Resampling-based simultaneous confidence intervals for location shift using medians

By: Scott J. Richter, Melinda H. McCann

Richter, S. J. and McCann M. H. (In press). Resampling-based simultaneous confidence intervals for location shift using medians. *AStA Advances in Statistical Analysis*, 100(2), 189-205.

The final publication is available at Springer via http://dx.doi.org/10.1007/s10182-015-0258-4.

Abstract:

A method for computing simultaneous pairwise confidence intervals for location shift is presented, based on the permutation distribution of the maximum absolute pairwise difference among all pairs. The method guarantees strong control of familywise confidence, and does not require assumptions about the form of the population distribution. Simulations compare the permutation procedure to a bootstrap procedure, as well as to the Tukey–Kramer procedure. Simulation results suggest the proposed permutation method produces intervals that maintain simultaneous coverage, and that can be more precise for heavy-tailed distributions compared to competing methods. The permutation intervals may be preferred for data from heavy-tailed distributions.

Keywords: Simultaneous intervals | Permutation test | Median difference | Location shift

Article:

1 Introduction

In single factor designs, estimating all pairwise location differences is frequently of interest. Typically mean differences are estimated, using a method such as Tukey’s (1949) procedure for constructing simultaneous confidence intervals. Tukey’s procedure is optimal for equal size samples from normally distributed populations with equal variance, but there often arise instances when the normal assumption is not plausible. An example is presented in Sect. 5 where the goal is to estimate the magnitude of pairwise differences between four groups. However, the data appear to come from highly skewed distributions and contain extreme outliers, so Tukey’s procedure is not valid. Although methods based on means can often claim “robustness” to nonnormality when samples are large, they may no longer be optimal, and nonparametric procedures with greater precision can often be found. Further, robust measures of location, such as medians, may be more meaningful than means when populations are known to be skewed, and thus inference using medians might be preferred in these situations. In fact, since in most cases the shape of the population distribution will not be known, comparing medians instead of means might be considered a safer choice, since the median is a meaningful parameter to estimate in virtually any distribution.

Procedures that have been proposed for estimating median differences have generally focused on deriving approximate standard errors, and have been based on large sample approximations for resulting Wald-type statistics (see Price and Bonnett 2001). Recently, Bonnett and Price (2002) proposed an approximate procedure for estimating an arbitrary contrast of medians, employing a modification of the variance estimator of McKean and Schrader (1984). Simulations suggested that their procedure had more consistent (close to nominal confidence level) coverage probabilities for estimating all pairwise differences in a four group design, compared to other competitors based on medians. Wilcox (2006) found in further simulations, however, that the method of Bonnett and Price (2002), when employed for multiple testing, had inflated Type I error probabilities for certain discrete distributions where tied observations are likely, which suggests that coverage probabilities for simultaneous intervals would be lower than expected. Wilcox (2006) compared several interval estimators based on medians, and found a percentile bootstrap procedure based on median differences to be the only method to perform well in terms of Type I error probability for all situations. However, bootstrap intervals are only asymptotically exact, and can suffer from lack of precision for small samples (Good 2000). Alternatively, permutation tests can provide distribution-free exact *p* values for multiple testing, and can be inverted to obtain interval estimates, without the need to derive approximate standard errors.

Several methods for using permutation tests for multiple pairwise tests of mean differences have been discussed. For one and two-sample settings, Wheldon et al. (2007) adapted a technique due to Manly et al. (1986) for testing for group mean differences at multiple time points. For their method, a reference distribution is built by taking the minimum *p* value across all time points, for each permutation. Statistical significance is determined by comparing the separate permutation *p* value at each time point to the $1-\alpha 1-\alpha$percentile of the reference distribution. This method is similar to the testing procedure of Miller (1981), who proposed using the maximum absolute mean difference across all pairs to build the reference distribution. However, both the method of Miller (1981) and that of Wheldon et al. (2007) suffer from the fact that observations are permuted freely across all groups or time points, and thus can only control the familywise error rate (FWER) in the weak sense (Hochberg and Tamhane 1987). Richter and McCann (2007) proposed instead permuting separately within each pair of groups being compared, with the reference distribution constructed in fashion similar to Miller (1981), and showed that this technique controlled the FWER in the strong sense (Hochberg and Tamhane 1987).

A method for simultaneous estimation of the location shift parameters, $\Delta ij\Delta ij$, for all pairwise comparisons, based on inverting the test procedure of Richter and McCann (2007), is presented in Sect. 2.

2 Simultaneous confidence intervals using permutation tests

2.1 Testing procedure of Richter and McCann (2007)

Consider a one-way layout with k groups, where F_i is the common continuous distribution function for the ith group, n_i is the sample size of the ith group, and $N=n_1+n_2+\cdots+n_k$. Further, let μi 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. That is, for $i, j = 1, 2, \ldots, k$ with $i < j$, $F_i(x) = F_j(x - \Delta_{ij})$, where Δ_{ij} represents the location difference between groups i and j. The reference distribution 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. Richter and McCann (2007) showed that this procedure provides strong FWER control (Hochberg and Tamhane 1987).

2.2 Individual confidence intervals based on inverting the permutation distribution

Consider first the case where $k = 2$, and suppose a permutation test to compare the locations of the two distributions, using the sample median difference as the test statistic, is conducted. A confidence interval for the pairwise location difference, Δ_{12}, can be constructed as follows. Find two constants, d_l and d_u that, when subtracted from the observed median difference, lead to the smallest p value of the permutation test greater than or equal to $\alpha/2$. Then the $100(1-\alpha)\,\%$ confidence interval for Δ_{12} is $d_l \leq \Delta_{12} \leq d_u$. That this is indeed a $100(1-\alpha)\,\%$ confidence interval for Δ_{12} is verified by Good (2000, p. 210). Note that this amounts to finding a set of values for Δ_{12} that would not be rejected by the corresponding hypothesis test from Sect. 2.1 with $k = 2$.

The values d_l and d_u can be found by first finding the percentiles of the permutation distribution, and then determining d_l and d_u as $d_l = \Delta_{12} - p_{1-\alpha/2}$ and $d_u = \Delta_{12} + p_{1-\alpha/2}$, where $\hat{\Delta}_{12} = \hat{\mu}_1 - \hat{\mu}_2$ is the observed median difference. To see this, consider two independent, identically distributed observations, X_i and X_j, from distribution. Then $P(X_i - X_j \leq c) = P(X_j - X_i \leq c) = P(-(X_i - X_j) \leq c) = P(X_i - X_j \geq -c)$. Thus, the distribution of $X_i - X_j$ is symmetric about 0. Consequently, when using a statistic based directly on differences of statistics from a common distribution, to find a confidence interval it suffices to estimate an upper percentile from the sampling distribution, as the lower percentile will simply be the negative of the respective upper percentile, and thus the confidence interval will be the observed difference plus or minus the appropriate upper percentile point.

2.3 Simultaneous confidence intervals when *k*>2

First, the reference distribution described in Sect. 2.1 is derived, consisting of the maximum absolute median difference, $\max_{1 \leq i < j \leq k} |\hat{\mu}_i - \hat{\mu}_j|$, across all pairs for each permutation. Next, the $1 - \alpha/2$ percentile of the reference distribution, $p(\max)_{1-\alpha/2}$, is determined. Then for each pairwise difference, the constants $d(\max)_l$ and $d(\max)_u$ are found as described in Sect. 2.2. That is, $d(\max)_l = \hat{\Delta}_{ij} - p(\max)_{1-\alpha/2}$ and $d(\max)_u = \hat{\Delta}_{ij} + p(\max)_{1-\alpha/2}$, where $\hat{\Delta}_{ij} = \hat{\mu}_i - \hat{\mu}_j$ is the observed median difference for groups *i* and *j*.

Richter and McCann (2007) proved that for the case of testing all pairwise comparisons, using the distribution of the absolute maximum median difference across all pairs, where randomization is performed within each pair, provides strong control of the FWER. Since the confidence interval procedure is found by inverting the test procedure, the same percentiles used for the testing procedure will be used to construct the confidence intervals. Thus, since the acceptance region of the testing procedure is used to construct the confidence intervals, the simultaneous confidence level must also be controlled. This follows from Good (2000, Theorem 4.1, p. 210) when we make the appropriate generalization to a set of tests with their corresponding confidence intervals and consider FWER instead of individual coverage. Specifically, let $A(\mathbf{\Delta})$, $\mathbf{\Delta} = \{\Delta_{ij}, i, j = 1, \ldots, k, i < j\}$ be the set of all values in the sample space where the simultaneous testing procedure of Richter and McCann (2007), detailed in Sect. 2.1, would not be rejected if the values Δ_{ij} were simultaneously tested. Now let $\mathbf{S}(\mathbf{X})$ be the set of simultaneous confidence intervals described above that result from a specific $\mathbf{X}$ in the sample space. Note that a vector $\mathbf{\Delta}$ of Δ_{ij} values will be included in $\mathbf{S}(\mathbf{X})$ if and only if this $\mathbf{\Delta}$ would not be rejected by the simultaneous procedure described above. Consequently, $P(\mathbf{\Delta} \in \mathbf{S}(\mathbf{X})) = P(\mathbf{X} \in A(\mathbf{\Delta})) \geq 1 - \alpha$, and thus simultaneous coverage at the $1 - \alpha$ level is guaranteed.

2.4 Alternative estimators for location shift

The procedure described in Sect. 2.3 is based on the sample median difference as an estimator of the location shift. A similar procedure may be employed for forming confidence intervals based on other estimators, such as the Hodges–Lehmann estimator of location shift, which is the midpoint of all pairwise differences among observations. The reference distribution may be derived, using permutation tests, as the distribution of the maximum median pairwise difference for each permutation.

Since the sample median is known to be relatively inefficient for light-tailed distributions, the Harrell–Davis quantile estimator (Harrell and Davis 1982) is sometimes considered as an alternative. However, Wilcox (2006) found that replacing the sample median with the Harrell–Davis estimator worsened the performance of the best performing median procedures. For this reason, the Harrell–Davis estimator is not considered here.

3 Simulation

3.1 Simulation details

A simulation study was used to estimate simultaneous coverage probabilities for the permutation confidence interval method discussed in Sect. 2.3, using the median difference (MED) as well as the median of pairwise differences (PWD) to estimate location difference. To compare to existing methods, intervals were also computed using a percentile bootstrap procedure (MEDB) based on median differences (Wilcox 2006, 2012), as well as to the parametric Tukey–Kramer procedure (TK) based on mean differences.

Wilcox (2006, 2012) recommended a percentile bootstrap procedure based on median differences for comparing medians. Wilcox (2006) considered the pairwise testing case, and found that pairwise bootstrap distributions, combined with Rom 's (1990) method of p value adjustment, provided a testing procedure that worked well over all cases considered. For pairwise confidence intervals, however, Rom's method cannot be applied, and thus a Bonferroni correction is used here to achieve desired simultaneous coverage for the pairwise bootstrap intervals.

The additive model, $y_{ij} = \mu_i + \varepsilon_{ij},\ i = 1, 2, \ldots, k, j = 1, \ldots, n_i$, was assumed to generate the data, where μ_i is the location parameter associated with the *i*th treatment. Several different *g*-and-*h* (Hoaglin 1985) error distributions were considered. The five different distributions were the standard normal distribution (g=h=0)(g=h=0), symmetric non-normal distributions with moderately heavy (g=0,h=0.4)(g=0,h=0.4) and very heavy (g=0,h=0.8)(g=0,h=0.8) tails, and skewed distributions with light (g=0.8,h=0)(g=0.8,h=0) and moderately heavy (g=0.8,h=0.4)(g=0.8,h=0.4) tails.

Equal and unequal sample size cases were examined, for different configurations of location difference. For the unequal sample size cases, sample sizes were randomly assigned to each distribution to avoid potential bias in power estimates due to sample size.

3.2 Simulation results

3.2.1 Simultaneous coverage

Estimated simultaneous coverage probabilities are given in Tables 1, 2, 3 and 4. The MEDB bootstrap intervals had coverage very close to the nominal level over all conditions. The MED and PWD permutation intervals had simultaneous coverage of 95 % or higher for all cases considered, and tended to be conservative in most cases, with a higher estimated coverage probability than the advertised 0.95.

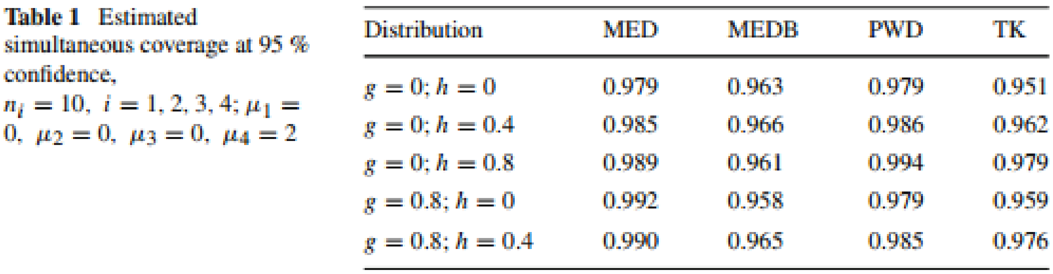

Table 1 Estimated simultaneous coverage at 95 % confidence, $n_i = 10,\ i = 1, 2, 3, 4; \mu_1 = 0,\ \mu_2 = 0,\ \mu_3 = 0,\ \mu_4 = 2$

Distribution	MED	MEDB	PWD	TK
$g = 0; h = 0$	0.979	0.963	0.979	0.951
$g = 0; h = 0.4$	0.985	0.966	0.986	0.962
$g = 0; h = 0.8$	0.989	0.961	0.994	0.979
$g = 0.8; h = 0$	0.992	0.958	0.979	0.959
$g = 0.8; h = 0.4$	0.990	0.965	0.985	0.976

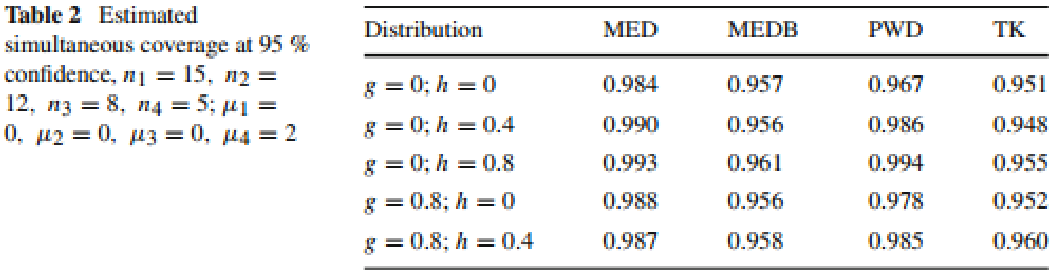

Table 2 Estimated simultaneous coverage at 95 % confidence, $n_1 = 15,\ n_2 = 12,\ n_3 = 8,\ n_4 = 5; \mu_1 = 0,\ \mu_2 = 0,\ \mu_3 = 0,\ \mu_4 = 2$

Distribution	MED	MEDB	PWD	TK
$g = 0; h = 0$	0.984	0.957	0.967	0.951
$g = 0; h = 0.4$	0.990	0.956	0.986	0.948
$g = 0; h = 0.8$	0.993	0.961	0.994	0.955
$g = 0.8; h = 0$	0.988	0.956	0.978	0.952
$g = 0.8; h = 0.4$	0.987	0.958	0.985	0.960

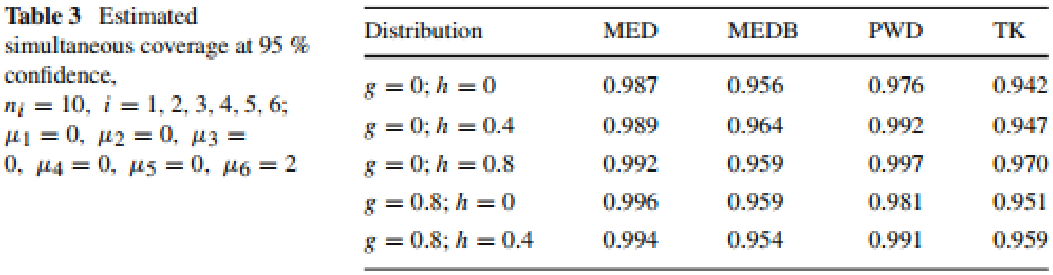

Table 3 Estimated simultaneous coverage at 95 % confidence, $n_i = 10,\ i = 1, 2, 3, 4, 5, 6; \mu_1 = 0,\ \mu_2 = 0,\ \mu_3 = 0,\ \mu_4 = 0,\ \mu_5 = 0,\ \mu_6 = 2$

Distribution	MED	MEDB	PWD	TK
$g = 0; h = 0$	0.987	0.956	0.976	0.942
$g = 0; h = 0.4$	0.989	0.964	0.992	0.947
$g = 0; h = 0.8$	0.992	0.959	0.997	0.970
$g = 0.8; h = 0$	0.996	0.959	0.981	0.951
$g = 0.8; h = 0.4$	0.994	0.954	0.991	0.959

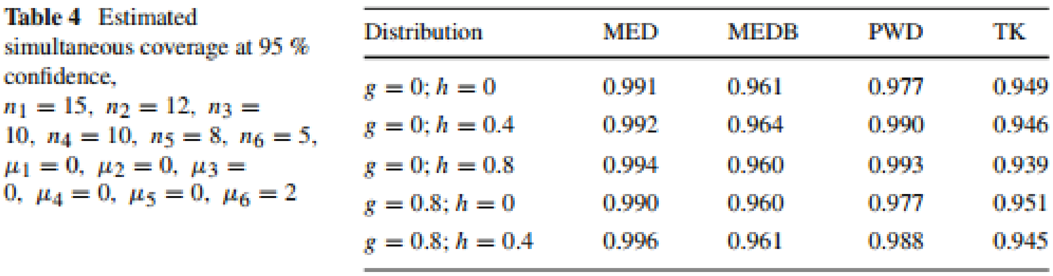

Table 4 Estimated simultaneous coverage at 95 % confidence, $n_1 = 15,\ n_2 = 12,\ n_3 = 10,\ n_4 = 10,\ n_5 = 8,\ n_6 = 5, \mu_1 = 0,\ \mu_2 = 0,\ \mu_3 = 0,\ \mu_4 = 0,\ \mu_5 = 0,\ \mu_6 = 2$

Distribution	MED	MEDB	PWD	TK
$g = 0; h = 0$	0.991	0.961	0.977	0.949
$g = 0; h = 0.4$	0.992	0.964	0.990	0.946
$g = 0; h = 0.8$	0.994	0.960	0.993	0.939
$g = 0.8; h = 0$	0.990	0.960	0.977	0.951
$g = 0.8; h = 0.4$	0.996	0.961	0.988	0.945

3.2.2 Interval width

For each simulated data set, the mean interval length for all intervals for a particular method was calculated. Then, for all simulated data sets under a given condition, minimum, maximum and quartiles of mean interval widths for each method were calculated, and are given in Tables 5, 6, 7 and 8. The MEDB procedure tended to produce wider intervals than the MED and PWD methods for all non-normal distributions, and could become much wider when distributions were heavier-tailed.

TK intervals were always the narrowest when data were generated by a normal distribution, and usually narrowest for the lighter-tailed skewed distribution (g=0.8,h=0)(g=0.8,h=0). For the heavier-tailed distributions (g=0,h=0.8g=0,h=0.8; g=0.8,h=0.4g=0.8,h=0.4), the median-based procedures always had smaller maximum length, and sometimes narrower third quartile and median lengths. The TK intervals could become very wide for all of the heavier-tailed distributions, and this problem was especially severe for the distribution with g=0,h=0.8g=0,h=0.8.

Among MED and PWD intervals, MED tended to have slightly narrower intervals when distributions were symmetric but heavy-tailed (g=0g=0, h=0.4h=0.4; g=0g=0, h=0.8h=0.8), while PWD tended to produce narrower intervals when distributions were skewed. However, the difference in interval length was never substantial.

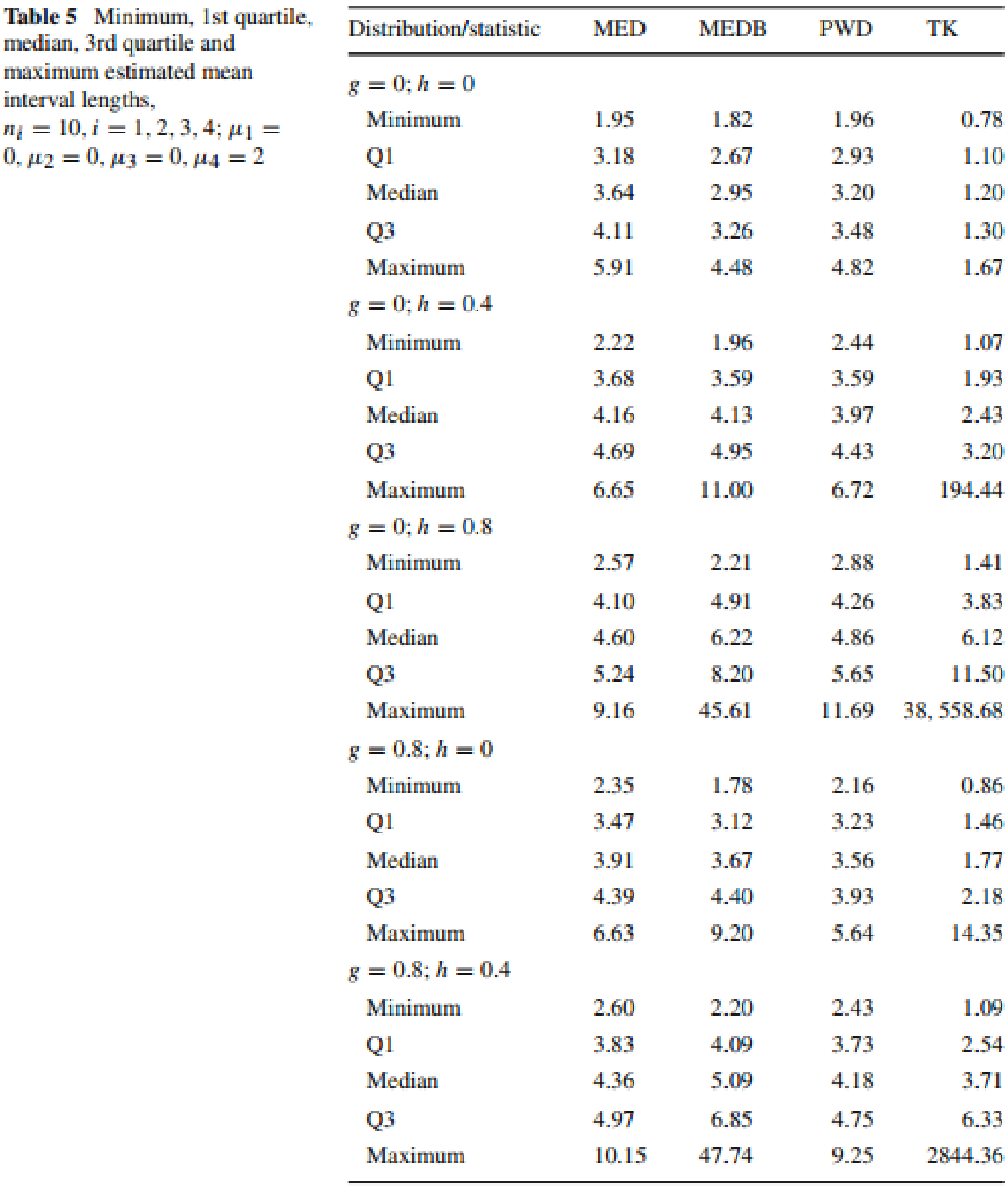

Table 5 Minimum, 1st quartile, median, 3rd quartile and maximum estimated mean interval lengths, $n_i = 10, i = 1, 2, 3, 4; \mu_1 = 0, \mu_2 = 0, \mu_3 = 0, \mu_4 = 2$

Distribution/statistic	MED	MEDB	PWD	TK
$g = 0; h = 0$				
Minimum	1.95	1.82	1.96	0.78
Q1	3.18	2.67	2.93	1.10
Median	3.64	2.95	3.20	1.20
Q3	4.11	3.26	3.48	1.30
Maximum	5.91	4.48	4.82	1.67
$g = 0; h = 0.4$				
Minimum	2.22	1.96	2.44	1.07
Q1	3.68	3.59	3.59	1.93
Median	4.16	4.13	3.97	2.43
Q3	4.69	4.95	4.43	3.20
Maximum	6.65	11.00	6.72	194.44
$g = 0; h = 0.8$				
Minimum	2.57	2.21	2.88	1.41
Q1	4.10	4.91	4.26	3.83
Median	4.60	6.22	4.86	6.12
Q3	5.24	8.20	5.65	11.50
Maximum	9.16	45.61	11.69	38, 558.68
$g = 0.8; h = 0$				
Minimum	2.35	1.78	2.16	0.86
Q1	3.47	3.12	3.23	1.46
Median	3.91	3.67	3.56	1.77
Q3	4.39	4.40	3.93	2.18
Maximum	6.63	9.20	5.64	14.35
$g = 0.8; h = 0.4$				
Minimum	2.60	2.20	2.43	1.09
Q1	3.83	4.09	3.73	2.54
Median	4.36	5.09	4.18	3.71
Q3	4.97	6.85	4.75	6.33
Maximum	10.15	47.74	9.25	2844.36

Table 6 Minimum, 1st quartile, median, 3rd quartile and maximum estimated mean interval lengths, $n_1 = 15,\ n_2 = 12,\ n_3 = 8,\ n_4 = 5, \mu_1 = 0,\ \mu_2 = 0,\ \mu_3 = 0,\ \mu_4 = 2$

Distribution/statistic	MED	MEDB	PWD	TK
$g = 0; h = 0$				
Minimum	2.16	1.82	2.09	0.85
Q1	3.39	2.86	3.06	1.19
Median	3.85	3.19	3.35	1.30
Q3	4.33	3.57	3.74	1.40
Maximum	6.85	5.09	5.69	1.84
$g = 0; h = 0.4$				
Minimum	2.63	2.23	2.47	1.08
Q1	4.08	3.94	3.94	2.09
Median	4.66	4.86	4.47	2.63
Q3	5.28	6.28	5.12	3.46
Maximum	9.74	43.88	9.56	212.15
$g = 0; h = 0.8$				
Minimum	3.09	2.69	2.97	1.42
Q1	4.76	5.79	4.86	4.10
Median	5.51	8.43	5.89	6.61
Q3	6.66	13.48	7.28	12.43
Maximum	22.12	856.62	23.10	42061.61
$g = 0.8; h = 0$				
Minimum	2.42	1.79	2.35	0.94
Q1	3.79	3.31	3.46	1.60
Median	4.37	4.04	3.93	1.91
Q3	4.99	5.08	4.45	2.35
Maximum	10.46	16.64	9.17	15.67
$g = 0.8; h = 0.4$				
Minimum	2.56	2.17	2.60	1.18
Q1	4.31	4.51	4.22	2.76
Median	5.13	6.12	4.90	3.99
Q3	6.28	9.09	5.92	6.82
Maximum	23.79	286.57	23.40	3102.79

Table 7 Minimum, 1st quartile, median, 3rd quartile and maximum estimated mean interval lengths, $n_i = 10,\ i = 1, 2, 3, 4, 5, 6;\ \mu_1 = 0,\ \mu_2 = 0,\ \mu_3 = 0,\ \mu_4 = 0,\ \mu_5 = 0,\ \mu_6 = 2$

Distribution/statistic	MED	MEDB	PWD	TK
$g = 0; h = 0$				
Minimum	2.31	2.31	2.41	0.94
Q1	3.44	3.10	3.15	1.24
Median	3.84	3.33	3.42	1.31
Q3	4.36	3.55	3.71	1.40
Maximum	6.07	4.63	5.19	1.80
$g = 0; h = 0.4$				
Minimum	2.47	2.98	2.70	1.24
Q1	4.01	4.65	3.89	2.27
Median	4.49	5.46	4.35	2.75
Q3	4.95	6.52	4.74	3.52
Maximum	7.28	49.54	9.38	45.77
$g = 0; h = 0.8$				
Minimum	2.60	3.36	3.07	1.60
Q1	4.52	7.75	4.80	4.94
Median	5.06	10.44	5.53	7.81
Q3	5.75	15.37	6.49	13.86
Maximum	10.48	730.67	27.63	2712.98
$g = 0.8; h = 0$				
Minimum	2.63	2.32	2.62	1.06
Q1	3.79	3.81	3.48	1.65
Median	4.19	4.37	3.79	1.95
Q3	4.71	5.20	4.19	2.31
Maximum	7.37	9.71	6.14	6.05
$g = 0.8; h = 0.4$				
Minimum	2.66	3.02	2.64	1.45
Q1	4.22	5.60	4.09	3.07
Median	4.75	7.34	4.58	4.34
Q3	5.38	10.05	5.19	6.85
Maximum	10.35	98.74	8.83	185.65

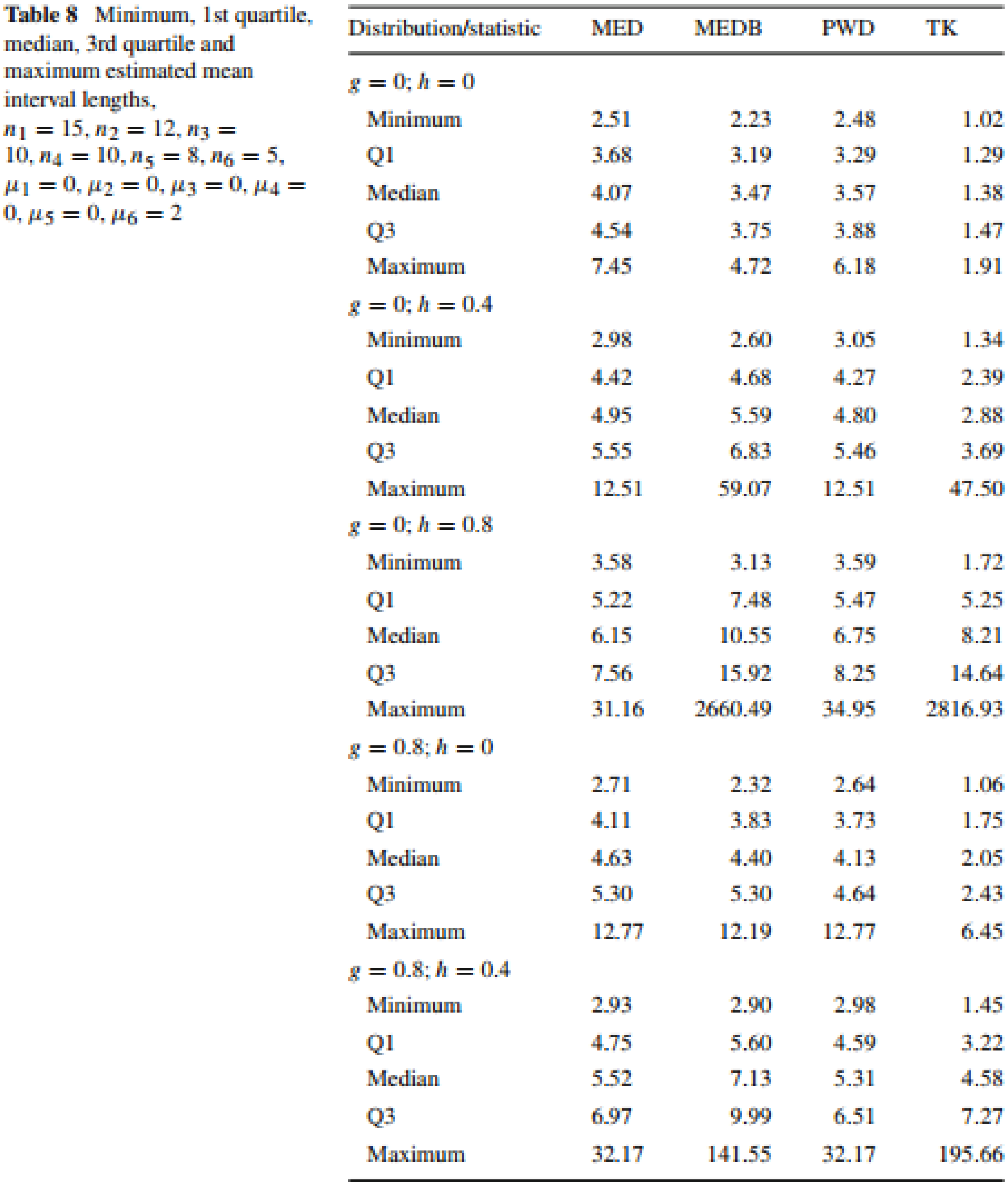

Table 8 Minimum, 1st quartile, median, 3rd quartile and maximum estimated mean interval lengths, $n_1 = 15, n_2 = 12, n_3 = 10, n_4 = 10, n_5 = 8, n_6 = 5, \mu_1 = 0, \mu_2 = 0, \mu_3 = 0, \mu_4 = 0, \mu_5 = 0, \mu_6 = 2$

Distribution/statistic	MED	MEDB	PWD	TK
$g = 0; h = 0$				
Minimum	2.51	2.23	2.48	1.02
Q1	3.68	3.19	3.29	1.29
Median	4.07	3.47	3.57	1.38
Q3	4.54	3.75	3.88	1.47
Maximum	7.45	4.72	6.18	1.91
$g = 0; h = 0.4$				
Minimum	2.98	2.60	3.05	1.34
Q1	4.42	4.68	4.27	2.39
Median	4.95	5.59	4.80	2.88
Q3	5.55	6.83	5.46	3.69
Maximum	12.51	59.07	12.51	47.50
$g = 0; h = 0.8$				
Minimum	3.58	3.13	3.59	1.72
Q1	5.22	7.48	5.47	5.25
Median	6.15	10.55	6.75	8.21
Q3	7.56	15.92	8.25	14.64
Maximum	31.16	2660.49	34.95	2816.93
$g = 0.8; h = 0$				
Minimum	2.71	2.32	2.64	1.06
Q1	4.11	3.83	3.73	1.75
Median	4.63	4.40	4.13	2.05
Q3	5.30	5.30	4.64	2.43
Maximum	12.77	12.19	12.77	6.45
$g = 0.8; h = 0.4$				
Minimum	2.93	2.90	2.98	1.45
Q1	4.75	5.60	4.59	3.22
Median	5.52	7.13	5.31	4.58
Q3	6.97	9.99	6.51	7.27
Maximum	32.17	141.55	32.17	195.66

4 Robustness to scale heterogeneity

For the simulations discussed in Sect. 3, distributions were assumed to be identical except for possible shifts in location. However, Romano (1990) showed that two-sample permutation tests for comparing medians were generally invalid when scale parameters were unequal. Thus, it was of interest to investigate the robustness of the procedures to scale heterogeneity.

Simulations were conducted under similar conditions as for the location shift model in Sect. 3, but allowing the scales of the distributions to vary by as much as a 4:1 ratio between the largest and smallest scales. Tables 9 and 10 show representative results of estimated coverage under various settings of heterogeneity. As expected, the permutation procedures can fail to maintain coverage at the nominal level for both 4- and 6-group cases when there is a large amount of scale heterogeneity. For all cases considered, coverage for the permutation tests eventually dropped below nominal level as the ratio between the smallest and largest scale increased. The magnitude of the effect depended on several factors, including the number of groups, the distribution that generated the data, whether or not the sample sizes were equal, and, when sample sizes were unequal, on the pattern of unequal scales and whether the larger scale parameter was associated with the smaller sample size. However, the lack of coverage was in general not due to narrower intervals, as interval widths tended to be larger when scales were unequal (Tables 11, 12).

The effect of heterogeneity was less severe when data came from heavy-tailed distributions, when sample sizes were equal, and when there were several nonzero differences in magnitude of scale between pairs of groups. For example, Table 10 shows that coverage was poorest when sample sizes were unequal, there was only one group with scale different from the others, and the group with the larger scale had the smallest sample size. For unequal sample sizes, while the effect was more severe when the largest scale parameter was associated with the smallest sample, when the largest scale parameter was associated with the largest sample the permutation methods actually became quite conservative.

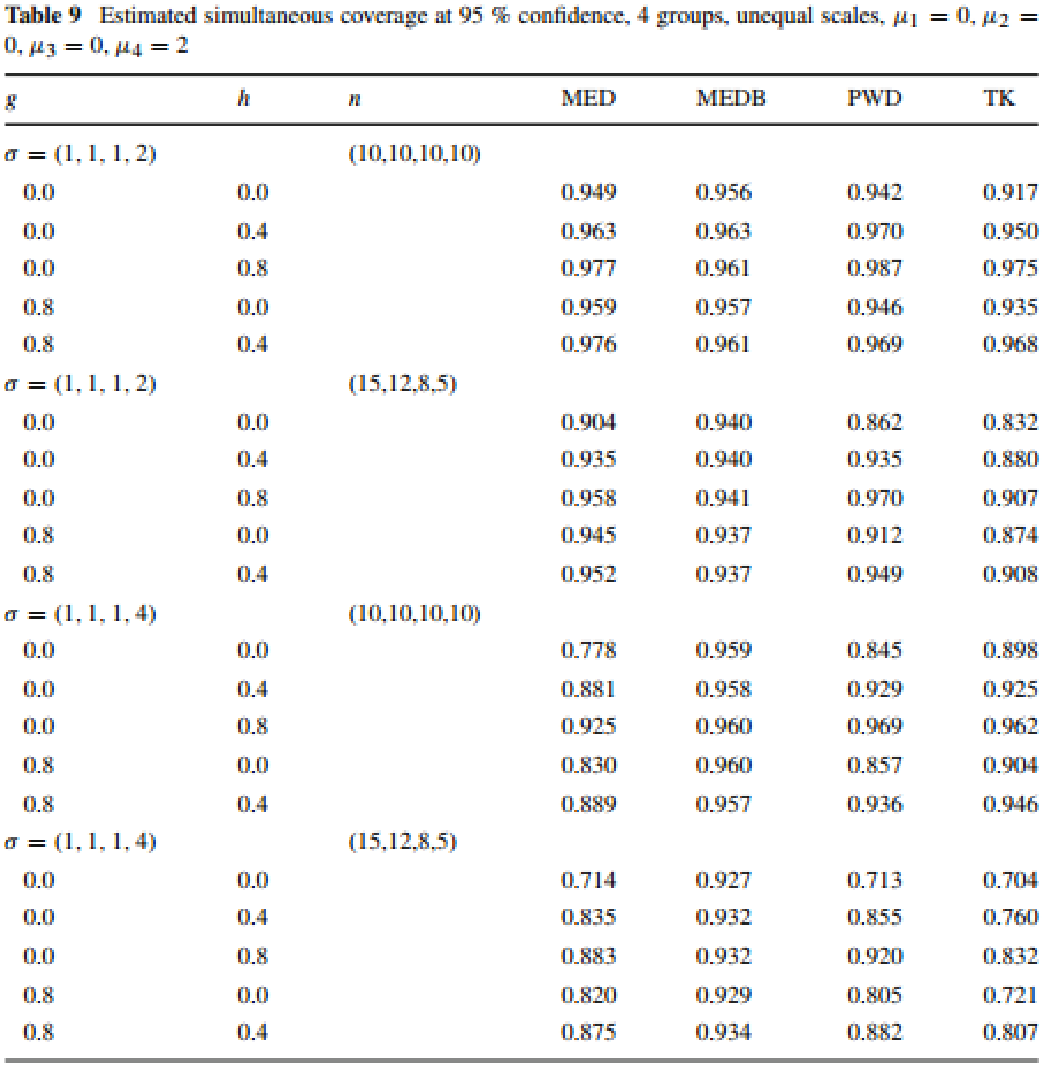

Table 9 Estimated simultaneous coverage at 95 % confidence, 4 groups, unequal scales, $\mu_1 = 0, \mu_2 = 0, \mu_3 = 0, \mu_4 = 2$

g	h	n	MED	MEDB	PWD	TK
$\sigma = (1, 1, 1, 2)$		(10,10,10,10)				
0.0	0.0		0.949	0.956	0.942	0.917
0.0	0.4		0.963	0.963	0.970	0.950
0.0	0.8		0.977	0.961	0.987	0.975
0.8	0.0		0.959	0.957	0.946	0.935
0.8	0.4		0.976	0.961	0.969	0.968
$\sigma = (1, 1, 1, 2)$		(15,12,8,5)				
0.0	0.0		0.904	0.940	0.862	0.832
0.0	0.4		0.935	0.940	0.935	0.880
0.0	0.8		0.958	0.941	0.970	0.907
0.8	0.0		0.945	0.937	0.912	0.874
0.8	0.4		0.952	0.937	0.949	0.908
$\sigma = (1, 1, 1, 4)$		(10,10,10,10)				
0.0	0.0		0.778	0.959	0.845	0.898
0.0	0.4		0.881	0.958	0.929	0.925
0.0	0.8		0.925	0.960	0.969	0.962
0.8	0.0		0.830	0.960	0.857	0.904
0.8	0.4		0.889	0.957	0.936	0.946
$\sigma = (1, 1, 1, 4)$		(15,12,8,5)				
0.0	0.0		0.714	0.927	0.713	0.704
0.0	0.4		0.835	0.932	0.855	0.760
0.0	0.8		0.883	0.932	0.920	0.832
0.8	0.0		0.820	0.929	0.805	0.721
0.8	0.4		0.875	0.934	0.882	0.807

Table 10 Estimated simultaneous coverage at 95 % confidence, 6 groups, unequal scales, $\mu_1 = 0$, $\mu_2 = 0$, $\mu_3 = 0$, $\mu_4 = 0$, $\mu_5 = 0$, $\mu_6 = 2$

g	h	n	MED	MEDB	PWD	TK
$\sigma = (1, 2, 2, 3, 3, 4)$		(10,10,10,10,10,10)				
0.0	0.0		0.945	0.950	0.901	0.913
0.0	0.4		0.961	0.961	0.970	0.932
0.0	0.8		0.977	0.957	0.989	0.966
0.8	0.0		0.970	0.945	0.923	0.927
0.8	0.4		0.978	0.954	0.975	0.953
$\sigma = (1, 2, 2, 3, 3, 4)$		(15,12,10,10,8,5)				
0.0	0.0		0.927	0.929	0.865	0.864
0.0	0.4		0.956	0.936	0.948	0.883
0.0	0.8		0.973	0.934	0.978	0.883
0.8	0.0		0.972	0.929	0.929	0.863
0.8	0.4		0.979	0.933	0.968	0.884
$\sigma = (1, 1, 1, 1, 1, 4)$		(10,10,10,10,10,10)				
0.0	0.0		0.883	0.958	0.895	0.872
0.0	0.4		0.931	0.959	0.953	0.898
0.0	0.8		0.957	0.959	0.982	0.954
0.8	0.0		0.932	0.960	0.908	0.870
0.8	0.4		0.957	0.961	0.954	0.930
$\sigma = (1, 1, 1, 1, 1, 4)$		(15,12,10,10,8,5)				
0.0	0.0		0.757	0.933	0.737	0.688
0.0	0.4		0.874	0.932	0.885	0.757
0.0	0.8		0.922	0.934	0.942	0.814
0.8	0.0		0.862	0.940	0.833	0.731
0.8	0.4		0.910	0.935	0.908	0.813

Table 11 Minimum, 1st quartile, median, 3rd quartile and maximum estimated mean interval lengths, unequal scales ($\sigma = (1, 1, 1, 4)$), $n_i = 10, i = 1, 2, 3, 4$; $\mu_1 = 0$, $\mu_2 = 0$, $\mu_3 = 0$, $\mu_4 = 2$

Distribution/Statistic	MED	MEDB	PWD	TK
$g = 0; h = 0$				
Minimum	1.92	2.84	2.24	1.27
Q1	3.46	4.84	3.95	2.22
Median	4.05	5.61	4.69	2.56
Q3	4.78	6.45	5.51	2.93
Maximum	8.78	10.53	9.06	4.38
$g = 0; h = 0.4$				
Minimum	2.05	3.44	2.69	1.80
Q1	4.07	6.23	5.03	3.64
Median	4.95	7.73	6.12	4.75
Q3	6.06	9.82	7.46	6.82
Maximum	13.13	80.44	14.51	194.46

Table 11 continued

Distribution/Statistic	MED	MEDB	PWD	TK
$g = 0; h = 0.8$				
Minimum	2.14	3.94	2.85	2.22
Q1	4.70	8.37	6.26	6.79
Median	5.86	11.13	7.88	11.23
Q3	7.48	15.79	10.11	22.02
Maximum	21.69	166.32	24.80	38558.68
$g = 0.8; h = 0$				
Minimum	1.81	2.38	2.00	1.35
Q1	3.66	5.39	4.22	2.73
Median	4.50	6.62	4.94	3.41
Q3	5.60	8.32	5.86	4.70
Maximum	15.12	25.33	13.27	20.47
$g = 0.8; h = 0.4$				
Minimum	1.93	3.43	2.48	1.94
Q1	4.20	6.87	5.16	4.61
Median	5.36	9.19	6.28	6.82
Q3	6.93	12.61	7.97	11.84
Maximum	25.03	122.52	24.40	2844.36

Table 12 Minimum, 1st quartile, median, 3rd quartile and maximum estimated mean interval lengths, unequal scales ($\sigma = (1, 1, 1, 1, 1, 4)$), $n_1 = 15, n_2 = 12, n_3 = 10, n_4 = 10, n_5 = 8, n_6 = 5$, $\mu_1 = 0, \mu_2 = 0, \mu_3 = 0, \mu_4 = 0, \mu_5 = 0, \mu_6 = 2$

Distribution/statistic	MED	MEDB	PWD	TK
$g = 0; h = 0$				
Minimum	2.63	2.61	4.41	1.12
Q1	4.42	4.88	4.41	1.70
Median	4.98	5.66	5.05	1.93
Q3	6.18	6.56	6.23	2.22
Maximum	16.51	9.71	16.07	3.71
$g = 0; h = 0.4$				
Minimum	3.12	3.16	2.96	1.70
Q1	5.73	7.07	6.00	3.04
Median	6.87	8.96	7.37	3.87
Q3	8.56	12.08	9.17	5.21
Maximum	45.13	90.65	45.13	47.52
$g = 0; h = 0.8$				
Minimum	3.40	3.98	3.31	2.54
Q1	7.35	11.11	8.27	6.40
Median	9.18	16.15	10.40	10.43
Q3	12.05	26.72	14.04	18.99
Maximum	129.64	5311.65	129.64	2816.93

Table 12 continued

Distribution/statistic	MED	MEDB	PWD	TK
$g = 0.8; h = 0$				
Minimum	2.73	2.79	2.55	1.22
Q1	5.17	5.68	5.12	2.20
Median	6.20	6.86	6.19	2.68
Q3	7.74	8.78	7.46	3.39
Maximum	30.30	37.23	28.94	16.81
$g = 0.8; h = 0.4$				
Minimum	3.21	3.61	3.07	1.76
Q1	6.31	8.22	6.65	4.04
Median	7.96	11.04	8.30	5.73
Q3	10.75	16.63	10.86	9.38
Maximum	57.67	546.51	55.16	274.22

Table 13 Dry biomass (mg) of ants for 24 adult males and yearling females, taken in four months in 1980

Month	Dry biomass (mg)
June	13, 105, 242
July	2, 8, 20, 59, 245
August	40, 50, 52, 82, 88, 233, 488, 515, 600, 1889
September	0, 5, 6, 18, 21, 44

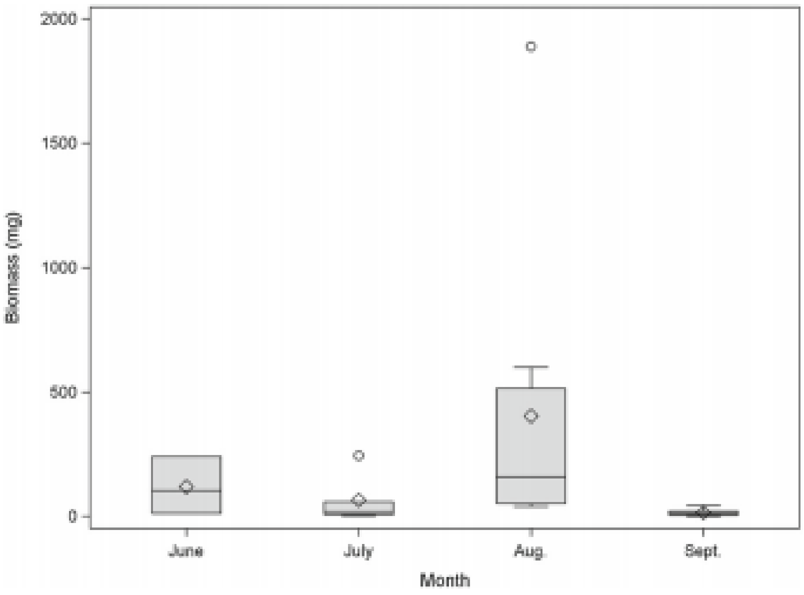

Fig. 1 Distributions of biomass (mg) of ants for 24 adult males and yearling females, taken in 4 months in 1980

Table 14 Simultaneous 95 % confidence limits for location difference, using methods MED, PWD, and TK, for all pairs of four months for the data in Table 13

Months	Median difference	Mean difference	MED		PWD		MEDB		TK	
			95 % Lower	95 % Upper	95 % Lower	95 % Upper	95 % Lower	95 % Upper	95 % Lower	95 % Upper
June, July	85.0	53.2	−359.5	529.5	−409.5	501.5	−232.0	240.0	−731.4	837.8
June, Aug.	−55.5	−283.7	−500.0	389.0	−527.5	383.5	−544.5	190.0	−991.0	423.6
June, Sept.	93.0	104.3	−351.5	537.5	−362.5	548.5	−19.5	239.5	−655.4	864.1
July, Aug.	−140.5	−336.9	−585.0	304.0	−538.5	372.5	−580.0	178.0	−925.4	251.6
July, Sept.	8.0	51.1	−436.5	452.5	−444.5	466.5	−30.5	240.0	−599.5	701.7
Aug., Sept.	148.0	388.0	−296.0	593.0	−317.0	594.0	31.0	579	−166.8	942.9

5 Example

Powell and Russell (1984, 1985) and Linton et al. (1989) collected data (Table 13) on the stomach contents of eastern horned lizards for each of four summer months. It was desired to estimate the magnitude of consumption differences between different pairs of months. Since the sample distributions were skewed with a few extreme outliers (see Fig. 1), medians might be more meaningful measures of location.

Simultaneous 95 % confidence intervals are presented in Table 14 using methods MED, MEDB, PWD, and TK. Note that for these data, which represent skewed distributions with extreme values, the Tukey–Kramer method can produce extremely wide intervals, in comparison to the permutation and bootstrap methods. However, even if the TK intervals were narrower, they would necessarily be centered at the mean difference, which for heavy-tailed data may not be appropriate as the estimate of the location shift. Thus, the median-based intervals, which are centered at the sample median difference and are not susceptible to extremely wide intervals due to extreme outliers, may provide a better alternative for such distributions. The MEDB bootstrap intervals were generally narrower than the MED intervals, but were also not centered on the sample median difference. This is likely due to skewness in the bootstrap distribution, which can result in biased intervals (see Manly 1997). Thus, the MED intervals may be preferred for these data.

6 Discussion

Permutation-based methods were presented for simultaneous estimation of location difference. The results of the simulation study suggest that the permutation-based intervals may be preferred when data are expected to come from heavy-tailed distributions. While the permutation methods can be less precise in some situations, they are also not susceptible to producing extremely wide intervals, as are the MEDB and TK methods. It may seem surprising that the median-based MEDB intervals could be adversely affected by extreme data points. However, the bootstrap intervals can have samples where extreme values are oversampled, resulting in large variability in the bootstrap distribution and thus wide intervals. Thus, even though they have shown good coverage estimates for heavy-tailed distributions, this appears to be at the expense of precision.

The conservativeness of coverage levels of MED in certain situations may be due to the discreteness of the permutation distributions, especially for small sample sizes. For permutation tests, one suggestion to alleviate the effect of discreteness on the *p* value is the mid-*p* value (Lancaster 1961). That is, if t_0 is the observed test statistic, then the *p* value is $mid\text{-}p = \frac{1}{2}P(T = t_0) + P(T > t_0)$. Thus, we proposed an analogous adjustment, the *mid-critical point*, for simultaneous confidence intervals. The adjustment works as follows: if $P(D \leq d_c) = 1 - \alpha/2$, then the confidence level is exact, and $p_{1-\alpha/2} = d_c$ is used; however, if $P(D \leq d_c) > 1 - \alpha/2$, then the critical value is chosen to be the midpoint between d_{c-1} and d_c: $p^*_{1-\alpha/2} = (d_c + d_{c-1})/2$. Unfortunately, except for very small sample sizes $(n_i < 5)$ simulation results showed little or no gain in precision for the scenarios considered, and for this reason results using this adjustment were not included.

The permutation procedures considered in this paper can be recommended as robust alternatives for data from heavy-tailed distributions. Since the bootstrap procedure does not require the assumption of equal distributions, it may be preferred when variance heterogeneity is expected.

References

Bonnett, D.G., Price, R.M.: Statistical inference for a linear function of medians: confidence intervals, hypothesis testing, and sample size requirements. Psychol. Methods 7(3), 370–383 (2002)

Good, P.: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, 2nd edn. Springer, New York (2000)

Harrell, F.E., Davis, C.E.: A new distribution-free quantile estimator. Biometrika 69, 635–640 (1982)

Hoaglin, D.C.: Summarizing shape numerically: the g-and-h distribution. In: Hoaglin, D., Mosteller, F., Tukey, J. (eds.) Exploring Data Tables Trends and Shapes. Wiley, New York (1985)

Hochberg, Y., Tamhane, A.C.: Multiple Comparison Procedures. Wiley, New York (1987)

Lancaster, H.O.: Significance tests in discrete distributions. J. Am. Stat. Assoc. 56, 223–234 (1961)

Linton, L.R., Edgington, E.S., Davies, R.W.: A view of niche overlap amenable to statistical analysis. Can. J. Zool. 67, 55–60 (1989)

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

Manly, B.F.J., Mcalevey, L., Stevens, D.: A randomization procedure for comparing group means on multiple measurements. Br. J. Math. Stat. Psychol. 39, 183–189 (1986)

McKean, J.W., Schrader, R.M.: A comparison of methods for studentizing the sample median. Commun. Stat. Simul. Comput. 13, 751–773 (1984)

Miller, R.G.: Simultaneous Statistical Inference, 2nd edn. Springer, New York (1981)

Powell, G.L., Russell, A.P.: The diet of the eastern short-horned lizard (*Phrynosoma douglassi brevirostre*) in Alberta and its relationship to sexual size dimorphism. Can. J. Zool. 62, 428–440 (1984)

Powell, G.L., Russell, A.P.: Growth and sexual size dimorphism in Alberta populations of the eastern short-horned lizard, *Phrynosoma douglassi brevirostre*. Can. J. Zool. 63, 139–154 (1985)

Price, R.M., Bonnett, D.G.: Estimating the variance of the sample median. J. Stat. Comput. Simul. 68, 295–305 (2001)

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

Rom, D.M.: A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika 77, 663–666 (1990)

Romano, J.P.: On the behavior of randomization tests without a group invariance assumption. J. Am. Stat. Assoc. 85(411), 686–692 (1990)

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

Wheldon, M.C., Anderson, M.J., Johnson, B.W.: Identifying treatment effects in multi-channel measurements in electroencephalographic studies: multivariate permutation tests and multiple comparisons. Aust. N. Z. J. Stat. 49(4), 397–413 (2007)

Wilcox, R.R.: Comparing medians. Comput. Stat. Data Anal. 51, 1934–1943 (2006)

Wilcox, R.R.: Introduction to Robust Estimation and Hypothesis Testing, 3rd edn. Academic Press, Burlington (2012)