Journal of Modern Applied Statistical Methods

Volume 5 | Issue 1 Article 4

5-1-2006

The Effect On Type I Error And Power Of Various Methods Of Resolving Ties For Six Distribution-Free Tests Of Location

Bruce R. Fay
Wayne County Regional Educational Service Agency, Michigan

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 22-40

1538 – 9472/06/$95.00

REGULAR ARTICLES

The Effect On Type I Error And Power Of Various Methods Of Resolving Ties For Six Distribution-Free Tests Of Location

Bruce R. Fay
Wayne County Regional Educational Service Agency, Michigan

The impact on Type I error robustness and power for nine different methods of resolving ties was assessed for six distribution-free statistics with four empirical data sets using Monte Carlo techniques. These statistics share an underlying assumption of population continuity such that samples are assumed to have no equal data values (no zero difference–scores, no tied ranks). The best results across all tests and combinations of simulation parameters were obtained by randomly resolving ties, although there were exceptions. The method of dropping ties and reducing the sample size performed poorly.

Key words: Distribution-free, ties, location-shift, Monte Carlo, Rosenbaum's test, Tukey's quick test, Kolmogorov-Smirnov test, Wilcoxon rank-sum test, Kruskal-Wallis test, Terpstra-Jonckheere test.

Introduction

Distribution-free tests are important in the context of social and behavioral science research because they have less stringent assumptions than parametric statistics. Micceri (1986, 1989) showed that many variables studied in the social and behavioral sciences clearly do not meet distributional assumptions of parametric tests, such as normality or homoscedasticity.

In terms of hypotheses of a pure shift in location parameter combined with a violation of the normality assumption, nonparametric statistics are much more powerful than their parametric counterparts. In many layouts, these advantages are evident with very small samples and improve dramatically as sample sizes increase (Blair & Higgins, 1980, van den Brink & van den Brink, 1989, Sawilowsky, 1990, Sawilowsky & Blair, 1992, Kelley, Sawilowsky, & Blair, 1994, MacDonald, 1999).

Many distribution-free statistics lose efficiency when there is a violation of their underlying assumption of population continuity. In practice, this means the samples are assumed to have no equal data values (no zero difference–scores, no tied ranks), either within groups or between groups. Data in the social and behavioral sciences almost never meet this assumption either because of the inherently discrete nature of the data (Micceri, 1986, 1989) or because of a lack of precision in measurement (Cliff, 1996a, 1996b).

Sparks (1967) conducted one of the few empirical studies to have specifically examined violation of continuity. He investigated Student's *t*-test (Student, 1908) and the Wilcoxon Rank-sum (Mann-Whitney *U*) test (Wilcoxon, 1945, Mann & Whitney, 1947) using discrete approximations to the normal, rectangular, and exponential distributions. Results were similar for both Student's *t*-test and the Wilcoxon-Mann-Whitney test when ties were randomly resolved. The Wilcoxon-Mann-Whitney test, however, produced very conservative results when ties were resolved using mid-ranks.

The practical consequence of violating the assumption of population continuity is that samples will contain equal data values resulting

Dr. Fay is an Assessment Consultant in Wayne County, Michigan where he works with the state and local education agencies in the areas of school improvement, accountability, accreditation, and assessment. His research interests include the study of the properties of statistics through computer-intensive Monte Carlo methods using Fortran.

in zero difference–scores or tied ranks. A useful distinction can be made, however, between consequential (critical, meaningful) and inconsequential (non-critical) ties. Ties can occur in such a way that regardless of how they are resolved they have no effect on the calculation of the test statistic or the resulting inference. Such ties are clearly inconsequential. Ties that occur only within a group, when looking for between group effects, are often of this type. By definition, inconsequential ties may be resolved by any simple procedure that maintains the integrity of the ranks, such as arbitrary assignment in sequence of the set of ranks for which the group of scores is tied. Other ties occur in such a way that different resolutions result in different values of the statistic that may, in turn, result in different inferential decisions. Such ties are clearly consequential.

Purpose of the Study

Even though the less stringent underlying assumptions of distribution-free tests are rarely met in practice, the effects of violation of assumptions on robustness of Type I error rates and power have not been studied extensively. Given the potentially deleterious effects of ties on these tests, and the necessity of dealing with them in some way, a careful investigation of the impact of different methods of resolution is warranted. This is especially true given the subtle nature of robustness (Bradley, 1978, Wilcox, 1998). Therefore, nine methods were used, as applicable, to resolve consequential ties prior to the computation of six statistics.

Fahoome (1999, 2002) studied the Type I error properties of large-sample approximation formulas for twenty nonparametric and/or distribution-free statistics, including the six presented here, using the theoretical standard Normal distribution and four of the Micceri (1986) data sets. Ties, however, were either ignored or resolved in one specific way on a test-by-test basis. These same data sets served as pseudo-population models for the present study.

Tests

The following distribution-free tests were investigated:

1. Kolmogorov-Smirnov Test of General Differences for Two Independent Samples (Kolmogorov, 1933).

2. Rosenbaum's Test of Location for Two Independent Samples (Rosenbaum, 1953, 1954, 1965).

3. Tukey's Quick Test of Location for Two Independent Samples (Tukey, 1959).

4. Wilcoxon-Mann-Whitney Test for Two Independent Samples (Wilcoxon, 1945, Mann & Whitney, 1947, Kruskal, 1957).

5. Kruskal-Wallis Test for k Independent Samples (k = 3 to 6) (Kruskal, 1952, Kruskal & Wallis, 1952).

6. Terpstra-Jonckheere Test of an Ordered Alternative Hypothesis for k Independent Samples (k = 3 to 6) (Terpstra, 1952, Jonckheere, 1954).

Resolution of Ties

The nine methods for dealing with consequential ties (zero difference–scores or tied ranks) were:

1. (M-1) Resolve consequential ties in the manner least favorable to rejection of the null hypothesis and in the manner most favorable to rejection of the null hypothesis, calculate the statistic for each of these resolutions, and then calculate the mid-range (mean) value of these two statistics and use it to conduct the test.

2. (M-2) Count ties as 1/2 (Rosenbaum's Test and Tukey's Quick Test only).

3. (M-3) Alternately resolve each set of tied-for ranks.

4. (M-4) Randomly resolve each set of tied-for ranks.

5. (M-5) Delayed increment (Kolmogorov-Smirnov Test of General Differences only).

6. (M-6) Assign the mid-rank of a set of tied ranks to each score without further correction.

7. (M-7) Weighted average of all possible resolutions (Rosenbaum's Test only).

8. (M-8) Drop matching tied-for ranks and reduce *N* accordingly.

9. (M-9) Drop all tied-for ranks (if possible) and reduce *N* accordingly.

Methods 3, 4, 6, and 9 were described by Bradley (1968) as well as Gibbons and Chakraborti (1992). Methods 1, 2, 5, and 7 were described by Neave and Worthington (1988). Method 1 is related to a method described by Bradley (1968). Method 9 is widely mentioned in textbooks. Method 8 was not encountered in the literature but was added to the study as a variation of Method 9 that preserved equal sample sizes when dropping tied values.

Bradley (1968) also described methods involving calculation of statistics for all possible resolutions of consequential ties, the results being used to establish probability bounds for the test or to calculate a mean probability. Although theoretically attractive, these methods are often impractical, requiring the calculation of very large numbers of statistics and/or the availability of the probabilities (see, however, Fay, 2002, for a discussion of methods for generating critical values and associated probabilities for some of these tests). For many tests, the calculation of an average statistic, based on all possible resolutions of ties, turns out to be equivalent to resolving each set of tied-for ranks using the mid-rank (Neave & Worthington, 1988). Bradley (1968) warned, however, that under some circumstances the use of mid-ranks might give a statistic something closer to its minimum or maximum value rather a median or mean value. This might account for the results in Sparks (1967).

Many of the methods involve schemes for eliminating ties, either by: (a) breaking them, that is, by somehow assigning the available ranks to the tied observations, or (b) dropping them. Other methods, such as mid-ranks, result in modified samples that still contain duplicate (and perhaps non-integer) ranks, even though this cannot happen when all assumptions of the test are met. Averaging the statistics from the least and most likely to reject resolutions can also result in non-integer values of statistics that are normally integer-valued. Such statistics were still referred to a standard table of critical values, for example, Neave (1981), as the performance when used in this manner was a major point of this study. The test/method combinations investigated are shown in Table 1.

Data Sets

A theoretical distribution and four empirical data sets were used as sources of samples. The theoretical standard Normal distribution ($\mu = 0$, $\sigma = 1$) did not produce samples with significant numbers of duplicate data values and thus served as a baseline for the performance of these tests under conditions meeting their underlying continuity assumption. The four empirical data sets, due to Micceri (1986), were (a) Extreme Asymmetric (EA), (b) Extreme Bi-modal (EB), (c) Multi-modal Lumpy (ML), and (d) Smooth Symmetric (SS).

The four Micceri (1986) data sets are inherently discrete and decidedly non-normal (see Appendix, Figures A1 through A4). They were also discussed in Micceri (1989), Sawilowsky, Blair and Micceri (1990), Sawilowsky and Blair (1992), and Fahoome (1999, 2002). With regard to the extreme bimodal data set, Fahoome (1999) concluded:

> [B]ecause of the small number (6) of data points, there were an extremely large number of ties, even for relatively small sample sizes. This data is Likert-type data. The performance by most tests was extremely poor. Most of the tests had inflated Type I error rates, some as high as 0.99999. A few had very low Type I error rates. (p. 462)

In spite of this finding, the extreme bimodal data set was retained for this study because of the widespread existence of such data. Properties of these data sets are given in Table 2.

Table 1. Tests and Applicable Methods of Resolving Ties

	Test					
Method	K-S[a]	R[b]	TQ[c]	W-M-W[d]	K-W[e]	T-J[f]
M-1[g]	X	X	X	X	X	X
M-2[h]	na	X	X	na	na	na
M-3[i]	X	X	X	X	X	X
M-4[j]	X	X	X	X	X	X
M-5[k]	X	na	na	na	na	na
M-6[l]	na	na	na	X	X	X
M-7[m]	na	X	na	na	na	na
M-8[n]	X	X	X	X	X	X
M-9[o]	X	X	X	X	X	X

Note: Cells marked 'na' indicate that the method does not apply to the test.
[a]K-S = Kolmogorov-Smirnov Test, [b]R = Rosenbaum's Test, [c]TQ = Tukey's Quick Test,
[d]W-M-W = Wilcoxon-Mann-Whitney Test, [e]K-W = Kruskal-Wallis Test,
[f]T-J = Terpstra-Jonckheere Test, [g]M-1 = Average of least and most likely to reject,
[h]M-2 = Count ties as ½, [i]M-3 = Alternating, [j]M-4 = Random, [k]M-5 = Delayed Increment,
[l]M-6 = Mid-ranks, [m]M-7 = Weighted average, [n]M-8 = Drop matching, [o]M-9 = Drop all.

Table 2. Properties of Selected Micceri (1986,1989) Data Sets

Data Set	Parameter				
	μ[a]	ϕ[b]	σ[c]	γ_3[d]	γ_4[e]
Extreme Asymmetric	24.50	27.00	5.79	−1.33	4.11
Extreme Bi-modal	2.97	4.00	1.69	−0.08	1.30
Multi-modal Lumpy	21.15	18.00	1.90	0.19	1.80
Smooth Symmetric	13.19	13.00	4.91	0.01	2.66

Note: Excerpted from "A more realistic look at robustness and type II error properties of the t test to departures from population normality," by S. S. Sawilowsky & R. C. Blair, 1992, *Psychological Bulletin, 111*(2), 352-360, Table I, p. 353, copyright 1992 by Psychological Bulletin. Adapted with permission.
[a]μ = mean, [b]ϕ = median; [c]σ = variance, [d]γ_3 = skewness, [e]γ_4 = kurtosis.

Methodology

The simulations were programmed in Fortran 90/95. A main program was built for each of the six tests to conduct both the Type I error and power studies by controlling the combinations of simulation parameters and making calls to the appropriate modules. For each unique combination of distribution, sample size, number of groups (for *k*-sample tests only), and effect size (for power studies only), 1 million samples were drawn. For each sample one- and two-sided tests where conducted at both nominal alpha .01 and .05 for each applicable method of resolving ties (Table 1). Counts were maintained of significant and non-significant results, as well as un-testable trials, until the end of the simulation cycle when they were converted to proportions and written to output files.

Separate programs were written for each of the six tests to conduct the simulations for the drop ties and reduce N methods of resolving ties as these methods often led to tests on unequal sample sizes for which the test statistic could either not be computed or for which critical values were unavailable. This necessitated a modified approach to the simulations in which un-testable samples were discarded and additional samples were drawn until: (a) 10,000 testable samples were obtained, or (b) the program reached its 10,000,000th cycle, whichever came first.

All sample sizes from 3 to 30 [3(1)30] were examined, limited only by the availability of critical values. Because the method of dropping ties and reducing *N* often resulted in unequal sample sizes, this method was only studied for tests where tables of critical values for unequal sample sizes were available (Neave, 1981, Neave & Worthington, 1988) or could be generated (Fay, 2002). Power studies were conducted for equal initial per-group sample sizes of 3(3)30 if Type I error results were satisfactory and critical values were available.

One of the most widely suggested methods for dealing with (consequential) ties is to resolve them in all possible ways, obtaining a value of the statistic (or its associated

probability) for each resolution. A mean value of the statistic is then obtained and tested, or a mean value of the probability established. This method was only implemented for Rosenbaum's test as there was a practical method for doing so. It was not otherwise used in this study because of the practical difficulties involved in implementing it for even moderate sample sizes when there are numerous ties at several different values. Also, comprehensive tables of exact probabilities are even more difficult to obtain than critical value tables for some of these tests.

Bradley (1978) recommended conservative bounds for robust Type I error of nominal alpha ± 10% and liberal bounds of nominal alpha ± 50%. Many distribution-free tests, however, cannot achieve nominal alpha at small sample sizes. The entries in critical value tables are typically best conservative values that may fall below Bradley's recommended 10% lower bound. As the main interest in the Type I error studies was the ability of each test to resist inflation of Type I error rate the conservative and liberal criteria were combined such that Type I error rates were considered acceptable if they fell in the range of .5α to 1.1α or were no more conservative than the results obtained when sampling from the standard Normal distribution.

The power of a test was of no interest if the Type I error rate was not robust to violations of assumptions. A priori, it was expected that those combinations of test conditions that produced Type I error rates well below nominal alpha would also have attenuated power.

For the power studies, a one-sided test was made in the direction of the simulated effect, while significant results in the wrong tail constituted Type III errors (MacDonald, 1999). Pure shift-effects of known size were simulated by shifting one or more of the groups relative to a base group. Nominal effect size multipliers of 0.2, 0.5, 0.8 and 1.2 were planned following Cohen (1988) and Sawilowsky and Blair (1992). Because of the necessity of generating integral shifts with the empirical data sets in order to obtain between-group ties, actual effect size multipliers for each empirical data set differed slightly from these targets, as shown in Table 3. The performance of the six tests with respect to the various methods of resolving ties, when used with such data, was of primary interest in this study.

Statistical Tests

All six tests share the assumptions of: (a) random and independent sampling of continuous populations, with sufficient precision of measurement to avoid tied observations (Bradley, 1968, Conover, 1999), (b) independence of sample observations both within and between groups (Hollander & Wolfe, 1999). All the tests have null hypotheses that assume all samples are drawn from identical populations. Assumptions about the populations under the alternative hypothesis differ for each test. The tests can be used successfully with discrete populations, but become approximate with the tabled critical values generally providing best conservative estimates.

Kolmogorov-Smirnov Test

Background. Neave and Worthington (1988) and Conover (1999) identified this as Smirnov's (1939) application of Kolmogorov's (1933) goodness-of-fit test. Everitt (1998) described it as "A distribution free method that tests for any difference between two population probability distributions. The test is based on the absolute maximum difference between the cumulative distribution functions of the samples from each population" (p. 179). The maximum distance referred to is the vertical distance between the cumulative probability distributions.

Hypotheses. The null hypothesis for the two-sided test is that the two sampled populations have identical distributions. The two-sided alternative hypothesis is simply that the two sampled populations are different in some way. In the case of a one-sided test, the alternative hypothesis is that one population is stochastically greater than the other. Neave (1981) suggested that the test only be used in the two-sided situation, the Wilcoxon–Mann–Whitney test being more powerful for the directional hypothesis.

Table 3. Actual Shifts and Effect Sizes for Nominal Effect Sizes

Data Set (σ[a])	Nominal Effect Size S[b](.2σ)	M[c](.5σ)	L[d](.8σ)	VL[e](1.2σ)
Extreme Asymmetric (5.79)				
NS[f]	1.158	2.895	4.632	6.948
AS[g]	1	3	5	7
AES[h]	0.173σ	0.518σ	0.864σ	1.209σ
Extreme Bi-modal (1.69)				
NS	0.338	0.845	1.352	2.028
AS	n/a	1	n/a	2
AES	n/a	0.592σ	n/a	1.183σ
Multi-modal Lumpy (11.90)				
NS	2.380	5.950	9.520	14.280
AS	2	6	10	14
AES	0.168σ	0.504σ	0.840σ	1.176σ
Smooth Symmetric (4.91)				
NS	0.982	2.455	3.982	5.892
AS	1	2	4	6
AES	0.204σ	0.407σ	0.815σ	1.222σ
Standard Normal (1.00)				
NS	0.200	0.500	0.800	1.200
AS	0.200	0.500	0.800	1.200
AES	0.200σ	0.500σ	0.800σ	1.200σ

Note: Developed based on Cohen (1988) and Sawilowsky and Blair (1992).
[a]σ = Standard deviation, [b]S = Small, [c]M = Medium, [d]L = Large, [e]VL = Very Large.
[f]NS = Nominal Shift, [g]AS = Actual Shift, [h]AES = Actual Effect Size (rounded).

Procedure and Test Statistic.

The following procedure was described in Neave and Worthington (1988). Let there be $N = n_A + n_B$ ranked observations, each designated as an A or B. For the A observations, maintain a count above the letter sequence, starting from zero and incremented by n_B each time an A is encountered. For the B observations, maintain a count below the letter sequence, starting from zero and incremented by n_A each time a B is encountered. The final count for both A's and B's should be $M = n_A \times n_B$. Compute the differences, $d_i = B_i - A_i$, by subtracting the A counts from the B counts for each letter position. Find the absolute value of these differences. For the two-sided test, take $D^* = \max|d_i|$. For a one-sided test take $D_+^* = \max|pos(d_i)|$ or $D_-^* = \max|neg(d_i)|$ depending on what is expected under H_1. Conover (1999) defined the test statistic, T, in terms of two empirical distribution functions, S_A and S_B, using the supremum. For the two-sided test, $T = \sup_x |S_A(x) - S_B(x)|$. For the one-sided test that A < B (stochastically), $T^+ = \sup_x [S_A(x) - S_B(x)]$. Thus, for the one-sided test that A > B (stochastically), $T^- = \sup_x [S_B(x) - S_A(x)]$.

Rejection Region.

Critical regions are usually tabulated as $D^* \geq critical\ value$. Note that $D^* = n_A n_B D$, where D is the statistic derived from a direct comparison of the sample cdf's, is more convenient to work with as it takes only integer values (Neave & Worthington, 1988).

Rosenbaum's Test

Background.

This test first appeared in its current form in Rosenbaum (1954), which was based on Rosenbaum (1953). In both articles, Rosenbaum cited Wilks (1942) as the original source of the formulas for deriving the critical value tables. Rosenbaum (1965) reiterated this earlier work. The test is classified as a runs test. It is a quick and easy test, but is not routinely included in textbooks on nonparametric statistics. Neave and Worthington (1988) presented it as a test for general differences between two sampled populations where spread tends to increase with an increase in the mean, consistent with Rosenbaum (1954). They claimed that under the conditions of an increase in spread with an increase in the median tests such as the Wilcoxon-Mann-Whitney test and Tukey's Quick test have almost no power because of the change in spread. Likewise, tests for spread, such as the Siegel-Tukey test (Siegel & Tukey, 1960), have little or no power because of the change in location. If more general differences were suspected, or needed to be protected against, the Kolmogorov-Smirnov test was suggested as a better choice. Processes that are known to be exponential or Poisson in nature, where the standard deviation is related to the mean, would be excellent candidates for analysis by Rosenbaum's test. Thus, Rosenbaum's test appears to occupy a somewhat unique place among its better-known peers.

Hypotheses.

The null hypothesis is that there is no difference in the two sampled populations. The alternative hypothesis can be two-sided or one-sided. The two-sided alternative hypothesis is simply that the two sampled populations are different in some way. In the case of a one-sided test, the alternative hypothesis is that one population is stochastically greater than the other.

Procedure and Test Statistic.

The following procedure was described in Neave and Worthington (1988). For the two-sided test, determine which sample has the overall greatest value and then count the number of observations in that sample that are greater than the greatest value in the other sample and let this be the test statistic R. For the one-sided test, determine if the greatest overall value comes from the sample whose population is hypothesized under H_1 to have the greater mean. If it does, proceed as for the two-sided test, if not, set $R = 0$.

Rejection Region.

Critical regions are of the form $R \geq critical\ value$. The table of critical values must be entered with n_1 as the size of the sample from which R is calculated and n_2 as the size of the other sample (Neave & Worthington, 1988).

Tukey's Quick Test

Background.

This test first appeared in Tukey (1959). It is a two-sample test constructed according to Duckworth's (1958) portability specifications. It is a quick test because it only requires a few of the sample observations to be ordered. It is also compact, in the sense that tables of critical values are not generally needed for most applications, as only a limited number of critical values occur in practice. These two characteristics combine to make the test portable. Like Rosenbaum's test, Tukey's Quick test is based on extreme runs and is not routinely included in applied textbooks.

Hypotheses.

The test is primarily a test for differences in location of the medians of the two sampled populations and is most appropriate when there is reason to believe that the sampled populations have the same spread, or better, the same shape (Neave & Worthington, 1988). The null hypothesis is that there is no difference in the two sampled populations or no difference in the medians of the populations. The alternative hypothesis can be two-sided or one-sided. The two-sided alternative hypothesis is simply that the two sampled populations are different in some way, or have different medians. In the case

of a one-sided test the alternative hypothesis is that one population is stochastically greater than the other, or that there is a directional difference in the medians.

Procedure and Test Statistic.

The following procedure was described in Neave and Worthington (1988). It begins by arranging the sample observations in a single combined array from least to greatest, keeping track of original sample membership, say A and B, and then ranking them. For a two-sided test, if the minimum and maximum observed values come from the same sample then the test statistic is $T_y = 0$. If the minimum and maximum observed values come from different samples, then the test statistic is the sum of the extreme runs, that is, if the minimum value comes from sample A and the maximum from sample B, then count the number of A's from the beginning of the array until the first B is reached, say C_L, and count the number of B's from the end of the array back until the first A is reached, say C_U, and set $T_y = C_L + C_U$. For a one-sided test, if the minimum and maximum observed values come from the same sample, set $T_y = 0$. If the minimum and maximum observed values come from different samples, determine if the maximum observation comes from the sample that is expected to have the greater median. If not, set $T_y = 0$. If so, calculate T_y just as for the two-sided.

Rejection Region.

Critical regions are of the form $T_y \geq$ *critical value* and tables are available in Neave and Worthington (1988). However, for one-sided tests with sample sizes that are not too small and not too dissimilar, the .05 and .01 critical values are generally 6 and 9, respectively. For a two-sided test under the same conditions, the .05 and .01 critical values are generally 7 and 10, respectively. These critical values are reported to work well for ratios of sample sizes from 1 to 1.5. Equal sample sizes are not required, although tables of critical values should be employed when the ratio of larger to smaller sample exceeds 1.5 (Tukey, 1959).

Wilcoxon-Mann-Whitney Test

Background.

Wilcoxon (1945) introduced the rank-sum version of this test for equal sample sizes in the same article as the signed-rank test, while Mann and Whitney (1947) independently developed the Mann–Whitney *U* test. The two versions are procedurally different but mathematically equivalent and are often referred to jointly in the literature as the Wilcoxon-Mann-Whitney test (Sprent & Smeeton, 2001). The test is applied to ordinal data. Tables of critical values are more commonly available for the Mann-Whitney version of the test. In either form this is one of the better-known distribution-free tests, and is the one that corresponds most directly to Student's *t*-test for two independent samples (Student, 1908). It is also a powerful test, with an asymptotic relative efficiency that never falls below 0.864 with respect to the *t*-test (Lehmann, 1998), although it is often much more powerful under conditions that violate the assumptions of the *t*-test, yet respect its own assumptions (Blair & Higgins, 1980).

The Wilcoxon-Mann-Whitney test is generally regarded as a test of whether two independent samples represent the same population versus populations that differ in location, either of their medians or with respect to the rank ordering of their scores (Sheskin, 1997). Bergmann, Ludbrook, and Spooren (2000) described it as a test of group mean ranks or, equivalently, rank sums, for testing two different hypotheses: (a) a shift in otherwise identical populations, or (b) a difference in mean ranks between randomized groups. A detailed theoretical treatment of the test was given in Lehmann (1998). Kruskal (1957) detailed the history of the test from 1941 to 1957.

Hypotheses.

The alternative hypothesis under the population model assumes that the populations have identical probability distributions other than a constant shift (Sheskin, 1997), also known as a translation, or location–shift, model. If F and G are the population distribution functions, the location-shift model requires $G(x) = F(x - \Delta), \forall x$. The null hypothesis is then H_0: $[\Delta = 0]$ (Hollander & Wolfe, 1999). The null hypothesis can also be

stated as no difference in the medians of the populations, or H_0: [$\phi_1 = \phi_2$] (Neave & Worthington, 1988). With equal sample sizes, this is equivalent to the hypothesis that the sum of ranks for each group is the same, or H_0: [$\sum R_1 = \sum R_2$]. For unequal sample sizes this generalizes as the mean rank of the groups being equal, or H_0: [$\bar{R}_1 = \bar{R}_2$] (Sheskin, 1997). The parallel to Student's *t*-test is most evident in this form.

The test can be one-sided or two-sided. The two-sided alternative hypothesis for shift is H_1: [$\Delta \neq 0$] (Hollander & Wolfe, 1999) and the alternative hypothesis for medians is H_1: [$\phi_1 \neq \phi_2$] (Neave & Worthington, 1988). The alternative hypotheses for ranks are H_1: [$\sum R_1 \neq \sum R_2$] or H_1: [$\bar{R}_1 \neq \bar{R}_2$] (Sheskin, 1997). For a one-sided test, the alternative hypotheses for shift are either H_1: [$\Delta < 0$], or H_1: [$\Delta > 0$] (Hollander & Wolfe, 1999). The alternative hypotheses for medians are H_1: [$\phi_1 < \phi_2$] or H_1: [$\phi_1 > \phi_2$] (Neave & Worthington, 1988). The alternative hypotheses for ranks are H_1: [$\sum R_1 < \sum R_2$], H_1: [$\sum R_1 > \sum R_2$], H_1: [$\bar{R}_1 < \bar{R}_2$] or H_1: [$\bar{R}_1 > \bar{R}_2$] (Sheskin, 1997).

Procedure and Test Statistic.

Siegel and Castellan (1988) and Neave and Worthington (1988) described the Wilcoxon version of the test. Given two samples, A and B, with $N = n_A + n_B$, combine the observations in a single array, keeping track of original sample membership, and then rank them from 1 to *N*. Compute R_A as the sum of the ranks of the observations from sample A and R_B as the sum of the ranks of the observations from sample B. The test statistic, *W*, is the rank sum that would be expected to be smaller if H_1 were true.

Rejection Region.

Tables of critical values are usually given for the Mann-Whitney *U* test (Neave & Worthington, 1988, Sheskin, 1997), with critical regions of the form $U_{min} \leq$ *critical value* representing best conservative values. The test can be applied to unequal sample sizes with appropriate critical value tables. Because they are mathematically equivalent, the results of the Wilcoxon procedure can be converted to values of *U*. Neave and Worthington (1988) gave the conversion for a two-sided test as: $U = \min\left[U_A,\ U_B\right]$, with $U_A = R_A - \frac{1}{2} n_A \left(n_A + 1\right)$ and $U_B = n_A n_B - U_A = R_B - \frac{1}{2} n_B \left(n_B + 1\right)$. For a one-sided test, use either U_A or U_B according to which one is expected to have the smaller value under H_1. Converting to values of *U* also accounts for the effect of unequal sample sizes.

Kruskal-Wallis Test

Background.

This test was introduced in Kruskal (1952) and Kruskal and Wallis (1952). Vogt (1999) described it as, "A nonparametric test of statistical significance used when testing more than two independent samples. It is an extension of the Mann-Whitney *U* test, and of the Wilcoxon [rank-sum test], to three or more independent samples. It is a nonparametric one-way ANOVA for rank order data" (p. 151).

Everitt (1998) described the test as a "distribution free method that is the analogue of the analysis of variance of a one-way design. It tests whether the groups to be compared have the same population median" (p. 180). The test is applied to ordinal (rank ordered) data (Sheskin, 1997). Power comparisons with the *F*-test are very favorable. Conover (1999) gave the following asymptotic relative efficiencies for the Kruskal-Wallis test relative to the *F*-test: (a) For distributions that differ only in their means, never less than 0.864, but as high as infinity, (b) for Normal populations, 0.955, (c) for uniform distributions, 1.0, and (d) for exponential distributions, 1.5.

Hypotheses.

For *k* groups, the population distribution functions, $F_1,\ldots,F_k$ are assumed to have the relationship $F_j\left(x\right) = F\left(x - \tau_j\right), -\infty < j < \infty$ over all *j* (*j* = 1 to *k*) where *F* is a continuous distribution function with unknown median and τ_j is the unknown treatment effect for the *j*th population (Hollander & Wolfe, 1999). The null hypothesis can be stated as no difference in the medians of the populations, H_0: [$\phi_1 = \phi_2 = \ldots = \phi_n$] (Neave & Worthington,

1988, Siegel & Castellan, 1988), identical populations, H_0: [All of the *k* population distribution functions are identical] (Conover, 1999) or identical treatment effects, H_0: [$\tau_1 = \tau_2 = \cdots = \tau_k$] (Hollander & Wolfe, 1999). The alternative hypothesis assumes that the populations differ only in location (Sprent & Smeeton, 2001) and that at least one of the populations, medians or treatment effects is different from the others.

Vargha and Delaney (1998) took exception to the use of the Kruskal-Wallis test with the foregoing assumptions on the grounds that the attendant hypotheses, while mathematically correct, were too narrow to be of practical value to researchers. They claimed that the Kruskal-Wallis test "cannot detect with consistently increasing power any alternative other than exceptions to stochastic homogeneity" (p.170). This, in turn, is mathematically equivalent to the "equality of expected values of the rank sample means" (p.170). They argued that the requirement for identical distributions under H_0 is too strict, and that only variance homogeneity is needed. Further, they asserted that the H_1 to which the test is actually sensitive is "the tendency for observations in at least one of the populations to be larger (or smaller) than all the remaining populations together" (p. 186).

The test is two-sided with an omnibus alternative hypothesis for shift of H_1: [$\tau_1,\ldots,\tau_k$ not all equal] (Hollander & Wolfe, 1999), H_1: [not all of ϕ_1, ϕ_2, …, ϕ_k are equal] (Neave & Worthington, 1988, Siegel & Castellan, 1988) or H_1: [At least one of the populations tends to yield larger observations than at least one of the other populations] (Conover, 1999). All of these hypotheses can be formulated in terms of rank-sums (for the equal sample size case) or mean ranks (for the general case) as H_0: [$\sum R_1 = \sum R_2 = \ldots = \sum R_k$] or H_0: [$\overline{R}_1 = \overline{R}_2 = \ldots = \overline{R}_k$], with the alternative hypothesis of H_1: [not H_0] (Sheskin, 1997). The alternative hypothesis is stated in this way because it only requires that some pair of groups be different, not that all groups are different, consistent with Conover (1999).

Procedure and Test Statistic.

The general procedure, which does not assume equal sample sizes, is to combine the samples and rank the observations while keeping track of original group membership. For each of the *k* groups, let the number of observations be n_i $(i = 1,2,\ldots,k)$ such that the total number of observations is $N = \sum_{i=1}^{k} n_i$. Calculate the rank-sum for each group as $s_i = \sum_{j=1}^{n_i} r_{ij}$, where r_{ij} is the rank assigned to the jth observation in the *i*th group. The sum of the mean squared ranks is calculated as $S_k = \sum_{i=1}^{k}\left(\frac{s_i^2}{n_i}\right)$. The statistic is then calculated as $H = \frac{12}{N(N+1)}S_k - 3(N+1)$. This is the common computational formulation (Sprent & Smeeton, 2001, Neave & Worthington, 1988, Feir-Walsh & Toothaker, 1974, Siegel & Castellan, 1988, Conover, 1999).

Conover (1999) defined the test statistic as $T = \frac{1}{S^2}\left(S_k - \frac{N(N+1)^2}{4}\right)$ where S_k and N are as defined above and $S^2 = \frac{1}{N-1}\left(\sum_{\substack{all\\ranks}} R(X_{ij})^2 - N\frac{(N+1)^2}{4}\right)$. He noted that S^2 simplified to $N(N+1)/12$ in the absence of ties such that $T = H$ as defined above. H can also be defined as $H = \frac{12}{N(N+1)}\sum_{i=1}^{k} n_i\left(\overline{R}_i - \overline{R}\right)^2$, where n_i is as above, $\overline{R}_i$ is the mean rank of group *i*, and $\overline{R}$ is the overall mean rank of the *N* total observations (Neave & Worthington, 1988, Siegel & Castellan, 1988). In this form it can be seen most clearly that the statistic is a weighted sum of squared deviations. Post-hoc procedures using pairwise comparisons are available (Conover, 1999, Sheskin, 1997, Siegel & Castellan, 1988), but are not considered further here.

Rejection Region.

Critical regions are of the form $H \geq \textit{critical value}$ (Neave & Worthington, 1988). Approximate critical values can be obtained from a chi-squared distribution with $k - 1$ degrees-of-freedom, but see Fahoome (1999, 2002). The test will work with unequal sample sizes since the calculation of the statistic involves a weighted sum of squares of differences between group mean ranks and the overall mean rank, although critical value tables tend to be limited (Neave, 1981).

Terpstra-Jonckheere Test

Background.

The Terpstra-Jonckheere test was developed independently by Terpstra (1952) and Jonckheere (1954). Like the Kruskal-Wallis test, it is an extension of the Wilcoxon-Mann-Whitney test on ranks for the one-way design. It differs from the Kruskal-Wallis test in that it postulates a specific ordering of the groups under the alternative hypothesis based on prior knowledge, that is, that the situation being tested supports an a priori expectation of a specific, identifiable order of the population medians based on the experimental design, not on the observed data (Hollander & Wolfe, 1999, Siegel & Castellan, 1988). A general assumption is that all of the possible assignments of joint ranks are equally possible (Hollander & Wolfe, 1999).

Hypotheses.

For k groups, the population distribution functions, $F_1,\ldots,F_k$ are assumed to have the relationship $F_j(x) = F(x - \tau_j), -\infty < x < \infty$ over all j, (j = 1 to k), where F is a continuous distribution function with unknown median and τ_j is the unknown treatment effect for the jth population (Hollander & Wolfe, 1999). The null hypothesis can be stated in terms of medians as H_0: [$\phi_1 = \phi_2 = \ldots = \phi_k$] (Neave & Worthington, 1988, Siegel & Castellan, 1988), identical populations as H_0: [$F_1(x) = F_2(x) = \cdots = F_k(x), \forall x$] (Sprent & Smeeton, 2001), or treatment effects as H_0: [$\tau_1 = \tau_2 = \cdots = \tau_k$] (Hollander & Wolfe, 1999). If the k groups are numbered to correspond to the expected order, the alternative hypothesis is one-sided and given by H_1: [$\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k$, with at least one strict inequality] (Hollander & Wolfe, 1999), H_1: [$F_1(x) \leq F_2(x) \leq \cdots \leq F_k(x)$, at least one inequality strict for some x] (Sprent & Smeeton, 2001), or H_1: [$\phi_1 \leq \phi_2 \leq \ldots \leq \phi_k$, at least one of the inequalities is strict] (Neave & Worthington, 1988, Siegel & Castellan, 1988).

Procedure and Test Statistic.

The procedure calculates the Mann-Whitney U statistic for all pairs of samples and then combines the results. If the Wilcoxon rank-sum procedure is used the resulting statistics must be converted to Mann-Whitney U statistics before being combined. For the alternative hypothesis, as stated above, the test statistic was given by Neave and Worthington (1988) as

$$J = U_{21} + U_{31} + \ldots + U_{k1} + U_{32} + \ldots + U_{ij} + \ldots + U_{k(k-1)} = \sum_{j=1}^{k-1} \sum_{i=j+1}^{k} U_{ij},$$

where U_{ij} represents the Mann-Whitney U statistic for each pair of samples, computed in the order dictated by H_1 to give the least value of each U_{ij}. This is consistent with Siegel and Castellan (1988) and others. To the extent that H_1 tends to be true, each of the U_{ij} will tend to be small and thus their sum will tend to be small.

For k groups there will be $k(k - 1)/2$ values of U. Hollander and Wolfe (1999) gave the Mann-Whitney procedure for calculating the values of U directly, including an adjustment for ties (equivalent to using mid-ranks in the Wilcoxon version of the procedure) as

$$U_{uv} = \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \phi^*\left(X_{iu}, X_{jv}\right), 1 \leq u < v \leq k,$$

where

$$\phi^*(a,b) = \begin{cases} 1 \text{ if } a < b \\ \frac{1}{2} \text{ if } a = b \\ 0 \text{ if } a > b \end{cases}.$$

This is consistent with Siegel and Castellan

Table 4. Test / Method Combinations with Acceptable Type I Error Results

	Test					
Method	K-S[a]	R[b]	TQ[c]	W-M-W[d]	K-W[e]	T-J[f]
M-1[g]	EA,-- ML,SS	EA,-- ML,SS	--,-- ML,--	EA,EB ML,SS	--,-- ML,(SS)	EA,EB ML,SS
M-2[h]	na	EA,-- ML,SS	--,-- ML,--	na	na	na
M-3[i]	EA,-- ML,SS	--,-- ML,SS	--,-- ML,--	--,EB ML,SS	--,EB ML,SS	EA,EB ML,SS
M-4[j]	EA,EB ML,SS	EA,EB ML,SS	EA,-- ML,--	EA,EB ML,SS	EA,EB ML,SS	EA,EB ML,SS
M-5[k]	--,-- ML,--	na	na	na	na	na
M-6[l]	na	na	na	EA,EB ML,SS	EA,EB ML,SS	EA,EB ML,SS
M-7[m]	na	--,-- ML,SS	na	na	na	na

Note. EA = Extreme Asymmetric Data Set, EB = Extreme Bi-modal Data Set, ML = Multi-modal Lumpy Data Set, SS = Smooth Symmetric Data Set.
[a]K-S = Kolmogorov-Smirnov Test, [b]R = Rosenbaum's Test, [c]TQ = Tukey's Quick Test,
[d]W-M-W = Wilcoxon-Mann-Whitney Test, [e]K-W = Kruskal-Wallis Test,
[f]T-J = Terpstra-Jonckheere Test.
[g]M-1 = Average of least and most likely to reject, [h]M-2 = Count ties as ½, [i]M-3 = Alternating,
[j]M-4 = Random, [k]M-5 = Delayed Increment, [l]M-6 = Mid-ranks, [m]M-7 = Weighted average.

(1988). The test is approximate when ties are present.

Rejection Region.

Critical regions are of the form $J \leq \textit{critical value}$. The test supports unequal samples sizes and more extensive critical value tables are available as Table *R* in Neave and Worthington (1988). As the sample size increases, the null distribution of *J* becomes asymptotically normal. Formulas exist for obtaining approximate critical values (Neave & Worthington, 1988, Siegel & Castellan, 1988), but see Fahoome (1999, 2002).

Results

Type I Error Results

Question 1: For samples drawn from the same population, is the Type I error rate maintained between .5α and 1.1α for each combination of test, method, number of groups, directionality, sample size, and distribution?

Combinations of tests, methods and Micceri (1986) data sets that demonstrated acceptable Type I Error rates are shown in Table 4. Results for the theoretical standard Normal distribution are not shown, as it did not produce ties. Note, however, that the performance of these tests with the theoretical Normal distribution was not always acceptable due to the

Table 5. Preferred Methods[k, l, m, n, o, p] by Test and Micceri (1986) Data Set

	Test					
Data Set	K-S[a]	R[b]	TQ[c]	W-M-W[d]	K-W[e]	T-J[f]
EA[g]	M-4, M-1	M-1/ M-2/ M-4	na	M-4	M-4, M-6	M-4
EB[h]	na	na	na	M-4	M-4/ M-6	M-4
ML[i]	M-4	M-3	M-4	M-3	M-4/ M-6, M-1	M-4
SS[j]	M-4	M-3	M-4	M-4, M-3	M-4/ M-6, M-1	M-4

Note. A/B indicates very similar results, A, B indicates A better than B.
[a]K-S = Kolmogorov-Smirnov Test, [b]R = Rosenbaum's Test, [c]TQ = Tukey's Quick Test.
[d]W-M-W = Wilcoxon-Mann-Whitney Test, [e]K-W = Kruskal-Wallis Test.
[f]T-J = Terpstra-Jonckheere Test.
[g]EA = Extreme Asymmetric Data Set, [h]EB = Extreme Bi-modal Data Set,
[i]ML = Multi-modal Lumpy Data Set, [j]SS = Smooth Symmetric Data Set.
[k]M-1 = Average of least and most likely to reject, [l]M-2 = Count ties as ½, [m]M-3 = Alternating,
[n]M-4 = Random, [o]M-5 = Delayed Increment, [p]M-6 = Mid-ranks.

discrete nature of the statistics and the use of best conservative critical values whose probabilities were sometimes less than 0.5α. Following Bradley (1978), Type I error performance was judged to be acceptable if it was not inflated beyond 1.1α and was not more conservative than the corresponding performance with the theoretical Normal distribution. As shown in Table 4, the random method provided acceptable Type I error rates for the largest combination of tests and distributions. Most of the other methods provided acceptable results for specific combinations of test and data set with the exception of Methods 8 and 9 (not shown).

Method 9, the drop all ties and reduce N method, is one of the most widely recommended, especially in textbooks, for situations where there are not too many ties. But how many is too many? Methods 8 and 9 are absent from Table 4 because the Type I error results were unacceptable across all combinations of tests and simulation parameters.

Power Results

The remaining research questions were only studied for those combinations of test, method, number of groups, directionality, sample size and distribution for which Question 1 was answered in the affirmative as shown in Table 4. In order to answer the 3rd and 4th research questions it was necessary to analyze the power results from a large number of simulation runs in a manner that might permit determination of the order of preference of methods across various combinations of simulation parameters for each test.

Table 6. Best Method[g, h, i, j] By Test Across Distributions

K-S[a]	R[b]	TQ[c]	W-M-W[d]	K-W[e]	T-J[f]
M-4[i]	M-3[h]	M-4	M-4	M-4, M-3	M-4, M-6[j], M-1[g]

Note. A, B indicates A better than B.
[a]K-S = Kolmogorov-Smirnov Test, [b]R = Rosenbaum's Test, [c]TQ = Tukey's Quick Test,
[d]W-M-W = Wilcoxon-Mann-Whitney Test, [e]K-W = Kruskal-Wallis Test,
[f]T-J = Terpstra-Jonckheere Test.
[g]M-1 = Average of least and most likely to reject, [h]M-3 = Alternating, [i]M-4 = Random.
[j]M-6 = Mid-ranks.

Question 2: For samples drawn from populations differing only in location, is there a preferred method of resolving tied ranks for each combination of test and data set, irrespective of the number of groups, directionality, and sample size?

As shown in Table 5, the random method was the preferred method (13 of 20), or tied for first (4 of 20), for the vast majority of combinations of test and data set (17 of 20). The method of analysis employed for this purpose involved ranking the power results across methods for each specific combination of test, number of groups, nominal alpha level and distribution at each combination of nominal effect size multiplier and initial sample size. Mean ranks were then calculated in three ways: (a) by summing across nominal effect size multipliers at each initial sample size, (b) by summing across initial sample sizes at each nominal effect size multiplier, and (c) by summing across both nominal effect size multipliers and initial sample sizes.

Question 3: For samples drawn from populations differing only in location, is there a preferred method of resolving tied ranks for each test, irrespective of the number of groups, directionality, sample size, and data set?

This question requires a conclusion to be drawn about the relative behavior of the methods across data sets. The results of the preceding analysis were used to determine the number of first place finishes for each test for each combination of method and distribution across nominal alpha and number of groups. If a particular method consistently had the most first place finishes for a particular test, across data sets, then it could in some sense be considered the best method for that test/data set combination. As shown in Table 6, random resolution of ties was clearly superior for four of the six tests, and a close second for another.

Question 4: Is there a best method for resolving ties across all tests and data sets in the study?

Given the results presented in Tables 4, 5, and 6, random resolution of ties performs best across the set of tests, data sets and methods examined in this study.

Conclusion

This study examined various methods of resolving equal data values (tied ranks) in a set of distribution-free statistical tests of location or general difference for *k* independent samples using Monte Carlo simulations with theoretical Normal and discrete, non-normal data. These tests were all based on the assumption of continuity in the underlying population. As such, the presence of ties—which occurred frequently with the discrete, non-normal data sets—and the

efficacy of various methods of resolving them were of theoretical and practical interest.

Of the methods investigated for resolving ties, random resolution seemed to perform best, in the sense of guarding against inflation of Type I error rates while maintaining power, for the majority of combinations of simulation parameters, but not all. This is of interest both theoretically and practically. First, although random resolution might be expected to produce the best results on theoretical grounds, it does not always do so. There are also strong objections in practice to resolving ties at random as the outcome of any particular test then depends on a secondary random event. But what are the consequences of the alternatives if random resolution is rejected on these grounds? How well do the common alternatives, such as mid-ranks or dropping tied values, work?

The often-recommended method of dropping tied values and reducing the sample size performed very poorly across all combinations of simulation parameters. Based on the results of this study, this method should not be used. All of these tests and methods also performed poorly with Likert scale data (i.e., Micceri, 1986, Extreme Bi-modal data set). They should not be used with discrete population data sets that contain relatively few distinct values.

References

Bergmann, R., Ludbrook, J., & Spooren, W. P. J. M. (2000). Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages. *The American Statistician, 54*(1), 72-77.

Blair, R. C., & Higgins, J. J (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distributions. *Journal of Educational Statistics, 5*(4), 309-335.

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

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

Cliff, N. (1996a). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research, 31*(3), 331-350.

Cliff, N. (1996b). *Ordinal methods for behavioral data analysis.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: John Wiley and Sons Inc.

Duckworth, W. E. & Wyatt, J. V. (1958). Rapid statistical techniques for operations research workers. *Operations Research Quarterly, 9,* 218-233.

Everitt, B. S. (1998). *The Cambridge dictionary of statistics*. Cambridge, England: Cambridge University Press.

Fahoome, G. F. (1999). A Monte Carlo study of twenty-one nonparametric statistics with normal and nonnormal data. Unpublished doctoral dissertation, Wayne State University, Detroit, MI.

Fahoome, G. F. (2002). Review of twenty nonparametric statistics and their large sample approximations. *Journal of Modern Applied Statistical Methods, 1*(2), 248-268.

Fay, B. R. (2002). JMASM4: Critical values for four nonparametric and/or distribution-free tests of location for two independent samples. *Journal of Modern Applied Statistical Methods, 1*(2), 489-517.

Feir-Walsh, B. J. & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement, 34*(4), 789-799.

Gibbons, J. D., & Chakraborti, S. (1992). *Nonparametric statistical inference* (3rd ed. Vol. 131). New York: Marcel Dekker, Inc.

Hollander, M., & Wolfe, D. (1999). *Nonparametric statistical methods* (2nd ed.). New York: John Wiley and Sons, Inc.

Jonckheere, A. R. (1954). A distribution free k-sample test against ordered alternatives. *Biometrika, 41*, 133-145.

Kelley, D. L., Sawilowsky, S. S., & Blair, R. C. (1994, October). Comparison of ANOVA, McSweeney, Bradley, Harwell-Serlin, and Blair-Sawilowsky tests in the balanced 2x2x2 layout. Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago, IL.

Killian, C. R., & Hoover, H. D. (1974, April). An investigation of selected two-sample hypothesis testing procedures when sampling from empirically based test score models. Paper presented at the 59th annual meeting of the American Educational Research Association, Chicago, IL.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari, 4*, 83-91.

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

Kruskal, W. H. (1957). Historical notes on the Wilcoxon unpaired two-sample test. *Journal of American Statistical Association, 52*, 356-360.

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

Lehmann, E. L. (1998). *Nonparametrics: Statistical methods based on ranks* (1st (Revised) ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

MacDonald, P. (1999). Power, Type I, and Type III error rates of parametric and nonparametric statistical tests. *Journal of Experimental Education, 67*(4), 369-379.

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

Micceri, T. (1986, November). A futile search for that statistical chimera of normality. Paper presented at the annual meeting of the Florida Educational Research Association, Tampa, FL.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

Neave, H. R. (1981). *Elementary statistics tables*. London: Unwin Hyman Ltd.

Neave, H. R. & Worthington, P. L. B. (1988). *Distribution-free tests.* London: Unwin Hyman Ltd.

Rosenbaum, S. (1953). Tables for a nonparametric test of dispersion. *Annals of Mathematical Statistics, 24*, 663-668.

Rosenbaum, S. (1954). Tables for a nonparametric test of location. *Annals of Mathematical Statistics, 25*, 146-150.

Rosenbaum, S. (1965). On some two-sample non-parametric tests. *Journal of American Statistical Association, 60*, 1118-1126.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60*(1), 91-126.

Sawilowsky, S. S. & Blair, R. C. (1992). A more realistic look at robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin, 111*(2), 352-360.

Sawilowsky, S. S., Blair, R. C., & Micceri, T. (1990). A PC FORTRAN subroutine library of psychology and education data sets. *Psychometrika, 55*(4), 729.

Sheskin, D. J. (1997). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, Fl: CRC Press.

Siegal, S. & Tukey, J. W. (1960). A nonparametric sum of ranks procedure for relative spread in unpaired samples. *Journal of American Statistical Association, 55*, 429-445.

Siegel, S. & Castellan, N. J., Jr., (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). Boston: McGraw-Hill.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin of Mathematics University of Moscow, 2*(2), 3-14.

Sparks, J. N. (1967). Effects of Inapplicability of the continuity condition upon the probability distributions of selected statistics and their implications for research in education (Final report for project no. BR-6-8467 PA-24). Pennsylvania State Univ. (RIE SYN71840) (ERIC Document Reproduction Service No. ED021317)

Sprent, P. & Smeeton, N. C. (2001). *Applied nonparametric statistical methods* (3rd ed.). Boca Raton, Fl: Chapman and Hall / CRC.

Student [W. S. Gosset], (1908). The probable error of a mean. *Biometrika, 6*, 1-25.

Terpstra, T. J. (1952). A non-parametric k sample test and its connection with the H-test (S-92, VP2). Amsterdam: Mathematical Center.

Tukey, J. W. (1959). A quick, compact, two-sample test to Duckworth's specifications. *Technometrics, 1*, 31-48.

van den Brink, W. P. & van den Brink, S. G. (1989). A comparison of the power of the t test, Wilcoxon's test, and the approximate permutation test for the two-sample location problem. *British Journal of Mathematical and Statistical Psychology, 42*(2), 183-189.

Vargha, A. & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics, 23*(2), 170-192.

Vogt, P. W. (1999). Dictionary of statistics and methodology: *A nontechnical guide for the social sciences* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Wilcox, R. R. (1998). The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology, 51*, 1-39.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics, 1*, 80-83.

Wilks, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *Annals of Mathematical Statistics, 13*, 400-409.

Appendix

Micceri (1986) data sets (see Sawilowsky & Blair, 1992):

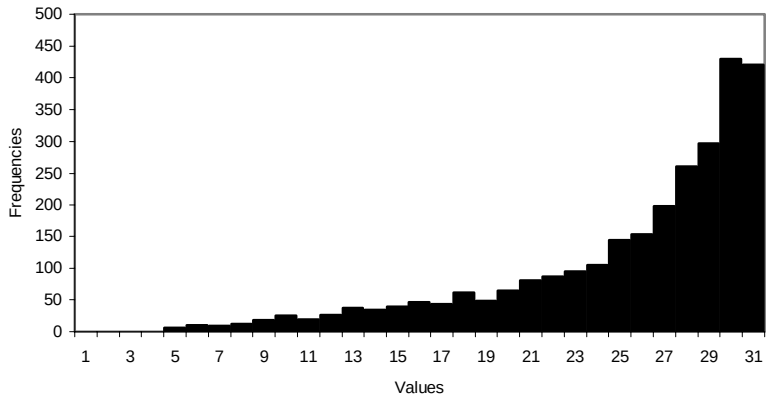

Figure A1. Micceri (1986) extreme asymmetric data set. See Sawilowsky & Blair (1992).

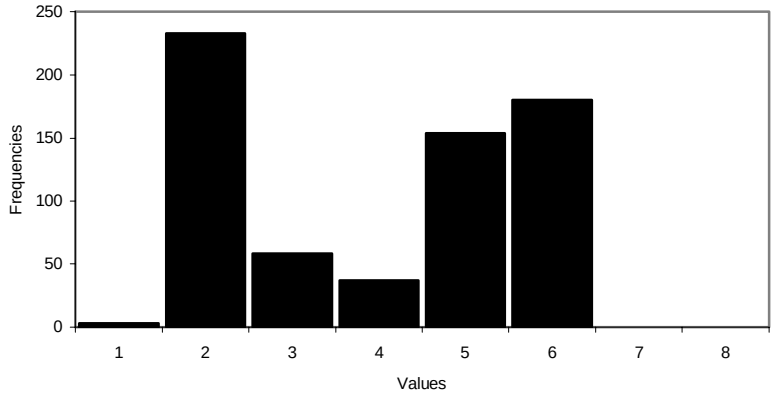

Figure A2. Micceri (1986) extreme bi-modal data set. See Sawilowsky & Blair (1992).

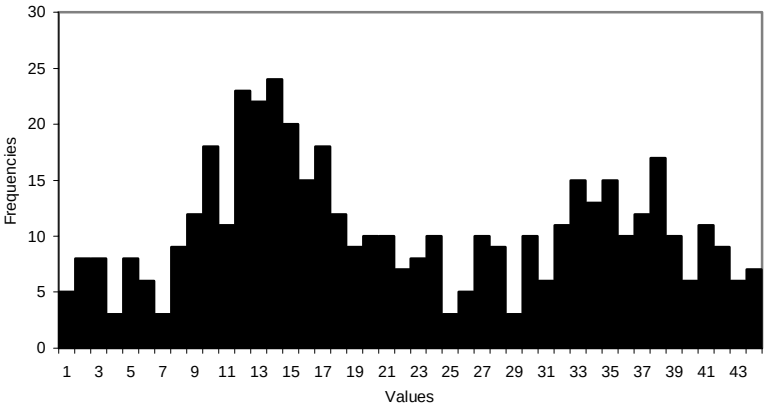

Figure A3. Micceri (1986) multi-modal lumpy data set. See Sawilowsky & Blair (1992).

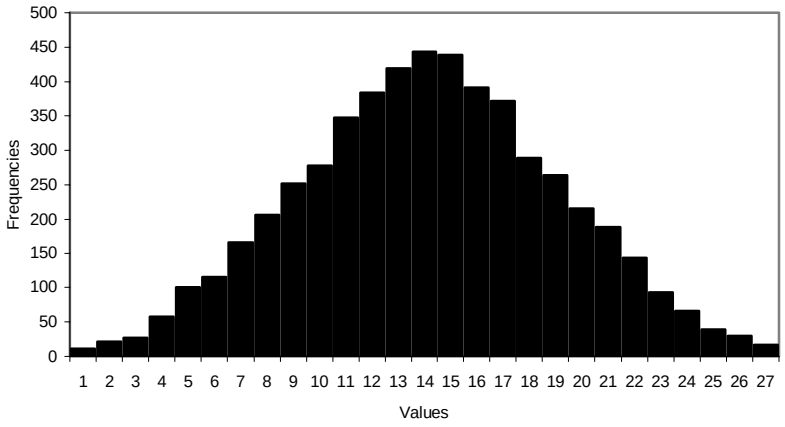

Figure A4. Micceri (1986) smooth symmetric data set. See Sawilowsky & Blair (1992).