Simultaneous multiple comparisons with a control using median differences and permutation tests

By: Scott J. Richter, Melinda H. McCann

Richter, S. J. and McCann, M. H. (2013). Simultaneous Multiple Comparisons with a Control Using Medians and Permutation Tests. *Statistics and Probability Letters*, 83(4) 1167-1173. doi:10.1016/j.spl.2013.01.014

Abstract:

Permutation methods using median differences for simultaneous pairwise comparisons with a control are investigated. Simulation results suggest that the permutation methods are generally more powerful than the Dunnett procedure when data are from nonnormal distributions. A new procedure is shown to provide strong control of the familywise error rate, and have highest power for detecting the treatment that differs most from the control, for certain nonnormal distributions. Step-down permutation procedures, which have greater power to detect treatment differences with the control, are also proposed and examined. The procedures are illustrated using an example from the applied literature.

Keywords: Nonparametric simultaneous inference | Multiple comparisons with a control | Median difference | Permutation test

Article:

1. Introduction

Many applied research studies involve comparing two or more treatments to a control treatment. In cases where the data are skewed or contain extreme outliers, medians may be more appropriate than means for describing central tendency of the distributions, and may be more effective for detecting location differences.

In this paper, permutation methods for comparing all treatments to a control using median differences are presented and compared.

2. Methods for comparing medians

Much of the recent research on using medians to compare distributions has focused on approximating the asymptotic distribution of the sample median difference. However, these methods require an estimator of the asymptotic standard error of the sample median difference, and can have Type I error rates that are either much lower or higher than the nominal level (Wilcox, 2006). In addition, among these methods none are guaranteed to control the familywise error rate (FWER), the probability of making at least one false declaration of inequality. Alternatively, permutation methods can be used to determine exact reference distributions for comparing medians. Richter and McCann (2007) proposed a permutation procedure, using median differences and restricted randomization, for testing all pairwise comparisons. They showed that restricted randomization guarantees strong FWER control, and that the permutation procedure can have greater power for heavier-tailed distributions than the Tukey‑Kramer (Tukey, 1949 and Kramer, 1956) testing procedure.

In this paper, the permutation method of Richter and McCann (2007) is extended to the case of multiple pairwise comparisons with a control (MCC). The method is shown to control the FWER, and a simulation study is used to investigate power properties of the procedure, for various distributions and sample size configurations. The new procedure is compared to the method of using separate two-group median-difference permutation tests employing a Bonferroni correction, and also to Dunnett's (1955) procedure, the optimal procedure for making all pairwise comparisons to a control for means of normal distributions with equal variances.

3. Methodology

3.1. Notation and assumptions

Consider a one-way layout with $c-1$ independent treatment groups plus a control group, where F_i is the common continuous distribution function for the ith group, $i=1,\ldots,c-1,c$, where $i=c$ represents the control group, n_i is the sample size of the ith group, and $N=n_1+n_2+\cdots+n_{c-1}+n_c$. Further, let μ_i be the location parameter associated with the ith distribution and $\hat{\mu}_i$ be the sample median for the ith group. Distributions are assumed identical for all treatments except for possible location differences.

3.2. Permutation-based median difference procedure (MED)

To compute a p-value to assess each pairwise hypothesis, the absolute observed median difference for each pair will be compared to the permutation distribution of the statistic, $\max_{1\le i\le c-1}|\hat{\mu}_c-\hat{\mu}_i|$, the maximum of all pairwise median differences. This statistic will be calculated for a large number of random reassignments of observations to groups, where the

maximum is computed over the c−1treatment/control pairs, and where randomization is performed separately within each pair. This will be referred to as the *MED* procedure.

3.3. Restricted randomization provides strong FWER control

Strong FWER control for the MCC case can be established as follows. Consider c−1 independent samples from distributions that differ from a control distribution by at most a location parameter. That is, forj=1,…,c−1,$F_c(x)=F_j(x-\Delta_j)$. The null hypothesis then involves c−1 pairwise hypotheses of the form, $H_{0j}:\Delta_j=0$. Now consider the permutation distribution of median differences from samples c and j, and let $D_j(\alpha)$ be the 1−α percentile of this permutation distribution. Similarly, define $D_{\max}(\alpha)$ to be the1−α percentile of the permutation distribution for the maximum median difference among all c−1 pairs of a treatment with the control.

First consider the case under the complete null hypothesis where all $\Delta_j=0$. Let the calculated median difference from samples c and j be denoted by $\tilde{D}_j$. 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 1/(c−1). Consequently, the probability that any observed difference from a particular pair exceeds $D_{\max}(\alpha)$, the comparisonwise error rate, is α/(c−1). Alternatively, the familywise error rate is given by

$$P(\text{declare at least one location different from control} \mid \text{all pairs have equal location}) \leq \sum_{j=1,\ldots,c-1} P(\tilde{D}_j \geq D_{\max}(\alpha)) = (c-1)(\alpha/(c-1)) = \alpha$$

. This shows that using the permutation distribution of the maximum difference controls the FWER in the*weak* sense (Hochberg and Tamhane, 1987). Now consider the case where only t<(c−1) 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 (c−1)−t pairs where $\Delta_j\neq 0$. Consequently, the comparisonwise error rate is $P(\tilde{D}_j \geq D_{\max}(\alpha)) \leq \alpha/(c-1)$. The familywise error rate, the probability of rejecting at least one of the t true null hypotheses, is $P(\text{reject at least one true null hypothesis} \mid t \text{ true null hypotheses}) \leq t(\alpha/(c-1)) < \alpha$. Thus, the FWER is controlled at level α for any combination of t true and (c−1)−t false hypotheses, and the FWER is controlled in the *strong* sense (Hochberg and Tamhane, 1987). This proves strong FWER control using median differences, but a similar argument can be used for any other statistic of interest.

4. Increasing power

4.1. Step-down procedure based on MED (MEDSD)

Richter and McCann (2009) proposed a step-down procedure, using medians and permutation tests, for testing all pairwise differences. This method can be adapted to the case of all pairwise comparisons to a control to increase power as follows:

1. Employ the procedure (*MED*) described in Section 3.2 to compute p-values for testing each of the treatments to the control. If the smallest p-value is less than the specified level α, declare the treatment associated with that p-value to have different location from the control, and proceed to step 2. If the smallest p-value is not less than α, stop, and no treatment locations can be declared different from the control at FWER α.

2. Again employ the *MED* procedure, but when constructing the reference distribution of the maximum median differences, exclude the treatment declared different from the control in step 1. Again, if the smallest p-value is less than the specified level α, declare the treatment associated with that p-value to have different location from the control, and proceed to step 3. If the smallest p-value is not less than α, stop, and no further treatment locations can be declared different from the control at FWER α.

3. Continue until the smallest p-value at a step is larger than α.

Richter and McCann (2009) showed that this step-down procedure provides strong control of the FWER while increasing the likelihood that other differences, in addition to the largest observed difference, will be declared statistically significant.

4.2. Step-down procedure based on BON (BONSD)

Two-group permutation tests based on median differences, for all treatments compared to the control, using a Bonferroni adjustment to control the FWER, may also be used. This procedure will be referred to as*BON*. A step-down procedure based on *BON*, denoted BONSD, due to Holm (1979), can also be used as follows:

(1) Employ separate, pairwise permutation tests, based on median differences. Multiply all pairwise p-values by $c-1$, the number of comparisons. If the smallest p-value is less than the specified level α, declare the treatment associated with that p-value to have different location from the control, and proceed to step 2. If the smallest p-value is not less than α stop, and no treatment locations can be declared different from the control at FWER α.

(2) Again employ separate, pairwise permutation tests, but multiply all p-values by $c-2$, and so on, until the last step, when only one comparison is left, where the unadjusted p-value is used.

5. Simulation study

5.1. Simulation setup

A small simulation study was conducted to compare FWER and power properties of the procedures discussed in Sections 3 and 4:

(1) *MED* : The test based on the randomization distribution of $\max|\hat{\mu}_c - \hat{\mu}_j|, j = 1, \ldots, c-1$, the maximum median difference, over all comparisons of treatments to the control, using restricted randomization.

(2) *MEDSD*: The step-down procedure based on *MED*.

(3) *BON*: Separate two-group permutation tests, using median differences, with a Bonferroni correction based on the number of comparisons.

(4) *BONSD*: The step-down procedure based on *BON*.

(5) *DUN*: Dunnett's (1955) procedure.

The following model was assumed to generate the data:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, c-1, c, \; j = 1, \ldots, n_i,$$

where y_{ij} is the jth observation for the ith treatment, μ_i is the location parameter for the ith treatment, and ε_{ij} is the random error associated with the jth observation for the ith treatment. The errors, ε_{ij}, are assumed independent and identically distributed. Several different distributions were considered for ε_{ij}. Normal, Laplace, and Cauchy distributions were chosen to represent symmetric distributions with progressively heavier tails. Similarly, exponential and lognormal (σ=1.5) distributions were chosen to represent lighter and heavier-tailed skewed distributions, respectively. While methods based on means (such as Dunnett's procedure) are not consistent for the Cauchy distribution, this comparison is included to get a sense of the maximum power advantage of the median-based methods.

The FWER, power for each individual comparison, and average power over all false hypotheses were estimated. Models containing three, four and five treatments, in addition to the control, with both equal and unequal sample sizes per group were examined. In most cases the total number of permutations possible was prohibitive, and thus a random sample of permutations was used to estimate the p-value for any given test. Each permutation test was based on a reference distribution estimated using 2000 randomly sampled permutations, and the estimated proportion of rejections in each case was based on 1000 randomly generated samples. For cases with unequal sample sizes, the estimated values were averaged over the different arrangements of sample sizes with treatment locations.

5.2. Simulation results

5.2.1. Type I error

All procedures controlled the FWER in the strong sense for all scenarios examined. FWER for *BONSD*tended to be closest to the nominal level of 0.05, while others, especially those

for *MED* and *MEDSD*, tended to be much smaller than 0.05. Tables 1‑2 show representative results for FWER estimates.

Table 1. FWER − proportion of times at least one true null hypothesis was rejected at α=0.05, three treatments, n_i=10, locations$\mu_C=\mu_1=\mu_2=0;\mu_3=2$.

Procedure	Distribution				
	Normal	Laplace	Cauchy	Exponential	Lognormal
MED	0.013	0.007	0.015	0.009	0.018
MEDSD	0.013	0.007	0.015	0.010	0.018
BON	0.023	0.030	0.027	0.034	0.029
BONSD	0.034	0.037	0.030	0.039	0.036
DUN	0.021	0.019	0.008	0.020	0.014

Table 2. FWER − proportion of times at least one true null hypothesis was rejected at α=0.05, four treatments, n_i=10, locations$\mu_C=\mu_1=0;\mu_2=0.5,\mu_3=1,\mu_4=1.5$.

Procedure	Distribution				
	Normal	Laplace	Cauchy	Exponential	Lognormal
MED	0.002	0.005	0.004	0.007	0.011
MEDSD	0.002	0.006	0.006	0.009	0.011
BON	0.008	0.012	0.010	0.012	0.010
BONSD	0.020	0.031	0.019	0.033	0.021
DUN	0.009	0.016	0.006	0.013	0.008

5.2.2. Power

Table 3, Table 4, Table 5, Table 6, Table 7, Table 8 and Table 9 show representative results for estimated power. As expected, *DUN* always showed highest power (of any kind) for normally distributed data. *MED* always had the highest power for detecting the largest difference for non-normal distributions. However, when more than one treatment location differed from the control, the power for detecting smaller location differences suffered, resulting in lower average power for *MED* compared to *BON* (although *MED* did have highest average power for data from a Cauchy distribution, when only one treatment differed substantially from the control− see Table 7, Table 8 and Table 9). For lighter-tailed non-normal distributions, however, *MED* tended to have the lowest average power, with *DUN* and *BON* obtaining higher, but often similar, power (*DUN* could have slightly higher power than *BON* for the symmetric distribution (Laplace), while *BON* usually had slightly higher power for the skewed distribution (exponential)). With small sample sizes (n_i=5), *BON* showed little or no power (see Table 7), due to the small number of permutations and the discreteness of the two-sample

permutation distributions. The step-down extensions *MEDSD* and *BONSD* had only slightly higher power than the single-step methods.

Table 3. Power‑power to detect the superior treatment, when all other treatments have equal location with the control, using α=0.05, three treatments, n_i=10, locations $\mu_C=\mu_1=\mu_2=0;\mu_3=2$.

Procedure	Distribution				
	Normal	Laplace	Cauchy	Exponential	Lognormal
MED	0.920	0.814	0.433	0.957	0.427
MEDSD	0.920	0.815	0.433	0.957	0.427
BON	0.919	0.668	0.308	0.957	0.437
BONSD	0.919	0.670	0.310	0.959	0.438
DUN	0.973	0.758	0.073	0.951	0.108

Table 4. Power − average power and power to detect the largest location difference, using α=0.05, three treatments, n_i=10, locations $\mu_C=0,\mu_1=0.5,\mu_2=1,\mu_3=1.5$.

Procedure	Distribution									
	Normal		Laplace		Cauchy		Exponential		Lognormal	
	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power
MED	0.709	0.321	0.526	0.163	0.228	0.101	0.830	0.376	0.279	0.144
MEDSD	0.711	0.344	0.529	0.166	0.231	0.105	0.832	0.415	0.288	0.156
BON	0.689	0.360	0.451	0.236	0.184	0.099	0.841	0.520	0.297	0.162
BONSD	0.702	0.392	0.466	0.258	0.196	0.109	0.852	0.573	0.310	0.183
DUN	0.815	0.457	0.483	0.252	0.038	0.024	0.798	0.471	0.242	0.034

Table 5. Power − average power and power to detect the largest location difference, using α=0.05, three treatments, n_i=20, locations $\mu_C=0,\mu_1=0.5,\mu_2=1,\mu_3=1.25$.

Procedure	Distribution									
	Normal		Laplace		Cauchy		Exponential		Lognormal	
	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power

MED	0.784	0.457	0.761	0.401	0.372	0.203	0.922	0.594	0.345	0.211
MEDSD	0.785	0.458	0.762	0.401	0.372	0.203	0.926	0.595	0.347	0.212
BON	0.820	0.512	0.714	0.432	0.318	0.194	0.971	0.710	0.494	0.293
BONSD	0.833	0.519	0.733	0.442	0.329	0.200	0.973	0.711	0.499	0.295
DUN	0.918	0.631	0.654	0.386	0.021	0.014	0.910	0.620	0.036	0.029

Table 6. Power − average power and power to detect the largest location difference, using α=0.05, four treatments, n_i=10, locations $\mu_C=\mu_1=0;\mu_2=0.5,\mu_3=1,\mu_4=1.5$.

Procedure	Distribution									
	Normal		Laplace		Cauchy		Exponential		Lognormal	
	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power
MED	0.649	0.286	0.539	0.224	0.197	0.088	0.797	0.351	0.233	0.113
MEDSD	0.651	0.286	0.540	0.225	0.197	0.088	0.799	0.352	0.233	0.113
BON	0.606	0.313	0.412	0.215	0.147	0.081	0.792	0.466	0.224	0.127
BONSD	0.616	0.318	0.425	0.221	0.157	0.085	0.806	0.473	0.235	0.132
DUN	0.779	0.420	0.488	0.256	0.026	0.027	0.791	0.443	0.052	0.031

Table 7. Power − average power and power to detect the largest location difference, using α=0.05, four treatments, n_i=5, locations$\mu_C=0;\mu_1=\mu_2=\mu_3=0.5,\mu_4=3$.

Procedure	Distribution									
	Normal		Laplace		Cauchy		Exponential		Lognormal	
	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power
MED	0.797	0.266	0.629	0.211	0.218	0.076	0.756	0.252	0.277	0.101
MEDSD	0.797	0.266	0.629	0.211	0.218	0.076	0.756	0.252	0.277	0.101
BON	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
BONSD	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DUN	0.974	0.358	0.762	0.281	0.09	0.035	0.947	0.354	0.138	0.053

Table 8. Power − average power and power to detect the largest location difference, using α=0.05, four treatments, n_i=10, locations μ_C=0;μ_1=μ_2=μ_3=0.5,μ_4=2.

Procedure	Distribution									
	Normal		Laplace		Cauchy		Exponential		Lognormal	
	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power
MED	0.906	0.309	0.809	0.274	0.376	0.132	0.960	0.323	0.379	0.141
MEDSD	0.906	0.309	0.809	0.274	0.376	0.132	0.960	0.323	0.379	0.141
BON	0.873	0.331	0.619	0.236	0.229	0.092	0.921	0.384	0.346	0.137
BONSD	0.874	0.335	0.626	0.241	0.234	0.095	0.924	0.389	0.346	0.139
DUN	0.961	0.372	0.773	0.296	0.048	0.022	0.959	0.387	0.090	0.039

Table 9. Power − average power and power to detect the largest location difference, using α=0.05, four treatments,n_i=5,8,12,15, locations μ_C=0;μ_1=μ_2=0.5,μ_3=2.

Procedure	Distribution									
	Normal		Laplace		Cauchy		Exponential		Lognormal	
	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power	Largest difference	Average power
MED	0.820	0.286	0.671	0.233	0.266	0.098	0.839	0.290	0.246	0.098
MEDSD	0.820	0.290	0.672	0.236	0.266	0.099	0.839	0.294	0.247	0.101
BON	0.768	0.300	0.568	0.219	0.231	0.095	0.773	0.333	0.285	0.115
BONSD	0.772	0.316	0.574	0.231	0.237	0.100	0.778	0.365	0.288	0.123
DUN	0.939	0.376	0.697	0.272	0.072	0.041	0.921	0.388	0.104	0.050

6. Example

Researchers collected data on the levels of osteopontin (OPN), a glycoprotein that has been associated with inflammation and fibrosis (Delimpoura et al., 2010) (See Table 10). The primary focus was to compare the OPN levels in patients with severe refractory asthma (SRA) to levels of those with milder forms of the disease. Since the observed OPN levels for all four groups were highly skewed, medians were reported and used to compare the groups.

Table 10. Means and medians of the OPN levels for the three treatment groups and control group (SRA).

Treatment	n	Mean	Median
SRA	33	6601.2	1840
Moderate asthma	29	188.1	130
Naïve asthma	21	104.8	100
Normal subjects	20	57.5	50

There are four groups, including the control, and thus three comparisons to be made. The estimated p-values for the *MED* , *BON*, *MEDSD* and *BONSD* tests are based on 10,000 random permutations and are presented in Table 11. Both methods declare the SRA group different from all three of the milder condition groups, using α=0.05. Note, however, that *MED* provides stronger evidence for differences from the Mild and Normal groups, which were the two largest observed median differences, while *BON*provides stronger evidence for a difference from the Moderate group, which was the smallest observed median difference.

Table 11. P-values for testing for location difference between SRA and less severe groups, using the single step and step-down median-difference procedures.

Group	Single-step method		Step-down method	
	MED	*BON*	*MEDSD*	*BONSD*
SRA vs. Moderate	0.0122	0.0030	0.0001	0.0001
SRA vs. Mild	0.0097	0.0111	0.0037	0.0074
SRA vs. Normal	0.0094	0.0177	0.0094	0.0177

The first step of the step-down procedure *MEDSD* is to find the smallest p-value from the *MED* procedure, 0.0094, which is associated with the comparison of SRA vs. Normal. Since this p-value is less than α=0.05, the remaining p-values are recomputed by reapplying the *MED* procedure, excluding the Normal group. The smaller of these two p-values is 0.0037, which is associated with the comparison of SRA vs. Mild. Since this p-value is less than α=0.05, recompute the p-value for comparing the SRA vs. Normal, using a two-group permutation test, resulting in 0.0001.

For the step-down procedure *BONSD* , multiply all pairwise permutation test p-values by the number of comparisons (in this example, three), resulting in a smallest p-value of 0.0177 for the SRA vs. Normal comparison. Since this p-value is less than α=0.05, multiply the original p-values from the pairwise permutation tests for the remaining two comparisons by two, yielding a smallest p-value of 0.0074, for the test of SRA vs. Mild. Finally, since this p-value is less than α=0.05, the original pairwise permutation testp-value of 0.001 is used for the test of SRA vs. Moderate. Table 11 summarizes the results of the two step-down tests. Note that the p-values

for Steps 2 and 3 are both smaller than those of the single-step procedure, and also that all p-values using *MEDSD* yield at least as strong evidence as those using*BONSD*.

7. Discussion

Two permutation test procedures were investigated as more robust alternatives to Dunnett's procedure for all pairwise comparisons to a control. The simulation results suggest that the procedures based on medians (*MED* , *BON*) are preferred for data from heavy-tailed distributions. However, there does not appear to be a clear choice between the median procedures for all scenarios. In the cases where it is expected that only one treatment is clearly better than the control, *MED* is the preferred choice to detect the superior treatment, especially for symmetric, non-normal distributions. However, for skewed, non-normal distributions and situations where two or more treatments are expected to be substantially better than the control, *BON* is preferred due to higher power for detecting differences in addition to the largest. With small sample sizes (e.g., $n_i \leq 5$), *BON* may have little or no power, due to the discreteness of the two-sample permutation distributions. Thus, *MED* may also be preferred when sample sizes are small, since the distribution of maximum differences will generally be less discrete.

Step-down extensions of *MED* and *BON* were also presented. While these methods are guaranteed to have at least as much power as the single-step procedures, the power advantage may not be substantial, since only one comparison can be eliminated at each step, resulting in only a nominal power advantage over the single step procedure.

References

Delimpoura, V., et al., 2010. Increased levels of osteopontin in sputum supernatant in severe refractory asthma. Thorax 65, 782–786.

Dunnett, C.W., 1955. A multiple comparison procedure for comparing several treatments with a control. J. Amer. Statist. Assoc. 509, 1096–1121.

Hochberg, Y., Tamhane, A.C., 1987. Multiple Comparison Procedures. John Wiley and Sons, New York.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

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

Richter, S.J., McCann, M.H., 2007. Multiple comparisons using medians and permutation tests. J. Modern Appl. Statist. Methods 6 (2), 399–412.

Richter, S.J., McCann, M.H., 2009. Step-down multiple comparison procedures based on medians and permutation tests. Comm. Statist. Simulation Comput. 38 (8), 1551–1561.

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

Wilcox, R.R., 2006. Comparing medians. Comput. Statist. Data Anal. 51, 1934–1943.