Biometrical Journal **44** (2002) 4, 466–477

Multiple Comparisons Among Dependent Groups Based on a Modified One-Step M-Estimator

RAND R. WILCOX

Dept. of Psychology
University of Southern California
U.S.A.

Summary

Currently, among multiple comparison procedures for dependent groups, a bootstrap-t with a 20% trimmed mean performs relatively well in terms of both Type I error probabilities and power. However, trimmed means suffer from two general concerns described in the paper. Robust M-estimators address these concerns, but now no method has been found that gives good control over the probability of a Type I error when sample sizes are small. The paper suggests using instead a modified one-step M-estimator that retains the advantages of both trimmed means and robust M-estimators. Yet another concern is that the more successful methods for trimmed means can be too conservative in terms of Type I errors. Two methods for performing all pairwise multiple comparisons are considered. In simulations, both methods avoid a familywise error (FWE) rate larger than the nominal level. The method based on comparing measures of location associated with the marginal distributions can have an actual FWE that is well below the nominal level when variables are highly correlated. However, the method based on difference scores performs reasonably well with very small sample sizes, and it generally performs better than any of the methods studied in WILCOX (1997*b*).

Key words: M-estimators; Asymmetric trimming; Bootstrap; Depth.

1. Introduction

Currently, when comparing dependent groups based on some measure of location, a relatively successful approach is to use a bootstrap-t method in conjunction with a 20% trimmed mean. Probability coverage has been found to be relatively accurate when sample sizes are small – even in situations where methods based on means are known to be unsatisfactory – power competes well with methods based on means when sampling from a normal distribution, and power can be much higher than methods based on means under slight departures from normality (e.g., WILCOX, 1997*a*).

However, 20% trimmed means in particular, and trimmed means in general, suffer from at least two practical concerns. First, the amount of trimming is assumed to be fixed in advance. If the amount of trimming is set at 20%, efficiency is reasonably good versus the mean under normality, but when sampling from a

 0323-3847/02/0406-0466 $ 17.50+.50/0

sufficiently heavy-tailed distribution, efficiency can be poor versus using more trimming or switching to some robust M-estimator of location. A second general concern is that typically trimmed means assume symmetric trimming. That is, the same proportion of observations are trimmed from both tails of an empirical distribution. When sampling from a reasonably symmetric distribution, symmetric trimming seems appropriate, but asymmetric trimming seems more appropriate as the degree of skewness increases. Well known theoretical results indicate how to estimate the standard error of a trimmed mean when asymmetric trimming is used (e.g., HUBER, 1981), but now unsatisfactory probability coverage can result when sample sizes are small (e.g., WILCOX, 1997*a*). Also, if the amount of trimming is empirically determined, and the standard error is estimated by conditioning on this amount of trimming, even poorer control over probability coverage can result.

Robust M-estimators are more flexible in the sense that they empirically determine whether a value is unusually large or small and then such values are down weighted in some manner. Letting $X_1, \ldots, X_n$ be a random sample, a well-known example is the one-step M-estimator based on Huber's Ψ:

$$\frac{1.28(\text{MADN})\,(i_2 - i_1) + \sum_{i=i_1+1}^{n-i_2} X_{(i)}}{n - i_1 - i_2}\,, \tag{1}$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the values written in ascending order, M is the usual median, MAD is the median of the values $|X_1 - M|, \ldots, |X_n - M|$, MADN = MAD/.6745, i_1 is the number of observations X_i such that $(X_i - M) < -K(\text{MADN})$ and i_2 is the number of observations X_i such that $(X_i - M) > K(\text{MADN})$. So basically this estimator empirically determines whether an observation is an outlier, trims it, averages the values that remain, but with asymmetric trimming an adjustment is made based on a measure of scale, MAD. The adjustment based on MAD is a consequence of how the population value of the one-step M-estimator is defined. It is the value θ satisfying

$$E\left[\Psi\left(\frac{X - \theta}{\text{MADN}}\right)\right] = 0\,, \tag{2}$$

where $\Psi(x) = \max\,[-K, \min\,(K, x)]$. Equation (2) can be solved with the Newton-Raphson method and a single iteration of this technique yields (with $K = 1.28$) equation (1). The choice $K = 1.28$ provides good efficiency under normality and its finite sample breakdown point is .5, the highest possible value. (The finite sample breakdown point of an estimator is the smallest proportion of observations, which when altered, can drive the value of an estimator to plus or minus infinity.) However, when performing all pairwise comparisons among J dependent groups based on this one-step M-estimator, none of the techniques examined by WILCOX (1997*b*) performed well in simulations. Moreover, situations arise where even the most successful method can have Type I error probabilities well below the nominal level.

The goal in this paper is to suggest that a modified one-step M-estimator (MOM) be used instead of (1), which performs much better in simulations, in terms of probability coverage or controlling the probability of a Type I error. The MOM estimator belongs to the class of skipped estimators originally proposed by Tukey and studied by Andrews BICKEL, HAMPEL, HUBER, ROGERS and TUKEY (1972). The basic idea is simple: Check for outliers, discard any that are found, and then average the values that remain. But the class of skipped estimators they studied is based on a boxplot outlier detection rule which has a finite sample breakdown point of only .25. Here an outlier detection rule based on M and MADN is used instead resulting in a location estimator having a finite sample breakdown point of .5 as well. (HUBER, 1993, argues that at a minimum, an estimator should have a finite sample breakdown point of at least .1.) A practical disadvantage of skipped estimators is that expressions for their standard errors are very complicated when sampling from an asymmetric distribution. One of the main points in this paper is that a variation of the percentile bootstrap method not only circumvents this problem, it provides good probability coverage in simulations where no effective method based on a robust M-estimator has been found. Moreover, the choice between comparing measures of location associated with the marginal distributions, versus using a measure of location based on difference scores, can make a practical difference in the probability of a Type I error.

2. A Modified One-Step M-estimator

The proposed modified one-step M-estimator (MOM) begins by declaring X_i an outlier if

$$\frac{.6745|X_i - M|}{\text{MAD}} > K\,,$$

where K is adjusted so that efficiency is good under normality. Then MOM is given by

$$\hat{\theta} = \sum_{i=i_1+1}^{n-i_2} \frac{X_{(i)}}{n - i_1 - i_2}\,, \tag{3}$$

where now $i_1(i_2)$ is the number of observations less (greater) than the median that are declared outliers. In this paper the main concern is with small sample sizes, so K is chosen so that efficiency is reasonably good under normality for $n \leq 100$. In particular, using simulations with 10,000 replications, it was found that with $K = 2.24$, the standard error of the sample mean divided by the standard error of $\hat{\theta}$ is approximately .9 for $n = 20(5)100$. For $n = 10$ and 15, this ratio is .88. So in effect, simply drop the term containing MAD in equation (1) and adjust K to retain good efficiency under normality. (Note that 2.24 is approximately equal to the square root of the .975 quantile of a chi-square distribution with one degree of

freedom and that the outlier detection rule given used here is basically a special case of the general method suggested by ROUSSEEUW and VAN ZOMEREN, 1990.)

It is noted that the population analog of $\hat{\theta}$ is given by

$$M(X) = \frac{1}{\xi} \int_{\eta - K\tau}^{\eta + K\tau} x dF(x) .$$

where for any function of X, $U(X)$,

$$M(U(X)) = \frac{1}{\xi} \int_{\eta - K\tau}^{\eta + K\tau} U(x)\, dF(x) ,$$

η is the population median, τ is the population value estimated by MADN, and

$$\xi = \int_{\eta - K\tau}^{\eta + K\tau} dF(x) .$$

Note that θ is scale equivariant and location invariant. That is, for constants a and b, $a + bX$ has a population MOM of $a + b\theta$.

3. Pairwise Comparisons

Several methods were considered for both multiple comparisons and an omnibus test for equal measures of location. This section describes the percentile bootstrap methods for performing all pairwise comparisons that performed reasonably well in simulations.

Let X_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, J$ be a random sample of n observations from some J-variate distribution. The first approach is based on testing

$$H_0 : \theta_j = \theta_k , \tag{4}$$

for each $j < k$ such that the familywise error rate (FWE) is α. (That is, the probability of at least one Type I error is to be α.) The second strategy uses difference scores instead. That is, let θ_{djk} be the population value of MOM corresponding to $X_{ij} - X_{ik}$. Then the goal is to test

$$H_0 : \theta_{djk} = 0 , \tag{5}$$

for each $j < k$. As is evident, (4) and (5) are equivalent when distributions are identical, but it is readily verified that this is not necessarily true otherwise, and it will be illustrated that the choice between these two approaches can lead to different conclusions in applied work.

First consider testing (5) based on the difference scores. For the special case $J = 2$, the strategy is to use a percentile bootstrap method that is based on the general results in LIU and SINGH (1997) and then extend the method to $J > 2$

using the sequentially rejective method derived by ROM (1990). More precisely, for the $L = (J^2 - J)/2$ pairwise comparisons to be made, denote the collection of all pairwise differences by

$$\begin{aligned} D_{i1} &= X_{i1} - X_{i2} \\ D_{i2} &= X_{i2} - X_{i3} \\ &\vdots \\ D_{iL} &= X_{i,J-1} - X_{iJ} \end{aligned}$$

Next, generate a bootstrap sample by randomly resampling with replacement n rows from

$$\begin{pmatrix} D_{11}, \ldots, D_{1,L} \\ \vdots \\ D_{n1}, \ldots, D_{n,L} \end{pmatrix}$$

yielding

$$\begin{pmatrix} D^*_{11}, \ldots, D^*_{1,L} \\ \vdots \\ D^*_{n1}, \ldots, D^*_{n,L} \end{pmatrix}.$$

Let $\hat{\theta}^*_\ell$ be the value of $\hat{\theta}$ applied to the ℓth column of the D^* values. For any ℓ, let

$$p^*_\ell = P(\hat{\theta}^*_\ell > 0).$$

That is, p^*_ℓ is the probability that a bootstrap estimate based on $D^*_{1\ell}, \ldots, D^*_{n\ell}$ is greater than zero. From LIU and SINGH (1997), if (5) is true, then p^*_ℓ has, asymptotically, a uniform distribution (cf. HALL, 1986). So if p^*_ℓ is estimated to be less than or equal to $\alpha/2$ or greater than $1 - \alpha/2$, reject $H_0: \theta_{d\ell} = 0$.

Here, an estimate of p^*_ℓ is obtained simply by repeating the process just described B times yielding $\hat{\theta}^*_{\ell b}$, $b = 1, \ldots, B$; $\ell = 1, \ldots, L$. With $J \leq 4$, $B = 1,000$ is used and with $5 \leq J \leq 10$, B = 2000 is used. Let $I_{\ell b} = 1$ if $\hat{\theta}^*_{\ell b} > 0$, otherwise $I_{\ell b} = 0$. Then an estimate of $p^*_{\ell b}$ is

$$\hat{p}^*_\ell = \frac{1}{B} \sum_{b=1}^{B} I_{\ell b},$$

($\ell = 1, \ldots, L$). So for fixed ℓ, reject $H_0: \theta_{d\ell} = 0$ if $\hat{p}^*_\ell \leq \alpha/2$ or if $\hat{p}^*_\ell \geq 1 - \alpha/2$.

There remains the problem of controlling FWE. This problem is approached by using the sequentially rejective method derived by ROM (1990). In particular, let

$$\hat{p}^*_{m\ell} = \min(\hat{p}^*_\ell, 1 - \hat{p}^*_\ell)$$

and let $\hat{p}^*_{m[L]} \geq \ldots \geq \hat{p}^*_{m[1]}$ be the $\hat{p}^*_{m\ell}$ values written in descending order. Then reject $H_0: \theta_{d\ell} = 0$ if for any $\hat{p}^*_{m[c]} \geq \hat{p}^*_{m[\ell]}$,

$$2\hat{p}^*_{m[c]} < d_c, \tag{6}$$

where d_c is read from Table 1 which was taken from ROM (1990). For other values of α or $c > 10$, HOCHBERG's (1988) method can be used where $d_c = \alpha/c$. This will be called method D.

Note that unlike non-bootstrap sequentially rejective methods, confidence intervals for all pairs can be computed such that, asymptotically, the simultaneous probability coverage is $\geq 1 - \alpha$. In particular, when comparing the ℓth pair of groups with critical value d_c, set $t = [d_c B]$, where $[\,.\,]$ is the greatest integer function, let $u = B - t$, in which case the confidence interval for $\theta_{d\ell}$ is $(\hat{\theta}^*_{\ell(t+1)}, \hat{\theta}^*_{\ell(u)})$, where $\hat{\theta}^*_{\ell(1)} \leq \ldots \leq \hat{\theta}^*_{\ell(B)}$ are the B bootstrap estimates of the ℓth difference written in ascending order.

Table 1
Values of d_ℓ for $\alpha = .05$ and .01

ℓ	$\alpha = .05$	$\alpha = .01$
1	.05000	.01000
2	.02500	.00500
3	.01690	.00334
4	.01270	.00251
5	.01020	.00201
6	.00851	.00167
7	.00730	.00143
8	.00639	.00126
9	.00568	.00112
10	.00511	.00101

Now consider testing H_0 given by equation (4). Then bootstrap samples are obtained by resampling with replacement n rows from the n by J matrix of X_{ij} values. Repeat this process B times and let $\hat{\theta}^*_{bj}$ be the bootstrap value of MOM based on the jth group and bth bootstrap sample, $b = 1, \ldots, B$; $j = 1, \ldots, J$. Let

$$p^*_{jk} = P(\hat{\theta}^*_j > \hat{\theta}^*_k)$$

based on a random bootstrap sample. Here this probability is estimated in the obvious way: the proportion of bootstrap samples having $\hat{\theta}^*_{bj} > \hat{\theta}^*_{bk}$. Again for convenience, set

$$\hat{p}^*_{mjk} = \min\left(\hat{p}^*_{jk}, 1 - \hat{p}^*_{mjk}\right).$$

Then from general results by HALL (1988), if H_0 is true, $\hat{p}^*_{mjk}$ has, asymptotically, a uniform distribution, so again reject if $\hat{p}^*_{mjk} \leq \alpha/2$. To control FWE, again Rom's method is applied. This will be called method M.

4. A Comment on Omnibus Tests

It is evident that the results in LIU and SINGH (1997) can be used to test the hypothesis that for C linear contrasts, all C linear contrasts are equal to zero. The

goal in this section to comment briefly about a practical issue related to one of the variations considered in their paper when comparing groups based on MOM. The point is that when using the Mahalanobis measure of depth, applied to the bootstrap values, the choice of a center for these bootstrap values can make a substantial difference regarding control over the probability of a Type I error.

To elaborate, suppose $J = 4$, let

$$\Psi_c = \theta_c - \theta_{c+1}$$

and consider the problem of testing

$$H_0 : \Psi_1 = \Psi_2 = \Psi_3 = 0 .$$

Generate bootstrap values as before and let

$$\hat{\Psi}^*_{cb} = \hat{\theta}^*_{cb} - \hat{\theta}^*_{c+1,b} ,$$

($c = 1, 2, 3$; $b = 1, \ldots, B$). The basic idea is to determine how deeply $\mathbf{0} = (0, 0, 0)$ is nested within the cloud of bootstrap values. Suppose we measure the depth of $\mathbf{0}$ with a Mahalanobis distance. A natural variation of this approach is to let

$$\bar{\Psi}_c = \frac{1}{B} \sum \hat{\Psi}^*_{cb} ,$$

$$s_{ck} = \frac{1}{B-1} \sum (\hat{\Psi}^*_{cb} - \bar{\Psi}_c)(\hat{\Psi}^*_{kb} - \bar{\Psi}_k) ,$$

in which case the depth of $\mathbf{0}$ is

$$T = (\mathbf{0} - \bar{\mathbf{\Psi}})\, \mathbf{S}^{-1} (\mathbf{0} - \bar{\mathbf{\Psi}})' .$$

Let

$$G_b = (\hat{\mathbf{\Psi}}_b - \bar{\mathbf{\Psi}})\, \mathbf{S}^{-1} (\hat{\mathbf{\Psi}}_b - \bar{\mathbf{\Psi}})' .$$

From LIU and SINGH (1997), reject H_0 if $T \geq G_{(u)}$, where $G_{(1)} \leq \cdots \leq G_{(B)}$ are the G_b values written in ascending order and $u = (1 - \alpha) B$, rounded to the nearest integer.

But notice that when generating bootstrap values, the mean of these values (when $B = \infty$) is known and simply $\hat{\mathbf{\Psi}} = (\hat{\Psi}_1, \hat{\Psi}_2, \hat{\Psi}_3)$. So rather than use $\boldsymbol{S}$ and T as defined above, let

$$v_{ck} = \frac{1}{B-1} \sum (\Psi^*_{cb} - \hat{\Psi}_c)(\Psi^*_{kb} - \hat{\Psi}_k) ,$$

$$H_b = (\hat{\mathbf{\Psi}}_b - \hat{\mathbf{\Psi}})\, \boldsymbol{V}^{-1} (\hat{\mathbf{\Psi}}_b - \hat{\mathbf{\Psi}})' ,$$

and now reject if $F \geq H_{(u)}$, where

$$F = (\mathbf{0} - \hat{\mathbf{\Psi}})\, \boldsymbol{U}^{-1} (\mathbf{0} - \hat{\mathbf{\Psi}})' .$$

Here is the point. Among the situations considered in Section 5, it was found that F resulted in Type I error probabilities less than or equal to the nominal .05

level, but that T had Type I error probabilities exceeding .15 in some situations, so the choice of the center of the bootstrap values makes a practical difference.

5. Design of the Simulation Study

The small-sample properties of methods M and D were studied for $J = 4$ with simulations where observations were generated from a multivariate normal distribution via the IMSL (1987) subroutine RNMVN. Nonnormal distributions were generated using the g-and-h distribution (HOAGLIN, 1985). That is, first generate Z_{ij} from a multivariate normal distribution and set

$$X_{ij} = \frac{\exp(gZ_{ij}) - 1}{g} \exp(hZ_{ij}^2/2).$$

For $g = 0$ this last expression is taken to be

$$X_{ij} = Z_{ij} \exp(hZ_{ij}^2/2).$$

The case $g = h = 0$ corresponds to a normal distribution. Setting $g = 0$ yields a symmetric distribution, and as g increases, skewness increases as well. Heavy-tailedness increases with h. The values for g and h were taken to be $(g, h) = (0, 0)$, $(0, .5)$, $(.5, 0)$ and $(.5, .5)$. Table 2 contains skewness (κ_1) and kurtosis (κ_2) values for the four g-and-h distributions used in the simulations.

Table 2
Some properties of the g-and-h distribution

g	h	κ_1	κ_2	$\hat{\kappa}_1$	$\hat{\gamma}_2$
0.0	0.0	0.00	3.00	0.00	3.0
0.0	0.5	0.00	–	0.00	11,896.2
0.5	0.0	1.75	8.9	1.81	9.7
0.5	0.5	–	–	120.10	18,393.6

When $h > 1/k$, $E(X - \mu)^k$ is not defined and the corresponding entry in Table 2 is left blank. A possible criticism of simulations performed on a computer is that observations are generated from a finite interval, so the moments are finite even when in theory they are not, in which case observations are not being generated from a distribution having the theoretical skewness and kurtosis values listed in Table 2. In fact, as h gets large, there is an increasing difference between the theoretical and actual values for skewness and kurtosis. Accordingly, Table 2 also lists the estimated skewness ($\hat{\kappa}_1$) and kurtosis ($\hat{\kappa}_2$) values based on 100,000 observations generated from the distribution. Simulations were also run where the marginal distributions are lognormal or exponential.

Simulations were run where the marginal distributions had equal and unequal variances. When working with skewed distributions, the marginal distributions were first shifted so that they have a θ value of zero, and for the unequal variance case the ith observation in the jth group was multiplied by σ_j, $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (4, 1, 1, 1)$. That is, for skewed distributions, before multiplying the X_{ij} by σ_j, the observations were shifted by subtracting the population value of θ so that when multiplying by σ_j, the null hypothesis remains true. The values of θ were determined by computing $\hat{\theta}$ with one million observations generated from the distribution under study. For $(g, h) = (.5, 0)$ and (.5, .5) it was found that $\theta = -.33$ and .25, respectively.

Four patterns of correlations were used. Three of the four correlation matrices have a common correlation, ρ, with $\rho = .1, .5$ and .8. The fourth correlation matrix had $\rho_{12} = .8$, $\rho_{13} = .5$, $\rho_{14} = .2$, $\rho_{23} = .5$, $\rho_{24} = .2$ and $\rho_{34} = .2$. These four correlation matrices are labeled C1, C2, C3 and C4, respectively.

WILCOX (1997*b*) notes that even with $n = 21$, when using the one-step M-estimator given by equation (1), the actual probability of at least one Type I error, when testing at the .05 level, can exceed .08. Moreover, with $n \leq 15$, situations arise where the bootstrap value of the one-step M-estimator cannot be computed due to division by zero. For this reason, simulation results based on this one-step M-estimator are not given.

6. Results

Table 3 contains the estimated Type I error probabilities when using the multiple comparison procedure described in Section 3 with marginal distributions having a common variance. The results are based on 2,000 replications. For unequal variances, the estimates changed by only a few units in the third decimal place, so for brevity they are not reported. The most successful method found in WILCOX 1997*b*), based on controlling the probability of at least one Type I error, was a bootstrap-*t* method with 20% trimmed means. For comparative purposes, results on this method are included and reported in Table 3 under the column headed by B. Both of the multiple comparison methods in section 3 have estimated Type I error probabilities less than or equal to .05 in all situations considered. The main difference is that when using method M, situations arise where the estimated Type I error probability drops below .01, while for method D, estimated Type I error probabilities are very stable and range between .023 and .035.

As for power, methods M and D test different hypotheses in some situations, so in some sense power comparisons are meaningless. However, there are some indications that method D will typically have more power. One is that in some instances, method M becomes extremely conservative in terms of Type I errors. But even when the two methods have comparable Type I error probabilities, there are

Table 3
Estimated Type I error probabilities for g-and-h distributions, $n = 11$, $J = 4$

g	h	Correlation	$\boldsymbol{\sigma}$	Method		
				M	D	B
0.0	0.0	C1	$(1, 1, 1, 1)$	.038	.035	.036
		C2		.025	.030	.038
		C3		.005	.024	.027
		C4		.009	.024	.014
0.0	0.5	C1	$(1, 1, 1, 1)$	.026	.032	.026
		C2		.016	.022	.019
		C3		.004	.029	.036
		C4		.007	.029	.010
0.5	0.0	C1	$(1, 1, 1, 1)$	.034	.029	.031
		C2		.022	.023	.022
		C3		.010	.034	.035
		C4		.009	.025	.009
0.5	0.5	C1	$(1, 1, 1, 1)$	.022	.026	.023
		C2		.011	.028	.015
		C3		.003	.032	.033
		C4		.008	.023	.009

indications that method D will have more power. For example, when sampling from normal distributions, with the correlation matrix C1, both methods have very similar Type I error probabilities. If $\delta = 1$ is added to every observation in the first group, the probability of one or more rejections is .56 when using method D. However, for method M, it is only .43.

7. An Illustration

It is illustrated that the choice between methods D and M can make a practical difference in the conclusions reached. RAO (1948) presented data on the weight of cork borings from the north, east, west and south sides of 28 trees. Here attention is focused on comparing the typical weights associated with the first two sides. First consider method M where the goal is to compare θ_1 to θ_2. Figure 1 shows a plot of the bootstrap estimates where the polygon is the convex hull of the central 95% of the bootstrap samples based on their Mahalanobis depth, which provides an approximate confidence region for (θ_1, θ_2). If $H_0 : \theta_1 = \theta_2$ is true, then (θ_1, θ_2) must fall on the line shown in Figure 1 which has slope one and intercept zero. Note that a portion of this line lies within the approximate .95 confidence region. The p-value is .18. In contrast, if method D based on difference scores is used instead, the significance level is .026.

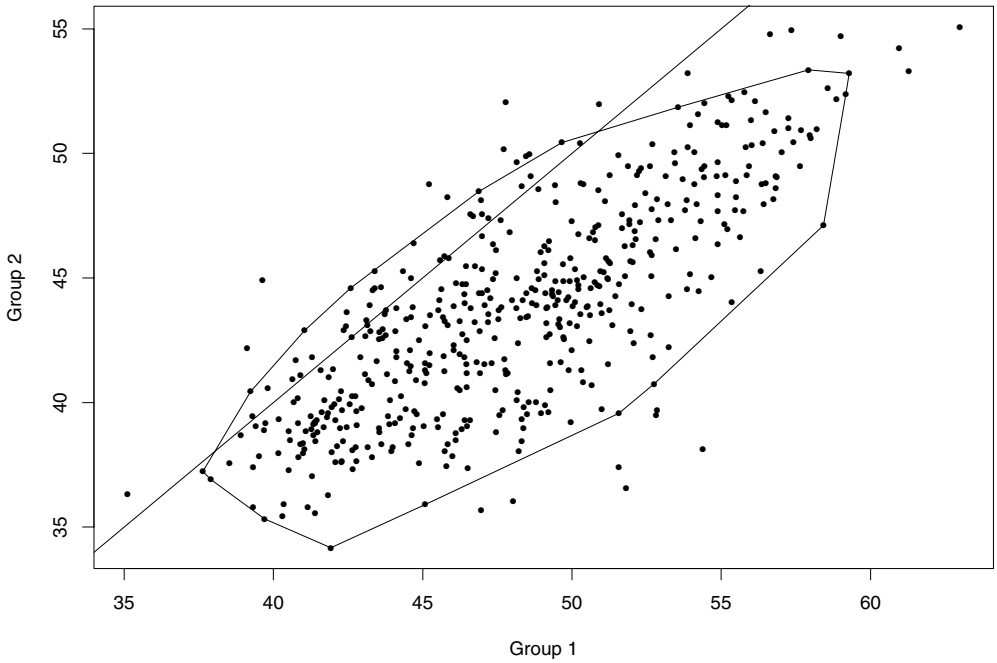

Fig. 1. A plot of the bootstrap estimates of MOM for the north and east sides of the trees. The polygon is the convex hull of the central 95% of the bootstrap values; it provides an approximate .95 confidence region for the population values. If the null hypothesis is true, the population values must lie somewhere on the straight line.

8. Concluding Remarks

Why is method M too conservative in terms of the probability of a Type I error? It might seem that this has to do with the conservative nature of Rom's method, but even when comparing only two groups, method M becomes too conservative as the correlation between the random variables gets close to one. The exact reason for this remains unclear. In contrast, method D is fairly stable in terms of Type I errors as the correlation increases. Also, compared to the methods studied by WILCOX (1997*b*) for means, trimmed means and an M-estimator based on Huber's Ψ, method D is much more satisfactory in terms of Type I errors, even with the smaller sample size considered here. Moreover, when using an M-estimator, no method has been found to be reasonably satisfactory when sample sizes are small.

Of course, it is not being suggested that comparing groups based on MOM is optimal in all situations. Given the goal of relatively high power, for example, the best method will depend on how the groups differ, which of course is unknown. The main advantages of methods M and D are that, unlike M-estimators, they control Type I error probabilities for the situations considered in Section 5. More-

over, MOM provides a certain amount of flexibility not enjoyed by trimmed means in general, and the mean and median in particular. So as a general tool for comparing dependent groups based on some measure of location, methods D and M appear to have practical value. There are indications that method D will often have more power than method M, but more experience with actual data is needed to resolve this issue.

References

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H., and TUKEY, J. W., 1972: *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, NJ.

HALL, P., 1986: On the bootstrap and confidence intervals. *Annals of Statistics* **14**, 1431–1452.

HOAGLIN, D. C., 1985: Summarizing shape numerically: The *g*-and-*h* distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Exploring data tables, trends, and shapes.* (pp. 461–515). New York: Wiley.

HOCHBERG, Y., 1988: A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802.

HUBER, P. J., 1981: *Robust Statistics.* New York: Wiley.

HUBER, P., 1993: Projection pursuit and robustness. In S. Morgenthaler, E. Ronchetti & W. Stahel (Eds.) *New Directions in Statistical Data Analysis and Robustness*. Boston: Birkhäuser Verlag.

IMSL, 1987: Library I, vol. II Houston: International Mathematical and Statistical Libraries.

LIU, R. Y. and SINGH, K., 1997: Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association* **92**, 266–277.

RAO, C. R., 1948: Tests of significance in multivariate analysis. *Biometrika* **35**, 58–79.

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

ROUSSEEUW, P. J., and VAN ZOMEREN, B. C., 1990: Unmasking multivariate outliers and leverage points (with discussion). *Journal of the American Statistical Association* **85**, 633–639.

STAUDTE, R. G., and SHEATHER, S. J., 1990: *Robust Estimation and Testing* New York: Wiley.

WILCOX, R. R., 1997a: *Introduction to Robust Estimation and Hypothesis Testing.* San Diego, CA: Academic Press.

WILCOX, R. R., 1997b: Pairwise comparisons using trimmed means or M-estimators when working with dependent groups. *Biometrical Journal* **39**, 677–688.

Received, April 2001
Revised, October 2001
Accepted, January 2002