Bulletin of the Psychonomic Society
1979, Vol. 14 (2), 95-98

A Monte Carlo study of nonparametric multiple-comparison tests for a two-way layout

JAMES D. CHURCH and EDWARD L. WIKE
University of Kansas, Lawrence, Kansas 66045

A Monte Carlo study of two overall tests, the Friedman and Doksum, and five nonparametric multiple-comparison tests, the McDonald-Thompson, Nemenyi, Wilcoxon, Rhyne-Steel, and sign test, was done to determine the tests' Type I error rates in a two-way layout with k = 3, 5, and 7 treatments and n = 8, 11, and 15 blocks. It was found that (1) the Friedman test was superior to the Doksum test as an overall test, and (2) the Wilcoxon signed-ranks test, when protected by a significant Friedman test, was the best pairwise multiple-comparison procedure. In a second Monte Carlo study, a modified sign test, termed the stepped-down sign test, was found to be superior to the Wilcoxon test as a protected test; it is recommended for pairwise comparisons in a two-way layout.

Experimenters ask a variety of questions regarding the significance of k treatments, and multiple-comparison tests (MCTs) have been devised by statisticians to answer these diverse questions. Multiple-comparison problems are complicated by (1) the diversity of MCTs that are available and (2) the fact that errors of inference can be indexed both in terms of numbers of comparisons and number of experiments. To assess and compare MCTs empirically, Monte Carlo studies have been performed (e.g., Petrinovich & Hardyck, 1969). Recently, Bernhardson (1975) and Carmer and Swanson (1973) did simulations in which they asked a different important question: How do various pairwise MCTs perform when overall F tests are significant? This question is important because (1) MCTs are usually applied after a significant overall test, and (2) MCTs are derived without respect to the outcome of overall tests. Both Bernhardson and Carmer and Swanson found that Fisher's l.s.d. test had excellent Type I error rates and power when applied after a significant F. With the restriction that a MCT can reject equality of pairs of treatments only after a significant overall test, the MCT is referred to as a "protected" test. Thus, when used in this limited way, the Fisher test has also been termed the "protected t test."

Bernhardson (1975) described six different observed Type I error rates. Four of these error rates are used in the present study and are defined as follows: (1) α_1 is the total number of significant comparisons divided by the total number of comparisons, that is, the product of the total number of experiments, N, and the number of comparisons per experiment, k(k – 1)/2. (2) α_2 is the total number of experiments with one or more significant comparisons divided by the total number of experiments, N. (3) α_C is the number of experiments with one or more significant comparisons in which the overall test is significant divided by the total number of experiments, N. (4) α_D is the number of experiments with one or more significant comparisons when the overall test is significant divided by the number of experiments in which the overall test is significant. α_1 and α_2 are the conventional comparisonwise and experimentwise error rates, respectively. α_C and α_D are experimentwise error rates that are obtained when MCTs are employed in a protected manner. If a protected MCT demonstrates good Type I error performance, then α_C should be near the nominal significance level (such as .05 or .01) and α_D near 1; that is, the MCT almost always identifies some pair of treatments as different when the overall test rejects the hypothesis of equality of the k treatments.

More recently, Wike and Church (1978) did a Monte Carlo study of four nonparametric MCTs for a one-way layout using Bernhardson's (1975) Type I error rates. In accord with the parametric MCT investigations above, it was found that Wilcoxon's rank-sum test was superior to the other tests when it was protected by an overall Kruskal-Wallis H test.

This paper presents the results of a Monte Carlo study of the Type I errors for five nonparametric MCTs of equality of pairs of treatment distributions in a two-way layout with k treatments, n blocks, and one observation per cell, both when the MCTs are protected by a Friedman (1937) rank-sum test of simultaneous equality of the k treatment distributions and when the MCT are not so protected.[1]

The specific aims of the study were (1) to evaluate and compare the Type I errors for Friedman's (1937) rank-sum test and Doksum's (1967) signed-rank test of the simultaneous equality of the k distributions for small samples (k = 3, 5, and 7; n = 8, 11, and 15); (2) to determine the above four Type I error rates for five MCTs: McDonald and Thompson's (1967) rank-sum test, Wilcoxon's (1945) signed-rank test, Nemenyi's (1963) signed-rank test, the sign test, and Rhyne and Steel's (1967) k-sample sign test; (3) to assess the relative

0090-5054/79/080095-04$00.65/0

effects of three independent factors, MCTs, number of treatments, and number of blocks, on the various error rates; and (4) to offer a recommendation as to the best protected nonparametric MCT for pairwise treatment comparisons in a complete two-way layout with one observation per cell. Except for Rhyne and Steel's test, all of the above tests are described and discussed by Hollander and Wolfe (1973) (see also Miller, 1966).

METHOD

Monte Carlo Procedures

For each of the nine k,n combinations, k sets of n random numbers, uniformly distributed on the interval from 0 to 1, were generated on a Honeywell 66/60 computer. A total of 1,000 runs were made for each k,n combination, thus 9,000 simulated experiments were conducted. For each experiment, the two overall tests and the five MCTs were performed at significance levels of .05 and .01. For each k,n pair, counts were obtained for (1) the total number of experiments in which each overall test was significant, and, for each MCT, (2) the total number of significant pairwise comparisons, (3) the number of significant comparisons when Friedman's (1937) test was significant, (4) the number of experiments with one or more significant comparison, and (5) the number of experiments with one or more significant comparison and a significant Friedman test. Since the random numbers were generated from the same distribution throughout, each significant comparison is a Type I error. The study yielded no information regarding the nonnull case, that is, the occurrence of Type II errors and the power of the tests.

The original intention was to include Monte Carlo studies in which the errors within each block were positively and symmetrically correlated, which might occur when the k observations within each block represented the exposure of a single subject to all k treatments. A standard method of generating positively and symmetrically correlated variables $Y_1, Y_2, \ldots, Y_k$ (within a block) is to first generate a list of independent variables, $X_1, X_2, \ldots, X_k$, and to define

$$Y_i = X_i + \lambda \sum_{j=1}^{k} X_j,$$

for $i = 1, 2, \ldots, k$, where λ is a small positive constant. However, it can be shown that the rank sums, signed ranks, and sign counts, on which the tests are based, are the same for these correlated observations and the underlying independent observations. Thus the Monte Carlo results presented here for the case of independent errors are also representative of cases in which the errors are correlated by the transformation described above.

Tests

Three kinds of nonparametric tests were studied: rank-sum tests, signed-rank tests, and sign tests. In the first kind, the test statistics are based on the rank sums produced by ranking the k observations within each block across treatments, and then summing the n ranks down the blocks for each treatment. The signed-rank statistics are calculated by ranking the absolute differences of pairs of observations in the same block, for each pair of treatments, and summing the resulting ranks associated with differences of fixed sign. The sign test statistics are based on counts of the total number of differences, of fixed sign, of pairs of observations in the same block, for each pair of treatments.

For the rank-sum tests and the signed-rank tests, most critical values were obtained from formulas based on asymptotic approximations given by Hollander and Wolfe (1973), and thus the actual significance levels only approximate the nominal levels of .05 and .01. The exceptions were the critical values for Friedman's (1937) test for $k = 3$, $n = 8$, and for $k = 3$, $n = 11$, which were taken from a table in Hollander and Wolfe, and critical values for Wilcoxon's (1945) test, taken from Wilcoxon, Katti, and Wilcox (1963). Critical values for Rhyne and Steel's (1967) sign test were taken from a table in Miller (1966). Critical values for the simple sign test were obtained from MacKinnon's (1964) table.

RESULTS AND DISCUSSION

Trial runs of the computer program suggested that the Doksum (1967) test would perform poorly with small samples. Accordingly, it was decided to continue collecting information on the Type I errors for the Doksum test, but to protect the MCTs only with Friedman's (1937) test, which did well in trial runs. Results of the full study confirmed the poor performance of Doksum's test. At significance level $\alpha = .05$, the observed proportions of Type I errors ranged from .064 to .134 for the nine k,n combinations, with an average proportion of .094. At level $\alpha = .01$, the rejection proportions ranged from .005 to .038, with an average of .018. The rejection proportions were most excessive for $k = 3$, and near the nominal levels .05 and .01 for $k = 7$. The performance of Friedman's test was much better. At level $\alpha = .05$, proportions of Type I errors ranged from .037 to .054, with an average of .045. At level $\alpha = .01$, observed rejection proportions ranged from .004 to .011, with an average of .007. We may conclude that Friedman's test is superior to Doksum's test as an overall test of k treatments in a two-way layout with small samples.

Earlier (Wike & Church, 1977), we advocated the factorial design of Monte Carlo studies and the assessment of the importance of the factors in the design by analysis of variance and percents of accounted-for variance. The present study had an orthogonal design, a 3 (k) by 3 (n) by 5 (t, MCT) plan. The four Type I error rates for $\alpha = .05$ and .01 were subjected to a series of ANOVAs that disclosed that 52 out of 72 possible effects were significant ($p < .05$). To deal with this large number of significant effects and to estimate their relative magnitudes, Hays' (1963) formulas for percentages of accounted-for variance were computed for the different error rates. Every ANOVA revealed that two effects, t and nt, accounted for substantial portions of the variance in error rates. Together, these two effects accounted for from 55% of the variance for α_2 at $\alpha = .01$ to 91% of α_1 at $\alpha = .05$, with an average of 72% over the four α rates and two α levels. However, since k, n, and kt sometimes accounted for significant portions of variance, it seemed judicious to display the error rate data in both t by k (Table 1) and t by n arrays (Table 2). By examining these tables, we may evaluate further the performance of five MCTs.

The McDonald-Thompson (1967) rank-sum test is designed to control the experimentwise rate α_2, and it performs fairly well in this respect. At level $\alpha = .05$,

Table 1
Error Rates for the Six Tests for Different Numbers of Groups (k)

	k											
	α_1			α_2			α_C			α_D		
Test	3	5	7	3	5	7	3	5	7	3	5	7
							$\alpha = .05$					
McDonald-Thompson	.021	.005	.003	.054	.041	.044	.045	.034	.028	.964	.729	.697
Nemenyi	.014	.002	.001	.038	.024	.010	.023	.014	.004	.486	.293	.090
Wilcoxon	.051	.048	.050	.128	.300	.486	.041	.045	.041	.871	.964	1.000
Rhyne-Steel	.008	.003	.003	.023	.026	.057	.020	.014	.021	.421	.300	.516
Sign	.017	.018	.018	.045	.138	.247	.030	.038	.038	.650	.807	.934
Step-Down Sign	.251	.252	.250	.209	.498	.719	.047	.047	.041	1.000	1.000	1.000
							$\alpha = .01$					
McDonald-Thompson	.003	.001	.000	.008	.012	.009	.007	.004	.004	.690	.765	.579
Nemenyi	.001	.000	.000	.003	.000	.000	.001	.000	.000	.138	.000	.000
Wilcoxon	.009	.009	.009	.025	.077	.136	.008	.006	.006	.828	1.000	1.000
Rhyne-Steel	.003	.003	.003	.010	.023	.050	.005	.003	.004	.552	.471	.579
Sign	.005	.005	.005	.016	.047	.087	.007	.004	.005	.690	.765	.789
Step-Down Sign	.114	.115	.110	.096	.262	.424	.010	.006	.006	1.000	1.000	1.000

Table 2
Error Rates for the Six Tests for Different Sample Sizes (n)

	n											
	α_1			α_2			α_C			α_D		
Test	8	11	15	8	11	15	8	11	15	8	11	15
							$\alpha = .05$					
McDonald-Thompson	.004	.005	.005	.040	.051	.048	.034	.037	.037	.778	.806	.818
Nemenyi	.001	.002	.003	.012	.025	.035	.007	.015	.018	.159	.324	.401
Wilcoxon	.051	.053	.045	.310	.322	.281	.040	.044	.043	.944	.942	.942
Rhyne-Steel	.007	.002	.002	.069	.019	.018	.028	.014	.013	.659	.309	.277
Sign	.007	.011	.034	.069	.102	.259	.028	.033	.045	.659	.712	.993
Step-Down Sign	.069	.066	.115	.421	.421	.583	.042	.046	.046	1.000	1.000	1.000
							$\alpha = .01$					
McDonald-Thompson	.001	.001	.001	.010	.010	.009	.004	.006	.005	.520	.773	.778
Nemenyi	.000	.000	.000	.000	.001	.002	.000	.001	.001	.000	.091	.111
Wilcoxon	.007	.010	.009	.069	.087	.082	.008	.006	.005	1.000	.864	.889
Rhyne-Steel	.007	.001	.000	.069	.012	.002	.008	.002	.001	1.000	.318	.167
Sign	.007	.001	.007	.069	.012	.069	.008	.002	.005	1.000	.318	.889
Step-Down Sign	.069	.011	.034	.421	.102	.259	.008	.007	.006	1.000	1.000	1.000

the observed α_2 error rates ranged from .031 to .063, with an average of .046. At level $\alpha = .01$, the α_2 values ranged from .004 to .015, with an average of .010. However, as shown in Tables 1 and 2, its obtained α_C error rates were consistently below the nominal α = .05 and .01 levels, and the obtained α_D rates were below the ideal value of unity. An example of the latter deficiency is at level $\alpha = .05$ for k = 5 and 7, where it was found that in about 30% of the experiments when the Friedman (1937) test was significant, the McDonald-Thompson test failed to identify any pairs of treatments as different.

The Nemenyi (1963) and Rhyne-Steel (1967) tests may be quickly dismissed as MCTs because their performance was obviously deficient on all four error rates. In fact, the Nemenyi could be classed as unsuitable as a small-sample MCT because of its low incidence of Type I errors throughout.

Wilcoxon's (1945) signed-ranks test is a two-sample procedure, but it was used as a MCT. As shown in the tables, it performed better than all other tests in terms of its observed α_1, α_C, and α_D error rates. Its α_1 rates ranged from .042 to .056, with an average of .050, at the $\alpha = .05$ level; at the $\alpha = .01$ level, α_1 ranged from .006 to .013, with an average of .009. As a protected test, its α_C values fell only a little short of the nominal $\alpha = .05$ and .01 levels, and its α_D values were nearer to unity than the other MCTs.

If Friedman's (1937) test is applied to two treatments, the resulting rank-sum statistics are equivalent to the sign counts for the sign test. This suggests that the sign test might function well as a MCT when protected by Friedman's test. However, the Monte Carlo results suggested otherwise.

Although the sign test was the second-best test among the MCTs, it was clearly too conservative with respect to its Type I error rates (cf. Tables 1 and 2). The source of the difficulty appeared to reside in the conservativeness of the required values for significance in the sign test table. For example, for $n = 8$ at $\alpha = .05$, a critical value of 8 (all like signs) is required for significance. The actual two-sided binomial probability in this instance is .008. (The binomial probability associated with the critical value of 7 is about .07, which exceeds the nominal .05.) These considerations led to an expansion of the study to include what we may call the stepped-down sign test, which is the ordinary sign test except that the critical value corresponded to the smallest actual (binomial) level which is just greater than or equal to the nominal significance level. (The usual sign test critical values are those with the largest binomial probabilities not exceeding the nominal level of significance.) The critical values for the stepped-down sign test and their binomial probabilities were: for $n = 8$, 7 ($p = .070$) at both $\alpha = .05$ and $\alpha = .01$; for $n = 11$, 9 ($p = .066$) at $\alpha = .05$, and 10 ($p = .012$) at $\alpha = .01$; and for $n = 15$, 11 ($p = .118$) at $\alpha = .05$, and 12 ($p = .036$) at $\alpha = .01$. Each of the nine k,n combinations was replicated 1,000 times with these stepped-down values and with the same random starting digits ("seeds") as in the original study.

It is apparent from Tables 1 and 2 that the stepped-down sign test is a decided improvement over the simple sign test when these tests are employed as protected tests. In fact, the performance of the stepped-down sign test equals or exceeds the Wilcoxon (1945) signed-ranks test with respect to its α_C errors, and its α_D levels achieves the desired level of unity in every k,n combination.

What is the recommended protected MCT for pairwise comparisons in a two-way layout? Our preference is the stepped-down sign test because of its excellent α_C and α_D error rates and because it is a distribution-free test. The Wilcoxon (1945) signed-ranks test, which is nearly as good, is a strong alternative test but is not always a distribution-free test (see Hollander & Wolfe, 1973).

Finally, it is worthy of note that the α_C error rate for the stepped-down sign test converges upon the Type I error rate of the overall Friedman (1937) test. Similar relationships have been observed by Bernhardson (1975) and by Carmer and Swanson (1973) for the protected t test and by Wike and Church (1978) for the protected Wilcoxon rank-sum test. In each of these situations, one of the competing MCTs was clearly superior as a protected MCT.

REFERENCES

Bernhardson, C. S. Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. *Biometrics,* 1975, **31**, 229-232.

Carmer, S. G., & Swanson, M. R. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association,* 1973, **68**, 66-74.

Doksum, K. Robust procedures for some linear models with one observation per cell. *Annals of Mathematical Statistics*, 1967, **38**, 878-883.

Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 1937, **32**, 675-701.

Hays, W. L. *Statistics for psychologists.* New York: Holt, Rinehart, & Winston, 1963.

Hollander, M., & Wolfe, D. A. *Nonparametric statistical methods.* New York: Wiley, 1973.

MacKinnon, W. J. Table for both the sign test and distribution-free confidence intervals for the median for sample sizes to 1,000. *Journal of the American Statistical Association*, 1964, **59**, 935-956.

McDonald, B. J., & Thompson, W. A., Jr. Rank sum multiple comparisons in one- and two-way classifications. *Biometrika*, 1967, **54**, 487-497.

Miller, R. G., Jr. *Simultaneous statistical inference.* New York: McGraw-Hill, 1966.

Nemenyi, P. Distribution-free multiple comparisons. Unpublished doctoral thesis, Princeton University, 1963.

Petrinovich, L. F., & Hardyck, C. D. Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. *Psychological Bulletin*, 1969, **71**, 43-54.

Rhyne, A. L., & Steel, R. G. D. A multiple comparisons sign test: All pairs of treatments. *Biometrics*, 1967, **23**, 539-549.

Wike, E. L., & Church, J. D. Analysis of variance methods for the design and analysis of Monte Carlo statistical studies. *Bulletin of the Psychonomic Society,* 1977, **10**, 131-133.

Wike, E. L., & Church, J. D. A Monte Carlo investigation of four nonparametric multiple-comparison tests for k independent groups. *Bulletin of the Psychonomic Society,* 1978, **11**, 25-28.

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

Wilcoxon, F., Katti, S. K., & Wilcox, R. A. *Critical values and probability levels for the Wilcoxon rank sum test and the signed rank test.* American Cyanamid Company & Florida State University, 1963.

NOTE

1. Actually, rejections of the null hypothesis by the Friedman, Doksum, and various MCTs are regarded as indicants of location differences as a result of treatment effects.

(Received for publication April 19, 1979.)