WITHIN GROUPS ANOVA WHEN USING A ROBUST MULTIVARIATE MEASURE OF LOCATION

Rand R. Wilcox & Timothy Hayes
Dept of Psychology
University of Southern California

January 27, 2014
ABSTRACT

Let $\theta_1, \ldots, \theta_J$ be measures of location associated with $J$ dependent groups. Various methods have been proposed that are aimed at testing the global hypothesis $H_0$: $\theta_1 = \cdots = \theta_J$ based on a robust measure of location applied to the marginal distributions. An alternative approach to comparing dependent groups is in terms of difference scores. Let $X_{ij}$ ($i = 1, \ldots, n; j = 1, \ldots, J$) be a random sample of $n$ vectors from some $J$-variate distribution. Let $D_{ijk} = X_{ij} - X_{ik}$, $j < k$, and let $\hat{\Theta}(D)$ be some multivariate location estimator based on the $D_{ijk}$ values. Then another way of comparing the $J$ dependent groups is to test $H_0$: $\delta_1 = \cdots = \delta_L = 0$, where $L = (J^2 - J)/2$ and $\delta_1, \ldots, \delta_L$ are the population parameters being estimated by $\hat{\Theta}(D)$. Methods have been studied based on robust estimators applied to the marginal distributions associated with the $D_{ijk}$ values. A criticism of these methods is that they do not deal with outliers in a manner that takes into account the overall structure of the data. Location estimators have been derived that deal with outliers in this manner, but evidently there are no simulation results regarding how well they perform when the goal is to test the two hypotheses just described. The paper compares four bootstrap methods in terms of their ability to control the Type I error probability when the sample size is small, two of which were found to perform poorly. The choice of location estimator was found to be important as well. Indeed, for several of the estimators considered here, control over the Type I error probability was very poor. Only one estimator performed well when testing the first hypothesis. It is based on a variation of the (affine equivalent) Donoho–Gasko trimmed mean. As for the other hypothesis, only a skipped estimator performed reasonably well. (It removes outliers via a projection method and averages the remaining data.) Only one bootstrap method was found to perform well when testing the first hypothesis. A different bootstrap method is recommended when testing the second hypothesis.

Keywords: Bootstrap methods, outliers, within groups ANOVA, skipped estimator, Donoho–Gasko trimmed mean.
1 Introduction

It is well known that methods for comparing dependent groups, based on the usual sample mean, are not robust under general conditions. A fundamental concern with any inferential technique based on the mean is that it can result in relatively low power when dealing with heavy-tailed distributions (e.g., Marrona et al., 2006; Staudte & Sheather, 1990; Wilcox, 2012). Roughly, heavy-tailed distributions are characterized by outliers that inflate the standard error of the sample mean. Even an arbitrarily small departure from normality can result in poor power. Another concern is that the breakdown point of the sample mean is only $1/n$, where $n$ is the sample size. That is, the minimum proportion of points that must be altered to completely destroy the sample mean (make it arbitrarily large or small) is $1/n$.

Various methods for comparing $J \geq 2$ dependent groups have been derived and studied that are based on replacing the marginal means with some robust estimator (e.g., Wilcox, 2012, Ch. 8). That is, if $X_{ij}$ ($i = 1, \ldots, n; j = 1, \ldots, J$) is a random sample of $n$ vectors from some $J$-variate distribution, for each $j$, a robust measure of location is computed. These methods deal with outliers among the marginal distributions, but they do not deal with outliers in a manner that takes into account the overall structure of the data. As a simple example of what this means, it is not unusual to be young, it is not unusual to have heart disease, but it is very unusual to be both young and have heart disease. Situations are encountered where there are no outliers among the marginal distributions based on, for example, a boxplot or the MAD-median rule, yet there are outliers when using a multivariate outlier detection technique that does take into account the overall structure (e.g., Wilcox, 2012).

Another possible criticism of simply applying a robust estimator to each of the marginal distributions is that the resulting measure of location is not affine equivariant (e.g., Rousseeuw & Leroy, 1986). To elaborate, note that a basic requirement for $\hat{\theta}_j$ to qualify as a location estimator is that it be both scale and location equivariant. That is, if $\hat{\theta}_j = T(X_{1j}, \ldots, X_{nj})$ is some estimate of $\theta_j$, then for $\hat{\theta}_j$ to qualify as a location estimator, it should be the case that for constants $a$ and $b$,

$$T(aX_1 + b, \ldots, aX_n + b) = aT(X_1, \ldots, X_n) + b.$$ 

In the multivariate case, a generalization of this requirement, affine equivariance, is that for
a $J$-by-$J$ nonsingular matrix $\mathbf{A}$ and vector $\mathbf{b}$ having length $J$,

$$
T(X_1\mathbf{A} + \mathbf{b}, \ldots, X_n\mathbf{A} + \mathbf{b}) = T(X_1, \ldots, X_n)\mathbf{A} + \mathbf{b}.
$$

(1)

In particular, the estimate is transformed properly under rotations of the data as well as changes in location and scale.

The goal in this paper is to report simulation results on several methods for comparing dependent groups with an emphasis on situations where the sample size is small. Several multivariate estimators were considered that take into account the overall structure of the data when dealing with outliers. All of them are location and scale equivariant, but one is not affine equivariant.

Here, two types of global hypotheses are considered. To describe them, let $\hat{\Theta}(\mathbf{X})$ represent one of the multivariate location estimators to be considered. Letting $\Theta = (\theta_1, \ldots, \theta_J)$ represent the estimand associated with $\hat{\Theta}$ (the population analog of $\hat{\Theta}$), the first global hypothesis is

$$
H_0 : \theta_1 = \cdots = \theta_J.
$$

(2)

To describe the second hypothesis, let $D_{ijk} = X_{ij} - X_{ik}$, $j < k$ and let $\hat{\Theta}(\mathbf{D})$ be some multivariate location estimator based on the $D_{ijk}$ values. There are $L = (J^2 - J)/2$ parameters being estimated, which are labeled $\Delta = (\delta_1, \ldots, \delta_L)$. Now the goal is to test

$$
H_0 : \delta_1 = \cdots = \delta_L = 0.
$$

(3)

From basic principles, when dealing with means, there is no distinction between (2) and (3). But under general conditions, this is not the case when using a robust estimator. (It is readily verified, for example, that the difference between the marginal medians is not necessarily equal to the median of the difference scores.)

Two bootstrap methods for testing (2) were considered here, and another two methods were considered when testing (3). As will be seen, the choice of estimator, as well as the bootstrap method that is used, is crucial in terms of controlling the Type I error probability, at least when the sample size is small.
2 Description of the Methods

Details of the more successful location estimators, in terms of controlling the Type I error probability, are described in Section 2.1 followed by a list of the estimators that were considered but which performed poorly in simulations. The methods for testing (2) and (3) are described in section 2.2.

2.1 The Location Estimators

The first estimator is based on a particular variation of an affine equivariant estimator derived by Donoho and Gasko (1992), which will be labeled the DG estimator henceforth. Roughly, it begins by quantifying how deeply each point is nested within the cloud of points. Here, this is done using a projection-type method, which provides an approximation of half-space depth (Wilcox, 2012, section 6.2.5). To elaborate, let \( \hat{\tau} \) be some initial affine equivariant location estimator. Here, the minimum covariance determinant estimator (MCD) is used (e.g., Wilcox, 2012, section 6.3.2). Briefly, the MCD estimator searches for a subset of half the data that minimizes the generalized variance. The mean of this subset is the MCD measure of location. Let

\[
U_i = X_i - \hat{\tau}
\]

\((i = 1, \ldots, n)\),

\[
B_i = U_i U_i'
\]

and for any \( j \) \((j = 1, \ldots, n)\), let

\[
W_{ij} = \sum_{k=1}^{J} U_{ik} U_{jk}
\]

and

\[
T_{ij} = \frac{W_{ij}}{B_i} (U_{i1}, \ldots, U_{ij}).
\]

(4)

The distance between \( \hat{\theta} \) and the projection of \( X_j \) (when projecting onto the line connecting \( X_i \) and \( \hat{\tau} \)) is

\[
H_{ij} = \text{sign}(W_{ij}) \|T_{ij}\|,
\]

where \( \|T_{ij}\| \) is the Euclidean norm associated with the vector \( T_{ij} \). Let \( d_{ij} \) be the depth of \( X_j \) when projecting points onto the line connecting \( X_i \) and \( \hat{\theta} \). That is, for fixed \( i \) and \( j \), the
depth of the projected value of $X_j$ is
\[ d_{ij} = \min(\#\{H_{ij} \leq H_{ik}\}, \#\{H_{ij} \geq H_{ik}\}), \]
where $\#\{H_{ij} \leq H_{ik}\}$ indicates how many $H_{ik}$ ($k = 1, \ldots, n$) values satisfy $H_{ij} \leq H_{ik}$. Then the depth of $X_j$ is taken to be
\[ L_j = \min d_{ij}, \]
the minimum being taken over all $i = 1, \ldots, n$.

The Donoho–Gasko $\gamma$ trimmed mean associated with the $X_{ij}$ values is the average of all points that are at least $\gamma$ deep in the sample. That is, points having depth less than $\gamma$ are trimmed and the mean of the remaining points is computed. If the maximum depth among all $n$ points is at least $\gamma$, the breakdown point of the DG estimator is $\gamma/(1 + \gamma)$, where the breakdown point refers to the minimum proportion of points that must be altered to completely destroy an estimator. Here, $\gamma = .2$ is used.

The other estimator considered here, which performed well in simulations when testing (3), is a skipped estimator based on a projection method for detecting outliers, which will be labeled the SP estimator. Fix $i$, and for the point $X_i$ let
\[ A'_i = X_i - \hat{\tau}, \]
\[ B'_j = X_j - \hat{\tau}, \]
\[ C_j = \frac{A'_i B'_j}{B'_i B'_j}, \quad j = 1, \ldots, n. \]
Then when projecting the points onto the line between $X_i$ and $\hat{\tau}$, the distance of the $j$th point from $\hat{\tau}$ is
\[ V_{ij} = \|C_j\|. \]
The $j$th point is declared an outlier if
\[ V_{ij} > M_V + \sqrt{\chi^2_{.975,J}(q_2 - q_1)}, \quad (5) \]
where $M_V$, $q_1$ and $q_2$ are the usual sample median and estimates of the lower and upper quartiles, respectively, based on the $V_{i1}, \ldots, V_{in}$ values, and $\chi^2_{.95,J}$ is the .95 quantile of a chi-squared distribution with $J$ degrees of freedom. (Here, the quartiles are estimated via the ideal fourths; see Frigge et al., 1989.)
The process just described is for a single projection. Repeating this process for each
\(i (i = 1, \ldots, n)\), \(X_j\) is declared an outlier if for any of these projections, \(V_{ij}\) satisfies Eq.
(5). Removing any points declared an outlier, the mean of the remaining data is taken
to be the SP estimator of location. Its small-sample efficiency compares well to the DG
estimator (Wilcox, 2012). Note that the estimate of interquartile range, \(q_2 - q_1\), based on
the ideal fourths, has a breakdown point of .25 indicating that the breakdown point of the
SP estimator is .25 as well. The small-sample efficiency of the SP estimator compares well
to several other robust estimators that have been derived (Ng & Wilcox, 2010).

Several other affine equivariant estimators were considered but which performed poorly
in simulations in terms of controlling the Type I error probability. So computational details
related to these other estimators are not provided. They included the minimum volume
elipsoid (MVE) estimator (Rousseeuw & van Zomeren, 1990), the minimum covariance
determinant (MCD) estimator (Rousseeuw & van Driessen, 1999), the translated-biweight
S-estimator proposed (Rocke, 1996), the median ball algorithm (Olive, 2004) and the or-
thogonal Gnanadesikan-Kettenring (OGK) estimator (Maronna & Zamar, 2002).

2.2 Testing (2) and (3)

Two bootstrap methods for testing (2), as well as (3), were considered. The first, which
is designed to test (2) and corresponds to the RMPB3 in Wilcox (2012, section 8.2.5), is
applied as follows. Compute the test statistic
\[
Q = \sum (\hat{\theta}_j - \bar{\theta})^2,
\]
where \(\bar{\theta} = \sum \hat{\theta}_j / J\). An appropriate critical value is estimated by first setting \(Z_{ij} = X_{ij} - \hat{\theta}_j\).
That is, shift the empirical distributions so that the null hypothesis is true. Next, a bootstrap
sample is obtained by resampling, with replacement, \(n\) rows from the matrix \(Z\) yielding
\[
\begin{pmatrix}
Z_{11}^{*}, \ldots, Z_{1J}^{*} \\
\vdots \\
Z_{n1}^{*}, \ldots, Z_{nJ}^{*}
\end{pmatrix}.
\]
Compute the measure of location that is of interest based on this bootstrap sample yielding \( \hat{\theta}^*_j \) and

\[
Q^* = \sum (\hat{\theta}^*_j - \bar{\theta}^*_j)^2,
\]

where \( \bar{\theta}^*_j = \frac{\sum \hat{\theta}^*_j}{J} \). Repeat this process \( B \) times yielding \( Q^*_1, \ldots, Q^*_B \). Put these \( B \) values in ascending order yielding \( Q^*_{(1)} \leq \cdots \leq Q^*_{(B)} \). Then reject the hypothesis of equal measures of location if \( Q > Q^*_{(u)} \), where again \( u = (1 - \alpha)B \) rounded to the nearest integer.

The second method for testing (2) is based in part on bootstrap samples obtained from the \( X_{ij} \) values rather than the \( Z_{ij} \) values. The strategy is based on determining how deeply the grand mean is nested within the resulting bootstrap cloud. Details about this strategy can be found in Wilcox (2012, pp. 392-393). Because this approach performed poorly for the situation at hand, no details are provided.

The two bootstrap methods for testing (3) can be roughly described as follows. Take \( B \) bootstrap samples by resampling with replacement from the matrix \( X \), compute a measure of location based on the resulting difference scores and determine how deeply the null vector, \( 0 \), is nested within the bootstrap cloud. Here, two methods used to measure the depth of a point in data cloud: Mahalanobis distance and projection distance. In general this approach did not perform well. But when coupled with the DG estimator, it did perform reasonably well when testing (3).

To provide more details, let \( \hat{\Delta}^*_b \) \( (b = 1, \ldots, B) \) indicate the location estimate of \( \Delta \) based on the \( b \)th bootstrap sample and for convenience let \( \hat{\Delta}^*_0 \) denote the null vector. Let \( P^*_d(\hat{\Delta}^*_0) \) be the projection distance of \( \hat{\Delta}^*_b \) based on the \( B + 1 \) points \( \hat{\Delta}^*_0, \hat{\Delta}^*_1, \ldots, \hat{\Delta}^*_B \). So \( P^*_d(0) \) reflects how far the null vector is from the center of the bootstrap cloud. Then, from general theoretical results in Liu and Singh (1997), a p-value is

\[
1 - \frac{1}{B} \sum_{b=1}^B I(P^*_d(\hat{\Delta}^*_0) \geq P^*_d(\hat{\Delta}^*_b)),
\]

where the indicator function \( I(P^*_d(\hat{\Delta}^*_0) \geq P^*_d(\hat{\Delta}^*_b)) = 1 \) if \( P^*_d(\hat{\Delta}^*_0) \geq P^*_d(\hat{\Delta}^*_b) \); otherwise \( I(P^*_d(\hat{\Delta}^*_0) \geq P^*_d(\hat{\Delta}^*_b)) = 0 \). This will be called method D-P. When the projection distance is replaced by Mahalanobis distance, this will be called method D-M.
3 Simulation Results

Simulations were used to study the small-sample properties of the methods described in the previous section. The simulations were run using the software R, with much of the code written in C++. In addition, the R functions took advantage of a multi-core processor via the R package parallel. Despite this, execution time was relatively high, particularly when using the DG estimator in conjunction with method D-P. Consequently, estimated Type I error probabilities, \( \hat{\alpha} \), were based on 2000 replications. Four types of distributions were used: normal, symmetric and heavy-tailed, asymmetric and light-tailed, and asymmetric and heavy-tailed. More precisely, the marginal distributions were taken to be one of four g-and-h distributions (Hoaglin, 1985) that contain the standard normal distribution as a special case. (The R function rmul, in Wilcox, 2012, was used to generate observations.) If 

\[
W = \begin{cases} 
\frac{\exp(gZ) - 1}{g} \exp(hZ^2/2), & \text{if } g > 0 \\
Z \exp(hZ^2/2), & \text{if } g = 0 
\end{cases}
\]

has a g-and-h distribution where \( g \) and \( h \) are parameters that determine the first four moments. The four distributions used here were the standard normal \( (g = h = 0.0) \), a symmetric heavy-tailed distribution \( (h = 0.2, g = 0.0) \), an asymmetric distribution with relatively light tails \( (h = 0.0, g = 0.2) \), and an asymmetric distribution with heavy tails \( (g = h = 0.2) \). Table 1 shows the skewness \( (\kappa_1) \) and kurtosis \( (\kappa_2) \) for each distribution. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985). The number of bootstrap samples was taken to be \( B = 500 \). This choice generally seems to perform well in other settings, in terms of controlling the Type I error probability (Wilcox, 2012). But a possibility is that a larger choice for \( B \) might yield more power (e.g., Racine & MacKinnon, 2000). The correlation among the variables was taken to be \( \rho = 0 \) or \( \rho = .5 \).

As a partial check on the impact of heteroscedasticity on the Type I error probability, the \( X_{i1} \) values were taken to be \( \lambda X_{i1} \) \( (i = 1, \ldots, n) \). The two choices for \( \lambda \) were 1 and 4. For symmetric g-and-h distributions \( (g = 0) \), all of the measures of location considered here are equal to zero, so for \( \lambda = 4 \) the null hypothesis remains true. But when dealing with skewed distributions \( (g > 0) \), this is not the case. To deal with this, the expected value of an estimator was determined by generating 4000 samples of size \( n \) from a specified g-and-h
Table 1: Some properties of the g-and-h distribution.

<table>
<thead>
<tr>
<th>g</th>
<th>h</th>
<th>κ₁</th>
<th>κ₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.00</td>
<td>3.0</td>
</tr>
<tr>
<td>0.0</td>
<td>0.2</td>
<td>0.00</td>
<td>21.46</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0</td>
<td>0.61</td>
<td>3.68</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>2.81</td>
<td>155.98</td>
</tr>
</tbody>
</table>

Table 2: Estimated Type I error probabilities when testing (3), \( n = 20, \alpha = .05 \) using the SP estimator

<table>
<thead>
<tr>
<th>( \lambda = 1 )</th>
<th>D-M</th>
<th>( \lambda = 4 )</th>
<th>D-M</th>
<th>( \lambda = 1 )</th>
<th>D-P</th>
<th>( \lambda = 4 )</th>
<th>D-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>g h ρ = 0.0 ρ = 0.5</td>
<td>ρ = 0.0 ρ = 0.5</td>
<td>ρ = 0.0 ρ = 0.5</td>
<td>ρ = 0.0 ρ = 0.5</td>
<td>ρ = 0.0 ρ = 0.5</td>
<td>ρ = 0.0 ρ = 0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0 0.0 .069 .065</td>
<td>.096 .083</td>
<td>.055 .063</td>
<td>.075 .065</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0 0.2 .052 .047</td>
<td>.055 .049</td>
<td>.033 .042</td>
<td>.041 .043</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2 0.0 .070 .071</td>
<td>.039 .046</td>
<td>.054 .070</td>
<td>.054 .056</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2 0.2 .044 .044</td>
<td>.030 .040</td>
<td>.035 .039</td>
<td>.028 .040</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

distribution (with \( \lambda = 1 \)) and then averaging the resulting estimates. So with \( p = 4 \), in essence 16,000 estimates are being used. Then the marginal distributions were shifted so that, based on the expected value of an estimator, the null hypothesis is true.

Table 2 shows the results when using the SP estimator with methods D-M and D-P. Although the seriousness of a Type I error depends on the situation, Bradley (1978) has suggested that as a general guide, when testing at the .05 level, at a minimum the actual level should be between .025 and .075. As can be seen, this criterion is generally met when using D-M. But under normality, with \( ρ = .5 \), is this not the case, the largest estimate being .098. In contrast, when using D-P, the largest estimate is .075.

Table 3 reports simulation results when using method Q to test (2) with the DG estimator and \( n = 30 \). For \( n = 20 \), estimated Type I error probabilities exceed .075. But as indicated in Table 3, with \( n = 30 \), the estimates ranged between .025 and .061 when testing at the .05 level. When testing (3) instead via methods D-M or D-P, control over the Type I error probability was poor.
Table 3: Estimated Type I error probabilities, $n = 30$, $\alpha = .05$ using method Q to test (2) with the DG estimator

<table>
<thead>
<tr>
<th>$g$</th>
<th>$h$</th>
<th>$\rho = 0.0$</th>
<th>$\rho = 0.5$</th>
<th>$\rho = 0.0$</th>
<th>$\rho = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>.056</td>
<td>.057</td>
<td>.053</td>
<td>.060</td>
</tr>
<tr>
<td>0.0</td>
<td>0.2</td>
<td>.031</td>
<td>.034</td>
<td>.040</td>
<td>.041</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0</td>
<td>.054</td>
<td>.060</td>
<td>.057</td>
<td>.061</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>.026</td>
<td>.025</td>
<td>.038</td>
<td>.040</td>
</tr>
</tbody>
</table>

4 Concluding Remarks

In summary, when using a location estimator that takes into account the overall structure of data when dealing outliers, finding a method for testing (2) and (3) appears to be nontrivial when the sample size is small. The bulk of the methods considered here performed poorly in terms of controlling the Type I error probability, particularly when using an affine equivariant estimator. Only one method performed well in simulations when testing (2) and an affine equivariant estimator is used. No method based on an affine equivariant estimator was found to perform reasonably well when testing (3). Moreover, several bootstrap methods that perform reasonably well using a robust estimator applied to each of marginal distributions did not perform well for the situations considered here. However, the skipped estimator studied here, which is location and scale equivariant, was found to perform reasonably well when testing (3) via a percentile bootstrap method that measures the depth of null vector using projection distances. Another possible appeal of the SP estimator over the DG estimator is that for light-tailed distributions, including normal distributions, the DG estimator has relatively poor efficiency (e.g., Masse & Plante, 2003; Wilcox, 2012, p. 251). In contrast, the SP estimator performs nearly as well as the usual sample mean.

R functions are available for applying the methods that performed well in the simulations. The R function bd1GLOB tests (2). The DG estimator can be used by setting the argument est=dmean. Setting the argument MC=TRUE takes advantage of multi-core processor, if one multiple cores are available, via the R package parallel, which can be installed via R command install.packages. The R function rmdzD applies method D-P in conjunction with the SP estimator. Again, setting the argument MC=TRUE will take advantage of a
multi-core processor if one is available and the R package parallel has been installed. These functions can be installed with the R command install.packages("WRS",repos="http://R-Forge.R-project.org"). They are also stored in the file Rallfun-v23, which can be downloaded from the first author’s web page.

REFERENCES


Olive, D. J. (2004). A resistant estimator of multivariate location and dispersion *Com-


