DOCUMENT RESUME

ED 384 622 TM 023 615

AUTHOR Lix, Lisa M.; Keselman, H. J.
TITLE To Trim or Not To Trim: Tests of Location Equality under Heteroscedasticity and Nonnormality.
PUB DATE Apr 95
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Estimation (Mathematics); Foreign Countries; *Least Squares Statistics; *Research Design; *Statistical Distributions
IDENTIFIERS *Equality (Mathematics); *Heteroscedasticity (Statistics); Mean (Statistics); Nonnormal Distributions; Population; Trimmed Means; Type I Errors

ABSTRACT

Tests of mean equality proposed by Alexander and Govern (1994) and Tsakok (1978) were compared to the well-known procedures of Brown and Forsythe (1974), James (1951), and Welch (1951) for their ability to limit the number of Type I errors in one-way designs where the underlying distributions were nonnormal, variances were nonhomogeneous, and group sizes were unequal. These tests were compared when the usual method of least squares was applied to estimate group means and variances and when adopting Yuen's (1974) trimmed means and winsorized variances. In the former case, the procedures can be used to test for population mean equality, while in the latter case they can be used to test for equality of the population trimmed means. Based on the variables examined in this investigation, which included numbers of treatment groups, degree of population skewness, and type of pairing of variances and group sizes, it is recommended that applied researchers utilize trimmed means and winsorized variances with Tsakok's test, since its rates of Type I error were closest to the nominal level of significance, ranging in value from 4.5% to 6.62%. However, it must be remembered that by adopting this strategy, one is testing for equality of population trimmed means, not the equality of population means. (Contains 30 references and 4 tables.) (Author/SLD)

TO TRIM OR NOT TO TRIM: TESTS OF LOCATION EQUALITY UNDER HETROSCEDASTICITY AND NONNORMALITY

Lisa M. Lix and H. J. Keselman
Department of Psychology
University of Manitoba

Paper presented at the Annual Meeting of The American Educational Research Association, 1995, San Fransisco

Abstract

Tests of mean equality proposed by Alexander and Govern (1994) and Tsakok (1978) were compared to the well-known procedures of Brown and Forsythe (1974), James (1951) and Welch (1951) for their ability to limit the number of Type I errors in one-way designs where the underlying distributions were nonnormal, variances were nonhomogeneous, and groups sizes were unequal. These tests were compared when the usual method of least squares was applied to estimate group means and variances and when adopting Yuen's (1974) trimmed means and windsorized variances. In the former case the procedures can be used to test for population mean equality, while in the latter case they can be used to test for equality of the population trimmed means. Based on the variables examined in this investigation, which included number of treatment groups, degree of population skewness, and type of pairing of variances and group sizes, we recommend that applied researchers utilize trimmed means and winsorized variances with Tsakok's test, since its rates of Type I error were closest to the nominal level of significance, ranging in value from 4.50% to 6.62%. However, it must be remembered that by adopting this strategy one is testing for equality of population trimmed means not the equality of population means.

TO TRIM OR NOT TO TRIM: TESTS OF LOCATION EQUALITY UNDER HETEROSCEDASTICITY AND NONNORMALITY

Introduction

Testing for mean equality in the presence of unequal variances has a long history in the statistical literature dating back to the time of Behrens (1929) and Fisher (1935). Since this early work, numerous authors have offered potential solutions to the problem. Welch (1951) presented an approximate degrees of freedom solution for a nonpooled statistic in the one-way completely randomized design. Two other solutions that are frequently recommended in the literature are the James (1951) second order and Brown and Forsythe (1974) approximation methods. The empirical literature indicates that all of these procedures generally control the rate of Type I error when group variances are heterogeneous and the data are normally distributed (e.g., Dijkstra & Werter 1981; Oshima & Algina, 1992; Wilcox, 1988). However, the literature also indicates that these tests can become liberal when variance heterogeneity exists in combination with unequal group sizes and the data are nonnormal in form (e.g., Oshima & Algina, 1992). Thus, these statistics have limitations, namely their sensitivity to the nature of the population distributions.

Recently, Alexander and Govern (1994) proposed another statistic that may be applied to test for mean equality in the presence of variance heterogeneity; their solution, like that of James (1951), is based on large sample theory and utilizes a χ^2 statistic. A less well known alternative was suggested by Tsakok (1978). Tsakok's approach, which involves the computation of multiple one sample t statistics is appealing because it represents an exact solution. To date, neither of these procedures has been investigated for the effects of nonnormality, nor for the combined effects of nonnormality and variance heterogeneity, particularly when group sizes are unequal.

With regard to the effects of nonnormality, numerous authors have suggested ways in which treatment groups may be compared when the underlying distributions are nonnormal and variances are heterogeneous (See Wilcox, 1990; 1994). Specifically, censoring or trimming the data, as suggested by Yuen (1974), can be utilized when comparing groups for treatment effects when the treament group populations are not normal in form (Yuen, 1974; Yuen & Dixon, 1973; Wilcox, 1992, 1993). This procedure has generally been recommended in order to obtain a more powerful test for group differences as compared to the traditional procedures which lose power when distributions are nonnormal. However, little is known about the Type I error control characteristics of Yuen's method, particularly when applied with the previously enumerated statistical tests, and to treatment group data that is neither normal in form nor equal in variability.

Therefore, the purpose of our investigation was to determine whether it may be possible to use trimmed means and winsorized variances with the Alexander and Govern (1994), Brown and Forsythe (1974), James (1951), Tsakok (1978), and Welch (1951)

statistics in order to obtain a robust test for mean equality when the data are both heterogeneous and nonnormal in form and group sizes are unequal.[1]

Definition of the Test Statistics

Suppose n_j independent random observations $X_{i1}, X_{i2}, \ldots, X_{in_j}$ are sampled from population j (j = 1, ..., J). We assume that the X_{ij}'s are obtained from a normal population with mean μ_j and unknown variance σ_j^2, with $\sigma_j^2 \neq \sigma_{j'}^2$ ($j \neq j'$). Then, let $\bar{X}_j = \Sigma_i X_{ij}/n_j$ and $s_j^2 = \Sigma_j(X_{ij} - \bar{X}_j)^2/(n_j - 1)$, where $\bar{X}_j$ is the estimate of μ_j and s_j^2 is the usual unbiased estimate of the variance for population j. Further, let the standard error of the mean be denoted as $S_j = (s_j^2/n_j)^{\frac{1}{2}}$ and let $w_j = 1/S_j^2/(\Sigma_j 1/S_j^2)$.

The procedures presented by Alexander and Govern (1994), Brown and Forsythe (1974), James (1951), Tsakok (1978), and Welch (1951) for testing the null hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_J$ in the presence of variance heterogeneity may all be obtained from a single general result. That is, if we denote the variance weighted estimate of the grand mean as $\hat{\mu} = \sum_{j=1}^{J} w_j \bar{X}_j$, a one − sample t statistic can be computed for each group,

$$t_j = \frac{\bar{X}_j - \hat{\mu}}{S_j}. \qquad (1)$$

This statistic is distributed as a t variable with $\nu = n_j - 1$ degrees of freedom. In order to test the null hypothesis of mean equality, Welch (1951), James (1951), and Brown and Forsythe (1974) derived statistics which relate to $\Sigma_j t_j^2$ (see Alexander & McGovern, 1994 for the definition of these approximate statistics). These test statistics reference either the chi square or F distributions.

In Alexander and Govern's (1994) solution, a normalizing transformation is first applied to each t_j. These normalized values (i.e., z scores) are then used to derive a statistic ($\Sigma_j z_j^2$) that is distributed as a chi square variable.

Tsakok (1978), on the other hand, demonstrated how one can obtain an exact test within this context using the statistic presented in Equation 1. To test H_0, consider the following j sub-hypotheses:

$$H_{0j}: \mu_j = \mu, \ [H_{A_j}: \mu_j \neq \mu] \qquad j = 1, \ldots, J.$$

While, Tsakok did not indicate how μ should be estimated, we use the variance weighted estimate of the grand mean. The statistic t_j is distributed as a t variable so that $\alpha_j = P(|t_j| > t)$, and the probability that all H_{0j} will be accepted (assuming that they are

true) is $\prod_j(1-\alpha_j)$. The probability of making a Type I error in testing the set of hypotheses that the μ_j have a common value μ, is $\alpha = 1 - \prod_j(1-\alpha_j) \doteq \sum_j \alpha_j$. Similarly, the power of the exact test is given by $1 - \beta = 1 - \prod_j \beta_j$, where $\beta_j = 1 - P(|t_j| < t \mid H_{0_j}$ is false). Type I error protection for the set of J tests can be obtained by adopting a Bonferroni method. In particular, in this investigation, the method due to Sidak (1967) was employed.

Another consideration in this paper was the application of robust estimates of the group means and variances to these various test procedures. Since it is well known that the traditional methods of analysis, e.g., the analysis of variance (ANOVA) F test or *t* test, will lose power when the data are obtained from nonnormal distributions, numerous authors have recommended that robust estimators be substituted for the least squares estimators (Gross, 1976; Yuen, 1974; Yuen & Dixon, 1973; Wilcox, 1992, 1993, 1995). However, little information is available regarding either the robustness of these estimators to assumption violations or the effect of these assumption violations on rates of Type I error. In the present investigation, trimmed means and windsorized standard deviations were utilized when computing the test statistics (See Wilcox, 1994).[2] Specifically, we utilized asymmetric trimming, trimming only in the upper tail (i.e., 20%) associated with a particular group. Asymmetric trimming was examined since it is theorized to be potentially advantageous when the distributions are known to be skewed (see Tiku, 1980, 1982; De Wet & van Wyk, 1979).

Method

In summary, twelve tests for mean equality were compared for their rates of Type I error under conditions of nonnormality and variance heterogeneity in independent groups designs. The twelve tests resulted from crossing the Alexander and Govern (1994), Brown and Forsythe (1974), James (1951) second order, Tsakok (1978), Welch (1951), and usual ANOVA F statistics with two methods for estimating the group means and variances, Yuen's method, which uses a trimmed mean and winsorized variance (see Yuen & Dixon, 1973; Wilcox, 1993), and the usual least squares estimators for the mean and variance. The ANOVA F test was included only to serve as a baseline measure for comparison of Type I error rates.

Three factors were varied in the study: (a) number of groups (2, 4, and 6), (b) population distribution (χ^2_3 and χ^2_6), and (c) pairing of unequal variances and group sizes (positive and negative).

We chose to investigate completely randomized designs containing two, four, and six groups since previous research looked at these conditions(e.g., Wilcox, 1988). Most of the investigated conditions were selected because they were employed in previous studies (e.g., Dijkstra & Werter, 1981; Wilcox, Charlin & Thompson, 1986; Oshima & Algina, 1992) and thus allowed us to compare the procedures under conditions which are

known to highlight the strength and weaknesses of tests for location equality. Table 1 contains the numerical values of the sample sizes and variances investigated in this study and the studies from which these conditions were obtained.

Insert Table 1 About Here

With respect to the effects of distributional shape on Type I error, we chose to investigate conditions in which the data were obtained from chi-square distributions. To investigate the effects of skewness, we generated χ^2_3 and χ^2_6 variates. These particular types of nonnormal distribution were selected since educational and psychological research data typically have skewed distributions (Micceri, 1989). Furthermore, Sawilowsky and Blair (1992) investigated the effects of eight nonnormal distributions identified by Micceri on the robustness of Student's t test and found that only distributions with extreme skewness (e.g., $\gamma_1 = 1.64$) were found to affect the Type I error control of the independent sample t statistic. For the χ^2_3 distribution, skewness and kurtosis values are $\gamma_1 = 1.63$ and $\gamma_2 = 4.00$, respectively. The χ^2_6 distribution was included in our investigation in order to examine the effects of sampling from a distribution with moderate skewness. For this distribution, $\gamma_1 = 1.16$ and $\gamma_2 = 2.00$.

The third factor manipulated was the nature of the pairing of the group sizes and variances. Specifically, we choose to investigate both positive and negative pairings. For positive (negative) pairings, the group having the fewest (greatest) number of observations was associated with the population having the smallest (largest) variance, while the group having the greatest (fewest) number of observations was associated with the population having the largest (smallest) variance. These conditions were chosen since they typically produce conservative (liberal) results.

We investigated asymmetric trimming since symmetric trimming is based on the removal of outliers from symmetric distributions, while asymmetric trimming has been recommended by Tiku (1908, 1982) and others (e.g., See De Wet & van Wyk, 1979) for nonsymmetric skewed distributions as a means of reducing the effects of deviant observations in the longer tail.

To generate pseudorandom variates having a χ^2 distribution with three (six) degrees of freedom, three (six) standard normal variates, generated using the SAS (SAS Institute, 1989) generator RANNOR, were squared and summed. The variates were standardized, and then transformed to χ^2_3 or χ^2_6 variates having mean μ_j (when comparing the tests based on the least squares estimates) or μ_{jt} (when comparing the tests based on trimed means) and variance σ^2_j. [See Hastings & Peacock (1975), pp. 46-51, for further details on the generation of data from these distributions].

Five thousand replications of each condition were performed using a .05 significance level. For all tests, with the exception of Tsakok's (1978), a Type I error occurred when the value of the observed statistic exceeded its .05 critical value. For the Tsakok test, a Type I error occurred if at least one of the subhypothesis tests exceeded its Bonferroni critical value.

Results

To evaluate the particular conditions under which a test was robust to assumption violations, we use the criterion that the empirical rate must be contained in the 99% confidence interval for α. According to this criterion, based on $\alpha = .05$, in order for a test to be considered robust, its empirical rate of Type I error ($\hat{\alpha}$) must be contained in the interval $.042 \leq \hat{\alpha} \leq .058$. We adopted this criterion as a means of identifying only those procedures which could provide strict control of the error rate. Correspondingly, a test was considered to be nonrobust if, for a particular condition, its Type I error rate was not contained in this interval. In the tables, bolded entries are used to denote these latter values. In discussing the results, the tests due to Alexander and Govern (1994), Brown and Forsythe (1974), James (1951), Tsakok (1978), and Welch (1951) will be referred to as the AG, BF, J, T, and W tests, respectively and the ANOVA F test will be denoted by the abbreviation F.

The J = 2 results are contained in Table 2. As expected, the error rates associated with F deviated greatly from α under both methods of estimation of means and variances, and were predictably conservative and liberal for positive and negative pairings of variances and group sizes, respectively. Among the remaining procedures, when the simulated data were obtained from the χ^2_6 distribution and least squares estimates of central tendency and variability were employed, the rates associated with all but the T test were within the 99% interval when the variances and group sizes were positively paired (Condition A); the T value was conservative. For negative pairings of variances and group sizes (Condition B), the rates all exceeded the upper bound of 5.80%, with the T procedure having the lowest (5.86%) of these liberal rates. Not surprisingly, the corresponding rates using trimmed means and winsorized variances were closer to α. Specifically, for positive pairings of variances and group sizes the T test was again conservative, while for the negative pairing case, all values were within the interval of 4.20% to 5.80%.

Insert Table 2 About Here

The more extreme case of skewness investigated (i.e., χ^2_3) resulted in rates that were larger than those associated with a moderate degree of skewness. Thus, when variances and group sizes were positively paired and least squares estimation was employed, the rates associated with all of the F alternatives, were within the bounds of the 99% interval. On the other hand, all values were liberal when variances and group sizes were negatively paired; the minimum value was 7.16% (T). The values obtained when employing trimmed means and winsorized variances were, with the exception of

the T (3.42%) value, within the 99% interval for positive pairings of variances and group sizes, while for the negative pairing case all values, except for T, were liberal. The BF, W, J, AG, and T values were 6.26%, 6.26%, 6.24%, 6.26%, and 5.32%, respectively.

Table 3 contains the percentages of error for the four group design. When sampling from the χ^2_6 distribution the BF, W, J, and AG values were liberal when least squares estimation was employed and variances and group sizes were positively paired (Condition C). Excluding the value for the BF test, all values were again liberal when the variances and group sizes were negatively paired (Condition D). Once again, the rates obtained when employing trimmed means and winsorized variances were generally smaller than their least squares counterparts. That is, for positive pairings of variances and group sizes only the BF procedure was liberal (6.76%), though the F and T resulted in conservative tests, with rates of 4.06% and 3.74%, respectively. For the negative pairing case, only the BF and T values were within the 99% interval. The F, BF, W, J, AG, and T values were 12.44%, 5.68%, 6.44%, 6.04%, 6.42%, and 5.06%, respectively.

Insert Table 3 About Here

When the simulated data were from the χ^2_3 distribution, the majority of empirical values fell outside the bounds of the 99% interval. Thus, when least squares estimates were utilized all but two liberal values [ANOVA F and T (Condition C)] were obtained for both cases of pairings of variances and group sizes. This liberalness of values was also generally present when trimmed means and winsorized variances were utilized; however, for positive pairings, the F and T tests resulted in conservative rates of 3.64% and 3.94%, respectively. Of all the liberal rates the T test produced the smallest empirical value (i.e., 6.08%). For negative pairings of variances and group sizes (Condition D) the F, BF, W, J, and AG values were 12.80%, 5.86%, 8.10%, 7.80%, and 8.08%, respectively.

Table 4 contains the empirical rates of error when there were six treatment groups. The pattern of values in Table 4 is similar to that found and enumerated in Tables 2 and 3. That is, when least squares estimation was employed, the values were rarely contained within the 99% confidence interval and typically were very liberal. On the other hand, the values obtained when utilizing trimmed means and winsorized variances were considerably smaller. Additionally, the error rates obtained when sampling from the χ^2_6 distribution generally deviated less from α than those obtained when the simulated data were distributed as χ^2_3.

Insert Tables 4 About Here

With regard to the performance of the tests when utilizing trimmed means and winsorized variances, with the exception of BF, all had rates which were well controlled

when variances and group sizes were positively paired (Condition E) and the data were only moderately skewed. However, for the negative pairing case and this same degree of nonnormality, only T was not liberal (Condition F). Specifically, for condition F the F, BF, W, J, AG, and T values were 13.76%, 6.80%, 7.00%, 6.48%, 6.50%, and 5.76%, respectively.

When sampling from the more extreme chi-square distribution, the rates were typically outside the 99% interval and were generally liberal in value regardless of the nature of the pairing of variances and group sizes. The exceptions were the ANOVA F and T values, which were within the interval for positive pairings of variances and group sizes. In addition, of all the condition F liberal values, the T value (6.62%) was smallest. The remaining tests, F, BF, W, J, and AG had values of 14.30%, 6.86%, 8.34%, 7.54%, and 7.54%, respectively.

Discussion

This investigation compared six procedures that can be used to test for location equality among two or more groups when population variances are heterogeneous. When utilizing group means and variances, these procedures test for the equality of population mean equality, while the use of trimmed means and winsorized variances results in tests of equality of population trimmed means.

Results from our study indicate that when the homogeneity of variances and normality assumptions are not satisfied and the design is unbalanced, the use of trimmed means and winsorized variances results in better Type I error control as compared with the use of the usual least squares estimates of the mean and variance.

Based upon the reported findings and the conditions investigated, we recommend Tsakok's (1978) procedure since its rates of Type I error were closest to the nominal level of significance. According to the criterion of robustness employed in this investigation, all of the procedures generally resulted in liberal rates of error when variances and group sizes were negatively paired, a condition known to adversely affect a test's rates of Type I error. However, the rates associated with Tsakok's procedure were, for this condition, always closer to the nominal five percent value than the other procedures' rates, ranging in value from 4.50% to 6.62%. Thus, if a researchers goal is to employ a test statistic that can limit the number of Type I errors across conditions known to produce conservative or liberal tests (positive and negative pairings of variances and group sizes) then Tsakok's test with trimmed means and winsorized variances seems, at this time, best to achieve this goal.

Footnotes

1. Wilcox's (1989) alternative (H_m) to the James (1951) second order test was not included in this investigation as Hsiung, Olejnik, & Huberty (1994) show that the procedure is not always valid.

2. When trimmed means are being compared the null hypothesis(ses) pertain to the equality of population trimmed means, i.e., the μ_ts. That is, $H_0: \mu_{t1} = \mu_{t2} = \ldots = \mu_{tJ}$ and $H_{0j}: \mu_{tj} = \mu_t$, $[H_{A_j} : \mu_{tj} \neq \mu_t]$. Further, $\widehat{\mu} = \sum_{j=1}^{J} w_j \bar{X}_{tj}$, and $w_j = 1/S^2_{wj}/(\Sigma_j 1/S^2_{wj})$, where S^2_{wj} is the standard error of the trimmed mean based upon the winsorized variance (See Wilcox, 1994, p. 61).

References

Alexander, R.A., & Govern, D.M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. Journal of Educational Statistics, 19, 91 – 101.

Behrens, W.V. (1929). Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. Landwirtsch Jahrbucher, 68, 807 – 837.

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.

De Wet, T., & van Wyk, J.W.J. (1979). Efficiency and robustness of Hogg's adaptive trimmed means. Communications in Statistics. Theory and Methods, A8(2), 117 – 128.

Dijkstra, J.B., & Werter, P.S.P.J. (1981). Testing the equality of several means when the population variances are unequal. Communications in Statistics, Simulation and Computation, B10(6), 557 – 569.

Fisher, R.A. (1935). The fiducial argument in statistical inference. Annals of Eugenics, 6, 391 – 398.

Gross, A. M. (1976). Confidence interval robustness with long – tailed symmetric distributions. Journal of the American Statistical Association, 71, 409 – 416.

Hsiung, Tung-Hsing, Olejnik, S., & Huberty, C.J. (1994). Comment on a Wilcox test statistic comparing means when variances are unequal. Journal of Educational Statistics, 19, 111 – 118.

Hastings, N. A. J., & Peacock, J. B. (1975). Statistical distributions: A handbook for students and practioners. New York: Wiley.

James, G.S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. Biometrika, 38, 324 – 329.

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

Oshima, T.C., & Algina, J. (1992). Type I error rates for James's second – order test and Wilcox's H_m test under heteroscedasticity and non – normality. British Journal of Mathematical and Statistical Psychology, 45, 255 – 263.

Sawilowsky, S.S., & Blair, R.C. (1992). A more realistic look at the robustness and Type II error probabilities of the *t* test to departures from population normality. Psychological Bulletin, 111, 352360.

SAS Institute Inc. (1989). SAS/IML software: Usage and reference, version 6 (1st ed.). Cary, NC: Author.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association, 62, 626 – 633.

Tabatabai, M.A., & Tan, W.Y. (1988). A robust test for multi – sample location problems with unequal group variances and nonnormality. Biometrical Journal, 30, 275 – 281.

Tiku, M.L. (1980). Robustness of MML estimators based on censored samples and robust test statistics. Journal of Statistical Planning and Inference, 4, 123 – 143.

Tiku, M.L. (1982). Robust statistics for testing equality of means and variances. Communications in Statistics, Theory and Methods, 11(22), 2543 – 2558.

Tsakok A.D. (1978). A solution to the generalized Behrens – Fisher problem. Metron, 36(3 – 4), 79 – 91.

Welch, B.L. (1961). On the comparison of several mean values: An alternative approach. Biometrika, 38, 330 – 336.

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

Wilcox, R.R. (1989). Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA models. Journal of Educational Statistics, 14, 269 – 278.

Wilcox, R.R. (1990). Comparing the means of two independent groups. Biometrics Journal, 32, 771 – 780.

Wilcox, R.R. (1992). An improved method for comparing variances when distributions have non-identical shapes. Computational Statistics & Data Analysis, 13, 163 – 172.

Wilcox, R.R. (1993). Robustness in ANOVA. In L.K. Edwards (Ed.), Applied analysis of variance in behavioral science (pp. 34374). New York: Marcel Dekker.

Wilcox, R.R. (1994). ANOVA: A paradigm for low power and misleading measures of effect size? Review of Educational Research, 65, 51 – 77.

Wilcox, R.R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? Review of Educational Research, 65(1), 51 – 77.

Wilcox, R.R., Charlin, V.L., & Thompson, K.L. (1986). New Monte Carlo results on the robustness of the ANOFA F, W and F* statistics. Communications in Statistics. Simulation, 15(4), 933 – 943.

Yuen, K.K. (1974). The two – sample trimmed t for unequal population variances. Biometrika, 61, 165 – 170.

Yuen, K.K., & Dixon, W.J. (1973). The approximate behaviour and performance of the two – sample trimmed t. Biometrika, 60, 369 – 374.

Table 1. Investigated Conditions		
Condition[a]	Sample Sizes	Population Variances
A	11, 21	1, 16
B	11, 21	16, 1
C	8, 10, 12, 14	1, 4, 9, 16
D	8, 10, 12, 14	16, 9, 4, 1
E	8(2), 12(3), 14	1(2), 4, 9(2), 16
F	8(2), 12(3), 14	16, 9(2), 4, 1(2)
[a]Note:A,B,C,D-investigated by Wilcox, Charlin & Thompson (1986)		
E,F-investigated by Dijkstra & Werter (1981)		

Table 2. Empirical Percentages of Type I Error (J = 2)						
	χ^2_6					
Condition	F	BF	W	J	AG	T
	Least Squares Estimation					
A	**1.90**	5.62	5.62	5.62	5.48	**3.44**
B	**16.50**	**6.80**	**6.80**	**6.80**	**6.80**	**5.86**
	Yuen's Trimmed Means					
A	**1.62**	4.46	4.46	4.52	4.44	**3.28**
B	**15.72**	5.22	5.22	5.22	5.22	4.50

	χ^2_3					
Condition	Least Squares Estimation					
A	**2.40**	6.16	6.16	6.18	6.06	4.22
B	**17.52**	**8.32**	**8.32**	**8.32**	**8.32**	**7.16**
	Yuen's Trimmed Means					
A	**2.00**	5.64	5.64	5.64	5.38	**3.42**
B	**16.80**	**6.26**	**6.26**	**6.24**	**6.26**	5.32

Table 3. Empirical Percentages of Type I Error (J = 4)						
	F	BF	W	J	AG	T
Condition	χ^2_6					
Least Squares Estimation						
C	4.84	**6.68**	**7.00**	**6.94**	**6.78**	5.32
D	**11.74**	5.72	**7.76**	**7.52**	**7.52**	**6.66**
Yuen's Trimmed Means						
C	**4.06**	**6.76**	5.68	5.60	5.52	**3.74**
D	**12.44**	5.68	**6.44**	**6.04**	**6.42**	5.06
	χ^2_3					
Least Squares Estimation						
C	4.50	**6.08**	**7.14**	**7.10**	**7.02**	5.36
D	**11.94**	**6.64**	**10.16**	**9.86**	**9.96**	**8.22**
Yuen's Trimmed Means						
C	**3.64**	5.76	**6.12**	**6.02**	**6.04**	**3.94**
D	**12.80**	**5.86**	**8.10**	**7.80**	**8.08**	**6.08**

Table 4. Empirical Percentages of Type I Error (J = 6)						
	F	BF	W	J	AG	T
Condition	χ^2_6					
Least Squares Estimation						
E	4.64	**6.70**	**6.50**	**6.28**	**6.50**	**6.08**
F	**12.58**	**6.74**	**8.54**	**8.14**	**8.14**	**8.26**
Yuen's Trimmed Means						
E	4.60	**6.46**	5.80	5.44	5.68	5.12
F	**13.76**	**6.80**	**7.00**	**6.48**	**6.50**	5.76
	χ^2_3					
Least Squares Estimation						
E	5.06	**7.06**	**8.70**	**8.40**	**8.42**	**7.80**
F	**13.34**	**7.10**	**11.50**	**10.88**	**10.42**	**10.96**
Yuen's Trimmed Means						
E	4.88	**6.68**	**6.76**	**6.34**	**6.46**	5.20
F	**14.30**	**6.86**	**8.34**	**7.54**	**7.54**	**6.62**