

## Step-Down Multiple Comparison Procedures Using Medians and Permutation Tests

By: [Scott J. Richter](#), Melinda H. McCann

Richter, S. J. and McCann, M. H. (2009). Step-down Multiple Comparison Procedures based on Medians and Permutation Tests. *Communications in Statistics: Simulation and Computation*, 38 (8), 1551-1561. doi: 10.1080/03610910903039465

This is an Accepted Manuscript of an article published by Taylor & Francis Group in *Communications in Statistics: Simulation and Computation* on 07 July 2009, available online at: <http://www.tandfonline.com/10.1080/03610910903039465>

\*\*\*© Taylor & Francis. Reprinted with permission. No further reproduction is authorized without written permission from Taylor & Francis. This version of the document is not the version of record. Figures and/or pictures may be missing from this format of the document. \*\*\*

### Abstract:

Richter and McCann (2007) presented a median-based multiple comparison procedure for assessing evidence of group location differences. The sampling distribution was based on the permutation distribution of the maximum median difference among all pairs, and provides strong control of the FWE. This idea is extended to develop a step-down procedure for comparing group locations. The new step-down procedure exploits logical dependencies between pairwise hypotheses and provides greater power than the single-step procedure, while still maintaining strong FWE control. The new procedure can also be a more powerful alternative to existing methods based on means, especially for heavy-tailed distributions.

**Keywords:** Median difference | Pairwise comparisons | Permutation test | Step-down procedure | 62G09

### Article:

#### 1. Introduction

Richter and McCann (2007) presented a single-step multiple pairwise testing procedure, using median differences, for assessing evidence of group location differences. The sampling distribution was based on the permutation distribution of the maximum median difference among all pairs, and provides strong control of the familywise error rate (FWE), which is defined here as the probability of at least one false rejection among all true hypotheses. This article develops a more powerful step-down procedure utilizing the permutation distribution considered by Richter and McCann (2007) and incorporating logical constraints between hypotheses.

Holm (1979) introduced a sequential (step-down) procedure applicable to multiple pairwise comparisons, and showed it to be more powerful than using a simple Bonferroni adjustment. It is

based on the idea that after a particular hypothesis has been rejected, it is not necessary to adjust the significance level for that comparison in subsequent steps. Shaffer (1986) later showed that Holm's method may be modified to give more powerful tests by incorporating logical dependencies between hypotheses. Holland and Copenhaver (1987) suggested a slight improvement on Shaffer's method under a certain type of dependence of the test statistics. Westfall (1997) noted further improvements in power are possible by also incorporating the dependencies between the means themselves. He presented an approach based on resampling the  $p$ -value distribution, which he showed can be more powerful than Shaffer's method. Westfall and Tobias (2007) further improved Shaffer's method by developing a closed testing procedure based on resampling the  $p$ -value distribution.

Other authors have considered nonparametric permutation tests to address multiplicity issues. Finos and Salmaso (2005) considered a nonparametric permutation approach to adjust for multiplicity based on optimal subsets. Their method, however, provides only weak FWE control. Finos and Salmaso (2006) developed weighted nonparametric permutation methods which strongly control the FWE. However, these methods are designed to encompass situations where certain comparisons are considered more important *a priori*, and specifically focus on cases where the number of variables is much higher than the number of observations. Finos and Salmaso (2007) considered weighted nonparametric permutation methods that in some situations allow the researcher to select weights *a posteriori* without sacrificing multiplicity control. Their procedure is developed for multivariate data with dependencies where the tests that are adjusted for multiplicity each involve only one variable.

For the specific case of all pairwise comparisons, Miller (1981) proposed a permutation multiple pairwise testing procedure based on the range of the sample means. However, Petrondas and Gabriel (1983) showed that the FWE may not be controlled when randomization is over all groups, including those not in the subset being tested. Nemenyi (1963) proposed a similar procedure based on the maximum of differences of Mood's (1950) two-sample median statistics. Hochberg and Tamhane (1987) pointed out that Nemenyi's procedure is not based on a testing family, and suggested instead that the test be based on the maximum of the separate median statistics for each pairwise comparison. Richter and McCann (2007) proposed using median differences instead of differences of Mood statistics, with the sampling distribution based on the permutation distribution of the maximum median difference among all pairs, and randomization restricted to within each pair. Richter and McCann (2007) showed that their method provided strong control of the FWE, and that using median differences is preferred over using differences of Mood statistics.

This article extends the procedure of Richter and McCann (2007) and presents step-down pairwise testing procedures for comparing group locations that incorporate logical dependencies between hypotheses. Simulation results suggest that the step-down procedures can be more powerful alternatives to existing analogous normal theory methods, especially for heavy-tailed distributions.

## 2. Methodology

Throughout the remainder, 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  $N = n_1 + n_2 + \dots + n_k$ . Further, let  $\mu_i$  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. That is, for  $i, j = 1, 2, \dots, k$  with  $i < j$ ,  $F_i(x) = F_j(x - \Delta_{ij})$ , where  $\Delta_{ij}$  represents the location difference between groups  $i$  and  $j$ . The reference distribution for testing the single comparison,  $H_{0(ij)}: \Delta_{ij} = 0$ , is

based on the distribution of  $\max_{1 \leq i < j \leq k} |\hat{\mu}_i - \hat{\mu}_j|$ , the maximum of all pairwise median differences, calculated for a large set of random reassignments of observations to groups. It should be noted that although we consider only the special case of a location shift, the proposed methods are valid also in the more general case where distributions are allowed to differ in shape under the alternative hypothesis. In the general case, the hypotheses are  $H_{0(ij)}: F_i(x) = F_j(x)$ ,  $i, j = 1, 2, \dots, k$  with  $i < j$ ,  $H_{a(ij)}: F_i(x) \leq F_j(x)$  (or  $H_{a(ij)}: F_i(x) \geq F_j(x)$ .)

Let  $C_j$  denote the number of possible true hypotheses at Step  $j$ . In order to ensure strong control of the FWE, rather than randomize over all groups remaining at each step, randomization is done separately within each pair, and the maximum absolute difference over all pairs obtained.

### 2.1. Maximal Subsets Procedure

**Step 1.** Compute the permutation reference distribution using all  $C_1 = \binom{k}{2}$  pairs. For each pairwise median difference, compute a  $p$ -value using the permutation reference distribution. Denote the ordered  $p$ -values as  $p_{11}, p_{12}, \dots, p_{1C_1}$ . (i) If all  $p$ -values are greater than the specified  $\alpha$ -level, then no group location differences can be declared and the procedure stops. (ii) If  $p_{11} \leq \alpha$ , declare the locations associated with the corresponding groups different, and proceed to Step 2.

**Step 2.** Consider the pairwise hypothesis associated with  $p_{12}$  from Step 1. The maximum number of possible true hypotheses, given that the hypothesis rejected in Step 1 cannot be true, is

now  $C_2 = \binom{k-1}{2}$ . Thus, compute a reference distribution for each set,  $S_{i2}$ ,

of  $C_2 = \binom{k-1}{2}$  pairs containing the pairwise hypothesis associated with  $p_{12}$ , and a

corresponding  $p$ -value,  $p_{2S_{i2}}$ . Then set  $p_2 = \max(p_{2S_{i2}^*}, p_{11})$ ,

where  $p_{2S_{i2}^*} = \max_i(p_{2S_{i2}})$  and  $\max(p_{2S_{i2}^*}, p_{11})$  is used to maintain monotonicity.

If  $p_2 > \alpha$ , no further group location differences can be declared and the procedure stops.  
 If  $p_2 \leq \alpha$ , declare the locations associated with the corresponding groups different, and proceed to Step 3. Note that in Step 2, permutation is over sets of size  $\binom{k-1}{2}$  rather than  $\binom{k}{2} - 1$ , as would be the case using Holm's method, which does not incorporate logical dependencies.

**Step  $j$ .** Consider the pairwise hypothesis associated with  $p_{1j}$  from Step 1. Compute a reference distribution for each set,  $S_{ij}$ , of  $C_j$  pairs containing this pair, where again  $C_j$  is the maximum number of possible true hypotheses, given the hypotheses rejected in Step 1 through  $j$  cannot be true. Compute a  $p$ -value,  $p_{jS_{ij}}$ , using each reference distribution, and

set  $p_j = \max(p_{jS_j^*}, p_{j-1})$ , where  $p_{jS_j^*} = \max_i(p_{jS_{ij}})$  is the maximum  $p$ -value calculated at Step  $j$ . If  $p_j > \alpha$ , no further group location differences can be declared and the procedure stops. If  $p_j \leq \alpha$ , declare the locations associated with the corresponding groups different, and proceed to Step  $j + 1$ .

The following theorem shows that this procedure provides strong control of the FWE.

### Theorem 2.1

The permutation maximal subsets procedure using the maximum absolute median difference as the test statistic has familywise error rate no larger than  $\alpha$ .

### Proof

Our procedure is performed and described in terms of  $p$ -values for ease in computation and explanation, but could equivalently be carried out in terms of test statistics. The proof is more transparent in terms of test statistics, and thus this formulation of the procedure will be utilized here.

Let  $T_i$  be the absolute value of the median difference for the  $i$ th hypothesis,  $i = 1, 2, \dots, p$ , and  $T_i^*$  be the  $i$ th largest order statistic of this set of median differences. Assume that some set  $I$  of  $m$  hypotheses is actually true. Let  $d_{\alpha, m}$  be the permutation distribution percentile with an area to the right of  $\alpha$  when comparing a set of  $m$  hypotheses. In situations where the set of hypotheses that can be true in any particular application can be any subset of the indexes of the  $p$  hypotheses, then we have a situation of "free combinations". In our situation, however, this is not true.

At each stage  $j$ ,  $j = 1, 2, \dots, p$  of our procedure, consider the set of remaining hypotheses to be possibly rejected as  $J_j$  and let  $m_j$  be the maximum number of hypotheses among the set  $J_j$  that could possibly be true. Notice that when  $I = J_j$ , then  $m_j = m$  as obviously all of these true hypotheses can be true simultaneously. Our procedure will begin with the largest absolute median difference and reject at each stage if the largest remaining absolute median difference is

greater than the permutation critical point when the number of hypotheses to compare is equal to  $m_j$ .

Let  $j^*$  be such that  $\min \{m_j \mid m_j \geq m\} = m_{j^*}$ . Now  $d_{\alpha, m}$  is nondecreasing in  $m$  since for  $m_1 < m_2$  we have that

$$P(\max\{T_1, T_2, \dots, T_{m_1}\} \geq d_{\alpha, m_2}) \leq P(\max\{T_1, T_2, \dots, T_{m_2}\} \geq d_{\alpha, m_2}) = \alpha,$$

which implies that  $d_{\alpha, m_1} \leq d_{\alpha, m_2}$ . As  $m_j$  is non increasing in  $j$ , we have that  $d_{\alpha, m_{j+1}} \leq d_{\alpha, m_j}$ . Thus,  $d_{\alpha, m} \leq d_{\alpha, m_{j^*}}$ . Now consider the case where  $T_i < d_{\alpha, m} \forall i \in I$ .

Then  $T_{p-m+1}^* < d_{\alpha, m} \leq d_{\alpha, m_{j^*}}$  and the test will terminate in step  $p - m + 1$  or earlier. But then obviously we will also have failed to reject the set of true hypotheses.

As  $P(T_i < d_{\alpha, m} \forall i \in I) = 1 - \alpha$ , the proof is complete.

The maximum possible numbers of true hypotheses at each step are given for up to ten groups by Holland and Copenhaver (1987) and Shaffer (1986) provided a recursive formula to calculate these for any number of groups. The following example illustrates the maximal subsets procedure.

## 2.2. Example

Consider the hypothetical data for four groups in Table 1. Suppose the interest is in determining if there are group location differences. The maximal subset procedure can be utilized to address this question.

**Table 1.** Data for four groups, with median and mean for each group

Group	Observations	Median	Mean
1	11, 13, 14, 33, 84	14	31
2	19, 21, 107, 108, 184	107	87.8
3	1, 1, 8, 9, 33, 39, 65	9	22.3
4	1, 15, 16, 16, 26, 56, 100	16	32.9

**Step 1.** Perform the single step procedure where each pairwise difference is compared to the permutation distribution of the maximum difference among all  $C_1 = 6$  pairs. The resulting  $p$ -values, based on 10,000 randomizations, are given in Table 2.

**Table 2.** Pairwise hypotheses and Step 1  $p$ -values

Null hypothesis	Step 1 $p$ -value
H 0(12): $\Delta_{12} = 0$	0.059 = $p_{12}$
H 0(13): $\Delta_{13} = 0$	1.000 = $p_{16}$

H <sub>0(14)</sub> : $\Delta_{14} = 0$	1.000 = p <sub>15</sub>
H <sub>0(23)</sub> : $\Delta_{23} = 0$	0.012 = p <sub>11</sub>
H <sub>0(24)</sub> : $\Delta_{24} = 0$	0.072 = p <sub>13</sub>
H <sub>0(34)</sub> : $\Delta_{34} = 0$	1.000 = p <sub>14</sub>

Since  $p_{11} = 0.012 < 0.05$ , we would reject  $H_{0(23)}$ :  $\Delta_{23} = 0$  and declare the locations of groups 2 and 3 statistically different and proceed to Step 2.

**Step 2.** We are now interested in testing  $H_{0(12)}$ :  $\Delta_{12} = 0$ , corresponding to  $p_{12} = 0.059$ . There is now a maximum of  $C_2 = 3$  hypotheses that can be simultaneously true. However, only one set of three hypotheses,  $\Delta_{12} = \Delta_{14} = \Delta_{24} = 0$ , does not imply  $\Delta_{23} = 0$ . Thus, the single-step procedure is performed, where the randomization is done within each of the three above pairs, and the maximum difference over the three pairs is used to determine the reference distribution. The corresponding  $p$ -value, based on 10,000 randomizations, is 0.047. Thus, the  $p$ -value for testing  $H_{0(12)}$ :  $\Delta_{12} = 0$  is  $\max(0.047, 0.012) = 0.047$ , and the locations of groups 1 and 2 are declared different, and we proceed to Step 3.

**Step 3.** We are now interested in testing  $H_{0(24)}$ :  $\Delta_{24} = 0$ , corresponding to  $p_{13} = 0.072$ . There is still a maximum of  $C_3 = 3$  hypotheses that can be simultaneously true. However, all remaining sets of three hypotheses containing  $H_{0(24)}$ :  $\Delta_{24} = 0$  imply either  $\Delta_{12} = 0$  or  $\Delta_{23} = 0$ . In fact, there is only one set of two hypotheses, (2,4), (1,3), for which  $\Delta_{24} = 0$  that can be simultaneously true. The corresponding  $p$ -value is 0.014. Then  $p_3 = \max(0.014, 0.047) = 0.047 < 0.05$ , and we declare groups 2 and 4 different and proceed to Step 4.

**Step 4.** We are now interested in testing  $H_{0(34)}$ :  $\Delta_{34} = 0$ , corresponding to  $p_{14} = 1.000$  (Since three hypotheses had  $p$ -values of 1.000, the tiebreaker used was the largest observed median difference). There is still a maximum of  $C_4 = 3$  hypotheses that can be simultaneously true. Now only  $\Delta_{13} = \Delta_{14} = 0$  can additionally be assumed. The  $p$ -value for the test of  $H_{0(34)}$ :  $\Delta_{34} = 0$  is 0.928, so  $p_4 = \max(0.928, 0.047) = 0.928 > 0.05$ ; thus, we cannot declare groups 3 and 4 different at  $\alpha = 0.05$ , and the procedure stops.

For completeness, we also report the  $p$ -values associated with the tests for the remaining hypotheses, as well as the  $p$ -values computed using Westfall's (1997) procedure comparing means, in Table 3.

**Table 3.** Pairwise hypotheses and  $p$ -values for permutation step-down and Westfall's procedures

Null hypothesis	$p$ -value	Westfall
H <sub>0(12)</sub> : $\Delta_{12} = 0$	0.047	0.054
H <sub>0(13)</sub> : $\Delta_{13} = 0$	0.939	0.899
H <sub>0(14)</sub> : $\Delta_{14} = 0$	0.966	0.849
H <sub>0(23)</sub> : $\Delta_{23} = 0$	0.012	0.048
H <sub>0(24)</sub> : $\Delta_{24} = 0$	0.047	0.061

H <sub>0</sub> (34): Δ <sub>34</sub> = 0	0.928	0.849
------------------------------------------	-------	-------

The result of the procedure, with familywise  $\alpha = 0.05$ , is that we have statistical evidence that group 2 has different location from groups 1, 3, and 4, but not that the locations of any other pairs differ. Note that the single-step median and Westfall's procedure declares only that group 2 differs in location from groups 1 and 3, but not group 4.

### 2.3. Conservative Maximal Subsets Procedure

Implementing the above procedure requires determining the possible true hypotheses at each step. This is possible, but can be a difficult and computationally intensive task when the number of groups is not small. For example, Proc Glimmix in SAS will determine the logical alternatives for pairwise testing of means, but this becomes computationally infeasible for more than ten groups (SAS Institute, 2005). A conservative alternative is to simply form all possible sets of size  $C_j$ , the maximum number of subsets that can be simultaneously true at Step  $j$ , including sets containing pairs declared different at a previous step. These sets must result in a  $p$ -value at least as large as sets containing only non rejected pairs. This is because previously rejected pairs will contribute differences to the reference distribution that are larger than those contributed by the non rejected pairs, thus creating more values at least as large as the one under consideration than would be the case for sets containing only non rejected pairs. Thus, the  $p$ -value calculated,  $p_j$ , will be conservative.

To illustrate the conservative maximal subsets procedure, we return to Example 2. For each of Steps 2–4, there are a maximum of  $C_j = 3$  pairs that can be simultaneously true, and for each

hypothesis, there is  $\binom{5}{2} = 10$  possible sets of three hypotheses. Thus, the  $p$ -value for testing a hypothesis in these steps will be, for each hypothesis, the maximum  $p$ -value across all ten possible sets of three including that pair. At Step 2, we wish to test  $H_{0(12)}: \Delta_{12} = 0$ , and thus consider tests of this hypothesis for all sets of three containing the pair (1,2). The  $p$ -values for all possible permutation tests are given in Table 4.

**Table 4.**  $p$ -values for the test of  $H_{0(12)}: \Delta_{12} = 0$  for all possible subsets of size  $C = 3$

Other parameters assumed equal to 0	p-value
$\Delta_{13} = \Delta_{14} = 0$	0.047
$\Delta_{13} = \Delta_{23} = 0$	0.059
$\Delta_{13} = \Delta_{24} = 0$	0.047
$\Delta_{13} = \Delta_{34} = 0$	0.047
$\Delta_{14} = \Delta_{23} = 0$	0.047
$\Delta_{14} = \Delta_{24} = 0$	0.047
$\Delta_{14} = \Delta_{34} = 0$	0.047
$\Delta_{23} = \Delta_{24} = 0$	0.059
$\Delta_{23} = \Delta_{34} = 0$	0.059

$\Delta_{24} = \Delta_{34} = 0$	0.047
---------------------------------	-------

The maximum  $p$ -value is 0.059, and thus the  $p$ -value for testing  $H_{0(12)}: \Delta_{12} = 0$  is  $\max(0.059, 0.012) = 0.059$ , and locations of groups 1 and 2 cannot be declared different and the procedures stops.

A summary of the procedure for the more conservative method compared to the method based on the actual hypotheses is given in Table 5.

**Table 5.** Pairwise hypotheses and  $p$ -values for maximal subsets and conservative maximal subsets procedures

Null hypothesis	$p$ -value—maximal subsets	$p$ -value—conservative maximal subsets
H 0(12): $\Delta_{12} = 0$	0.047	0.059
H 0(13): $\Delta_{13} = 0$	0.939	1.000
H 0(14): $\Delta_{14} = 0$	0.966	1.000
H 0(23): $\Delta_{23} = 0$	0.012	0.012
H 0(24): $\Delta_{24} = 0$	0.047	0.072
H 0(34): $\Delta_{34} = 0$	0.928	1.000

Note that all  $p$ -values are slightly higher for the more conservative procedure. Thus, the power is slightly less, and for this example the overall conclusions reached by the two procedures are not the same, as only groups 2 and 3 can be declared different.

## 2.4. Two-Step Procedure

The conservative maximal subset procedure, while eliminating the need to determine the specific subsets that can be simultaneously equal, requires potentially very many permutation tests at each step, and the procedure may also become computationally infeasible if there are a large number of groups. Shaffer (1986) observed that when incorporating logical restrictions between hypotheses, the greatest single improvement in power occurs between Steps 1 and 2, and suggested that a conservative short-cut to the full step-down method would be to use the number of possible true hypotheses at Step 2, not only for Step 2, but also for all remaining steps. This suggestion can also be used as a conservative short-cut for the proposed permutation procedure,

by using  $C_2 = \binom{k-1}{2}$  for Steps 2 through  $\binom{k}{2}$ . Since when  $k = 4$ ,  $C_2 = C_3 = C_4 = 3$ , in Example 2 the two-step procedure would be identical to the maximal subset procedure, and result in the same conclusions. Note that the two-step procedure may also be based on the conservative approach, where the pairs rejected at previous steps are included in the sets over which the maximum absolute difference is computed. Simulation results in the next section support that there is both a substantial gain in power of the two-step over the single-step method, and little loss of power compared to the maximal subset procedure.

### 3. Simulation

#### 3.1. Simulation Details

A small simulation study was performed to assess the power advantage of the step-down procedure over the single-step method, as well as to compare the power of the two-step method to the maximal subset procedure (Note: we consider here only the conservative maximal subset procedure, due to computational complexity of the maximal subset procedure). Power was also calculated for the normal-based means comparison procedure of Westfall (1997), to determine if there are situations where the median-based method may be more powerful than an analogous method using means.

Four groups were considered, and the additive model

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, 2, \dots, k, \quad j = 1, \dots, n_i,$$

employed, where  $\mu_i$  is the location parameter associated with the  $i$ th population. Several different distributions were considered for  $\varepsilon_{ij}$ . Normal, Laplace, and Cauchy distributions were chosen to represent symmetric distributions with progressively heavier tails, and, similarly, exponential and lognormal ( $\sigma = 1.5$ ) represented light and heavy-tailed skewed distributions, respectively. While methods based on means (such as Westfall's procedure) are not consistent for the Cauchy distribution, we make this comparison to get a sense of the maximum power advantage of the median-based methods.

The single step procedure (*1-step*) as well as four variations of the step-down procedure were examined: the full maximal subsets procedure (*full*); the full step-down procedure using conservative maximal subsets (*c-full*); the two-step procedure (*2-step*); and the two-step procedure using conservative maximal subsets (*c2-step*).

All permutation tests were based on 10,000 random permutations, and for each case considered, average and all-pairs power and FWE estimates were based on 1,000 randomly generated samples.

#### 3.2. Simulation Results

The results of the simulation suggest that the step-down procedures have substantially higher average, or per-pair, power than the one-step procedure, especially for the very heavy-tailed Cauchy and lognormal distributions (see Tables 6-8). In addition, there seems little to be gained by using the full step-down procedure over the two-step method, particularly when comparing the conservative procedures. In fact, the two-step procedure (*2-step*) often had the same or better power than the conservative full step-down procedure (*c-full*), although the two-step procedure (*2-step*) was not substantially better than the conservative two-step (*c2-step*). The median-based procedures also showed higher power than Westfall's method for very heavy-tailed (Cauchy, lognormal) distributions. As would be expected, the median-based procedures had lower power

for lighter-tailed (normal, exponential, Laplace) distributions. With respect to all-pairs power, the median-based procedures performed best when all location differences were the same size (see Tables 6 and 7), but had virtually no all-pairs power when there were several magnitudes of location differences present (see Table 9). Westfall's procedure had much higher all-pairs power for the lighter-tailed distributions, but virtually none for the heavier-tailed distributions.

**Table 6.** Estimated average power, based on 1,000 random data sets. Location configuration:  $\Delta_1 = 3, \Delta_2 = \Delta_3 = \Delta_4 = 0$

Method	Error distribution				
	Normal	Laplace	Cauchy	Exponential	Lognormal
Perm 1-step	0.996	0.987	0.732	0.962	0.414
Perm c2-step	0.997	0.993	0.818	0.980	0.485
Perm 2-step	0.998	0.993	0.834	0.982	0.494
Perm c-full	0.997	0.993	0.818	0.980	0.485
Perm full	0.998	0.993	0.834	0.982	0.494
Westfall	1.000	1.000	0.130	1.000	0.261

**Table 7.** Estimated all-pairs power, based on 1,000 random data sets. Location configuration:  $\Delta_1 = 3, \Delta_2 = \Delta_3 = \Delta_4 = 0$

Method	Error distribution				
	Normal	Laplace	Cauchy	Exponential	Lognormal
Perm 1-step	0.991	0.965	0.497	0.895	0.184
Perm c2-step	0.994	0.981	0.700	0.946	0.263
Perm 2-step	0.996	0.983	0.656	0.952	0.283
Perm c-full	0.994	0.981	0.656	0.946	0.263
Perm full	0.996	0.983	0.700	0.952	0.283
Westfall	1.000	1.000	0.074	1.000	0.176

**Table 8.** Estimated average power, based on 1,000 random data sets. Location configuration:  $\Delta_1 = 3, \Delta_2 = 1, \Delta_3 = \Delta_4 = 0$

Method	Error distribution				
	Normal	Laplace	Cauchy	Exponential	Lognormal
Perm 1-step	0.537	0.492	0.330	0.467	0.260
Perm c2-step	0.560	0.525	0.374	0.493	0.299
Perm 2-step	0.569	0.530	0.378	0.496	0.302
Perm cfull	0.560	0.525	0.374	0.493	0.299
Perm full	0.577	0.535	0.378	0.496	0.304
Westfall	0.915	0.774	0.043	0.916	0.153

**Table 9.** Estimated all pairs power, based on 1,000 random data sets. Location configuration:  $\Delta_1 = 3, \Delta_2 = 1, \Delta_3 = \Delta_4 = 0$

Method	Error distribution				
	Normal	Laplace	Cauchy	Exponential	Lognormal
Perm 1-step	0	0	0	0	0
Perm c2-step	0.001	0.001	0	0	0
Perm 2-step	0.017	0.011	0	0.002	0.002
Perm cfull	0.001	0.002	0	0	0
Perm full	0.04	0.022	0	0.003	0.003
Westfall	0.733	0.333	0.006	0.721	0

#### 4. Discussion

The example in Sec. 2 and the simulation results of Sec. 3 suggest that the step-down procedures using medians can substantially improve the power to detect pairwise differences in location, compared to the single step median procedure. In addition, using medians can be a more powerful alternative to methods based on means, especially for heavy-tailed distributions. The results can depend upon which characterization of power is used, however, with the median-based procedure generally having good average power but low all-pairs power, especially when median differences of different magnitudes were present. This suggests that the median-based methods will be better at detecting larger location differences for heavy-tailed distributions, but may not be as useful for detecting smaller location differences when larger differences are also present.

Only Westfall's (1997) method was considered as a representative procedure for comparing means of normal distributions. Thus, results may vary for different step-down procedures for comparing means. Since Westfall's procedure exploits both dependencies between means and logical relationships between hypotheses, and has been shown to be more powerful than several competitors (Westfall, 1997), it seems reasonable to assume that it is one of the most powerful step-down procedures for comparing means, and a good benchmark against which to compare the median-based procedures.

In addition, only a relatively small number of groups was considered (four) in the simulations. It would be expected that the power of the permutation methods may increase as the number of groups increase, as with more pairs in a subset, the larger median differences would have less influence on the distribution of the maximum differences across pairs. It has been shown that for two groups, rank-based methods have increased power advantages over the  $t$ -test with larger sample sizes (e.g., Blair and Higgins, 1980), and that there is little difference in power between rank and median-based methods (Keller-McNulty and Higgins, 1987). Thus, it might be expected that a similar relationship exists between the median-based pairwise comparison procedures and analogous procedures for means of normal distributions. However, it is still likely that the most

substantial gain in power will occur at the second step, and given the computational complexity of the full step-down procedure, the two-step procedure may serve as a viable, and only slightly less powerful alternative.

## References

1. Blair , R. C. , Higgins , J. J. ( 1980 ). A comparison of the power of Wilcoxon'S rank-sum test to that of the student's *T*statistic under various nonnormal distributions . *Journal of Educational Statistics* 5 ( 4 ): 309 – 335 .
2. Einot , I. , Gabriel , K. R. ( 1975 ). A study of the powers of several methods of multiple comparisons . *Journal of the American Statistical Association* 70 ( 351 ): 574 – 583 .
3. Finos , L. , Salmaso , L. ( 2005 ). A new nonparametric approach for multiplicity control: Optimal subset procedures . *Computational Statistics* 20 ( 4 ): 643 – 654 .
4. Finos , L. , Salmaso , L. ( 2006 ). Weighted methods controlling the multiplicity when the number of variables is much higher than the number of observations . *Journal of Nonparametric Statistics* 18 ( 2 ): 245 – 261 .
5. Finos , L. , Salmaso , L. ( 2007 ). FDR- and FWE-controlling methods using data-driven weights . *Journal of Statistical Planning and Inference* 137 : 3859 – 3870 .
6. Hochberg , Y. , Tamhane , A. C. ( 1987 ). *Multiple Comparison Procedures* . New York : John Wiley and Sons .
7. Holland , B. S. , Copenhaver , M. D. ( 1987 ). An improved sequentially rejective test procedure . *Biometrics* 43 : 417 – 423 .
8. Holm , S. ( 1979 ). A simple sequentially rejective multiple test procedure . *Scandinavian Journal of Statistics* 6 : 65 – 70 .
9. 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 . *Communication in Statistics: Simulation and Computation* 16 ( 1 ): 17 – 36 .
10. Mood , A. M. ( 1950 ). *Introduction to the Theory of Statistics* . New York : McGraw-Hill .
11. Nemenyi ( 1963 ). Distribution-Free Multiple Comparisons. Unpublished Doctoral Dissertation . Princeton University , Princeton , NJ .
12. Petrondas , D. A. , Gabriel , K. R. ( 1983 ). Multiple comparisons by rerandomization tests . *Journal of the American Statistical Association* 78 ( 384 ): 949 – 957 .
13. Richter , S. J. , McCann , M. H. ( 2007 ). Multiple comparisons using medians and permutation tests . *Journal of Modern Applied Statistical Methods* 6 ( 2 ): 399 – 412 .

- 14.** SAS Institute ( 2005 ). The GLIMMIX Procedure, Nov. 2005. SAS Institute, Inc., Cary NC .
- 15.** Shaffer , J. P. ( 1986 ). Modified sequentially rejective multiple test procedures . *Journal of the American Statistical Association* 81 ( 395 ): 826 – 831 .
- 16.** Westfall , P. H. ( 1997 ). Multiple testing of general contrasts using logical constraints and correlations . *Journal of the American Statistical Association* 92 : 299 – 306 .
- 17.** Westfall , P. H. , Young , S. S. ( 1993 ). *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*. New York : John Wiley & Sons, Inc.
- 18.** Westfall , P. H. , Tobias , R. D. ( 2007 ). Multiple testing of general contrasts: truncated closure and the extended Shaffer-Royen method . *Journal of the American Statistical Association* 102 : 487 – 494 .