Journal of Modern Applied Statistical Methods

Volume 6 | Issue 2 Article 6

11-1-2007

Multiple Comparison of Medians Using Permutation Tests

Scott J. Richter
University of North Carolina at Greensboro, sjricht2@uncg.edu

Melinda H. McCann
Oklahoma State University

DOI: 10.22237/jmasm/1193889900

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 399- 412

1538 – 9472/07/$95.00

Multiple Comparison Of Medians Using Permutation Tests

Scott J. Richter
University of North Carolina at Greensboro

Melinda H. McCann
Oklahoma State University

A robust method is proposed for simultaneous pairwise comparison using permutation tests and median differences. The new procedure provides strong control of familywise error rate and has better power properties than the median procedure of Nemenyi/Levy. It can be more powerful than the Tukey-Kramer procedure using mean differences, especially for nonnormal distributions and unequal sample sizes.

Key words: Simultaneous inference, pairwise comparisons, median difference, permutation test.

Introduction

The technique of using permutation methods for multiple comparisons has received relatively little attention in the literature. Nemenyi (1963) and later Levy (1979) proposed a procedure using medians, with the maximum of the differences of pairwise Mood statistics used to construct the reference distribution. Miller (1966, 1981), and more recently Higgins (2004), proposed a permutation version of the Tukey-Kramer method (Tukey, 1949; Kramer, 1956), where the range of the sample means is calculated for each permutation of observations among the k groups to obtain the reference distribution. The mean difference for each pair of means is then compared to this reference distribution to determine statistically significant differences. However, when distributions are skewed or there are outliers in the data, it may be desirable to make comparisons of medians rather than means. Thus, a logical extension of Miller's procedure is to replace means by medians. Consider the following example.

Scott J. Richter is Associate Professor and Director of the Statistical Consulting Center. His research interests are in robust methods and applied statistics. Email him at sjricht2@uncg.edu. Melinda McCann is Associate Professor of Statistics. Her research interests involve multiple comparisons procedures and their applications.

Example

Manly (1997) reported the data in Table 1 based on articles by Powell & Russell (1984, 1985) and Linton et al (1989). The data represent dry biomass (in mg) of ants for 24 eastern horned lizards, taken in three months in 1980.

It is desired to determine which, if any, of the months have different consumptions. The relation between the means and medians for each month suggests that the distributions of biomass are skewed, and that the means may not be representative of monthly consumption. Thus, comparisons based on medians may be more appropriate.

Both the median procedure of Nemenyi and Levy and Miller's procedure permute freely across all groups (unrestricted randomization). However, this unrestricted randomization scheme has been criticized. Petrondas and Gabriel (1983) contend that Miller's approach does not control the familywise error rate (FWE): the probability of making at least one false declaration of inequality, since the test for any subset hypothesis that a pair of means is equal should be based on permuting observations only among the groups whose distributions are assumed equal under the null hypothesis. The FWE actually is controlled under the overall null hypothesis that all k distributions have the same location—that is, in the weak sense (Hochberg & Tamhane, 1987), but not necessarily under a subset pairwise null hypothesis that requires only the two distributions being considered to have equal

Table 1. Dry biomass of ants for 24 eastern horned lizards, taken in three months in 1980.

Month	Dry biomass (mg)	Median	Mean
June	13, 242, 105	105.0	120.0
July	8, 59, 20, 2, 245	20.0	66.8
August	515, 488, 88, 233, 50, 600, 82, 40, 52, 1889	160.5	403.7

location, that is, in the strong sense (Hochberg & Tamhane, 1987). Accordingly, both Petrondas and Gabriel (1983) and Hochberg and Tamhane (1987) suggest performing each pairwise test separately using a Bonferroni adjustment. Similarly, Hochberg and Tamhane (1987) and Ryan and Ryan (1980) note that the median procedure of Nemenyi/Levy is not based on a joint testing family, and thus does not control the FWE. Hochberg and Tamhane (1987) instead suggest permuting separately within each pair (restricted randomization) and utilizing the maximum of pairwise Mood statistics to derive the reference distribution.

A new testing procedure is proposed based on the procedure of Nemenyi/Levy, using median difference statistics instead of differences between Mood statistics, and Type I error and power properties are compared to the new procedure to those of the Nemenyi/Levy procedure, pairwise tests using a Bonferroni adjustment, and also to the Tukey-Kramer procedure based on mean differences, which assumes normally distributed populations.

Methodology

Throughout, consider a one-way layout with k groups, where F_i is the common continuous distribution function for the i^{th} group, n_i is the sample size of the i^{th} group, and $N = n_1 + n_2 + \cdots + n_k$. Further, let μ_i be the location parameter associated with the i^{th} distribution and $\hat{\mu}_i$ be the sample median for the i^{th} group. Distributions are assumed identical for all treatments except for possible location differences.

Permutation-based Multiple Comparison Procedures:

Miller (1966, 1981) proposed a permutation analog to the Tukey-Kramer procedure for multiple pairwise comparison of several means. The reference distribution for Miller's method was based on the statistic, $\max_{1\le i<j\le k}\left|\overline{Y}_i - \overline{Y}_j\right|$, where $\overline{Y}_i$ and $\overline{Y}_j$ are the respective sample means of groups i and j. The reference distribution consists of the values of this statistic for all $\dfrac{N!}{n_1!n_2!\cdots n_k!}$ possible permutations of the observed data. Each pairwise absolute difference is compared to this distribution to determine statistical significance. Bonferroni-adjusted pairwise tests suggested by Hochberg and Tamhane (1987) and Petrondas and Gabriel (1983) will also be considered.

Nemenyi (1963) and later Levy (1979) also proposed an analog to the Tukey-Kramer procedure, but based on Mood's (1950) median test, as follows. First, calculate the grand median for the pooled sample of $N = n_1 + n_2 + \cdots + n_k$ observations. Then determine M_i, the number of observations in the i^{th} sample that exceed the grand median. The test statistic for comparing any pair is $\left|\dfrac{M_i}{n_i} - \dfrac{M_j}{n_j}\right|$. The reference distribution is based on the distribution of $\max_{1\le i<j\le k}\left|\dfrac{M_i}{n_i} - \dfrac{M_j}{n_j}\right|$, the maximum value of the test statistic over all pairs, which is calculated for a large set of random reassignments of observations to groups. As with Miller's method, an observation may be

reassigned to any of the k groups to form a new permutation. Hochberg and Tamhane (1987) suggest computing a separate grand median for each pair and calculating the test statistic above. The maximum over all pairs is then found for a large set of random reassignments, where reassignments are restricted to within each pair, and these values form the reference distribution.

A New Method Using Median Differences:

In situations involving skewed distributions or outliers it may be more appropriate to consider medians instead of means. Thus, we propose multiple comparison procedures based on median differences. The method of Nemenyi/Levy, based on Mood statistics, does utilize medians, but does not incorporate the magnitude of the difference between medians. It is believed that there may be situations when incorporating this information could lead to a more sensitive procedure.

Analogous to the mean-based procedure of Miller, the reference distribution for our new procedure is based on the distribution of $\max_{1\le i<j\le k}\left|\hat{\mu}_i-\hat{\mu}_j\right|$, the maximum of all pairwise *median* differences, calculated for a large set of random reassignments of observations to groups. Each pairwise absolute median difference is compared to this reference distribution to determine statistical significance. Both methods of permuting discussed in Section 2.1, namely restricted and unrestricted, are investigated.

Restricted Randomization Guarantees FWE Control:

The strongest argument against unrestricted permuting is that it does not necessarily provide strong control of the FWE. Restricted permuting, however, does provide strong control.

Consider k independent samples from distributions that differ by at most a location parameter. That is, for $i, j = 1, 2, ..., k$ with $i < j$, $F_i(x) = F_j\left(x - \Delta_{ij}\right)$. (Throughout Section 2.3 let $i, j = 1, 2, ..., k$ with $i < j$.) The null hypothesis then involves $\binom{k}{2}$ pairwise hypotheses of the form $H_{0ij}: \Delta_{ij} = 0$. Now consider the permutation distribution of median differences from samples i and j, and let $D_{ij}(\alpha)$ be the $1-\alpha$ percentile of this permutation distribution. Similarly, define $D_{\max}(\alpha)$ to be the $1-\alpha$ percentile of the permutation distribution for the maximum median difference among all $\binom{k}{2}$ pairs.

First consider the case under the complete null hypothesis where all $\Delta_{ij} = 0$. Let the calculated median difference from samples i and j be denoted by $\tilde{D}_{ij}$. Under the complete null hypothesis the probability that a calculated median difference from a particular pair of samples in a given permutation is the maximum difference is $\binom{k}{2}^{-1}$. Thus, each pair of samples will contribute $\alpha\binom{k}{2}^{-1}$ of the values from the pairwise difference permutation distribution to the maximum difference permutation distribution.

Consequently, the probability that any observed difference from a particular pair exceeds $D_{\max}(\alpha)$, the comparisonwise error rate, is $\alpha\binom{k}{2}^{-1}$. Alternatively, the familywise error rate is given by

$$P(\text{declare at least one pair different in location} \mid \text{all pairs have equal location})$$
$$= \sum_{\left[i,j=1,\ldots,\binom{k}{2},\, i<j\right]} P\left(\tilde{D}_{ij} \ge D_{\max}(\alpha)\right) = \binom{k}{2}\left(\alpha / \binom{k}{2}\right)$$
$$= \alpha.$$

This shows that using the permutation distribution of the maximum difference controls

the FWE in the weak sense (Hochberg & Tamhane, 1987).

Now consider the case where only $t < \binom{k}{2}$ of the pairwise null hypotheses are indeed true. For any permutation, a difference from one of these t pairs with a true pairwise null hypothesis is less likely to be the maximum difference than differences from the $\binom{k}{2} - t$ pairs where $\Delta_{ij} \neq 0$. Consequently, the comparisonwise error rate is $P\left(\tilde{D}_{ij} \geq D_{\max}(\alpha)\right) \leq \alpha \binom{k}{2}^{-1}$. Thus, the familywise error rate, the probability of rejecting at least one of the t true null hypotheses, is

$$P(\text{reject at least one true null hypothesis} \mid t \text{ true null hypotheses}) \leq t\left(\alpha / \binom{k}{2}\right) < \alpha.$$

Thus, the FWE is controlled at level α for any combination of t true and $\binom{k}{2} - t$ false hypotheses, and the FWE is controlled in the *strong* sense (Hochberg & Tamhane, 1987).

Alternatively, the FWE may be controlled by performing separate two-sample permutation tests and utilizing $\alpha \binom{k}{2}^{-1}$, a Bonferroni adjustment, as the significance level for each individual comparison. Based on their performance in the normal theory setting, it is expected that a Tukey-type permutation procedure will generally be less conservative than a procedure utilizing pairwise permutation tests with a Bonferroni adjustment.

Simulation Study

A simulation was conducted to evaluate five permutation procedures:

1. A modification of Miller's (1966, 1981) procedure, using medians instead of means and unrestricted randomization (MEDUR);
2. A modification of (1) using restricted randomization (MEDR);
3. Separate Bonferroni-adjusted pairwise permutation tests for median differences (MEDBON);
4. The procedure of Nemenyi (1963)/Levy (1979) based on differences between Mood statistics and unrestricted randomization (MOODUR);
5. A modification of (4), using restricted randomization (MOODR).

The following model was assumed to generate the data:

$$y_{ij} = \mu_i + e_{ij},$$

where $y_{ij} =$ the j^{th} observation for the i^{th} treatment $\mu_i =$ the location parameter for the i^{th} treatment $e_{ij} =$ the random error associated with the j^{th} observation for the i^{th} treatment. The e_{ij} are assumed independent and identically distributed.

Several different error distributions were examined:

- Normal ($\mu = 0, \sigma^2 = 1$);
- Uniform [-3,3];
- Exponential ($\lambda = 3$);
- Double exponential (Exp($\lambda = 3$) – Exp($\lambda = 3$));
- Location-contaminated normal (N (0,1) with 10% contamination from N (9,1)).

These choices encompass two symmetric, nonnormal distributions: the uniform (lighter-tailed than normal) and the double exponential (heavier-tailed than normal); and two skewed distributions: the exponential and contaminated normal. Models contained either three or five groups, and both equal and unequal sample sizes were examined. In most cases the total number of permutations possible is prohibitive, and thus a random sample of permutations was used to estimate the p-value for any given test. Keller-McNulty and Higgins (1987) examined the issue

of randomly sampling the permutations, and concluded that little is to be gained by taking more than 1600 randomly sampled permutations. Thus, each permutation test was based on a reference distribution estimated via a slightly conservative 2000 randomly sampled permutations, and the estimated proportions of rejections were based on 2000 randomly generated samples. The simulations were implemented using Resampling Stats version 5.0 (Resampling Stats Inc., 2000).

The familywise error rate (FWE) and any-pair power (Shaffer, 1995), the probability of detecting at least one true difference, are reported in the Tables 2-12. For the Tukey-type procedures based on medians, in cases where either all groups have identical locations or all groups had different locations, these were estimated by comparing the maximum pairwise difference from among the samples to the respective reference distribution, and counting the number of random samples where this maximum was in the top 5% of the reference distribution. In cases where some pairs had identical locations while others pairs differed in location, the FWE was estimated as the proportion of permutations where at least one of the true null hypotheses was rejected (strong FWE).

Results

Comparison of Median-based Procedures

Type I Error

All median-based procedures controlled the FWE in the strong sense (See Tables 2-4). In fact, in the cases where some pairs had equal locations and some did not, the probability of at least one false rejection was usually lower than the case where all locations were equal. As Petrondas and Gabriel (1983) admitted, their counterexample was very small, and, "for realistic, larger examples the corresponding tests (using unrestricted permuting) may be both valid and useful." It is also worth noting, however, that even though the unrestricted permuting method did not exhibit inflated FWE rates for either the median difference statistic or the Mood statistic, in cases where there was a difference between unrestricted and restricted FWE rates, the unrestricted FWE was almost always higher. This was true especially with unequal sample sizes, where error rates more than twice as large for unrestricted permuting were not uncommon. As we shall see in the next section, however, higher FWE rates did not typically lead to more powerful tests. In light of this evidence and the earlier cited criticisms of unrestricted randomization, as well as the fact that power is generally at least as good under restricted randomization, only procedures using restricted randomization will be considered in the remainder of the discussion.

Power

Consider first the case of equal sample sizes. With small group sample size ($n = 5$) and small location differences ($\Delta_1 = \Delta_2 = 0, \Delta_3 = 2$ or $\Delta_1 = \Delta_2 = 2,\ \Delta_3 = \Delta_4 = \Delta_5 = 0$), MEDR always had the highest power among the median procedures (See Tables 5 and 7). When there were larger location differences ($\Delta_1 = \Delta_2 = 2, \Delta_3 = 5$ or $\Delta_1 = \Delta_2 = 2,\ \Delta_3 = 3,\ \Delta_4 = \Delta_5 = 0$),MOODR often had highest power for normal and contaminated normal data (e.g., see Table 6). On the other hand, MEDBON had no power with $n = 5$ (See Tables 5-7). With group sample size $n = 10$ (e.g., see Table 8), MEDR was often most powerful for heavier-tailed distributions (exponential, double exponential), especially with larger location differences and more groups (e.g., 3 groups, $n = 10$, $\Delta_1 = \Delta_2 = 2, \Delta_3 = 5$; 5 groups, $n = 10,\ \Delta_1 = \Delta_2 = 2,\ \Delta_3 = \Delta_4 = \Delta_5 = 0$) while MOODR was most powerful for the latter five group scenarios for contaminated normal data. MEDBON often had higher power than MOODR, but always trailed MEDR. For $n = 20$, MEDBON was most powerful for uniform and exponential data, and all three median-based procedures had similar power for the other distributions (See Table 9). MEDR performed most consistently across different scenarios, was never much less powerful than any other procedure for nonnormal data, and was often substantially more powerful. For example, in Table 11, MEDR had power almost 200 times the power of MOODR (0.591 versus 0.003), while the largest power advantage for

Table 2. FWE – Proportion of times at least one true null hypothesis was rejected at $\alpha = 0.05$, three groups, $n_i = 5$, locations $\Delta_1 = \Delta_2 = \Delta_3 = 0$.

	Distribution				
	Normal	Uniform	Double-Exp.	Exponential	Cont.-Normal
Procedure					
MEDR	0.053	0.046	0.047	0.037	0.027
MEDUR	0.035	0.041	0.054	0.040	0.019
MOODR	0.013	0.018	0.017	0.019	0.007
MOODUR	0.009	0.013	0.011	0.013	0.003
TUKEY	0.053	0.059	0.060	0.044	0.026

Table 3. FWE – Proportion of times at least one true null hypothesis was rejected at $\alpha = 0.05$, five groups, $n_i = 5$, locations $\Delta_1 = \Delta_2 = 2; \Delta_3 = \Delta_4 = \Delta_5 = 0$.

	Distribution				
	Normal	Uniform	Double-Exp.	Exponential	Cont.-Normal
Procedure					
MEDR	0.000	0.009	0.009	0.014	0.000
MEDUR	0.000	0.023	0.017	0.021	0.001
MOODR	0.001	0.008	0.005	0.003	0.001
MOODUR	0.001	0.008	0.005	0.003	0.001
TUKEY	0.024	0.025	0.025	0.023	0.025

Table 4. FWE – Proportion of times at least one true null hypothesis was rejected at $\alpha = 0.05$, five groups, $n_1 = 3, n_2 = 4, n_3 = 5, n_4 = 6, n_5 = 7$, locations $\Delta_1 = \Delta_2 = 2; \Delta_3 = \Delta_4 = \Delta_5 = 0$.

	Distribution				
	Normal	Uniform	Double-Exp.	Exponential	Cont.-Normal
Procedure					
MEDR	0.001	0.005	0.008	0.006	0.003
MEDUR	0.003	0.013	0.025	0.014	0.026
MOODR	0.001	0.005	0.007	0.001	0.002
MOODUR	0.001	0.005	0.007	0.001	0.002
TUKEY	0.000	0.000	0.000	0.001	0.001

Table 5. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, three groups, $n_i = 5$, locations $\Delta_1 = \Delta_2 = 0, \Delta_3 = 2$.

	Distribution				
	Normal	Uniform	Double-Exp.	Exponential	Cont.-Normal
Procedure					
MEDR	0.579	0.269	0.098	0.151	0.336
MEDUR	0.487	0.256	0.095	0.113	0.297
MEDBON	0.000	0.000	0.000	0.000	0.000
MOODR	0.238	0.064	0.049	0.080	0.133
MOODUR	0.131	0.045	0.039	0.055	0.070
TUKEY	0.818	0.342	0.125	0.186	0.478

Table 6. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, three groups, $n_i = 5$, locations $\Delta_1 = 0, \Delta_2 = 2, \Delta_3 = 5$.

	Distribution				
	Normal	Uniform	D-Exp	Exponential	Cont-Normal
Procedure					
MEDR	0.786	0.707	0.262	0.410	0.455
MEDUR	0.976	0.716	0.220	0.422	0.581
MEDBON	0.000	0.000	0.000	0.000	0.000
MOODR	0.888	0.469	0.156	0.302	0.537
MOODUR	0.820	0.377	0.127	0.248	0.499
TUKEY	1.000	0.979	0.350	0.620	0.590

Table 7. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, five groups, $n_i = 5$, locations $\Delta_1 = \Delta_2 = 2; \Delta_3 = \Delta_4 = \Delta_5 = 0$.

	Distribution				
	Normal	Uniform	Double-Exp.	Exponential	Cont.-Normal
Procedure					
MEDR	0.637	0.369	0.059	0.137	0.396
MEDUR	0.400	0.293	0.078	0.104	0.245
MEDBON	0.000	0.000	0.000	0.000	0.000
MOODR	0.477	0.112	0.096	0.135	0.303
MOODUR	0.477	0.112	0.096	0.135	0.303
TUKEY	0.886	0.422	0.000	0.186	0.540

Table 8. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, three groups, $n_i = 10$, locations $\Delta_1 = 0, \Delta_2 = 2, \Delta_3 = 5$.

	Distribution				
	Normal	Uniform	Double-Exp.	Exponential	Cont.-Normal
Procedure					
MEDR	1.000	0.996	0.661	0.949	0.923
MEDUR	1.000	0.990	0.635	0.904	0.911
MEDBON	1.000	1.000	0.574	0.947	0.854
MOODR	0.888	0.469	0.156	0.302	0.537
MOODUR	0.820	0.377	0.127	0.248	0.499
TUKEY	1.000	1.000	0.627	0.890	0.940

Table 9. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, three groups, $n_i = 20$, locations $\Delta_1 = \Delta_2 = 0, \Delta_3 = 2$.

	Distribution				
	Normal	Uniform	Double-Exp.	Exponential	Cont-Normal
Procedure					
MEDR	1.000	0.664	0.374	0.664	0.991
MEDUR	1.000	0.676	0.361	0.676	0.979
MEDBON	1.000	0.776	0.342	0.776	0.983
MOODR	0.998	0.569	0.384	0.648	0.996
MOODUR	0.997	0.529	0.352	0.614	0.992
TUKEY	1.000	0.550	0.278	0.550	0.436

Table 10. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, three groups, $n_1 = 4, n_2 = 5, n_3 = 6$, locations $\Delta_1 = \Delta_3 = 0, \Delta_2 = 2$.

	Distribution				
	Normal	Uniform	Double–Exp.	Exponential	Cont.-Normal
Procedure					
MEDR	0.607	0.260	0.090	0.129	0.287
MEDUR	0.558	0.262	0.093	0.121	0.264
MEDBON	0.332	0.108	0.047	0.100	0.203
MOODR	0.147	0.041	0.060	0.070	0.125
MOODUR	0.147	0.041	0.060	0.070	0.125
TUKEY	0.220	0.035	0.005	0.012	0.051

Table 11. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, three groups, $n_1 = 4, n_2 = 5, n_3 = 6$, normally distributed data.

	Location pattern		
	$\Delta_1 = 2, \Delta_2 = \Delta_3 = 0$	$\Delta_1 = \Delta_3 = 0, \Delta_2 = 2$	$\Delta_1 = \Delta_2 = 0, \Delta_3 = 2$
Procedure			
MEDR	0.591	0.607	0.711
MEDUR	0.656	0.558	0.478
MEDBON	0.302	0.332	0.458
MOODR	0.003	0.147	0.654
MOODUR	0.003	0.147	0.654
TUKEY	0.219	0.220	0.228

Table 12. Power – Proportion of times at least one difference detected at $\alpha = 0.05$, five groups, $n_1 = 3, n_2 = 4, n_3 = 5, n_4 = 6, n_5 = 7$, normally distributed data.

	Location pattern			
	$\Delta_1 = \Delta_2 = 2;$ $\Delta_3 = \Delta_4 = \Delta_5 = 0$	$\Delta_1 = 0; \Delta_2 = \Delta_3 = 2;$ $\Delta_4 = \Delta_5 = 0$	$\Delta_1 = \Delta_2 = 0;$ $\Delta_3 = \Delta_4 = 2; \Delta_5 = 0$	$\Delta_1 = \Delta_2 = \Delta_3 = 0;$ $\Delta_4 = \Delta_5 = 2$
Procedure				
MEDR	0.546	0.451	0.556	0.702
MEDUR	0.516	0.372	0.322	0.298
MEDBON	0.003	0.000	0.041	0.002
MOODR	0.001	0.001	0.416	0.832
MOODUR	0.001	0.001	0.430	0.831
TUKEY	0.000	0.032	0.025	0.024

MOODR was less than 1.2 times that of MEDR, 0.537 versus 0.455 See Table 6). Table 8 shows, however, that when the sample size increased from $n = 5$ to $n = 10$, MOODR no longer had a power advantage over MEDR (in fact had substantially less power) for the same location pattern as in Table 6.

When sample sizes were unequal and group locations were different, the power of all tests depended on the pattern of location parameters. MOODR was by far the most affected by the pattern of differences, with virtually no power in the most extreme case (smallest samples with nonzero location parameters and largest with zero location parameters), while sometimes having the highest power with the situation reversed. In contrast, MEDR maintained respectable power for all location patterns (See Tables 11 and 12). MEDBON displayed low power when sample sizes were small, especially with five groups (10 comparisons). Power was higher with larger sample sizes, but still generally trailed the other two procedures. Many other scenarios were examined. These results are available at www.uncg.edu/~sjricht2/Research.html.

Table 13. *P*-values for pairwise comparisons.

Comparison	Median difference	Procedure				
		MEDR	MOODR	MEDUR	MOODUR	TUKEY
1vs2	85.0	0.950	1.000	0.794	0.974	0.985
1vs3	55.5	0.996	0.566	0.834	0.534	0.605
2vs3	140.5	0.691	0.295	0.645	0.345	0.372

Table 14. Average times to complete an interview for four interviewers.

Interviewer	Average time (min.)	Median	Mean
1	10.0, 25.0, 40.1, 29.2, 4.1	25.0	21.6
2	15.0, 5.2, 55.3, 15.1, 23.2	15.1	22.8
3	19.1, 25.4, 8.3	19.1	17.6
4	5.1, 9.2, 14.1	9.2	9.5

Table 15. *P*-values for pairwise comparisons.

Comparison	Median difference	Procedure				
		MEDR	MOODR	MEDUR	MOODUR	TUKEY
1vs2	9.9	0.851	1.000	0.920	1.000	0.999
1vs3	5.9	1.000	1.000	0.978	0.915	0.980
1vs4	15.8	0.211	0.450	0.525	0.362	0.666
2vs3	4.0	1.000	1.000	1.000	0.915	0.961
2vs4	5.9	1.000	0.450	0.978	0.362	0.607
3vs4	9.9	0.851	0.824	0.920	0.915	0.900

Power Advantages of Median-based Procedures

The power of the median-based procedures was compared to that of the Tukey-Kramer procedure using means. For normally distributed data and equal sample sizes, TUKEY always had higher power than the median-based procedures (See Tables 4-6). However, with unequal sample sizes, the median based procedures often had higher power even for normally distributed data (See Tables 10, 11 and 12). This may not be surprising, since the Tukey-Kramer procedure has been shown to be conservative for unequal sample sizes (Hayter, 1984). For nonnormally distributed data, the median-based procedures often had higher power, especially with larger sample sizes.

Conclusion

The maximum median difference test (MEDR) is recommended as a robust pairwise comparison procedure when strong control of FWE is desired. The maximum Mood difference test (MOODR) is not recommended, due to poor power properties, especially for unequal sample sizes. Likewise, the procedure of using separate median difference tests with a Bonferroni adjustment (MEDBON) generally had less power and no power in some cases with small sample sizes. Tukey's HSD (TUKEY) is preferred when groups have small and equal samples sizes ($n = 5$), even for nonnormal data, and also with normal data, regardless of the sample size. In all other cases, the maximum median difference test (MEDR) is preferred. With nonnormal data and large ($n \geq 20$) equal

sample sizes, and in all cases with unequal sample sizes, MEDR had higher power than TUKEY. MEDR never performed poorly with regard to power, and was often much more powerful than the other median-based procedures considered.

Example 1

The first example is based on the data in the Introduction (See Table 1.) Table 13 gives *p*-values for the three pairwise comparisons, for the MEDR, MEDUR, MOODR, MOODUR and TUKEY procedures. Notice that the Mood tests yield the most evidence for a difference between months two and three. This is an example of a scenario studied in the simulations, namely small samples with differences between all pairs, with larger differences associated with the larger samples, a case where the Mood tests often had the highest power.

Example 2:

Consider data reported by Gibbons (1985, p. 202) in Table 14. The data represent average times spent to complete an interview for four interviewers.

It is desired to test if there is evidence that certain interviewers tend to have longer interview times. Table 15 gives *p*-values for the six pairwise comparisons. Here MEDR provides the strongest evidence of location difference between the pair with the largest observed difference, interviewers 1 and 4. Resampling Stats code for calculating the permutation *p*-values in this example is provided in the Appendix.

References

Gibbons, J. D. (1985). *Nonparametric Methods for Quantitative Analysis, 2nd edition.* Columbus, OH: American Sciences Press, Inc.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Statistics*, *12*, 61–75.

Higgins, J. J. (2004). *Introduction to Modern Nonparametric Statistics*. Pacific Grove, CA: Brooks/Cole.

Hochberg Y. & Tamhane, A. C. (1987). *Multiple Comparison Procedures.* New York: John Wiley and Sons.

Keller-McNulty, S. & Higgins, J. J. (1987). Effect of tail weight and outliers on power and type-I error of robust permutation tests for location. *Communications in Statistics: Simulation and Computation*, *16*(1):17-36.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, *12*, 309–310.

Levy (1979). Pairwise Comparisons associated with the k independent sample median test. *The American Statistician*, *33*(3), 138-139.

Linton, L. R., Edgington, E. S. & Davies, R. W. (1989). A view of niche overlap amenable to statistical analysis. *Canadian Journal of Zoology*, *67*, 55-60.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology,* (2nd ed.). London: Chapman & Hall.

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

Miller, R. G. (1981). *Simultaneous statistical inference.* (2nd ed.). New York: Springer-Verlag.

Mood, A. M. (1950). *Introduction to the Theory of Statistics*. New York: McGraw-Hill.

Nemenyi, P. (1963). *Distribution-free multiple comparisons*. Unpublished doctoral dissertation, Princeton University, Princeton, NJ.

Petrondas, D. A. & Gabriel, K. R. (1983). Multiple comparisons by rerandomization tests. *Journal of the American Statistical Association*, *78*, 949-957.

Powell, G. L. & Russell, A. P. (1984). The diet of the eastern short-horned lizard (*Phrynosoma douglassi breviroste*) in Alberta and its relationship to sexual size dimorphism. *Canadian Journal of Zoology*, *62*, 428-440.

Powell, G. L. & Russell, A. P. (1985). Growth and sexual size dimorphism in Alberta populations of the eastern short-horned lizard, *Phrynosoma douglassi breviroste*. *Canadian Journal of Zoology*, *63*, 139-154.

Resampling Stats (2000). *Resampling Stats Inc*., Arlington, Virginia.

Ryan, T. A. & Ryan, T. A., Jr. (1980). *K* Independent sample median test. A letter to the Editor, *The American Statistician*, *34*, 123.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561-584.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*, 99–114.

Appendix

Below is Resampling Stats® code to calculate the permutation *p*-values in Example 2. The program can be modified to handle different numbers of groups.

```
'set maximum vector size
        maxsize default 500000
        seed 1234

'create data vectors
        data (10 25 40.1 29.2 4.1) d1
        data (15 5.2 55.3 15.1 23.2) d2
        data (19.1 25.4 8.3) d3
        data (5.1 9.2 14.1) d4

'combine data vectors for unrestricted randomization
        concat d1 d2 d3 d4 dat

'create pairwise data vectors for restricted randomization
        concat d1 d2 dat12
        concat d1 d3 dat13
        concat d1 d4 dat14
        concat d2 d3 dat23
        concat d2 d4 dat24
        concat d3 d4 dat34

'obtain permutation distribution
        let nrand=2000
        repeat nrand

'unrestricted randomization
        shuffle dat sdat
        take sdat 1,5 sdat1
        take sdat 6,10 sdat2
        take sdat 11,13 sdat3
        take sdat 14,16 sdat4

'restricted randomization
        shuffle dat12 sdat12
        take sdat12 1,5 sdat121
        take sdat12 6,10 sdat122
        shuffle dat13 sdat13
        take sdat13 1,5 sdat131
        take sdat13 6,8 sdat133
        shuffle dat14 sdat14
        take sdat14 1,5 sdat141
        take sdat14 6,8 sdat144
        shuffle dat23 sdat23
        take sdat23 1,5 sdat232
        take sdat23 6,8 sdat233
        shuffle dat24 sdat24
        take sdat24 1,5 sdat242
        take sdat24 6,8 sdat244
        shuffle dat34 sdat34
        take sdat34 1,3 sdat343
        take sdat34 4,6 sdat344

'compute medians of shuffled data
        median sdat1 med1
        median sdat2 med2
        median sdat3 med3
        median sdat4 med4
        median sdat121 med121
        median sdat122 med122
        median sdat131 med131
        median sdat133 med133
        median sdat141 med141
        median sdat144 med144
        median sdat232 med232
        median sdat233 med233
        median sdat242 med242
        median sdat244 med244
        median sdat343 med343
        median sdat344 med344

'compute median differences of shuffled data, unrestricted randomization
        subtract med1 med2 med12
        subtract med1 med3 med13
        subtract med1 med4 med14
        subtract med2 med3 med23
        subtract med2 med4 med24
        subtract med3 med4 med34

'create one vector, take absolute values
        concat med12 med13 med14 med23 med24 med34
```

```
        medvec abs medvec medvec

'compute median differences of shuffled data,
restricted randomization
        subtract med121 med122 med12r
        subtract med131 med133 med13r
        subtract med141 med144 med14r
        subtract med232 med233 med23r
        subtract med242 med244 med24r
        subtract med343 med344 med34r

'create one vector, take absolute value
        concat med12r med13r med23r medvecr
        abs medvecr medvecr

'compute maximum absolute difference
        max medvec qmedsim
        max medvecr qmedsimr

'compute Mood statistics, unrestricted
randomization
        median sdat grndmed
        count sdat1 >= grndmed m1
        count sdat2 >= grndmed m2
        count sdat3 >= grndmed m3
        count sdat4 >= grndmed m4
        median sdat12 gm12
        count sdat1 >= gm12 m121
        count sdat2 >= gm12 m122
        median sdat13 gm13
        count sdat1 >= gm13 m131
        count sdat3 >= gm13 m133
        median sdat14 gm14
        count sdat1 >= gm14 m141
        count sdat4 >= gm14 m144
        median sdat23 gm23
        count sdat2 >= gm23 m232
        count sdat3 >= gm23 m233
        median sdat24 gm24
        count sdat2 >= gm24 m242
        count sdat4 >= gm24 m244
        median sdat34 gm34
        count sdat3 >= gm34 m343
        count sdat4 >= gm34 m344
        subtract m1 m2 m12
        subtract m1 m3 m13
        subtract m1 m4 m14
        subtract m2 m3 m23
        subtract m2 m4 m24
        subtract m3 m4 m34
```

```
'Mood statistics are m12-m34

'create one vector, take absolute values
        concat m12 m13 m14 m23 m24 m34
mood
        abs mood mood

'compute maximum absolute difference
        max mood maxmood

'Compute Mood statistics, restricted
randomization
        subtract m121 m122 m12r
        subtract m131 m133 m13r
        subtract m141 m144 m14r
        subtract m232 m233 m23r
        subtract m242 m244 m24r
        subtract m343 m344 m34r

'Mood statistics are m12r-m34r

'create one vector, take absolute values
        concat m12r m13r m14r m23r m24r
m34r
        moodr abs moodr moodr

'compute maximum absolute difference
        max moodr maxmoodr

'save statistic values for reference distributions
        score qmedsim qmddist
        score qmedsimr qmddistr
        score maxmood qmood
        score maxmoodr qmoodr
    end

'compute medians and differences of observed
data
        median d1 obsmed1
        median d2 obsmed2
        median d3 obsmed3
        median d4 obsmed4

        subtract obsmed1 obsmed2 mddiff12
        abs mddiff12 mddiff12
        subtract obsmed1 obsmed3 mddiff13
        abs mddiff13 mddiff13
        subtract obsmed1 obsmed4 mddiff14
        abs mddiff14 mddiff14
        subtract obsmed2 obsmed3 mddiff23
        abs mddiff23 mddiff23
```

```
        subtract obsmed2 obsmed4 mddiff24
        abs mddiff24 mddiff24
        subtract obsmed3 obsmed4 mddiff34
        abs mddiff34 mddiff34

'compute Mood statistic for observed data
        median dat grndmed
        count d1 >= grndmed obsm1
        count d2 >= grndmed obsm2
        count d3 >= grndmed obsm3
        count d4 >= grndmed obsm4
        subtract obsm1 obsm2 obsm12
        abs obsm12 obsm12
        subtract obsm1 obsm3 obsm13
        abs obsm13 obsm13
        subtract obsm1 obsm4 obsm14
        abs obsm14 obsm14
        subtract obsm2 obsm3 obsm23
        abs obsm23 obsm23
        subtract obsm2 obsm4 obsm24
        abs obsm24 obsm24
        subtract obsm3 obsm4 obsm34
        abs obsm34 obsm34

'compute p-values
**************************************
'MEDUR
        count qmddist >=  mddiff12 mdsg12q
        divide mdsg12q nrand medur12
        count qmddist >=  mddiff13 mdsg13q
        divide mdsg13q nrand medur13
        count qmddist >=  mddiff14 mdsg14q
        divide mdsg14q nrand medur14
        count qmddist >=  mddiff23 mdsg23q
        divide mdsg23q nrand medur23
        count qmddist >=  mddiff24 mdsg24q
        divide mdsg24q nrand medur24
        count qmddist >=  mddiff34 mdsg34q
        divide mdsg34q nrand medur34
'MEDR
        count qmddistr >= mddiff12 mdsg12qr
        divide mdsg12qr nrand medr12
        count qmddistr >= mddiff13 mdsg13qr
        divide mdsg13qr nrand medr13
        count qmddistr >= mddiff14 mdsg14qr
        divide mdsg14qr nrand medr14
        count qmddistr >= mddiff23 mdsg23qr
        divide mdsg23qr nrand medr23
        count qmddistr >= mddiff24 mdsg24qr
        divide mdsg24qr nrand medr24
        count qmddistr >= mddiff34 mdsg34qr
        divide mdsg34qr nrand medr34
'MOODUR
        count qmood >= obsm12 mood12q
        divide mood12q nrand moodur12
        count qmood >= obsm13 mood13q
        divide mood13q nrand moodur13
        count qmood >= obsm14 mood14q
        divide mood14q nrand moodur14
        count qmood >= obsm23 mood23q
        divide mood23q nrand moodur23
        count qmood >= obsm24 mood24q
        divide mood24q nrand moodur24
        count qmood >= obsm34 mood34q
        divide mood34q nrand moodur34
'MOODR
        count qmoodr >= obsm12 mood12qr
        divide mood12qr nrand moodr12
        count qmoodr >= obsm13 mood13qr
        divide mood13qr nrand moodr13
        count qmoodr >= obsm14 mood14qr
        divide mood14qr nrand moodr14
        count qmoodr >= obsm23 mood23qr
        divide mood23qr nrand moodr23
        count qmoodr >= obsm24 mood24qr
        divide mood24qr nrand moodr24
        count qmoodr >= obsm34 mood34qr
        divide mood34qr nrand moodr34

**************************************
'print output here
print medur12  medur13  medur14  medur23
medur24  medur34
print medr12  medr13  medr14  medr23
medr24  medr34
print moodur12 moodur13 moodur14 moodur23
moodur24 moodur34
print moodr12  moodr13  moodr14  moodr23
moodr24  moodr34
```