Update

Learning and Understanding the Kruskal-Wallis One-Way Analysis-of-Variance-by-Ranks Test for Differences Among Three or More Independent Groups

When several treatment methods are available for the same problem, many clinicians are faced with the task of deciding which treatment to use. Many clinicians may have conducted informal "mini-experiments" on their own to determine which treatment is best suited for the problem. These results are usually not documented or reported in a formal manner because many clinicians feel that they are "statistically challenged." Another reason may be because clinicians do not feel they have controlled enough test conditions to warrant analysis. In this update, a statistic is described that does not involve complicated statistical assumptions, making it a simple and easy-to-use statistical method. This update examines the use of two statistics and does not deal with other issues that could affect clinical research such as issues affecting credibility. For readers who want a more in-depth examination of this topic, references have been provided.[1–5]

The Kruskal-Wallis one-way analysis-of-variance-by-ranks test (or H test) is used to determine whether three or more independent groups are the same or different on some variable of interest when an ordinal level of data or an interval or ratio level of data is available.[1] A hypothetical example will be presented to explain when and how to use this statistic, how to interpret results using the statistic, the advantages and disadvantages of the statistic, and what to look for in a written report. This hypothetical example will involve the use of ratio data to demonstrate how to choose between using the nonparametric H test and the more powerful parametric F test.

[Chan Y, Walmsley RP. Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. *Phys Ther.* 1997;77:1755–1762.]

Key Words: *Analysis of variance, Kruskal-Wallis, Ordinal data, Ranks.*

Yvonne Chan

Roy P Walmsley

Hypothetical Example

Suppose that a person wanted to know whether a difference existed in the effectiveness of three different exercise programs in increasing the range of knee flexion after cast immobilization, the outcome of which was measured in degrees by use of an inclinometer. In this single-factor design, 18 subjects were randomly assigned to one of three different groups, with each subject undergoing only one type of treatment. The purpose of this hypothetical study was to determine whether there was a significant difference at an alpha level of .05 in the effectiveness of the three treatment types in increasing the range of knee flexion. Looking at the purpose more carefully, the question asked is: Are the three samples really different, or are the differences found merely reflecting the variations to be expected from random sampling from the same population? That is, are any differences found between the groups genuine, or are they occurring by chance? If these differences are genuine, which treatment is superior to the other treatment methods? The null hypothesis (H_0) stipulates that there are no differences among the three samples. The Kruskal-Wallis statistic answers these questions by comparing the form of the sample curve with the form of the population curve. This concept of comparing curve forms is the basis of the H test. Neither the nonparametric H test nor the parametric F test, however, demonstrates than an obtained difference is meaningful or worthwhile.

The Concept of Comparing Curve Forms

The concept of comparing sample and population curves to determine whether they have the same form is important to understanding the hypotheses guiding the H test and is illustrated in Figures 1 through 3. Figure 1 illustrates the null hypothesis that there is no difference between the sample distribution and the distribution of the populations from which the sample was derived. Both distributions have the same curve form, differing only by a translation. This sample variation has been compared with the population variation without making any assumptions about what the real population curve is like with respect to its definition on the x and y axes. Figure 2 exemplifies the case when significant differences are detected among three treatment groups, but these differences could still be in the same form as the population differences. Figure 3 shows the distributions when significant differences are detected among three treatment groups and these differences are significantly different from those of the originating population. Figure 3 exemplifies what the Kruskal-Wallis statistic aims to detect: genuine differences between the sample curve and the population curve without making any assumptions as to the original distribution of the population. Not knowing the exact form of the original population is the most basic assumption behind the H test and is the fundamental difference between the H test and the F test.

The purpose of the H test is to look for the same form of distribution between samples and the population from which they came.

The Usual Solution: The Parametric F Test

The usual technique of solving the hypothetical problem presented is through an analysis of variance with a single criterion of classification, better known as the F test. For a comprehensive guide on this technique, the reader is referred to the article by Norton and Strube.[6] Although the H test is an analog to the F test, the two tests are based on different assumptions and have different purposes. Although the F test uses the variation among the sample means to estimate the variation among individuals, the H test uses variation among ranked sample means. The F test assumes that the population is approximately normally distributed, with the population variance being σ^2. With the H test, only general assumptions are

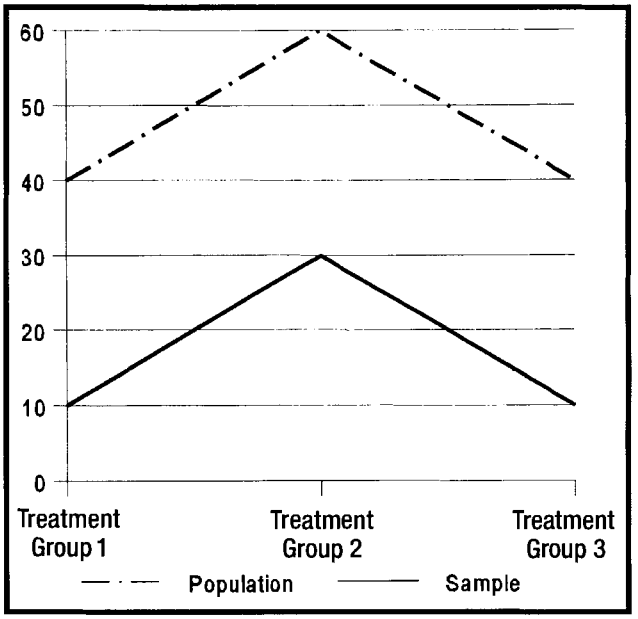

Figure 1.
The distribution of the population and the distribution of the sample have approximately the same form.

Y Chan, BSc(PT), is a student in the Master of Science degree program, School of Rehabilitation Therapy, Queen's University, Kingston, Ontario, Canada K7L 3N6 (ychan@cbi.ca). Address all correspondence to Ms Chan.

RP Walmsley, PhD, is Professor, Division of Physical Therapy, School of Rehabilitation Therapy, Queen's University.

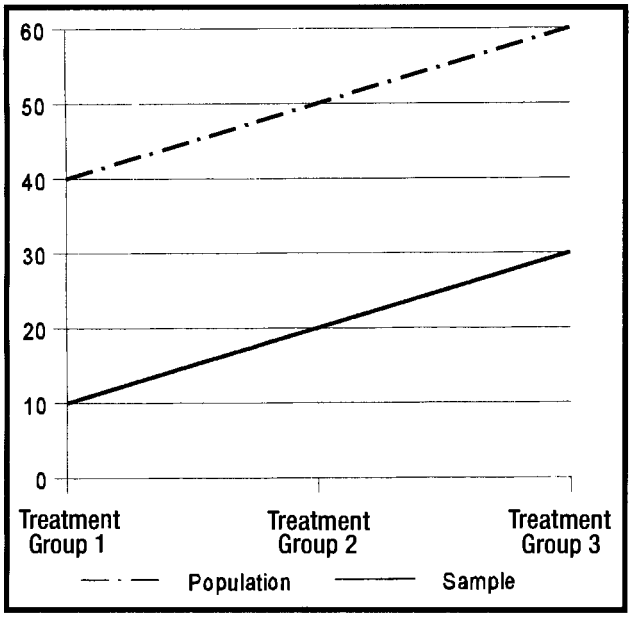

Figure 2.
Differences observed in the sample are of the same form as differences observed in the population.

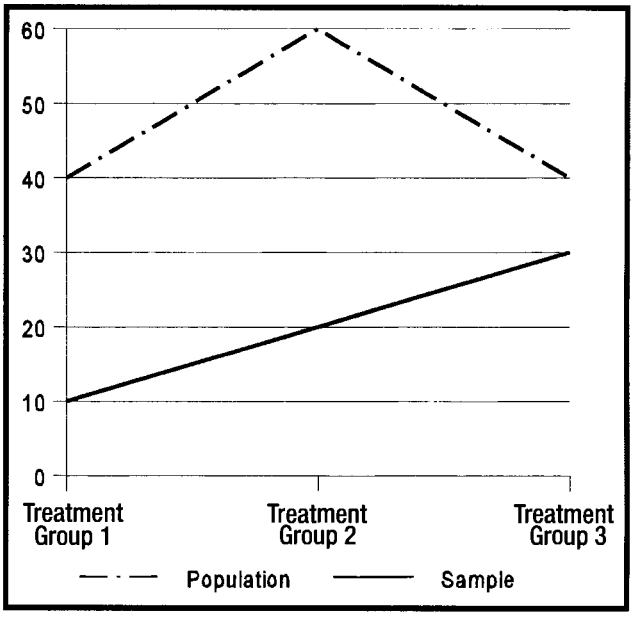

Figure 3.
Differences observed between treatment groups 2 and 3 reflect true differences between the sample and the population.

made regarding the population distribution, and this distribution can be arbitrary.[7] An arbitrary population distribution can be assumed because the null hypothesis related to the H test only looks for the same form and not specific definitions. The F-test calculations all depend on the assumption of a normal distribution, whereas the H-test calculations do not depend, in general, on the population distribution.

Finally, the F ratio tests whether the group means for a dependent variable (eg, outcome) differ significantly after exposing each group to a single factor or independent variable (eg, treatment). This test determines whether the independent variable contributed to the variability in the dependent variable, thus effecting differences among the group means.[3] In contrast, the H test assumes that the response to the independent variable is unaffected by some factor that can be disregarded.[2] For example, in a study of tire wear, it may not be necessary to take into account the make of the car on which the tires are mounted. In our example of three different treatment programs for knee flexion where each program contains a number of different exercises, it may not be necessary to distinguish among the exercises within each group, the characteristics of the patients, or their initial available range of knee flexion. This example reiterates the fact that the researcher does not attempt to define the population distribution beforehand, and this feature is what makes the H test an attractive alternative to the clinician who may not be able to fully define the population but who still wants to make general conclusions about three (or more) groups.

The differences between the parametric F test and the nonparametric H test can be summarized as follows. The F test attempts to isolate the source of variance (difference), whereas the H test disregards this variable. The F test assumes known population variances of approximately normal distribution and the population variances are homogeneous, whereas the H test makes only very general assumptions related to the distributions' source. The F test depends on the assumption that the population is normally distributed, whereas the H test does not depend on the shape of the population distribution, which can be arbitrary. Besides being easier to use and understand, the H test makes fewer assumptions about the population being studied than does the F test.

Advantages of Ranks

The H test can be used to answer all of the questions in the hypothetical example by replacing the actual data obtained from the clinical observations with rankings. Four advantages of ranking data in statistical analyses instead of using the original observations are (1) the calculations are simplified, (2) only very general assumptions are made about the kind of distributions from which the observations arise, (3) data available only in ordinal form often may be used, and (4) when the assumptions of the parametric test procedure are too far from reality, not only is there a problem of distribution theory if the usual test is used but it is possible that the usual test may not have as good a chance as a rank test of detecting the kinds of differences of real interest. If there are multiple samples, the mean ranks for any of them are jointly distributed approximately according to a multivariate normal distribution, provided that the sample sizes are not too small. The Appendix provides a

Table 1.
Hypothetical Increases in Range of Knee Flexion (in Degrees) of Subjects (N=18) After Completing One Treatment Regimen

Treatment Group 1	Treatment Group 2	Treatment Group 3
44	70	80
44	77	76
54	48	34
32	64	80
21	71	73
28	75	80

Table 2.
Ranked Hypothetical Increases in Range of Knee Flexion (in Degrees) of Subjects (N=18) After Completing One Treatment Regimen, with R Being the Sum of the Ranks in the ith Sample

	Treatment Group 1	Treatment Group 2	Treatment Group 3
	5.5	10	17
	5.5	15	14
	8	7	4
	3	9	17
	1	11	12
	2	13	17
ΣR_i	25	65	81
$\bar{R}_i$	4.17	10.83	13.50

glossary of the terms and abbreviations used in this article.

When only an ordinal level of data is available, the Kruskal-Wallis technique can be used if data are obtained in the form of ranks. However, the advantages of ranks can explain why the H test may be chosen over the F test when data are of the interval or ratio type. Because only very general assumptions are made, the H test assumes that the observations are all independent, that those observations within a given sample arise from a single population, and most importantly, that the multiple samples are of approximately the same distribution. Although the disadvantage to using ranks is a loss of information related to the spread of the data, the use of ranks suits the H-test assumptions very well in that it does not seek to define a population.

The Nonparametric H Test

Given multiple samples (C), with n_i observations in the ith sample, the H statistic tests the null hypothesis that the samples come from identical population distributions. This hypothesis is tested by ranking the observations from 1 to N (giving each observation in a group of ties the mean of the ranks tied), finding the C sum of ranks, and computing an H statistic. This statistic is then compared with a tabled value for the H statistic. This comparison will determine whether the null hypothesis is accepted or rejected. Visual inspection of the raw data (Tab. 1) shows that the patients in treatment group 1 appear to have less average knee flexion than the patients in the other two groups and that the patients in treatment group 3 appear to have the greatest average knee flexion. By computing the H statistic, it can be determined whether the alternate hypothesis that the three treatment groups are different is correct.

Computing the H Statistic

The data for the hypothetical example are first placed into a two-way table (Tab. 1) and then ranked (Tab. 2). In this example, the total number of samples is 3 and N is 18. Each of the N observations is replaced by a rank relative to all the observations in all of the samples. The lowest score is often replaced by rank 1, the next lowest score is replaced by rank 2, and so on. Ranking from high to low is also possible if that is more appropriate to the question being asked. For ties in the scores, the tied observations are assigned the average of the ranks that would be assigned if there were no ties. The sum of ranked scores (ΣR_i) in each column is then found. The mean of the ranks in each group ($\bar{R}_i$) is found by dividing the sum of ranks by n_i, the number of observations in each group. The R_i and $\bar{R}_i$ are now compared for their overall closeness because if the treatments differ widely among each other, large differences among the values would be expected.[2] If adjacent ranks are well distributed among all of the samples, which would be true for a random sample from a single population, the total sum of ranks would be divided proportionally according to sample size among the multiple samples.[8] The criterion for measuring the closeness of R_i and $\bar{R}_i$ is a weighted sum of the squared differences $[R_i - \frac{1}{2}(N+1)]^2$, which is incorporated into the defining equation to compute the H statistic (Tab. 3). The weights in this statistic were chosen to provide a simple approximation to the null distribution when the n_i are large.[1,2] Following the defined H-test equation, the H statistic is calculated to be 9.73 (Tab. 3).

Correction for Ties

In the matter of the correction factor for ties (Tab. 3), it should be noted that if the statistic is significant at the desired level of alpha (.05) without the adjustment, there is no point in using this correction factor. With 10 or fewer samples, an H-statistic value of 0.01 or more does not change more than 10% when the adjusted value is computed, provided that not more that one fourth of the observations are involved in ties.[1] If H_o is rejected with the first value obtained, it will also be rejected with the corrected value. Thus, even though 5 of the 18 observations were involved in ties, the correction factor produced a very small change in the final value of the H statistic: the uncorrected value of 9.73 was corrected to 8.76 (Tab. 3).

Interpretation of the Results

The final value of the H statistic is now used to determine whether to accept or reject the null hypothesis. Depending

Table 3.
Computations Based on Defining Equations[a]

The H-Test Equation:

$$H = \frac{12}{N(N+1)} \Sigma_{i=1}^{C} \frac{R_i^2}{n_i} - 3(N+1)$$

$$H = \frac{12}{18(18+1)}\left[\frac{25^2}{6} + \frac{65^2}{6} + \frac{81^2}{6}\right] - 3(18+1)$$

$$= 9.730952$$

Correction Factor for Ties:

$$1 - \frac{\Sigma_{i=1}^{g}(t_i^3 - t_i)}{N^3 - N} = 1 - \frac{(2^3-2)+(3^3-3)}{(18^3-18)} = 0.995$$

Corrected Value of the H Statistic for Ties:

$$H = \frac{H}{1 - \Sigma T/(N^3-N)}$$

$$= \frac{9.73 - 0.995}{0.995}$$

$$= \frac{8.73}{0.995}$$

$$= 8.76$$

Tabled Value of the H Statistic as Compared With Computed Value of the H Statistic:

In this example, reference to the χ^2 table indicates that a value of the H statistic of ≥5.99 with $df = 3-1=2$ has a probability of occurrence when H_0 is true of $P < .05$. Thus, because the observed value of the H statistic (9.73) exceeds the tabled value of the H statistic (5.99), the null hypothesis is rejected.

Conclusion:

It can be concluded that there are differences among the three treatment groups relative to the change in the dependent variable (outcome).

[a] Refer to Appendix for explanation of symbols.

on the conditions of the study, a decision rule is made by comparing the computed value of the H statistic with the tabled value of the H statistic (Tab. 4). If there are more than five observations in each sample, the H statistic has been shown to be distributed approximately as a chi-square distribution (with $df = C-1$) and therefore chi-square tables are used for the comparison.[7,9,10] If the samples have fewer than five observations, special approximations through exact tables, called the "critical values" for the H statistic, are used.[1,8] Both tables can be found in statistic textbooks.[2–5]

In this example, because there are more than five observations in each group, chi-square tabled values have been used for the comparison. Results are significant and the null hypothesis is rejected if the computed value of the H statistic is larger than the tabled value of the H statistic. Reference to the chi-square table indicates that an H-statistic value of ≥5.99 (with $df = 3-1=2$) has a probability of occurrence when H_0 is true of $P < .05$. Both the corrected value of the H statistic (9.73) and the uncorrected value of the H statistic (8.76) exceed 5.99, and the hypothesis of no differences is thus rejected. It can therefore be concluded that there are differences in increasing the range of knee flexion among the three treatment groups.

Table 4.
Indications for Making a Decision Rule on the Computed H Statistic[a]

Condition	Table to Use	Decision Rule
$i=3$ or more groups Number of observations in each group exceeds 5	Chi-square tabled values for $df = C-1$	If observed value of the H statistic is ≥ tabled value, reject H_0
$i=3$ or more groups Number of observations in each group is less than or equal to 5	Kruskal-Wallis critical values table	If observed value of the H statistic is ≥ tabled value, accept H_0

[a] Refer to Appendix for explanation of symbols.

Comparison After the H Statistic

When the obtained value of the H statistic is statistically significant, it indicates that at least one of the groups is different from the others. It does not indicate, however, which groups are different or whether the difference is meaningful, nor does it specify how many of the groups are different from each other. This next procedure, called "multiple comparisons between treatments," constructs pair-wise multiple comparisons to locate the source of significance.[3] This procedure tests the null hypothesis that some groups u and v are the same against the alternate hypothesis that some groups u and v are different (Tab. 5). When the sample size is large, these differences are approximately normally distributed. Because there are a large number of differences and because the differences are not independent, however, the comparison procedure must be adjusted appropriately. An inequality is used, and the hypothesis of no difference among the three groups is tested at the alpha level of significance of .05. The null hypothesis is rejected if the calculated difference among groups is greater than the critical difference. In this example, the difference between treatment groups 1 and 3 (ie, 9.33) is the only difference that is greater than the critical difference (ie, 7.38). It can therefore be concluded that treatment 3 led to a different, and in this case better, result because it provided a greater increase in the range of knee flexion than did treatment 1. Furthermore,

Table 5.
Computation of Multiple Comparisons Among Treatments[a]

The hypotheses: *For some groups u and v,*

$$H_0: \theta_u = \theta_v$$
$$H_A: \theta_u \neq \theta_v$$

The number of comparisons possible in this case is computed by:

$$Number\ of\ comparisons = \frac{i(i-1)}{2} = \frac{3(3-1)}{2} = 3$$

Equality used to test the significance of individual pairs of differences:

$$|\overline{R_u} - \overline{R_v}| \geq z_{\alpha/i(i-1)} \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_u} + \frac{1}{n_v}\right)}$$

The critical difference for comparison is:

Using the table of normal distribution:

$$z_{\alpha/i(i-1)} = z_{0.05/3(3-1)} = z_{0.0083} \approx 2.394$$

The critical difference is then calculated:

$$z_{\alpha/i(i-1)} \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_u} + \frac{1}{n_v}\right)}$$

$$= 2.394 \sqrt{\frac{18(18+1)}{12}\left(\frac{1}{6} + \frac{1}{6}\right)} = 7.38$$

The differences are obtained for the three pairs of groups:

$$|\overline{R_1} - \overline{R_2}| = |4.17 - 10.83| = 6.66$$
$$|\overline{R_1} - \overline{R_3}| = |4.17 - 13.50| = 9.33$$
$$|\overline{R_2} - \overline{R_3}| = |10.83 - 13.50| = 2.67$$

The comparison:

The critical difference is compared with the differences among the average rankings for the three treatment groups. Because only the difference between treatment groups 1 and 3 (9.33) exceeds the critical value of 7.38, that difference is considered to be significant. It may be concluded that the medians between treatment groups 1 and 3 are different.

Conclusion:

Treatment 1 results in a different outcome from that of treatment 3.

[a] Refer to Appendix for explanation of symbols.

there were no real differences between the outcomes of treatments 1 and 2 and treatments 2 and 3.

Only a single critical difference was calculated for this example. This calculation was possible because the groups were equal in size (eg, each group had six observations). If the sample sizes had been unequal, each of the observed differences would have to be compared against different critical differences. In addition, if a person wanted to compare specific treatment groups with a control group, the inequality used to compare treatments should be slightly adjusted to account for the smaller number of comparisons (see statistical texts[2–5] for these methods).

Assumptions Regarding Use

The hypothetical example presented covered the four main assumptions underlying the use of the H test. First, it was assumed that the dependent variable under study (ie, range of knee flexion) had an underlying continuous distribution to avoid the problem of ties,[8] and it was measured on at least an ordinal scale. Second, the scores of the patients in one treatment group were independent of the scores of patients in the other treatment groups. Third, the patients' scores within each type of treatment setting were not influenced by any other patients' scores within the same treatment setting. The fourth assumption states that the null hypothesis is true. This assumption is rejected if variability among the means of the summed ranks is sufficiently large, as it was in this example.

Advantages of the H Test

The H test is simply an analog to the F test in that the statistic is calculated using ranked data rather than the original observations. Practical advantages of the H test are that it is simple to use, it does not require a computer to calculate, and it is widely available in applied texts.[3] Compared with the F test, the H test is quicker and easier to apply and it makes fewer assumptions of the population under study. An H statistic can be calculated for interval- or ratio-level data by transforming it to the ordinal scale through ranking. In essence, actual measurements are not required.[8] The H test may perform better than the F test if the F test's assumptions are not satisfied.[1,8]

Disadvantages of the H Test

The H test is not able to single out differences if the null hypothesis is rejected. In addition, this technique tests only for differences that are collectively significant. If two samples are then singled out for comparison, the usual problem of unknown overall probabilities for Type I and II errors results.[8] This technique can only be used when looking for differences among three or more independent samples. The H test is distributed approximately as a chi-square distribution (with $df = C-1$) only when the null hypothesis holds[9] or when there are large numbers of observations (more than five) in each sample.[7,9,10]

Power—Efficiency and Consistency

An important aspect of any statistic is its power identify when to reject the hypothesis tested when the given alternative is true (ie, a Type II error). Compared with the most powerful parametric test for comparison among several

means (ie, the F test), the H test has an asymptotic relative efficiency of $3/\pi=0.955$.[7] *Asymptotic relative efficiency* means that, when compared with the F test, the H test has a 95.5% chance of choosing a sequence of alternative hypotheses that vary with the sample sizes in such a manner that the powers of the two tests for this sequence of alternatives have a common limit of less than 1.[7]

In terms of consistency, the H test is consistent only if the variables from at least one population tend to be either larger or smaller than the other variables.[9] Furthermore, the test based on large values of the H statistic has been shown to be consistent against the given alternative.[9]

Compared with using the extension of the median test in testing ordinal data for three or more groups, the H test is more efficient because it utilizes more of the information in the observations, converting the scores into ranks rather than simply dichotomizing them as above or below the median.[5,8] For other alternatives to the H test, the reader is referred to a discussion in the textbook by Daniel.[11]

Essentials of Written Reports

In reporting the results of a study analyzed with the H test, we believe that the introductory section of the report should include the null hypothesis that the three or more groups examined are considered to be from identical populations. If the data presented are interval- or ratio-scale data, the author should explain why the parametric F test was not used. The method section should clearly outline that the study involved the use of random sampling and independent groups. If there were fewer than five observations in each group, we contend that the author needs to explain this limitation, because more than five observations increase the power of the test. There should not be repeated measures on the subjects, nor should the author attempt to partition the score variations on the dependent variable into components. The purpose of the H test is to look for the same form of distribution between the samples and the population from which they came. If this distribution is different, then a comparison must be made afterward to determine where the differences in the medians lie. Although there is no consistent way of summarizing the computations in a table, the results section should include the mean of the summed ranks of each group, the exact H statistic, the table used for comparison of values of the H statistic, the degrees of freedom, the results of comparison analysis, and a concluding statement as to which groups were found to be significantly different.

Summary

Through the use of a hypothetical example in this article, we reviewed the rationale, indications, method, and interpretation of the Kruskal-Wallis one-way analysis-of-variance-by-ranks test. A hypothetical example was used to clarify the instances in which a person would choose to use the nonparametric H test over the parametric F test. The Kruskal-Wallis technique, based on ranks, tests and estimates whether the samples and the population from which they came have the same curve form. A method was presented to show clinicians how to compare sample and population distributions in order to test for differences among three or more treatment groups when at least an ordinal level of data is available.

References

1 Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association.* 1952;47:583–621. Addendum. 1953;48:907–911.

2 Lehmann EL. *Nonparametrics: Statistical Methods Based on Ranks.* New York, NY: McGraw-Hill Inc; 1975.

3 Hettmansperger TP. *Statistical Inference Based on Ranks.* New York, NY: John Wiley & Sons Inc; 1984.

4 Hollander M, Wolfe DA. *Nonparametric Statistical Methods.* New York, NY: John Wiley & Sons Inc; 1973.

5 Siegel S, Castellan NJ. *Nonparametric Statistics for the Behavioral Sciences.* New York, NY: McGraw-Hill Inc; 1988.

6 Norton BJ, Strube MJ. Guide for the interpretation of one-way analysis of variance: readings tips. *Phys Ther.* 1985;65:1888–1896.

7 Andrews FC. Asymptotic behaviour of some rank tests for analysis of variance. *Annals of Mathematical Statistics.* 1954;25:724–736.

8 Gibbons JD. *Nonparametric Statistical Inference.* New York, NY: McGraw-Hill Inc; 1971.

9 Kruskal WH. A nonparametric test for the several sample problem. *Annals of Mathematical Statistics.* 1952;23:535–540.

10 Gabriel KR, Lachenbruch PA. Non-parametric ANOVA in small samples: a Monte Carlo study of the adequacy of the asymptotic approximation. *Biometrics.* 1969;25:593–596.

11 Daniel WW. *Applied Nonparametric Statistics.* Boston, Mass: Houghton Mifflin Co; 1978:204–205.

Appendix.

Glossary of Terms and Abbreviations

Term	Abbreviation
Null hypothesis	H_0
Alternate hypothesis	H_A
Population variance	σ^2
Total number of samples in the study	C
The sample number in the study (eg, 1, 2, or 3)	i
The number of observations in the ith sample	n_i
The sum of all observations in all samples combined	N
Sum of ranks	R_i
Mean of the sum of ranks in each group	$\bar{R}_i$
Sigma: the instruction to sum scores	Σ
Number of groups of ties	g
Number of tied ranks in the ith grouping	t_i
$(t_i^3-t_i)$	T
Chi square	χ^2
Degrees of freedom: number of elements free to vary	df
Probability value	P