Tests for Treatment Group Equality When Data are Nonnormal and Heteroscedastic

Robert A. Cribbie
York University

Rand R. Wilcox
University of Southern California, rwilcox@usc.edu

Carmen Bewell
York University

H. J. Keselman
University of Manitoba

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm
Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation
DOI: 10.22237/jmasm/1177992660
Available at: http://digitalcommons.wayne.edu/jmasm/vol6/iss1/12

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.
Several tests for group mean equality have been suggested for analyzing nonnormal and heteroscedastic data. A Monte Carlo study compared the Welch tests on ranked data and heterogeneous, nonparametric statistics with previously recommended procedures. Type I error rates for the Welch tests on ranks and the heterogeneous, nonparametric statistics were well controlled with a slight power advantage for the Welch tests on ranks.

Key words: Welch test on ranks, nonparametric statistics, nonnormality, heteroscedasticity, mean equality

Robert A. Cribbie is Associate Professor of Psychology, Department of Psychology, Toronto, Ontario, M3J 1P3 (cribbie@yorku.ca), specializing in robust test statistics and multiplicity control. Rand. R. Wilcox is Professor of Psychology, Department of Psychology, (rwilcox@usc.edu). Carmen Bewell (cbewell@yorku.ca) is a Ph. D. student in Clinical Psychology. H. J. Keselman, Department of Psychology, is Professor of Psychology (hj_keselman@umanitoba.ca).
and power for some patterns of nonnormality (with unequal variances); however, for other patterns the Type I error rates can deviate considerably from the nominal rate (e.g., Cressie & Whitford, 1986; Keselman, Lix & Kowalchuk, 1998). Another potential solution when variances are heterogeneous and distribution shapes are nonnormal is to use a heteroscedastic statistic, such as Welch’s (1938, 1951) tests, with sample estimators that are intended to be robust to the biasing effects of nonnormality, e.g., trimmed means and Winsorized variances (see Yuen & Dixon, 1973; Wilcox, 1995, 1997). By minimizing the effects of extreme observations the trimmed mean can provide a more accurate representation of the central tendency of the majority of the distribution. An increase in power may also be experienced if eliminating the extreme observations reduces the standard error of the mean. However, Keselman, Lix, et al. (1998) reported that under some patterns of nonnormality power could be depressed relative to utilizing the usual means and variances.

Nonparametric test statistics (e.g., Wilcoxon, Mann-Whitney, Kruskal-Wallis) have been studied for unequal variances and nonnormal data. Zimmerman (1987; 1996) and Zimmerman and Zumbo (1993a), among others, showed that nonparametric test statistics are not robust to unequal variances, regardless of whether the data are normal or nonnormal. Zimmerman and Zumbo (1993a) explained that, “an attractive hypothesis is that both problems [nonnormality and variance heterogeneity] can be solved at once by the Welch t test performed on the ranks of measures instead of the measures themselves” (p. 507). Thus, with this approach, researchers would convert nonnormal, heteroscedastic data to ranks, and analyze the data with the Welch two-sample or omnibus tests. Zimmerman and Zumbo (1993a; 1993b) conducted simulation studies with several patterns of nonnormality and variance heterogeneity and report that the Welch test on ranks “counteracts effects of nonnormality and unequal variances at the same time” (p. 535). More specifically, for many patterns of nonnormality and variance homogeneity, the Welch test on ranks provided better overall Type I and Type II error control relative to the two-sample t and Welch t on unranked data or the two-sample t on ranks. However, it should be noted that for some patterns of nonnormality (e.g., lognormal) Type I error rates were not controlled within Bradley’s (1978) liberal criterion (+/- .5 α).

Another potential solution is the heteroscedastic rank-based test statistics proposed by Brunner and Munzel (2000) and Brunner, Dette and Munk (1997). Specifically these authors presented two-sample and omnibus, respectively, heteroscedastic rank-based test statistics that, unlike the traditional Kruskal-Wallis nonparametric statistic, consider the variance heterogeneity of the group distributions in the computational procedure. Munzel and Hothorn (2001) presented findings on the Type I error and power properties of the Brunner and Munzel two-sample procedure for nonnormal distributions with unequal variances, indicating that Type I error and power rates were considerably better than those of the parametric and nonparametric competitors. However, results were only reported for a many-to-one multiple comparisons setting for the discretized normal distribution.

The purpose of this article is to compare the Type I error control and power of the above strategies under several conditions of nonnormality and/or heteroscedasticity. It extends the conditions investigated by Zimmerman and Zumbo (1993a; 1993b) and Munzel and Hothorn (2001) to independent groups designs with more than two levels of the independent variable and, with respect to nonnormality, investigates skewed distributions not previously investigated and that have been reported to be representative of many behavioral science variables (Micceri, 1989; Wilcox, 1995). The Type I error control and power of the procedures in a multiple comparisons setting is also examined.

Test Statistics

Five omnibus test statistic and data configuration combinations were evaluated and compared in this study. These included: a) Welch’s (1951) test statistic on unranked data (Welch); b) Welch’s test statistic on trimmed means and Winsorized variances (20% symmetric trimming) (Welch-t); c) Welch’s test
statistic on ranked data (Welch-r); d) the Kruskal-Wallis (Kruskal & Wallis, 1952) omnibus nonparametric test statistic (which utilizes ranked data) (KW); and e) the Brunner, Dette and Munk (1997) heterogeneous nonparametric test statistic (BDM).

Welch’s (1938) two-sample test statistic can be expressed as:

\[
t_w = \frac{\bar{X}_j - \bar{X}_j'}{\sqrt{\frac{S_j^2}{n_j} + \frac{S_j'^2}{n_j'}}},
\]

which is distributed as a t variable with degrees of freedom due to Satterthwaite (1946),

\[
v_w = \frac{\left(\frac{S_j^2}{n_j} + \frac{S_j'^2}{n_j'}\right)^2}{\frac{S_j^4}{n_j^2(n_j-1)} + \frac{S_j'^4}{n_j'^2(n_j'-1)}}
\]

where \(\bar{X}_j, S_j\) and \(n_j\) represent the sample means, variances, and sample sizes, respectively, for the jth group (\(j \neq j', j = 1, ..., J\)).

Welch’s (1951) omnibus test can be expressed as:

\[
F_w = \frac{\sum_j w_j \left(\bar{X}_j - \bar{X}_j'\right)^2}{\frac{J-1}{\sum_j \left(1 - w_j / \sum_j w_j^2\right)^2} \left\{\sum_j \frac{1 - w_j / \sum_j w_j^2}{n_j - 1}\right\}^{1/2}}
\]

which is distributed as an F variable with J-1 and \(v_w\) degrees of freedom, where

\[
w_j = \frac{n_j}{S_j^2},
\]

\[
\bar{X}_j = \frac{\sum_j w_j \bar{X}_j}{\sum_j w_j},
\]

and

\[
v_w = \frac{J^2 - 1}{\sum_j \left(1 - w_j / \sum_j w_j^2\right)^2}.
\]

Kruskal-Wallis
The Kruskal-Wallis nonparametric procedure begins by ranking the observations in the combined sample. Let the rank of the ith observation in the jth group be represented by \(r_{ij}\) and the sum of the ranks for the jth group be represented by \(a_j = \sum_i r_{ij}\). The statistic tests the null hypothesis \(H_0: \lambda_1 = ... = \lambda_j\) (where \(\lambda\) represents the population mean only under the assumption that the population shapes are identical) and rejects \(H_0\) if \(KW \geq \chi^2_{(J-1)}\) where:
TESTS FOR TREATMENT GROUP EQUALITY

\[
KW = 12 \left( \frac{\sum_j a_j^2 / n_j}{N(N + 1)} \right) - 3(N + 3),
\]

and \( N = \sum_i n_i \). Multiple comparisons are performed with a modified two-sample version of the omnibus Kruskal-Wallis test (see Sprent & Smeeton, 1993). The null hypothesis \( H_0: \lambda_j = \lambda_j' \) is rejected if \(|t_{KW}| \geq t_{\alpha, N-1}\), where:

\[
t_{KW} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_r - C}{(N-1-KW)(n_j + n_{j'})} \frac{n_j n_{j'}(N-J)(N-1)}}}
\]

\[
S_r = \sum_i \sum_j r_{ij}^2
\]

and

\[
C = \frac{N(N+1)^2}{4}.
\]

**Welch-t**

Trimmed means are computed by removing a percentage of observations from each of the tails of a distribution. Let \( g_j = [\gamma n_j] \), where \( \gamma \) represents the proportion of observations to be trimmed from each tail of the distribution and \([x]\) is the largest integer less than or equal to \( x \). Further, let \( h_j \) represent the remaining (effective) sample size following removal of the trimmed observations. Recommendations have been made in the literature for 15% symmetric trimming (Mudholkar, Mudholkar & Srivastava, 1991) and 20% symmetric trimming (Wilcox, 1995). The jth sample trimmed mean can be represented as:

\[
\bar{X}_{ij} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j} X_{ij}
\]

and the jth sample Winsorized mean as

\[
\bar{X}_{wuj} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}
\]

where:

\[
Y_{ij} = X_{(g_j+1)j} \quad \text{if} \quad X_{ij} \leq X_{(g_j+1)j},
\]

\[
= X_{ij} \quad \text{if} \quad X_{(g_j+1)j} < X_{ij} < X_{(\gamma_j-g_j+1)j},
\]

\[
= X_{(\gamma_j-g_j+1)j} \quad \text{if} \quad X_{ij} \geq X_{(\gamma_j-g_j+1)j}.
\]

An associated Winsorized variance is computed by replacing the censored observations from the lower tail with the lowest uncensored observation and the censored observations from the upper tail with the highest uncensored observation. The Winsorized variance is:

\[
s_{wuj}^2 = \frac{1}{h_j-1} \sum_{i=1}^{n_j} (Y_{ij} - \bar{X}_{wuj})^2.
\]

The sample trimmed means and Winsorized variances can then be substituted into Welch’s (1938; 1951) two-sample and omnibus test statistics. For example, substituting the trimmed means and Winsorized variances into the Welch (1938) two-sample test yields the statistic:

\[
t_w = \frac{\bar{X}_{ij} - \bar{X}_{ij'} }{\sqrt{\frac{s_{wuj}^2}{h_j} + \frac{s_{wuj'}^2}{h_{j'}}}}
\]

\[
\sqrt{h_j \cdot h_{j'}}.
\]
with error degrees of freedom,

\[
R_j = \frac{1}{n_j} \sum_{i=1}^{n_j} R_{ij},
\]

and \( R_{ij} \) is the rank of \( X_{ij} \) after the data are pooled. The null hypothesis, \( H_0: \lambda_1 = \ldots = \lambda_J \), is rejected if \( F_w \geq F_{\alpha, v_1, v_2} \) where:

\[
v_1 = \frac{M_{11}}{\text{tr}(V)} \frac{\text{tr}(V^2)}{\text{tr}(V^2 \Lambda)}
\]

\[
v_2 = \frac{\left[ \text{tr}(V) \right]^2}{\text{tr}(V^2 \Lambda)}
\]

and \( \Lambda = \text{diag} \{(n_1 - 1)^{-1}, \ldots, (n_J - 1)^{-1}\} \). Multiple comparisons are performed with the two-sample version of the Brunner, Dette and Munk (1997) procedure (see Brunner & Munzel, 2000).

It is important to note that the null hypotheses associated with the above tests differ based on the characteristic(s) of the data that each test is sensitive to. The Welch test evaluates the null hypothesis that all population means are equal (i.e., \( H_0: \mu_1 = \ldots = \mu_J \)). The Welch-t evaluates the null hypothesis that all population trimmed means are equal (i.e., \( H_0: \mu_{t1} = \ldots = \mu_{tJ} \)). The K-W, Welch-r and BDM procedures evaluate the null hypothesis that all distribution functions are equal (i.e., \( H_0: \lambda_1 = \ldots = \lambda_J \)). It is important to note that with the K-W, Welch-r and BDM procedures that the null hypotheses only relate to a test of location when population distribution shapes and variances are equal, where the procedures are sensitive to differences in the mean ranks (see Brunner, Dette & Munk, 1997, p. 1498; Kruskal & Wallis, 1952; Sprent & Smeeton, 2001). Hence, an important component of this article is to evaluate the rates of rejection for these procedures when variances are unequal.

**Pairwise Multiple Comparison Procedures (MCPs)**

**Tukey**

The Tukey (1953) procedure rejects \( H_0: \mu_i = \mu_j \) if \( |t_i| \geq q(a, J, v) / (2)^{1/2} \), where \( q \) is a value from the Studentized range distribution with \( J \) groups and \( v \) degrees of freedom.
freedom, and \( t \) and \( \nu \) represents the appropriate two-sample \( t \)-distributed test statistic and associated degrees of freedom, respectively.

**REGWQ**

Ryan (1960) proposed a modification to the Newman-Keuls (Newman, 1939; Keuls, 1952) procedure that ensures that the familywise (overall) Type I error rate is maintained at \( \alpha \), even in the presence of partial null hypotheses. Ryan’s original procedure became known as the REGWQ after modifications to the procedure proposed by Einot and Gabriel (1975) and Welsch (1977). The REGWQ MCP sequentially tests all ordered mean differences for stretch sizes (inclusive ranges between rank-ordered means) \( p = J, J - 1, \ldots, 2, \) and rejects \( H_c: \mu_j = \mu_j' \) \((j \neq j')\) if an associated omnibus test has been rejected and:

\[
|t| \geq q \left( \alpha_p, p, \nu \right) / (2)^{1/2},
\]

where \( \alpha_p = \alpha \) for \( p = J, J - 1, \) and \( \alpha_p = 1 - (1 - \alpha)^{p / 2} \), for \( p = J - 2, \ldots, 2 \). If any \( H_c \)s are retained for \( p = p' \) then all \( H_c \)s contained in that stretch are retained and not tested at later stages (i.e., \( p < p' \)). If all \( H_c \)s are retained for \( p = p' \) then all \( H_c \)s with \( p \leq p' \) are retained.

**Methodology**

A Monte Carlo study was used to compare the Type I error and power rates of the Welch test on ranks and the Brunner heteroscedastic rank-based statistics with that of the Welch test on unranked data, the Welch test with trimmed means and Winsorized variances and the Kruskal-Wallis (Kruskal & Wallis, 1952) nonparametric test in a one-way independent groups design. In addition, the procedures were compared in a pairwise comparison framework, with the Tukey (1953) and REGWQ (Ryan, 1960; Einot & Gabriel, 1975; Welsch, 1977) procedures.

Seven variables were manipulated in this study: a) number of levels of the independent variable; b) total sample size; c) degree of sample size imbalance; d) degree of variance inequality; e) pairings of group sizes and variances; f) configuration of population means; and g) population distribution shape.

The number of levels of the independent variable was set at \( J = 4 \) and \( J = 7 \), resulting in 6 and 21 pairwise comparisons, respectively. This permits evaluation of the effect of the number of pairwise comparisons computed on Type I error control and power.

In order to investigate the effects of sample size, the total sample size (\( N \)) was manipulated by setting the average \( n_j = 10, 15, \) and 20 resulting in \( N = 40, 60 \) and 80 for \( J = 4 \), and \( N = 70, 105 \) and 140 for \( J = 7 \). The sample sizes were selected to be similar to those used by Zimmerman and Zumbo (1993a, b) in their investigations of the two-sample Welch (1938) test on ranked data. For the nonnull mean configurations used in this study, the group sizes 10, 15 and 20 result in a priori omnibus (ANOVA F statistic) power estimates of approximately .80, .95, and .98, respectively (assuming equal group sizes and variances).

Sample size balance or imbalance was also manipulated. Keselman et al. (1998) reported that unbalanced designs were more common than balanced designs in a review of studies published in educational and psychological journals. In addition, the effects of variance heterogeneity can be exacerbated when paired with unequal sample sizes. Therefore, three sample size conditions were examined (equal, moderately unequal and extremely unequal). The sample sizes used are enumerated in Table 1.

Degree of variance heterogeneity was also manipulated. According to Keselman et al. (1998), ratios of largest to smallest variances of 8:1 are not uncommon in educational and psychological studies and can have deleterious effects on the performance of many test statistics, especially when paired with unequal sample sizes. Therefore, three levels of variance equality/inequality were examined in this study: a) equal variances; b) largest to smallest variance ratio of 4:1; and c) largest to smallest variance ratio of 8:1. See Table 1 for group variances.

Pairings of variances and sample sizes can have differing effects on the Type I error and power rates of many test statistics. Specifically, when variances and sample sizes are directly (positively) paired Type I error estimates for the usual t/F tests can be
conservative (with correspondingly deflated power). On the other hand, when variances and sample sizes are inversely (negatively) paired Type I error estimates for the usual t/F tests can be liberal (with correspondingly inflated power). Therefore, both positive and negative pairings were examined.

Several configurations of nonnull population means were investigated, in addition to the complete null case. Following Toothaker’s (1991) definitions of mean configuration, equally spaced, minimum variability and maximum variability configurations were utilized. See Table 2 for a listing of the mean configurations.

Another factor examined in this study was population distribution shape. The three distribution shapes investigated were: 1) normally distributed data; 2) moderately skewed data from the g- and h- distribution (Hoaglin, 1985), where g = .5 and h = 0 (Skewness = 1.75, Kurtosis = 8.90); and 3) substantially skewed data from the g- and h- distribution, where g = 1 and h = 0 (Skewness = 6.20, Kurtosis = 114).

Empirical Type I error rates were recorded for all procedures, with familywise error rates reported for the MCPs. In this paper, the robustness of a procedure, with respect to Type I error control, will be determined using Bradley’s (1978) liberal criterion. That is, a procedure is deemed robust with respect to Type I errors if the empirical rate of Type I error falls within the range +/- .5 α. Power rates were also recorded for all the procedures, with power rates for the MCPs quantified with respect to average per-pair power (where per-pair power is the probability of rejecting a false pairwise null
hypothesis) and all-pairs power (the probability of rejecting all false pairwise null hypothesis).

The simulation program was written in SAS/IML (SAS Institute, Inc., 1999). Pseudorandom normal variates were generated with the SAS generator RANNOR. If \( Z_{ij} \) is a standard normal deviate, then \( X_{ij} = \mu_j + (\sigma_j Z_{ij}) \) is a normal variate with mean \( \mu_j \) and variance \( \sigma_j^2 \).

To generate data from the \( g \)- and \( h \)-distributions, standard unit normal variables were converted to the random variable:

\[
X_{ij} = \left[ \frac{\exp(gZ_{ij})}{g} \right] \left[ \exp \left( \frac{hZ_{ij}^2}{2} \right) \right].
\]

To obtain a distribution with standard deviation \( \sigma_j \), each \( X_{ij} \) was multiplied by a value of \( \sigma_j \). When \( g > 0 \) the \( g \)- and \( h \)-distribution population mean is not 0 and therefore the population mean was subtracted from \( X_{ij} \) before being multiplied by \( \sigma_j \). When working with trimmed means, the population trimmed mean for the \( j \)th group was also subtracted from the variate before multiplying by \( \sigma_j \). In order to ensure that the null hypothesis associated with the rank-based procedures was true when distribution shapes were nonnormal and variances were unequal, the Nelder and Mead (1965) minimization function was implemented through an S-Plus version of the FORTRAN code in Olsson (1974. See also Olsson & Nelson, 1975).

Distributions were shifted accordingly. Specifically, the S-Plus function 'nelder' was used, which is available in the library of R and S-Plus functions described in Wilcox (2005). Five thousand replications were performed for each condition, using a nominal significance level of .05.

<table>
<thead>
<tr>
<th>Population Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_1 )</td>
</tr>
<tr>
<td>( J = 4 )</td>
</tr>
<tr>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
</tr>
<tr>
<td>( J = 7 )</td>
</tr>
<tr>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
</tr>
</tbody>
</table>
Results

The pattern of Type I error and power results were consistent across sample size inequality, variance inequality, and nonnull mean configurations, and were therefore averaged over these conditions. Further, the pattern of results was similar across sample size conditions and therefore only the results for the largest sample size condition are presented and discussed (except when noted otherwise). For the pairwise MCPs, partial null familywise error rates were controlled within Bradley’s limits in all cases where complete null Type I error rates were controlled, and therefore are not reported.

Omnibus Tests

Type I error Control

Type I error rates (%) for J = 4 and J = 7 are presented in Table 3. When the distribution shapes were normal, Type I error rates were maintained within Bradley’s liberal bounds (2.5%-7.5%) by all, but one, procedure for J = 4 and J = 7; the Kruskal-Wallis procedure was liberal (7.8%) for J = 4 when sample sizes and variances were negatively paired. When the distribution shapes were skewed, the Welch and Kruskal-Wallis tests did not always maintain Type I error rates within Bradley’s bounds when J = 4 and sample sizes and variances were negatively paired. The Welch test in particular became very liberal (e.g., 16.9%), whereas the Kruskal-Wallis test exhibited some inflation (e.g., 7.9%). The remaining procedures were able to maintain Type I error rates within Bradley’s bounds under all conditions.

Power

Power rates (%) for J = 4 and J = 7 are presented in Table 4. When the variances were equal there was very little difference between the procedures, with the exception that the Welch test had reduced power for the g=1, h=0 distribution. In general, the power for the Welch test on ranks, the Brunner heteroscedastic nonparametric procedure and the Kruskal-Wallis procedure was slightly larger than that for either of the other Welch statistics. With unequal variances, the usual Welch test and the Welch test with trimmed means had deflated power relative to the remaining procedures for both nonnormal distributions, although the Brunner heteroscedastic nonparametric procedure had especially low power with negatively paired sample sizes and variances, particularly for J = 7. There was very little difference between the power rates of the Kruskal-Wallis and the Welch test on ranks. Caution, however, should be taken in interpreting the power rates of the Welch and Kruskal-Wallis procedures with negatively paired sample sizes and variances given that the Type I error rates were not adequately controlled in some of these conditions.

Pairwise MCPs

The pattern of familywise error and average per-pair power results for the MCPs were consistent across J = 4 and J = 7 and therefore only results for J = 7 are displayed and discussed. The all-pairs power rates for J = 4 are displayed and discussed. (The J = 7 rates were too low for meaningful comparisons.)

Type I error Control

Complete null familywise error rates (%) for the REGWQ and Tukey pairwise MCPs are presented in Table 5. The REGWQ procedure maintained rates within Bradley’s bounds under all conditions, with the exception that the procedure became conservative (i.e., empirical familywise error rates less than 2.5%) when it was used with either the Welch test or the Welch test on trimmed means and the data were g=1, h=0 distributed. The Tukey procedure maintained rates within Bradley’s limits when applied with Welch’s statistic on trimmed means, the Welch on ranks, or the Brunner-Munzel heteroscedastic statistic, although the Type I error rates became liberal when the Tukey procedure was applied with the usual Welch test or the Kruskal-Wallis test when sample sizes and variances were negatively paired.

Power

Average per-pair and all-pairs power rates for the REGWQ and Tukey pairwise MCPs are presented in Tables 6 and 7, respectively. Power rates overall were very low given the strict familywise error control and the inflated variances in the heteroscedastic conditions. There was very little difference in the overall
pattern of results for the Tukey and REGWQ procedures so given that the power was generally slightly larger for the REGWQ procedure (especially all-pairs power) only its results will be discussed. When the variances were equal, there was very little difference between the procedures in terms of per-pair or all-pairs power across all distributions, although the REGWQ procedure when applied with the Kruskal-Wallis statistic was generally the most powerful. When variances were unequal, the Welch test with trimmed means had less power than the Welch test on ranks or the Brunner-Munzel procedure across all distributions, with a slight advantage going to the Welch test on ranks (the usual Welch and Kruskal-Wallis procedures are not discussed because, when the variances were not equal, the Type I error rates were not controlled).

Table 3. Type I Error Percentages for n = 20 for the Welch test (Welch), the Welch test with trimmed means and Winsorized variances (Welch-t), the Welch test with ranked data (Welch-r), the Kruskal-Wallis nonparametric test (K-W) and the Brunner, Dette and Munk (1997) heteroscedastic nonparametric test (BDM).

<table>
<thead>
<tr>
<th>Normal Distribution</th>
<th>g=.5, h=0 Distribution</th>
<th>g=1, h=0 Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>= $\sigma^2_j$ PP  NP</td>
<td>= $\sigma^2_j$ PP  NP</td>
<td>= $\sigma^2_j$ PP  NP</td>
</tr>
<tr>
<td>J = 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Welch</td>
<td>5.3 5.1 5.3</td>
<td>5.9 5.7 7.2</td>
</tr>
<tr>
<td>Welch-t</td>
<td>5.7 5.5 5.8</td>
<td>5.6 5.3 6.2</td>
</tr>
<tr>
<td>Welch-r</td>
<td>5.7 6.4 6.6</td>
<td>5.7 5.9 6.4</td>
</tr>
<tr>
<td>K-W</td>
<td>4.8 4.0 7.8</td>
<td>4.8 3.9 7.9</td>
</tr>
<tr>
<td>BDM</td>
<td>6.8 7.0 7.0</td>
<td>6.8 7.2 7.1</td>
</tr>
</tbody>
</table>

| J = 7               |                        |                      |
| Welch               | 5.0 4.9 5.0            | 6.7 6.5 8.4          | 9.6 10.0 16.9         |
| Welch-t             | 6.2 5.7 6.4            | 6.0 5.8 6.7          | 6.0 6.1 7.5           |
| Welch-r             | 5.6 5.7 6.2            | 5.6 5.8 6.4          | 5.7 5.8 6.6           |
| K-W                 | 4.3 3.6 7.0            | 4.3 3.7 7.2          | 4.3 3.8 7.4           |
| BDM                 | 5.7 5.6 6.5            | 5.7 5.6 6.5          | 5.7 6.0 6.6           |

Note: $\sigma^2_j =$ equal population variances; PP and NP = positive and negative pairings of sample sizes and variances, respectively. Values exceeding Bradley's liberal limits (2.5% - 7.5%) are presented in bold.
Table 4. Power percentages for n = 20 for the Welch test (Welch), the Welch test with trimmed means and Winsorized variances (Welch-t), the Welch test with ranked data (Welch-r), the Kruskal-Wallis nonparametric test (K-W) and the Brunner, Dette and Munk (1997) heteroscedastic nonparametric test (BDM).

Normal Distribution  \( \sigma^2_j \) PP NP  g=.5, h=0 Distribution  \( \sigma^2_j \) PP NP  g=1, h=0 Distribution  \( \sigma^2_j \) PP NP

\[
\begin{array}{cccccc}
J &=& 4 & & & \\
Welch & 98.3 & 57.8 & 66.8 & 93.5 & 45.4 & 62.8 & 69.8 & 22.0 & \textbf{52.4} \\
Welch-t & 95.8 & 54.2 & 60.0 & 94.2 & 47.8 & 60.0 & 88.7 & 40.3 & 56.4 \\
Welch-r & 98.0 & 57.5 & 66.5 & 97.7 & 63.0 & 65.3 & 96.9 & 73.4 & 65.0 \\
K-W & 98.2 & 49.5 & \textbf{65.8} & 98.2 & 57.4 & \textbf{64.1} & 97.4 & 71.3 & \textbf{63.5} \\
BDM & 97.8 & 60.7 & 51.3 & 95.8 & 63.8 & 47.1 & 93.1 & 70.8 & 45.3 \\
\end{array}
\]

J = 7

\[
\begin{array}{cccccc}
Welch & 98.6 & 54.4 & 69.5 & 95.3 & 43.3 & \textbf{68.4} & \textbf{75.1} & \textbf{24.9} & \textbf{59.7} \\
Welch-t & 96.3 & 47.7 & 63.4 & 95.5 & 44.9 & 65.1 & 91.7 & 38.9 & 63.2 \\
Welch-r & 98.5 & 54.5 & 70.0 & 98.7 & 59.6 & 71.8 & 98.7 & 69.5 & 74.4 \\
K-W & 98.6 & 46.7 & 66.8 & 98.9 & 53.1 & 68.5 & 98.9 & 65.1 & 70.5 \\
BDM & 98.1 & 54.8 & 49.4 & 97.1 & 56.9 & 48.0 & 95.6 & 63.3 & 48.1 \\
\end{array}
\]

Note: \( \sigma^2_j \) = equal population variances; PP and NP = positive and negative pairings of sample sizes and variances, respectively. Conditions for which values exceeding Bradley's liberal limits (2.5% - 7.5%) are presented in bold.
Table 5. Type I Error Percentages for J = 7 and n = 20 for the Tukey and REGW MCPs with the Welch test (W), the Welch test with trimmed means and Winsorized variances (WT), the Welch test with ranked data (WR), the Kruskal-Wallis nonparametric test (KW) and the Brunner and Munzel (2000) heteroscedastic nonparametric test (BM).

<table>
<thead>
<tr>
<th>Normal Distribution</th>
<th>g=.5, h=0 Distribution</th>
<th>g=1, h=0 Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2_j$</td>
<td>PP</td>
<td>NP</td>
</tr>
<tr>
<td>Tukey-W</td>
<td>5.3</td>
<td>5.0</td>
</tr>
<tr>
<td>REGW-W</td>
<td>3.6</td>
<td>3.2</td>
</tr>
<tr>
<td>Tukey-WT</td>
<td>6.2</td>
<td>6.0</td>
</tr>
<tr>
<td>REGW-WT</td>
<td>4.3</td>
<td>3.5</td>
</tr>
<tr>
<td>Tukey-WR</td>
<td>5.9</td>
<td>6.0</td>
</tr>
<tr>
<td>REGW-WR</td>
<td>4.5</td>
<td>4.5</td>
</tr>
<tr>
<td>Tukey-KW</td>
<td>4.5</td>
<td>4.2</td>
</tr>
<tr>
<td>REGW-KW</td>
<td>3.3</td>
<td>3.0</td>
</tr>
<tr>
<td>Tukey-BM</td>
<td>5.7</td>
<td>5.2</td>
</tr>
<tr>
<td>REGW-BM</td>
<td>3.6</td>
<td>3.3</td>
</tr>
</tbody>
</table>

Note: $\sigma^2_j$ = equal population variances; PP and NP = positive and negative pairings of sample sizes and variances, respectively. Values exceeding Bradley's liberal limits (2.5% - 7.5%) are presented in bold.
Table 6. Per-Pair Power Percentages for $J = 7$ and $n = 20$ for the Tukey and REGW MCPs with the Welch test (W), the Welch test with trimmed means and Winsorized variances (WT), the Welch test with ranked data (WR), the Kruskal-Wallis nonparametric test (KW), and the Brunner and Munzel (2000) heteroscedastic nonparametric test (BM).

<table>
<thead>
<tr>
<th>Normal Distribution</th>
<th>g=.5, h=0 Distribution</th>
<th>g=1, h=0 Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_j^2$</td>
<td>PP</td>
<td>NP</td>
</tr>
<tr>
<td>Tukey-W</td>
<td>44.3</td>
<td>10.4</td>
</tr>
<tr>
<td>REGW-W</td>
<td>45.8</td>
<td>10.4</td>
</tr>
<tr>
<td>Tukey-WT</td>
<td>37.1</td>
<td>8.5</td>
</tr>
<tr>
<td>REGW-WT</td>
<td>36.4</td>
<td>7.8</td>
</tr>
<tr>
<td>Tukey-WR</td>
<td>44.1</td>
<td>10.0</td>
</tr>
<tr>
<td>REGW-WR</td>
<td>46.5</td>
<td>10.5</td>
</tr>
<tr>
<td>Tukey-KW</td>
<td>46.6</td>
<td>8.5</td>
</tr>
<tr>
<td>REGW-KW</td>
<td>50.9</td>
<td>9.1</td>
</tr>
<tr>
<td>Tukey-BM</td>
<td>43.7</td>
<td>9.1</td>
</tr>
<tr>
<td>REGW-BM</td>
<td>44.8</td>
<td>9.1</td>
</tr>
</tbody>
</table>

Note: $\sigma_j^2$ = equal population variances; PP and NP = positive and negative pairings of sample sizes and variances, respectively. Conditions for which values exceeding Bradley's liberal limits (2.5% - 7.5%) are presented in bold.

Table 7. All-Pairs Power Percentages for $J = 4$ and $n = 20$ for the Tukey and REGW MCPs with the Welch test (W), the Welch test with trimmed means and Winsorized variances (WT), the Welch test with ranked data (WR), the Kruskal-Wallis nonparametric test (KW), and the Brunner and Munzel (2000) heteroscedastic nonparametric test (BM).

<table>
<thead>
<tr>
<th>Normal Distribution</th>
<th>g=.5, h=0 Distribution</th>
<th>g=1, h=0 Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_j^2$</td>
<td>PP</td>
<td>NP</td>
</tr>
<tr>
<td>Tukey-W</td>
<td>28.8</td>
<td>4.2</td>
</tr>
<tr>
<td>REGW-W</td>
<td>36.0</td>
<td>7.4</td>
</tr>
<tr>
<td>Tukey-WT</td>
<td>22.4</td>
<td>3.1</td>
</tr>
<tr>
<td>REGW-WT</td>
<td>29.9</td>
<td>5.5</td>
</tr>
<tr>
<td>Tukey-WR</td>
<td>27.9</td>
<td>3.3</td>
</tr>
<tr>
<td>REGW-WR</td>
<td>35.1</td>
<td>6.1</td>
</tr>
<tr>
<td>Tukey-KW</td>
<td>31.4</td>
<td>3.1</td>
</tr>
<tr>
<td>REGW-KW</td>
<td>38.4</td>
<td>5.7</td>
</tr>
<tr>
<td>Tukey-BM</td>
<td>26.8</td>
<td>3.2</td>
</tr>
<tr>
<td>REGW-BM</td>
<td>33.5</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Note: $\sigma_j^2$ = equal population variances; PP and NP = positive and negative pairings of sample sizes and variances, respectively. Conditions for which values exceeding Bradley's liberal limits (2.5% - 7.5%) are presented in bold.
Conclusion

This article addressed the problem of testing for differences in the central tendency of independent groups with nonnormal (skewed) data and heterogeneous variances. This is an especially important issue for researchers in the behavioral sciences because these assumptions are rarely satisfied (e.g., Micceri, 1989; Keselman et al., 1998; Wilcox, 1988).

Of the omnibus tests evaluated in this paper, the Welch (1951) test with trimmed means and Winsorized variances, the Welch (1951) test on ranked data (Zimmerman & Zumbo, 1993a), and the Brunner heteroscedastic rank-based procedures (Brunner, Dette & Munk, 1997; Brunner & Munzel, 2000) provided superior Type I error control relative to the remaining procedures. The Type I error rates of the omnibus Welch test became liberal when distributions were skewed, and the Kruskal-Wallis test had liberal Type I error rates when variances were unequal (specifically when sample sizes and variances were negatively paired). These results concerning the liberal Type I error control of the Welch test with skewed and heteroscedastic data, and the Kruskal-Wallis procedure with unequal variances are consistent with previous reports (e.g., Algina, Oshima & Lin, 1994; Zimmerman & Zumbo, 1993a, 199b). With respect to power, there was very little difference between the procedures when the distributions were normal, although the power rates of the Welch test on ranks, the Brunner heteroscedastic nonparametric procedure, and the Kruskal-Wallis procedure were generally the largest.

These tests were also contrasted when they were applied to the set of all possible pairwise comparisons. In this case, the REGWQ MCP was able to maintain Type I error rates below Bradley’s upper liberal bound (7.5%) with all of the tests investigated. The test statistics with a Tukey critical value also maintained their empirical Type I error rates below Bradley’s upper liberal bound under most conditions; however, the Kruskal-Wallis statistic became slightly liberal when sample sizes and variances were negatively paired. These results are not unexpected given that the omnibus Kruskal-Wallis procedure also became liberal under these conditions. Adopting an REGWQ critical value generally resulted in more powerful tests than adopting a Tukey critical value, especially with respect to all-pairs power. Further, when the distributions were nonnormal, adopting an REGWQ critical value resulted in the largest power when used with one of the ranked data procedures (Welch on ranks, Kruskal-Wallis, or the Brunner & Munzel heteroscedastic nonparametric procedure).

In summary, when treatment distributions were skewed and variances heterogeneous, both the Welch (1938; 1951) tests with ranked data and the heteroscedastic, nonparametric procedures proposed by Brunner and colleagues (Brunner, Dette & Munk, 1997; Brunner & Munzel, 2000) provided good Type I error control (in both omnibus and pairwise multiple comparison settings). However, the Welch tests on ranked data are recommended as they were generally more powerful than the Brunner procedures. Further, the Welch tests on ranked data can easily be implemented in any software program that allows the user to rank the observations and run the Welch heteroscedastic procedures (e.g., SAS, SPSS, R).

References


Welch, B. L. (1938). The significance of the difference between two means when population variances are unequal. *Biometrika, 29*, 350-362.


