International Journal of Statistics and Applications 2015, 5(3): 113-119
DOI: 10.5923/j.statistics.20150503.03

Kruskal-Wallis Test: A Graphical Way

Elsayed A. H. Elamir

Department of Statistics and Mathematics, Benha University, Egypt & Management & Marketing Department, College of Business, University of Bahrain, Kingdom of Bahrain

Abstract The Kruskal-Wallis is a non-parametric method for testing whether samples originate from the same distribution. When the null hypothesis is rejected, at least one sample stochastically dominates at least one other sample. The test does not identify where this stochastic dominance occurs. Consequently, a decision limit for Kruskal-Wallis test is derived based on the gamma distribution and Bonferonni approximation that shows graphically where this stochastic dominance occurs. Simulation studies confirm the validity and robustness of the decision limit in small and large samples. An application is given to illustrate the method.

Keywords Bonferonni approximation, Chi square distribution, Gamma distribution, Nonparametric tests

1. Introduction

Kruskal and Wallis (1952) introduced a rank-based test for comparison of several medians or means, complementing the parametric approaches. Kruskal-Wallis (KW) test is to be among the most useful of available hypothesis testing procedures for behavioural and social research, though it is also one of the many under-utilized nonparametric procedures. Parametric methods, along with the requirement for a stronger set of assumptions, continue to dominate the research landscape despite convincing studies that call into question the wisdom of making such assumptions; see, [2], and [3]. Replacing original scores with ranks does not inherently lead to lower power, as one might suppose, but rather can result in a power increase at best and a slight power loss, at worst; see, for example, [4], [5], [6], [7] and [8].

Kruskal and Wallis [1] said that "...One of the most important applications of the test is in detecting differences among the population means" (p.584). Also they suggested that "… in practice the H test may be fairly insensitive to differences in variability, and so may be useful in the important 'Behrens-Fisher problem' of comparing means without assuming equality of variances" (p.599). Furthermore, Iman [9] formulated the null hypothesis of Kurskal-Wallis test in terms of the expected values (p. 726).

When the null hypothesis is rejected, at least one sample stochastically dominates at least another one. The KW test does not identify where this stochastic dominance occurs. Therefore, the Kruskal-Wallis test is rewritten in the form of the sum of independent chi square random variables. Using the idea of Bonferonni approximation, a decision limit for Kruskal-Wallis test is obtained based on the gamma distribution that shows graphically which groups out of the decision limit. A simulation study is conducted using the normal and exponential distributions to confirm the validity and robustness of the decision limit in small and large samples for the new method in comparison with KW test.

The graphical presentation of the KW test is introduced in Section 2. Simulation results to compare between the KW test and the new method are given in Section 3. An application is presented in Section 4. Section 5 is devoted for conclusion.

2. Graphical Presentation of Kruskal-Wallis Test

Suppose independent random observations $Y_{gi}(g = 1, \dots, G, i = 1, \dots, n_g,$ and $n_1 + \cdots + n_g = n)$ are obtained from a continuous population with mean μ_g and variance σ_g^2. G is the number of groups or treatments and n_g is the sample size in each group. The model is

$$Y_{gi} = \mu + \alpha_g + \epsilon_{gi}$$

μ is the global location of the data, α_g the difference to the location of the g-th group and ϵ_{gi} is the residual error; see, for example, [2]. Thus the null hypothesis can be expressed as

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_G = \theta$$

versus at least two medians or means are not equal.

2.1. Kruskal-Wallis test

The ranks of the observations Y_{gi} are

r_{gi} = the rank of Y_{gi} in the combined sample

The Kruskal-Wallis test is defined as

* Corresponding author:
shabib40@gmail.com (Elsayed A. H. Elamir)
Published online at http://journal.sapub.org/statistics

$$KW = \frac{(n-1)\sum_{g=1}^{G} n_g\left(\bar{r}_g - \bar{r}\right)^2}{\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(r_{gi} - \bar{r}\right)^2}$$

$$= \frac{12}{n(n+1)}\sum_{g=1}^{G} n_g\left(\bar{r}_g - \frac{n+1}{2}\right)^2$$

where $\bar{r}_g = \frac{\sum_{g=1}^{n_g} r_{gi}}{n_g}$ and $\bar{r} = (n+1)/2$ and it is assumed that the ties are handled by random method.

This is distributed as $\chi^2(G-1)$; See, [10] and [1] and [11]. Also, KW can be written as

$$KW = \sum_{g=1}^{G}\left(\frac{\bar{r}_g - 0.5(n+1)}{\sqrt{\frac{n(n+1)}{12n_g}}}\right)^2$$

The contribution of each standardized group mean rank in the test is defined as

$$H_g = \left(\frac{\bar{r}_g - 0.5(n+1)}{\sqrt{\frac{n(n+1)}{12n_g}}}\right)^2, \qquad g = 1, \dots, G$$

Therefore, the Kruskal-Wallis test could be plotted as

$$x\_axis = g \text{ versus } y\_axis = H_g, \text{ for } g = 1,2, \dots, G$$

This is called H-graph. To do this the sampling distribution of H_g is needed. The first thinking is gamma distribution where [1] used in small sample sizes and the chi square is a special case from it.

2.2. Sampling Distribution of H

To investigate if the sampling distribution of H is a gamma distribution or not, a simulation study is conducted to obtain the first four moments of H for $G = 3$ and $G = 4$ using simulated data from normal and exponential distributions with different sample sizes. Three sets of data are investigated (a) small ($n_g = 10$), (b) medium ($n_g = 25$) and large ($n_g = 50$). The following steps are used in simulation:

1. Simulate data from a distribution with the same mean for the required design.
2. Rank the combined sample, compute H for each group and moments for each H.
3. Repeat this r times and compute average for each moments.

Tables 1 and 2 gives the moments of H for $G = 3$ and $G = 4$.

From Tables 1 and 2 it can conclude that

$$E(H_g) = 1 - \frac{1}{G} \text{ and } Var(H_g) = 2 - \frac{3.5}{G}$$

Table 1. Empirical first four moments (mean, variance, skewness and kurtosis) of H_g using simulated data from normal and exponential distributions, $G = 3$ and the number of replications is 10000

		$G = 3$							
n_g		Normal				Exponential			
(10,10,10,10)		Mean	Var.	Skew	Kurt	Mean	Var.	Skew	Kurt
	H_1	0.659	0.800	2.542	11.973	0.679	0.871	2.539	11.856
	H_2	0.659	0.818	2.582	12.277	0.663	0.826	2.590	12.799
	H_3	0.653	0.804	2.604	12.747	0.675	0.851	2.499	11.232
(25,25,25,25)									
	H_1	0.658	0.829	2.802	15.048	0.673	0.860	2.629	13.087
	H_2	0.667	0.885	2.688	12.989	0.673	0.871	2.788	15.291
	H_3	0.663	0.844	2.658	13.315	0.670	0.867	2.599	12.302
(50,50,50,50)									
	H_1	0.661	0.858	2.699	13.567	0.667	0.891	2.820	14.655
	H_2	0.667	0.896	2.915	16.121	0.663	0.862	2.694	13.311
	H_3	0.676	0.888	2.692	13.987	0.663	0.857	2.596	12.469

Table 2. Empirical first four moments (mean, variance, skewness and kurtosis) of H_g using simulated data from normal and exponential distributions, $G = 4$ and the number of replications is 10000

$G = 4$									
n		Normal				Exponential			
(10,10,10,10)		Mean	Var.	Skew	Kurt	Mean	Var.	Skew	Kurt
	H_1	0.753	1.061	2.508	11.47	0.759	1.072	2.553	12.135
	H_2	0.762	1.082	2.571	12.01	0.737	1.014	2.510	11.533
	H_3	0.759	1.072	2.517	11.59	0.746	1.043	2.573	12.430
	H_4	0.750	1.043	2.516	11.47	0.752	1.076	2.590	12.383
(25,25,25,25)									
	H_1	0.758	1.139	2.856	15.35	0.748	1.050	2.505	11.449
	H_2	0.758	1.112	2.625	12.451	0.758	1.119	2.875	15.192
	H_3	0.758	1.137	2.803	14.274	0.754	1.101	2.698	13.361
	H_4	0.763	1.121	2.773	13.041	0.747	1.041	2.596	12.419
(50,50,50,50)									
	H_1	0.746	1.126	2.833	14.562	0.757	1.128	2.748	14.18
	H_2	0.732	1.057	2.801	14.934	0.749	1.088	2.686	13.289
	H_3	0.752	1.136	2.829	14.877	0.742	1.068	2.782	14.808
	H_4	0.731	1.081	2.926	16.199	0.742	1.065	2.747	14.319

Therefore the chi square distribution with one degree of freedom does not fit H for small G. The gamma distribution is used to fit the sampling distribution of H by matching the first two moments; see, [1]. Since the first two moments for gamma distribution are

$$\text{Mean} = k\theta \quad \text{and} \quad \text{Var} = k\theta^2$$

where k is the shape and θ is the scale. Therefore,

$$H_g \approx \text{gamma}\left(k = \frac{(1-1/G)^2}{2-3.5/G}, \theta = \frac{2-3.5/G}{1-1/G}\right)$$

Using the shape and the rate=$1/\theta$ parametrization then

$$H_g \approx \text{gamma}\left(\text{shape} = \frac{(1-1/G)^2}{2-3.5/G}, \text{rate} = \frac{1-1/G}{2-3.5/G}\right)$$

It is clear that for large G, the sampling distribution of H approaches chi square distribution with one degree of freedom.

2.3. Graphical Presentation

The Kruskal-Wallis test is plotted as

$x_{axis} = g$ versus $y_{axis} = H_g$, with decision limit DL

If any point outside the decision limit, H_0 is rejected and this will identify where stochastic dominance occurs.

Since the test is written as sum of independent chi square random variables

$$KW = \sum_{g=1}^{G}\left(\frac{\bar{r}_g - 0.5(n+1)}{\sqrt{\frac{n(n+1)}{12n_g}}}\right)^2$$

and each term has almost the same distribution. Rather than working with whole distribution, it can do the test based on

$$H_g = \left(\frac{\bar{r}_g - 0.5(n+1)}{\sqrt{\frac{n(n+1)}{12n_g}}}\right)^2 \quad \text{for} \quad g = 1,..,G$$

with adjusted α using Bonferonni approximation. The advantage of this (a) it can tell which group is out of the limit (b) it can easily be used to compute the effect size.

To find the decision limit for H_g, there are multiple tests ($G-$tests) and it is needed to distinguish between two meanings of α when performing multiple tests:

1. The probability of making a Type I error when dealing only with a specific test. This probability is denoted $\alpha[PT]$ ("alpha per test"). It is also called the test-wise alpha.
2. The probability of making at least one Type I error for the whole family of tests. This probability is denoted $\alpha[PF]$ ("alpha per family of tests"). It is also called the family-wise or the experiment-wise alpha.

The probability of making at least one Type I error for a family of G tests is

$$\alpha(PF) = 1 - (1 - \alpha(PT))^G$$

This equation can be rewritten as

$$\alpha(PT) = 1 - (1 - \alpha(PF))^{1/G}$$

For more details; see, for example, [12] and [13].

A simpler approximation which is known as the Bonferonni approximation is

$$\alpha(PT) \approx \frac{\alpha(PF)}{G}$$

For example, to perform $G = 4$, and the risk of making at least one Type I error to an overall value of $\alpha(PF)$= 0.05, with the Bonferonni approximation, a test reaches significance if its associated probability is smaller than

$$\alpha(PT) \approx \frac{\alpha(PF)}{G} = \frac{0.05}{4} = 0.0125$$

By using the quantile function of gamma distribution (for example, R-software), the decision limit is

$$DL = \text{qgamma}\left(1 - \frac{\alpha}{G}, k, \theta\right)$$

Therefore,

$$\text{if any } H_g > DL, \text{ for } g = 1,2, \dots, G$$

H_0 is rejected. Note that this technique is related to analysis of means; see, for example, [14], [15] and [16].

Figure 1 shows the graphical presentation of Kurskal-Wallis test for simulated data from normal and exponential distributions using $G = 4$ and total sample sizes 40.

Figures 1 (a) and (c) show the H graph for data simulated from normal and exponential distributions with no shift in mean and it clear that no points outside the decision limit while Figures 1 (b) and (d) show the H graph for data simulated from normal and exponential distributions with shift in mean and is clear that there are points outside the decision limit.

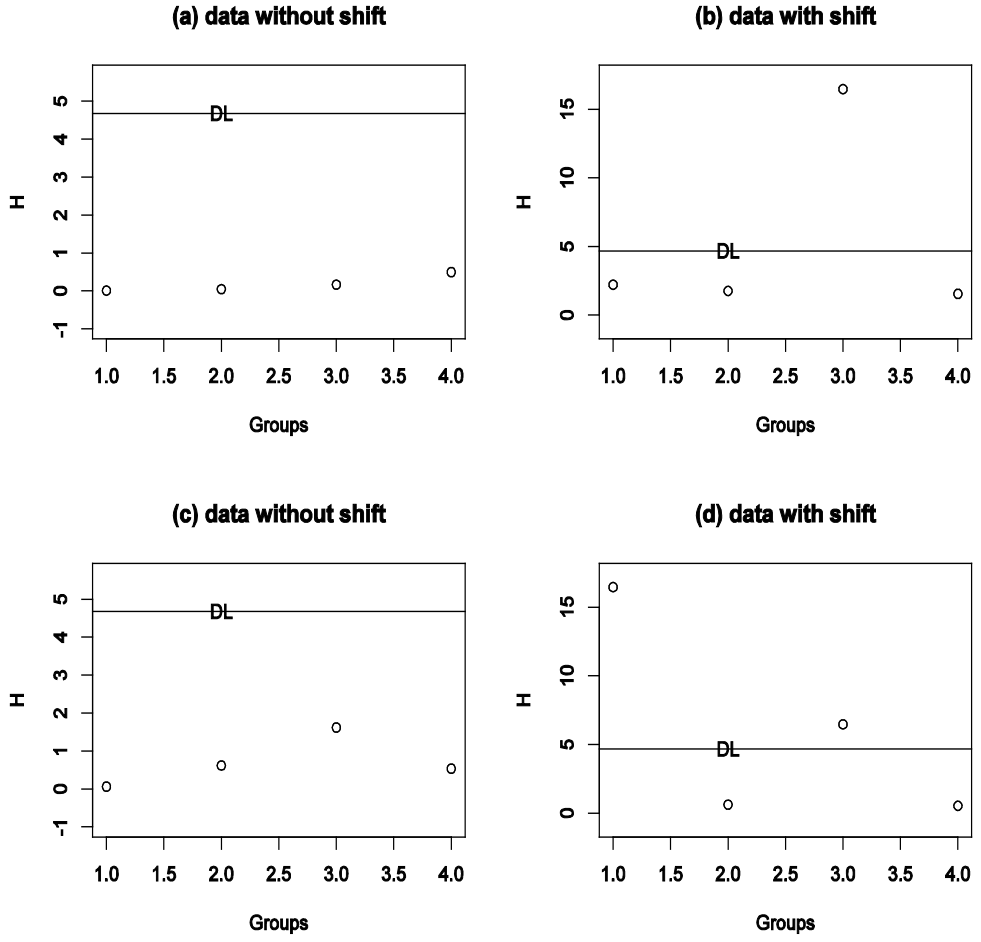

Figure 1. H-graph with decision line for $G = 4$ and $n = 40$ using simulated data: (a) normal with equal means (100,100,100,100), (b) normal with shift in mean (100,100,90,100), (c) exponential with means (1,1,1,1,1) and (d) exponential with shift in mean (9,1,1,1)

3. Comparisons

The new method is compared with KW test in terms of Type I error. Two variables were manipulated in the study: (a) number of groups (3 and 4) and (b) sample size (small-medium-large). For each design size, three sample size cases were investigated. In our designs, the smallest of the three cases investigated for each design has an average group size of less than 10, the middle has an average group size less than 20 while the larger case in each design had an average group size less than 30. With respect to the effects of distributional shape on Type I error, the normal and exponential distributions were selected.

To evaluate the particular conditions under which a test was insensitive to assumption violations, the idea of [17] of robustness criterion was employed. According to this criterion, in order for a test to be considered robust, its empirical rate of Type I error α must be contained in the interval $\alpha \pm \varepsilon$. The choice of Bradley was $\varepsilon = 0.025$ and this makes the interval is liberal. Therefore, in this study the choice of $\varepsilon = 0.015$ a something in the middle between nothing and 0.025. Therefore, for the five percent level of significance, a test was considered robust in a particular condition if its empirical rate of Type I error fell within the interval $0.035 \leq \hat{\alpha} \leq 0.065$ and for the one percent level of significance the choice of $= 0.01$, a test was considered robust if its empirical rate of Type I error fell within the interval $0 \leq \hat{\alpha} \leq 0.02$. Correspondingly, a test was considered to be nonrobust if, for a particular condition, its Type I error rate was not contained in these intervals. Nonetheless, there is no one universal standard by which tests are judged to be robust, so different interpretations of the results are possible.

Tables 3 and 4 contain empirical rates of Type I error for a design containing three and four groups, respectively. The tabled data indicates that

1. When the observations were obtained from normal distributions, rates of Type I error were controlled by KW and H methods for equal and non equal sample sizes.
2. When the observations were obtained from non-normal distributions, rates of Type I error were controlled by KW and H methods for equal and non equal sample sizes.

Table 3. Empirical Type I error for Kruskal-Wallis (KW) and H_g tests for $G = 3$ using simulated data from normal and exponential distributions and the number of replications is 10000

$G = 3$								
	Normal				Exponential			
α	0.05		0.01		0.05		0.01	
n_g	KW	H	KW	H	KW	H	KW	H
(10,10,10)	0.044	0.043	0.0078	0.0083	0.046	0.045	0.0072	0.0083
(20,20,20)	0.048	0.047	0.0087	0.0103	0.047	0.046	0.0080	0.0091
(50,50,50)	0.051	0.048	0.0088	0.0095	0.049	0.048	0.0093	0.0106
(5,8,9)	0.047	0.046	0.0061	0.0075	0.044	0.045	0.0057	0.0083
(10,13,19)	0.048	0.049	0.0080	0.0100	0.045	0.048	0.0080	0.0095
(20,25,30)	0.051	0.051	0.0095	0.0101	0.050	0.049	0.0086	0.0110

Table 4. Empirical Type I error for Kruskal-Wallis (KW) and H_g tests for $G = 4$ using simulated data from normal and exponential distributions and the number of replications is 10000

$G = 4$								
	Normal				Exponential			
	α							
	0.05		0.01		0.05		0.01	
n_g	KW	H	KW	H	KW	H	KW	H
(10,10,10,10)	0.045	0.044	0.0070	0.0070	0.047	0.047	0.0073	0.0072
(20,20,20,20)	0.046	0.044	0.0102	0.0095	0.048	0.045	0.0080	0.0071
(50,50,50,50)	0.050	0.049	0.0101	0.0111	0.051	0.049	0.0101	0.0102
(5,7,9,10)	0.041	0.040	0.0059	0.0060	0.042	0.040	0.0056	0.0055
(10,13,16,20)	0.044	0.041	0.0070	0.0078	0.049	0.048	0.0082	0.0081
(20,25,28,30)	0.049	0.048	0.0090	0.0089	0.051	0.051	0.0102	0.0103

4. Application

From [18] a study of the use of a semi automated method for measuring the amount of chlopheniramine maleate in tablets, for each of four manufacturers, composites were prepared by grinding and mixing together tablets that had nominal dosage levels of 4 mg. Seven labs were asked to make 10 determinations on each composite; each determination was made on a portion of composite whose weight was equivalent to that of one tablet. The purpose of the experiment: are the differences in the means or medians of the measurement from the various labs significant, or might they be due to chances?

Table 5 gives the amount of chlopheniramine maleate in tablets for seven labs.

Using the Kruskal-Wallis it is found that

$$KW = 23.68, \quad \chi^2_{1-0.05}(df = 6) = 12.59, \text{ and } \chi^2_{1-0.01}(df = 6) = 16.81$$

This indicates that there is a systematic difference among the labs at 0.05 and 0.01 level of significance without telling anything about the differences.

While H-graph in Figure 2 shows that there are two points outside the decision limit that indicates there is a systematic difference among the labs at 0.05 and 0.01 level of significance. Moreover identify the labs 1 and 4 as different from the overall mean.

Table 5. The amount of chlopheniramine maleate in tablets for seven labs

Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	Lab 6	Lab 7
4.13	3.86	4.01	3.87	4.13	4.12	4.00
4.07	3.85	4.02	3.86	3.95	3.86	4.07
4.04	4.08	4.03	3.91	4.02	3.96	4.09
4.07	4.11	4.04	3.95	3.89	3.97	4.08
4.05	4.19	3.99	3.92	3.91	4.00	4.10
4.04	4.01	4.04	3.97	4.01	3.82	3.81
4.02	4.02	4.03	3.90	3.89	3.98	3.91
4.06	4.04	3.96	3.89	3.99	3.99	3.96
4.10	3.97	3.97	3.98	4.00	4.02	4.05
4.04	3.95	3.98	3.93	3.88	3.93	4.06

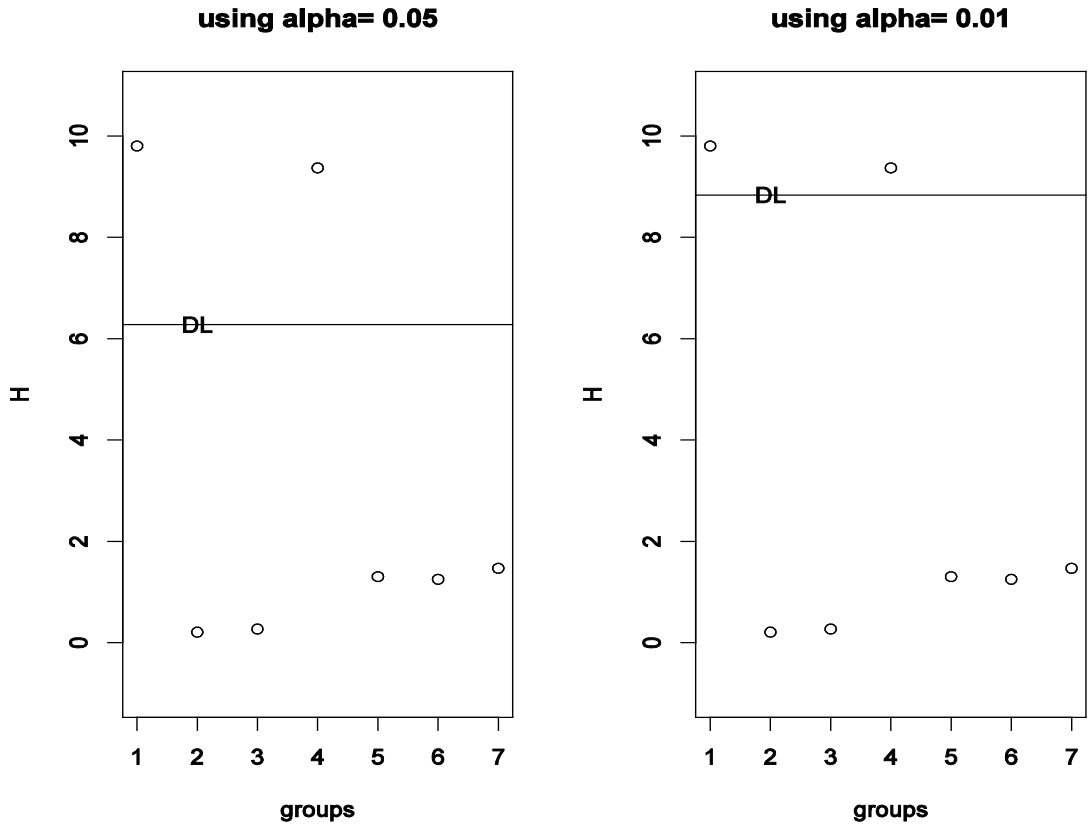

Figure 2. H-graph for the amount of chlopheniramine maleate in tablet data

5. Conclusions

A graphical presentation of Kruskal Wallis test is studied by re-expressing the test as sum of independent chi square random variables. The sampling distribution for this function was obtained and found that the gamma distribution had given a very good fit for this function until for small sample sizes.

A decision limit is obtained using the gamma distribution and Bonferonni approximation that enabled us to show Kruskal-Wallis test graphically. The main advantage of the new method is not only provided us about the differences in medians or means but also where these differences had been happened. Moreover, a simulation study confirmed the validity and robustness of the decision limit in small and large samples in comparison with Kruskal Wallis test.

REFERENCES

[1] Kruskal, W. H. & Wallis, W. A., 1952, Use of ranks. *Journal of American Statistical Association,* 47*,* 583-621.

[2] Micceri, T., 1989, The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin,105,* 156-166.

[3] Brown, M. B. & Forsythe, A. B., 1974, The small sample behavior of some statistics which test the equality of several means. *Technometrics 16,* 129-132.

[4] Iman, R. L. & Davenport, J. M., 1976, New approximations to the exact distribution of the Kruskal-Wallis test statistic. *Communications in Statistics - Theory and Methods, 5,* 1335-1348.

[5] Conover, W. J. & Iman, R. L., 1981, Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124-129.

[6] Hettmansperger, T. P., 1984, *Statistical inference based on ranks.* Wiley: New York.

[7] Sugiura, N., Murakami, H., Lee, S.K. and Maeda, Y., 2006, Biased and unbiased two-sided Wilcoxon tests for equal sample sizes. Annals of Institute Statistics and Mathematics, 58, 93–100.

[8] Acar, E., F. and Sun L., 2013, A Generalized Kruskal–Wallis test incorporating group uncertainty with application to genetic association studies. Biometrics, 69, 427-35.

[9] Iman, R.L., 1994, *A data-based approach to statistics*, Belmont, California: Duxbury Press, Wadsworth Publishing Company.

[10] Kruskal, W. H., 1952, A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics, 23,* 525-540.

[11] Rice, A.J., 1995, *Mathematical Statistics and Data Analysis*. 2nd Edition, Duxbury Press.

[12] Dunn, O.J., 1964, Multiple comparisons using rank sums. *Technometrics*, 6, 241-252.

[13] Abdi, H., 2007, The Bonferonni and Šidák corrections for multiple comparisons. In: Neil Salkind. *Encyclopedia of Measurement and Statistics*. Thousand Oaks (CA): Sage.

[14] Ramig, P. F., 1983, Applications of the analysis of means. *Journal of Quality Technology*, 15, 399–406.

[15] Nelson, L. S., 1983, Exact critical values for use with the analysis of means. *Journal of Quality Technology,* 15(1):40–44.

[16] Nelson, P.R., Wludyka, P.S., and Copeland, A.F., 2005, *The analysis of means: A graphical method for comparing means, rates, and proportions*. Society for Industrial and Applied Mathematics.

[17] Bradley, J.V., 1978, Robustness?, *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.

[18] Kirchoefer, R., 1979, Semi automated method for the analysis of chlopheniramine maleate tables: collaborative study. *Journal of Association of Official Analysis Chemistry*, 62, 1197-1120.