Biom. J. 38 (1996) 5, 613–630 Akademie Verlag

Estimation of the Power of the Kruskal-Wallis Test

MICHELLE MAHONEY

Medical Center
Rochester, Minnesota, U.S.A.

RHONDA MAGEL

Department of Statistics
North Dakota State University
Fargo, North Dakota, U.S.A.

Summary

Power calculations of a statistical test require that the underlying population distribution(s) be completely specified. Statisticians, in practice, may not have complete knowledge of the entire nature of the underlying distribution(s) and are at a loss for calculating the exact power of the test. In such cases, an estimate of the power would provide a suitable substitute. In this paper, we are interested in estimating the power of the Kruskal-Wallis one-way analysis of variance by ranks test for a location shift.

We investigated an extension of a data-based power estimation method presented by COLLINGS and HAMILTON (1988), which requires no prior knowledge of the underlying population distributions other than necessary to perform the Kruskal-Wallis test for a location shift. This method utilizes bootstrapping techniques to produce a power estimate based on the empirical cumulative distribution functions of the sample data. We performed a simulation study of the extended power estimator under the conditions of $k = 3$ and $k = 5$ samples of equal sizes $m = 10$ and $m = 20$, with four underlying continuous distributions that possessed various location configurations. Our simulation study demonstates that the Extended Average X & Y power estimation method is a reliable estimator of the power of the Kruskal-Wallis test for $k = 3$ samples, and a more conservative to a mild overestimator of the true power for $k = 5$ samples.

Key words: Nonparametric; Unknown distributions; Simulation study.

1. Introduction

The power of a statistical hypothesis test is defined as the probability of rejecting the null hypothesis, in favor of the alternative hypothesis, given the alternative hypothesis is true. In essence, it is the probability that a statistical test will lead to a correct decision, only if the alternative hypothesis is indeed true. Practitioners are particularly interested in the power of a test. The power of a test aids in the determination of use or non-use of which statistical test to apply, when there is more than one test applicable. Clearly, a practitioner would generally choose the

statistical test which has the highest power. Practitioners frequently use the power of a test in research proposals. They often determine the sample size(s) necessary to conduct an experiment, on the basis of the power of a statistical test and a predetermined level of significance. Unfortunately, the power of a statistical test may be unattainable. In the event that the exact power of a test cannot be calculated, estimates of the power may provide suitable substitutes. In this paper, we are interested in estimating the power of the Kruskal-Wallis oneway analysis of variance by ranks test.

The Kruskal-Wallis test was formally introduced, in its entirety, by KRUSKAL and WALLIS (1952). Let $F_1, F_2, \ldots, F_k$ denote $k \geq 3$ unknown cumulative population distributions (cdf's), from which k independent random samples of sizes $n_1, n_2, \ldots, n_k$ are drawn, respectively. Suppose that it is known that $F_1, F_2, \ldots, F_k$ differ with respect to location only, i.e., shape and dispersion of the population distributions are indentical. Therefore, there exists a vector, $\underline{\delta} = (\delta_1, \delta_2, \ldots, \delta_k)'$, such that $F_i(x) = F(x - \delta_i)$, for all x and $i = 1, 2, \ldots, k$. Without loss of generality, assume that $0 \leq \delta_1 \leq \ldots \leq \delta_k$. The null and alternative hypotheses, that we are interested in testing are, respectively:

H_0: The k populations are identical;

H_a: At least one of the k populations differ with respect to location.

Equivalently, in terms of the vector, $\underline{\delta} = (\delta_1, \delta_2, \ldots, \delta_k)'$, and without loss of generality, assume that $\delta_1 = 0$:

H_0: Each $\delta_i = 0$, for $i = 1, 2, \ldots, k$.

H_a: At least one δ_i differs from 0, for $i = 1, 2, \ldots, k$.

One requirement needed to calculate the power of a statistical hypothesis test, is complete specification, or knowledge, of the underlying population distribution(s). The Kruskal-Wallis test is a nonparametric statistical method, that usually is applied when the underlying population distributions are unknown. Thus, we cannot readily calculate the exact power of the Kruskal-Wallis test for the case of unknown underlying populations. Presently, no method exists for estimation of the power of the Kruskal-Wallis test in the unknown case. Our goal then, is to try to find a reliable estimator of the power of the Kruskal-Wallis test, that is free from any underlying distributional assumptions.

Power calculations for the Kruskal-Wallis test have posed a dilemma for statisticians and practitioners in the unknown case. Power investigations for nonparametric techniques in general, were difficult when the Kruskal-Wallis test was formally introduced in 1952. LEHMANN (1953, 1975) expressed the notion that power calculations were complicated by the infinite number of combinations of location configurations, sample sizes, and number of samples. Furthermore, he suggested that even if a formula for the exact power for a particular situation could be derived, it would be too complex to be of any practical use. In any event, practitioners still desire a method for calculating or estimating the power of the Kruskal-Wallis test. In this section, we will present a brief survey pertaining to past power discussions of the Kruskal-Wallis test.

Power discussions of statistical hypotheses tests often include the notion of the asymptotic relative efficiency (ARE). The ARE of test A relative to test B, de-

noted *are*(A, B), is defined as the limiting value of the ratio N_B/N_A where N_A and N_B are the sample sizes necessary for the two tests to conform to the same testing conditions. Test B is considered more efficient than test A if *are*(A, B) < 1. That is, test B requires fewer subjects than test A to obtain the same results.

KRUSKAL (1952) suggested that the asymptotic relative efficiency of the Kruskal-Wallis test relative to its parametric counterpart, the F-test, should be investigated under normal theory. ANDREWS (1954) has calculated the ARE's of the Kruskal-Wallis (KW) test and Wilcoxon (W) test to that of their parametric competitors, which are the F-test (F) and t-test (t), respectively. He found that when the underlying distributions are normal, *are*(KW, F) ≈ 0.955 and *are*(W, t) ≈ 0.955, and when sampling from the Uniform (0, 1) distribution, *are*(KW, F) $= 1$. Another competitor of the Kruskal-Wallis test is the Median (M) test (DANIEL, 1990). ANDREWS (1954) determined that *are*(M, KW) $= 1/3$ when the underlying distributions are normally distributed. HODGES and LEHMANN (1956) obtained a lower limit to *are*(W, t) and *are*(KW, F) as 0.864, provided that the underlying distributions are symmetric and continuous. They also concluded that *are*(W, t) is 1, when the underlying distribution is Uniform (0, 1). Furthermore, HODGES and LEHMANN (1956) calculated *are*(W, t) $=$ *are*(KW, F) $= 81/64$, when the underlying distribution is Gamma($\alpha = 3$, $\beta = 1$). GUENTHER (1982) reported that *are*(KW, F) ≈ 1.096 when the underlying distribution is logistic. Formulas for *are*(KW, F), as well as asymptotic properties under a random effects model, are presented by SHIRAHATA (1985). Shirahata concluded that these formulas correspond directly with the formulas for a fixed effects model. BLAIR and HIGGINS (1985) comment that ARE's, although insightful, often are useless from a practitioners perspective, since ARE's are large sample results. In practice, we are generally concerned with small samples due to cost-effectiveness and availability of subjects. In our study, we are primarily interested in small sample properties.

BLAIR and HIGGINS (1985) also have compared the small sample and large sample powers of the two-sample Wilcoxon test and the two-sample t-test under various distributions and location configurations. They concluded that the Wilcoxon test is generally more powerful than the t-test, with rare and isolated exceptions. Also, power advantages of the t-test over the Wilcoxon test were small, even under conditions most favorable to the t-test, which are normal theory conditions with equal population variances. In cases where the Wilcoxon test attained higher power than the t-test, the differences were vast for nonnormal distributions. LEAVERTON and BIRCH (1969) generated small sample power curves under normal theory with equal population variances and small sample sizes. They determined that the Wilcoxon test compared remarkably well to the t-test, in all cases. DIXON (1954) also concluded that the Wilcoxon test has high power efficiency relative to the t-test, being never less than the large sample limiting value of 0.955, under conditions favorable to the t-test. The above mentioned authors question the widespread use of the t-test, when the equality of population variances is not valid. It should be noted, however, that the Mann-Whitney-Wilcoxon test also requires the

assumption of equal variances. FLIGNER and POLICELLO (1981) proposed a modification to the Mann-Whitney-Wilcoxon test when variances cannot be assumed to be equal. The effect of nonequal population variances on the Mann-Whitney-Wilcoxon test has been investigated by PRATT (1964) and VAN DER VAART (1961). These authors found that the significance level of this test is not preserved for differences in variance.

CONOVER, WEHNANEN, and RAMSEY (1978) suggest using the Wilcoxon test when a nonparametric test is applicable, except when the underlying distribution is double exponential. If a practitioner has reason to believe that the underlying cdf's belong to a particular family of distributions, LEHMANN (1975) and MILTON (1970) provide tables, formulas, and approximations for the exact power of the Wilcoxon test. CHANDA (1963) calculated the power efficiency of the Mann-Whitney, or Wilcoxon, test for a class of discrete distributions of the exponential type. He calculated the power efficiencies of the Mann-Whitney test to be at least 0.955, 1, and at least 0.75 for the Poisson, Binomial, and Geometric distributions, respectively. Although the above mentioned authors gave small sample power discussions and comparisons of the Wilcoxon test under a variety of known distributions, they did not provide a method for calculating or estimating the power of the Wilcoxon test in the unknown case.

COLLINGS and HAMILTON (1988) presented a method of estimating the power of the Wilcoxon test. Their method, called the Average X & Y method, is calculated using the empirical cdf's of the sample data as estimates of the unknown underlying cdf's. The Average X & Y power estimate is produced using a bootstrapping technique presented by EFRON (1982). BERAN (1986), and BICKEL and FREEDMAN (1981), discuss the consistency of bootstrapping estimators in the one and two sample cases. HINKLEY (1988), and DICICCIO and ROMANO (1988), provide surveys of recent developments in bootstrapping methodology for calculating point estimates and interval estimates of unknown parameters.

COLLINGS and HAMILTON (1988) investigated the reliability of their estimator in the small sample case, and concluded that their estimator was superior to other estimators investigated under the same conditions. The Average X & Y estimator performed well even for equal sample sizes as small as $m = 5$, producing negligible negative mediam bias particularly when the simulated true power was larger than 0.6. They recommended using the Average X & Y estimator for small sample sizes even when the underlying distribution is known to be normal. Collings and Hamilton suggested that their estimator has a natural extension to the Kruskal-Wallis test, and could be used to estimate the power of this test.

IBRAHIM (1991) presented a method for estimating the power of the Mann-Whitney test, which also is referred to as the Wilcoxon test. Ibrahim concluded that the method that he investigated was very reliable, and may be preferred over traditional parametric methods. Ibrahim's method was similar to the Average X & Y method presented by COLLINGS and HAMILTON (1988) in the sense that it utilized the same bootstrapping methodology and is a data-based point estimate of the

power. Ibrahim's method calculated the empirical cdf's producing an estimate of the power by resampling from these empirical cdf's simultaneously. In contrast, the Average X & Y method assumes that we do not know which sample most accurately represents the unknown underlying cdf under the null hypothesis. Therefore the Average X & Y method treats **each** empirical cdf as if it were the unknown underlying cdf, requiring a power estimate from each of the empirical cdf's to be averaged in the overall power estimate of the Wilcoxon test.

We are inclined to believe statisticians have attempted to estimate the power of the Kruskal-Wallis test; however, no method has formally been presented in the unknown case at this time, other than for the Wilcoxon test. In the known case, ANDREWS (1954) provided asymptotic power approximation formulas that ultimately required non-central chi-square tables, such as those given by FIX (1949), to find the approximate power of the Kruskal-Wallis test. In the unknown case, one possible approximation of a lower bound for the power of the Kruskal-Wallis test, at times, would be the power of the *F*-test under the same testing conditions. LACHENBRUCH and CLEMENTS (1991) have determined that the Kruskal-Wallis test generally has greater power, when population distributions are not normal, than its parametric counterpart, the *F*-test, which requires the stronger assumption of normality. Furthermore, they argued that the Kruskal-Wallis test is more robust than the *F*-test, being less sensitive to departures from the equality of variance assumption. Therefore, at times, it does seem reasonable to use the power of the *F*-test to approximate the power of the Kruskal-Wallis test. Formulas and examples of power calculations for the *F*-test have been revisited by KOELE (1982) and NICHOLSON (1954). TIKU (1967) provides tables for the power of the *F*-test under the fixed effects model under various levels of significance. This type of approximation will require more time, calculations, and may provide a poor lower bound for distributions, that are not similar in shape to the normal distribution. Therefore, at the present time, an extension of the Average X & Y power estimator is the only reasonable estimator that exists, and warrants investigation with respect to its reliability.

2. An Example of the Kruskal-Wallis Test Procedure

Often, a patient is required to have an x-ray dye injected into their veins, in order to obtain an adequate x-ray on a particular area of his or her anatomy. In the medical field it is known that accidental injection of x-ray dye into subcutaneous tissues causes serious inflammation, and may cause severe irreversible soft tissue damage. A radiologist is interested in studying the soft tissue damage resulting from intravenous injection of ionic and nonionic x-ray dyes, and a saline solution. The saline solution is included in the study as a control. The radiologist hypothesizes that the soft tissue damage resulting from the use of the nonionic x-ray dye is less than that of the soft tissue damage resulting from the ionic x-ray dye. Hence, the null and alternative hypotheses that the radiologist is interested in testing are, respectively:

H_0: The median soft tissue damage for the three solutions are identical;

H_a: At least one solution differs with respect to median soft tissue damage.

To test the radiologist's hypotheses, subcutaneous injections of 0.3 cc of each of the three solutions were injected into the thighs of a total of eighteen white mice. Both posterior thighs of each of six mice were injected with one of the three solutions. Therefore, we have twelve injected thighs for each of the three solutions. The mice were sacrificed and the injected sites were removed and examined at one week. The degree of subcutaneous inflammation was measured. The measurements ranged from 0 to 3, where a measurement of 0 corresponds to no inflammation being present, and a measurement of 3 corresponds to the presence of the most severe inflammation. The data and their corresponding combined sample rankings, as required for the calculation of the Kruskal-Wallis test statistic, appear in Table 1. It should be noted that this is not actual data.

Table 1

Subcutaneous inflammation measurements on 36 white mouse thighs, taken one week after injection with one of three solutions. This is not actual data

Thigh	Hypaque (Ionic)	Omnipaque (Nonionic)	NaCl (Saline)
1	.75 (17)*	.43 (9)	.02 (3)
2	.79 (19.5)	.39 (6)	.85 (25.5)
3	1.57 (34)	.46 (11)	0 (1)
4	.40 (7)	1.75 (36)	.80 (21)
5	.82 (23)	1.02 (30)	.79 (19.5)
6	.67 (15)	1.18 (32)	.81 (22)
7	.55 (13.5)	.55 (13.5)	.01 (2)
8	1.70 (35)	.49 (12)	.93 (29)
9	.77 (18)	1.25 (33)	.89 (27)
10	.90 (28)	1.10 (31)	.06 (5)
11	.85 (25.5)	.42 (8)	.05 (4)
12	.68 (16)	.45 (10)	.84 (24)
R_i	$R_1 = 251.5$	$R_2 = 231.5$	$R_3 = 183$

* Values in parentheses represent the combined sample rankings.

In this example, the total sample size is $N = 36$, where we have $k = 3$ samples of equal sizes, $n_1 = n_2 = n_3 = m = 12$. The total of the combined sample rankings for the ith sample is denoted by R_i. From the information contained in Table 1, we have that $R_1 = 251.5$, $R_2 = 231.5$, and $R_3 = 183$. The Kruskal-Wallis test statistic, H, is calculated as 1.8637. At $\alpha = 0.05$ level of significance, the critical value obtained from the chi-square table with 2 degrees of freedom is 5.99. In this example, we are not rejecting H_0 in favour of H_a. That is, at $\alpha = 0.05$ level of significance, there is insufficient evidence to conclude that at least one of the solutions differs with respect to median soft tissue damage.

Since H_0 has not been rejected, a natural question to consider is whether or not the radiologist can conclude that the three solutions are the same with respect to the median soft tissue damage produced. If the test applied has high power at a given alternative, this is a safe conclusion. Suppose that the radiologist views a difference of 0.5 tissue damage between the nonionic x-ray dye and the saline solution, and also a difference of 0.5 between the ionic and nonionic x-ray dyes, to be clinically significant. Therefore, the radiologist would like to know the probability that the Kruskal-Wallis test will accurately detect the aforementioned differences, given the sample data. In statistical terminology, the radiologist would like to know the power of the Kruskal-Wallis test for the location shift vector $\underline{\delta} = (0, 0.5, 1.0)'$, for the saline solution, nonionic x-ray dye, and ionic x-ray dye, respectively. Without prior knowledge of the underlying distributions, his question cannot be adequately answered. We will revisit this problem in Section 5.

3. The Monte Carlo Simulation Study

As discussed in Section 1, we need to produce a data-based estimate of each F_i, the underlying unknown population distribution, from which the k samples have been drawn. In practice, a researcher may use many different procedures based on pilot studies, or actual collected data. The following method for calculating empirical cdf's was found most efficient by COLLINGS and HAMILTON (1988) for calculating empirical cdf's. We utilized the same method in our simulation study.

Let us consider only one empirical cdf calculation. Suppose there are L distinct data points in a random sample, from an unknown probability distribution, say F. Without loss of generality, let us denote these data points by $Z_1, Z_2, \ldots, Z_L$. Order the data points, from smallest to largest, and denote these ordered data points by $Z_{(1)}, Z_{(2)}, \ldots, Z_{(L)}$. These ordered data points, commonly are referred to as the order statistics associated with the values $Z_1, Z_2, \ldots, Z_L$. Define $Z_{(0)} = 2Z_{(1)} - Z_{(2)}$, and $Z_{(L+1)} = 2Z_{(L)} - Z_{(L-1)}$. Let G be the continuous empirical cdf obtained by assigning the probability $1/(L+1)$, uniformly to each interval $(Z_{(i)}, Z_{(i+1)})$ for $i = 0, 1, \ldots, L$. Power estimation is performed with the empirical cdf, G, as an estimate of the unknown population distribution, F.

3.1 *Power Estimation Using the Extended Average X & Y Method*

The Average X & Y power estimator for two samples presented by COLLINGS and HAMILTON (1988) is a weighted average of the power estimates $\hat{\pi}(G_i, \underline{\delta}, \alpha)$, of the empirical cdf's, G_i, for location shift vector $\underline{\delta}$, at significance level α, for $i = 1, 2$. The Average X & Y power estimator is calculated by the following equation:

$$\hat{\pi} = \frac{n_1\hat{\pi}(G_1, \underline{\delta}, \alpha) + n_2\hat{\pi}(G_2, \underline{\delta}, \alpha)}{n_1 + n_2}. \tag{3.1.1}$$

The natural extension of the Average X & Y power estimator, which we will refer to as the Extended Average X & Y Power estimator, is calculated by the following equation:

$$\hat{\pi} = \frac{n_1\hat{\pi}(G_1, \underline{\delta}, \alpha) + n_2\hat{\pi}(G_2, \underline{\delta}, \alpha) + \ldots + n_k\hat{\pi}(G_k, \underline{\delta}, \alpha)}{n_1 + n_2 + \ldots + n_k}. \tag{3.1.2}$$

As discussed by COLLINGS and HAMILTON (1988), $\hat{\pi}(G_i, \underline{\delta}, \alpha)$ may be obtained through computer bootstrapping simulations. We have incorporated the same bootstrapping procedure in our study. The procedure for equal sample sizes is as follows:

(1) Calculate the k data-based empirical cdf's using the technique presented in Section 3. Denote these empirical cdf's as G_i, for $i = 1, 2, \ldots, k$.

(2) Use IMSL subroutines, to draw a random sample of size km from G_i, the ith empirical cdf. Denote the values contained in this random sample, by $Y_1, Y_2, \ldots, Y_{km}$. It is crucial, that we notice that the km observations are from the same empirical cdf.

(3) To produce the k simulated random samples, from different population distributions, we need to add the components of a $\underline{\delta}$ location shift vector to the $Y_1, Y_2, \ldots, Y_{km}$ values. That is, the ith component of $\underline{\delta}$, is added to the ith set of m, Y values, for $i = 1, 2, \ldots, k$. Without loss of generality, denote these shifted values as X_{ij}, the jth observation from the ith sample, for $i = 1, 2, \ldots, k$ and $j = 1, \ldots, m$. In essence, the Y values have been divided into the k random samples of size m. Hence, we have k random samples of size m, from populations that differ only with respect to location.

(4) Perform the Kruskal-Wallis test on the k simulated random samples, namely the X_{ij} values, and obtain a value of H. If the value of H is greater than or equal to the critical value that is obtained from the chi-square distribution with $k - 1$ degrees of freedom, a success is counted.

(5) Repeat steps 2, 3, and 4, for $J = 1000$ times. The power, $\pi(G_i, \underline{\delta}, \alpha)$, for the ith empirical cdf is estimated by calculating the proportion of successes among the $J = 1000$ repetitions. Thus, the estimated power is given by:

$$\hat{\pi}(G_i, \underline{\delta}, \alpha) = \frac{\text{number of rejections of } H_0}{J}. \tag{3.1.3}$$

Let us reiterate two significant facts at this moment. First, this procedure will produce $\hat{\pi}(G_i, \underline{\delta}, \alpha)$ for **one** empirical distribution. In other words, we need to perform the procedure on **each** of the k empirical cdf's. Second, in our study, we have considered α values of 0.01, 0.05, and 0.10. Therefore, in (4), we have three success counts to record. Again, we are primarily interested in $\alpha = 0.05$.

3.2 *Simulation of the True Power*

We needed to simulate the true power, $\pi(F, \underline{\delta}, \alpha)$ of the Kruskal-Wallis test under specified distributions, k, m, and $\underline{\delta}$, so as to have a yardstick for comparisons with

the Extended Average X & Y power estimates. We simulated 10000 experiments for each combination of distribution, k, m, and $\underline{\delta}$, using the same method discussed in COLLINGS and HAMILTON (1988). The computer draws a random sample of size km, from a specified distribution. Denote the values contained in this random sample by $Y_1, Y_2, \ldots, Y_k$. It is crucial, at this stable, that we notice that all of the observations drawn have originated from the **same** population distribution. In the same manner as in (3) of the previous section, each component of $\underline{\delta}$, is then added to the ith set of m, Y values, to produce the desired observations for the random samples that are used in the simulation. Let us denote these shifted Y values, as X_{ij}, where $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, m$. The k random samples, that are based on the X_{ij}'s, have distributions that differ only with respect to location. The Kruskal-Wallis test is then conducted, and the test statistic, H, is calculated. This procedure is repeated 10000 times, and the number of times H_0 is rejected based on a chi-square critical value with $k-1$ degrees of freedom, for each of the three α levels, is counted. The simulated true power for location shift vector $\underline{\delta}$, at significance level α, is calculated as follows:

$$\pi(F, \underline{\delta}, \alpha) = \frac{\text{number of rejections of } H_0}{10000}. \tag{3.2.1}$$

Notice that when $\underline{\delta} = \underline{0}$, the simulated true power should approximately equal the α level of significance. In the least favorable case, which corresponds to the simulated true power equal to 0.5, we are 95% confident that we are within 0.01 of the true power.

3.3 *The Design of the Study*

Our study examined the performance of the Extended Average X & Y power estimator with respect to the estimation of the Kruskal-Wallis test for four underlying continuous distributions: standard normal; Cauchy; uniform; and exponential. We have considered $k = 3$ and $k = 5$ samples for our simulation study. Equal samples sizes, say m, were drawn from each sample population. That is, samples sizes of the k populations are equal to $n_1 = n_2 = \ldots = n_k = m$. The values that we have chosen for our sample sizes, m, are 10 and 20. Recall, that the Extended Average X & Y power estimator, for a given δ location shift vector, is calculated by the following equation:

$$\hat{\pi} = \frac{n_1\hat{\pi}(G_1, \underline{\delta}, \alpha) + n_2\hat{\pi}(G_2, \underline{\delta}, \alpha) + \ldots + n_k\hat{\pi}(G_k, \underline{\delta}, \alpha)}{n_1 + n_2 + \ldots + n_k}. \tag{3.3.1}$$

Under the conditions of our study, the above equation becomes:

$$\hat{\pi} = \frac{\hat{\pi}(G_1, \underline{\delta}, \alpha) + \hat{\pi}(G_2, \underline{\delta}, \alpha) + \ldots + \hat{\pi}(G_k, \underline{\delta}, \alpha)}{k}. \tag{3.3.2}$$

where, $\hat{\pi}(G_i, \underline{\delta}, \alpha)$ is the estimated power at a given $\underline{\delta} = (\delta_1, \delta_2, \ldots, \delta_k)$ and α, obtained from a data-based empirical cdf, $G_I(X)$, for $i = 1, 2, \ldots, k$.

Eight different location shift configurations were considered for each of the 16 combintions of the population distribution, k, and m. That is, we had eight different $\underline{\delta}$ vectors of location shifts, denoted by $\underline{\delta}_t$, for $t = l, \ldots, 8$. The location shift $\underline{\delta}_t$ vectors, were defined as follows:

$K = 3$:

($\underline{\delta}_1$) Median 1 = median 2 < median 3;
($\underline{\delta}_2$) Median 1 < median 2 < median 3, with equal spacing between the medians;
($\underline{\delta}_3$) Median 1 < median 2 < median 3, with distances of 1/3 and 2/3 between medians 1 and 2, and medians 2 and 3, respectively;
($\underline{\delta}_4$) Median 1 < median 2 < median 3, with distances of 1/6 and 5/6 between medians 1 and 2, and medians 2 and 3, respectively.

$K = 5$:

($\underline{\delta}_5$) Median 1 = median 2 = median 3 = median 4 < median 5;
($\underline{\delta}_6$) Median 1 < median 2 < ... < median 5, with equal spacing between the medians;
($\underline{\delta}_7$) Median 1 = median 2 = median 3 < median 4 < median 5, with the distance between medians 3 and 4 half as much as the distance between medians 4 and 5.
($\underline{\delta}_8$) Median 1 < median 2 < ... < median 5, with the distances between medians 2, 3, 4 and 5, as 1/10, 2/10, 3/10, and 4/10; respectively.

Extended Average X & Y power estimates have been calculated, on eleven simulated data sets, for each combination of distribution type, k, m, and the above mentioned location shift vectors, at eleven equally spaced alternatives. That is, at eleven equally spaced $\underline{\delta}_t$, shift vectors depending on whether $k = 3$ or $k = 5$, and $t = 1, 2, \ldots, 8$. True powers have been simulated under the same conditions. The simulated true powers range from 0.05 to 0.95, for $\alpha = 0.05$, approximately. Our primary interest is in $\alpha = 0.05$; however, we have included $\alpha = 0.01$, and $\alpha = 0.10$, in our study. Comparison between the simulated true powers and the Extended Average X & Y power estimates, have been obtained. The results and a discussion of these comparisons, are the subject of Section 4.

The two basic simulation programs that were used in this study were written in FORTRAN77. The two basic programs have incorporated the following IMSL subroutines: RNGCT, RNGCS, RNNOF, RNCHY, RNEXP, RNUN, and RANKS. The programs were written such that only certain parameters, subroutines, and location vectors needed to be changed. Therefore, we were able the use the same two basic programs repeatedly in our simulations, with only minor extensions required as we moved from $k = 3$ to $k = 5$ samples.

4. Results and Discussion

We have utilized a method of analysis similar to one of the methods employed by COLLINGS and HAMILTON (1988). Collings and Hamilton created median curves and 90% envelope curves, at a given alternative, based on the differences between the Average X & Y power estimate at the corresponding 5th, 50th and 95th percentiles of 100 simulated data sets, and the simulated "true" powers. The median curve should reflect the amount of median bias present in the power estimates; whereas, the envelope curves should behave similar to confidence intervals. Roughly, the envelope curves should capture a certain percentage of the estimates. We have created median curves and 80% envelope curves, at a given alternative, for our eleven simulated data sets. Our curves, are based on the differences between the Extended Average X & Y power estimates at the corresponding 10th, 50th and 90th percentiles of eleven simulated data sets, and the simulated "true" powers.

The values that we are referring to as the 10th and 90th percentiles in our study, are actually the second smallest and second largest Extended Average X & Y power estimates for eleven simulated data sets. We will refer to them as percentiles, with this distinction having been made transparent to the reader. The 50th percentile, in our study, is the actual median of the Extended Average X & Y power estimates for eleven simulated data sets, at a given alternative.

In the following paragraphs, we have provided a discussion of the results of our simulation study. Graphs of a few of our median curves and 80% envelope curves, are included for illustrative purposes. In these graphs, the horizontal axis represents the simulated "true" powers, and the vertical axis represents the differences between the Extended Average X & Y power estimates and the "true" powers at the 10th, 50th, and 90th percentiles. Let us denote these differences, $\hat{\pi} - \pi$. An indication that $\hat{\pi}$ is close to π, is an envelope curve, that is tightly centered about the horizontal line $\hat{\pi} - \pi = 0$. Our median curves and envelope curves, also should convey whether or not the extended Average X & Y estimator is overestimating, or underestimating, the true power.

When there were $k = 3$ samples and $\alpha = 0.05$, the Extended Average X & Y estimator performed reasonably well for all the distributions considered. For samples of size 20, the width of the envelope curves tended to be at most 0.3. The 50th percentile was usually within 0.10 of the "true" power for the Cauchy, normal, and exponential distributions with sample sizes of 20. For the uniform distribution, the estimator tended to mildly underestimate when the "true" power was less than 0.5 and mildly overestimate when the "true" power was greater then 0.5. The amount of discrepancy between the estimate and the "true" power in this case was generally between 0 and 0.2. Table 2 and Figure 1 illustrate the findings in one case for the Cauchy distribution. Table 3 and Figure 2 illustrate the findings in one case for the uniform distribution.

When there were $k = 5$ samples and $\alpha = 0.05$, the power estimates tended more towards underestimating the "true" power. This was seen less in samples of

size 20 than in samples of size 10. For samples of size 20, with a "true" power greater than .5, the 50th percentile of the estimates was usually within 0.1 of the "true" power for all distributions considered.

Table 2

Simulation study results for $k = 3$ amples of size $m = 20$, with location shift vector $\underline{\delta}_4$, when the underlying distribution is Cauchy (0, 1) and $\alpha = 0.05$

Population Distribution: Cauchy (0, 1)	δ_4	$k = 3$	$m = 20$	$\alpha = 0.05$
$\underline{\delta} = (\delta_1, \delta_2, \delta_3)'$ $\delta = (0.000, 0.050, 0.300)'$	"True" Power	10-th Percentile	50-th Percentile	90-th Percentile
0.000 0.000 0.000	0.0443	0.0000	0.0000	0.0000
		(−0.0443)*	(−0.0443)	(−0.0443)
0.000 0.050 0.300	0.0695	0.0000	0.0003	0.0103
		(−0.0695)	(−0.0692)	(−0.0592)
0.000 0.100 0.600	0.1457	0.0033	0.0330	0.2323
		(−0.1424)	(−0.1127)	(0.0866)
0.000 0.150 0.900	0.2710	0.0583	0.1463	0.3230
		(−0.2127)	(−0.1247)	(0.0520)
0.000 0.200 1.200	0.4222	0.2043	0.3953	0.4687
		(−0.2179)	(−0.0269)	(0.0465)
0.000 0.250 1.500	0.5881	0.3647	0.5080	0.6373
		(−0.2234)	(−0.0801)	(0.0492)
0.000 0.300 1.800	0.7186	0.5163	0.7013	0.7787
		(−0.2023)	(−0.0173)	(0.0601)
0.000 0.350 2.100	0.8156	0.6150	0.8217	0.8680
		(−0.2006)	(0.0061)	(0.0524)
0.000 0.400 2.400	0.8855	0.7130	0.8770	0.9173
		(−0.1725)	(−0.0085)	(0.0318)
0.000 0.450 2.700	0.9320	0.7910	0.9030	0.9603
		(−0.1410)	(−0.0290)	(0.0283)
0.000 0.500 3.000	0.9569	0.8510	0.9313	0.9677
		(−0.1059)	(−0.0256)	(0.0108)

* Values in parentheses represent the difference between the percentile and the "true" power at a given shift.

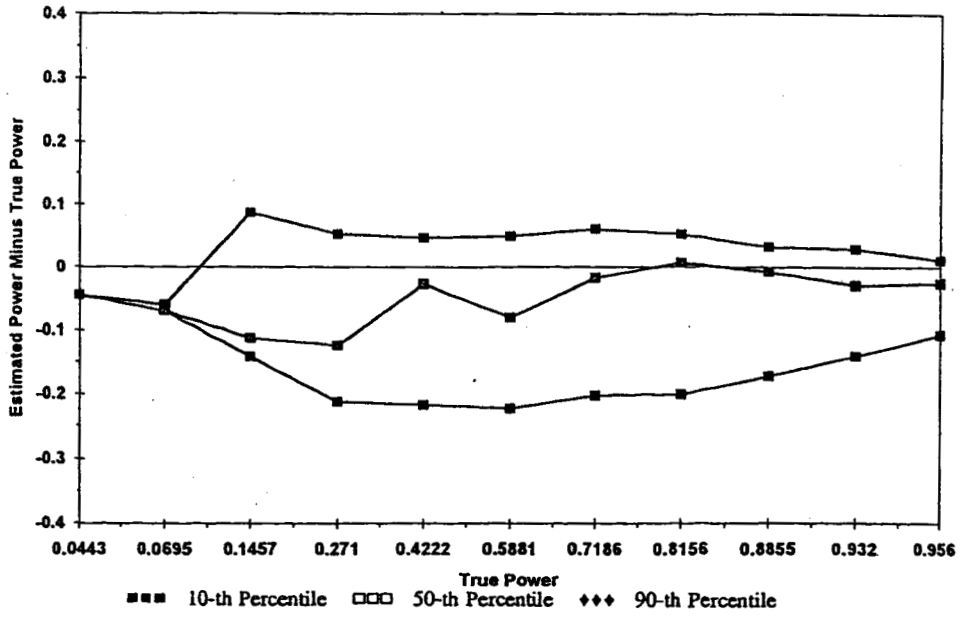

Fig. 1: Illustration of Table 2. The median curve and 80% envelope curves for the differences between the Extended Average X & Y estimates and the simulated "true" power, based on eleven simulated data sets

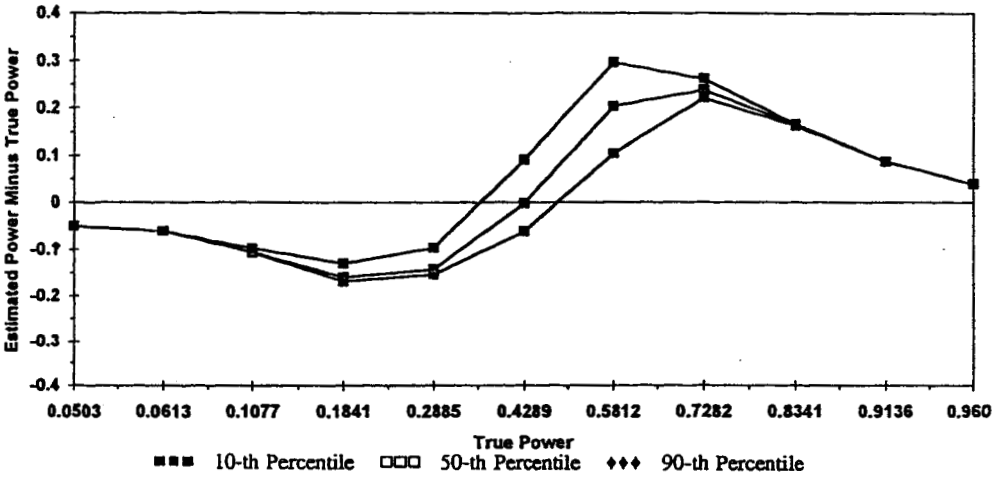

Fig. 2: Illustration of Table 3. The median curve and 80% envelope curves for the differences between the Extended Average X & Y estimates and the simulated "true" power, based on eleven simulated data sets

Table 3

Simulation study results for $k = 3$ samples of size $m = 20$, with location shift vector $\underline{\delta}_4$, when the underlying distribution is Uniform (0, 1) and $\alpha = 0.05$

Population Distribution: Uniform (0, 1) δ_4		$k = 3$	$m = 20$	$\alpha = 0.05$
$\underline{\delta} = (\delta_1, \delta_2, \delta_3)'$ $\delta = (0.000, 0.006, 0.038)'$	"True" Power	10-th Percentile	50-th Percentile	90-th Percentile
0.000 0.000 0.000	0.0503	0.0000	0.0000	0.0000
		(−0.0503)*	(−0.0503)	(−0.0503)
0.000 0.006 0.038	0.0613	0.0000	0.0000	0.0003
		(−0.0613)	(−0.0613)	(−0.0610)
0.000 0.013 0.077	0.1077	0.0010	0.0020	0.0103
		(−0.1067)	(−0.1057)	(−0.0974)
0.000 0.0119 0.115	0.1841	0.0157	0.0250	0.0540
		(−0.1684)	(−0.1591)	(−0.1301)
0.000 0.026 0.154	0.2885	0.1350	0.1463	0.1917
		(−0.1535)	(−0.1422)	(−0.0968)
0.000 0.032 0.192	0.4289	0.3667	0.4273	0.5197
		(−0.0622)	(−0.0016)	(0.0908)
0.000 0.038 0.230	0.5812	0.6857	0.7847	0.8777
		(0.1045)	(0.2035)	(0.2965)
0.000 0.045 0.269	0.7282	0.9493	0.9667	0.9907
		(0.2211)	(0.2385)	(0.2625)
0.000 0.051 0.307	0.8341	0.9970	0.9997	1.0000
		(0.1629)	(0.1656)	(0.1659)
0.000 0.058 0.346	0.9136	1.0000	1.0000	1.0000
		(0.0864)	(0.0864)	(0.0864)
0.000 0.064 0.384	0.9606	1.0000	1.0000	1.0000
		(0.0394)	(0.0394)	(0.0394)

* Values in parentheses represent the differences between the percentile and the "true" power at a given shift.

5. General Conclusions

As discussed in Section 1, the goal of this research was to estimate the power of the Kruskal-Wallis test, using an Extended Average X & Y estimator as suggested by COLLINGS and HAMILTON (1988). Through the use of computer simulations and bootstrapping methods, we have been able to observe the reliability of the Extended Average X & Y estimator, under various underlying distributions and location conditions. Our study focused on $k = 3$ and $k = 5$ samples, of equal sizes $m = 10$ and $m = 20$. Our simulation study demonstrates that the Extended Average X & Y power estimator is a reliable estimator of the true power of the Kruskal-Wallis test for $k = 3$ samples, and a more conservative to a mild overestimator of the true power for $k = 5$ samples.

In general, for $k = 3$ samples, the Extended Average X & Y estimator tended to have negligible to moderate negative median bias, with approximately equal amounts of discrepancies from the "true" power. Samples of size $m = 10$ produced the widest envelope curves when the underlying distribution was Cauchy; however, the envelope curves were well centered, with small amounts of median bias. The uniform distribution produced primarily moderate underestimates to moderate overestimates, for more distant alternatives, when $\alpha = 0.05$. These estimates were less varying for distant alternatives and when α was increased to 0.10. The Extended Average X & Y estimator tended to underestimate the "true" power, in all cases, when the level of significance was equal to 0.01. The Extended Average X & Y estimator performed the most favorable, for the normal distribution, for both samples investigated.

In general, for $k = 5$ samples, the Extended Average X & Y estimator primarily underestimated the true power when the significance level was equal to 0.05, producing very mild amounts of overestimation for more distant alternatives. The estimates were less varying for samples of size $m = 20$. In particular, when the underlying distribution was uniform, the envelope curves possessed almost zero width when the "true" power was greater than 0.6; however, the 90th percentiles often depicted small amounts of overestimation. The Extended Average X & Y estimator tended to underestimate the "true" power, in all cases, when the level of significance was equal to 0.01. The Extended Average X & Y power estimates often slightly improved, when the level of significance was increased from 0.05 to 0.10.

Based of the results of our simulation study, and the work of COLLINGS and HAMILTON (1988) that was discussed in Section 1, we feel that the Extended Average X & Y power estimator is a reliable estimator of the true power of the Kruskal-Wallis test for $k = 3$ samples, and a more conservative to a mild overestimator of the true power for $k = 5$ samples. It is our opinion, that the Extended Average X & Y estimator can be applied successfully to estimate the power for $k = 3$ samples of equal size greater than $m = 10$, and $\alpha = 0.05$ level of significance. In this case, we are considering any discrepancies between the estimator and the true power to fall between -0.2 and 0.2. It also is our opinion, that the Extended Average X & Y estimator can be applied successfully to estimate the power for $k = 5$ samples of equal size greater than

$m = 10$ with $\alpha = 0.05$ to $\alpha = 0.10$ levels of significance. In this case, we are considering any discrepancies between the estimator and the true power to fall between -0.25 and 0.1 for $\alpha = 0.05$ level of significance, in particular.

In reference to the soft tissue inflammation example, given in Section 2, we calculated a power estimate for the location configuration $\underline{\delta} = (0, 0.5, 1.0)'$. The Extended Average X & Y power estimate at $\alpha = 0.05$ level of significance produced was $\hat{\pi} = 1.0000$. Based of the results of our simulation study and given that our specified location configuration could be considered a distant alternative, we would say that $\hat{\pi} = 1.0000$ is a mild overestimate of the true power in this example; however, the amount of overestimation could be as roughly as high as 0.1. Therefore, we can be quite confident that our test is accurately detecting a difference between the amounts of soft tissue damage of the three solutions, for the sample data. It should be noted that had we not rejected the null hypothesis, our power estimate of at least 0.90 to 1.0000 would indicate that the null hypothesis would have a good chance of being true.

We have the impression that the Extended Average X & Y estimator will tend to produce entirely underestimates of the true power, as k increases, at 0.05 level of significance. Therefore, further simulation research studies on larger numbers of samples, k, may be of interest. It should be noted, that k greater than five samples may be impractical from a practitioners point of view..

We have the impression that the manner in which the empirical cdf's are calculated and resampled, may have an impact on the Extended Average X & Y power estimates produced. Therefore, further simulation studies changing the calculation of the empirical cdf may be of interest. Bootstrap samples may be drawn from either a smoothed empirical cdf or by sampling, with replacement, from the observed data. The Extended Average X & Y method uses a smoothed empirical cdf technique for bootstrapping; however, the calculation of the empirical cdf as suggested COLLINGS and HAMILTON (1988) is quite different from the classical definition of an empirical cdf. The classical method of calculating empirical cdf's is presented by BAIN and ENGLEHARDT (1992). This method defines the range of the empirical cdf as the range of the order statistics contained in the sample, and assigns probabilities uniformly over the intervals that are defined by the order statistics. COLLINGS and HAMILTON (1988) utilized an empirical cdf method in which they redefined the lower and upper limits of the range of the classical empirical cdf; assigning probabilities uniformly over the intervals determined by the lower limit, the order statistics, and the upper limit. We assume that this redefinition is taking into consideration the fact that the smallest and largest order statistics may not represent the actual lowest and largest values that may be obtained from the range of the unknown cdf. Therefore, we are inclined to believe that the redefinition of endpoints, as well as the smoothing of the empirical cdf's, may have an effect on the estimates produced. It should be noted that in comparison, IBRAHIM (1991) resampled with replacement from the observed data points to obtain the bootstrap samples. Our work is based on the smoothed empirical cdf technique described by COLLINGS and HAMILTON (1988).

As a third and final suggestion for future research, sample size determination using the Extended Average X & Y estimator is the next logical step. GUENTHER (1982) concluded that asymptotic power formulas, given by ANDREWS (1954), produced poor sample size estimates. Guenther derived a method for estimating the sample size of the Kruskal-Wallis test, based on normal theory and asymptotic relative efficiencies with reject to the F-test; however, not all distributions provided reliable results. HAMILTON and COLLINGS (1991) have adapted their bootstrap procedure to use their Average X & Y power estimator in determining the adequate sample sizes necessary to attain a fixed power at a given shift alternative, for the Wilcoxon test. They compared their method to a traditional Assume Normal, Pool Variance (ANPV) method and a method developed specifically for the Wilcoxon test by NOETHER (1987). Their simulation study demonstrated that their method was at least as good as Noether's method, and that the ANPV method was inferior to both methods. They preferred their method over Noether's, since their method easily can be extended to other statistical tests, such as the Kruskal-Wallis test. Therefore, we feel that sample size determination for the Kruskal-Wallis test using the Extended Average X & Y power estimator is logical.

References

ANDREWS, F. C., 1954: Asymptotic Behavior of Some Rank Tests for Analysis of Variance. *Annals of Mathematical Statistics* **25**, 724–736.

BAIN, L. J. and ENGLEHARDT, M., 1992: *Introduction to Probability and Mathematical Statistics, Second Edition.* PWS-KENT, Boston, Mass.

BERAN, R., 1986: Simulated Power Functions. *Annals of Statistics* **14**, 151–173.

BICKEL, P. J. and FREEDMAN, D. A., 1981: Some Asymptotic Theory for the Bootstrap. *Annals of Statistics* **9**, 1196–1217.

BLAIR, R. C. and HIGGINS, J. J., 1985: Comparison of the Power of Paired Samples *t*-Test to That of Wilcoxon's Signed-Ranks Test Under Various Population Shapes. *Psychological Bulletin* **97**, 119–128.

CHANDA, K. C., 1963: On the Efficiency of Two-Sample Mann-Whitney Test for Discrete Populations. *Annals of Mathematical Statistics* **34**, 612–617.

COLLINGS, B. J. and HAMILTON, M. A., 1988: Estimating the Power of the Two-Sample Wilcoxon Test for Location Shift. *Biometrics* **44**, 847–860.

CONOVER, W. J., WEHMANEN, O., and RAMSEY, F. L., 1978: A Note on Small-Sample Power Functions for Nonparametric Tests of Location in the Double Exponential Family. *Journal of the American Statistical Association* **73**, 188–190.

DANIEL, W. W., 1990: *Applied Nonparametric Statistics.* PWS-KENT, Boston, Mass.

DICICCIO, T. J. and ROMANO, J. P., 1988: A Review of Bootstrap Confidence Intervals. *Journal of the Royal Statistical Society (Series B)* **50**, 338–354.

DIXON, W. J., 1954: Power Under Normality of Several Nonparametric Tests. *Annals of Mathematical Statistics* **25**, 310–314.

EFRON, B., 1982: *The Jackknife, the Bootstrap, and Other Resampling Plans.* National Science Foundation-Conference Board of the Mathematical Sciences Monograph 38. Philadelphia: Society for Industrial and Applied Mathematics.

FIX, E., 1949: Tables of the Noncentral κ^2. *University of California Publications in Statistics* **1**, 15–19.

FLIGNER, M. and POLICELLO, G., 1981: Robust Rank Procedures for the Behrens-Fisher Problem. *Journal of the American Statistical Association* **76**, 162–168.

GUENTHER, W. C., 1982: Normal Theory Sample Size Formulas for Some Nonnormal Distributions. *Communications in Statistic: Simulation and Comutation* **11(6)**, 727–732.

HAMILTON, M. A. and COLLINGS, G. J., 1991: Determining the Appropriate Sample Size for Nonparametrics Tests for Location Shift. *Technometrics* **33**, 327–337.

HINKLEY, D. V., 1988: Bootstrap Methods. *Journal of the Royal Statistic Society (Series B)* **50**, 321–337.

HODGES, J. L. and LEHMANN., E. L., 1956: The Efficiency of Some Nonparametrics Competitors of the t-Test. *Annals of Mathematical Statistics* **27**, 324–335.

IBRAHIM, I. H., 1991: Evaluating the Power of the Mann-Whitney Test Using the Bootstrap Method. *Communications in Statistics: Theory and Methods* **20(9)**, 919–931.

KOELE, P., 1982: Calculating Power in Analysis of Variance. *Psychological Bulletin* **92**, 513–516.

KRUSKAL, W. H., 1952: A Nonparametric Test for the Several Sample Problem. *Annals of Mathematical Statistics* **23**, 525–540.

KRUSKAL, W. H. and WALLIS, W. A., 1952: Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association* **47**, 583–621.

LACHENBRUCH, P. A. and CLEMENTS, P. J., 1991: ANOVA, Kruskal-Wallis, Normal Scores and Unequal Variance. *Communications in Statistics: Theory and Methods* **20(1)**, 107–126.

LEHMANN, E. L., 1953: The Power of Rank Tests. *Annals of Mathematical Statistics* **24**, 23–43.

LEHMANN, E. L., 1975: *Nonparametrics: Statistical Methods Based on Ranks.* Holden Day, Inc., San Francisco, CA.

LEAVERTON, P. and BIRCH, J. J., 1969: Small Sample Power Curves for the Two Sample Location Problem. *Technometrics* **11**, 299–307.

MILTON, R. C., 1970: *Rank Order Probabilities: Two-Sample Normal Shift Alternatives.* Wiley, New York, NY.

NICHOLSON, W. L., 1954: A Computing Formula for the Power of the Analysis of Variance Test. *Annals of Mathematical Statistics* **25**, 607–610.

NOETHER, G. E., 1987: Sample Size Determination for Some Common Nonparametric Tests. *Journal of the American Statistical Association* **82**, 645–647.

PRATT, J. W., 1964: Robustness of Some Procedures for the Two-Sample Location Problem. *Journal of the American Statistical Association* **59**, 665–680.

SHIRAHATA, S., 1985: Asymptotic Properties of the Kruskal-Wallis Test and the Friedman Test in the Analysis of Variance Models with Random Effects. *Communications in Statistics: Theory and Methods* **14(7)**, 1685–1692.

TIKU, M. L., 1967: Tables of the Power of the F-Test. *Journal of the American Statistical Association* **62**, 525–539.

VAN DER VAART, H. R., 1961: On the Robustness of Wilcoxon's Two-Sample Test. *Quantitative Methods in Pharmacology*, ed. H. de Jonege. Interscience, New York, New York, 140–158.

Received, January 1995
Revised, July 1995
Revised, October 1995
Accepted, November 1995

MICHELLE MAHONEY
Medical Center
Rochester, Minnesota 55905
U.S.A.

RHONDA MAGEL
Department of Statistics
North Dakota State University
P.O. Box 5575, Waldron 201
Fargo, North Dakota 58105
U.S.A.