British Journal of Mathematical and Statistical Psychology (2000), **53**, 69–82 *Printed in Great Britain*

Repeated measures ANOVA: Some new results on comparing trimmed means and means

Rand R. Wilcox*

Department of Psychology, University of Southern California, USA

H. J. Keselman

Department of Psychology, University of Manitoba, Canada

Jan Muska

Department of Psychology, University of Southern California, USA

Robert Cribbie

Department of Psychology, University of Manitoba, Canada

This paper considers the common problem of testing the equality of means in a repeated measures design. Recent results indicate that practical problems can arise when computing confidence intervals for all pairwise differences of the means in conjunction with the Bonferroni inequality. This suggests, and is confirmed here, that a problem might occur when performing an omnibus test of equal means. The problem is that the probability of rejecting is not minimized when the means are equal and the usual univariate F test is used with the Huynh-Feldt correction ($\tilde{\varepsilon}$) for the degrees of freedom. That is, power can actually decrease as the mean of one group is lowered, although eventually it increases. A similar problem is found when using a multivariate method (Hotelling's T^2). Moreover, the probability of a Type I error can exceed the nominal level by a large amount. The paper considers methods for correcting this problem, and new results on comparing trimmed means are reported as well. In terms of both Type I errors and power, simulations reported here suggest that a percentile t bootstrap used with 20% trimmed means and an analogue of the $\tilde{\varepsilon}$-adjusted F gives the best results. This is consistent with extant theoretical results comparing methods based on means with trimmed means.

1. Introduction

Let $\mu_1, \ldots, \mu_J$ be the means corresponding to the marginal distributions of some J-variate distribution. As is well known, repeated measures designs play an important role in many areas of research where a common goal is to test

$$H_0: \mu_1 = \ldots = \mu_J \tag{1}$$

* Requests for reprints should be addressed to Professor Rand R. Wilcox, Department of Psychology, University of Southern California, Seeley G. Mudd Building, Room 501, Los Angeles, CA 90089-1061, USA.

versus H_1: at least one mean differs from the others. Recent results related to performing all pairwise comparisons hint that a previously undetected problem might arise when testing (1). (Details are given later in this section.) One goal in this paper is to confirm that a problem does indeed arise and to consider how this problem might be addressed.

Let $\mu_{t1}, \ldots, \mu_{tJ}$ be the population trimmed means corresponding to the J marginal distributions. Another goal is to report new results on testing

$$H_0\colon \mu_{t1} = \ldots = \mu_{tJ} \tag{2}$$

which extend results in Wilcox (1997b).

There are many known reasons for possibly preferring the goal of comparing trimmed means versus means (e.g. Huber, 1981, 1993; Keselman, Lix, & Kowalchuk, 1998; Staudte & Sheather, 1990; Wilcox, 1997a, 1997b; Wilcox, Keselman, & Kowalchuk, 1998). Briefly, trimmed means typically have smaller standard errors in applied work, suggesting that comparing trimmed means will usually have higher power. Theory, simulations and experience with actual data indicate that power can be greatly increased by comparing trimmed means versus means. Moreover, when sampling from normal distributions, little power is lost versus methods based on means. As is evident, exceptions occur – no single method is perfect in terms of providing the highest amount of power – but often comparing means can result in substantially lower power. In fact, very slight departures from normality (in the Kolmogorov sense) can mean relatively high power when using trimmed means versus low power when comparing means. In contrast, if sampling is from normal distributions, the advantage of using means, in terms of power, is small.

Another problem with conventional methods for comparing means is that control over the probability of a Type I error can be poor and in some situations they are biased, meaning that power can go down as we move away from the null hypothesis, although eventually it goes up. (An illustration is given later.). Extant theoretical and simulation results indicate that switching to trimmed means reduces this problem (Wilcox, 1997a). Another general argument for comparing population trimmed means is that they satisfy three fundamental robustness properties not enjoyed by the population mean: quantitative robustness, qualitative robustness and infinitesimal robustness (e.g. Huber, 1981; Staudte & Sheather, 1990). Despite these results, presumably arguments can still be made for comparing means, but it is not the goal of this paper to debate this issue. Rather, the goal is to expand on the implications of results recently reported by Wilcox (1997b).

Let $(X_{i1}, \ldots, X_{iJ})$, be a random sample from some J-variate distribution, where X_{ij} represents the ith observation from the jth group, $(i = 1, \ldots, n; j = 1, \ldots, J)$. Let F_j be the marginal distribution corresponding to the jth group, and let μ_{tj} be the γ-trimmed mean where, for any random variable X having distribution F,

$$\mu_t = \frac{1}{1-2\gamma} \int_{F^{-1}(\gamma)}^{F^{-1}(1-\gamma)} x \, \mathrm{d}F(x). \tag{3}$$

With no trimming, $\gamma = 0$, we have $\mu_t = \mu$, the population mean. An estimate of μ_{tj} is

$$\bar{X}_{tj} = \frac{1}{n-2g} \sum_{i=g+1}^{n-g} X_{(i)j},$$

where $X_{(1)j} \leq X_{(2)j} \leq \ldots \leq X_{(n)j}$ are the n values in the jth group written in ascending order, and $g = [\gamma n]$, where $[\gamma n]$ is the greatest integer less than or equal to γn.

Here attention is focused on μ and the 20% trimmed mean, $\gamma = 0.2$. The reason for using $\gamma = 0.2$ stems from published papers examining efficiency, which is of course related to achieving high power. Briefly, if no trimming is done ($\gamma = 0$), efficiency can be very poor under arbitrarily small departures from normality towards a heavy-tailed distribution. However, if too much trimming is done (for example, medians are used), then efficiency and power are poor when sampling from a normal distribution. The choice $\gamma = 0.2$ provides a good compromise between the mean and median. The choice $\gamma = 0.2$, versus no trimming, is also motivated by asymptotic results relating to controlling the probability of a Type I error (Wilcox, 1997a). It is noted that $\gamma = 0.2$ is consistent with recommendations made by Huber (1993).

Recent results on multiple comparisons (Wilcox, 1997a) suggest that practical problems might arise when using the usual F test with the Huynh and Feldt (1976) ($\tilde{\varepsilon}$) correction for the degrees of freedom to test (1). The simulations in Section 3 verify that problems do indeed arise: the probability of a Type I error can be substantially larger than the nominal level and the probability of rejecting is *not* minimized when the null hypothesis of equal means is true. More precisely, when the marginal distributions are skewed with unequal variances, the probability of rejecting can decrease as the means become unequal, although eventually it increases. For example, Table 3 in Section 3 describes a situation where the means are equal and the probability of rejecting is 0.067 when testing at the $\alpha = 0.05$ level. Decreasing the mean of the first group by about a half standard deviation, so that it differs from the means of the other three groups, the probability of rejecting *drops* to 0.041. Decreasing it by a standard deviation, the probability of rejecting is 0.066. (This result is not surprising in light of general results in Wilcox, 1997a, on how skewness affects the paired t test for means.) For another situation considered in Section 3, the probability of rejecting is 0.106 when H_0 is true and drops to 0.054 when the mean of the first group is decreased by a half standard deviation. Obviously this is an undesirable power property, and there is interest in whether some alternative method might be more satisfactory.

Switching to the multivariate method for comparing dependent groups (Hotelling's T^2), similar problems arise. For the first situation considered in the previous paragraph, the probability of a Type I error is 0.227 when H_0 is true, again testing at the 0.05 level. Decreasing the first mean by a half standard deviation, the probability of rejecting drops to 0.096, and decreasing it by a standard deviation, the probability of rejecting is only 0.145. That is, there is a substantially higher probability of rejecting when the null hypothesis is true versus a situation where the first mean differs from the other three by one standard deviation. The purpose of this paper is to consider two alternative methods for testing (1), plus three new methods for testing (2). It is already known that when comparing means, the multivariate method can be unsatisfactory in terms of Type I errors (e.g. Keselman & Keselman, 1990), but perhaps a bootstrap analogue of this approach gives better results, and one of the goals here is to investigate this possibility. A related goal is to compare a bootstrap analogue of the multivariate method with a bootstrap analogue of the usual ε-adjusted F test.

2. Description of the procedures to be compared

For convenience and brevity, we describe the methods for testing the null hypothesis of equal trimmed means. The methods we consider for testing (1) are obtained by setting the amount of trimming to $\gamma = 0$.

To begin, first consider a random example, $X_1, \ldots, X_n$, from a single population of participants. Again, let $g = [\gamma n]$, and let

$$Y_i = \begin{cases} X_{(g+1)} & \text{if } X_i \leq X_{(g+1)}, \\ X_i & \text{if } X_{(g+1)} < X_i < X_{(n-g)}, \\ X_{(n-g)} & \text{if } X_i \geq X_{(n-g)}. \end{cases}$$

In other words, Winsorize the X_i values and label the results Y_i, $i = 1, \ldots, n$. The statistic

$$\bar{Y} = \sum Y_i/n$$

is the usual γ-Winsorized mean (e.g. Staudte & Sheather, 1990).

Now for fixed j, define Y_{ij} in an analogous fashion – that is, Winsorize each of the marginal distributions. The estimated Winsorized covariance matrix is $\mathbf{V} = (\upsilon_{jk})$, where

$$\upsilon_{jk} = \frac{1}{n-1}\sum (Y_{ij} - \bar{Y}_j)(Y_{ik} - \bar{Y}_k).$$

Let $\bar{X}_{tj}$ be the trimmed mean for the jth group, and let

$$Q_c = (n - 2g)\sum_{j=1}^{J} (\bar{X}_{tj} - \bar{X}_t)^2$$

where $\bar{X}_t = \sum \bar{X}_{tj}/J$. Also let

$$Q_e = \sum\sum (Y_{ij} - \bar{Y}_{.j} - \bar{Y}_{i.} + \bar{Y}_{..})^2,$$

where $\bar{Y}_{.j} = \sum Y_{ij}/n$ is the Winsorized mean corresponding to the jth group, $\bar{Y}_{i.} = \sum Y_{ij}/J$, and $Y_{..} = \sum\sum Y_{ij}/(nJ)$ is the grand Winsorized mean. The Winsorized sum of squares provides an asymptotically correct estimate of the standard error of the trimmed mean (e.g., Staudte & Sheather, 1990). Following Wilcox (1997a, 1993), a test statistic for H_0 given by (2) is

$$F = \frac{R_c}{R_e}, \tag{4}$$

where $R_c = Q_c/(J-1)$ and $R_e = Q_e/[(n - 2g - 1)(J-1)]$.

The null distribution of F is approximated with an F distribution using an analogue of Box's (1954) ($\hat{\varepsilon}$) correction of the degrees of freedom when dealing with the violation of the usual sphericity (or circularity) assumption. (For a description of this assumption, see Kirk, 1995, pp. 275–277; or Rogan, Keselman, & Mendoza, 1979.) Let

$$\varepsilon = \frac{J^2}{J-1}\frac{(\bar{\xi}_{jj} - \bar{\xi}_{..})^2}{\sum\sum \xi_{jk}^2 - 2J\sum \bar{\xi}_{j.}^2 + J^2\bar{\xi}_{..}^2},$$

where ξ_{jk} is the population Winsorized covariance between X_{ij} and X_{ik}, $\bar{\xi}_{..} = \sum\sum \xi_{jk}/J^2$, $\bar{\xi}_{j.} = \sum \xi_{jk}/n$ and

$$\hat{\varepsilon} = \frac{J^2}{J-1}\frac{(\bar{\upsilon}_{jj} - \bar{\upsilon}_{..})^2}{\sum\sum \upsilon_{jk}^2 - 2J\sum \bar{\upsilon}_{j.}^2 + J^2\bar{\upsilon}_{..}^2},$$

where $\upsilon_{..} = \sum\sum \upsilon_{jk}/J^2$, and $\bar{\upsilon}_{j.} = \sum \upsilon_{jk}/n$. Using an obvious modification of the approach

in Huynh and Feldt (1976), let

$$\tilde{\varepsilon} = \frac{n(J-1)\hat{\varepsilon} - 2}{(J-1)\,[n-1-(J-1)\hat{\varepsilon}]}.$$

The degrees of freedom are estimated to be $\nu_1 = (J-1)\tilde{\varepsilon}$ and $\nu_2 = (J-1)\,(n-2g-1)\tilde{\varepsilon}$. Thus, reject H_0 if $F > f_{1-\alpha}$, where $f_{1-\alpha}$ is the $1-\alpha$ quantile of an F distribution with ν_1 and ν_2 degrees of freedom. Setting $\gamma = 0$, this method reduces to the Huynh–Feldt method for means.

Proceeding along the lines of Wilcox (1997a), an alternative method of testing the hypothesis of equal trimmed means can be derived using a simple generalization of the usual multivariate method for means. Let

$$U_j = \bar{X}_{tj} - \bar{X}_{tJ}, \qquad j = 1, \ldots, J-1.$$

Then a test statistic for H_0 given by (2) is

$$H = \frac{(n-2g)\,(n-2g-J+1)}{(n-1)\,(J-1)}\mathbf{U}\mathbf{V}^{-1}\mathbf{U}^{\mathrm{T}},$$

where $\mathbf{U} = (U_1, \ldots, U_{J-1})$, and H_0 is rejected if $H > f_{1-\alpha}$, the $1-\alpha$ quantile of an F distribution with $\nu_1 = J-1$ and $\nu_2 = n-2g-J+1$ degrees of freedom. When $\gamma = 0$, H reduces to the usual multivariate T^2 test statistic for means.

Hall and Padmanabhan (1992) report theoretical (asymptotic) results on a percentile t bootstrap method for trimmed means. The basic strategy is to estimate an appropriate critical value using the data at hand versus assuming that the null distribution has a particular form. Extending their method in an obvious way, a critical value for the test statistic F can be estimated as follows. First, generate a bootstrap sample by randomly sampling, with replacement, n rows of observations from the matrix

$$\begin{pmatrix} X_{11}, \ldots X_{1J} \\ \vdots \\ X_{n1}, \ldots, X_{nJ} \end{pmatrix}.$$

Label the results

$$\begin{pmatrix} X^*_{11}, \ldots, X^*_{1J} \\ \vdots \\ X^*_{n1}, \ldots, X^*_{nJ} \end{pmatrix}.$$

Next, set

$$C_{ij} = X^*_{ij} - \bar{X}_{tj}.$$

That is, shift the bootstrap samples so that, in effect, the bootstrap sample is obtained from a distribution for which the null hypothesis of equal trimmed means is true. Put another way, approximate the distribution of $X_{ij} - \mu_t$ with the distribution of $X^*_{ij} - \bar{X}_{tj}$. Next, compute F^*, the value of the F statistic based on the C_{ij} values. Repeat this process B times yielding F^*_b, $b = 1, \ldots, B$. Let $F^*_{(1)} \leq \cdots \leq F^*_{(B)}$ be the B values written in ascending order and set $m = (1-\alpha)B$. Then an estimate of an appropriate critical value is $F^*_{(m)}$. That is, reject the hypothesis of equal trimmed means if $F > F^*_{(m)}$. (For more details about the percentile t bootstrap method, see Efron & Tibshirani, 1993.) Again, setting $\gamma = 0$ yields a method for comparing means. A bootstrap analogue of the multivariate method is obtained in a similar

manner. That is, compute H^* using the C_{ij} values, repeat this process B times yielding H_b^*, $b = 1, \ldots, B$, and reject if $H > H_{(m)}^*$.

Some additional comments about the bootstrap method just described might be helpful. Consider the ten observations 1, 1, 1, 1, 1, 1, 1, 1, 1 and 100. The trimmed mean is 1, so centering these ten values yields 0, 0, 0, 0, 0, 0, 0, 0, 0 and 99. As is evident, the trimmed mean of the centred values is zero, as desired. However, the expected value of the bootstrap trimmed mean, with respect to the bootstrap distribution, is not equal to zero. This feature has already been discussed by Wilcox (1998) and found to be negligible in terms of probability coverage and Type I error probabilities when comparing independent groups. Based on the simulations described here, we find this to be unimportant for the problem at hand.

Here $B = 599$ is used because it gives good results in related situations (Wilcox, 1997a). The reason for using $B = 599$ rather than $B = 600$ stems from results in Hall (1986) showing that it is advantageous to choose B such that $1 - \alpha$ is a multiple of $(B + 1)^{-1}$, and here attention is focused on $\alpha = 0.05$.

3. Simulation results

Simulations were used to check the small-sample properties of the methods described in the previous section when $J = 4$. The general procedure was to generate observations from a multivariate normal distribution with a particular correlation matrix, and variances all equal to one, and then transform the observations when considering non-normal distributions. The variances of the marginal distributions were also varied in a manner to be described. Three of the four correlation matrices considered here have a common correlation, ρ, with $\rho = 0.1$, 0.5 and 0.8. The fourth correlation matrix is $\rho_{12} = 0.8$, $\rho_{13} = 0.5$, $\rho_{14} = 0.2$, $\rho_{23} = 0.5$, $\rho_{24} = 0.2$ and $\rho_{34} = 0.2$. The first three matrices correspond to $\varepsilon = 1$, while the last matrix corresponds to $\varepsilon = 0.43$. In our $J = 4$ design, the possible values for ε range between 1 and $1/(J - 1) = 0.33$. Henceforth, these four correlation matrices will be called C1, C2, C3 and C4, respectively. One reason for reporting the results for these four matrices is to illustrate what happens when all of the correlations are small, all of them are moderately large, all are close to one, and when the correlations are more or less uniformly distributed between 0 and 1.

Simulations were run by generating observations from a multivariate normal distribution via the IMSL (International Mathematical and Statistical Library, 1987) subroutine RNMVN. Non-normal distributions were generated using the g-and-h distribution (Hoaglin, 1985). That is, 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 reason for using the g-and-h distribution is that it provides a convenient method for considering a very wide range of situations corresponding to both symmetric and asymmetric distributions. The case $g = h = 0$ corresponds to a normal distribution. The case $g = 0$ corresponds to a symmetric distribution, and as g increases, skewness increases as well. The parameter h determines heavy-tailedness. As h increases, heavy-tailedness increases as well.

Table 1. Some properties of the g-and-h distribution

g	h	κ_1	κ_2	$\hat{\kappa}_1$	$\hat{\kappa}_2$
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

Table 1 summarizes the skewness (κ_1) and kurtosis (κ_2) for the four g-and-h distributions used in the simulations. Again, the non-normal distributions considered here might seem extreme, but it is unclear how non-normal distributions might be in practice, so if a method performs well under seemingly extreme departures from normality, it would seem preferable to a method that does not.

When $h > 1/k$, $E(X - \mu)^k$ is not defined and the corresponding entry in Table 1 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 1. In fact, as h gets large, there is an increasing difference between the theoretical and actual values for skewness and kurtosis. Accordingly, Table 1 also lists the estimated skewness ($\hat{\kappa}_1$) and kurtosis ($\hat{\kappa}_2$) based on 100 000 observations generated from the distribution.

The g-and-h distributions with $g = 0.5$ deserve a special comment because the resulting levels of kurtosis or skewness might seem extreme and unrealistic. Empirical attempts at determining realistic ranges for skewness and kurtosis in psychological research have been published, but doubt remains as to what constitutes a satisfactory range of values in a simulation study because estimated ranges vary drastically among published papers (e.g., Wilcox, 1997a). Yet, a practical issue is whether a hypothesis testing method can be found that performs well over all realistic values for skewness and kurtosis. The g-and-h distribution is used with the idea that if two methods perform well under normality, and the first breaks down as we increase skewness or kurtosis, but the second does not, even under seemingly unrealistic departures from normality, this suggests that the second method is more satisfactory. (For a more detailed discussion of designing simulation studies, see Wilcox, 1995.)

Because of results in Wilcox (1997a), additional simulations were also run where the marginal distributions had a lognormal or exponential distribution. This was accomplished by generating X_{ij} from a multivariate normal distribution, computing $U_{ij} = \Phi(X_{ij})$, where $\Phi(x)$ is the standard normal distribution, and then transforming the U_{ij}s using an appropriate quantile function. (A description of these quantile functions can be found in Parzen, 1979.)

Simulations were also 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 mean (or trimmed mean when appropriate) of zero, and then the ith observation in the jth group was multiplied by σ_j, $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (1, 3, 4, 5)$. (The means and trimmed means of the g-and-h distributions used here are listed in Wilcox, 1997, p. 73. The lognormal has $\mu_t = 1.111$, and the exponential has $\mu_t = 0.761$.) When examining power, the mean (or trimmed mean) of the first group was shifted by δ.

For each replication in the simulations, both F and H were used to test the hypothesis of equal means (or trimmed means) with $\alpha = 0.5$ and $n = 21$. When not using the bootstrap method, 10 000 replications were used to estimate the actual probability of a Type I error, the estimate being the proportion of times H_0 was rejected among all the replications used. When using the bootstrap method, 1000 replications were used instead. Results in Robey and Barcikowski (1992) suggest that using 1000 replications is adequate. For example, suppose we follow Bradley's (1978) so-called liberal criterion where, when testing at the 0.05 level, the actual probability of rejecting should not drop below 0.025 or be above 0.075. Further suppose simulation results are used to test at the 0.05 level the hypothesis that when using F, for example, the actual probability of a Type I error is $\alpha = 0.05$. In symbols, if the actual probability of a Type I error is π when using F, simulations can be used to test H_0: $\pi = 0.05$. The number of replications needed to achieve power equal to 0.9, when $\pi = 0.025$ or 0.075, is 976.

Table 2 reports the estimated Type I error probabilities for g-and-h distributions when the bootstrap is not used. Estimates that do not satisfy Bradley's liberal criterion are shown in bold. For normal distributions ($g = h = 0$) where the marginal distributions have a common variance, both F and H provide good control over the probability of a Type I error when using means ($\gamma = 0$), as would be expected. With unequal variances, both methods again give good results. However, for heavy-tailed distributions ($h = 0.5$), both become too conservative in terms of Type I errors, meaning that the estimated probability of a Type I error drops below Bradley's criterion of 0.025, a result that was expected based on related simulation studies (Wilcox, 1997a). In contrast, comparing trimmed means ($\gamma = 0.2$) with F results in a Type I

Table 2. Estimated Type I error probabilities for means (parametric)

			C1		C2		C3		C4	
h	γ	σ	F	H	F	H	F	H	F	H
Symmetric distributions ($g = 0$)										
0.0	0.0	(1,1,1,1)	0.049	0.052	0.048	0.054	0.048	0.051	0.049	0.050
0.5	0.0		**0.022**	0.032	**0.019**	0.030	**0.017**	0.031	**0.022**	0.029
0.0	0.2		0.046	0.029	0.041	0.028	0.029	0.028	0.047	0.028
0.5	0.2		0.038	**0.020**	0.034	**0.020**	**0.023**	**0.023**	0.037	**0.020**
0.0	0.0	(1,3,4,5)	0.051	0.050	0.051	0.050	0.050	0.048	0.049	0.051
0.5	0.0		**0.012**	**0.022**	0.032	**0.018**	0.026	**0.023**	0.027	**0.023**
0.0	0.2		0.054	0.030	0.047	0.033	0.048	0.041	0.051	0.037
0.5	0.2		0.042	**0.018**	0.035	**0.021**	0.031	0.027	0.031	0.027
Asymmetric distributions ($g = 0.5$)										
0.0	0.0	(1,1,1,1)	0.038	0.055	0.040	0.054	0.041	0.049	0.039	0.056
0.5	0.0		**0.016**	0.037	**0.013**	0.030	**0.013**	0.025	**0.016**	0.033
0.0	0.2		0.041	0.030	0.035	**0.023**	0.027	**0.022**	0.042	0.025
0.5	0.2		0.034	**0.019**	0.028	**0.019**	**0.022**	**0.020**	0.035	**0.020**
0.0	0.0	(1,3,4,5)	0.047	**0.082**	0.054	**0.095**	0.061	**0.105**	0.047	**0.098**
0.5	0.0		0.054	**0.216**	**0.092**	**0.274**	**0.152**	**0.332**	0.065	**0.308**
0.0	0.2		0.049	0.037	0.045	0.042	0.047	0.056	0.049	0.044
0.5	0.2		0.037	**0.022**	0.035	0.031	0.038	0.039	0.037	0.030

error probability reasonably close to the nominal level, even when sampling from a heavy-tailed distribution, except for the C3 condition. Using H is a bit less satisfactory, the estimated probability of a Type I error dropping below 0.025 in some cases. For asymmetric distributions ($g = 0.5$), H can result in an estimated Type I error probability exceeding 0.075 and even 0.2 when comparing means. Using F to compare means typically results in better control over the probability of a Type I error, but in two cases the estimate exceeds 0.1. Generally, comparing trimmed means with F provides the best control over the probability of a Type I error.

Next, consider the situation where the marginal distributions are lognormal. Table 3 shows the estimated probability of rejecting when not using the bootstrap. Results with $\delta = -1$ reflect power when the first distribution is shifted a half standard deviation. In terms of Type I errors ($\delta = 0$), both methods for means are unsatisfactory, meaning that the estimated Type I error probabilities typically do not satisfy Bradley's liberal criterion. Particularly interesting are the results when the marginal distributions have unequal variances. In some instances, power actually decreases as we move away from the null hypothesis, although eventually it goes up. Comparing means with H is especially unsatisfactory. Switching to trimmed means ($\gamma = 0.2$), F yields much better power properties and generally provides better control over the Type I error probability. The main difficulty is that it can be too conservative in some situations, meaning that the estimated probability of a Type I error can drop below 0.025.

Table 3. Estimated Type I error probabilities and power for lognormal and exponential distributions (parametric)

			C1		C2		C3		C4	
δ	γ	σ	F	H	F	H	F	H	F	H
Lognormal distribution										
0	0.0	(1,1,1,1)	**0.018**	0.062	**0.021**	0.044	**0.019**	0.030	**0.023**	0.052
−1	0.0		0.386	0.426	0.542	0.614	0.792	0.854	0.489	0.779
0	0.2		0.025	**0.021**	**0.022**	**0.015**	**0.015**	**0.012**	0.027	**0.017**
−1	0.2		0.705	0.542	0.864	0.736	0.978	0.941	0.835	0.885
0	0.0	(1,3,4,5)	0.042	**0.177**	0.067	**0.227**	**0.106**	**0.250**	0.054	**0.227**
−1	0.0		0.033	0.076	0.041	0.096	0.054	0.117	0.054	0.075
−2	0.0		0.054	0.157	0.066	0.145	0.116	0.259	0.058	0.248
0	0.2		0.037	0.055	0.040	0.066	0.040	**0.087**	0.049	0.071
−1	0.2		0.049	0.060	0.060	0.046	0.080	0.072	0.053	0.077
−2	0.2		0.165	0.457	0.287	0.464	0.493	0.618	0.195	0.738
Exponential distribution										
0	0.0	(1,1,1,1)	0.031	0.060	0.037	0.052	0.036	0.047	0.036	0.058
0	0.0	(1,3,4,5)	0.046	**0.103**	0.072	**0.136**	0.072	**0.136**	0.050	**0.128**
−0.5	0.0		0.044	0.066	0.049	0.067	0.055	0.085	0.047	0.066
0	0.2	(1,1,1,1)	0.035	0.029	0.034	**0.018**	**0.024**	**0.015**	0.037	**0.022**
0	0.2	(1,3,4,5)	0.044	0.051	0.043	0.058	0.049	**0.077**	0.048	0.062
−0.5	0.2		0.050	0.041	0.053	0.036	0.056	0.051	0.051	0.048

As for exponential distributions, similar results are obtained, only the problems just noted are less severe. (Now $\delta = -0.5$ is used because this reflects a shift of a half standard deviation.) Simulations were also run with $\delta = -1$, but nothing interesting was found when comparing means.

Next attention is turned to the bootstrap method. Table 4 shows the estimated Type I error probabilities when sampling from the g-and-h distributions. For normal distributions ($g = h = 0$) where the marginal distributions have a common variance, both F and H provide a good control over the probability of a Type I error when using means ($\gamma = 0$). With unequal variances, both methods again give reasonably good results, with the F statistic being perhaps a bit better. However, for heavy-tailed distributions ($h = 0.5$), both become too conservative in terms of Type I errors, a result that was expected based on related simulation studies (Wilcox, 1997a). In some cases the estimated probability of a Type I error drops below 0.01. In contrast, comparing trimmed means ($\gamma = 0.2$) with F results in a Type I error probability reasonably close to the nominal level, even when sampling from a heavy-tailed distribution. Using H is a bit less satisfactory, the probability of a Type I error dropping below 0.025 in some cases. For asymmetric distributions ($g = 0.5$), H results in an estimated Type I error probability exceeding 0.075 in some cases, when comparing means. Using F to compare means usually results in better control over the probability of a Type I error, but in one instance the estimate exceeds 0.1. Generally, comparing trimmed means ($\gamma = 0.2$) with F provides the best control over the probability of a Type I error, with the other three methods being unsatisfactory in some situations.

Table 4. Estimated Type I error probabilities for g-and-h distributions (using the bootstrap)

			C1		C2		C3		C4	
h	γ	σ	F	H	F	H	F	H	F	H
Symmetric distributions ($g = 0$)										
0.0	0.0	(1,1,1,1)	0.044	0.038	0.046	0.040	0.050	0.034	0.054	0.044
0.5	0.0		**0.006**	**0.007**	**0.012**	**0.008**	**0.011**	**0.006**	**0.017**	**0.006**
0.0	0.2		0.053	0.029	0.051	0.031	0.046	0.030	0.064	0.035
0.5	0.2		0.058	0.025	0.059	0.026	0.042	**0.021**	0.065	0.032
0.0	0.0	(1,3,4,5)	0.056	**0.024**	0.053	**0.022**	0.040	**0.020**	0.060	0.027
0.5	0.0		**0.012**	**0.006**	**0.007**	**0.005**	**0.013**	**0.004**	**0.017**	**0.009**
0.0	0.2		0.060	0.034	0.055	0.032	0.048	0.025	0.057	0.033
0.5	0.2		0.056	**0.024**	0.053	**0.022**	0.040	**0.020**	0.060	0.027
Asymmetric distributions ($g = 0.5$)										
0.0	0.0	(1,1,1,1)	0.038	0.031	0.035	0.029	0.045	–	0.029	0.030
0.5	0.0		**0.008**	**0.003**	**0.008**	**0.002**	**0.010**	–	**0.006**	**0.005**
0.0	0.2		0.053	0.026	0.047	**0.021**	0.043	–	0.043	0.026
0.5	0.2		0.050	**0.021**	0.045	**0.022**	0.037	–	0.057	**0.022**
0.0	0.0	(1,3,4,5)	0.048	0.051	0.044	0.050	0.053	–	0.050	0.047
0.5	0.0		0.038	**0.086**	0.062	**0.103**	**0.116**	–	0.050	**0.100**
0.0	0.2		0.063	0.041	0.057	0.044	0.050	–	0.054	0.036
0.5	0.2		0.054	0.037	0.054	0.034	0.047	–	0.058	0.034

Note that in Table 4, no results are given for matrix C3 when $g = 0.5$. This is because on very rare occasions the bootstrap method breaks down when trying to invert the covariance matrix.

Finally, Table 5 shows estimated Type I error probabilities and power for lognormal and exponential distributions when using the bootstrap. Note that when comparing means ($\gamma = 0$), sampling is from a lognormal distribution and the marginal distributions have a common variance, both F and H can be too conservative in terms of Type I errors. In fact, the bootstrap seems to be a bit less satisfactory than using F with no bootstrap at all. As for comparing trimmed means ($\gamma = 0.2$) with H, again the probability of a Type I error can be less than 0.025, but reasonably good control is obtained with F, the estimates ranging between 0.029 and 0.045.

Particularly interesting are the results for the lognormal distribution when the marginal distributions have unequal variances. For means, the bootstrap does not correct the problem noted in Table 3: power goes down when the first mean is decreased by a half standard deviation, and with a shift of one standard deviation the probability of rejecting is about the same as when the null hypothesis is true. In contrast, when using F with trimmed means, the probability of rejecting increases with a half standard deviation shift. Moreover, power can be substantially higher versus using means. Again, H is not quite as satisfactory as F. Also, using F to compare trimmed means with a bootstrap critical value provides slightly better control over the probability of a Type I error versus using no bootstrap at all.

Table 5. Estimated Type I error probabilities and power for lognormal and exponential distributions (using the bootstrap method)

			C1		C2		C3		C4	
δ	γ	σ	F	H	F	H	F	H	F	H
Lognormal distribution										
0	0.0	(1,1,1,1)	**0.019**	**0.007**	**0.013**	**0.005**	**0.011**	**0.004**	**0.018**	**0.007**
−1	0.0		0.317	0.232	0.454	0.338	0.700	0.612	0.381	0.531
0	0.2		0.045	**0.020**	0.038	**0.015**	0.029	**0.012**	0.041	**0.016**
−1	0.2		0.721	0.534	0.877	0.709	0.975	0.905	0.862	0.875
0	0.0	(1,3,4,5)	0.042	0.070	0.046	0.072	0.068	0.061	0.048	0.063
−1	0.0		0.029	0.020	0.028	0.018	0.036	0.026	0.035	0.012
−2	0.0		0.042	0.023	0.037	0.015	0.062	0.029	0.049	0.033
0	0.2		0.051	0.047	0.042	0.046	0.047	0.037	0.050	0.043
−1	0.2		0.062	0.038	0.074	0.031	0.079	0.031	0.075	0.045
−2	0.2		0.207	0.313	0.291	0.248	0.443	0.255	0.217	0.416
Exponential distribution										
0	0.0	(1,1,1,1)	0.036	0.027	0.033	**0.018**	0.025	**0.014**	0.040	**0.021**
0	0.0	(1,3,4,5)	0.048	0.058	0.046	0.050	0.053	0.044	0.050	0.047
−0.5	0.0		0.044	0.030	0.038	0.021	0.043	0.026	0.045	0.018
0	0.2	(1,1,1,1)	0.051	**0.021**	0.043	**0.018**	0.036	**0.016**	0.047	**0.018**
0	0.2	(1,3,4,5)	0.061	0.048	0.049	0.046	0.048	0.035	0.055	0.039
−0.5	0.2		0.064	0.025	0.057	0.025	0.061	0.029	0.064	0.028

It is noted that power was also checked when observations were generated from skewed g-and-h distributions: no new insights were obtained, so the results are not reported.

Finally, we compare the power of the percentile t bootstrap method based on F and trimmed means with the power of Friedman's test as well as the parametric F test based on means. First consider situations where the marginal distributions are normal with a common variance of 1, three of the marginal distributions have means equal to zero and the fourth has mean $\delta = -1$. For the four covariance matrices considered here, C1–C4, power using the bootstrap method was estimated to be 0.855, 0.880, 0.988 and 0.982, respectively. As for Friedman's test, power was estimated to be 0.853, 0.984, 1.0 and 0.867 based on 10 000 replications. For the parametric F test using means, the estimates are 0.942, 0.997, 1 and 0.998. Thus, comparing means has more power, as would be expected. For $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (1, 3, 4, 5)$ and $\delta = -2$, power estimates were 0.302, 0.312, 0.483, 0.361 using the bootstrap with trimmed means versus 0.283, 0.399, 0.628, 0.299 using Friedman's test. For the parametric F test for means the estimates are 0.383, 0.603, 0.744 and 0.441. For symmetric heavy-tailed distributions ($g = 0$, $h = 0.5$) having equal marginal variances, the estimates were 0.673, 0.710, 0.896, 0.871 versus 0.559, 0.761, 0.972, 0.568 with Friedman's test and 0.188, 0.299, 0.528, 0.241 with means. This illustrates the well-known result that when comparing means, power can be relatively poor even under slight departures from normality. For unequal marginal variances the estimates were 0.495, 0.551, 0.691, 0.590 using trimmed means and the bootstrap versus 0.379, 0.513, 0.756, 0.395 using Friedman. Thus, situations arise where the bootstrap with trimmed means has more power versus Friedman's test, but the reverse happens as well. For asymmetric distributions, power advantages will depend on how the groups differ simply because the tests are designed to be sensitive to different features of the data (Vargha and Delaney, 1998, p. 185, provide a useful description of the alternative hypothesis of Friedman's test.)

4. An illustration

A portion of a study by M. Earleywine collected data on hangover symptoms after individuals are given a specific amount of alcohol in a laboratory setting. Each individual was measured at three different times. For the control group, the hypothesis of equal means for the three times is not rejected with the $\tilde{\varepsilon}$-adjusted F test, the significance level being 0.51. Switching to trimmed means, the test statistic is $F = 2.69$ and the bootstrap 0.05 critical value is 3.16. Thus, again we do not reject, but we do reject with $\alpha = 0.1$. Friedman's test has a significance level of 0.11. If we pool the controls with the experimental group, again comparing measures taken at three different times, the hypothesis of equal means has a significance level of 0.41. In contrast, for trimmed means the test statistic is $F = 5.89$ with an $\alpha = 0.05$ bootstrap critical value of 3.26, so the hypothesis of equal trimmed means is rejected. In fairness, the data are highly skewed, so comparing means is not the same as comparing trimmed means. The only point is that the choice of method can give a substantially different result.

5. Concluding remarks

In summary, poor control over the probability of a Type I error was illustrated when comparing repeated measures means with either the univariate $\tilde{\varepsilon}$-adjusted F or multivariate T^2 statistic, H, and an undesirable power property was illustrated as well. Switching to the

percentile *t* bootstrap method does not necessarily eliminate these problems when attention is restricted to means. However, comparing trimmed means with a bootstrap method provides good control over the probability of a Type I error, it eliminates the undesirable power property associated with means, and high power is achieved for non-normal distributions versus low power when using means. The main concern with comparing trimmed means with the $\tilde{\varepsilon}$-adjusted F and no bootstrap is that in some cases the probability of a Type I error can be substantially less than nominal level.

For completeness, it is noted that other bootstrap methods (e.g., Efron & Tibshirani, 1993; Westfall & Young, 1993) might give better control over the probability of a Type I error when comparing means. However, even if such a method could be found, slight departures from normality can drastically reduce power (e.g. Wilcox, 1997a). In contrast, comparing trimmed means maintains high power in these same situations for the general reasons summarized in Wilcox (1997a). Consequently, comparing trimmed means with a percentile *t* bootstrap method appears to have practical value.

References

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effect of inequality of variance and correlation of errors in the two-way classification. *Annals of Statistics*, *25*, 484–498.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to bootstrap*. New York: Chapman & Hall.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, *14*, 1431–1452.

Hall, P., & Padmanabhan, A. R. (1992). On the bootstrap and the trimmed mean. *Journal of Multivariate Analysis*, *41*, 132–153.

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

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Huber, P. J. (1993). Projection pursuit and robustness. In S. Morgenthaler, E. Ronchetti, & W. Stahel (Eds.), *New directions in statistical data analysis and robustness*. Boston: Birkhäuser Verlag.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, *1*, 69–82.

International Mathematical and Statistical Library (1987). *Library I* (Vol. II). Houston, TX: IMSL.

Keselman, J. C., & Keselman, H. J. (1990). Analysing unbalanced repeated measures designs. *British Journal of Mathematical and Statistical Psychology*, *43*, 265–282.

Keselman, H. J., Lix, L. M., & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. *Psychological Methods*, *3*, 123–141.

Kirk, R. E. (1995). *Experimental design* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Parzen, E. (1979). Nonparametric statistical data modeling (with discussion). *Journal of the American Statistical Association*, *74*, 105–120.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*, 283–288.

Rogan, J. C., Keselman, H. J., & Mendoza, J. L. (1979). Analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology*, *32*, 269–286.

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

Vargha, A., & Delaney, H. D. (1998). The Kruskal–Wallis test and stochastic homogeneity. *Journal of Educational and Behavioural Statistics*, *23*, 170–192.

Westfall, P. H., & Young, S. S. (1993). *Resampling based multiple testing*. New York: Wiley.

Wilcox, R. R. (1993). Analysing repeated measures or randomized block designs using trimmed means. *British Journal of Mathematical and Statistical Psychology*, *46*, 63–76.

Wilcox, R. R. (1995). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. *British Journal of Mathematical and Statistical Psychology*, *48*, 99–114.

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.

Wilcox, R. R. (1998). The goals and strategies of robust methods (with discussion). *British Journal of Mathematical and Statistical Psychology*, *51*, 1–61.

Wilcox, R. R., Keselman, H. J., & Kowalchuk, R. K. (1998). Can tests for treatment group equality be improved?: The bootstrap and trimmed means conjecture. *British Journal of Mathematical and Statistical Psychology*, *51*, 123–134.

Received 13 March 1998; revised version received 9 February 1999