	*	*	*	*

Table 9: Probability of rejecting H_0 for row- and column-ordered matrices with continuity correction

cause a rejection of the null hypothesis for the Kruskal-Wallis statistic using the chi-square approximation. Notice that although Table 8 is symmetrical in m and n , Table 9 is not. Comparing the values in Table 9 with those in Table 6, where we looked at the column rank-sums, we observe that the probabilities in Table 9 tend to be higher when n is small or m is large, but smaller otherwise. This is due to two countervailing effects. One is the plausibility argument we gave in a previous section for why the probabilities in Table 6 are so large, due to the fact that every column rank-sum is equally weighted, whether it only occurs for one rank combination or for many. This is not true for the row- and column- ordered matrices, many of which have the same column rank sum. On the other hand, the matrices that lead to the column rank sums that are most evenly distributed, and which have the largest p -values, are not row ordered, such as the following example

	$\frac{A_1}{1}$	$\frac{A_2}{2}$	
for $m = 2, n = 4$:	4	3	. This latter effect is more pronounced for
	5	6	
	<u>8</u>	<u>7</u>	
	18	18	

when n is small or m is large, while the effect of multiple rank combinations leading to the same column rank sum is more pronounced otherwise.

6 Conclusions

The overall goal of the paper is to explore what types of data lead the Kruskal-Wallis test to reject the null. As a first step, using simulations in R and in Mathematica, we offer a correction to the Kruskal-wallis statistic that improves its performance. Once the correction is made, we then use combinatorial efforts to compute the number of possible data sets producing unique rank-sums for the general case of m alternatives and n observations per alternative. In this manner, we can note the conditions the rank sums must meet in order for the Kruskal-Wallis to reject.

Our findings reveal that in practice when collecting data, one will tend to obtain data sets that only lead to a very small portion of all the possible rank-sums. In fact, the most common occurring data lead to an uncommonly small set of possible rank-sums. Finally, we extend prior results about row-ordered data structures as well as column-ordered data structures to understand that these criteria do not ensure rejection of the null hypothesis.

References

- Bargagliotti, A. E., Saari, D. G. (2011). Symmetry of nonparametric statistical tests on three samples, *Journal of Mathematics and Statistics*. In press.
- Bargagliotti, A. E., Greenwell, R. N. (2011). Statistical significance of ranking paradoxes, *Communications in Statistics–Theory and Methods* 40:916-928.
- Choi, W., Lee, J. W., Huh, M. H., Kang, S. H. (2003). An algorithm for computing the exact distribution of the Kruskal-Wallis test, *Communications in Statistics–Simulation and Computation* 32:1029-1040.
- Frame, J. S., Robinson, G. de B., Thrall, R. M. (1954). The hook graphs of the symmetric group, *Canadian Journal of Mathematics* 6:316-325.
- Greenwell, R. N., Finch, S. J. (2004). Randomized rejection procedure for the two-sample Kolmogorov-Smirnov statistic, *Computational Statistics and Data Analysis* 46:257-267.
- Griffiths, M., Lord, N. (2011). The hook-length formula and generalized Catalan numbers, *The Mathematical Gazette* 95:23-30.
- Haunsperger, D. B. (2003). Aggregated statistical ranks are arbitrary, *Social Choice and Welfare*, 20:261-272.
- Iman, R. L., Quade, R., Alexander, D. A. (1975). Exact probability levels for the Kruskal-Wallis test, *Selected Tables in Mathematical Statistics, Vol. III*, 20:261-272.

Kruskal, W. H., Wallis, W. A. (1952). Use of ranks on one-criterion variance analysis, *Journal of the American Statistical Association* Providence:American Mathematical Society, 329-384.

Kim, P. J. (1969). On the exact and approximate sampling distribution of the two sample Kolmogorov-Smirnov criterion $D_{mn}, m \leq n$, *Journal of the American Statistical Association* 64:16251637.

MacMahon, P. A. (1979). Chess tournaments and the like treated by the calculus of symmetric functions, In: Andrews, G. E., ed. *P. A. MacMahon, Collected Papers, Vol. I*, Cambridge:MIT Press, 353-384.

The On-Line Encyclopedia of Integer Sequences (2011), oeis.org.

7 Appendix

$m \setminus n$	2	3	4	5	6	7	8
2	0	0	0	0.00781	0.00427	0.00686	0.00678
3	0	0	0.00050	0.00336	0.00428	0.00544	0.00606
4	0	0	0.00101	0.00261	0.00401	0.00477	0.00549
5	0	0	0.00120	0.00269	0.00408	0.00500	0.00582
6	0	0.00001	0.00128	0.00284	0.00395	0.00466	0.00569
7	0	0.00009	0.00147	0.00261	0.00439	0.00488	0.00579
8	0	0.00018	0.00163	0.00300	0.00404	0.00462	0.00514

Table 10: Fraction of times H_0 rejected with $\alpha = 0.01$.

$m \setminus n$	2	3	4	5	6	7	8
2	0	0	0.02847	0.01559	0.01542	0.01098	0.01076
3	0	0.00377	0.01052	0.01113	0.01106	0.01042	0.01103
4	0	0.00793	0.01046	0.01155	0.01119	0.01084	0.01015
5	0	0.00999	0.01200	0.01195	0.01173	0.01136	0.01223
6	0.00062	0.01206	0.01267	0.01287	0.01162	0.01169	0.01206
7	0.00138	0.01341	0.01412	0.01424	0.01388	0.01288	0.01260
8	0.00264	0.01647	0.01506	0.01537	0.01431	0.01286	0.01218

Table 11: Fraction of times H_0 rejected with $\alpha = 0.01$ using continuity correction

$m \setminus n$	2	3	4	5	6	7	8
2	0	0.09918	0.11403	0.09511	0.09379	0.09729	0.10644
3	0	0.10047	0.09667	0.09229	0.09874	0.09616	0.09718
4	0.00938	0.08603	0.08848	0.09178	0.09251	0.09470	0.09459
5	0.02525	0.07295	0.08602	0.08864	0.08986	0.09253	0.09359
6	0.02791	0.06989	0.08064	0.08686	0.09014	0.09132	0.09147
7	0.02752	0.06611	0.07975	0.08528	0.08782	0.09081	0.09150
8	0.02860	0.06574	0.07767	0.08411	0.08813	0.08898	0.08966

Table 12: Fraction of times H_0 rejected with $\alpha = 0.10$.

$m \setminus n$	2	3	4	5	6	7	8
2	0.33270	0.09918	0.11403	0.09511	0.09379	0.09729	0.10644
3	0.06609	0.13954	0.10389	0.10544	0.10569	0.10166	0.10366
4	0.12387	0.11742	0.10929	0.10581	0.10293	0.10300	0.10167
5	0.11000	0.11124	0.10735	0.10368	0.10096	0.10174	0.10114
6	0.09690	0.10656	0.10405	0.10328	0.10234	0.10110	0.09971
7	0.09374	0.10610	0.10436	0.10300	0.10137	0.10179	0.10083
8	0.09196	0.10642	0.10434	0.10265	0.10336	0.10035	0.09904

Table 13: Fraction of times H_0 rejected with $\alpha = 0.10$ using continuity correction