Pairwise comparison of scale using deviances

By: Scott J. Richter and Melinda H. McCann

Scott J. Richter & Melinda H. McCann (2019) Pairwise comparison of scale using deviances, *Journal of Statistical Computation and Simulation*, 89:9, 1730-1739, DOI:10.1080/00949655.2019.1593986

This is an Accepted Manuscript of an article published by Taylor & Francis in *Journal of Statistical Computation and Simulation* on 24 March 2019, available online: http://www.tandfonline.com/10.1080/00949655.2019.1593986.

Abstract:

Permutation tests based on medians are examined for pairwise comparison of scale. Tests that have been found in the literature to be effective for comparing scale for two groups are extended to the case of all pairwise comparisons, using the Tukey-type adjustment of Richter and McCann [Multiple comparison of medians using permutation tests. J Mod Appl Stat Methods. 2007;6(2):399–412] to guarantee strong Type I error rate control. Power and Type I error rate estimates are computed using simulated data. A method based on the ratio of deviances performed best and appears to be the best overall test.

Keywords: Deviance | scale parameter | permutation test

Article:

1. Introduction

Pairwise comparison of scale parameters may be of interest in many areas, including industrial quality control, agricultural production and experimental education [1]. However, it is well known that the parametric *F*-test for comparing variances of two treatments, as well as parametric tests for more than two treatments (e.g. tests due to Bartlett [2], Cochran [3], Hartley [4]) are generally not robust to non-normality (see [5]). Consequently, more robust tests of scale parameters are of interest.

There are many comparative studies of tests for comparing scale differences in the literature. Levene [6] proposed using the ANOVA *F*-test on the absolute deviations from the mean. Brown and Forsythe [7] proposed instead using absolute deviations from the median, which they referred to as the '*W50*' test. While no uniformly best test for scale has been demonstrated in the literature, the *W50* test has been recommended as showing good overall performance with respect to power and robustness to non-normality in several comparative studies, including

Keselman et al. [8], Conover et al. [9], and Balakrishnan and Ma [10]. O'Brien [11] proposed a modification of Levene's test (*OB50*) which has been recommended over the *W50* test for lighter-tailed distributions [11,12]. Marozzi [13] considered the *W50* and *OB50* tests, as well as permutation versions of these tests, and found that the permutation versions of these tests tended to be more robust and have higher power. They recommended the permutation *W50* test as a computationally simple robust test, but also the permutation version of *OB50*, which had higher power for symmetric and lighter-tailed skewed distributions. Richter and McCann [14] found that the permutation *RMD* test due to Higgins [15] was generally superior to *W50* and *OB50*, especially for heavier-tailed distributions.

In this paper, we extend the *RMD*, *W50* and *OB50* tests to the problem of simultaneous pairwise comparison of scale. The method of Richter and McCann [16] is used to control the familywise error rate (FWER), and estimated Type I error rates and power of the methods are compared.

2. Methods

Consider a one-way layout with *t* treatments and n_i observations per treatment. We assume a location-scale model, $y_{ij} = \mu_i + \sigma_i e_{ij}, i = 1, \ldots, t, j = 1, \ldots, n_i$ where μ_i and σ_i are the location and scale parameters, respectively, of treatment *i*, and ε_{ij} are independent and identically distributed with median 0. It is desired to test $H_0 : \sigma_i = \sigma_j$ versus $H_a : \sigma_i = \sigma_j$ for all pairs $i \neq j$.

2.1. *LEV* and *W50* tests

Levene [6] proposed a robust test (*LEV*) to compare scale parameters, using the ANOVA *F*-statistic computed on the absolute deviations from the treatment means, $\bar{z}_{ij} = \left|y_{ij} - \bar{y}_i\right|$. Brown and Forsythe [7] suggested a statistic (*W50*) that instead used absolute deviations from the treatment medians, $\tilde{z}_{ij} = \left|y_{ij} - \tilde{y}_i\right|$, which we will refer to in the remainder of the paper as *deviances*. Conover et al. [9] found *W50* to have better power and size properties than *LEV*. Both Levene [6] and Brown and Forsythe [7] suggested that *p*-values be based on the *F*-distribution with $t - 1$ and $n - t$ degrees of freedom. Marrozzi [13], however, examined permutation versions of these tests, and found the permutation versions to be more robust and more powerful than those based on the *F*-distribution.

2.2. *OB50* tests

O'Brien [11] proposed a modification to *LEV*, suggesting the scores $r_{ij}(w) = \left[(w + n_j - 2)n_i\left(y_{ij} - \bar{y}_j\right)^2 - ws_j^2(n_j - 1)\right]/\left[(n_i - 1)(n_j - 2)\right]$, where $0 \leq w \leq 1$. At one extreme, when $w = 0$, the statistic reduces to $r_{ij}(0) = \bar{z}_{ij}^2 = n_i\left(y_{ij} - \bar{y}_i\right)^2/(n_i - 1)$, which is a modification of *LEV* using the scores, $\bar{z}_{ij}^2 = \left(y_{ij} - \bar{y}_i\right)^2$. When using the permutation distribution, the test based on $\bar{z}_{ij}^2$ will be equivalent to that based on $\bar{z}_{ij}$. At the other extreme, when $w = 1, r_{ij}(1) = q_{ij} = \left[n_i\left(y_{ij} - \bar{y}_i\right)^2 - s_j^2\right]/(n_i - 2) = n_j s_j^2 - \left(n_j - 1\right)s_{j-1}^2$, which O'Brien [11] referred to as a 'jackknife pseudovalue of $s_j^{2'}$. Tests based on $\bar{z}_{ij}^2$ have been shown to have inflated Type I error rates [9], while those based on q_{ij} tend to have low power [12].

Since $r_{ij}(w)$ is a weighted average of the two tests, it provides a way to balance their drawbacks. O'Brien [11] suggested that a 'utility' value of $w = 0.5$ would work satisfactorily for a majority of situations, and this is the version studied here. Marozzi [11] studied the permutation version of *OB50*, and found it to be robust in all situations, and to have the highest power for symmetric and lighter-tailed distributions.

2.3. *RMD* test

Higgins [15] also suggested a test based on the deviances, or the scores $\tilde{z}_{ij} = |y_{ij} - \tilde{y}_i|$. Higgins' test statistic for the two-sided alternative was the ratio of the mean sample deviances, $RMD = \max(\bar{\tilde{z}}_i, \bar{\tilde{z}}_j)/\min(\bar{\tilde{z}}_i, \bar{\tilde{z}}_j)$, where $\bar{\tilde{z}}_{ij}$ is the mean of the scores $\tilde{z}_{ij}$ for treatment *i*. The *p*-value was determined using the permutation distribution. Richter and McCann [14] compared the performance of *RMD* to the permutation versions of *W50* and *OB*, and found that *RMD* often had more power and never much less power, and thus recommended *RMD* as the best overall choice for comparing scales of two treatments. They also noted the intuitive appeal of the *RMD* statistic as a ratio, and thus a closer, more robust analogue to the normal-theory variance ratio *F*-test.

3. Strong familywise error rate control for pairwise comparisons

The familywise error rate (FWER) will be controlled using two methods: the technique of Richter and McCann [16], and a Bonferroni correction. Richter and McCann [16] proposed a restricted permutation method to provide strong control of the familywise error rate (FWER) for pairwise comparison of location parameters. This method will be extended to the present case of comparing scale parameters as follows. First, the two-sample test statistic for a given method will be calculated for each of the possible $t(t-1)/2$ pairs of treatments. Then the maximum value of the test statistic across all $t(t-1)/2$ pairs will be calculated. Next, observations will be reassigned at random to treatments within each pair of treatments, and a test statistic calculated for each pair of treatments, and the maximum value determined. This will be repeated many times to build the permutation distribution, and the *p*-value for comparing each pair of treatments will be calculated as the proportion of values in the permutation distribution that are at least as extreme as the observed value.

A Bonferroni correction will also be employed to provide strong familywise error rate control by making comparisons using an adjusted significance level of $\alpha/[t(t-1)/2]$.

4. Simulation

A simulation was conducted to compare estimated power and FWER for the procedures described in Section 2, controlling FWER using both the Tukey-type adjustment of Richter and McCann [16] as well as with a Bonferroni correction:

1. *LEV* – Levene's method, using Tukey-type adjustment;
2. *LEVB* – Levene's method, using Bonferroni adjustment;
3. *W50* – Brown and Forsyth *W*50 test, using Tukey-type adjustment;
4. *W50B* – Brown and Forsyth *W*50 test, using Bonferroni adjustment;

5. *OB50* – O'Brien's method, using Tukey-type adjustment;
6. *OB50B* – O'Brien's method, using Bonferroni adjustment;
7. *RMD* – Higgins' method, using Tukey-type adjustment;
8. *RMDB* – Higgins' method, using Bonferroni adjustment.

4.1. Sample sizes and scale parameter settings

Several combinations of sample sizes were chosen to investigate large, small, equal and unequal sizes. Settings at both three and five treatments were examined. For each, equal sample sizes of $n_i = 10$ and $n_i = 20$ were used. For three treatments, unequal sample size settings of $n_1 = 5, n_2 = 10, n_3 = 15$ and $n_1 = 15, n_2 = 20, n_3 = 25$ were used, while for five treatments $n_1 = 5, n_2 = 5, n_3 = 10, n_2 = 15, n_3 = 15$ and $n_1 = 10, n_2 = 10, n_3 = 15, n_2 = 20, n_3 = 20$ were considered. For three treatments the scale parameter patterns $(\sigma, 1,1)$ and $(\sigma, (\sigma + 1)/2,1)$ were used, while for five treatments the patterns $(\sigma, 1,1,1,1)$ and $(\sigma, (\sigma + 1)/2,1,1,1)$ were used. In each case, the first pattern is referred to as the 'single extreme scale' parameter setting, while the second pattern has an intermediate scale value midway between the minimum and maximum.

For the unequal sample size scenarios, first considered was the case where the sample sizes were randomly assigned to the scale values. Next, two special cases were considered, where (1) the sample size was positively associated with scale magnitude, and (2) the sample size was negatively associated with scale magnitude.

4.2. Distributions and permutation test

Several different *g* and *h* distributions [17] were used to simulate data from distributions with different characteristics. The *g* and *h* distributions are monotonic functions of normal distributions, and allow investigation of non-normal distributions with specific characteristics. The *g*-and-*h* random variable is defined as $Y_{g,h}(Z) = ((\exp(gZ) - 1)/g)\exp(hZ^2/2)$, where $Z \sim N(0,1)$. When $g = h = 0$, $Y_{g,h}(Z) \sim N(0,1)$. Nonzero values of *g* increase the skewness and positive values of *h* increase the elongation (tail heaviness) of the distribution. Changing the values of *g* and *h* does not affect the location of the distribution. The following cases were considered:

1. $g = 0$, $h = 0$ – normally distributed (symmetric, light tails);
2. $g = 0$, $h = 0.4$ – symmetric, moderately heavy tails;
3. $g = 0$, $h = 0.8$ – symmetric, very heavy tails;
4. $g = 0.8$, $h = 0$ – skewed, light tails;
5. $g = 0.8$, $h = 0.4$ – skewed, moderately heavy tails.

Type I error rate and power were estimated based on 1000 randomly selected data sets from each distribution, for each setting of sample sizes and scale parameter pattern. Two types of power were estimated. The first was *any-pair power*, or the probability of at least one rejection among all false pairwise hypotheses (detecting at least the largest scale difference). The second was *all-pairs power*, or the probability of rejecting all false hypotheses (detecting all scale differences).

Marozzi [1] suggested that at least 253 random permutations are necessary with 1000 random data sets if the goal of the simulation is to estimate the power of a test and only a 'rough' estimate of the permutation *p*-value is required, while Keller-McNulty and Higgins [18] recommended a random sample of at least 1600 permutations to estimate the exact *p*-value for a permutation test. Since precise estimation of the permutation test *p*-values was considered important, a conservative 2000 random permutations were utilized to estimate the *p*-value for each permutation test.

5. Results

Representative simulation results are presented in this section. Additional results are available from the corresponding author upon request.

5.1. FWER

All methods maintained estimated probabilities of at least one false rejection at or below the nominal rate of 0.05. With five groups, the estimates tended to be much lower than the nominal level, generally 0.02 or less. However, *RMD* tended to have estimates closest to 0.05.

5.2. Any-pair power

Any-pair power – the probability of detecting at least the largest scale difference – was estimated for all cases. There was no method that was most powerful in all situations.
However, *W50B* emerged as having higher power more often than any other method. We next give a detailed summary of the results, separately for the equal and unequal sample size cases.

5.2.1. Any-pair power – equal sample sizes

For normally distributed data, there was little difference in power between any of the methods. Power estimates were also generally similar for all methods for the light-tailed, skewed ($g = 0.8$, $h = 0$) distribution, with the exception of *OB50*, which tended to have the least power. With three groups and for heavier-tailed distributions, *RMD* most often had the highest power when $n_i = 10$, although *RMDB* sometimes had slightly higher power for the $(\sigma, 1, 1)$ scale pattern. When $n_i = 20$, *RMD* tended to have higher power for the $(\sigma, (\sigma + 1)/2, 1)$ pattern, with *RMDB* and *W50B* not far behind, while *RMDB* and *W50B* had similar and higher power for the $(\sigma, 1, 1)$ pattern with *RMD* not far behind (see Table 1).

With five groups and for heavier-tailed and skewed distributions, *W50B* nearly always had the highest power with *RMD* usually second for the $(\sigma, (\sigma + 1)/2, 1)$ pattern and *RMDB* usually second for the $(\sigma, 1, 1)$ pattern (see Table 2).

Table 1. Proportion at least one rejection, (FWER/any-pair power) at $\alpha = 0.05$, three treatments, equal sample sizes.

$n_i = 10$	Method								
Distribution	$\sigma_1\ \sigma_2\ \sigma_3$	***LEV***	***LEVB***	***W50***	***W50B***	***OB***	***OB50***	***RMD***	***RMDB***
$g = 0$, $h = 0$	111	0.035	0.034	0.037	0.032	0.030	0.028	0.040	0.036
	311	0.697	0.626	0.699	0.620	0.684	0.637	0.744	0.666
	321	0.579	0.596	0.537	0.587	0.481	0.446	0.663	0.611
	511	0.919	0.879	0.923	0.891	0.893	0.815	0.962	0.927
	531	0.887	0.897	0.845	0.899	0.718	0.654	0.945	0.921
$g = 0$, $h = 0.4$	111	0.030	0.042	0.028	0.039	0.011	0.011	0.047	0.041
	311	0.233	0.260	0.218	0.284	0.114	0.113	0.300	0.303
	321	0.173	0.21	0.112	0.218	0.068	0.072	0.258	0.228
	511	0.438	0.526	0.425	0.540	0.223	0.212	0.528	0.576
	531	0.307	0.404	0.223	0.430	0.125	0.112	0.462	0.456
$g = 0$, $h = 0.8$	111	0.034	0.043	0.018	0.039	0.008	0.008	0.053	0.042
	311	0.115	0.139	0.081	0.155	0.038	0.041	0.151	0.163
	321	0.111	0.122	0.049	0.123	0.028	0.024	0.164	0.132
	511	0.225	0.286	0.181	0.131	0.067	0.085	0.278	0.328
	531	0.175	0.219	0.084	0.236	0.043	0.040	0.272	0.239
$g = 0.8$, $h = 0$	111	0.036	0.033	0.037	0.034	0.028	0.015	0.048	0.036
	311	0.525	0.540	0.499	0.512	0.492	0.334	0.523	0.517
	321	0.379	0.445	0.299	0.425	0.296	0.166	0.427	0.412
	511	0.790	0.812	0.786	0.776	0.738	0.544	0.812	0.816
	531	0.636	0.748	0.507	0.707	0.486	0.291	0.729	0.723
$g = 0.8$, $h = 0.4$	111	0.034	0.039	0.024	0.035	0.010	0.007	0.053	0.037
	311	0.222	0.265	0.179	0.261	0.103	0.090	0.256	0.257
	321	0.173	0.200	0.095	0.192	0.062	0.054	0.236	0.196
	511	0.400	0.480	0.364	0.489	0.203	0.185	0.447	0.511
	531	0.292	0.366	0.187	0.369	0.099	0.082	0.400	0.390
$n_i = 20$									
$g = 0$, $h = 0$	311	0.982	0.967	0.978	0.968	0.984	0.979	0.987	0.976
	321	0.950	0.961	0.933	0.961	0.917	0.915	0.982	0.966
	511	1	1	1	1	1	0.998	1	1
	531	0.999	1	0.998	1	0.992	0.988	0.999	1
$g = 0$, $h = 0.4$	311	0.330	0.413	0.340	0.443	0.160	0.162	0.404	0.447
	321	0.237	0.309	0.179	0.331	0.097	0.095	0.360	0.336
	511	0.628	0.754	0.659	0.797	0.384	0.377	0.688	0.804
	531	0.442	0.579	0.370	0.629	0.175	0.168	0.642	0.634
$g = 0$, $h = 0.8$	311	0.105	0.157	0.090	0.170	0.034	0.051	0.160	0.174
	321	0.104	0.123	0.049	0.125	0.024	0.033	0.157	0.128
	511	0.227	0.317	0.195	0.377	0.075	0.099	0.321	0.381
	531	0.175	0.225	0.089	0.263	0.040	0.051	0.298	0.268
$g = 0.8$, $h = 0$	311	0.737	0.789	0.788	0.801	0.746	0.607	0.780	0.802
	321	0.554	0.625	0.523	0.664	0.471	0.301	0.672	0.650
	511	0.962	0.974	0.977	0.978	0.948	0.855	0.970	0.983
	531	0.843	0.914	0.825	0.944	0.710	0.538	0.945	0.946
$g = 0.8$, $h = 0.4$	311	0.221	0.327	0.230	0.358	0.128	0.115	0.316	0.366
	321	0.164	0.237	0.115	0.251	0.071	0.062	0.284	0.257
	511	0.448	0.641	0.485	0.682	0.277	0.260	0.570	0.688
	531	0.306	0.446	0.244	0.493	0.119	0.100	0.503	0.497

Table 2. Proportion at least one rejection, (FWER/any-pair power), $\alpha = 0.05$, five treatments, equal sample sizes.

$n_i = 10$		Method							
Distribution	$\sigma_1\ \sigma_2\ \sigma_3\ \sigma_4\ \sigma_5$	***LEV***	***LEVB***	***W50***	***W50B***	***OB***	***OB50***	***RMD***	***RMDB***
$g = 0$, $h = 0$	11111	0.006	0.001	0.004	0.018	0.006	0.008	0.006	0.008
	31111	0.688	0.291	0.681	0.637	0.788	0.755	0.669	0.437
	32111	0.606	0.214	0.583	0.610	0.586	0.553	0.578	0.312
	51111	0.912	0.623	0.922	0.916	0.937	0.888	0.962	0.817
	53111	0.891	0.964	0.882	0.946	0.778	0.962	0.936	0.730
$g = 0$, $h = 0.4$	11111	0.003	0	0.002	0.010	0.005	0.006	0.013	0.006
	31111	0.114	0.020	0.119	0.218	0.087	0.091	0.167	0.152
	32111	0.072	0.008	0.074	0.150	0.050	0.054	0.125	0.078
	51111	0.274	0.068	0.298	0.425	0.213	0.197	0.342	0.367
	53111	0.192	0.024	0.203	0.399	0.114	0.283	0.274	0.221
$g = 0$, $h = 0.8$	11111	0.005	0	0.002	0.014	0.005	0.008	0.019	0.004
	31111	0.034	0.004	0.028	0.252	0.023	0.028	0.078	0.070
	32111	0.028	0.003	0.016	0.072	0.024	0.017	0.056	0.037
	51111	0.068	0.012	0.065	0.305	0.045	0.090	0.135	0.176
	53111	0.051	0.011	0.037	0.164	0.031	0.096	0.102	0.093
$g = 0.8$, $h = 0$	11111	0.004	0.001	0.002	0.016	0.007	0.010	0.010	0.009
	31111	0.390	0.134	0.416	0.405	0.472	0.310	0.325	0.325
	32111	0.277	0.080	0.285	0.353	0.321	0.185	0.273	0.186
	51111	0.734	0.390	0.777	0.760	0.774	0.585	0.718	0.686
	53111	0.580	0.257	0.585	0.727	0.505	0.750	0.544	0.496
$g = 0.8$, $h = 0.4$	11111	0.004	0.001	0.004	0.020	0.007	0.005	0.024	0.009
	31111	0.081	0.014	0.086	0.217	0.071	0.072	0.141	0.132
	32111	0.050	0.005	0.052	0.141	0.049	0.035	0.109	0.085
	51111	0.202	0.059	0.225	0.379	0.146	0.130	0.269	0.319
	53111	0.114	0.020	0.128	0.325	0.084	0.230	0.222	0.184
$n_i = 20$									
$g = 0$, $h = 0$	31111	0.981	0.862	0.978	0.975	0.989	0.988	0.984	0.940
	32111	0.961	0.771	0.961	0.896	0.943	0.961	0.973	0.896
	51111	0.995	1	1	1	0.998	1	1	1
	53111	1	0.980	1	1	0.992	0.990	1	1
$g = 0$, $h = 0.4$	31111	0.186	0.053	0.215	0.360	0.150	0.147	0.298	0.331
	32111	0.150	0.024	0.162	0.243	0.068	0.069	0.237	0.223
	51111	0.458	0.217	0.531	0.703	0.332	0.334	0.543	0.678
	53111	0.343	.0119	0.358	0.641	0.149	0.146	0.466	0.482
$g = 0$, $h = 0.8$	31111	0.045	0.008	0.036	0.275	0.030	0.038	0.112	0.114
	32111	0.037	0.003	0.028	0.119	0.015	0.020	0.084	0.077
	51111	0.099	0.021	0.090	0.363	0.065	0.077	0.194	0.276
	53111	0.068	0.015	0.051	0.258	0.025	0.032	0.137	0.174
$g = 0.8$, $h = 0$	31111	0.678	0.394	0.762	0.769	0.732	0.601	0.692	0.687
	32111	0.546	0.262	0.603	0.714	0.51	0.363	0.605	0.520
	51111	0.959	0.854	0.979	0.979	0.945	0.869	0.960	0.973
	53111	0.854	0.639	0.903	0.969	0.695	0.568	0.908	0.894
$g = 0.8$, $h = 0.4$	31111	0.117	0.026	0.131	0.306	0.092	0.093	0.207	0.235
	32111	0.079	0.012	0.095	0.259	0.044	0.038	0.179	0.167
	51111	0.290	0.119	0.365	0.581	0.223	0.211	0.415	0.571
	53111	0.190	0.048	0.225	0.523	0.088	0.086	0.332	0.370

5.2.2. Any-pair power – unequal sample sizes, random

When unequal sample sizes were randomly assigned to three groups, *RMDB* always had the highest power, followed by *RMD*, especially as tails became heavier and scale differences increased in magnitude (see Table 3). For five groups, *W50B* usually had the highest power followed by *RMDB* (see Table 4).

Table 3. Proportion at least one rejection, (FWER/any-pair power), $\alpha = 0.05$, three treatments, unequal sample sizes, random *n*.

n_1, n_2, n_3 = 10,20,30	Method								
Distribution	$\sigma_1\ \sigma_2\ \sigma_3$	***LEV***	***LEVB***	***W50***	***W50B***	***OB***	***OB50***	***RMD***	***RMDB***
$g = 0, h = 0$	111	0.047	0.044	0.049	0.043	0.034	0.035	0.043	0.035
	311	0.866	0.977	0.900	0.972	0.629	0.555	0.976	0.990
	321	0.909	0.980	0.307	0.980	0.535	0.441	0.944	0.989
	511	0.999	1	0.995	1	0.772	0.720	1	1
	531	0.997	1	0.45	1	0.739	0.667	0.999	1
$g = 0, h = 0.4$	111	0.019	0.050	0.013	0.053	0.012	0.013	0.039	0.036
	311	0.034	0.099	0.048	0.141	0.002	0.001	0.425	0.535
	321	0.045	0.171	0.038	0.209	0.017	0.015	0.309	0.398
	511	0.093	0.220	0.119	0.311	0.002	0.002	0.696	0.822
	531	0.098	0.343	0.068	0.418	0.022	0.019	0.559	0.661
$g = 0, h = 0.8$	111	0.026	0.052	0.010	0.057	0.012	0.011	0.046	0.038
	311	0.010	0.032	0.006	0.032	0.004	0.004	0.236	0.260
	321	0.027	0.080	0.012	0.084	0.015	0.013	0.158	0.180
	511	0.013	0.034	0.008	0.041	0.004	0.002	0.388	0.466
	531	0.036	0.128	0.019	0.140	0.016	0.015	0.264	0.313
$g = 0.8, h = 0$	111	0.022	0.043	0.021	0.043	0.015	0.013	0.045	0.038
	311	0.351	0.475	0.355	0.592	0.064	0.037	0.719	0.833
	321	0.258	0.504	0.115	0.602	0.076	0.049	0.559	0.724
	511	0.566	0.769	0.644	0.906	0.111	0.067	0.959	0.991
	531	0.496	0.828	0.184	0.925	0.116	0.067	0.893	0.976
$g = 0.8, h = 0.4$	111	0.025	0.050	0.016	0.043	0.014	0.012	0.043	0.042
	311	0.017	0.058	0.025	0.084	0.003	0.002	0.356	0.434
	321	0.045	0.133	0.027	0.162	0.017	0.012	0.258	0.315
	511	0.047	0.114	0.059	0.180	0.005	0.001	0.602	0.720
	531	0.081	0.249	0.050	0.296	0.021	0.017	0.448	0.545

5.2.3. Any-pair power – unequal sample sizes, not random

Also investigated were cases where the sample size was related to the scaled magnitude. First considered was the case where the largest sample size was associated with the largest scale parameter and the smallest sample size with the smallest scale parameter. In these scenarios, *RMDB* always had the highest power for all non-normal distributions (see Table 5).

Table 4. Proportion of at least one rejection at $\alpha = 0.05$, five treatments, unequal sample sizes, random n.

n_i = 10,10,20,30,30	Method								
Distribution	σ_1 σ_2 σ_3 σ_4 σ_5	***LEV***	***LEVB***	***W50***	***W50B***	***OB50***	***OB50B***	***RMD***	***RMDB***
$g = 0$, $h = 0$	11111	0.003	0	0.004	0.008	0.001	0.002	0.004	0.008
	31111	0.903	0.451	0.910	0.938	0.854	0.823	0.882	0.800
	32111	0.856	0.382	0.860	0.958	0.725	0.921	0.846	0.869
	51111	0.997	0.797	1	1	0.810	1	1	1
	53111	0.996	0.760	0.998	1	0.786	1	0.999	1
$g = 0$, $h = 0.4$	11111	0	0	0.001	0.010	0.001	0.001	0.010	0.014
	31111	0.050	0.008	0.073	0.294	0.056	0.116	0.191	0.271
	32111	0.040	0.005	0.049	0.262	0.025	0.137	0.155	0.204
	51111	0.125	0.029	0.175	0.600	0.120	0.352	0.390	0.587
	53111	0.088	0.017	0.121	0.532	0.050	0.277	0.327	0.481
$g = 0$, $h = 0.8$	11111	0	0	0.002	0.013	0.002	0.002	0.013	0.018
	31111	0.009	0.001	0.009	0.103	0.019	0.046	0.073	0.101
	32111	0.012	0.001	0.008	0.097	0.009	0.041	0.064	0.069
	51111	0.028	0.003	0.027	0.206	0.025	0.087	0.147	0.219
	53111	0.024	0.001	0.018	0.185	0.018	0.070	0.114	0.155
$g = 0.8$, $h = 0$	11111	0.006	0.001	0.003	0.008	0.003	0.003	0.006	0.006
	31111	0.335	0.087	0.436	0.724	0.265	0.284	0.469	0.577
	32111	0.228	0.056	0.309	0.675	0.211	0.548	0.362	0.482
	51111	0.611	0.284	0.745	0.957	0.521	0.908	0.850	0.858
	53111	0.480	0.160	0.597	0.948	0.328	0.817	0.766	0.864
$g = 0.8$, $h = 0.4$	11111	0.002	0	0.004	0.009	0.003	0.002	0.011	0.010
	31111	0.028	0.004	0.042	0.224	0.037	0.133	0.151	0.210
	32111	0.024	0.001	0.027	0.196	0.023	0.105	0.110	0.147
	51111	0.078	0.018	0.118	0.472	0.077	0.266	0.301	0.478
	53111	0.063	0.006	0.077	0.410	0.043	0.206	0.243	0.355

The power advantage of *RMDB* over the other methods was often quite substantial, especially for the (σ,(σ+1)/2,1,1,1)(σ,(σ+1)/2,1,1,1) pattern. *RMD* was usually next most-powerful, although *W50B* was occasionally next most-powerful, especially for five groups. In some cases for normal and the skewed light-tailed distribution, *RMDB* was not most powerful, especially for smaller variance ratios, but the power disparity was always very small.

For the reverse cases where the largest sample size was associated with the smallest scale parameter and the smallest sample size with the largest scale parameter, *W50B* always had the highest power, with *RMD* or *RMDB* next most powerful.

5.3. All-pairs power

All-pairs power – the probability of detecting all scale differences – was estimated for all cases. All-pairs power is most interesting in cases where more than one scale parameter differ, since in the single extreme scale case only one scale difference magnitude exists among unequal scale parameters. There was generally little or no all-pairs power for the heavier-tailed distributions, especially for five groups, and thus those results are not reported here. However, the same relative power patterns emerged as for any-pair power, with *RMD*/*RMDB* or *W50*/*W50B* with the highest power.

Table 5. Proportion of at least one rejection at $\alpha = 0.05$, three treatments, unequal sample sizes, random n.

$n_1, n_2, n_3 = 10,20,30$	Method								
Distribution	$\sigma_1\ \sigma_2\ \sigma_3$	***LEV***	***LEVB***	***W50***	***W50B***	***OB***	***OB50***	***RMD***	***RMDB***
$g = 0, h = 0$	311	0.903	0.878	0.893	0.883	0.923	0.892	0.831	0.724
	321	0.832	0.957	0.773	0.965	0.733	0.694	0.835	0.845
	511	0.989	0.984	0.991	0.989	0.977	0.948	0.988	0.965
	531	0.985	1	0.951	1	0.908	0.885	0.995	0.998
$g = 0, h = 0.4$	311	0.246	0.332	0.257	0.376	0.139	0.140	0.224	0.226
	321	0.148	0.295	0.139	0.338	0.083	0.090	0.202	0.202
	511	0.480	0.614	0.529	0.687	0.307	0.295	0.448	0.464
	531	0.269	0.533	0.248	0.596	0.134	0.124	0.432	0.410
$g = 0, h = 0.8$	311	0.097	0.154	0.094	0.173	0.053	0.060	0.075	0.071
	321	0.081	0.143	0.061	0.164	0.037	0.043	0.075	0.069
	511	0.175	0.274	0.161	0.313	0.086	0.093	0.136	0.152
	531	0.110	0.227	0.096	0.259	0.054	0.060	0.141	0.119
$g = 0.8, h = 0$	311	0.578	0.644	0.632	0.675	0.590	0.469	0.483	0.426
	321	0.398	0.616	0.384	0.662	0.357	0.260	0.442	0.429
	511	0.884	0.915	0.907	0.926	0.850	0.747	0.850	0.802
	531	0.660	0.900	0.636	0.936	0.502	0.393	0.848	0.819
$g = 0.8, h = 0.4$	311	0.218	0.291	0.212	0.322	0.125	0.129	0.182	0.171
	321	0.140	0.244	0.117	0.272	0.076	0.074	0.153	0.148
	511	0.338	0.521	0.389	0.563	0.236	0.228	0.355	0.372
	531	0.228	0.427	0.209	0.487	0.112	0.112	0.328	0.298
$n_1, n_2, n_3 = 30,20,10$									
$g = 0, h = 0$	311	0.967	0.988	0.971	0.992	0.644	0.575	0.989	0.998
	321	0.882	0.769	0.893	0.78	0.521	0.453	0.939	0.891
	511	0.999	1	1	1	0.774	0.721	1	1
	531	0.996	0.979	0.999	0.988	0.738	0.678	1.000	0.999
$g = 0, h = 0.4$	311	0.040	0.085	0.073	0.128	0.002	0.001	0.408	0.521
	321	0.014	0.031	0.026	0.041	0.002	0.001	0.359	0.344
	511	0.097	0.217	0.173	0.313	0.006	0.004	0.717	0.834
	531	0.047	0.060	0.084	0.084	0.001	0	0.597	0.615
$g = 0, h = 0.8$	311	0.002	0.008	0.006	0.013	0.001	0.001	0.217	0.268
	321	0.007	0.013	0.004	0.014	0.001	0.002	0.214	0.203
	511	0.003	0.014	0.008	0.023	0	0.001	0.383	0.469
	531	0.006	0.010	0.004	0.011	0.001	0.001	0.359	0.331
$g = 0.8, h = 0$	311	0.391	0.504	0.500	0.610	0.079	0.048	0.745	0.843
	321	0.245	0.214	0.317	0.267	0.038	0.021	0.597	0.602
	511	0.597	0.773	0.741	0.918	0.155	0.082	0.975	0.995
	531	0.441	0.431	0.573	0.520	0.080	0.036	0.896	0.898
$g = 0.8, h = 0.4$	311	0.020	0.041	0.031	0.068	0.001	0.001	0.362	0.430
	321	0.013	0.023	0.013	0.028	0.001	0.001	0.315	0.299
	511	0.053	0.110	0.089	0.193	0.001	0.001	0.594	0.755
	531	0.027	0.039	0.034	0.051	0	0	0.520	0.545

6. Discussion

The power and Type I error rates of several permutation multiple comparison procedures based on deviances, or absolute median differences, were investigated. The procedures based on deviances had the highest power in virtually all cases considered, and using the Bonferroni

correction usually resulted in a higher power than using the Tukey-type adjustment. In most cases, the power difference between *RMDB/RMD* and *W50B/W50* was not large. However, exceptions were the three group, unequal sample size cases (see Tables 3 and 5), where *RMDB/RMD* had substantial power advantages. Since the *RMDB* statistic, as a ratio of average deviances, has intuitive appeal as a more robust analogue to the parametric variance ratio F-test, and since using the Bonferroni correction will be easier to implement than the Tukey-type adjustment, we recommend using the *RMDB* statistic in general as an alternative to parametric tests for pairwise comparison of scale.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

1. Marozzi M. Multivariate tests based on interpoint distances with application to magnetic resonance imaging. Stat Methods Med Res. 2016;25(6):2593–2610. doi: 10.1177/0962280214529104
2. Bartlett MS. Properties of sufficiency and statistical tests. Proc R Soc A. 1937;160(901):268–282.
3. Cochran WG. Problems arising in the analysis of a series of similar experiments. J R Statist Soc. 1937;4:102–118.
4. Hartley HO. The use of range in analysis of variance. Biometrika. 1950;37:271–280. doi: 10.1093/biomet/37.3-4.271
5. Sharma D, Kibria BM. On some test statistics for testing homogeneity of variances: a comparative study. J Stat Comput Simul. 2013;83(10):1944–1963. doi: 10.1080/00949655.2012.675336
6. Levene H. Robust tests for equality of variances. In: I Olkin, H Hotelling, editors. Contributions to probability and statistics. Palo Alto, CA: Stanford University Press; 1960. p. 278–292.
7. Brown MB, Forsythe AB. Robust tests for the equality of variances. J Am Stat Assoc. 1974;69:364–367. doi: 10.1080/01621459.1974.10482955
8. Keselman HJ, Games PA, Clinch JJ. Tests for homogeneity of variance. Commun Stat – Simul Comput. 1979;8:113–129. doi: 10.1080/03610917908812108
9. Conover WJ, Johnson ME, Johnson MM. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics. 1981;23:351–361. doi: 10.1080/00401706.1981.10487680
10. Balakrishnan N, Ma CW. A comparative study of various tests for the equality of two population variances. J Stat Comput Simul. 1990;35:41–89. doi: 10.1080/00949659008811234
11. O'Brien RG. A general ANOVA method for robust tests of additive models for variances. J Am Stat Assoc. 1979;74:877–880. doi: 10.1080/01621459.1979.10481047

12. Olejnik SF, Algina J. Tests of variance equality when distributions differ in form and location. Educ Psychol Meas. 1988;48:317–329. doi: 10.1177/0013164488482005

13. Marozzi M. Levene type tests for the ratio of two scales. J Stat Comput Simul. 2011;81(7):815–826. doi: 10.1080/00949650903499321

14. Richter SJ, McCann MH. Permutation tests of scale using deviances. Commun Stat – Simul Comput. 2017;46(7):5553–5565. doi: 10.1080/03610918.2016.1165844

15. Higgins JJ. Introduction to modern nonparametric statistics. Pacific Grove (CA): Duxbury; 2004.

16. Richter SJ, McCann MH. Multiple comparison of medians using permutation tests. J Mod Appl Stat Methods. 2007;6(2):399–412. doi: 10.22237/jmasm/1193889900

17. Hoaglin DC. Summarizing shape numerically: The g-and-h distributions. In: DC Hoaglin, F Mosteller, JW Tukey, editors. Exploring data tables, trends, and shapes. Hoboken, NJ: John Wiley & Sons, Inc; 1985.

18. Keller-McNulty S, Higgins JJ. Effect of tail weight and outliers on power and type-I error of robust permutation tests for location. Commun Stat – Simul Comput. 1987;16(1):17–35. doi: 10.1080/03610918708812575