3
4-19-79

LA-7677-MS
Informal Report

On Multiple-Comparisons Procedures

University of California
LOS ALAMOS SCIENTIFIC LABORATORY
Post Office Box 1663 Los Alamos, New Mexico 87545

This work was supported by the U.S. Geological Survey.

LA-7677-MS
Informal Report
Special Distribution
Issued: February 1979

On Multiple-Comparisons Procedures

W. J. Conover*
Ronald L. Iman**

*Visiting Staff Member. College of Business Administration, Texas Tech University, P. O. Box 4319, Lubbock, TX 79409.
**Division 1223, Sandia Laboratories, P. O. Box 5800, Albuquerque, NM 87115.

ON MULTIPLE-COMPARISONS PROCEDURES

by

W. J. Conover
Ronald L. Iman

ABSTRACT

Some of the more popular multiple-comparisons procedures are discussed and compared. Some new nonparametric methods are introduced. One procedure is an analog to the Fisher's least significant difference method for the completely randomized design. Some simulation studies indicate this procedure is a reasonable nonparametric method to use. A summary description is given for other nonparametric methods, which may be used with the completely randomized or randomized blocks designs.

I. INTRODUCTION

When two treatments are being compared, the situation is fairly simple. Either the two treatments are considered equivalent or they are not. The traditional theory of statistical hypothesis testing corresponds nicely to the experimenter's objectives. The Type I error and Type II error, with corresponding α and β, are easy to interpret.

When more than two treatments are being compared simultaneously, the situation is no longer simple. The experimenter is usually not as interested in knowing whether any differences among the treatments exist as in knowing which treatments are different. The traditional theory of hypothesis testing no longer corresponds so nicely to the experimenter's objectives. The theorist is often overly concerned with protecting against Type I error, so the resulting procedures have relatively little power. On the other hand, repeated use of the two-sample procedures, while rich in power, tends to boost the experimentwise α level to unacceptable heights. Although convincing arguments may be made to

justify either of the two extreme procedures, some sort of "middle ground" procedure has more practical appeal to most experimenters and consulting statisticians.

A procedure that has some popular appeal and falls between the two extremes mentioned above consists of two stages. The first stage is an overall test of the hypothesis "no treatment differences" at an acceptable α level, say $\alpha = 0.05$. If the hypothesis is accepted, no further comparisons are made. In this way an overall experimentwise α level is maintained at or below the specified level.

If the null hypothesis is rejected, then some acceptable two-sample procedure is applied to all pairs of treatments which may be of interest, or to any other contrasts of interest. The nominal α level used in this second-stage procedure has no real probabilistic meaning, since the tests are conditional on the result of the first stage, but the method preserves most of the desired power characteristic of the two-sample procedure.

The purpose of this report is to discuss and compare some procedures that fall into the various categories above. In the next section some parametric multiple-comparisons procedures are discussed. Section III is concerned with some nonparametric procedures. A nonparametric multiple-comparisons procedure based on the rank transform, apparently not previously considered in the literature, is introduced in Section IV, along with some justification for its use. A collection of useful equations is given in Section V.

A test is usually called parametric or nonparametric, depending on whether the α level of the test is or is not a function of untested assumptions concerning the form of the distribution function. Therefore, the two-stage procedures are classified as parametric or nonparametric, depending on whether the first-stage test is parametric or nonparametric. Since the α level of the second-stage test is usually not known, it is no longer a hypothesis test in the usual sense but rather merely a convenient yardstick for separating some treatments from others. Although this seemingly opens the door to all types of procedures for the second stage, there are intuitive reasons for selecting a second-stage procedure that agrees "in spirit" with the philosophy behind the choice for a first-stage procedure. That is, a second-stage procedure that is sensitive to the same types of differences as the first-stage procedure will tend to produce results that are more in agreement with the results of the first-stage procedure. For this reason, the second-stage procedure is often

a two-sample version of the first-stage procedure.

II. PARAMETRIC METHODS

Virtually all parametric multiple-comparisons procedures assume that some sort of linear model exists with a normally distributed error term. Let $\mu_1,\ldots,\mu_k$ denote the means of the treatments of interest, and let $\bar{X}_1,\ldots,\bar{X}_k$ be the corresponding maximum likelihood estimates of those means. Further, let MSE denote the denominator of the F statistic usually used in the analysis of variance test of the hypothesis that all k means are equal. Then, $(MSE)/n_i$ is an estimator of the variance of $\bar{X}_i$, where n_i is the number of observations associated with $\bar{X}_i$.

Some of the more popular parametric procedures are listed below, in the notation of Carmer and Swanson.[1]

LSD (Least Significant Difference). Consider populations i and j to be different if the inequality

$$|\bar{X}_i - \bar{X}_j| > t_{\alpha/2} \sqrt{MSE} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

is satisfied, where $t_{\alpha/2}$ is obtained from t-tables with the same degrees of freedom as MSE. The α level applies to each individual comparison.

FSD (Fisher's Significant Difference). The LSD test is used here, but only if a preliminary F test has rejected the null hypothesis of no differences.

TSD (Tukey's Significant Difference). This is the same as LSD except $t_{\alpha/2}$ is replaced by a larger value q_α, which may be obtained from tables of the studentized range. No preliminary test is necessary. The α level applies to all pairwise comparisons simultaneously.

SSD (Scheffe's Significant Difference). This is the same as LSD and TSD except $t_{\alpha/2}$ or q_α is replaced by a still larger critical value, which equals the square root of the critical value of the F statistic used in the preliminary test in FSD. No preliminary test is applied, however. The α level applies to all possible contrasts of the μ's simultaneously.

BET (Bayes' Exact Test). The same general procedure is used again here, except that $t_{\alpha/2}$ is replaced by a critical value that is a function of the F statistic used in FSD. In particular, if F is small (indicating homogeneous sample means), the multiple comparisons are still made but with a relatively large critical value, while if F is large (indicating heterogeneous sample means), the critical value is smaller and differences are more likely to be declared significant. Thus, BET provides an interesting compromise between the LSD and FSD methods. The interesting question of interpreting the α level is sidestepped, however, and replaced with a measure of "minimum average risk."

Two other procedures, SNK (Student-Newman-Keuls) and MRT (Multiple-Range Test), resemble LSD except the critical value $t_{\alpha/2}$ is replaced by various critical values, depending on how many of the $\overline{X}$'s are intermediate in value to the $\overline{X}_i$ and $\overline{X}_j$ being considered. These critical values lie somewhere between $t_{\alpha/2}$ and q_α, with the MRT critical values being, in general, less than the corresponding SNK values.

Some of these procedures are easy to compare directly. For example, MRT will tend to have more pairs declared significant than will SNK because of the inequality of their critical values, and both will have more significant differences noted than TSD or SSD. Comparisons with FSD and BET are not as clear, however. An extensive Monte Carlo study of the relative merits of the above procedures resulted in the following conclusions by Carmer and Swanson.[1]

1) If one agrees with the notion that an experimenter should want to use a procedure capable of detecting a real difference when it exists, then one should not use TSD, SSD, or SNK.

2) The LSD procedure appears to offer too little protection against a Type I error, and offers little advantage over FSD, BET, and MRT in detecting real differences when they exist, and, therefore, should not be used.

3) In agreement with Duncan, who says that it makes more sense for a critical value to depend on F than on the number of samples, BET should be preferred over his MRT procedure.

4) Although FSD and BET appear to be more practical than the other procedures studied, there is little basis for choosing between the two.

The interested reader is referred to Waller[2] and Waller and Duncan[3] for a presentation of the BET and some exact tables.

III. NONPARAMETRIC METHODS

Three of the more popular nonparametric multiple-comparisons procedures were compared by Lin and Haseman.[4] These methods apply only to the completely randomized design. The first procedure is a nonparametric analog to the FSD procedure and is recommended by Conover.[5] The first stage is a Kruskal-Wallis test for overall differences. If the test is significant, pairwise comparisons are made by using the Mann-Whitney test, which involves reranking the observations for each comparison.

The second procedure is due to Nemenyi[6] but is usually attributed to Dunn.[7] The same overall ranks that are used to replace the observations in the Kruskal-Wallis test are used here also, but no preliminary test is applied. Treatments i and j are considered different if the inequality

$$|\bar{R}_i - \bar{R}_j| > \sqrt{h_\alpha}\ \sqrt{MST}\ \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

is satisfied, where h_α is the critical value from a Kruskal-Wallis test, MST refers to the "mean square total"

$$MST = \frac{\text{Total sum of squares}}{\text{Total degrees of freedom}}$$

of the ranks, which equals N(N+1)/12 if there are no ties, and n_i and n_j are the respective sample sizes. The α level covers all possible contrasts in the spirit of Scheffe's (SSD) procedure. Actually, Dunn suggests using, instead of $\sqrt{h_\alpha}$, the $1-\alpha/(2p)$ quantile from the standard normal distribution, where p is the total number of contrasts to be considered. Thus, for all pairwise comparisons, p equals $\binom{k}{2}$, which may be quite large for a moderate number of samples, k.

The third procedure was proposed independently by Steel[8] and Dwass.[9] As in the Nemenyi-Dunn procedure, no preliminary overall test is performed. Rather, each pair of samples being compared is ranked between themselves, and the larger

of the two rank sums is compared against a critical value, which ensures one of an overall level of significance α applicable to all pairwise comparisons, in the spirit of the TSD procedure.

After extensive Monte Carlo simulations under both the null hypothesis and alternatives involving normal, uniform, and exponential distributions, Lin and Haseman[4] reach the following conclusions.

1) The Nemenyi-Dunn and Steel-Dwass procedures seem to unduly stress protection against Type I errors at the expense of power to detect real differences when they exist.

2) The Kruskal-Wallis-Mann-Whitney test seems to provide a better balance between Type I and Type II errors, in agreement with the corresponding results found by Carmer and Swanson for the FSD procedure.

IV. THE RANK TRANSFORM PROCEDURE

The Rank Transform (RT) procedure consists of ranking the observations from the smallest to largest and then applying a reasonable parametric procedure to the ranks. For the completely randomized design the rank transform procedure analogous to the FSD method has been compared with FSD using Monte Carlo simulation. Of course, the F test on the ranks is equivalent to the Kruskal-Wallis test, so the only difference between this rank transform procedure and the Kruskal-Wallis-Mann-Whitney test reported above is in the multiple-comparisons procedure following significance in the Kruskal-Wallis test. This procedure is much simpler than the repeated use of the Mann-Whitney test because the original ranks are used throughout the analysis instead of re-ranking for each pairwise comparison. In particular, if there are no ties the LSD analog indicates populations i and j to be significantly different if the inequality

$$|\bar{R}_i - \bar{R}_j| > t_{\alpha/2} \sqrt{\frac{N(N+1)}{12}} \sqrt{\frac{N-1-T}{N-k}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

is satisfied, where $\bar{R}_i$ and $\bar{R}_j$ are the average ranks for the corresponding samples, $t_{\alpha/2}$ is the same value used in LSD and FSD, N is the sum of all the sample sizes, and T is the Kruskal-Wallis statistic. Although this procedure is simply the rank transform counterpart to the FSD procedure, one can readily see

that the larger the value of T, the easier it is to obtain significant differences, much like the BET procedure. Also, one can see the difference between this method and the Nemenyi-Dunn procedure. Where the Nemenyi-Dunn procedure uses "mean square total" on the right side of the inequality, this method uses "mean square error," which may be larger or smaller than mean square total, depending on whether T is smaller or larger than its mean k-1. Of course, the only time the above inequality will be used is when T is significant, in which case T will be much larger than k-1.

This RT procedure was compared with the FSD procedure under the null hypothesis, with "medium" nonnull conditions, and with "strong" nonnull effects, as detailed in Table I. One thousand simulations were made for each of the three situations combined with four populations: normal, lognormal, exponential, and Cauchy. The Kruskal-Wallis test used $\alpha = 0.05$ in all cases. The second-stage results for $\alpha = 0.05$ and $\alpha = 0.10$ are given.

TABLE I

THE TWELVE CONDITIONS UNDER WHICH THE RT AND FSD PROCEDURES WERE COMPARED

Population (Sample Size)	1 (n_1=7)	2 (n_2=8)	3 (n_3=9)	4 (n_4=10)
a) No effects				
1.	N(0,1)[a]	N(0,1)	N(0,1)	N(0,1)
2.	$-\ell n$ U[b]	$-\ell n$ U	$-\ell n$ U	$-\ell n$ U
3.	exp{N(0,1)}	exp{N(0,1)}	exp{N(0,1)}	exp{N(0,1)}
4.	C(0,1)[c]	C(0,1)	C(0,1)	C(0,1)
b) Medium effects				
5.	N(0,1)	N(0,1)	N(.5,1.5)	N(1,2)
6.	$-\ell n$ U	$-\ell n$ U	$-\frac{4}{3}\ell n$ U	$-\frac{5}{3}\ell n$ U
7.	exp{N(0,1)}	exp{N(0,1)}	exp{N(.5,1)}	exp{N(.84,1)}
8.	C(0,1)	C(0,1)	C(.5,1.5)	C(1,2)
c) Strong effects				
9.	N(0,1)	N(0,1)	N(1,1.5)	N(2,2)
10.	$-\ell n$ U	$-\ell n$ U	$-2\ell n$ U	$-3\ell n$ U
11.	exp{N(0,1)}	exp{N(0,1)}	exp{N(1,1)}	exp{N(2,1)}
12.	C(0,1)	C(0,1)	C(1,1.5)	C(2,2)

[a] $N(\mu,\sigma^2)$ = normal random variable.
[b] U = uniform random variable on (0,1).
[c] C(a,b) = a + b tan (π(U-.5)) = Cauchy random variable with a=median,b=scale factor.

Table II shows the proportion of times the null hypothesis was rejected using the F test and using the Kruskal-Wallis test. With normal populations these results agree with the theory which says the Kruskal-Wallis test is not as powerful as the F test. For the other three distributions the Kruskal-Wallis test appears to have as much or more power than the F test, although the heterogeneity of variance present in the lognormal situation probably causes the lack of power in both tests.

When the null hypothesis was rejected, multiple comparisons were made with the FSD and RT procedures, as reported in Table III. The comparison of error rates and power rates were computed. It is interesting to note that in the case with normal populations, the power of the RT procedure matches the power of the FSD method even though the Kruskal-Wallis test rejected the null hypothesis fewer times than the F test, and therefore the RT method was applied fewer times than the FSD method. A larger number of Type I errors also accompanies the RT method, although the proportion of Type I errors is still well below the nominal value 0.05. For the nonnormal distributions reported in Table III, the RT method declared more population pairs to be different than the FSD procedure did, in situations where the populations were not different (Type I error) as well as when they were.

TABLE II

THE PROPORTION OF TIMES THE F TEST AND THE KRUSKAL-WALLIS TEST REJECTED THE HYPOTHESIS OF NO OVERALL DIFFERENCES AT $\alpha = 0.05$

		No Effects	Medium	Strong
Normal	F:	0.036	0.339	0.910
	K-W:	0.039	0.322	0.896
Exponential	F:	0.049	0.106	0.391
	K-W:	0.055	0.119	0.423
Lognormal	F:	0.040	0.038	0.034
	K-W:	0.064	0.062	0.048
Cauchy	F:	0.028	0.024	0.057
	K-W:	0.053	0.106	0.248

TABLE III

THE PROPORTION OF PAIRWISE COMPARISONS WHICH WERE DECLARED SIGNIFICANT

		I. No differences were present in the experiment	II. Some differences were present in the experiment: (a) among identical pairs	II. Some differences were present in the experiment: (b) among pairs with differences
First stage $\alpha = 0.05$; Second stage $\alpha = 0.05$.				
Normal	FSD	0.015	0.016	0.379
	RT	0.016	0.033	0.380
Exponential	FSD	0.021	0.003	0.131
	RT	0.022	0.023	0.139
Lognormal	FSD	0.018	0.014	0.016
	RT	0.027	0.022	0.023
Cauchy	FSD	0.011	0.005	0.019
	RT	0.022	0.019	0.087
First stage $\alpha = 0.05$; Second stage $\alpha = 0.10$.				
Normal	FSD	0.018	0.041	0.438
	RT	0.019	0.063	0.434
Exponential	FSD	0.025	0.005	0.148
	RT	0.027	0.044	0.166
Lognormal	FSD	0.020	0.018	0.018
	RT	0.033	0.025	0.028
Cauchy	FSD	0.013	0.006	0.022
	RT	0.027	0.033	0.106

Although this simulation study is not extensive, it provides some support for using the RT method as a multiple-comparisons procedure to follow the Kruskal-Wallis test. Because no reranking is necessary, the RT method is easier to use than the Mann-Whitney method. Perhaps further work comparing the power of the two procedures will be done.

V. A SUMMARY OF RECOMMENDED NONPARAMETRIC MULTIPLE-COMPARISONS PROCEDURES

The rank transform procedure described in the previous section may be used in any experimental situation for which a parametric procedure exists. Once the initial rank transformation is performed, with average ranks used in case of ties, the usual parametric procedures may be applied to the ranks, or to scores such as normal scores used in place of ranks if desired. Problems with ties are handled automatically and are no longer problems. No assumptions of continuity need be made. Evidence from past research indicates that these procedures are powerful and robust.

The primary disadvantage of the rank transform procedure is that, except in the completely randomized design, these procedures are not commonly in use for analysis of data from experimental designs. The Friedman test is commonly used for the randomized blocks design, so a multiple-comparisons procedure to follow the Friedman test, which has characteristics similar to the Friedman test, is needed. Similarly, a procedure is needed to follow the Durbin test for balanced incomplete block designs. Such procedures are being planned to appear in the forthcoming revision of Conover.[5] Equations for these procedures are given in this section for the interested reader. They are merely the rank analogs to the corresponding FSD procedure.

A. Kruskal-Wallis Test (Completely Randomized Design)

In the previous section no indication of how to handle ties was given, except to recommend assigning average ranks and using FSD or LSD formulas on the ranks. In case this explanation is not sufficient, more explicit instructions will now be given.

Consider the following notation.

X_{ij} = the i^{th} observation in the j^{th} sample,
$i=1,\ldots,n_j$; $j=1,\ldots,k$.

R_{ij} = the rank (or average rank in case of ties) of X_{ij}, from 1 to
$N = \sum_{j=1}^{k} n_j$.

R_j = the sum of the ranks assigned to the j^{th} sample.

$$S^2 = \frac{1}{N-1}\left(\sum_{\substack{\text{all}\\ \text{ranks}}} R_{ij}^2 - N\frac{(N+1)^2}{4}\right).$$

$$T = \frac{1}{S^2}\left(\sum_{j=1}^{k} R_j^2/n_j - \frac{N(N+1)^2}{4}\right).$$

If there are no ties S^2 reduces to $N(N+1)/12$. If T, the Kruskal-Wallis statistic exceeds the $1-\alpha$ quantile of a chi-square distribution with k-1 degrees of freedom, multiple comparisons are made using the inequality

$$\left| \frac{R_i}{n_i} - \frac{R_j}{n_j} \right| > t_{1-\alpha/2} \left(S^2 \frac{N-1-T}{N-k}\right)^{\frac{1}{2}} \left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{\frac{1}{2}}$$

for all pairs of samples, where $t_{1-\alpha/2}$ is the $1-\alpha/2$ quantile from a t distribution with N-k degrees of freedon.

As an alternative to the above procedure the ranks R_{ij} may be treated as data in the FSD procedure. The results of these two procedures are equivalent.

B. Van der Waerden Test (Completely Randomized Design)

If normal scores are used instead of ranks in the above analysis, the equations are as follows.

$A_{ij} = \Phi^{-1}(R_{ij}/(N+1))$, where $\Phi(x)$ is the standard normal distribution function.

A_j = the sum of the scores assigned to the j^{th} sample.

$$S_1^2 = \frac{1}{N-1} \sum_{\substack{\text{all} \\ \text{scores}}} A_{ij}^2 .$$

$$T_1 = \frac{1}{S_1^2} \sum_{j=1}^{k} A_j^2/n_j .$$

Multiple comparisons are based on the inequality

$$\left|\frac{A_i}{n_i} - \frac{A_j}{n_j}\right| > t_{1-\alpha/2} \left(S_1^2 \frac{N-1-T_1}{N-k}\right)^{\frac{1}{2}} \left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{\frac{1}{2}}$$

only if the statistic T_1 exceeds the $1-\alpha$ quantile of a chi-square distribution with k-1 degrees of freedom.

As an alternative to the above procedure, the scores A_{ij} may be treated as data in the FSD procedure. The results of these two procedures are equivalent.

C. Friedman Test (Randomized Complete Block Design)

The most popular nonparametric test for the randomized complete block design is the Friedman test, which is presented for the case with several observations per cell.

X_{ijn} = the n^{th} observation in block i, treatment j, i=1,...,b; j=1,...,k; n=1,...,m.

R_{ijn} = the rank (or average rank in case of ties) of X_{ijn} among those observations in block i only; from 1 to km.

R_j = the sum of all ranks assigned to treatment j.

$$S_2^2 = \frac{m}{(mk-1)} \left(\sum_{\substack{\text{all} \\ \text{ranks}}} R_{ijn}^2 - mkb(mk+1)^2/4 \right) .$$

$$T_2 = \frac{1}{S_2^2} \sum_{j=1}^{k} (R_j - bm(mk+1)/2)^2 .$$

If the Friedman test statistic T_2 exceeds the $1-\alpha$ quantile of a chi-square distribution with k-1 degrees of freedom, multiple comparisons are based on the inequality

$$|R_j - R_i| > t_{1-\alpha/2} \left(S_2^2 \frac{2b(mk-1)}{mbk-k-b+1}\right)^{1/2} \left(1 - \frac{T_2}{b(mk-1)}\right)^{1/2}$$

for all pairs of treatments i and j. If there are no ties S_2^2 reduces to $kbm^2(mk+1)/12$. The number of degrees of freedom is mbk-k-b+1.

As an alternative to the above procedure, the ranks R_{ijn} may be treated as data in an ordinary two-way analysis of variance, without interaction. The resulting F test for treatments is equivalent to the Friedman test. A significant value of F is then followed by the LSD procedure, still treating the ranks as data. These two procedures are equivalent.

D. Durbin Test (Balanced Incomplete Block Design)

The usual nonparametric test for the balanced incomplete block design and the appropriate multiple-comparisons procedure are as follows.

t = the number of treatments to be examined.

k = the number of experimental units per block ($k<t$).

b = the total number of blocks.

r = the number of times each treatment appears ($r<b$).

X_{ij} = the result of treatment j in block i, if treatment j appears in block i.

R_{ij} = the rank of X_{ij} within block i only, from 1 to k.

R_j = the sum of the r ranks assigned to treatment j; $j=1,\ldots,t$.

$$T_3 = \frac{12(t-1)}{rt(k-1)(k+1)} \sum_{j=1}^{t} \left[R_j - \frac{r(k+1)}{2}\right]^2 .$$

If T_3 exceeds the $1-\alpha$ quantile of a chi-square distribution with $t-1$ degrees of freedom, make pairwise comparisons using the inequality

$$|R_j - R_i| > t_{1-\alpha/2} \left(\frac{r(k+1)(k-1)(bk(t-1)-t\,T_3}{6(t-1)(bk-t-b+1)} \right)^{\frac{1}{2}} ,$$

where $t_{1-\alpha/2}$ is obtained from t tables with $bk-t-b+1$ degrees of freedom.

The above procedure is equivalent to the usual parametric analysis on the ranks if there are no ties. In case of extensive ties the above chi-square approximation may be inaccurate, and the parametric analysis on the ranks R_{ij} should be used instead, because of its built-in correction for ties.

ACKNOWLEDGEMENTS

The work of W. J. Conover was supported in part by the Los Alamos Scientific Laboratory.

REFERENCES

1. S. G. Carmer and M. R. Swanson, "An Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte Carlo Methods," Journal of the American Statistical Association 68 (No. 341), 66-74 (1973).

2. R. A. Waller, "A Bayes Solution to the Symmetric Multiple Comparisons Problem," Ph.D. Thesis, Johns Hopkins University, Baltimore, MD (1967).

3. R. A. Waller and D. B. Duncan, "A Bayes Rule for the Symmetric Multiple Comparisons Problem," Journal of The American Statistical Association, 64, 1484-1503 (1969). Corrigenda 67 (No. 337), 253-255 (1972).

4. F. A. Lin and J. K. Haseman, "An Evaluation of Some Nonparametric Multiple Comparison Procedures by Monte Carlo Methods," Communications in Statistics-Simulation and Computation B7 (No. 2), 117-128 (1978).

5. W. J. Conover, Practical Nonparametric Statistics (John Wiley and Sons, New York, 1971).

6. P. Nemenyi, "Distribution-Free Multiple Comparisons," Ph.D. Thesis, Princeton University, Princeton, NJ (1963).

7. Olive Jean Dunn, "Multiple Comparisons Using Rank Sums," Technometrics 6 (No. 3), 241-252 (1964).

8. R. G. D. Steel, "A Rank Sum Test for Comparing All Pairs of Treatments," Technometrics 2, 197-207 (1970).

9. M. Dwass, "Some k-sample Rank-Order Tests," Contributions to Probability and Statistics (I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Meadow, and H. B. Mann, Eds.), (Stanford University Press, Stanford, California) pp. 198-202.