

# Combinatorics and Statistical Issues Related to the Kruskal-Wallis Statistic

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

- Offer a continuity correction to the Kruskal-Wallis statistic
- Explore the number of combinations of column sums
- Explore the rejection probability when counted according to distinct column sums
- Look at special data structures: row-ordered data and Young tableaux

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

Ranks of  $1, \dots, mn$  allocated.

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

$N$  = number of data

$m$  = number of columns

$n_k$  = number of data in column  $k$

$\bar{r}_k$  = average rank of the data in column  $k$

$\bar{r}$  = average rank of the data.

$$p \approx P \left( \chi_{m-1}^2 \geq KW \right) < \alpha$$

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

where  $R_i$  = the sum of the ranks in column  $i$ .

$m \backslash n$	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

Fraction of times out of 100,000  $H_0$  rejected in simulation

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

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

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

$m \backslash n$	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

Fraction of times out of 100,000  $H_0$  rejected using continuity correction

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

**Example** ( $m = 3, n = 2$ )

$\frac{A_1}{5}$	$\frac{A_2}{6}$	$\frac{A_3}{4}$
$\underline{3}$	$\underline{1}$	$\underline{2}$
8	7	6

$$R_1 \geq R_2 \geq R_3$$

$\frac{A_1}{6}$	$\frac{A_2}{4}$	$\frac{A_3}{5}$
$\underline{2}$	$\underline{3}$	$\underline{1}$
8	7	6

$m \setminus n$	2	3	...	6	7	8
2	3	10		462	1,716	6435
3	15	280	2,858,856	66,512,160	1,577,585,295	
4	105	15,400	$9.620 \cdot 10^{10}$	$1.969 \cdot 10^{13}$	$4.148 \cdot 10^{15}$	
5	945	1,401,400	$1.142 \cdot 10^{16}$	$2.648 \cdot 10^{19}$	$6.381 \cdot 10^{22}$	
6	10,395	190,590,400	$3.709 \cdot 10^{21}$	$1.191 \cdot 10^{26}$	$4.013 \cdot 10^{30}$	
7	135,135	$3.621 \cdot 10^{10}$	$2.779 \cdot 10^{27}$	$1.461 \cdot 10^{33}$	$8.143 \cdot 10^{38}$	
8	2,027,025	$9.162 \cdot 10^{12}$	$4.263 \cdot 10^{33}$	$4.235 \cdot 10^{40}$	$4.505 \cdot 10^{47}$	

Number of Combinations of Rankings,  $r(m, n)$

$m \setminus n$	2	3	...	6	7	8
2	3	5		19	25	33
3	13	50		685	1,250	2,113
4	76	630		33,268	82,400	181,521
5	521	9,285	1,893,961	6,365,425	18,276,481	
6	3,996	151,652	119,298,580	543,960,010		*
7	32,923	2,658,131		*	*	*
8	286,202	49,061,128		*	*	*

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

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

## Theorem

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

where  $S = \sum_{i=1}^{mn} i = mn(mn+1)/2$  and  $\lceil \dots \rceil$  indicates the ceiling function.

## Conjecture

There exists an allocation of ranks leading to every set of rank column sums  $R_i$  satisfying the above inequalities.

## Theorem

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

## Theorem

$c(m, n)$  is the number of score sequences in round-robin tournaments with  $m$  players, when  $n^2$  points are awarded in each game. (Studied by P. A. MacMahon.)

$m \setminus n$	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	*	*	*	*	*

Fraction of the time  $H_0$  rejected with continuity correction when only distinct rank-sums are considered

Conjecture: This probability goes to 1 as  $n, m$  go to  $\infty$ . Proven when  $m = 2$ .

## Theorem

For  $n > 2$ , there exists a data set for all  $m$  such that the Kruskal-Wallis test will reject  $H_0$  for  $\alpha = 0.05$ .

## Theorem

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

## Theorem

$\sum R_i^2$  is minimized when  $R_1 = R_2 = \dots = R_m$ .

## Corollary

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

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

$y(m, n)$  = number of data sets that are row- and column-ordered (Young tableaux)

Using the hook length theorem,

## Theorem

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

$m \setminus n$	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	*	*	*	*	*	*

Probability of rejecting  $H_0$  for row- and column-ordered matrices with continuity correction