The Kruskal-Wallis Test and Stochastic Homogeneity

András Vargha
Eötvös Loránd University, Budapest, Hungary

Harold D. Delaney
The University of New Mexico, Albuquerque, New Mexico, USA

Running head:
KRUSKAL-WALLIS TEST AND STOCHASTIC HOMOGENEITY

Published
in the Journal of Educational and Behavioral Statistics
1998, Vol. 23, No. 2. pp. 170-192.

Author Notes

Much of the work reported herein resulted from the collaborative efforts of Drs. Vargha and Delaney while Vargha was at the University of New Mexico under the support of a Fulbright grant from the U.S. government, and while Vargha was supported by Hungarian grant OTKA No. T018353.

Abstract

For the comparison of more than two independent samples the Kruskal-Wallis H test is a preferred procedure in many situations. However, the exact null and alternative hypotheses as well as the assumptions of this test do not seem to be very clear among behavioral scientists. This paper attempts to bring some order to the inconsistent, sometimes controversial treatments of the Kruskal-Wallis test. First it is clarified that the H test cannot detect with consistently increasing power any alternative other than exceptions to stochastic homogeneity. It is then shown by a mathematical derivation that stochastic homogeneity is equivalent to the equality of the expected values of the rank sample means. This finding implies that the null hypothesis of stochastic homogeneity can be tested by an ANOVA performed on the rank transforms. The Kruskal-Wallis H test is a procedure of this kind. If the variance homogeneity condition does not hold then it is suggested that robust ANOVA alternatives performed on ranks be used for testing stochastic homogeneity. Generalizations are also made with respect to Friedman's G test.

Key words: measure of stochastic superiority, stochastic equality, stochastic homogeneity, ANOVA, nonparametric ANOVA, Kruskal-Wallis H test, Friedman's G test, Mann-Whitney test.

The Kruskal-Wallis H test (hereafter abbreviated as KWt) is a nonparametric statistical procedure frequently used to compare several populations. However, current statistical textbooks written for the behavioral sciences are quite inconsistent or unclear about what aspects of the populations can really be compared by the KWt and under what conditions.

Regarding the null and alternative hypotheses of the KWt, several authors (e.g., Lehman, 1991; Howell, 1992; Hurlburt, 1994; Pagano, 1994; Triola, 1995, Wilcox, 1996) consider the null hypothesis to be that the distribution of the dependent variable is the same in the different populations to be compared, and usually add that the test is most powerful against alternatives of different location parameters. According to Hurlburt (1994, pp. 436-437), the alternative hypothesis of the KWt is that the distributions of the scores in the different populations are not identical. Howell, however, asserts that nonparametric comparison tests are Amore sensitive to medians than to means (1992, p. 610). On the other hand, there are several authors (e.g., Lehman, 1991; Pagano, 1994; Welkowitz, Ewen & Cohen, 1991) who emphasize that the KWt can be used as a substitute for ANOVA, and thereby they seem to imply that the alternative hypothesis of the KWt pertains to the inequality of population means. For example, Pagano writes the KWt Ais used as a substitute for the parametric one-way ANOVA (1994, p. 465). Iman (1994) goes even further and formulates the null hypothesis in terms of the expected values:

$$H_{0:}\ \mu_1 = \mu_2 = \ldots = \mu_k$$

(1994, p. 726).

These types of textbooks are even more inconsistent regarding the statistical assumptions or side conditions of the KWt. Many books (like Jaccard & Becker, 1990; Pagano, 1994, Runyon & Haber, 1991; Spence, Cotton, Underwood, & Duncan, 1990, Wilcox, 1996) provide only

minimal discussion of assumptions. These authors introduce nonparametric comparison tests such as the KWt or Mann-Whitney by saying only that they do not make the same assumptions of the parametric alternative, sometimes with no hint of what assumptions remain critical. For example Jaccard and Becker comment AFor all of the tests to be described, we will assume that the assumptions of the parametric tests have been violated to the extent that nonparametric tests are required (1990, p. 388; see also Spence et al., 1990, p. 313).

Most troubling are introductory treatments that directly contradict each other regarding the assumption of heterogeneity of variances. For example, Pagano asserts, AThe Kruskal-Wallis test does not assume population normality, nor homogeneity of variance, as does parametric ANOVA, and requires only ordinal scaling of the dependent variable. It is used when violations of population normality and/or homogeneity of variance are extreme (1994, p. 465). Similarly, Hinkle, Wiersma and Jurs (1994) comment These tests are analogous to the parametric tests that were presented..., but they are distribution-free. That is, we do not have to be concerned about whether the distribution of the dependent variable for the population is a normal distribution. We will also not be concerned with the assumption of homogeneity of variance (1994, p. 559). By contrast, Welkowitz, Ewen and Cohen introduce nonparametric comparison tests as follows: AThese methods do generally assume, however, that the populations being compared have the same shape and variability (1991, p. 307).

In the next section we provide an overview of some early theoretical literature on nonparametric statistics that helped lead to the still existing controversies regarding the null hypothesis and the side conditions of the KWt we have just outlined. We also present the shift model approach of the recent theoretical nonparametric literature, which on the other hand may be so restrictive that it will not meet the practical needs of behavioral scientists. It will be shown that the alternative hypothesis of the KWt cannot be other than that of stochastic heterogeneity, that is, the opposite of stochastic homogeneity. If the distributions are asymmetric, and have different shapes, then stochastic homogeneity is a different state of affairs than the equality of population means or medians.

In a later section of the paper we explain in detail the concept of stochastic homogeneity and some related concepts (measure of stochastic superiority, stochastic equality, stochastic ordering), and derive an equivalent mathematical expression to stochastic homogeneity.

Finally we will show that an ANOVA performed on the rank transforms of the original scores (like the KWt) is asymptotically a valid procedure for testing stochastic homogeneity, provided that the variance homogeneity condition is met. Furthermore it will be argued that robust ANOVA alternatives carried out on rank scores should be used in the case of variance heterogeneity.

The Kruskal-Wallis test in nonparametric literatures

Suppose a certain test statistic S is appropriate for testing the null hypothesis of identical population distributions. If S is significant at level α we can conclude that the distributions to be compared are different from each other. Very often we are not satisfied with only such a general

statement. We would like also to know in what way are these distributions different. Do they have different means, medians, variances, or what? Many statisticians thought, mainly in the 50s and 60s, that an answer to this problem could be given by determining what alternative hypotheses are consistent for the S test statistic. A statistical test is called **consistent** for a given alternative to the null hypothesis if, for any level of significance α, when that alternative hypothesis is true, the probability of rejecting the false null hypothesis, i.e., the power of the test, approaches 1 as the sample size N on which the test is based approaches infinity (Kruskal, 1952; Bradley, 1968, p. 56). The consistency analysis of a test can reveal to which alternatives this test is most sensitive because if the sample size is large, under these alternatives the probability of rejecting the false null hypothesis will be the highest.

Some of those who are concerned with the consistency of a statistical test accept the following reasoning. If the test is not significant, accept the null hypothesis of identical distributions, but if the test is significant, accept the alternative for which the test is consistent.

A major problem with this argumentation is that if we always accept, whenever the test is significant, the special alternative for which the test is consistent, our conclusion may frequently be incorrect. Why? Because inconsistent alternatives can also lead to significant results (though with a lower probability).

An analogy can be drawn with the two-sample t test. It can easily be verified that the two-sample t test, when used to compare two distributions, is consistent for any alternative hypothesis which implies the inequality of the population means, irrespective of the possible inequality of population variances. But variance heterogeneity in itself can lead to significant t-values. For example, if the sampling distribution is normal, the sample sizes are 15 and 5, and the ratio of the corresponding population variances is 1 to 10, then the probability that the t test will be significant at the 5% level when the population means are equal, is as large as 23% (see Scheffé, 1959, p. 353, Table 10.4.1). Thus, if we use the t test to compare population means without any consideration of variances, just because the t test is consistent against mean differences, the probability of falsely rejecting the null hypothesis of equal population means (i.e., the Type I error rate) will be seriously inflated in certain situations.

We wish to emphasize that consistency theorems are important not so much because they indicate which alternative hypothesis should be adopted but because they indicate which null hypothesis can most legitimately be tested (in the case of the t test, the equality of expected values). At the same time, one must not forget about the side conditions that must be satisfied for the Type I error rate to be correct (kept at the nominal level). Unfortunately, many early treatments of the KWt ignored the importance of side conditions, as we will now review.

Kruskal (1952) formulated the null hypothesis H_0 of the H test in terms of the equality of the distributions of variable X in the *I* populations to be compared. He demonstrated that, assuming H_0 is true, the H test statistic follows asymptotically a chi-square distribution as the overall sample size, N, tends to infinity. Kruskal (1952) also was able to prove that an alternative hypothesis is consistent for the H test if and only if there exist Aat least one of the populations for which the limiting probability is not one-half that a random observation from this population is greater than an independent random member of the N sample observations (Kruskal & Wallis,

1952, p.). For the sake of simplicity let us use the term **stochastic heterogeneity** to describe a set of distributions (or populations) for which this consistency condition holds and the term **stochastic homogeneity** for the opposite (for a more precise definition see the next section). Kruskal and Wallis (1952, p. 598) conclude: AThus, what H really tests is a tendency for observations in at least one of the populations to be larger (or smaller) than all the observations together, when paired randomly. In many cases, this is practically equivalent to the mean of at least one population differing from the others.

Kruskal and Wallis (1952) also 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).

The first concise book on applied nonparametric statistics was published by S. Siegel in 1956. AIt was a great success, especially among behavioral scientists, and, in fact, it ranked second on the list of most cited mathematics and statistics books, 1961-1972, with 1824 citations (Hettmansperger, 1984, p. vii). Unfortunately, this book treated the KWt in the same way as Kruskal and Wallis, which turns out not to be totally correct: AThe Kruskal-Wallis technique tests the null hypothesis that the k samples come from the same population or from identical populations with respect to the averages. The test assumes that the variable under study has an underlying continuous distribution. It requires at least ordinal measurement of that variable (Siegel, 1956, pp. 184-185). This formulation implies that the KWt can be used to compare several population means and that it does so without requiring the more restrictive conditions of the ANOVA (namely normality and variance homogeneity). This is more explicitly stated when Siegel treats the Mann-Whitney U test, an equivalent of KWt in the two-group case: A[the Mann-Whitney test] is therefore an excellent alternative to the t test, and of course it does not have the restrictive assumptions and requirements associated with the t test (Siegel, 1956, p. 126).

Bradley (1968) introduced the KWt in a similar way: “Suppose that an experimenter wishes to test the hypothesis that C infinite populations ... are identical against the alternative that they are not, and wishes the test to be especially sensitive to differences in location” (Bradley, 1968, p. 129). He lists much the same conditions of the KWt as Siegel (Bradley, 1968, p. 131), and refers to the same consistency condition as Kruskal and Wallis (Bradley, 1968, p. 132).

Mosteller and Rourke's book on nonparametric statistics, published in 1973, still described the KWt and the Mann-Whitney test similarly to Siegel (1956). Concerning the Mann-Whitney test Mosteller and Rourke claim: ARecall that the t-distribution approach for small sample tests (used in Section 2-6) assumes that the two populations from which the samples are drawn have approximately equal variances and roughly the shape of normal distributions. No such assumptions are needed for the test discussed in this chapter (Mosteller & Rourke, 1973, p. 55). Mosteller introduced the KWt as follows: AGiven three or more independent samples of measurements, we may wish to discover whether or not these samples could reasonably have come from distributions with the same location (such as mean or median) (Mosteller & Rourke, 1973, p. 212). And no side condition or special assumption is mentioned in his book with regard to the KWt.

These authors are all mistaken when they assume that the KWt is appropriate to test the

null hypothesis of stochastic homogeneity [the opposite of which the test was proven to be consistent for] without any further side condition (e.g., equal population variances, identical shape of distributions, etc.). But they commit another error, when they treat stochastic homogeneity as practically equivalent to the equality of means or medians (see the next section).

We think that these influential works are largely responsible for the still existing misunderstandings concerning the null hypothesis and the side conditions of the KWt (and the Mann-Whitney test as well). But how can the KWt be used in a correct way? The simplest solution is to apply restrictions similar to those used in ANOVA. In ANOVA the side conditions of normality and variance homogeneity imply that the distributions to be compared can differ from each other only by a shift constant. It must be emphasized that, if one does not want to place any restrictions on the form of the distributions, as is common with nonparametric tests, but one still wants to make inferences about location parameters, then very restrictive additional assumptions are required. Variance homogeneity is no longer sufficient. One must assume that the distributions do not differ in *any* respect besides location parameters, which necessitates among other things that the distributions must be identical not only in their variances, but also in their skewness parameters, in their kurtosis parameters, etc.

This restrictive assumption, which is generally referred to as the **shift model** or **additivity model**, or **location model**, appeared in the mathematical statistics literature with regard to the KWt as early as 1969 (see Hájek, 1969, p. 97). This approach has become the standard treatment of the KWt in more mathematically oriented nonparametric statistics books since then (see Kendall & Stuart, 1973, p. 521; Lehmann, 1975, p. 231; Randles & Wolfe, 1979, p. 394; Hettmansperger, 1984, p. 179; Gibbons & Chakraborti, 1992, p. 299; Maritz, 1995, p. 190).

Within the shift model the null hypotheses of identical distributions, equal population means, equal population medians, and stochastic homogeneity are all equivalent to each other. Another preferable feature of this model is that it can be generalized to the general linear model where more complex problems (multiway and multivariate comparisons, simple and multiple regression, etc.) can be handled (Hettmansperger, 1984; Puri & Sen, 1985; Maritz, 1995). However, the shift model approach has some very disadvantageous features as well:

(1) No well known statistical tests are available to check whether the shift model is acceptable or not (none of the above cited books contains such a test). But if one uses the KWt without testing the tenability of the shift model, one commits the same mistake as performing an ANOVA without checking the equality of population variances.

(2) Those few tests of the shift model that exist (see, e.g., Doksum & Sievers, 1976) have not yet been proven to be valid and powerful.

(3) In many important situations the restrictive shift model, which requires equality of variance, of skewness, and of kurtosis, is definitely not acceptable. For example, with commonly used Likert scales, a change in central tendency generally implies a change in distribution shape.

In the current paper we provide a model for the use and interpretation of the KWt that is theoretically correct but much less restrictive than the shift model. Our logic is as follows:

A) The KWt should be viewed primarily as a test to evaluate the alternative hypothesis for which it is consistent, namely stochastic heterogeneity. Therefore, if we would like to use the

KWt for testing a more specific null hypothesis than that of identical distributions, the null hypothesis should only be stochastic homogeneity. For this reason our primary goal will be to find the conditions under which the KWt could be a valid test of this null hypothesis.

B) Contrary to Kruskal and Wallis (1952), who tried to neglect the difference between the equality of population means and stochastic homogeneity, we will argue that this latter characteristic of populations (or distributions) differs in many cases from the former, can be interpreted for any set of ordinal scale variables (not just for interval scale ones as the equality of means), and is just as meaningful as the former.

C) In contrast to the shift model, the proof in the current paper will not require homogeneity of variances of the original scores. Nonetheless, it will be seen that homogeneity of variances of the *ranks* will be required to make the KWt valid as a test of stochastic homogeneity. One is then faced with a problem similar to the Behrens-Fisher problem of how to test for equality of means in the unequal variance case. We propose to use robust alternatives to the KWt to test stochastic homogeneity when this condition of the KWt is violated.

Stochastic homogeneity and related concepts

Before explaining stochastic homogeneity we first introduce a concept that is closely related to it.

Definition. Let X and Y be two variables that are at least ordinally scaled. Then define A(X,Y) as follows:

$$A(X, Y) = P(X > Y) + \frac{1}{2}P(X = Y).$$

We will call A(X,Y) the **measure of stochastic superiority**[1] of variable X over variable Y.

It is readily seen from this definition that the larger A(X,Y) is, the greater the probability that a random X-value will be larger than a random Y-value. This measure can be used to compare two populations, say P and Q, with respect to a variable X as well. Just denote X with X_1 in P and X_2 in Q and let $A(P,Q) = A(X_1,X_2)$. A(P,Q) is the probability that a randomly sampled observation from population P is greater than a randomly sampled observation from Q plus one half times the probability that a randomly sampled observation from population P is equal to a randomly sampled observation from population Q. For the sake of simplicity we will sometimes use the simplified A_{XY}, A_{PQ}, A_{12}, or A_{ij} index notation.

The measure of stochastic superiority is not symmetric, since it can easily be verified that

[1] We chose not to use S or SS to designate the measure of stochastic superiority because these letters are so commonly used as abbreviations for other terms in statistics. As a mnemonic for A, we suggest that the measure of stochastic superiority indicates the likelihood that an observation on X will be above an observation on Y.

$$A(X, Y) = 1 - A(Y, X).$$

for any two variables X and Y

If A(X,Y) = A(Y,X) = 0.5, then we say that X and Y are **stochastically equal**. Furthermore, if A(X,Y) < 0.5 (or A(X,Y) > 0.5), then X is said to be **stochastically smaller** (or **larger**) than Y. Analogously, two populations, P and Q are said to be stochastically equal with respect to a variable X, if A(P,Q) = A(Q,P) = 0.5. The stochastically smaller and stochastically larger relations of two populations can be defined in the same way as they were defined for two variables.

It can easily be shown that the stochastic equality of variables X and Y is equivalent to

$$P(X < Y) = P(X > Y).$$

the following identity:

Now suppose that we compare *I* populations with respect to a variable X which is at least ordinally scaled. We say that these populations are **stochastically heterogeneous** with respect to the variable X if for these populations and variable X the KWt is consistent (i.e., if under this condition the power of the KWt tends to 1 when the total sample size, N, tends to infinity). Kruskal (1952) showed in a theorem (see also Noether, 1967, pp. 51-52) that the KWt is consistent against an alternative hypothesis if, and only if, for at least one group k (k = 1, ... , *I*)

$$\sum_{i=1}^{I} \frac{n_i}{N} A_{ki} = 0.5$$

the

identity does not hold asymptotically, provided the $\gamma_i = n_i/N$ proportions (here n_i is the i-th sample size) converge to positive constants while the total sample size, N, tends to infinity.

Here A_{ki} is the measure of stochastic superiority (of population k over population i) as was defined in (1). In order to make it obvious what identity (4) means, we will perform a small transformation upon it.

Reformulation of Kruskal's consistency condition

Because trivially $A_{XX} = 0.5$ for any variable X, the same is true for A_{kk}. Therefore (4) is readily seen to be equivalent to the following:

$$\frac{1}{N - n_k} \cdot \sum_{\substack{i=1 \\ i \neq k}}^{I} n_i A_{ki} = 0.5 \, (k = 1, \ldots, I).$$

In order to clarify the interpretation of equation (5), suppose that k = 1 and unite populations P_2, ..., P_I in such a way that each P_i (i = 2, ..., I) is weighted by the respective sample size n_i. Then, select randomly both an element x_1 from population P_1 and an element x_u from the united population, P_u. The extent to which population P_1 is larger than the united population can be assessed by the measure of stochastic superiority defined in (1):

$$A_{1u} = P(x_1 > x_u) + \frac{1}{2}P(x_1 = x_u).$$

Using a well-known theorem of probability theory, this expression for the united population can be expressed in terms of its component populations:

$$A_{1u} = \sum_{i=2}^{I} P(x_1 > x_u \_ x_u \varepsilon P_i) \bullet P(x_u \varepsilon P_i) + \frac{1}{2}\sum_{i=2}^{I} P(x_1 = x_u \_ x_u \varepsilon P_i) \bullet P(x_u \varepsilon P_i).$$

But realizing that

$$P(x_1 > x_u \_ x_u \varepsilon P_i) + \frac{1}{2}P(x_1 = x_u \_ x_u \varepsilon P_i) = A_{1i},$$

$$P(x_u \varepsilon P_i) = \frac{n_i}{\sum_{j=2}^{I} n_j} = \frac{n_i}{N - n_1}$$

$$A_{1u} = \frac{1}{N - n_1} \bullet \sum_{i=2}^{I} n_i A_{1i}.$$

$$A_{1u} = 0.5.$$

and that

we obtain

Now this identity entitles us to say that if (5) is true then

This means that in this situation the P_1 population is stochastically equal to the P_u population, the weighted union of the remaining populations. And as the independent samples play an entirely symmetrical role in the Kruskal-Wallis test, a similar statement can be formulated with respect to all of the other populations too. From this derivation it is now clear that the H test is consistent against an alternative hypothesis if, and only if, it implies that at least one of the *I* populations is not equal stochastically to the weighted united population of the remaining ones. From (5) it is also clear that in the two-group case ($I = 2$) stochastic equality and stochastic homogeneity are equivalent to each other and therefore the latter can be considered the generalization of the former.

In order to reveal the conditions under which the KWt will be a valid test of the null hypothesis of stochastic homogeneity, in the next section an expression equivalent to stochastic

homogeneity will be derived.

Stochastic homogeneity is equivalent to equality of rank mean expected values

Let us assume that we have *I* independent random samples, of size n_1, n_2, ..., n_I, from the populations P_1, P_2, ..., P_I respectively. Let X_{ij} (i = 1, ... , *I*; j = 1, ... , n_i) denote the j-th observation of the i-th sample. In the Kruskal-Wallis test all the observations have to be ranked (from low to high) in the entire set of the N data values. Let us denote the rank assigned to an original observation X_{ij} by R_{ij}, and the mean rank for the i-th sample by r_i. A one-way ANOVA performed

$$H_o : E(r_1) = E(r_2) = ... = E(r_I).$$

on the ranks is obviously a test of the following null hypothesis:

We will now prove that this identity implies stochastic homogeneity and vice versa. In this proof we will not assume that the different populations have identical distributions but only that the samples to be compared are independent, and the scores *within* a given sample are independently and identically distributed. The key step in the proof is the expression of the expected value of the rank means in terms of the pairwise stochastic superiority measures A_{ki}. We begin the proof by calculating $E(r_1)$, the expected value of the first rank sample mean. Let Z_{ij} be for each X_{ij} observation (i = 1, ..., *I*; j = 1, ..., n_i) a random variable with a value of 1, 0.5, or 0, depending on whether X_{11} (which is the first score in the first sample) is greater than, equal to, or less than X_{ij}, respectively. By means of these variables R_{11}, the rank value of X_{11}, can be written

$$R_{11} = 1 + \sum_{j=2}^{n_1} Z_{1j} + \sum_{i=2}^{I} \sum_{j=1}^{n_i} Z_{ij}.$$

as follows:

Consider first the Z_{ij} elements in the rightmost term above. Since the X_{ij} are themselves

$$A_{1i} = P(X_{11} > X_{i1}) + \frac{1}{2} P(X_{11} = X_{i1}).$$

random variables we can again use the A_{1i} measure of stochastic superiority, introduced in (1): A_{1i} equals the probability that the first element in sample 1 is greater than the first element in sample i, plus 1/2 times the probability that the first element in sample 1 is equal to the first element in sample i. But as the original data are identically distributed within each subsample, X_{11} can be replaced in formula (14) with any of the X_{1j} (j = 2, ..., n_1) elements and X_{i1} with any of the X_{ij} (j = 2, ..., n_i) elements; hence A_{1i} depends only upon the two samples selected (namely upon 1 and i) and indicates the extent to which population P_1 is stochastically greater than P_i. For this reason the expected value of Z_{ij} equals A_{1i} for any j.

Considering the Z_{1j} terms we can realize that for every X_{1j} (j > 1) element of sample 1

$$E(Z_{1j}) = P(X_{11} > X_{1j}) + \frac{1}{2}P(X_{11} = X_{1j}) = \frac{1}{2}$$

(recall that X_{11} and X_{1j} are independent and have the same distribution).

$$E(R_{11}) = 1 + \frac{n_1 - 1}{2} + \sum_{i=2}^{I} n_i A_{1i}.$$

Taking now the expected value of (13), we obtain

Because within each sample the individual ranks have the same distribution, the $E(R_{1j})$ (j = 1, ..., n_1) expected values of the ranks, as well as the $E(r_1)$ expected value of the r_1 rank mean, are all

$$E(r_1) = \frac{n_1 + 1}{2} + \sum_{i=2}^{I} n_i A_{1i},$$

equal. Hence we get from (16) by a little rearrangement:

$$E(r_k) = \frac{n_k + 1}{2} + \sum_{\substack{i=1 \\ i \neq k}}^{I} n_i A_{ki}.$$

and again, because of the interchangeability of the populations we have for any k (k = 1, ... , *I*)

Now let us assume that (12) holds, i.e., the expected value of the rank mean is the same for all *I* groups, and let us denote the common rank mean expected value by C. It can easily be verified that the sum of the *I* sample rank sums equals the sum of the first N natural numbers,

$$\sum_{i=1}^{I} n_i r_i = \frac{N(N+1)}{2}$$

namely N(N+1)/2. Therefore

$$C \cdot N = \frac{N(N+1)}{2},$$

from which, taking the expected value of the left side, we obtain

$$C = \frac{N+1}{2}.$$

and hence, for C we get

If we substitute the obtained value of $E(r_k) = E(R_{11}) = C$ in (18), we obtain

$$\frac{N+1}{2} = \frac{n_k + 1}{2} + \sum_{\substack{i=1 \\ i \neq k}}^{I} n_i A_{ki}.$$

After a little rearranging we get the identity (5) of stochastic homogeneity.

We have now shown that equality of the expected values of the rank means implies stochastic homogeneity. The opposite side of the equivalence, i.e., that stochastic homogeneity implies identity (12), is clearly seen from the above derivations. Identity (5) implies identity (22), and comparing (22) and (18), we can see that the $E(r_k)$ expected values must be equal.

Why is the obtained equivalence so important? Suppose that, instead of comparing the population means, one is interested in the stochastic homogeneity of the populations. Now the null hypothesis of stochastic homogeneity has been proven to be equivalent to (12). But for comparing the rank mean expected values in (12) one can apply an ANOVA on the rank scores, provided the following well-known conditions of the ANOVA are met:

(i) Normality: The rank scores will certainly not follow a normal distribution. However, it is widely known that the ANOVA is quite robust against nonnormality concerning Type I error (Scheffé, 1959, p. 345; Tan, 1982; Vargha, 1993), and hence no severe problem is expected to arise in applying it to rank data.

(ii) Variance homogeneity: Note that this condition refers to the ranks rather than to the original scores.

(iii) Independence: The ranking of the original scores lessens their degrees of freedom by one (the sum of the ranks is always a constant that depends only on the total sample size). From this it follows that the larger the samples are, the less the dependency among the rank scores will be. Thus, the independence of both the individual rank scores, and of the whole rank samples will be fulfilled asymptotically as the total sample size, N, tends to infinity.

From (i), (ii), and (iii) one can readily conclude that a rank ANOVA procedure should be an asymptotically valid method for testing the null hypothesis of stochastic homogeneity if the expected values of the rank sample variances are equal.

The KWt is often called an "ANOVA by Ranks". The reason for this is that the H statistic (after a division by df=I-1) is in its form very much like the F statistic of the one-way ANOVA performed on the ranks assigned to the original scores (Conover & Iman, 1981; Freund & Walpole, 1987, p. 537; Maxwell & Delaney, 1990, p. 704). Certain studies also show that using the F=H/df test statistic and the appropriate F-table instead of H in the KWt, yields at least as good results as H does with the chi-square table (Iman & Davenport, 1976). Thus, computationally the KWt can be regarded as an ANOVA-type procedure carried out on ranks. Combining this evidence with our obtained results, we are now justified to claim that the KWt should be regarded as an asymptotically valid test of the null hypothesis of stochastic homogeneity, under the condition of rank variance homogeneity (which is not so restrictive as the assumption of the shift model). When variances differ substantially, however, the KWt may prove inadequate, especially in the in the unequal sample size case (Keselman, Rogan & Fair-Walsh, 1977; Tomarken & Serlin, 1986; Oshima & Algina, 1992). For example Tomarken and Serlin (1986), working with normal distributions (where equality and stochastic homogeneity are equivalent states of affair), found that in direct pairing cases (when larger sample sizes are associated with larger variances) the KWt was extremely conservative, while in inverse pairing cases the KWt was consistently liberal.

The same results were obtained by Oshima and Algina (1992, Table 4) for three symmetric

distributions (normal, uniform, and t(5)). It is interesting, however, that for the two asymmetric distributions (beta and exponential) they used in their simulations, the Type I error rates were inflated not only for indirect, but also for direct pairing cases (though in these latter cases the inflation was lower). This and a similar finding with the test of equal medians led Oshima and Algina (1992, p. 261) to conclude that "the Kruskal-Wallis should not be used as a test of location if data are expected to be heteroscedastic". Based on the findings of our paper we have to note, however, that the found inadeqacy of the KWt with asymmetric heteroscedastic distributions was only partly due to the inequality of the variances. In these situations the equality of expected values (or medians) excludes stochastic homogeneity, which itself tends to increase the rejection rate of the KWt when the expected values (or medians) are equal.

Turning back to our theoretical reasoning, we are also entitled to assume that in the unequal variances case robust ANOVA alternatives (Welch test, Brown-Forsythe test, James test, etc.) should be appropriate for testing the null hypothesis of stochastic homogeneity. This claim is supported by a simulation study of Zimmerman and Zumbo (1992). In this study the authors sampled from the normal and an extremely long-tailed symmetric distribution, where equality of expected values and stochastic homogeneity were equivalent states of affair. The authors found that under violation of normality and variance homogeneity the Welch-Satterthwaite approximate t test performed on ranks exhibited fairly acceptable Type I and excellent power characteristics, outperforming both parametric Welch, and nonparametric Mann-Whitney (equivalent to KWt when I=2) tests.

Furthermore we can hypothesize that the KWt will be robust to some extent to violations of expected rank variance equality as long as sample sizes are equal, because the same is true with respect to violations of expected variance equality (see., e.g., Scheffé, 1959, p. 354; Clinch & Keselman, 1982).

More about stochastic equality and stochastic homogeneity

The concept of 'stochastically smaller' or 'larger' is not new in the statistical literature. Mann and Whitney used it when they analyzed the consistency characteristics of the two-sample rank test (Mann & Whitney, 1947). Specifically they have shown that having the random variables X and Y, the relation AX is stochastically smaller than Y is a sufficient condition for the consistency of the Mann-Whitney test.

However, they used a stronger form of the *stochastically smaller* relation than that defined by means of the measure of stochastic superiority (see (1)). According to their definition, X is

$$F_X(c) > F_Y(c), \text{ for all } c$$

stochastically smaller than Y if for the respective F_X and F_Y cumulative distribution functions holds.

It can be seen that (23) always implies our weak form, while the opposite is not necessarily true (see Randles & Wolfe, 1979, p. 132). Note that (23) is a stronger relation than the statement $\mu_X < \mu_Y$ in the sense that (23) always implies the analogous relation of the expected values, while the

opposite is not true.

It is also worth mentioning that in their paper, Mann and Whitney (1947) actually proved that their test was consistent against all alternatives for which the stochastic equality defined by (1) does not hold (see also Lehmann, 1951; Kruskal, 1952; Randles & Wolfe, 1979, p. 127).[2]

The relations of stochastic equality and stochastic ordering (stochastically larger or smaller) between populations or distributions have, however, a major shortcoming - they do not have the valuable property of transitivity. It may happen, for example, that for a given variable X and populations P, Q, and R, the relationships P = Q and Q = R stochastically hold, while P = R does not. From this weakness it also follows, that stochastic homogeneity is not equivalent to the pairwise stochastic equalities of the respective populations. The latter implies the former, while the opposite is not true, unless the number of the populations to be compared is only two. Extreme situations can occur; for example stochastic homogeneity may hold but Q > P, P > R and R > Q with $A_{QP} = A_{PR} = A_{RQ} = A > 0.5$.

Some of the various possibilities can be clearly illustrated by using examples constructed by using sets of three dice. Assume that we have three fair dice with integer numbers on their sides. A die P is called superior to die Q if throwing them simultaneously (and independently) P will show a greater number than Q in the majority of the cases (i.e., with a probability greater than 0.5). The following sets of numerical values for dice P, Q, and R yield numbers with equal means but where P is superior to Q, Q is superior to R, and R is superior to P:

Number on P: 1, 1, 1, 4, 7, 7;
Number on Q: 0, 0, 3, 6, 6, 6;
Number on R: 2, 2, 2, 5, 5, 5.

It can be easily verified that in this case

$$P(P>Q) = P(Q>R) = P(R>Q) = 21/36 > 0.5,$$

which shows that for each pair of dice one of them is superior to the other. However, the equation also implies that this set of distributions (defining the variable values as the results of the throws) is stochastically homogeneous.

In the above example the equality of expected values also holds (each of them equals 3.5). But if we define the distributions as follows:

Number on P: 0, 0, 0, 2, 4, 8;
Number on Q: 0, 0, 0, 4, 4, 4;
Number on R: 0, 0, 1, 2, 2, 5,

[2]For some results concerning the stochastically smaller relation, see Randles & Wolfe (1979, pp. 127-133), and for a test of whether one of two variables is stochastically larger than the other, see Wilcox (1990).

stochastic homogeneity will still hold (the pairwise stochastic superiority values are all equal to 0.5), but the expected values will be different (14/6, 12/6, and 10/6, respectively).

To show that the equality of expected values does not imply stochastic homogeneity, take three variables with each following a chi-square distribution with 2 degrees of freedom, which is a highly positively skewed distribution. Standardize the variables (i.e., subtract the expected value and divide them by the SD)[3], and multiply the first variable by -1 (to make it be oppositely skewed from the others). The obtained variables will then have identical zero expected values and identical SD=s of 1. It can be shown that in this case (rounding to two decimals) $A_{12} = A_{13} = .59$, and $A_{23} = .50$, and assuming equal sample sizes $A_{1u} = .59$, and $A_{2u} = A_{3u} = .45$, which is a case of stochastic heterogeneity (see identity (11)). Further, if we shift these three distributions to yield identical zero medians[4] while maintaining identical variances, the three distributions will still not be stochastically homogeneous, since in this case (again rounded to two decimals) $A_{1u} = .41$, and $A_{2u} = .A_{3u} = .55$.

Further examples could also be presented to illustrate that stochastic homogeneity, equality of expected values, and equality of medians can be really different states of affairs whenever distributions are not symmetric. It must be accepted that stochastic homogeneity implies only that none of the populations is stochastically larger or smaller than the union of the others. Nevertheless, if the distributions in question are symmetric, then stochastic homogeneity, pairwise stochastic equality, and equality of the expected values are equivalent (in the continuous case this list would also include the equality of the medians).

Another shortcoming of stochastic homogeneity is that the null hypothesis involves the observed sample sizes. This can be accepted only in two situations.

(i) If sample sizes are equal, since in this case we give equal weights to the populations to be compared and these weights are already independent of the concrete sample size values (see formula (4)).

(ii) If the sample sizes are - at least asymptotically - proportional to the sizes of the corresponding populations, since in this case the n_i/N proportions in formula (4) express real population characteristics. This can be afforded for example if the sampling is a large sampling from the whole population, yielding that the n_i/N sample size proportions are approximately equal to the corresponding subpopulation proportions.

Doubt may arise only in situations where one has unequal sample sizes that do not seem to reflect the real population size proportions. Should we equalize in such cases by throwing away some data? Though this can also be a solution simetimes, we are currently working on such a generalization of the KWt which will be able to test the null hypothesis of stochastic homogeneity with an equal weighting of the different populations even if sample sizes are different.

Although stochastic equality has the shortcoming of intransitivity, it has an advantage over

---

[3]For the mean and the SD of the chi-square distribution, see Evans, Hastings and Peacock (1993, p. 45).

[4]Using the fact that the median of a chi-square distribution with 2 degrees of freedom is 1.39 (see Siegel, 1956, p. 249), one can determine that the appropriate shift values are -.305, .305 and .305.

the comparison of expected values: if the observed variable is assumed to be dependent on an underlying latent trait in a nonlinear way (this may happen in many cases), then the relationship of the expected values is highly dependent on the item-response function, while stochastic equality and homogeneity are not (provided this function is of a monotone type, which is the usual case; see Maxwell & Delaney, 1985).

Closely related to the concept of stochastic equality is the common language effect size statistic (CL). According to McGraw and Wong (1992), for continuous data CL is the probability that a score sampled at random from one distribution will be greater than a score sampled from some other distribution. It is easily seen that our measure of stochastic superiority defined by (1) is a generalization of CL, which applies to both the discrete and the continuous cases. However, this measure is at the same time a key term in defining stochastic homogeneity by formal mathematics (see formulae (4) and (5)). The primary value of CL is "that it is better than the available alternatives for communicating effect size to audiences untutored in statistics" (McGraw & Wong, 1992).

Generalization to Friedman's G test

Following the same steps with respect to Friedman's G rank test (about G see, e.g., Randles & Wolfe, 1979, pp. 401-402) it can be shown that in the one-way repeated measures design, if the null

$$P(X_{ij} < X_{kj}) + \frac{1}{2}P(X_{ij} = X_{kj})$$

hypothesis (12) is true, and if for each (i,k) pair of treatments the
quantities do not depend on j (this situation holds, for example, when the person by treatment interaction is zero),
then the common expected rank mean will be C = (I+1)/2, and consequently we will have the

$$\frac{1}{I-1}\cdot\sum_{\substack{i=1\\ i\neq k}}^{I} A_{ki} = 0.5\ (k=1, ..., I).$$

following fundamental identity for the G test:
This identity can be shown to be equivalent to stochastic homogeneity in a way similar to (18). The Friedman test again happens to be consistent against an alternative hypothesis if, and only if, stochastic homogeneity of the distributions to be compared does not hold under this alternative (see Noether, 1967, pp. 52-54). These relations form the basis of the assertion that the Friedman test can be regarded in the same way as the KWt, the only difference being that the former is to be used for a repeated measures design, the latter for a set of independent samples.

Discussion

The literature we have reviewed reveals that many previous treatments of the KWt have been inadequate. Numerous authors introduce the test as a means of assessing location parameters while

disregarding the side conditions that must be satisfied for the test to be valid. This approach fails to meet the statistical criterion of keeping the Type I error rate close to the nominal level. The shift model approach that has come to be accepted in recent mathematical statistics literature, while mathematically correct, is so restrictive (not allowing any other difference between the distributions except for an additive constant) that it fails to meet the needs of behavioral scientists in realistic data analysis situations.

By the theorem proved in this paper we have been able to demonstrate that the KWt can be validly used to test a hypothesis about location under much less restrictive conditions than required by the recently accepted, mathematically correct shift model. We have shown that it is *not* necessary to presume that the distributions have the same shape, but only that a certain kind of variance homogeneity holds.

The specific null hypothesis for which the KWt can be validly used is stochastic homogeneity. As we have illustrated, this can hold even when the population means or medians are different. Stochastic homogeneity means that any of these distributions is stochastically equal to the weighted union of the rest, where the weights are proportional to the sample sizes. In simpler words, what H really tests is a tendency for observations in at least one of the populations to be larger (or smaller) than all the remaining populations together. In a situation where it is questionable to include the sample size considerations into the null hypothesis, it would be best to give equal weights to the distributions to be compared by using equal sample sizes. In some cases, however, as in nonorthogonal ANOVA, the unequal sample sizes may reflect important population characteristics.

In this paper stochastic homogeneity has been shown to be equivalent to the equality of the rank sample expected values (even for unequal n's). Considering that the KWt and the G test are practically ANOVA-type procedures performed on ranks, we may conclude that the KWt and the G test should be the test of choice to test stochastic homogeneity under the condition of variance homogeneity of ranks. Moreover, we assert that robust forms of ANOVA performed on the rank transforms are appropriate for testing stochastic homogeneity in the heterogeneous variance case.

If there are only two populations to be compared, then stochastic homogeneity reduces to stochastic equality, and the KWt becomes equivalent to the Mann-Whitney U test. Consequently, the Mann-Whitney test should be considered as an asymptotically valid test of stochastic equality if the condition of rank variance homogeneity holds. Under the violation of this condition a robust alternative to the t test (e.g., the Welch test) can be used on the rank scores (see Zimmerman & Zumbo, 1992).

Regarding these results the take-home message of our paper can be summarized as follows:

1) If you want to compare independent or matched samples, do not use nonparametric procedures (Mann-Whitney, Kruskal-Wallis or Friedman tests) just because the side conditions of the corresponding parametric comparison tests (two-sample t test and ANOVA) are violated. This is a highly incorrect way that still prevails in current statistical advisory systems (see, e.g., Silvers, Herrmann, Godfrey, Roberts, & Cerys, 1994; Herrmann, Silvers, Godfrey, Roberts, & Cerys, 1994). This approach is incorrect partly because parametric and nonparametric comparison procedures may have similar side conditions (for instance the homogeneity of variances), and partly because these latter techniques test a different null hypothesis, namely stochastic homogeneity, instead of the

equality of expected values. Once the dependent variable has a nonnormal, asymmetric distribution, these two null hypotheses are no longer equivalent to each other. And nonnormal distributions occur in behavior sciences far more frequently than one might think (see Micceri, 1989).

2) Before using any test, try to specify what kind of comparison you are mainly interested in. If mean levels can characterize appropriately the differences between populations and variables, use parametric comparison methods. Under the violation of their conditions use robust alternatives (see, e.g., Wilcox, 1996, sections 8.2, 9.2, 10.2).

3) Choose nonparametric procedures if you are interested in a holistic comparison and are aiming to discover if the scores in the different populations or situations to be compared are alike, or, on the contrary, there exists a population (condition) where the scores are generally smaller (or larger) than in the others. If there is any indication that the variance homogeneity condition of the Mann-Whitney test or the KWt is violated, use robust t test or ANOVA alternatives on the rank scores (preferably Welch and James tests; see Wilcox, 1987, 1988, 1989, 1996, pp. 133,182-184).

4) The discussed nonparametric methods are highly recommended if the dependent variable is only ordinally scaled (in this case the expected value is a meaningless measure), or if there is any doubt that the observable scores are not strictly linearly related to the real but latent scores (this might be the case with many personality scales and achievement measures), because the discussed nonparametric methods only need that the observable scores depend upon the latent scores in a monotonic way (yielding ordinally scaled observed scores).

It may be helpful to cite some concrete examples of when one might reasonably be more interested in a question of stochastic superiority than mean differences. The quarrel among sports fans from different regions of the country about which conference is the strongest overall is most reasonably a question of stochastic superiority, i.e., what would be the outcome of all possible pairings of a team from conference X and a team from conference Y. A single outstanding team might make conference X=s mean power rating, say, higher but not be indicative of whether that conference X=s teams would consistently beat teams from conference Y.

To conclude with a psychological example, the decision facing a prospective client about which therapist to choose from various categories of available therapists is really a question of stochastic superiority. Suppose a client in a large urban area has decided he wants to see a clinical psychologist who also holds a faculty position at a university in the area. The only available information he has is the professorial rank of the faculty member as Assistant, Associate or Full. The client should not be as concerned with the mean levels of quality of the therapists in the different rank categories as with the probability that he would be better off choosing a therapist from a particular rank category. After all, the client is likely going to experience therapy with a single individual, not the whole group of therapists within a professorial rank category. In general, when one is facing a decision between a single choice from one category and a single choice from another category, although other decision rules are also possible, a strategy based on the probability you would be better off with a selection from a particular category than from the other alternatives is not only defensible but is arguably one of the most rational decision rules that one could adopt.

Remaining questions

After the theoretical considerations of our paper we are still left with some unanswered questions.

1) It has been argued that the KWt, and the rank versions of the robust ANOVA alternatives should be asymptotically valid tests of stochastic homogeneity. The following questions arise, however. How large should samples be in order to insure the appropriateness of these tests? How do these tests work with relatively small samples?

2) Obviously, stochastic homogeneity and equality of expected values are theoretically different states of affairs. But can the difference between them be so large that it could cause severe inconsistencies between the results of KWt and ANOVA?

3) There is also an uncertainty concerning the choice among the possible robust rank alternatives in the variance heterogeneity case. Which is the best of them? How large a difference of the variances can they tolerate?

4) In the special case when stochastic homogeneity and equality of expected values are equivalent (this is the case when the distributions are symmetric or have identical shape, allowing only for differences in location parameter), the above mentioned robust rank methods may provide better alternatives to the ANOVA than their parametric counterparts in some cases. Under which circumstances are nonparametric tests better?

In order to answer these questions of great practical importance, a consize Monte Carlo analysis would be highly reccomended. Nevertheless there are some literatures which allow us to provide some guidance concerning these issues.

The KWt seems to be appropriately used to test stochastic homogeneity under the variance homogeneity condition with an average sample size of as low as 12 (see Tomarken & Serlin, 1986). This is likely to be true also for its robust alternatives.

If variances are unequal and the sampling distributions differ greatly from a normal distribution, parametric and nonparametric group comparison tests may lead to quite inconsistent test results (see Table 4 in Oshima & Algina, 1992, where rejection rates are summarized for five robust parametric ANOVA alternatives and the KWt under the hypothesis of equal expected values for five different distributions).

Since it is still unclear which method is the best among those proposed for comparing means that allow unequal variances (see, e.g., Wilcox, 1996, p. 182-184), we are also uncertain regarding the relative merits of their rank counterparts. There are, however, some indications that under several extreme variance heterogeneity conditions James' (1951) second order method outperforms both Welch's (1951) and the Brown-Forsythe (1974) procedure (see, e.g., Oshima and Algina, 1992).

In the special case when the distributions to be compared have identical shapes, the asymptotic relative efficiency (ARE) of the KWt relative to the ANOVA F-test can be as high as infinity, but cannot be lower than .864 (Hodges & Lehmann, 1956; Bradley, 1968, p. 132). There is some theoretical as well as empirical evidence that the more heavily tailed and more asymmetric a sampling distribution is, the better KWt is compared to ANOVA F-test in terms of power (Hodges & Lehmann, 1956; Zimmerman & Zumbo, 1992). Presumably a similar relationship may hold between robust KWt alternatives and their parametric counterparts.

References

Bradley, J. V. (1968). *Distribution-free statistical tests.* Englewood Cliffs, NJ: Prentice Hall.

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.

Clinch, J. J., & Keselman, H. J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics, 7,* 207-214.

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

Doksum, K. A. & Sievers, G. L. (1976). Plotting with confidence: Graphical comparison of two populations. *Biometrika, 63,* 421-434.

Evans, M., Hastings, N., & Peacock, B. (1993). *Statistical distributions* (2nd ed.). New York: Wiley.

Freund, J. E. & Walpole, R. D. (1987). *Mathematical statistics* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Gibbons, J. D. & Chakraborti, S. (1992). *Nonparametric statistical inference* (3rd ed). New York: Marcel Dekker.

Hájek, J. (1969). *A course in nonparametric statistics*. San Francisco: Holden-Day.

Herrmann, N., Silvers, A., Godfrey, K., Roberts, B., & Cerys, D. (1994). A prototype statistical advisory system for biomedical researchers II: Development of a statistical strategy. *Computational Statistics & Data Analysis,* 18, 357-369.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks.* Wiley: New York.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1994). *Applied statistics for the behavior sciences* (3rd ed.). Boston: Houghton Mifflin Company.

Hodges, J. L., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics, 27,* 324-335.

Howell, D. C. (1992). *Statistical methods for psychology.* Belmont, California: Duxbury Press.

Hurlburt, R. T. (1994). *Comprehending behavioral statistics.* Pacific Grove, California: Brooks/Cole Publishing Company.

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

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.

Jaccard, J., & Becker, M. A. (1990). *Statistics for the behavioral sciences* (2nd ed.). Belmont, CA: Wadsworth.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika, 38,* 324-329.

Kendall, M. G. & Stuart, A. (1973). *The Advanced Theory of Statistics, Vol. 2: Inference and Relationship* (3rd ed). London: Griffin.

Keselman, H. J., Rogan, J. C. & Feir-Walsh, B. J. (1977). An evaluation of some nonparametric and parametric tests for location equality. *British Journal of Mathematical and Statistical*

Psychology, 30, 213-221.

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

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

Lehman, R. S. (1991). *Statistics and research in the behavioral sciences.* Pacific Grove, California: Brooks/Cole Publishing Company.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics, 22,* 165-179.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks.* San Francisco: Holden-Day, New York: McGraw-Hill.

Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics, 18,* 50-60.

Maritz, J. S. (1995). *Distribution-free statistical methods* (2nd ed.). London, New York: Chapman & Hall.

Maxwell, S. E. & Delaney, H. D. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin, 97,* 85-93.

Maxwell, S. E. & Delaney, H. D. (1990). *Designing experiments and analyzing data. A model comparison perspective.* Belmont, California: Wadsworth Publishing Company.

McGraw, K. O. & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361-365.

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

Mosteller, F. & Rourke, R. E. K. (1973). *Sturdy statistics: Nonparametrics and order statistics.* Reading, Massachusetts: Addison-Wesley.

Noether, G. E. (1967). *Elements of nonparametric statistics.* New York: Wiley.

Oshima, T. C. & Algina, J. (1992). Type I error rates for the James's second order test and Wilcox's H_m test under heteroscedasticity and nonnormality. *British Journal of Mathematical and Statistical Psychology, 45,* 255-263.

Pagano, R. R. (1994). *Understanding statistics in the behavioral sciences* (4th ed.). St. Paul: West Publishing Company.

Puri, M. L., & Sen, P. K. (1985). *Nonparametric methods in general linear models.* New York: Wiley.

Randles, R. H. & Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics.* New York: Wiley.

Runyon, R. P., & Haber, A. (1991). *Fundamentals of behavioral statistics* (7th ed.). New York: McGraw-Hill.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Silvers, A., Herrmann, N., Godfrey, K., Roberts, B., & Cerys, D. (1994). A prototype statistical advisory system for biomedical researchers I: Overview. *Computational Statistics & Data Analysis,* 18, 341-355.

Spence, J. T., Cotton, J. W., Underwood, B. J., & Duncan, C. P. (1990). *Elementary statistics* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Tan, W. Y. (1982). Sampling distributions and robustness of t, F and variance-ratio in two samples and ANOVA models with respect to departures from normality. *Communications in Statistics - Theory and Methods, 11,* 2485-2511.

Tomarken, A. J. & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99,* 90-99.

Triola, M. F. (1995). *Elementary statistics* (6th ed.). Massachusetts, California, New York: Addison-Wesley.

Vargha, A. (1993). How to use ANOVA in case of dichotomous dependent variables. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.), *Psychometric Methodology. Proceedings of the 7th European Meeting of the Psychometric Society in Trier* (pp. 535-539). Stuttgart: Gustav Fischer Verlag.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38,* 330-336.

Welkowitz, J., Ewen, R. B., & Cohen, J. (1991). *Introductory statistics for the behavioral sciences* (4rd ed.). New York: Academic Press.

Wilcox, R. R. (1987). A heteroscedastic ANOVA procedure with specified power. *Journal of Educational Statistics, 12,* 271-281.

Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results on James' second order method. *British Journal of Mathematical and Statistical Psychology, 41,* 109-117.

Wilcox, R. R. (1989). Adjusting for unequal variances when comparing means in the one-way and two-way fixed effects ANOVA models. *Journal of Educational Statistics, 14,* 269-278.

Wilcox, R. R. (1990). Determining whether an experimental group is stochastically larger than a control. *British Journal of Mathematical and Statistical Psychology, 43,*327-333.

Wilcox, R. R. (1996). *Statistics for the social sciences.* San Diego, New York: Academic Press.

Zimmerman, D. W. & Zumbo, B. D. (1992). Parametric alternatives to the Student t test under violation of normality and homogeneity of variance. *Perceptual and Motor Skills, 74,* 835-844.