Journal of Modern Applied Statistical Methods

Volume 6 | Issue 1 Article 12

5-1-2007

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

Recommended Citation

Cribbie, Robert A.; Wilcox, Rand R.; Bewell, Carmen; and Keselman, H. J. (2007) "Tests for Treatment Group Equality When Data are Nonnormal and Heteroscedastic," *Journal of Modern Applied Statistical Methods*: Vol. 6 : Iss. 1 , Article 12.
DOI: 10.22237/jmasm/1177992660

Journal of Modern Applied Statistical Methods
May, 2007, Vol. 6, No. 1, 117-132

1538 – 9472/07/$95.00

Tests for Treatment Group Equality When Data are Nonnormal and Heteroscedastic

Robert A. Cribbie
York University

Rand R. Wilcox
University of Southern California

Carmen Bewell
York University

H. J. Keselman
University of Manitoba

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

Introduction

Researchers in the behavioral sciences are often interested in comparing the typical performance of subjects across independent groups, and they often select traditional test statistics (e.g., two-sample t, ANOVA F) without regard for their underlying assumptions, even though it has been pointed out that these assumptions may frequently be violated (e.g., Micceri, 1989; Keselman et al., 1998; Wilcox, 1988). Many authors have highlighted available procedures for analyzing data that violate either the assumption of normality or the assumption of variance homogeneity. Brown and Forsythe (1974), Kohr and Games (1974), and many others demonstrated the general effectiveness (i.e., Type I error control) of Welch's (1938, 1951) two-sample and omnibus test statistics with heterogeneous variances. In addition, Keselman, Cribbie and Zumbo (1997), Wilcox (1995; 1997), Yuen and Dixon (1973), and Zimmerman and Zumbo (1993a), among many others, have demonstrated the effectiveness of several alternatives to traditional parametric tests that can be used with nonnormal data, including nonparametric test statistics and tests with robust estimators (e.g., trimmed means).

However, there has been little success in discovering a test that is robust (with respect to Type I and Type II errors) to the simultaneous violations of both assumptions. That is, although procedures have been proposed for analyzing data that violate both the normality and variance heterogeneity assumptions concurrently (described below), there has not been a thorough investigation and comparison of the Type I error and power properties of these procedures. Therefore, the current article compares potential strategies for analyzing nonnormal and heteroscedastic data, with the goal of being able to recommend a procedure that provides a good balance between Type I error control and power.

One possibility for analyzing nonnormal and heteroscedastic data is to utilize the Welch two-sample and omnibus tests, which have been found to provide excellent Type I error control

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 non-normality 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 nonnnormal 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

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

$$t_w = \frac{\overline{X_j} - \overline{X_{j'}}}{\sqrt{\frac{s_j^2}{n_j} + \frac{s_{j'}^2}{n_{j'}}}}$$

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

$$v_w = \frac{\left(\frac{s_j^2}{n_j} + \frac{s_{j'}^2}{n_{j'}}\right)^2}{\left[\frac{s_j^4}{n_j^2(n_j - 1)} + \frac{s_{j'}^4}{n_{j'}^2(n_{j'} - 1)}\right]},$$

where $\overline{X}_j$, s^2_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:

$$F_w = \frac{\frac{\sum_j w_j \left(\overline{X_j} - \overline{X_j^*}\right)^2}{J - 1}}{1 + \left(\frac{2(J-2)}{J^2 - 1}\right)\left(\frac{\sum_j \left(1 - w_j / \sum_j w_j^2\right)^2}{n_j - 1}\right)},$$

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

$$w_j = \frac{n_j}{s_j^2},$$

$$\overline{X_j^*} = \frac{\sum_j w_j \overline{X_j}}{\sum_j w_j},$$

and

$$v_w = \frac{J^2 - 1}{3\sum_j \frac{\left(1 - w_j / \sum_j w_j\right)^2}{n_j - 1}}.$$

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 $a_j = \Sigma_i r_{ij}$. The statistic tests the null hypothesis H_o: $\lambda_1 = ... = \lambda_J$ (where λ represents the population mean only under the assumption that the population shapes are identical) and rejects H_o if $KW \geq \chi^2_{(J-1)}$ where:

$$KW = 12\left(\frac{\sum_{j} a_j^2 / n_j}{N(N+1)}\right) - 3(N+3),$$

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

$$t_{KW} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{(S_r - C)(N - 1 - KW)(n_j + n_{j'})}{n_j n_{j'}(N - J)(N - 1)}}},$$

$$S_r = \sum_i \sum_j r_{ij}^2$$

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 γ 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:

$$\overline{X_{tj}} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j - g_j} X_{ij}$$

and the jth sample Winsorized mean as

$$\overline{X_{wj}} = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$$

where:

$$\begin{aligned} Y_{ij} &= X_{(g_j+1)j} \quad \text{if } X_{ij} \leq X_{(g_j+1)j}, \\ &= X_{ij} \quad \text{if } X_{(g_j+1)j} < X_{ij} < X_{(n_j-g_j)j}, \\ &= X_{(n_j-g_j)j} \quad \text{if } X_{ij} \geq X_{(n_j-g_j)j}. \end{aligned}$$

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:

$$s_{wj}^2 = \frac{1}{h_j - 1} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{X_{wj}}\right)^2 .$$

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:

$$t_w = \frac{\overline{X_{tj}} - \overline{X_{tj'}}}{\sqrt{\dfrac{s_{wj}^2}{h_j} + \dfrac{s_{wj'}^2}{h_{j'}}}}$$

with error degrees of freedom,

$$v_w = \frac{\left(\frac{s_{wj}^2}{h_j}+\frac{s_{wj'}^2}{h_{j'}}\right)^2}{\left[\frac{s_{wj}^4}{h_j^2(h_j-1)}+\frac{s_{wj'}^4}{h_{j'}^2(h_{j'}-1)}\right]}.$$

Welch-r.

The Welch test can be performed on the ranked data, where ranks are established regardless of group membership. The null hypothesis, H_o: $\lambda_1 = ... = \lambda_J$, is rejected if $F_w \geq F_{\alpha, J-1, vw}$.

Brunner, Dette, and Munk

Brunner, Dette and Munk (1997) proposed the following heterogeneous, rank-based F statistic:

$$F_B = \frac{N}{tr(M_{11}V)} QMQ' ,$$

where

$$M = \underline{I} - \frac{1}{J}\underline{J} ,\ \underline{I} = \textit{identity matrix of rank J,}$$

$$\underline{J} = \textit{J by J matrix of 1s,}$$

$$V = N * diag\left(\frac{s_1^2}{n_1},\ldots,\frac{s_J^2}{n_J}\right),$$

$$s_j^2 = \frac{1}{N^2(n_j-1)}\sum_j (R_{ij}-\overline{R_j})^2 ,$$

$$Q = \frac{1}{N}\left(\overline{R_1}-\frac{1}{2},\ldots,\overline{R_j}-\frac{1}{2}\right),$$

$$\overline{R_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} R_{ij} ,$$

and R_{ij} is the rank of X_{ij} after the data are pooled. The null hypothesis, H_o: $\lambda_1 = ... = \lambda_J$ is rejected if $F_B \geq F_{\alpha, v1, v2}$ where:

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

$$v_2 = \frac{[tr(V)]^2}{tr(V^2\Lambda)}$$

and $\Lambda = diag\ \{(n_1 -1)^{-1}, ..., (n_J -1)^{-1}\}$. 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_o: $\mu_1 = ... = \mu_J$). The Welch-t evaluates the null hypothesis that all population trimmed means are equal (i.e., H_o: $\mu_{t1} = ... = \mu_{tJ}$). The K-W, Welch-r and BDM procedures evaluate the null hypothesis that all distribution functions are equal (i.e., H_o: $\lambda_1 = ... = \lambda_J$). 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_o: $\mu_j = \mu_{j'}$ ($j \neq j'$) if $|t| \geq q\ (\alpha, J, \nu) / (2)^{1/2}$, where q is a value from the Studentized range distribution with J groups and ν degrees of freedom, and t and ν 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 α, 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, ... , 2, and rejects H_o: $\mu_j = \mu_{j'}$ ($j \neq j'$) if an associated omnibus test has been rejected and:

$$|t| \geq q\,(\alpha_p, p, \nu) / (2)^{1/2},$$

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

Table 1. Sample Sizes and Population Variances Used in the Simulation Study.

J	Sample Sizes	Population Variances
4	10, 10, 10, 10	1, 1, 1, 1
	9, 10, 10, 11	1, 2, 4, 4
	5, 8, 12, 15	1, 3, 5, 8
	15, 15, 15, 15	
	13, 15, 15, 17	
	7, 12, 18, 23	
	20, 20, 20, 20	
	17, 20, 20, 23	
	9, 16, 24, 31	
7	10, 10, 10, 10, 10, 10, 10	1, 1, 1, 1, 1, 1, 1
	9, 9, 10, 10, 10, 11, 11	1, 1, 2, 2, 3, 3, 4
	5, 6, 8, 10, 12, 14, 15	1, 2, 2, 4, 7, 7, 8
	15, 15, 15, 15, 15, 15, 15	
	13, 14, 15, 15, 15, 16, 17	
	7, 9, 12, 15, 18, 21, 23	
	20, 20, 20, 20, 20, 20, 20	
	17, 18, 20, 20, 20, 22, 24	
	9, 12, 16, 20, 24, 28, 31	

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

Table 2. Population Mean Configurations Used in the Simulation Study.

		Population Means				
μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7
J = 4						
0.00	0.00	0.00	0.00			
0.00	0.00	0.00	1.28			
0.00	0.00	0.66	1.32			
0.00	0.00	1.09	1.09			
0.00	0.50	1.00	1.50			
J = 7						
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	1.30
0.00	0.00	0.00	0.00	0.00	1.04	1.04
0.00	0.00	0.00	0.51	1.02	1.02	1.02
0.00	0.00	0.62	0.62	0.62	1.24	1.24
0.00	0.23	0.46	0.69	0.92	1.15	1.38

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 μ_j and variance σ_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 σ_j, each X_{ij} was multiplied by a value of σ_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 σ_j. When working with trimmed means, the population trimmed mean for the jth group was also subtracted from the variate before multiplying by σ_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.

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).

	Normal Distribution			g=.5, h=0 Distribution			g=1, h=0 Distribution		
	= σ^2_j	PP	NP	= σ^2_j	PP	NP	= σ^2_j	PP	NP
				J = 4					
Welch	5.3	5.1	5.3	5.9	5.7	7.2	6.5	7.4	**13.5**
Welch-t	5.7	5.5	5.8	5.6	5.3	6.2	4.9	4.9	6.7
Welch-r	5.7	6.4	6.6	5.7	5.9	6.4	5.7	6.0	6.4
K-W	4.8	4.0	**7.8**	4.8	3.9	**7.9**	4.8	4.1	**7.9**
BDM	6.8	7.0	7.0	6.8	7.2	7.1	6.8	7.3	7.2
				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: = σ^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			g=.5, h=0 Distribution			g=1, h=0 Distribution		
	= σ^2_j	PP	NP	= σ^2_j	PP	NP	= σ^2_j	PP	NP
				J = 4					
Welch	98.3	57.8	66.8	93.5	45.4	62.8	69.8	22.0	**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	**65.8**	98.2	57.4	**64.1**	97.4	71.3	**63.5**
BDM	97.8	60.7	51.3	95.8	63.8	47.1	93.1	70.8	45.3
				J = 7					
Welch	98.6	54.4	69.5	95.3	43.3	**68.4**	**75.1**	**24.9**	**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

Note: = σ^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 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).

	Normal Distribution			g=.5, h=0 Distribution			g=1, h=0 Distribution		
	$= \sigma^2_j$	PP	NP	$= \sigma^2_j$	PP	NP	$= \sigma^2_j$	PP	NP
Tukey-W	5.3	5.0	5.0	4.2	4.5	7.2	2.5	4.8	**13.4**
REGW-W	3.6	3.2	2.6	3.2	2.9	3.3	1.3	1.4	3.6
Tukey-WT	6.2	6.0	6.3	5.1	5.1	6.4	3.0	3.5	6.5
REGW-WT	4.3	3.5	3.3	3.4	3.0	2.9	1.9	1.8	2.1
Tukey-WR	5.9	6.0	6.6	6.0	6.1	6.5	6.0	6.3	6.6
REGW-WR	4.5	4.5	4.2	4.5	4.5	4.3	4.5	4.6	4.5
Tukey-KW	4.5	4.2	**8.1**	4.5	4.2	**8.0**	4.5	4.3	**8.4**
REGW-KW	3.3	3.0	5.6	3.3	3.0	5.7	3.3	3.1	6.0
Tukey-BM	5.7	5.2	5.9	5.7	5.2	6.1	5.7	5.5	6.2
REGW-BM	3.6	3.3	3.2	3.6	3.4	3.2	3.6	3.6	3.4

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).

	Normal Distribution			g=.5, h=0 Distribution			g=1, h=0 Distribution		
	= σ^2_j	PP	NP	= σ^2_j	PP	NP	= σ^2_j	PP	NP
Tukey-W	44.3	10.4	14.3	34.5	6.2	14.7	17.0	1.6	**11.6**
REGW-W	45.8	10.4	9.7	34.7	5.8	10.1	13.2	1.1	5.7
Tukey-WT	37.1	8.5	12.1	35.1	7.1	13.5	29.7	4.9	13.4
REGW-WT	36.4	7.8	7.4	34.5	6.4	8.8	31.9	4.0	7.9
Tukey-WR	44.1	10.0	16.8	44.8	11.5	16.3	43.8	14.8	16.2
REGW-WR	46.5	10.5	14.8	46.8	12.0	14.2	45.6	15.5	14.0
Tukey-KW	46.6	8.5	**17.0**	47.7	10.0	**16.7**	46.8	13.9	**16.7**
REGW-KW	50.9	9.1	18.3	52.4	10.7	18.0	51.7	15.1	18.0
Tukey-BM	43.7	9.1	13.9	40.2	10.4	12.2	35.4	12.9	11.3
REGW-BM	44.8	9.1	8.8	39.5	9.8	7.4	33.5	11.9	6.7

Note: = σ^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 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).

	Normal Distribution			g=.5, h=0 Distribution			g=1, h=0 Distribution		
	= σ^2_j	PP	NP	= σ^2_j	PP	NP	= σ^2_j	PP	NP
Tukey-W	28.8	4.2	2.3	17.2	1.6	2.1	4.3	0.2	**1.2**
REGW-W	36.0	7.4	4.6	23.8	3.2	3.7	7.1	0.4	2.1
Tukey-WT	22.4	3.1	1.5	18.4	2.0	1.7	12.6	1.0	1.5
REGW-WT	29.9	5.5	3.1	25.3	3.9	3.2	18.0	2.2	2.7
Tukey-WR	27.9	3.3	3.8	26.7	3.7	3.1	23.8	5.5	2.7
REGW-WR	35.1	6.1	6.5	33.7	6.8	5.6	30.9	9.5	5.1
Tukey-KW	31.4	3.1	**6.0**	31.2	3.8	**5.1**	28.3	6.1	**4.6**
REGW-KW	38.4	5.7	9.7	38.2	6.8	8.5	35.7	10.4	7.9
Tukey-BM	26.8	3.2	1.8	21.6	3.0	1.1	15.8	3.7	0.8
REGW-BM	33.5	5.6	3.4	28.5	5.5	2.3	22.7	6.7	1.8

Note: = σ^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.

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

Algina, J., Oshima, T. C., & Lin, W.-Y. (1994). Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups. *Journal of Educational and Behavioral Statistics, 19,* 275-291.

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

Brown, M. B. & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics, 16*, 129-132.

Brunner, E., Dette, H. & Munk, A. (1997). Box type approximations in nonparametric factorial designs. *Journal of the American Statistical Association, 92*, 1494-1502.

Brunner, E. & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and a small sample approximation. *Biometrical Journal, 42*, 17-25.

Cressie, N. A. C. & Whitford, H. J. (1986). How to use the two sample *t*-test. *Biometrical Journal, 28*, 131-148.

Einot, I. & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association, 70*, 574-583.

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. 461-513). New York: Wiley.

Keselman, H. J., Cribbie, R. A. & Zumbo, B. D. (1997). Specialized tests for detecting treatment effects in the two-sample problem. *Journal of Experimental Education, 65*, 355-366.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C. & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.

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

Kohr, R. L. & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. *Journal of Experimental Education, 43*, 61-69.

Kruskal, W. H. & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*, 583-621.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Mudholkar, A., Mudholkar, G. S. & Srivastava, D. K. (1991). A construction and appraisal of pooled trimmed-t statistics. *Communications in Statistics: Theory and Methods, 20*, 1345-1359.

Munzel, U. & Hothorn, L. A. (2001). A unified approach to simultaneous rank test procedures in the unbalanced one-way layout. *Biometrical Journal, 43*, 553-569.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 71,* 308-313.

Olsson, D. M. (1974). A sequential simplex program for solving minimization problems. *Journal of Quality Control, 6,* 53-57.

Olsson, D. M. & Nelson, L. S. (1975). The Nelder-Mead simplex procedure for function minimization. *Technometrics, 17,* 45-51.

Ryan, T. A. (1960). Significance tests for multiple comparison of proportions, variances, and other statistics. *Psychological Bulletin, 57*, 318-328.

SAS Institute, Inc. (1999). *SAS/IML user's guide, Version 8*. Cary, NC: SAS Institute Inc..

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics, 2*, 110-114.

Sprent, P. (1993). *Applied nonparametric statistical methods (2nd ed.)*. London: Chapman & Hall.

Toothaker, L. E. (1991). *Multiple comparisons for researchers*. Newbury Park, CA: Sage Publications Inc.

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

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38*, 330-336.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association, 72*, 566-575.

Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results on James' second order method. *British Journal of Mathematical and Statistical Psychology, 41*, 109-117.

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. (1997). Three multiple comparison procedures for trimmed means. *Biometrical Journal, 37*, 643-656.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. New York: Elsevier Academic Press.

Yuen, K. K. & Dixon, W. J. (1973). The approximate behavior of the two-sample trimmed *t*. *Biometrika, 60*, 369-374.

Zimmerman, D. W. (1996). A note on homogeneity of variance of scores and ranks. *Journal of Experimental Education, 64*, 351-362.

Zimmerman, D. W. (1987). Comparative power of the Student t test and Mann-Whitney U-test for unequal sample sizes and variances. *Journal of Experimental Education, 55*, 171-174.

Zimmerman, D. W. & Zumbo, B. D. (1993a). The relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*. New Jersey: Lawrence Erlbaum Associates.

Zimmerman, D. W. & Zumbo, B. D. (1993b). Rank transformations and the power of the Student t test and Welch t' test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology, 47*, 523-529.