Comput Stat (2013) 28:1269–1297
DOI 10.1007/s00180-012-0353-x

ORIGINAL PAPER

Testing homogeneity of variances with unequal sample sizes

I. Parra-Frutos

Received: 28 February 2011 / Accepted: 14 July 2012 / Published online: 17 August 2012

Abstract When sample sizes are unequal, problems of heteroscedasticity of the variables given by the absolute deviation from the median arise. This paper studies how the best known heteroscedastic alternatives to the ANOVA *F* test perform when they are applied to these variables. This procedure leads to testing homoscedasticity in a similar manner to Levene's (1960) test. The difference is that the ANOVA method used by Levene's test is non-robust against unequal variances of the parent populations and Levene's variables may be heteroscedastic. The adjustment proposed by O'Neil and Mathews (Aust Nz J Stat 42:81–100, 2000) is approximated by the Keyes and Levy (J Educ Behav Stat 22:227–236, 1997) adjustment and used to ensure the correct null hypothesis of homoscedasticity. Structural zeros, as defined by Hines and O'Hara Hines (Biometrics 56:451–454, 2000), are eliminated. To reduce the error introduced by the approximate distribution of test statistics, estimated critical values are used. Simulation results show that after applying the Keyes–Levy adjustment, including estimated critical values and removing structural zeros the heteroscedastic tests perform better than Levene's test. In particular, Brown–Forsythe's test controls the Type I error rate in all situations considered, although it is slightly less powerful than Welch's, James's, and Alexander and Govern's tests, which perform well, except in highly asymmetric distributions where they are moderately liberal.

Keywords Homoscedasticity tests · Levene's test · Bartlett's test · Welch's test · Brown and Forsythe's test · James's second-order test · Alexander and Govern's test · Monte Carlo simulation · Small samples · Estimated critical values · Structural zeros

I. Parra-Frutos (✉)
Department of Quantitative Methods for Economics and Business, Economics and Business School, University of Murcia, Murcia, Spain
e-mail: ipf@um.es

1 Introduction

There is considerable statistical literature on testing homogeneity of variances that examines the various tests that have been proposed. A comprehensive study on tests of homogeneity of variances is given by Conover et al. (1981). A large number of tests have been examined and simulated in order to determine their robustness at nominal significance levels. The tests that have received the most attention are the F test (two samples), Bartlett's (1937) test, and Levene's (1960) test. It is widely known that the F test (two samples) is extremely sensitive to the normality assumption (Siegel and Tukey 1960; Markowski and Markowski 1990). Bartlett's test is extremely non-robust against non-normality (Conover et al. 1981; Lim and Loh 1996). Layard (1973) proposed a kurtosis adjustment for Bartlett's test that has been used by Conover et al. (1981) and Lim and Loh (1996). They find some improvement when using the modified Bartlett's test, although it is still not robust. Our simulation study includes Bartlett's, the modified Bartlett's and Levene's tests.

According to Boos and Brownie (2004), the procedures to test equal variances that aim to achieve robustness against non-normality follow three types of strategies: (1) using some type of adjustment based on an estimate of kurtosis (e.g. Layard 1973); (2) performing an ANOVA on absolute deviations from the mean, median or the trimmed mean (e.g. Levene 1960; Brown and Forsythe 1947a); (3) using resampling methods to obtain p values for a given statistic (e.g. Boos and Brownie 1989; Lim and Loh 1996). This paper focuses on the second strategy. In particular, we explore the performance of heteroscedastic alternatives to ANOVA, which is a test for comparing means of several populations. However, when the variables are the absolute deviations from the sample mean, the result is a test of homoscedasticity of the parent populations, and the procedure is known as Levene's test.

Levene's test continues to attract the attention of researchers. Recent studies use different approaches to improve it. Keselman et al. (2008) investigate other robust measures of location instead of the mean to calculate the absolute deviations. They recommend a Levene-type transformation based upon empirically determined 20 % asymmetric trimmed means. Neuhäuser (2007) studies the use of nonparametric alternatives to ANOVA on Levene's variables and finds that in some cases they are more powerful. Lim and Loh (1996), Wludyka and Sa (2004), Charway and Bailer (2007), Parra-Frutos (2009) and Cahoy (2010) focus on resampling methods and show that they may improve the Type I and Type II error rates. Iachine et al. (2010) propose an extension of Levene's method to dependent observations, consisting of replacing the ANOVA step with a regression analysis followed by a Wald-type test based on a clustered version of the robust Huber-White sandwich estimator of the covariance matrix. The problem of testing equality of variances against ordered alternatives, that is, detecting trends in variances, has been addressed by various authors, including Neuhäuser and Hothorn (2000), Hui et al. (2008) and Noguchi and Gel (2010).

Let $Y_{ij}, i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, denote the jth observation from the ith group. Levene's test is defined as the one-way analysis of variance (ANOVA) on the absolute deviation from the sample mean, $M_{ij} = \left|Y_{ij} - \overline{Y}_i\right|$, where $\overline{Y}_i$ is the sample mean of the ith group. Modifications given by Brown and Forsythe (1947a) show that calculating absolute deviations from the trimmed mean and from the median instead

of from the sample mean may improve the performance of the test in certain situations. The use of a robust estimator of location, like the median, instead of the sample mean to compute the absolute deviation, $Z_{ij} = \left|Y_{ij} - \tilde{Y}_i\right|$ where $\tilde{Y}_i$ is the ith group median, has been shown to be an effective modification (Conover et al. 1981; Carroll and Schneider 1985), and it is widely used in applied research.

Under the classical assumptions (normality, homoscedasticity and independence), the ANOVA F test is known to be an optimal test. However, when one or more of these basic assumptions is violated, it becomes overly conservative or liberal. The properties of the ANOVA F test under assumption violations and under various degrees of each violation have been extensively discussed in the literature (e.g., Scheffé 1959; Glass et al. 1972; Rogan and Keselman 1977; Keselman et al. 1977; Kenny and Judd 1986; Harwell et al. 1992; De Beuckelaer 1996; Akritas and Papadatos 2004; Bathke 2004).

De Beuckelaer (1996) argues that in a situation in which more than one basic assumption is violated, the ANOVA F test becomes very unreliable, especially for violations of the independence and the homoscedasticity assumptions. According to Lix et al. (1996), the only instance in which ANOVA may be a valid test under heteroscedasticity is when the degree of variance heterogeneity is small and group sizes are equal. So, it seems that a more appropriate procedure to test homoscedasticity may be to apply a heteroscedastic alternative to ANOVA on Z_{ij} and M_{ij}. These variables do not satisfy any of the standard assumptions of the ANOVA F test. They are neither independently nor normally distributed (note that the probability distribution is skewed even when Y_{ij} is symmetric) and homoscedasticity is not guaranteed. M_{ij} and Z_{ij} do not have constant variance unless the sample sizes are equal and Y_{ij} are homoscedastic (Loh 1987; Keyes and Levy 1997; O'Neill and Mathews 2000). To see this, if Y_{ij} is normally distributed with mean μ_i and variance σ_i^2 then

$$\begin{aligned}
E(M_{ij}) &= [(2/\pi)(1 - 1/n_i)\sigma_i^2]^{1/2}, \\
var(M_{ij}) &= (1 - 2/\pi)(1 - 1/n_i)\sigma_i^2, \\
E(Z_{ij}) &= \kappa_{n_i}\sigma_i, \\
var(Z_{ij}) &= \left(\frac{n_i - 2}{n_i}\sigma_i^2 + var\left(\tilde{Y}_i\right)\right) - \kappa_{n_i}^2\sigma_i^2, \\
var\left(\tilde{Y}_i\right) &\approx \frac{\pi}{2n_i}\sigma_i^2.
\end{aligned}$$

where κ_{n_i} is a constant depending only on the sample size n_i (O'Neill and Mathews 2000). Thus, the variances of M_{ij} and Z_{ij} depend on σ_i^2 and n_i. So, under the null hypothesis of equal σ_i^2, $\forall i = 1, \ldots, k$, the assumption of homoscedasticity of the M_{ij} and Z_{ij} is not guaranteed unless the sample sizes are equal.

On the other hand, applying ANOVA, or a heteroscedastic alternative, on M_{ij} to test the homogeneity of variances of Y_{ij} implies testing the hypothesis

$$H_0 : E\left(M_{1j}\right) = E\left(M_{2j}\right) = \cdots = E\left(M_{kj}\right).$$

Assuming that Y_{ij} are normally distributed with mean μ_i and variance σ_i^2, the H_0 above corresponds to the hypothesis

$$H_0 : (1 - 1/n_1)\,\sigma_1^2 = (1 - 1/n_2)\,\sigma_2^2 = \cdots = (1 - 1/n_k)\,\sigma_k^2$$

when M_{ij} is used. Similarly,

$$H_0 : \kappa_{n_1}^2 \sigma_1^2 = \kappa_{n_2}^2 \sigma_2^2 = \cdots = \kappa_{n_k}^2 \sigma_k^2$$

when Z_{ij} is used.

To obtain the correct hypothesis $H_0 : \sigma_1^2 = \cdots = \sigma_k^2$ an adjustment must be introduced. When using M_{ij}, Keyes and Levy (1997) suggest multiplying by $1/\sqrt{1 - 1/n_i}$. Let us denote $U_{ij} = M_{ij}/\sqrt{1 - 1/n_i}$. Using U_{ij}, we can test the homogeneity of variances with the desired hypothesis $H_0 : \sigma_1^2 = \cdots = \sigma_k^2$. On the other hand, $var\left(U_{ij}\right) = (1 - 2/\pi)\,\sigma_i^2$, and the effect of unequal sample size vanishes. That is, in normal populations under the null hypothesis the variables U_{ij} are homoscedastic.

For Z_{ij}, however, O'Neill and Mathews (2000) suggest multiplying by $1/\kappa_{n_i}$. The variance of Z_{ij}/κ_{n_i} is

$$\frac{1}{\kappa_{n_i}^2}\left(\frac{n_i - 2}{n_i}\sigma_i^2 + var\left(\tilde{Y}_i\right)\right) - \sigma_i^2$$

which is a function of n_i. Therefore, for unequal sample sizes we have heterogeneous variances of the variables Z_{ij}/κ_{n_i} and the correct null hypothesis. Since mean and median coincide for normal distribution, κ_{n_i} should be sufficiently close to $\sqrt{(2/\pi)(1 - 1/n_i)}$ (the Keyes–Levy adjustment for M_{ij}) even for moderate sample sizes. So, κ_{n_i} may be approximated by $\hat{\kappa}_{n_i} = \sqrt{(2/\pi)(1 - 1/n_i)}$. For example, when $(n_1, n_2, n_3, n_4) = (4, 10, 18, 22)$ the hypothesis is approximately

$$\begin{aligned} H_0 : (2/\pi)(1 - 1/4)\sigma_1^2 &= (2/\pi)(1 - 1/10)\sigma_2^2 \\ &= (2/\pi)(1 - 1/18)\sigma_3^2 = (2/\pi)(1 - 1/22)\sigma_4^2 \end{aligned}$$

that is,

$$H_0 : \sigma_1^2 = 1.20\sigma_2^2 = 1.26\sigma_3^2 = 1.27\sigma_4^2$$

and variances of $Z_{ij},\ i = 1, \ldots, 4$, would be approximately,

$$\begin{aligned} var\left(Z_{1j}\right) &\approx 1.11\, var\left(Z_{4j}\right) \\ var\left(Z_{2j}\right) &\approx 1.03\, var\left(Z_{4j}\right) \\ var\left(Z_{3j}\right) &\approx 1.01\, var\left(Z_{4j}\right) \end{aligned}$$

Using the Keyes–Levy adjustment, consisting of dividing by $\hat{\kappa}_{n_i}$, the hypothesis becomes $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ and the variances of $Z_{ij}/\hat{\kappa}_{n_i},\quad i = 1, \ldots, 4$, would be approximately

$$var\left(Z_{1j}/\hat{\kappa}_{n_1}\right) \approx 1.42\, var\left(Z_{4j}/\hat{\kappa}_{n_4}\right)$$
$$var\left(Z_{2j}/\hat{\kappa}_{n_2}\right) \approx 1.09\, var\left(Z_{4j}/\hat{\kappa}_{n_4}\right)$$
$$var\left(Z_{3j}/\hat{\kappa}_{n_3}\right) \approx 1.02\, var\left(Z_{4j}/\hat{\kappa}_{n_4}\right)$$

A maximum of $1.45 var\left(Z_{4j}/\hat{\kappa}_{n_4}\right)$ would be reached for $n_i = 4$ when the largest sample size is 30 observations ($n_4 = 30$). Thus, larger variances are associated with smaller sample sizes. These problems of heteroscedasticity motivate us to explore the performance of heteroscedastic alternatives to the ANOVA step in Levene's test. We focus on Z_{ij}, and hence on $Z_{ij}/\hat{\kappa}_{n_i}$, since several studies, including Conover et al. (1981), Carroll and Schneider (1985) and Lim and Loh (1996), confirm that absolute deviations from medians, rather than means, are preferable.

According to O'Neill and Mathews (2000), the various forms of Levene's test currently applied ignore the distributional properties of M_{ij} and Z_{ij} so the approximate distribution of the test statistic used is inadequate, which leads to a poor performance. An improvement may be found when using estimated critical values (Loh 1987) or using weighted least squares analysis of variance (O'Neill and Mathews 2000). Loh (1987) affirms that Levene's test can be made exact by computer simulation of the critical point of the statistic assuming normality and constant variance, but not otherwise. O'Neill and Mathews (2000) show that, in normal populations, empirical levels of significance are close to nominal values for several of the statistics they propose based on weighted least squares instead of on ordinary least squares. We include the former approach along with the heteroscedastic alternatives to ANOVA to deal with heteroscedasticity and its adverse consequences.

Lix et al. (1996) found that the parametric alternatives to the ANOVA F test were superior when the variance homogeneity assumption was violated. The heteroscedastic alternatives to the ANOVA F test that receive most attention are Welch's (1951) test, James's (1951) second-order method, Brown and Forsythe's (1947a) test, and Alexander and Govern's (1994) test. We study how these tests behave when they are applied to absolute deviations from the median.

All these procedures have been investigated in empirical studies. The evidence suggests that these methods can generally control the rate of Type I error when group variances are heterogeneous and the data are normally distributed (Dijkstra and Werter 1981; Wilcox 1990; Oshima and Algina 1992; Alexander and Govern 1994). However, the literature also indicates that these tests can become liberal when the data are both heterogeneous and non-normal, particularly when the design is unbalanced.

One of the best known parametric alternatives to the ANOVA is that given by Welch (1951). It has been widely used and is included in statistical packages. However, various simulation studies (Dijkstra and Werter 1981; Wilcox 1988, 1989; Alexander and Govern 1994; Hsiung et al. 1994; Oshima and Algina 1992) show that James's (1951) second-order test generally appears to be the most accurate method over a wide range of realistic conditions. One major drawback is its computational complexity. James (1951) proposed two methods for adjusting the critical value—first and second order-methods. However, James's first-order procedure does not control the rate of the Type I errors under variance heterogeneity for small sample sizes (Brown and Forsythe 1974b). Welch's (1951) and James's (1951) tests can be used whenever the

variance homogeneity assumption is not satisfied, but should be avoided if the data are moderately to highly skewed, even in balanced designs (Clinch and Keselman 1982; Wilcox et al. 1986; Lix et al. 1996).

One competitor of James's (1951) second-order and Welch's (1951) tests would seem to be Alexander–Govern's (1994) procedure (Lix et al. 1996), since it is reported to possess many characteristics which are similar to those of the James method. A second is the modification to the Brown–Forsythe (1974b) test suggested by Rubin (1983), and later by Mehrotra (1997).

A comparison of Alexander–Govern's, ANOVA, Kruskal–Wallis's, Welch's, Brown–Forsythe's, and James's second-order tests concluded that, under variance heterogeneity, Alexander–Govern's approximation was comparable to Welch's test and James's second-order test and, in certain instances, was superior (Schneider and Penfield 1997). The same study also finds that the Alexander–Govern test is liberal when distribution is extremely skew and conservative when it is platykurtic. Wilcox (1997) also reported similar findings. Schneider and Penfield recommend the Alexander–Govern procedure as the best alternative to the ANOVA F test when variances are heterogeneous, for three reasons: (1) it is computationally simpler; (2) its overall superiority under most experimental conditions; (3) the questionable results of Welch's test when more than four treatment groups are investigated (Dijkstra and Werter 1981; Wilcox 1988).

We use Bradley's (1978) liberal criterion of robustness to nominal significant level, which establishes that a test is considered robust if its empirical Type I error rate falls within the interval [0.025,0.075] for a nominal level $\alpha = 0.05$. Not all authors agree with this criterion. Cochran (1954) established the interval [0.04,0.06]. Conover et al. (1981) classify a test as robust if the maximum empirical Type I error rate is less than 0.10 for a 5 % test.

To test homoscedasticity we study how heteroscedastic alternatives of ANOVA F test perform when they are applied on the median-based Levene variables, that is, the absolute deviations from the median. We focus on tests given by Welch (1951), James (1951), Brown and Forsythe (1974b), Rubin (1983)—a correction to the Brown–Forsythe test also addressed by Mehrotra (1997), and Alexander and Govern (1994). In particular, we are interested in power levels (the ability to reject H_0 when it is false) and robustness of validity (whether the procedures have approximately the nominal significance level) under a variety of different settings: small, large, equal and unequal sample sizes; and for symmetric, asymmetric and heavy-tailed distributions. We compare these results with those obtained for the median-based Levene test and Bartlett's test (with and without kurtosis adjustment). The null hypothesis of Levene's test is non homoscedasticity of Y_i since the expected mean of Z_{ij} is not the variance of Y_i (Keyes and Levy 1997; O'Neill and Mathews 2000); thus an adjustment is needed, as suggested by O'Neil and Mathews. This adjustment can be well approximated by the Keyes–Levy adjustment for M_{ij}.

We include two refinements of tests, which are known to improve them. Tests applied on the median-based absolute deviations are too conservative for small, equal, and odd sample sizes. This is a consequence of using the median as the location measure. A remedy based on removing structural zeros was suggested by Hines and O'Hara Hines (2000). This method improves results in terms of the Type I error rate

and power. However, when the structural zero removal method is applied jointly with the Keyes–Levy adjustment a new procedure must be used consisting of a modification of the structural zero removal method, as shown in Noguchi and Gel (2010), in order to preserve the null hypothesis. The second refinement consists of using estimated critical values instead of the approximate distribution of the test statistics, as suggested by Loh (1987). When test statistics have an approximate distribution an additional error is introduced when taking a decision about the null hypothesis. The size of the error depends on the goodness of the approximation. In order to eliminate or, at least, reduce this error, empirical percentiles of the test statistic based on the standard normal are used as critical values in the rejection rule of the tests.

From a simulation study we obtain that, in the normal case, none of the tests studied improve Bartlett's test. However, it is well-known (and our results confirm this) that it is not a robust test, which implies that it must be used with caution. A much better control is observed with the kurtosis adjustment. Our simulation results also show a general good behaviour of heteroscedastic tests when applying the Noguchi–Gel procedure and using estimated critical values. In particular, the heteroscedastic Brown–Forsythe and the Levene tests control the Type I error rate for all the parent populations considered even when the samples are small and unequal. The first may be considered even more robust at the significant level than the second. The Levene test is a homoscedastic test that has been applied under mild heteroscedasticity so a little less control is observed. All the tests considered in this study perform similarly in large samples. When outliers are present some tests control the Type I error rate but the power achieved is very low.

2 Description of tests

Bartlett's (1937) *test* (B test). Its statistic is given by

$$B = \frac{M}{1 + C}, \tag{1}$$

where

$$M = (N - k) \ln S_a^2 - \sum_{i=1}^{k} (n_i - 1) \ln S_i^2$$

$$S_i^2 = \frac{\sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_i\right)^2}{n_i - 1}$$

$$S_a^2 = \frac{\sum_{i=1}^{k} (n_i - 1)\, S_i^2}{N - k}$$

$$C = \frac{1}{3\,(k-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{N - k} \right)$$

The Bartlett statistic is approximately distributed as a Chi-square variable with $k - 1$ degrees of freedom.

Bartlett's test with kurtosis adjustment (B2 test).

$$B2 = kB, \tag{2}$$

where

$$k = \frac{2}{\hat{\beta}_2 - 1}$$
$$\hat{\beta}_2 = \frac{N \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_i\right)^4}{\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_i\right)^2\right)^2}.$$

The $B2$ is approximately distributed as a Chi-square variable with $k-1$ degrees of freedom.

Levene's (1960) *test* (L50 test). Recall that $Z_{ij} = \left|Y_{ij} - \tilde{Y}_i\right|$, where $\tilde{Y}_i$ is the ith group median.

$$L = \frac{(N-k) \sum_{i=1}^{k} \left(\overline{Z}_i - \overline{Z}\right)^2 n_i}{(k-1) \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Z_{ij} - \overline{Z}_i\right)^2}, \tag{3}$$

where

$$N = \sum_{i=1}^{k} n_i,$$
$$\overline{Z}_i = \sum_{j=1}^{n_i} Z_{ij}/n_i,$$
$$\overline{Z} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Z_{ij}/N.$$

The L is approximately distributed as an F variable with $k-1$ and $N-k$ degrees of freedom.

2.1 Heteroscedastic tests

Welch's (1951) *test* (W test).

$$Q = \frac{\sum_{i=1}^{k} w_i \left(\overline{Z}_i - \overline{Z}^*\right)^2 / (k-1)}{1 + \frac{2(k-2)}{k^2-1} \sum_{i=1}^{k} \frac{(1-w_i/W)^2}{n_i - 1}}, \tag{4}$$

where

$$w_i = \frac{n_i}{S_{Z,i}^2},$$
$$W = \sum_{i=1}^{k} w_i,$$
$$S_{Z,i}^2 = \frac{\sum_{j=1}^{n_i} \left(Z_{ij} - \overline{Z}_i\right)^2}{n_i - 1},$$
$$\overline{Z}^* = \frac{\sum_{i=1}^{k} w_i \overline{Z}_i}{W}.$$

The Welch statistic is approximately distributed as an F variable with $k-1$ and v degrees of freedom where

$$v = \frac{k^2 - 1}{3 \sum_{i=1}^{k} \frac{(1 - w_i/W)^2}{n_i - 1}}.$$

James's (1951) *test* (J1 and J2 tests).

$$U = \sum_{i=1}^{k} w_i \left(\overline{Z}_i - \overline{Z}^*\right)^2. \tag{5}$$

A simple Chi-square approximation with $k-1$ degrees of freedom for U is known to be unsatisfactory when the sample sizes are small or even moderately large. Accordingly, James (1951) proposed two methods for adjusting the critical value. His second order-method is widely recommended and is as follows (J2 test). The null hypothesis is rejected if $U > h2(\alpha)$, where

$$\begin{aligned}
h2(\alpha) = {} & c(\alpha) + \frac{1}{2}(3\chi_4 + \chi_2) \sum_{i=1}^{k} \frac{(1 - w_i/W)^2}{n_i - 1} \\
& + \frac{1}{16}(3\chi_4 + \chi_2)^2 \left(1 - \frac{k-3}{c(\alpha)}\right) \left(\sum_{i=1}^{k} \frac{(1 - w_i/W)^2}{n_i - 1}\right)^2 \\
& + \frac{1}{2}(3\chi_4 + \chi_2) \left[\left(8R_{23} - 10R_{22} + 4R_{21} - 6R_{12}^2 + 8R_{12}R_{11} - 4R_{11}^2\right)\right. \\
& + \left(2R_{23} - 4R_{22} + 2R_{21} - 2R_{12}^2 + 4R_{12}R_{11} - 2R_{11}^2\right)(\chi_2 - 1) \\
& \left. + \frac{1}{4}\left(-R_{12}^2 + 4R_{12}R_{11} - 2R_{12}R_{10} - 4R_{11}^2 + 4R_{11}R_{10} - R_{10}^2\right)(3\chi_4 - 2\chi_2 - 1)\right] \\
& + (R_{23} - 3R_{22} + 3R_{21} - R_{20})(5\chi_6 + 2\chi_4 + \chi_2) \\
& + \frac{3}{16}\left(R_{12}^2 - 4R_{23} + 6R_{22} - 4R_{21} + R_{20}\right)(35\chi_8 - 15\chi_6 + 9\chi_4 + 5\chi_2) \\
& + \frac{1}{16}\left(-2R_{22} + 4R_{21} - R_{20} + 2R_{12}R_{10} - 4R_{11}R_{10} + R_{10}^2\right)(9\chi_8 - 3\chi_6 - 5\chi_4 - \chi_2) \\
& + \frac{1}{4}\left(-R_{22} + R_{11}^2\right)(27\chi_8 + 3\chi_6 + \chi_4 + \chi_2) \\
& + \frac{1}{4}(R_{23} - R_{12}R_{11})(45\chi_8 + 9\chi_6 + 7\chi_4 + 3\chi_2),
\end{aligned}$$

with $c(\alpha)$ denoting the $1-\alpha$ quantile of a χ^2_{k-1} distribution and with

$$\chi_{2s} = \frac{c\,(\alpha)^s}{(k-1)\,(k+1)\,(k+3)\cdots(k-2s-3)}, \qquad (6)$$
$$R_{st} = \sum_{i=1}^{k} \frac{(w_i/W)^t}{(n_i-1)^s}.$$

The first order-approximation to the critical value for U, $h1(\alpha)$, is given by (J1 test)

$$h1\,(\alpha) = c\,(\alpha) + \frac{1}{2}\,(3\chi_4 + \chi_2) \sum_{i=1}^{k} \frac{(1-w_i/W)^2}{n_i-1}.$$

Using (6), it can be rewritten as

$$h1\,(\alpha) = c\,(\alpha) + \frac{c\,(\alpha)}{2\,(k-1)}\left(1 + \frac{3c\,(\alpha)}{k+1}\right) \sum_{i=1}^{k} \frac{(1-w_i/W)^2}{n_i-1}.$$

This approximation is used in Alexander and Govern (1994) and is also derived by Johansen (1980).

Brown and Forsythe (1974b) (BF test).

$$F^* = \frac{\sum_{i=1}^{k} n_i \left(\overline{Z}_i - \overline{Z}\right)^2}{\sum_{i=1}^{k} (1-n_i/N)\, S^2_{Z,i}} \qquad (7)$$

The F^* is approximately distributed as an F variable with v_1 and v_2 degrees of freedom where

$$v_1 = k-1,$$
$$v_2 = \left(\sum_{i=1}^{k} \frac{f_i^2}{n_i-1}\right)^{-1}$$

and

$$f_i = \frac{(1-n_i/N)\, S^2_{Z,i}}{\sum_{i=1}^{k} (1-n_i/N)\, S^2_{Z,i}}$$

Brown, Forsythe and Rubin test (BFR test). Rubin (1983), and later Mehrotra (1997), show that the approximation given for F^* by Brown and Forsythe was inadequate and often leads to inflated Type I error rates. They found an improved approximation using Box's (1954) method, which involves modifying numerator degrees of freedom of F^*, as given here

$$v_1 = \frac{\left(\sum_{i=1}^{k} (1 - n_i/N)\, S_{Z,i}^2\right)^2}{\left(\sum_{i=1}^{k} S_{Z,i}^2 n_i/N\right)^2 + \sum_{i=1}^{k} (1 - 2n_i/N)^2\, S_{Z,i}^4}.$$

Alexander and Govern's (1994) *procedure* (AG test).

$$A = \sum_{i=1}^{k} g_i^2 \tag{8}$$

where

$$g_i = c_i + \frac{c_i^3 + 3c_i}{b_i} - \frac{4c_i^7 + 33c_i^5 + 240c_i^3 + 855c_i}{10b_i^2 + 8b_i c_i^4 + 1000b_i}$$

with

$$\begin{aligned} a_i &= n_i - 1.5, \\ b_i &= 48a_i^2, \\ t_i &= \frac{\left(\overline{Z}_i - \overline{Z}^*\right)\sqrt{n_i}}{S_{Z,i}}, \\ c_i &= \left[a_i \ln\left(1 + \frac{t_i^2}{n_i - 1}\right)\right]^{1/2}. \end{aligned}$$

The A is approximately distributed as a Chi-square variable with $k - 1$ degrees of freedom.

2.2 Estimated critical values

All the test statistics described have an approximate distribution which introduces an additional error when taking a decision on the null hypothesis. In order to eliminate or, at least, reduce this error, empirical percentiles of the test statistic are used as critical values in the rejection rule of a test. If the approximate distribution of the test statistic is not good enough, then an improvement is achieved by using estimated critical values. Otherwise, similar results would be obtained.

According to Loh (1987), empirical percentiles are obtained as follows. Given n_i, $i = 1, \ldots, k$, samples are generated from a standard normal population, absolute deviations from the group medians are computed and the test statistic calculated. This process is repeated $100M$ times (where M is an integer) and the $100M$ test statistic values are ordered from smallest to largest as $B(1), B(2), \ldots, B(100M)$, using the notation of Bartlett's statistic. The 5 % empirical critical value, then, is obtained as $C = \frac{1}{2}[B(95M) + B(95M + 1)]$. If the observed test statistic is higher than C the null hypothesis is rejected. We use $M = 100$, that is, 10,000 iterations.

Tests based on estimated critical values are denoted by adding an *e* at the end of the name. For example, *Be* test for Bartlett's test using estimated critical values. Empirical percentiles have been used by Loh (1987). We use the standard normal for generating the empirical percentiles in all cases, which obviously may introduce an error, depending on the sensitivity of the test statistic to non-normal distributions. However, when the underlying distribution of the data is known and a test is not robust for it, the test can be carried out accurately by using estimated critical values generated using the known parent distribution.

2.3 Structural zeros

Levene's test is extremely conservative for odd and small samples sizes. Hines and O'Hara Hines (2000) found that this is due to the presence of structural zeros, which should be removed before applying Levene's test in order to improve the performance.

When the sample size is odd, there will always be one $r_{ij} = Y_{ij} - \tilde{Y}_i$ that is zero since the median is one of the actual data values. According to Hines and O'Hara Hines, this particular r_{ij} is uninformative and labeled a structural zero.

When the sample size is even, $\tilde{Y}_i - Y_{i(m_i)} = Y_{i(m_i+1)} - \tilde{Y}_i$. Here, $Y_{i(k)}$ represents the kth order statistic for the ith set of data, and $m_i = \left[\frac{1}{2}n_i\right]$. Hines and O'Hara Hines consider then the following orthogonal rotation of the ordered vector $(r_{i,(1)}, \ldots, r_{i,(n_i)})$: Replace the pair of values $r_{i,(m_i)}$ and $r_{i,(m_i+1)}$ by the pair $(r_{i,(m_i+1)} - r_{i,(m_i)})/\sqrt{2}$ and $(r_{i,(m_i+1)} + r_{i,(m_i)})/\sqrt{2} \quad (= 0)$. Then delete the rotated deviation from the median in ordered location $(m_i + 1)$ since, after the indicated replacement it is a structural zero. We use these modifications in all tests applied on Z_{ij}, and rename them by adding -0. For example, L50 $-$ 0e denotes the median-based Levene test removing structural zeros and estimating critical values.

The Keyes–Levy adjustment leads to the right null hypothesis of homogeneity of variances. However, if the structural zero removal method is used after that, we are no longer testing that hypothesis. In this case the procedure to follow should be that described by Noguchi and Gel (2010). This procedure eliminates the structural zeros without altering the null hypothesis of homoscedasticity. Basically it consists of multiplying data by $\sqrt{1 - 1/n_i}$ and then applying a modified structural zero removal for even sample sizes and the original Hines-Hines method for odd sample sizes. For an even sample size $Z_{i(1)}$ and $Z_{i(2)}$ are transformed into $Z_{i(1)} - Z_{i(2)}(= 0)$ and $Z_{i(1)} + Z_{i(2)}(= 2Z_{i(1)})$, respectively, where $Z_{i(m)}$ denotes the mth order statistic of Z_{ij}, and the newly created structural zero $(Z_{i(1)} - Z_{i(2)}(= 0))$ is removed.

When the Noguchi–Gel method is used, tests are renamed by adding (*NG*). For example, BF(NG)e denotes the Brown–Forsythe test using the Noguchi–Gel procedure and estimating critical values.

3 Design of the simulation

The Type I error rate and the power of the tests are compared in a simulation based on six distributions and nine configurations of group sizes (n_1, n_2, n_3, n_4). The distributions

are: normal (symmetric); Student's t with 4 degrees of freedom (symmetric, long-tailed and low peakedness); mixed normal or contaminated normal (symmetric and heavy-tailed); uniform (symmetric and very low kurtosis); Chi-square with 4 degrees of freedom (skewed, long-tailed and high kurtosis); and exponential with mean 1/3 (skewed, heavy-tailed and high kurtosis). The nominal 5 % significance level is used throughout. The simulation results are based on 10,000 replications. The S language is used.

The mixed (or contaminated) normal may be described as $(1 - p)N(0, 1) + pN(0, \sigma_C)$, where $0 \leq p \leq 1$. This distribution is symmetric and quite similar to the normal distribution when p is close to 0. The distribution differs from the normal in that we see outliers more often than would be expected for a normal distribution. In our simulation study $p = 0.05$ and $\sigma = 3$.

For a general view of the behaviour of tests, five configurations of small samples and four of large samples are considered. Three of the configurations of small samples are of equal size: (5,5,5,5), (6,6,6,6), and (16,16,16,16). Two of them are of unbalanced design: (6,7,8,9) and (4,10,18,22). Four configurations of large samples are also studied: (30,30,30,30), (60,60,60,60), (35,40,45,52), and (30,65,90,150). When we focus on the behaviour of tests in small and unequal samples, fourteen configurations are considered:

(4, 5, 6, 7)	(4, 28, 28, 28)	(10, 14, 18, 20)
(6, 7, 8, 9)	(4, 4, 28, 28)	(10, 14, 18, 30)
(6, 9, 20, 30)	(4, 4, 4, 28)	(20, 22, 24, 26)
(10, 11, 12, 13)	(8, 12, 18, 20)	(15, 20, 25, 28)
(4, 10, 18, 22)	(8, 12, 18, 30)	

A null hypothesis of equal variances is studied along with three alternatives: $\left(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2\right) = (1, 6, 11, 16)$, (16, 11, 6, 1), and (1,1,1,16).

In order to obtain samples from populations with the desired variances σ_i^2, $i = 1, 2, 3, 4$, maintaining equal population means, the samples from Student's t (t_4) distributions are transformed multiplying by $\sigma_i/\sqrt{2}$. For data from the Chi-square distribution the transformation is $(\sigma_i/\sqrt{8})(X_i - 4)$, where X_i is a Chi-square variable with 4 degrees of freedom. Data from an exponential distribution are transformed by $3\sigma_i(G_i - 1/3)$, where G_i is an exponential variable with mean 1/3. For the uniform distribution, appropriate parameters are used with the same objective. Data were generated from uniform with minimum and maximum values given by $-\sqrt{3}\sigma_i$ and $\sqrt{3}\sigma_i$, respectively. Finally, in the case of the contaminated normal, the value of $\sigma_{C,i}^2$ to have a population variance σ_i^2 is given by $\sigma_{C,i}^2 = (\sigma_i^2 - 0.95)/0.05$.

4 Simulation results

A collection of figures is given to illustrate simulation results of the Type I error rate and estimated power. Some simulation results are also given in the tables in the "Appendix".

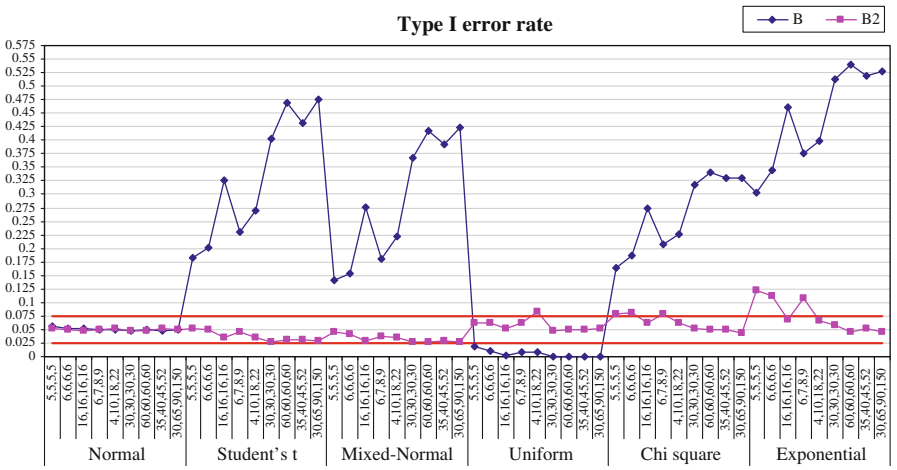

Fig. 1 Type I error rate of Bartlett's test (B) and Bartlett's test with kurtosis adjustment (B2)

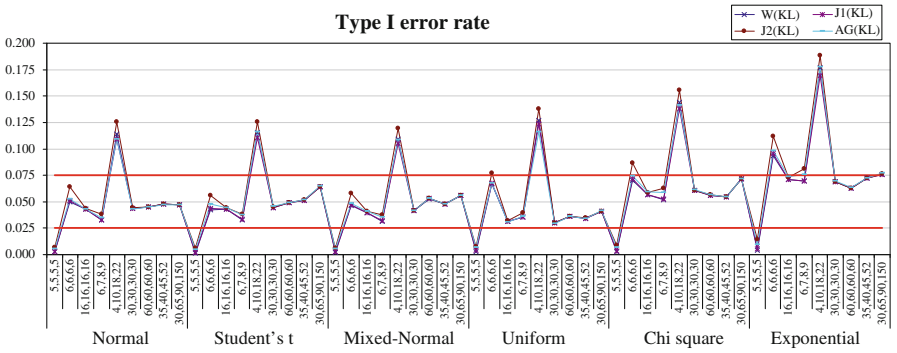

Fig. 2 Type I error rates of tests of group A using the Keyes–Levy adjustment

Power levels are very high for all tests if samples sizes are large, except when there are outliers (mixed normal distribution). In this case none of the tests have an acceptable power level under any condition. In contrast, Type I error rates may be controlled by some of the tests.

With respect to the Type I error rate, the simulation results show that B and B2 tests have their own behavior (see Fig. 1) and the remaining tests may be classified into two groups with similar performance. Group A would be included by the W(KL), J1(KL), J2(KL) and AG(KL) tests (see Figs. 2, 3, 4, 5), and group B by the L50(KL), BF(KL) and BFR(KL) tests (see Figs. 6, 7, 8, 9).

B test is extremely sensitive to non-normality, as reported in the literature. Having large samples does not lead to control of the Type I error rate under any distribution. However, B2 test is always robust at the significance level and powerful in large sample sizes, except when there are outliers. In small samples, the empirical Type I error rate of B2 test verifies Bradley's liberal criterion if distributions are symmetric and bell-shaped. Liberality problems are found for the uniform distribution and asymmetric distributions.

Fig. 3 Type I error rates of tests of group A using the Noguchi–Gel procedure

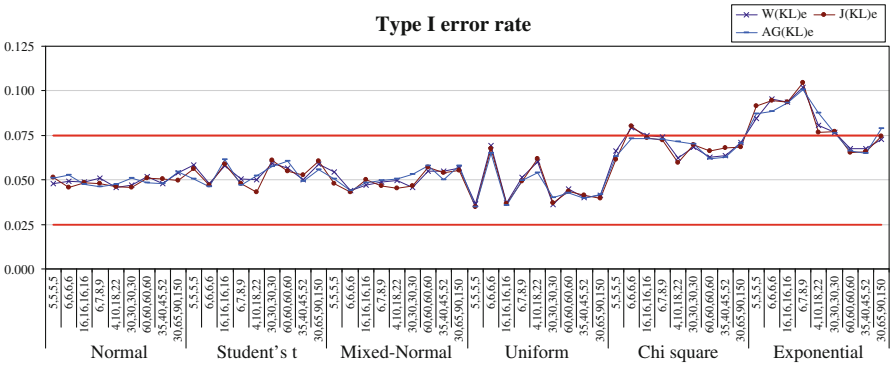

Fig. 4 Type I error rates of tests of group A using the Keyes–Levy adjustment and estimated critical values

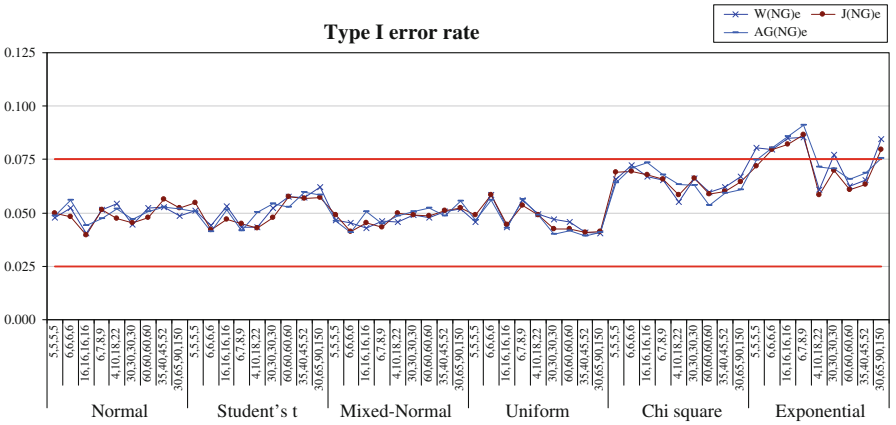

Fig. 5 Type I error rates of tests of group A using the Noguchi–Gel procedure and estimated critical values

Tests in group A: W(KL), J1(KL), J2(KL) and AG(KL). In large sample sizes all of them seem to control the Type I error rate. Serious problems are found in small sample sizes. In this case if they are equal and odd then tests are too conservative

Fig. 6 Type I error rates of tests of group B using the Keyes–Levy adjustment

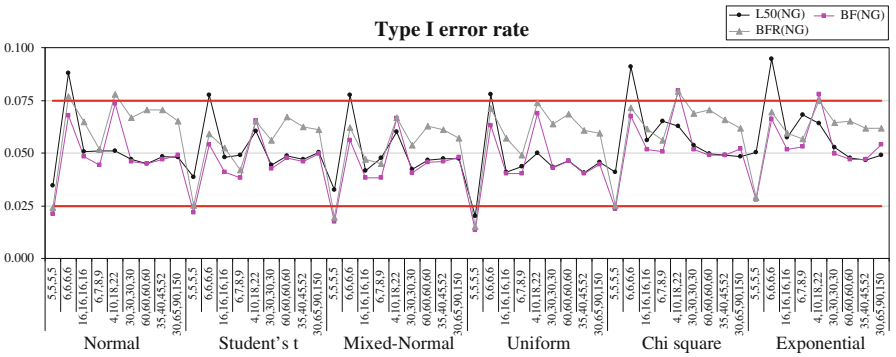

Fig. 7 Type I error rates of tests of group B using the Noguchi–Gel procedure

Fig. 8 Type I error rates of tests of group B using the Keyes–Levy adjustment and estimated critical values

(Fig. 2). This problem disappears (Fig. 3) when structural zeros are removed, but tests still present problems in controlling the Type I error rate. The Type I error rate is under control if decisions are based on estimated critical values and distributions are symmetric (Fig. 4). Problems of liberality seem to get worse as the degree of asymmetry increases. Removing structural zeros (Noguchi–Gel procedure) and using estimated critical values simultaneously lead to a slight improvement (Fig. 5).

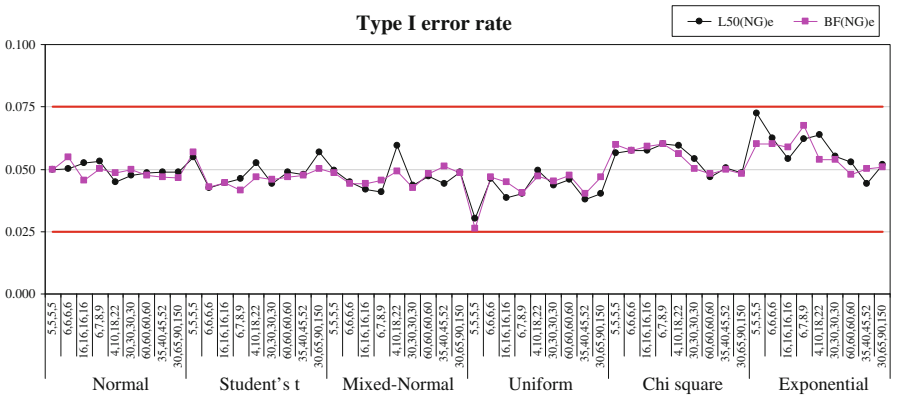

Fig. 9 Type I error rates of tests of group B using the Noguchi–Gel procedure and estimated critical values

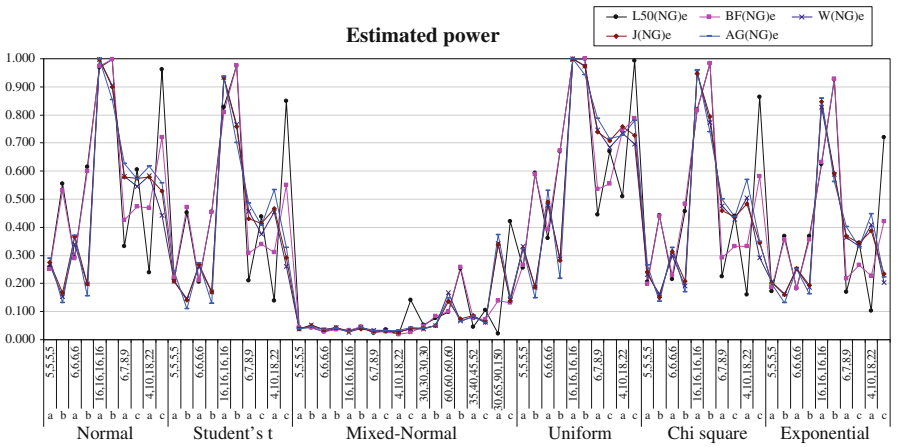

Fig. 10 Estimated power using the Noguchi–Gel procedure and estimated critical values, Note: $a = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 6, 11, 16)$; $b = (1, 1, 1, 16)$; $c = (16, 11, 6, 1)$

Tests in group B: L50(KL), BF(KL) and BFR(KL). They also perform well in large sample sizes. However, in small samples they are too conservative in two cases (Fig. 6): (1) when sample sizes are small, equal and odd [for example, $(n_1, n_2, n_3, n_4) = (5, 5, 5, 5)$]; (2) when they are unequal and very small [e.g. $(n_1, n_2, n_3, n_4) = (6, 7, 8, 9)$]. The latter seems to be only applicable in symmetric distributions.

Removing structural zeros (Noguchi–Gel procedure) does not lead to controlling the Type I error rate (Fig. 7). The use of estimated critical values seems to improve performance but problems of control are observed in highly asymmetric distributions and in the sample size combination (5,5,5,5) for the uniform distribution (Fig. 8). These results are only improved when both refinements are used (Fig. 9). In this case, both the L50(NG)e and BF(NG)e tests show a very good performance. They control the Type I error rate in any of the settings considered in this simulation study (Fig. 10).

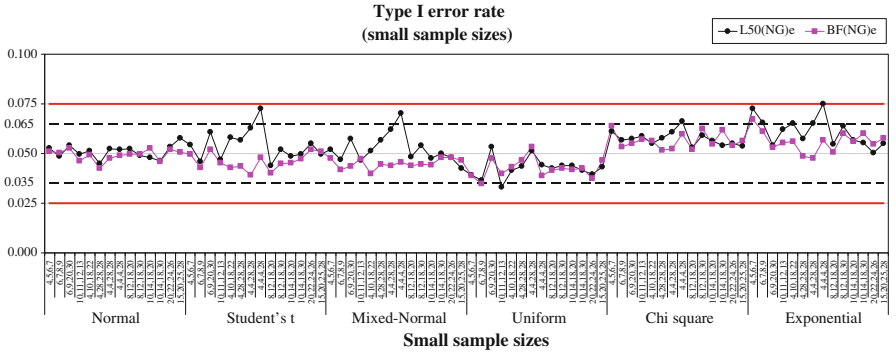

Fig. 11 Type I error rates of tests of group B in small sample sizes using the Noguchi–Gel procedure and estimated critical values

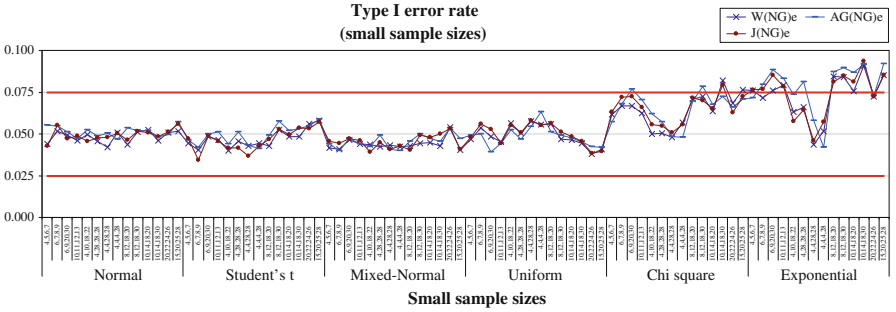

Fig. 12 Type I error rates of tests of group A in small sample sizes using the Noguchi–Gel procedure and estimated critical values

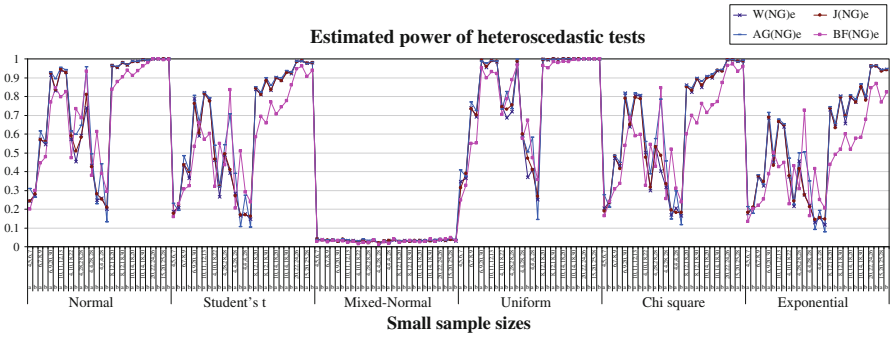

Fig. 13 Estimated power of heteroscedastic tests, Note: $a = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 6, 11, 16)$; $b = (16, 11, 6, 1)$

To gain further insights into the behaviour of tests in small samples, a new simulation has been made with a greater variety of small and unequal sample sizes (see Figs. 11, 12, 13, and Tables 1, 2 in the "Appendix"). The simulation results indicate that the BF(NG)e and L50(NG)e tests control the Type I error rate in small sample sizes (Fig. 11). However, the BF(NG)e test seems to show a better control than the L50(NG)e

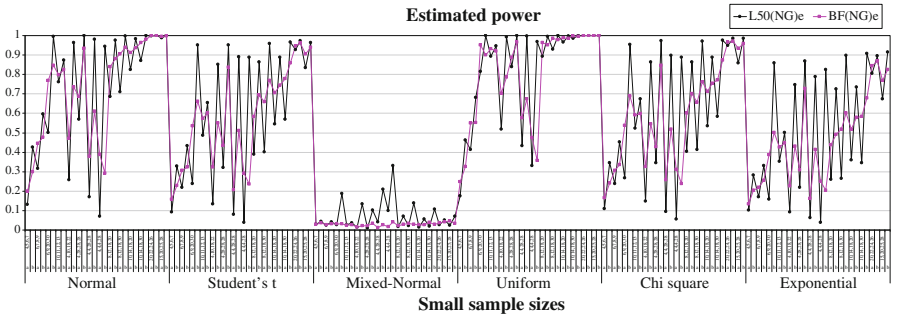

Fig. 14 Estimated power, Note: $a = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2) = (1, 6, 11, 16)$; $b = (16, 11, 6, 1)$

test. In small and unequal sample sizes problems of heteroscedasticity of Z_{ij} and $Z_{ij}/\hat{\kappa}_{n_i}$ may appear, so a heteroscedastic test is expected to perform better. However, the degree of heteroscedasticity may not be too high when σ_i^2, $i = 1, \ldots, 4$, are equal (the largest variance is 1.45 times the smaller one and is associated to the small sample size) as to give rise to serious problems of liberality when applying the homoscedastic Levene test L50(NG)e. Simulation results seem to confirm this extent (see Fig. 11). On the other hand, asymmetry and negative (or positive) correlation between sample sizes and population variances are problems for tests comparing means, the BF(NG)e test seems to be the most robust in these situations.

Power is slightly lower for the BF(NG)e test than for the remaining heteroscedastic tests (Fig. 13). This would suggest using W(NG)e, J(NG)e or AG(NG)e except when distributions are highly asymmetric (like the exponential distribution). For highly asymmetric distributions the BF(NG)e test shows a better control of the Type I error rate than the L50(NG)e test. However, the L50(NG)e test is more powerful than the BF(NG)e test when the unknown population variances σ_i^2 are negatively correlated with the sample sizes (Fig. 14).

Simulation results also show that tests have a very low power when there are outliers, although they may control the Type I error rate (Figs. 11, 13).

5 Conclusions

Problems of heteroscedasticity of the variables given by the absolute deviation from the median arise when sample sizes are unequal. To deal with the heterogeneity of these variables in Levene's test we propose substituting the ANOVA step for a heteroscedastic alternative. We focus on tests given by Welch (1951), James (1951), Brown and Forsythe (1974b), Rubin (1983) and Mehrotra (1997)—a correction to the Brown–Forsythe test, and Alexander and Govern (1994). The Keyes–Levy adjustment consisting of dividing the observations by $\hat{\kappa}_{n_i}$ is applied to get the correct null hypothesis of equal variances. We also consider removing structural zeros, according to the Noguchi–Gel procedure, and using estimated critical values.

None of the tests considered in this study show a good performance when there are outliers. In this case, the Type I error rate may be controlled by tests when using estimated critical values, but power levels are always too low.

Removing structural zeros implies a substantial improvement in the tests, but a better performance is achieved when estimated critical values are also applied. Moreover, simulation results show that for the variables $Z_{ij}/\hat{\kappa}_{n_i}$ the heteroscedastic tests perform better than the ANOVA step used by Levene's test in small and unequal sample sizes. In particular, the Brown–Forsythe test (BF(NG)e test) controls the Type I error rate in all settings studied here. The L50(NG)e test also seems to be robust to nominal significant level according to Bradley's criterion. However, only the empirical Type I error rate of the BF(NG)e test falls within the narrower interval $[0.035,\ 0.065]$. The James (J(NG)e), Alexander and Govern (AG(NG)e), and Welch (W(NG)e) tests show a good performance even in asymmetric distributions like the chi square, but they do not control the Type I error rate in highly asymmetric distributions, like the exponential. Nevertheless, they are slightly more powerful than the BF(NG)e test. In large samples, heteroscedastic tests (BF(NG)e, W(NG)e, J(NG)e and AG(NG)e) and Levene's test (L50(NG)e) perform similarly.

In conclusion, according to the results obtained in the simulation study, in small and unequal sample sizes it is recommendable to use W(NG)e, J(NG)e or AG(NG)e tests when testing homoscedasticity whenever the distributions are not highly asymmetric, otherwise the BF(NG)e test is preferable. The L50(NG)e test may be liberal.

Acknowledgments The author is sincerely grateful to two anonymous referees and the Associate Editor for their time and effort in providing very constructive, helpful and valuable comments and suggestions that have led to a substantial improvement in the quality of the paper.

6 Appendix

See Tables 1 and 2.

Table 1 Type I error rates in small and unequal sample sizes

Distribution	n_1, n_2, n_3, n_4	Test procedures						
		Be	B2e	L50(NG)e	W(NG)e	J(NG)e	BF(NG)e	AG(NG)e
Normal	4,5,6,7	0.053	0.048	0.053	0.044	0.043	0.051	0.055
	6,7,8,9	0.046	0.052	0.049	0.052	0.055	0.051	0.055
	6,9,20,30	0.046	0.048	0.054	0.049	0.047	0.053	0.051
	10,11,12,13	0.050	0.051	0.050	0.046	0.049	0.046	0.047
	4,10,18,22	0.047	0.048	0.051	0.050	0.046	0.049	0.052
	4,28,28,28	0.048	0.052	0.045	0.046	0.048	0.043	0.049
	4,4,28,28	0.050	0.045	0.052	0.042	0.048	0.048	0.051
	4,4,4,28	0.055	0.051	0.052	0.051	0.051	0.049	0.047
	8,12,18,20	0.053	0.048	0.052	0.044	0.046	0.050	0.054
	8,12,18,30	0.047	0.052	0.049	0.051	0.052	0.050	0.052
	10,14,18,20	0.045	0.046	0.048	0.053	0.051	0.053	0.052
	10,14,18,30	0.046	0.048	0.046	0.046	0.049	0.046	0.048

Table 1 continued

Distribution	n_1, n_2, n_3, n_4	Test procedures						
		Be	B2e	L50(NG)e	W(NG)e	J(NG)e	BF(NG)e	AG(NG)e
	20,22,24,26	0.047	0.051	0.053	0.051	0.051	0.052	0.050
	15,20,25,28	0.047	0.057	0.058	0.052	0.056	0.051	0.057
Mean of absolute deviations from 0.05		*0.003*	*0.002*	*0.003*	*0.003*	*0.003*	*0.002*	*0.003*
Student's t	4,5,6,7	**0.189**	0.047	0.054	0.045	0.047	0.050	0.047
	6,7,8,9	**0.219**	0.042	0.046	0.041	0.035	0.043	0.042
	6,9,20,30	**0.296**	0.034	0.061	0.049	0.049	0.052	0.050
	10,11,12,13	**0.284**	0.037	0.047	0.047	0.046	0.045	0.051
	4,10,18,22	**0.262**	0.035	0.058	0.040	0.042	0.043	0.044
	4,28,28,28	**0.310**	0.028	0.057	0.046	0.042	0.044	0.051
	4,4,28,28	**0.249**	0.029	0.063	0.043	0.037	0.039	0.043
	4,4,4,28	**0.174**	0.031	0.073	0.044	0.043	0.048	0.041
	8,12,18,20	**0.301**	0.033	0.044	0.043	0.047	0.040	0.049
	8,12,18,30	**0.323**	0.038	0.052	0.053	0.053	0.045	0.058
	10,14,18,20	**0.310**	0.039	0.049	0.049	0.050	0.045	0.052
	10,14,18,30	**0.316**	0.035	0.050	0.048	0.054	0.047	0.053
	20,22,24,26	**0.373**	0.035	0.055	0.056	0.053	0.052	0.055
	15,20,25,28	**0.353**	0.036	0.050	0.058	0.057	0.051	0.059
Mean of absolute deviations from 0.05		*0.233*	*0.014*	*0.006*	*0.005*	*0.006*	*0.005*	*0.004*
M. Norm.	4,5,6,7	**0.139**	0.044	0.052	0.042	0.046	0.048	0.044
	6,7,8,9	**0.173**	0.037	0.047	0.041	0.045	0.042	0.040
	6,9,20,30	**0.241**	0.035	0.058	0.046	0.047	0.044	0.047
	10,11,12,13	**0.245**	0.037	0.047	0.044	0.046	0.047	0.044
	4,10,18,22	**0.222**	0.032	0.051	0.044	0.039	0.040	0.043
	4,28,28,28	**0.284**	0.033	0.057	0.042	0.045	0.045	0.049
	4,4,28,28	**0.201**	0.034	0.062	0.043	0.041	0.044	0.041
	4,4,4,28	**0.137**	0.038	0.070	0.042	0.043	0.046	0.040
	8,12,18,20	**0.256**	0.031	0.048	0.043	0.041	0.044	0.046
	8,12,18,30	**0.260**	0.034	0.054	0.045	0.049	0.045	0.050
	10,14,18,20	**0.268**	0.030	0.048	0.045	0.048	0.044	0.047
	10,14,18,30	**0.273**	0.034	0.050	0.043	0.050	0.048	0.046
	20,22,24,26	**0.326**	0.031	0.048	0.054	0.053	0.048	0.054
	15,20,25,28	**0.296**	0.029	0.043	0.040	0.041	0.047	0.047
Mean of absolute deviations from 0.05		*0.187*	*0.016*	*0.005*	*0.007*	*0.005*	*0.005*	*0.005*
Uniform	4,5,6,7	0.014	0.066	0.039	0.047	0.048	0.039	0.049
	6,7,8,9	0.009	0.061	0.036	0.054	0.056	0.035	0.050
	6,9,20,30	0.007	0.068	0.053	0.048	0.053	0.048	0.039
	10,11,12,13	0.003	0.056	0.033	0.045	0.044	0.040	0.045
	4,10,18,22	0.006	**0.081**	0.042	0.057	0.055	0.043	0.052
	4,28,28,28	0.006	**0.085**	0.044	0.050	0.051	0.047	0.047
	4,4,28,28	0.013	**0.114**	0.051	0.058	0.058	0.053	0.055

Table 1 continued

Distribution	n_1, n_2, n_3, n_4	Test procedures						
		Be	B2e	L50(NG)e	W(NG)e	J(NG)e	BF(NG)e	AG(NG)e
	4,4,4,28	0.022	**0.112**	0.044	0.056	0.055	0.039	0.063
	8,12,18,20	0.005	0.058	0.043	0.056	0.057	0.041	0.051
	8,12,18,30	0.005	0.059	0.044	0.047	0.051	0.043	0.049
	10,14,18,20	0.003	0.059	0.044	0.047	0.048	0.042	0.048
	10,14,18,30	0.002	0.053	0.042	0.044	0.046	0.043	0.045
	20,22,24,26	0.001	0.046	0.040	0.038	0.039	0.037	0.042
	15,20,25,28	0.002	0.054	0.043	0.041	0.040	0.047	0.042
Mean of absolute deviations from 0.05		*0.043*	*0.020*	*0.008*	*0.005*	*0.005*	*0.008*	*0.005*
Chi square	4,5,6,7	**0.172**	**0.083**	0.061	0.062	0.063	0.064	0.057
	6,7,8,9	**0.203**	0.070	0.057	0.067	0.072	0.054	0.068
	6,9,20,30	**0.244**	0.056	0.058	0.067	0.072	0.055	**0.077**
	10,11,12,13	**0.239**	0.065	0.059	0.063	0.066	0.057	0.070
	4,10,18,22	**0.223**	0.061	0.055	0.050	0.056	0.057	0.062
	4,28,28,28	**0.258**	0.056	0.058	0.050	0.055	0.052	0.057
	4,4,28,28	**0.208**	0.049	0.061	0.048	0.051	0.052	0.049
	4,4,4,28	**0.160**	0.054	0.066	0.057	0.056	0.060	0.048
	8,12,18,20	**0.253**	0.061	0.053	0.071	0.072	0.052	0.070
	8,12,18,30	**0.276**	0.062	0.059	0.072	0.070	0.063	**0.078**
	10,14,18,20	**0.268**	0.061	0.056	0.064	0.065	0.055	0.068
	10,14,18,30	**0.269**	0.058	0.054	**0.082**	**0.079**	0.062	0.072
	20,22,24,26	**0.299**	0.061	0.055	0.068	0.063	0.054	0.066
	15,20,25,28	**0.298**	0.057	0.054	**0.076**	0.073	0.056	0.071
Mean of absolute deviations from 0.05		*0.191*	*0.011*	*0.008*	*0.014*	*0.015*	*0.007*	*0.016*
Exponential	4,5,6,7	**0.308**	**0.107**	0.073	**0.076**	**0.076**	0.067	0.072
	6,7,8,9	**0.372**	**0.109**	0.066	0.071	**0.077**	0.061	**0.080**
	6,9,20,30	**0.406**	0.060	0.054	**0.076**	**0.085**	0.053	**0.089**
	10,11,12,13	**0.419**	**0.077**	0.062	**0.079**	**0.078**	0.055	**0.083**
	4,10,18,22	**0.395**	0.070	0.065	0.063	0.058	0.056	0.074
	4,28,28,28	**0.432**	0.048	0.058	0.066	0.065	0.049	**0.081**
	4,4,28,28	**0.361**	0.052	0.065	0.044	0.046	0.048	0.058
	4,4,4,28	**0.293**	0.055	0.075	0.052	0.057	0.057	0.042
	8,12,18,20	**0.438**	0.074	0.055	**0.084**	**0.081**	0.051	**0.087**
	8,12,18,30	**0.435**	0.058	0.064	**0.084**	**0.085**	0.060	**0.090**
	10,14,18,20	**0.446**	0.074	0.057	**0.076**	**0.081**	0.056	**0.087**
	10,14,18,30	**0.454**	0.071	0.056	**0.091**	**0.094**	0.060	**0.092**
	20,22,24,26	**0.479**	0.061	0.050	0.073	0.072	0.055	0.073
	15,20,25,28	**0.466**	0.065	0.055	**0.085**	**0.085**	0.058	**0.092**
Mean of absolute deviations from 0.05		*0.357*	*0.020*	*0.011*	*0.024*	*0.025*	*0.007*	*0.030*

Nominal significance level 0.05; 10,000 iterations
Significance levels larger than 0.075 are in bold and those smaller than 0.025 are underlined

Table 2 Estimated power

Distribution	n_1, n_2, n_3, n_4	Variances	Test procedures						
			Be	B2e	L50(NG)e	W(NG)e	J(NG)e	BF(NG)e	AG(NG)e
Normal	4,5,6,7	1,6,11,16	0.430	0.309	0.132	0.243	0.243	0.201	0.308
		16,11,6,1	0.664	0.464	0.426	0.270	0.281	0.298	0.264
	6,7,8,9	1,6,11,16	0.741	0.567	0.317	0.573	0.568	0.444	0.616
		16,11,6,1	**0.864**	0.703	0.597	0.545	0.560	0.477	0.555
	6,9,20,30	1,6,11,16	**0.871**	0.788	0.501	**0.913**	**0.924**	0.770	**0.930**
		16,11,6,1	**0.999**	**0.992**	**0.995**	**0.831**	**0.835**	**0.847**	**0.894**
	10,11,12,13	1,6,11,16	**0.967**	**0.894**	0.762	**0.940**	**0.944**	0.797	**0.954**
		16,11,6,1	**0.985**	**0.930**	**0.874**	**0.930**	**0.926**	**0.826**	**0.942**
	4,10,18,22	1,6,11,16	0.654	0.554	0.258	0.571	0.589	0.473	0.615
		16,11,6,1	**0.995**	**0.960**	**0.963**	0.452	0.509	0.735	0.597
	4,28,28,28	1,6,11,16	**0.815**	0.744	0.570	0.593	0.583	0.685	0.646
		16,11,6,1	**1.000**	**0.999**	**0.999**	0.739	**0.811**	**0.935**	**0.956**
	4,4,28,28	1,6,11,16	0.564	0.471	0.171	0.432	0.426	0.381	0.494
		16,11,6,1	**0.996**	**0.975**	**0.981**	0.249	0.281	0.612	0.228
	4,4,4,28	1,6,11,16	0.532	0.412	0.072	0.259	0.254	0.390	0.441
		16,11,6,1	**0.948**	0.787	**0.945**	0.201	0.209	0.293	0.131
	8,12,18,20	1,6,11,16	**0.957**	**0.866**	0.686	**0.964**	**0.964**	**0.839**	**0.964**
		16,11,6,1	**0.998**	**0.981**	**0.976**	**0.954**	**0.954**	**0.880**	**0.960**
	8,12,18,30	1,6,11,16	**0.970**	**0.909**	0.709	**0.981**	**0.980**	**0.905**	**0.979**
		16,11,6,1	**1.000**	**0.995**	**0.997**	**0.967**	**0.967**	**0.938**	**0.968**
	10,14,18,20	1,6,11,16	**0.986**	**0.938**	**0.824**	**0.987**	**0.985**	**0.911**	**0.988**
		16,11,6,1	**0.999**	**0.989**	**0.983**	**0.987**	**0.986**	**0.938**	**0.987**
	10,14,18,30	1,6,11,16	**0.993**	**0.968**	**0.872**	**0.995**	**0.996**	**0.962**	**0.996**
		16,11,6,1	**1.000**	**0.998**	**0.999**	**0.989**	**0.990**	**0.981**	**0.993**
	20,22,24,26	1,6,11,16	**1.000**	**0.999**	**0.998**	**1.000**	**1.000**	**0.999**	**1.000**
		16,11,6,1	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	15,20,25,28	1,6,11,16	**1.000**	**0.997**	**0.989**	**1.000**	**1.000**	**0.996**	**1.000**
		16,11,6,1	**1.000**	**0.999**	**1.000**	**1.000**	**1.000**	**0.999**	**1.000**
	Mean (small samples)		***0.890***	***0.828***	*0.736*	*0.734*	*0.742*	*0.732*	*0.765*
	Mean (large samples)		***1.000***	***1.000***	***1.000***	***1.000***	***1.000***	***1.000***	***1.000***
Student's t	4,5,6,7	1,6,11,16	0.553	0.238	0.092	0.186	0.177	0.159	0.232
		16,11,6,1	0.681	0.365	0.329	0.206	0.224	0.229	0.192
	6,7,8,9	1,6,11,16	0.769	0.391	0.218	0.435	0.436	0.307	0.483
		16,11,6,1	**0.848**	0.504	0.433	0.365	0.400	0.324	0.405
	6,9,20,30	1,6,11,16	**0.889**	0.395	0.237	0.787	0.763	0.535	**0.804**
		16,11,6,1	**0.981**	**0.824**	**0.951**	0.589	0.606	0.661	0.663
	10,11,12,13	1,6,11,16	**0.940**	0.589	0.487	**0.820**	**0.816**	0.573	**0.819**
		16,11,6,1	**0.945**	0.667	0.655	0.785	0.775	0.602	0.793

Table 2 continued

Distribution	n_1, n_2, n_3, n_4	Variances	Test procedures						
			Be	B2e	L50(NG)e	W(NG)e	J(NG)e	BF(NG)e	AG(NG)e
	4,10,18,22	1,6,11,16	0.756	0.270	0.134	0.460	0.467	0.322	0.540
		16,11,6,1	**0.959**	0.726	**0.851**	0.265	0.322	0.550	0.319
	4,28,28,28	1,6,11,16	**0.849**	0.327	0.322	0.488	0.493	0.435	0.541
		16,11,6,1	**0.988**	**0.820**	**0.950**	0.392	0.409	**0.837**	0.705
	4,4,28,28	1,6,11,16	0.693	0.199	0.079	0.284	0.271	0.206	0.391
		16,11,6,1	**0.961**	0.723	**0.891**	0.164	0.169	0.510	0.107
	4,4,4,28	1,6,11,16	0.661	0.210	0.040	0.174	0.170	0.291	0.274
		16,11,6,1	**0.886**	0.591	**0.888**	0.148	0.163	0.237	0.103
	8,12,18,20	1,6,11,16	**0.937**	0.535	0.390	**0.838**	**0.844**	0.585	**0.852**
		16,11,6,1	**0.973**	0.776	**0.863**	**0.811**	**0.809**	0.694	**0.823**
	8,12,18,30	1,6,11,16	**0.954**	0.534	0.401	**0.895**	**0.883**	0.659	**0.899**
		16,11,6,1	**0.987**	**0.858**	**0.960**	**0.834**	**0.836**	0.769	**0.859**
	10,14,18,20	1,6,11,16	**0.960**	0.642	0.545	**0.902**	**0.899**	0.706	**0.903**
		16,11,6,1	**0.981**	0.781	**0.887**	**0.887**	**0.885**	0.744	**0.899**
	10,14,18,30	1,6,11,16	**0.971**	0.648	0.569	**0.931**	**0.927**	0.778	**0.936**
		16,11,6,1	**0.990**	**0.869**	**0.967**	**0.920**	**0.928**	**0.860**	**0.929**
	20,22,24,26	1,6,11,16	**0.993**	**0.834**	**0.928**	**0.987**	**0.986**	**0.948**	**0.988**
		16,11,6,1	**0.996**	**0.884**	**0.972**	**0.989**	**0.989**	**0.962**	**0.989**
	15,20,25,28	1,6,11,16	**0.991**	0.776	**0.836**	**0.977**	**0.976**	**0.906**	**0.976**
		16,11,6,1	**0.994**	**0.879**	**0.965**	**0.980**	**0.981**	**0.939**	**0.980**
	Mean (small samples)		***0.896***	*0.602*	*0.601*	*0.625*	*0.629*	*0.583*	*0.657*
	Mean (large samples)		***1.000***	***0.958***	***0.998***	***0.997***	***0.997***	***0.998***	***0.997***
M. Norm.	4,5,6,7	1,6,11,16	0.508	0.055	0.028	0.040	0.041	0.029	0.039
		16,11,6,1	0.439	0.093	0.044	0.038	0.037	0.036	0.035
	6,7,8,9	1,6,11,16	0.605	0.051	0.024	0.032	0.034	0.026	0.037
		16,11,6,1	0.563	0.069	0.041	0.034	0.035	0.032	0.037
	6,9,20,30	1,6,11,16	**0.877**	0.032	0.027	0.029	0.028	0.032	0.032
		16,11,6,1	0.706	0.084	0.187	0.036	0.040	0.031	0.041
	10,11,12,13	1,6,11,16	0.746	0.055	0.027	0.033	0.032	0.024	0.036
		16,11,6,1	0.711	0.068	0.037	0.032	0.031	0.027	0.033
	4,10,18,22	1,6,11,16	**0.827**	0.032	0.016	0.027	0.026	0.014	0.030
		16,11,6,1	0.694	0.086	0.135	0.033	0.037	0.023	0.042
	4,28,28,28	1,6,11,16	**0.922**	0.030	0.009	0.022	0.020	0.022	0.033
		16,11,6,1	**0.865**	0.075	0.102	0.032	0.033	0.035	0.041
	4,4,28,28	1,6,11,16	**0.850**	0.021	0.041	0.021	0.022	0.013	0.016
		16,11,6,1	0.685	0.075	0.209	0.030	0.030	0.026	0.030
	4,4,4,28	1,6,11,16	0.693	0.026	0.099	0.034	0.032	0.017	0.015
		16,11,6,1	0.401	0.112	0.330	0.036	0.039	0.041	0.040
	8,12,18,20	1,6,11,16	**0.834**	0.043	0.018	0.028	0.027	0.022	0.030
		16,11,6,1	0.747	0.076	0.069	0.032	0.031	0.028	0.033

Table 2 continued

Distribution	n_1, n_2, n_3, n_4	Variances	Test procedures						
			Be	B2e	L50(NG)e	W(NG)e	J(NG)e	BF(NG)e	AG(NG)e
	8,12,18,30	1,6,11,16	**0.889**	0.039	0.022	0.030	0.028	0.033	0.034
		16,11,6,1	0.758	0.089	0.139	0.032	0.034	0.029	0.034
	10,14,18,20	1,6,11,16	**0.849**	0.040	0.016	0.032	0.032	0.026	0.033
		16,11,6,1	0.788	0.070	0.056	0.031	0.034	0.028	0.034
	10,14,18,30	1,6,11,16	**0.891**	0.040	0.020	0.032	0.031	0.040	0.033
		16,11,6,1	0.786	0.082	0.107	0.031	0.033	0.029	0.038
	20,22,24,26	1,6,11,16	**0.923**	0.070	0.026	0.035	0.038	0.037	0.043
		16,11,6,1	**0.906**	0.078	0.050	0.032	0.033	0.041	0.034
	15,20,25,28	1,6,11,16	**0.926**	0.060	0.025	0.040	0.037	0.047	0.046
		16,11,6,1	**0.881**	0.084	0.071	0.031	0.033	0.033	0.038
	Mean (small samples)		*0.760*	*0.062*	*0.070*	*0.032*	*0.032*	*0.029*	*0.035*
	Mean (large samples)		***0.943***	***0.943***	***0.943***	***0.943***	***0.943***	***0.943***	***0.943***
Uniform	4,5,6,7	1,6,11,16	0.316	0.461	0.174	0.346	0.314	0.247	0.406
		16,11,6,1	0.622	0.596	0.464	0.365	0.389	0.327	0.376
	6,7,8,9	1,6,11,16	0.692	0.781	0.414	0.739	0.734	0.550	0.770
		16,11,6,1	**0.909**	**0.859**	0.680	0.692	0.701	0.552	0.725
	6,9,20,30	1,6,11,16	**0.902**	**0.992**	**0.815**	**0.992**	**0.993**	**0.951**	**0.988**
		16,11,6,1	**1.000**	**1.000**	**0.999**	**0.955**	**0.960**	**0.900**	**0.977**
	10,11,12,13	1,6,11,16	**0.994**	**0.990**	**0.892**	**0.991**	**0.991**	**0.931**	**0.991**
		16,11,6,1	**0.998**	**0.994**	**0.948**	**0.986**	**0.987**	**0.921**	**0.989**
	4,10,18,22	1,6,11,16	0.553	**0.893**	0.519	0.747	0.745	0.704	0.732
		16,11,6,1	**1.000**	**0.999**	**0.993**	0.685	0.731	0.787	**0.827**
	4,28,28,28	1,6,11,16	**0.842**	**0.987**	**0.840**	0.718	0.754	**0.889**	0.735
		16,11,6,1	**1.000**	**1.000**	**1.000**	**0.976**	**0.987**	**0.967**	**0.999**
	4,4,28,28	1,6,11,16	0.440	**0.879**	0.433	0.590	0.599	0.577	0.585
		16,11,6,1	**1.000**	**1.000**	**0.998**	0.369	0.472	0.673	0.504
	4,4,4,28	1,6,11,16	0.420	0.790	0.332	0.410	0.411	0.474	0.583
		16,11,6,1	**0.979**	**0.890**	**0.969**	0.250	0.268	0.356	0.145
	8,12,18,20	1,6,11,16	**0.985**	**0.996**	**0.893**	**0.998**	**0.998**	**0.964**	**0.998**
		16,11,6,1	**1.000**	**1.000**	**0.995**	**0.995**	**0.995**	**0.952**	**0.996**
	8,12,18,30	1,6,11,16	**0.991**	**0.999**	**0.929**	**1.000**	**0.999**	**0.985**	**0.999**
		16,11,6,1	**1.000**	**1.000**	**1.000**	**0.996**	**0.996**	**0.979**	**0.998**
	10,14,18,20	1,6,11,16	**0.998**	**1.000**	**0.966**	**1.000**	**1.000**	**0.989**	**1.000**
		16,11,6,1	**1.000**	**1.000**	**0.997**	**1.000**	**1.000**	**0.986**	**0.999**
	10,14,18,30	1,6,11,16	**0.999**	**1.000**	**0.986**	**1.000**	**1.000**	**0.998**	**1.000**
		16,11,6,1	**1.000**	**1.000**	**1.000**	**0.999**	**0.999**	**0.996**	**1.000**
	20,22,24,26	1,6,11,16	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		16,11,6,1	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	15,20,25,28	1,6,11,16	**1.000**	**1.000**	**0.999**	**1.000**	**1.000**	**1.000**	**1.000**
		16,11,6,1	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 2 continued

Distribution	n_1, n_2, n_3, n_4	Variances	Test procedures						
			Be	B2e	L50(NG)e	W(NG)e	J(NG)e	BF(NG)e	AG(NG)e
	Mean (small samples)		***0.880***	***0.932***	***0.830***	***0.814***	***0.822***	***0.809***	***0.833***
	Mean (large samples)		1.000	1.000	1.000	1.000	1.000	1.000	1.000
Chi square	4,5,6,7	1,6,11,16	0.561	0.299	0.109	0.206	0.190	0.166	0.277
		16,11,6,1	0.704	0.405	0.345	0.241	0.235	0.241	0.211
	6,7,8,9	1,6,11,16	0.785	0.459	0.239	0.470	0.484	0.307	0.486
		16,11,6,1	**0.851**	0.548	0.454	0.430	0.415	0.336	0.449
	6,9,20,30	1,6,11,16	**0.889**	0.474	0.267	**0.819**	0.789	0.539	**0.817**
		16,11,6,1	**0.990**	**0.872**	**0.953**	0.637	0.650	0.690	0.686
	10,11,12,13	1,6,11,16	**0.945**	0.673	0.524	**0.806**	0.795	0.591	**0.819**
		16,11,6,1	**0.964**	0.708	0.675	**0.804**	0.787	0.598	**0.808**
	4,10,18,22	1,6,11,16	0.755	0.335	0.148	0.501	0.475	0.327	0.559
		16,11,6,1	**0.974**	0.785	**0.863**	0.298	0.318	0.544	0.388
	4,28,28,28	1,6,11,16	**0.831**	0.415	0.345	0.532	0.530	0.428	0.575
		16,11,6,1	**0.997**	**0.902**	**0.973**	0.402	0.486	**0.846**	0.784
	4,4,28,28	1,6,11,16	0.711	0.258	0.096	0.316	0.334	0.257	0.455
		16,11,6,1	**0.979**	0.793	**0.898**	0.170	0.195	0.519	0.147
	4,4,4,28	1,6,11,16	0.656	0.282	0.055	0.205	0.181	0.313	0.296
		16,11,6,1	**0.905**	0.613	**0.887**	0.164	0.182	0.239	0.117
	8,12,18,20	1,6,11,16	**0.942**	0.611	0.405	**0.860**	**0.851**	0.600	**0.859**
		16,11,6,1	**0.983**	**0.818**	**0.865**	**0.822**	**0.833**	0.700	**0.845**
	8,12,18,30	1,6,11,16	**0.951**	0.615	0.414	**0.897**	**0.894**	0.657	**0.899**
		16,11,6,1	**0.996**	**0.897**	**0.971**	**0.849**	**0.861**	0.761	**0.881**
	10,14,18,20	1,6,11,16	**0.967**	0.679	0.536	**0.907**	**0.898**	0.713	**0.906**
		16,11,6,1	**0.991**	**0.845**	**0.888**	**0.897**	**0.904**	0.754	**0.919**
	10,14,18,30	1,6,11,16	**0.976**	0.723	0.584	**0.936**	**0.940**	0.771	**0.941**
		16,11,6,1	**0.998**	**0.911**	**0.975**	**0.939**	**0.933**	**0.874**	**0.941**
	20,22,24,26	1,6,11,16	**0.998**	**0.912**	**0.948**	**0.995**	**0.995**	**0.965**	**0.995**
		16,11,6,1	**1.000**	**0.941**	**0.985**	**0.995**	**0.995**	**0.972**	**0.996**
	15,20,25,28	1,6,11,16	**0.996**	**0.857**	**0.859**	**0.987**	**0.987**	**0.935**	**0.986**
		16,11,6,1	**0.999**	**0.940**	**0.986**	**0.991**	**0.991**	**0.960**	**0.992**
	Mean (small samples)		***0.903***	*0.663*	*0.616*	*0.646*	*0.648*	*0.593*	*0.680*
	Mean (large samples)		***1.000***	***0.991***	***1.000***	***0.999***	***0.999***	***1.000***	***0.998***
Exponential	4,5,6,7	1,6,11,16	0.646	0.281	0.101	0.178	0.181	0.134	0.214
		16,11,6,1	0.729	0.361	0.282	0.207	0.210	0.203	0.176
	6,7,8,9	1,6,11,16	**0.800**	0.389	0.171	0.376	0.376	0.220	0.383
		16,11,6,1	**0.844**	0.463	0.332	0.324	0.347	0.254	0.332
	6,9,20,30	1,6,11,16	**0.904**	0.373	0.158	0.685	0.688	0.388	0.713
		16,11,6,1	**0.978**	0.730	**0.859**	0.450	0.433	0.501	0.484
	10,11,12,13	1,6,11,16	**0.934**	0.524	0.354	0.669	0.666	0.426	0.678
		16,11,6,1	**0.940**	0.606	0.502	0.639	0.649	0.448	0.654

Table 2 continued

Distribution	n_1, n_2, n_3, n_4	Variances	Test procedures						
			Be	B2e	L50(NG)e	W(NG)e	J(NG)e	BF(NG)e	AG(NG)e
	4,10,18,22	1,6,11,16	0.792	0.252	0.093	0.373	0.378	0.227	0.470
		16,11,6,1	**0.953**	0.628	0.747	0.215	0.242	0.430	0.226
	4,28,28,28	1,6,11,16	**0.850**	0.275	0.219	0.455	0.416	0.308	0.500
		16,11,6,1	**0.991**	0.753	**0.869**	0.277	0.276	0.727	0.505
	4,4,28,28	1,6,11,16	0.767	0.195	0.064	0.220	0.214	0.163	0.350
		16,11,6,1	**0.950**	0.634	0.788	0.130	0.145	0.415	0.091
	4,4,4,28	1,6,11,16	0.729	0.217	0.040	0.153	0.155	0.250	0.193
		16,11,6,1	**0.873**	0.521	**0.824**	0.116	0.147	0.205	0.078
	8,12,18,20	1,6,11,16	**0.934**	0.460	0.261	0.728	0.737	0.438	0.745
		16,11,6,1	**0.973**	0.689	0.725	0.640	0.632	0.492	0.655
	8,12,18,30	1,6,11,16	**0.945**	0.468	0.264	**0.802**	0.795	0.519	**0.802**
		16,11,6,1	**0.987**	0.791	**0.899**	0.656	0.696	0.600	0.692
	10,14,18,20	1,6,11,16	**0.959**	0.538	0.361	0.794	**0.800**	0.518	**0.808**
		16,11,6,1	**0.980**	0.712	0.734	0.770	0.770	0.578	0.785
	10,14,18,30	1,6,11,16	**0.969**	0.540	0.345	**0.855**	**0.852**	0.583	**0.866**
		16,11,6,1	**0.992**	0.790	**0.907**	0.793	0.780	0.678	**0.803**
	20,22,24,26	1,6,11,16	**0.995**	0.778	**0.806**	**0.961**	**0.962**	**0.845**	**0.961**
		16,11,6,1	**0.997**	**0.822**	**0.896**	**0.961**	**0.963**	**0.868**	**0.964**
	15,20,25,28	1,6,11,16	**0.989**	0.685	0.674	**0.941**	**0.934**	0.768	**0.941**
		16,11,6,1	**0.994**	**0.825**	**0.916**	**0.945**	**0.941**	**0.824**	**0.948**
	Mean (small samples)		***0.907***	*0.546*	*0.507*	*0.547*	*0.550*	*0.465*	*0.572*
	Mean (large samples)		***1.000***	***0.961***	***0.994***	***0.992***	***0.992***	***0.995***	***0.990***

Structural zeros are removed and critical values are estimated
Estimated power levels larger than 0.8 are in bold

References

Akritas MG, Papadatos N (2004) Heteroscedastic one-way ANOVA and lack-of-fit tests. J Am Stat Assoc 99:368–382

Alexander RA, Govern DM (1994) A new and simpler approximation and ANOVA under variance heterogeneity. J Educ Stat 19:91–101

Bartlett MS (1937) Properties of sufficiency and statistical tests. Proc R Soc Lond A Mat A 160:268–282

Bathke A (2004) The ANOVA F test can still be used in some balanced designs with unequal variances and nonnormal data. J Stat Plan Inference 126:413–422

Boos DD, Brownie C (1989) Bootstrap methods for testing homogeneity of variances. Technometrics 31:69–82

Boos DD, Brownie C (2004) Comparing variances and other measures of dispersion. Stat Sci 19:571–578

Box GEP (1954) Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effect of inequality of variance in the one-way classification. Ann Math Stat 25:290–302

Bradley JV (1978) Robustness? Br J Math Stat Psych 31:144–152

Brown MB, Forsythe AB (1974a) Robust tests for equality of variances. J Am Stat Assoc 69:364–367

Brown MB, Forsythe AB (1974b) The small sample behavior of some statistics which test the equality of several means. Technometrics 16:129–132
Cahoy DO (2010) A bootstrap test for equality of variances. Comput Stat Data Anal 54:2306–2316
Carroll RJ, Schneider H (1985) A note on Levene's test for equality of variances. Stat Probab Lett 3:191–194
Charway H, Bailer AJ (2007) Testing multiple-group variance equality with randomization procedures. J Stat Comput Simul 77:797–803
Clinch JJ, Keselman HJ (1982) Parametric alternatives to the analysis of variance. J Educ Stat 7:207–214
Cochran WG (1954) Some methods for strengthening the common χ^2-tests. Biometrics 10:417–451
Conover WJ, Johnson ME, Johnson MM (1981) A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics 23:351–361
De Beuckelaer A (1996) A closer examination on some parametric alternatives to the ANOVA F-test. Stat Pap 37:291–305
Dijkstra JB, Werter PS (1981) Testing the equality of several means when the population variances are unequal. Commun Stat B-Simul 10:557–569
Glass GV, Peckham PD, Sanders JR (1972) Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev Educ Res 42:237–288
Harwell MR, Rubinstein EN, Hayes WS, Olds CC (1992) Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. J Educ Stat 17:315–339
Hui W, Gel YR, Gastwirth JL (2008) lawstat: an R package for law, public policy and biostatistics. J Stat Softw 28. http://www.jstatsoft.org/v28/i03/paper
Hines WGS, O'Hara Hines RJ (2000) Increased power with modified forms of the Levene (Med) test for heterogeneity of variance. Biometrics 56:451–454
Hsiung T, Olejnik S, Huberty CJ (1994) Comment on a Wilcox test statistic for comparing means when variances are unequal. J Educ Stat 19:111–118
Iachine I, Petersen HC, Kyvikc KO (2010) Robust tests for the equality of variances for clustered data. J Stat Comput Simul 80:365–377
James GS (1951) The comparison of several groups of observations when the ratios of the population variances are unknown. Biometrika 38:324–329
Johansen S (1980) The Welch-James approximation of the distribution of the residual sum of squares in weighted linear regression. Biometrika 67:85–92
Kenny DA, Judd CM (1986) Consequences of violating the independence assumption in analysis of variance. Psychol Bull 99:422–431
Keselman HJ, Rogan JC, Feir-Walsh BJ (1977) An evaluation of some nonparametric and parametric tests for location equality. Br J Math Stat Psychol 30:213–221
Keselman HJ, Wilcox RR, Algina J, Othman AR, Fradette K (2008) A comparative study of robust tests for spread: Asymmetric trimming strategies. Br J Math Stat Psychol 61:235–253
Keyes TK, Levy MS (1997) Analysis of Levene's test under design imbalance. J Educ Behav Stat 22: 227–236
Layard MWJ (1973) Robust large-sample tests for homogeneity of variances. J Am Stat Assoc 68:195–198
Levene H, Levene H (1960) Robust tests for equality of variances. Essays in Honor of Harold Hotelling. In: Olkin I, Ghurye SG, Hoeffding W, Madow WG, Mann HB (eds) Contributions to probability and statistics. Stanford University Press, Palo Alto, p 292
Lim TS, Loh WY (1996) A comparison of tests of equality of variances. Comput Stat Data Anal 22:287–301
Lix LM, Keselman JC, Keselman HJ (1996) Consequences of assumption violations revisited, a quantitative review of alternatives to the one-way analysis of variance F test. Rev Educ Res 66:579–619
Loh WY (1987) Some modifications of Levene's test of variance homogeneity. J Stat Comput Simul 28:213–226
Markowski CA, Markowski EP (1990) Conditions for the effectiveness of a preliminary test of variance. Am Stat 44:322–326
Mehrotra DV (1997) Improving the Brown-Forsythe solution to the generalizied Behrens-Fisher problem. Commun Stat-Simul 26:1139–1145
Neuhäuser M (2007) A comparative study of nonparametric two-sample tests after Levene's transformation. J Stat Comput Simul 77:517–526
Neuhäuser M, Hothorn LA (2000) Location-scale and scale trend tests based on Levene's transformation. Comput Stat Data Anal 33:189–200
Noguchi K, Gel YR (2010) Combination of Levene-type tests and a finite-intersection method for testing equality of variances against ordered alternatives. J Nonparametr Stat 22:897–913

O'Neill ME, Mathews K (2000) A weighted least squares approach to Levene's test of homogeneity of variance. Aust Nz J Stat 42:81–100
Oshima TC, Algina J (1992) Type I error rates for James's second-order test and Wilcox's Hm test under heteroscedasticity and non-normality. Br J Math Stat Psychol 45:255–263
Parra-Frutos I (2009) The behaviour of the modified Levene's test when data are not normally distributed. Comput Stat 24:671–693
Rogan JC, Keselman HJ (1977) Is the ANOVA F-test robust to variance heterogeneity when samples sizes are equal? Am Educ Res J 14:493–498
Rubin AS (1983) The use of weighted contrasts in analysis of models with heterogeneity of variance. P Bus Eco Stat Am Stat Assoc 347–352
Scheffé H (1959) The analysis of variance. Wiley, New York
Schneider PJ, Penfield DA (1997) Alexander and Govern's approximation, providing an alternative to ANOVA under variance heterogeneity. J Exp Educ 65:271–286
Siegel S, Tukey JW (1960) A nonparametric sum of ranks procedure for relative spread in unpaired samples. J Am Stat Assoc 55:429–444 (corrections appear in vol. 56:1005)
Welch BL (1951) On the comparison of several mean values, an alternative approach. Biometrika 38: 330–336
Wilcox RR (1988) A new alternative to the ANOVA F and new results on James' second-order method. Br J Math Stat Psychl 41:109–117
Wilcox RR (1989) Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA models. J Educ Stat 14:269–278
Wilcox RR (1990) Comparing the means of two independent groups. Biom J 32:771–780
Wilcox RR (1995) ANOVA: a paradigm for low power and misleading measures of effect size? Rev Educ Res 65:51–77
Wilcox RR (1997) A bootstrap modification of the Alexander-Govern ANOVA method, plus comments on comparing trimmed means. Educ Psychol Meas 57:655–665
Wilcox RR, Charlin VI, Thompson KL (1986) New Monte Carlo results on the robustness of the ANOVA F, W and F-statistics. Commun Stat Simul C 15:933–943
Wludyka P, Sa P (2004) A robust I-sample analysis of means type randomization test for variances for unbalanced designs. J Stat Comput Simul 74:701–726