Comput Stat (2016) 31:1607–1631
DOI 10.1007/s00180-016-0656-4

ORIGINAL PAPER

Preliminary tests when comparing means

I. Parra-Frutos[1]

Received: 13 February 2015 / Accepted: 4 April 2016 / Published online: 29 April 2016

Abstract The aim of this paper is to find a procedure to test equal means that is robust at the significance level. A simulation study is conducted to compare the performance of different strategies, including unconditionally applying the bootstrap ANOVA and twelve adaptive tests that take in pre-testing normality, homoscedasticity and skewness. Various final tests of equal means have been considered, like the ANOVA, bootstrap ANOVA, Welch, Brown–Forsythe and bootstrap James test. Our simulation results reveal that the usual adaptive test used by applied researchers (based on testing normality and homoscedasticity to choose from the ANOVA, Welch and Kruskal–Wallis tests) performs poorly. The simulation results show that preliminary tests may improve the performance of a test, and that this depends on the pre-tests chosen. In particular, we find that using decisions on normality to select the right homoscedasticity test and then choosing between the Brown–Forsythe test and the bootstrap ANOVA leads to controlling the Type I error rate in all of the settings studied.

Keywords Bootstrap ANOVA · Adaptive tests · Tests of equal means · Behrens–Fisher problem · Monte Carlo simulation

1 Introduction

It is widely known that most statistical tests require some assumptions regarding parent populations or their variances. The normality assumption is crucial in order to be able to choose between parametric and nonparametric methods. When comparing

✉ I. Parra-Frutos
ipf@um.es

[1] Department of Quantitative Methods for Economics and Business, Economics and Business School, University of Murcia, Murcia, Spain

two or more normal populations, homoscedasticity is the key assumption. Thus, it is a common practice to pre-test normality and homoscedasticity before applying a test of equal means or even only to pre-test homoscedasticity given the robustness of tests to non-normal populations. The use of preliminary tests to check the assumptions in order to choose the right test is a procedure that has received various names in the literature: two-step procedure, compound test, hierarchical test or adaptive test.

The problem of comparing means with unknown variances, even with normal populations, is referred to in the literature as the Behrens–Fisher problem, and researchers are still trying to improve the approximate solutions given so far. We address this problem and find a procedure that seems to control the Type I error rate in all the settings studied. We are also interested in showing the impact of preliminary testing of normality and homoscedasticity when comparing two or more population means against the alternative procedure of not doing so by simply applying the bootstrap ANOVA test. The bootstrap ANOVA and the bootstrap Brown–Forsythe tests show a good and similar behaviour in a wide variety of settings (Parra-Frutos 2014).

We investigate different procedures, some of which are based on preliminary tests. However, preliminary tests of equality of variances before a test of location are no longer widely recommended by some statisticians. A variety of authors including, Albers et al. (2000), Zimmerman (2004) and Rasch et al. (2011) offer various reasons for discontinuing the use of preliminary tests of equality of variances. Normality tests are not exempt from this criticism (Schoder et al. 2006; Rasch et al. 2011; Rochon and Kieser 2011). Nevertheless, in many applications the accuracy of the final inference has been improved by using preliminary test (Hall and Padmanabhan 1997; O'Gorman 1997; Freidlin et al. 2003; Miao and Gastwirth 2009).

When the violation of an assumption can severely affect the behavior of a test, an appropriate preliminary test enables one to choose a test that has high efficiency. Schucany and Ng (2006) noted that preliminary tests must be used with care, as at the second stage, the analysis is conditional on the results of the first-stage test. If preliminary tests are not good enough then they are not appropriate for an adaptive test, since they would lead to wrong decisions too frequently, and consequently to a bad selection of the final test. So, the appropriate selection of the preliminary tests as well as the final test seems to be fundamental. Thus, we investigate which could be the best tests of homoscedasticity, normality and equal means, according to the literature.

Recent investigations on selecting the best homoscedasticity test point out that the behaviour of the Levene test may be improved by other tests (e.g. Charway and Bailer 2007; Parra-Frutos 2009; Cahoy 2010). For normality testing, it seems that the Shapiro–Wilks test has a better performance than the Kolmogorov–Smirnov test (e.g. Mendes and Pala 2003; Keskin 2006; Farrel and Rogers-Stewart 2006; Razali and Wah 2010). On the other hand, the Welch (1951) test may also be improved by other tests under non-normality (Alexander and Govern 1994; Hsiung et al. 1994; Oshima and Algina 1992; Parra-Frutos 2014). However, the Levene, the Kolmogorov–Smirnov and the Welch tests have been included in adaptive tests in studies that conclude that preliminary tests should not be used (Zimmerman 2004; Schoder et al. 2006; Rasch et al. 2011). Our simulation study shows that an adaptive test may work well depending on the tests involved.

The tests for equal variances that have received the most attention are the F, the Levene (1960), and the Bartlett (1937) tests. The F test and the Bartlett tests are very sensitive to normality and behave extraordinarily badly in non-normal populations. The Levene test is often presented as an option for testing homoscedasticity in popular statistics packages. Recent studies use different approaches to improve this test (Keselman et al. 2008; Neuhäuser 2007; Lim and Loh 1996; Wludyka and Sa 2004; Charway and Bailer 2007; Parra-Frutos 2009; Cahoy 2010). Parra-Frutos (2013) shows that for small, unequal sample sizes the Levene test does not control the Type I error rate and it is recommendable to substitute the ANOVA-step by a heteroscedastic alternative. The best option seems to be the Welch test including the modification by Noguchi and Gel (2010) and estimation of critical values.

There are a significant amount of normality tests available in the literature. However, the most common normality test procedures available in statistical software are the Shapiro–Wilk test, Kolmogorov–Smirnov (1933, 1939) test, Anderson-Darling (1954) test and Lilliefors (1967) test. Razali and Wah (2010) concluded that of these four tests, the Shapiro–Wilk test is the most powerful for all types of distribution and sample sizes, whereas the Kolmogorov–Smirnov test is the least powerful. However, the power of the Shapiro–Wilk test is still low for small sample sizes. The performance of the Anderson–Darling test is quite comparable with that of the Shapiro–Wilk test, and the Lilliefors test always outperforms Kolmogorov–Smirnov test. According to D'Agostino and Stephens (1986, chap. 9) and D'Agostino et al. (1990), the Chi-squared test and the Kolmogorov test (1933) have poor power properties and should not be used when testing for normality. The Shapiro–Wilk test is an omnibus and the most powerful test in most situations (Shapiro et al. 1968; Mendes and Pala 2003; Keskin 2006; Farrel and Rogers-Stewart 2006; Razali and Wah 2010).

Various procedures can be used to test the equality of means. We focus on those that seem to control the Type I error rate in some settings (Parra-Frutos 2014): the ANOVA, the bootstrap ANOVA, the Brown–Forsythe, the bootstrap Brown–Forsythe, the Welch and the bootstrap James tests. According to Lix et al. (1996) the Welch test can be used to compare population means where variance heterogeneity exists. However, if the data appear to be highly skewed, this procedure should be avoided. Parra-Frutos (2014) shows that the Welch, the Brown–Forsythe, the bootstrap ANOVA and the bootstrap Brown–Forsythe tests control the Type I error rate in symmetric distributions. If there are one or two asymmetric distributions, when there are three populations involved, the bootstrap ANOVA and the bootstrap Brown–Forsythe tests seem to be the best choice. However, if all distributions are asymmetric the bootstrap James test shows the best behaviour.

Taking into account these results given in the literature, we construct some strategies to compare population means and compare them with the classical ones, used by applied researchers and still recommended in textbooks, and the bootstrap ANOVA. In particular, we study twelve adaptive tests for equal means and compare with the bootstrap ANOVA test. Basically, to construct adaptive tests we consider the Shapiro–Wilk test to check normality. To check homoscedasticity we use the Bartlett test, the Levene test and a modified Levene test (W(NG)e test) suggested by Parra-Frutos (2013) that outperforms the Levene test. We also consider pre-testing symmetry. As final tests of equal means we use the ANOVA, bootstrap ANOVA, Brown–Forsythe, Welch and

Bootstrap James tests. A simulation study is conducted to check the performance of tests in a wide variety of scenarios. The main objective is to find settings where the tests may fail to control the Type I error rate, if there is any. Thus, a great number of scenarios has been considered, adjusting the degree of heteroscedasticity, sample sizes, skewness and kurtosis of distributions, and positive and negative pairing of variances and sample sizes.

2 Description of adaptive tests

The performances, in terms of Type I error rate, of twelve adaptive tests are studied along with the bootstrap ANOVA test. Preliminary tests may draw erroneous conclusions about the normality or homoscedasticity of data. That is, they may give rise to false-positive decisions (reject normality with normal data or reject homoscedasticity with homoscedastic data) or false-negative decisions (not rejecting normality or homoscedasticity when they should). The first type of decision leads to applying an alternative method (more robust to violation of the particular assumption) that usually has less power or less control of the Type I error rate. The second leads to using a test that behaves badly under non-normality or heteroscedasticity. Thus, it is sometimes argued that generally using more robust methods whenever there are doubts about distributional assumptions is a simple and safe strategy. Therefore, we are interested in comparing the behaviour of the adaptive tests with that of the bootstrap ANOVA test used unconditionally, which seems to have a good behaviour in nearly all type of scenarios (Parra-Frutos 2014).

We propose some new strategies to test equal means and compare them to some common procedure used by applied researchers and to an unconditional and widely robust (at the significance level) test like the bootstrap ANOVA test (BA test).

The A1 test (see Fig. 1) is based on a common strategy used by applied researchers. It consists of testing first the normality, applying the Shapiro–Wilk test, solely to decide between parametric or non-parametric testing of equality of means. In the first case it is necessary to test homogeneity of variances to choose a homoscedastic test for equal means (ANOVA test) or a heteroscedastic alternative to it; the Welch test is performed as it has been one of the most popular heteroscedastic alternatives to ANOVA. The most used test for homogeneity of variances is the Levene test since it is offered by standard statistical software. If normality is rejected then the Kruskal-Wallis test is applied.

According to the general theory of hypothesis testing, the probability of Type I error P_I is fixed by the researcher and thus controlled at a significance level of, typically, 5 %. That is, the probability of rejecting the null hypothesis even though the data do verify is forced to be low. However, in preliminary tests, researchers are commonly more interested in the probability of the Type II error (data is accepted as proceeding from a normal population or from homoscedastic distributions even though it actually originates from different distributions) P_{II}. Researchers may prefer to have a low P_{II}, and hence a high power and sacrifice P_I. Many authors have pointed out that the statistical power of goodness-of-fit test is generally low, particularly for small sample sizes (Shapiro et al. 1968; Mendes and Pala 2003). Schoder et al. (2006) investigated the performance of the Kolmogorov–Smirnov test on non-normal data and concluded

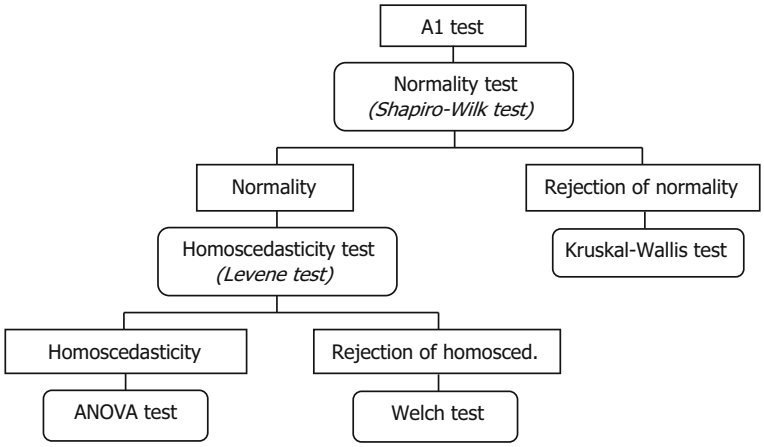

Fig. 1 Scheme of the A1 adaptive test

that preliminary testing for normality is not recommendable for small-to-moderate sample sizes due to its low power (high Type II error rates). From their point of view, it might be prudent to increase the probability of Type I error to 10 % in order to improve the statistical power of the test and hence its sensitivity to deviations from normality. Gnandesikan (1977) also stated that for p-values of 0.1 or higher, normality is a reasonable assumption.

Paull (1950) and Bancroft (1964) pointed out that the choice of the significance level of a preliminary test is restricted since one wishes the size of the final test to be close to the adopted standards of .05 or .01. The precise value of the final test will depend upon the chosen level of significance for the preliminary test. They recommended that the level of a preliminary test should be greater than 5 %, in particular, about 25 %. Paull argues that the use of the 25 % significance level for the preliminary test provides a reasonable amount of protection against an error in judgement regarding the true value. Gastwirth et al. (2009) found that the size of the Levene-type preliminary test should be between 15 and 25 %. In this case, they found that the adaptive ANOVA (which includes a Levene-type preliminary test to see whether the variances are equal or not) is a valid procedure and more robust to departures from the equal variance assumption than the usual ANOVA test.

In line with these arguments we propose the $A1_5$ test that uses 5 % as the nominal significance level in preliminary tests and the $A1_{20}$ test with 20 %. The nominal significance level of final tests is always 5 %. The subscript in the name of the procedures stands for the significance level of preliminary tests.

The A2 test (see Fig. 2) is also quite widely used by applied researchers and is based on the robustness of tests of equal means to the assumption of normality. It only focuses on detecting heteroscedasticity with the Levene test as preliminary test. If homoscedasticity is not rejected the ANOVA test is applied, otherwise the Welch test is used. We also propose two versions, the $A2_5$ and the $A2_{20}$ tests.

New strategies are proposed in the following adaptive tests. They are based on more recent results provided by the literature on testing equality of means and homoscedasticity. The A3 test (see Fig. 3) basically includes two modifications to the A1 test.

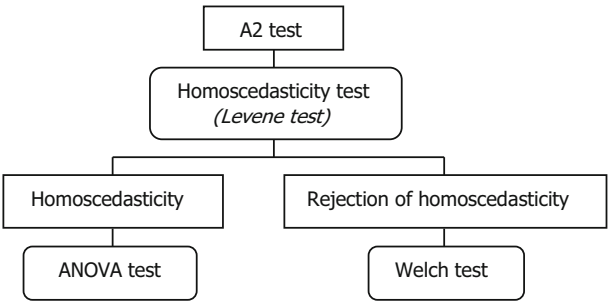

Fig. 2 Scheme of the A2 adaptive test

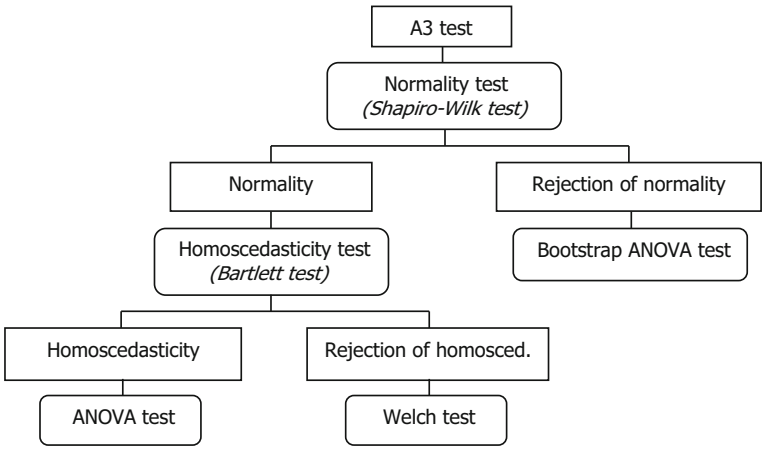

Fig. 3 Scheme of the A3 adaptive test

These are, on the one hand, a different homoscedasticity test in case of no rejection of normality (the Barlett test) and, on the other, in case of rejection of normality, it does not use a non-parametric test for equal means but the bootstrap ANOVA test, which is robust in a great variety of settings. We investigate the $A3_5$ test and the $A3_{20}$ test.

The A4 test (see Fig. 4) proposes changing the homoscedasticity test to the W(NG)e test in the case of rejecting the normality. According to Parra-Frutos (2014), to test equal means, the Brown–Forsythe test seems to control Type I error rate if distributions are homoscedastic or low heteroscedastic, and in the case of high heteroscedasticity it is the bootstrap ANOVA test. With this procedure we focus on using the right homoscedasticity test depending on the normality of the distributions, that is, the result of the normality test is only used to choose the appropriate homoscedasticity test. Therefore, this adaptive test may be considered a three-step procedure.

The Bartlett test is very sensitive to the normality assumption, so we consider it a priority to have a low probability of false negatives when testing normality. Therefore, a high power for the normality test is needed. Augmenting the significance level produces the desired result. On the other hand, the problem when testing homoscedasticity is different since the Brown–Forsythe test is robust in homoscedastic and low heteroscedastic distributions and the bootstrap ANOVA test may not control the Type

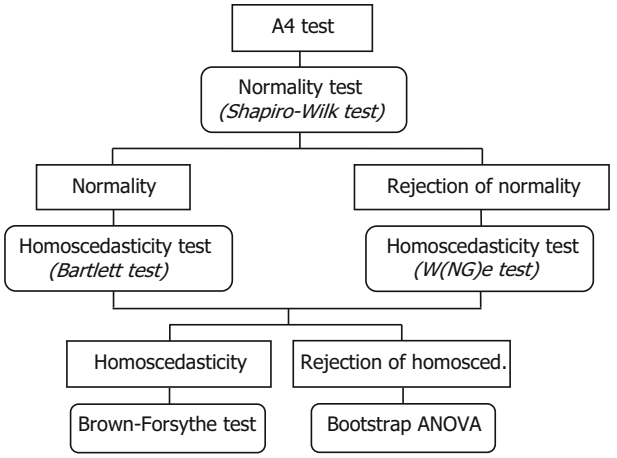

Fig. 4 Scheme of the A4 adaptive test

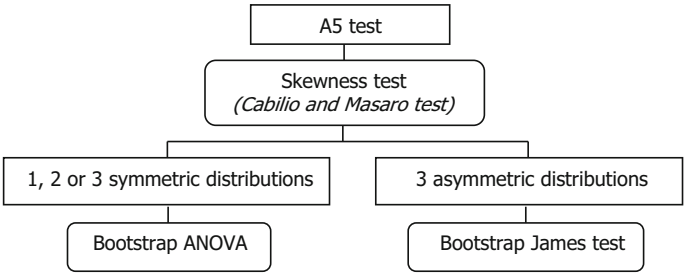

Fig. 5 Scheme of the A5 adaptive test

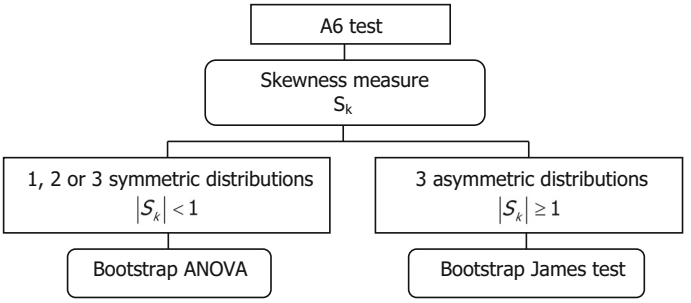

Fig. 6 Scheme of the A6 adaptive test

I error rate in those cases. Therefore, a low Type I error probability is more appropriate, that is, low probability of rejecting homoscedasticity when distributions are homoscedastic. Hence, we keep the significance level of homoscedasticity test at 5 %. The resulting procedure with different significance levels for each preliminary test is called the $A4_m$ test. We also investigate the $A4_5$ and the $A4_{20}$ tests.

Since the behaviour of tests of equal means depends on the symmetry of the distributions we also propose to study the skewness of distributions as a preliminary test,

which we do in the A5 and A6 tests. The A5 test (Fig. 5) begins by testing skewness with the Cabilio and Masaro test. If distributions are symmetric or there is one or two asymmetrics then one of the best tests for equal means is the bootstrap ANOVA (Parra-Frutos 2014) since it controls the Type I error rate. If the three distributions are asymmetric then the best option is the bootstrap James test, although it may fail to control the Type I error rate. We also propose to use a skewness measure based on the Cabilio and Masaro statistic in the A6 test (Fig. 6) to discriminate between symmetric or skewed distributions. For sample sizes equal or higher than 9, the 95th percentile of the Cabilio-Masaro statistic, S_k, is higher than 1 when sampling from a normal distribution, for others it is near to it (Cabilio and Masaro 1996). Thus, we propose that a distribution is asymmetric if the $|S_k|$ is equal or higher than 1.

3 Description of preliminary and final tests

3.1 The Shapiro–Wilk test

The test statistic is basically the square of the Pearson correlation coefficient between given data and their corresponding normal scores:

$$SW = \frac{\left(\sum_{i=1}^{m} a_i \left(Y_{(n-i+1)} - Y_{(i)}\right)\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}$$

where $Y_{(i)}$ are the order statistics of the random sample $Y_1, \ldots, Y_n$; m is the greatest integer in $n/2$ and a_i are the coefficients tabulated in Shapiro and Wilk (1965). A too small value of SW, that is, too far below 1, is indicating that the sample looks non-normal.

Let $y_{ij}, i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, denote the jth observation from the ith sample with size n_i. We have k samples with n_i observations each.

3.2 The Bartlett test

Its statistic is given by

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

where

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

$$S_i^2 = \frac{\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2}{n_i - 1} \qquad \bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$$

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

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

Thus the null hypothesis is rejected if $B > \chi^2_{k-1;\alpha}$, where $\chi^2_{k-1;\alpha}$ is the upper tail critical value for the χ^2_{k-1} distribution.

3.3 The W(NG)e test

This test has been proposed by Parra-Frutos (2013) and behaves well in most situations. It is a homoscedasticity test based on the Levene test but substituting the ANOVA-step by the Welch test. It includes the Noguchi and Gel (2010) procedure that consists of a modification of the structural zero removal method proposed by Hines and O'Hara Hines (2000) after applying the Keyes and Levy (1997) adjustment in order to preserve the null hypothesis. Empirical percentiles of the test statistic are used as critical values instead of using the approximate distribution. The test statistic is

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

where

$$Z_{ij} = \frac{\left| Y_{ij} - \tilde{Y}_i \right|}{\kappa_{n_i}} \quad \tilde{Y}_i \text{ is the } i\text{th group median}$$

$$\kappa_{n_i} \cong \sqrt{\frac{2}{\pi} \left(1 - \frac{1}{n_i}\right)}$$

$$w_i = \frac{n_i}{S^2_{Z,i}},$$

$$W = \sum_{i=1}^{k} w_i,$$

$$S^2_{Z,i} = \frac{\sum_{j=1}^{n_i} \left(Z_{ij} - \bar{Z}_i\right)^2}{n_i - 1},$$

$$\bar{Z}' = \frac{\sum_{i=1}^{k} w_i \bar{Z}_i}{W}.$$

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

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

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, Z_{ij} 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 $Q^*_{(1)}, Q^*_{(2)}, \ldots, Q^*_{(100M)}$. The 5 percent empirical critical value, then, is obtained as $C = [Q^*_{(95M)} + Q^*_{(95M+1)}]/2$. If the observed test statistic is higher than C then the null hypothesis is rejected. We use $M = 10{,}000$, that is, 1,000,000 iterations.

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 (2000), 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 $m_i = \left[\frac{1}{2}n_i\right]$. Noguchi and Gel (2010) propose a procedure to eliminate 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.

3.4 The one-way ANOVA

The test statistic is

$$F = \frac{(N-k)\sum_{i=1}^{k} (\bar{Y}_i - \bar{Y})^2 n_i}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}$$

where

$$\bar{Y}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \quad \bar{Y} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}}{N}$$

The null hypothesis, population means are equal, is rejected if the computed F statistic is higher than $F_{k-1,N-k;\alpha}$, the $(1-\alpha)$ percentile of the F distribution with $k-1$ and $N-k$ degrees of freedom.

3.5 The Welch test

It is given by the test statistic

$$Q = \frac{\sum_{i=1}^{k} w_i \left(\bar{Y}_i - \bar{Y}'\right)^2/(k-1)}{1 + \frac{2(k-2)}{k^2-1}\sum_{i=1}^{k} \frac{(1-w_i/W)^2}{n_i - 1}},$$

where

$$w_i = \frac{n_i}{S^2_{Y,i}},$$
$$W = \sum_{i=1}^{k} w_i,$$
$$S^2_{Y,i} = \frac{\sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2}{n_i - 1},$$
$$\bar{Y}' = \frac{\sum_{i=1}^{k} w_i \bar{Y}_i}{W}.$$

The null hypothesis -population means are equal-, is rejected if the computed Q statistic is higher than $F_{k-1,v;\alpha}$, the $(1-\alpha)$ percentile of the F distribution with $k-1$ and v degrees of freedom.

3.6 The bootstrap method

The idea is to approximate the distribution of the test statistic under the null hypothesis. Following the usual method of bootstrap hypothesis testing (Efron and Tibshirani 1993), the samples of $Y_i,\ i = 1, \ldots, k$, should be transformed so as to satisfy the null hypothesis; let them be $T_i,\ i = 1, \ldots, k,$. The null distribution of the test statistic is obtained by drawing B times k bootstrap samples, one of each pseudo-population $T_i,\ i = 1, \ldots, k$, and calculating the test statistic for each group of k bootstrap samples. This leads to B observations of the test statistic based on the bootstrap samples, that is, the bootstrap distribution of the test statistic.

A bootstrap procedure when comparing means would be as follows. Letting $\bar{y}$ be the mean of the combined sample, we can translate samples so that they have mean $\bar{y}$, and then resample each pseudo-population separately. The transformed observations are given by $t_{ij} = y_{ij} - \bar{y}_i + \bar{y}$. The T statistic is calculated for the original k samples, T_0, and for each group of k bootstrap samples, $T*^{(b)},\ b = 1, \ldots, B$. Let R be the number of times that $T*^{(b)} > T_0$, then the bootstrap p-value is given by R/B. We bootstrap the ANOVA test and note it as BA test.

3.7 The skewness test (Cabilio and Masaro 1996)

It is based on the test statistic

$$S_k = \frac{\sqrt{n}\left(\bar{Y} - \tilde{Y}\right)}{S}$$

The null hypothesis, the population is symmetric, is rejected if the absolute value of the computed S_k statistic is equal to or higher than $P_{\alpha/2}$, the $(1-\alpha/2)$-percentile of the distribution of S_k. The estimated percentiles when sampling from a normal distribution are given by Cabilio and Masaro (1996).

4 Design of the simulation

The simulation results are based on 5000 replications in non-bootstrap tests and on 1000 bootstrap replications for 5000 Monte Carlo replications in bootstrap tests. Tables 1 and 2 include the distributions used in the simulation study with their mean (μ), standard deviation (σ), asymmetry ($\gamma_1 = \mu_3/\mu_2^{3/2}$) and excess kurtosis ($\gamma_2 = \mu_4/\mu_2^2 - 3$) where $\mu_r = E[(Y - E(Y))^r]$. Appropriate data transformations are carried out to generate data sets from those populations with the desired variances σ_i^2, $i = 1, 2, 3$, and maintaining equal population means. To generate data from the g-and-h distribution, we use Cribbie et al. (2012) and Hoaglin (1985).

We consider three groups or populations. According to the notation of distributions given in Tables 1 and 2, a combination of distributions described by, for example, *n_exp_gh020* represents a combination of a normal, an exponential and a g-and-h (with $g = 0.2$ and $h = 0$) distributions. The notation for each distribution is included in the tables. We consider 16 combinations of distributions (see Table 3) which represents different combinations of symmetry and kurtosis.

The main objective of the simulation is to study the behaviour of each test in nearly all possible scenarios that may be found in applied research. So a great variety of scenarios are considered. Nine configurations of sample sizes are studied, seven of

Table 1 Distributions, and their characteristics, used in the simulation study

Notation	Symmetric			Asymmetric	
	$N(0, 1)$	t_4	$U(-a, a)$	χ_4^2	$Exp(3)$
	n	t	u	chi	exp
μ	0	0	0	4	1/3
σ	1	1.41	$a/\sqrt{3}$	2.83	1/3
γ_1	0	0	0	1.41	2
γ_2	0	–	1.8	3	6

Table 2 g-and-h distributions, and their characteristics, used in the simulation study

	g-and-h distributions					
	Symmetric				Asymmetric	
g	0	0	0	0	0.2	0.81
h	0.043	0.14	0.2	0.22	0	0
Notation	gh00043	gh0014	gh002	gh0022	gh020	gh0810
μ	0	0	0	0	0.1	0.48
σ	1.07	1.28	1.47	1.54	1.03	1.65
γ_1	0	0	0	0	0.61	3.8
γ_2	0.7	5.7	33.2[a]	103[a]	0.7	33.3[a]

[a] Heavy-tailed distributions

Table 3 Combinations of distributions

Symmetric	1 Asymmetric	2 Asymmetric	3 Asymmetric
3 × Normal	n_gh0014_gh0810	exp_n_gh020	3 × Chi
3 × t		gh0810_gh020_gh0014	3 × Exp
3 × Unif			3 × gh0810
3 × gh002			exp_exp_chi
n_t_u			exp_exp_gh0810
gh0014_gh002_gh0022			exp_gh0810_gh0810
gh00043_gh0014_gh002			

small sizes and two of large samples: (10, 10, 10), (15, 15, 15), (25, 25, 25), (10, 10, 15), (10, 10, 25), (10, 15, 15), (15, 20, 25), (30, 30, 30) and (30, 40, 60). The same sample sizes were also used in Parra-Frutos (2014).

We use a homoscedastic setting with all variances equal to one, and different heteroscedastic settings. In particular, the following combinations of standard deviations $(\sigma_1, \sigma_2, \sigma_3)$ from mild to extreme heteroscedasticity are applied: (1, 1.1, 1.2), (1, 1.5, 1.75) and (1, 5, 8). The reverse of the above has also been included, so pairing between variances and sample sizes is considered. The balancedness of the design is a problem in heteroscedastic settings, in particular positive/negative pairings of unequal sample sizes and unequal variances (Keselman et al. 1977). High heteroscedasticity is represented by $(\sigma_1, \sigma_2, \sigma_3) = (1, 5, 8)$ and $(\sigma_1, \sigma_2, \sigma_3) = (8, 5, 1)$. The first combination is studied with all sample size combinations and the second with unequal sample size combinations.

Fifty-one different settings have been described for each combination of distributions. That is, five unequal sample size combinations with each of the seven combinations of variances (35 cases) and four equal sample size combinations with four variance combinations (16 cases). As a result, each test has been studied in 816 different settings, that is, 51 different cases for each of the 16 distribution combinations.

Empirical Type I error rates were recorded for all tests. The robustness of a procedure, with respect to Type I error control, is determined using Bradley's (1978) liberal criterion. That is, a procedure is deemed robust with respect to Type I error if the empirical percentages of the Type I error falls within the range $\alpha \pm \alpha/2$. If $\alpha = 5\,\%$, then the interval is given by [2.5, 7.5]. An empirical significance level over 7.5 % would indicate a liberal test and one below 2.5 % a conservative test. We also calculate the percentage of times a test does not control the Type I error rate in a set of simulations, and denominate it the *failure rate*. A robust test will show $\hat{\alpha} \in [2.5, 7.5]$ in all the cases simulated, so its failure rate will be 0 %.

We obtain the percentage of times the estimated significance level observed in our simulation results was under 2.5 % and call it the lower failure rate. We also provides the minimum $\hat{\alpha}$. Thus, if minimum $\hat{\alpha}$ was 2 % and there were 1.6 % of cases where $\hat{\alpha}$ is under 2.5 % then it would be noted as 2 (1.6). If minimum $\hat{\alpha}$ was over 2.5 % then it would not be accompanied by the lower failure rate, since it would be 0. We also

obtain the percentage of times $\hat{\alpha}$ was above 7.5 % (the upper failure rate) along with the maximum $\hat{\alpha}$ observed. If the largest $\hat{\alpha}$ was, for example, 9.6 % and there were 2.3 % of simulated $\hat{\alpha}$ over 7.5 % then we would describe it as 9.6 (2.3). The failure rate may be obtained adding the lower and upper rates of failure. In the example it would be 3.9 (1.6 + 2.3).

5 Simulation results

The simulation results are presented in Tables 4, 5, 6, 7, 8, 9, 10, 11 and 12. In these tables we include the failure rate and the minimum and maximum $\hat{\alpha}$ with the lower and upper failure rates, respectively, if they were different from 0. In Tables 4 and 5 we include the general simulation results for all the 816 settings considered. These tables show that the A4 test, in particular the $A4_5$ and the $A4_m$, always controls the Type I error rate. A more detailed study of the simulation results of $A4_5$ test shows that in the 84.4 % of all the cases studied the $\hat{\alpha}$ takes values in the interval $[\alpha - 1.25, \alpha + 1.25) = [3.75, 6.25)$, 12.5 % in [2.5, 3.75) and 3.1 % in [6.25, 7.4].

The worse behaviour observed corresponds to the A1 test, with the highest failure rate (33.9) and the highest maximum Type I error rate (74.7 %). The A1 test is the procedure still recommended in textbooks and web pages for applied researchers. However, according to our simulation results it is the least recommendable. A ranking of tests according to their failure rate would be A4, A5 (3.3), BA (3.6), A6 (3.6), A3 (3.9), A2 (12.9) and A1 (33.9).

Tables 6 and 7 show the simulation results according to some characteristic of the parent populations. We may observe that the classical A1 test does not show control of the Type I error rate in any situation, even when distributions are normal or

Table 4 Failure rate, number of cases where the test does not control the Type I error rate and the minimum and maximum estimated significance level for the A1, A2 and A3 tests

	$A1_5$	$A1_{20}$	$A2_5$	$A2_{20}$	$A3_5$	$A3_{20}$
Failure rate	33.9	35.5	13.1	12.9	4.8	3.9
No. of cases	277	290	107	105	39	32
Min $\hat{\alpha}$	3.3	3	2.9	3.4	1.8 (2.7)	1.8 (3.4)
Max $\hat{\alpha}$	74.7 (33.9)	74.7 (35.5)	16.6 (13.1)	16.6 (12.9)	9.4 (2.1)	8 (0.5)

Table 5 Failure rate, number of cases where the test does not control the Type I error rate and the minimum and maximum estimated significance level for the A4, A5, A6 and BA tests

	$A4_5$	$A4_{20}$	$A4_m$	$A5_{2.5}$	$A5_{10}$	A6	BA
Failure rate	0.0	1.6	0.0	3.3	3.6	3.6	3.6
No. of cases	0	13	0	27	29	29	29
Min $\hat{\alpha}$	2.5	2.1 (1.6)	2.5	1.8 (3.3)	1.8 (3.4)	1.7 (3.6)	1.8 (3.6)
Max $\hat{\alpha}$	7.4	7.4	7.4	7.4	7.7 (0.1)	7.4	7.4

Table 6 Failure rates, minimum and maximum estimated significance level for the A1, A2 and A3 tests according to some characteristics of the parent populations

	$A1_5$	$A1_{20}$	$A2_5$	$A2_{20}$	$A3_5$	$A3_{20}$
Normal distrib.						
Failure rate	2.0	11.8	2.0	0.0	2.0	0.0
Min $\hat{\alpha}$	3.7	3.6	3.6	4.1	4	4.2
Max $\hat{\alpha}$	9.4 (2)	10.9 (11.8)	9.4 (2)	7.3	8.1 (2)	6.1
Homoscedasticity						
Failure rate	2.1	2.1	0.0	0.0	6.9	9.0
Min $\hat{\alpha}$	4.3	3.8	3.1	3.4	2.3 (6.9)	1.9 (9)
Max $\hat{\alpha}$	9.2 (2.1)	9.2 (2.1)	6.2	6.4	5.9	6.2
Low heteros						
Failure rate	22.3	20.8	3.1	2.5	3.3	3.3
Min $\hat{\alpha}$	3.5	3	2.9	3.4	1.8 (2.7)	1.8 (3.3)
Max $\hat{\alpha}$	32.4 (22.3)	32.4 (20.8)	9.7 (3.1)	9.4 (2.5)	8.1 (0.7)	6.5
High heteros.						
Failure rate	50.4	59.8	36.2	36.2	3.6	1.3
Min $\hat{\alpha}$	3.3	3	3.5	3.7	3.2	2.9
Max $\hat{\alpha}$	74.7 (50.4)	74.7 (59.8)	16.6 (36.2)	16.6 (36.2)	9.4 (3.6)	8 (1.3)
Low kurtosis						
Failure rate	28.9	30.8	12.9	12.0	2.0	0.8
Min $\hat{\alpha}$	3.6	3.1	3.2	3.7	2.9	2.3 (0.6)
Max $\hat{\alpha}$	61.4 (28.9)	61.4 (30.8)	13.5 (12.9)	13.4 (12)	8.7 (2)	7.7 (0.3)
High kurtosis						
Failure rate	41.2	42.5	14.4	14.6	7.8	6.5
Min $\hat{\alpha}$	3.3	3	2.9	3.4	1.8 (4.8)	1.8 (5.7)
Max $\hat{\alpha}$	74.7 (41.2)	74.7 (42.5)	16.6 (14.4)	16.6 (14.6)	9.4 (3.1)	8 (0.9)
Symmetric						
Failure rate	8.4	12.6	2.0	0.8	0.6	0.0
Min $\hat{\alpha}$	3.3	3	2.9	3.4	3	2.8
Max $\hat{\alpha}$	11.4 (8.4)	15.4 (12.6)	9.7 (2)	8.1 (0.8)	8.1 (0.6)	6.2
1 Asymmetric						
Failure rate	90.2	88.2	13.7	15.7	3.9	0.0
Min $\hat{\alpha}$	6.4	6.6	4.5	4.5	4.1	3.8
Max $\hat{\alpha}$	46.2 (90.2)	46.2 (88.2)	10.6 (13.7)	11.5 (15.7)	9 (3.9)	7.4
2 Asymmetric						
Failure rate	32.4	34.3	12.7	12.7	8.8	3.9
Min $\hat{\alpha}$	4	3.6	4	4.6	4	3.4
Max $\hat{\alpha}$	45.7 (32.4)	45.8 (34.3)	12.2 (12.7)	12.1 (12.7)	9.4 (8.8)	8 (3.9)
3 Asymmetric						
Failure rate	54.9	53.9	26.1	26.5	8.5	9.2
Min $\hat{\alpha}$	3.7	3.2	3.1	3.4	1.8 (7.2)	1.8 (9.2)
Max $\hat{\alpha}$	74.7 (54.9)	74.7 (53.9)	16.6 (26.1)	16.6 (26.5)	7.9 (1.3)	7.2

Table 7 Failure rates, minimum and maximum estimated significance level for the A4, A5, A6 and BA tests according to some characteristics of the parent populations

	$A4_5$	$A4_{20}$	$A4_m$	$A5_{2.5}$	$A5_{10}$	A6	BA
Normal distributions							
Failure rate	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min $\hat{\alpha}$	3.9	3.9	4	3.9	3.9	4	4
Max $\hat{\alpha}$	5.8	5.8	5.8	5.7	5.7	5.8	5.8
Homoscedasticity							
Failure rate	0.0	2.8	0.0	8.3	9.0	9.0	9.0
Min $\hat{\alpha}$	2.5	2.1 (2.8)	2.5	1.8 (8.3)	1.8 (9)	1.8 (9)	1.8 (9)
Max $\hat{\alpha}$	5.5	5.6	5.5	6.1	5.9	6.1	6.2
Low heteros.							
Failure rate	0.0	2.0	0.0	3.3	3.3	3.6	3.6
Min $\hat{\alpha}$	2.5	2.2 (2)	2.6	1.9 (3.3)	1.8 (3.3)	1.7 (3.6)	1.8 (3.6)
Max $\hat{\alpha}$	6	5.8	6	6	6	6	6
High heteros.							
Failure rate	0.0	0.0	0.0	0.0	0.4	0.0	0.0
Min $\hat{\alpha}$	3.1	3	3.1	3	3	3	3
Max $\hat{\alpha}$	7.4	7.4	7.4	7.4	7.7 (0.4)	7.3	7.4
Low kurtosis							
Failure rate	0.0	0.0	0.0	0.6	0.6	0.6	0.6
Min $\hat{\alpha}$	2.9	2.6	3.1	2.2 (0.6)	2.3 (0.6)	2.2 (0.6)	2.2 (0.6)
Max $\hat{\alpha}$	7.4	7.4	7.4	7.2	6.7	7.4	7.4
High kurtosis							
Failure rate	0.0	2.8	0.0	5.4	5.9	5.9	5.9
Min $\hat{\alpha}$	2.5	2.1 (2.8)	2.5	1.8 (5.4)	1.8 (5.7)	1.7 (5.9)	1.8 (5.9)
Max $\hat{\alpha}$	7.4	7.4	7.4	7.4	7.7 (0.2)	7.4	7.4
Symmetric							
Failure rate	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min $\hat{\alpha}$	3.1	3	3.1	2.8	2.9	2.8	2.8
Max $\hat{\alpha}$	5.8	5.8	5.8	6.1	5.9	6.1	6.2
1 Asymmetric							
Failure rate	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min $\hat{\alpha}$	4.1	4	4.1	3.6	3.6	3.7	3.7
Max $\hat{\alpha}$	6.5	6.5	6.6	6.4	6.5	6.5	6.5
2 Asymmetric							
Failure rate	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Min $\hat{\alpha}$	3.7	3.6	3.7	3.1	3.1	3.2	3.2
Max $\hat{\alpha}$	7.4	7.4	7.4	7.2	6.7	7.4	7.4
3 Asymmetric							
Failure rate	0.0	4.2	0.0	8.8	9.5	9.5	9.5
Min $\hat{\alpha}$	2.5	2.1 (4.2)	2.5	1.8 (8.8)	1.8 (9.2)	1.7 (9.5)	1.8 (9.5)
Max $\hat{\alpha}$	7.4	7.1	7.4	7.4	7.7 (0.3)	7	7.1

Table 8 Results according to the size of samples for the A1, A2 and A3 tests

		$A1_5$	$A1_{20}$	$A2_5$	$A2_{20}$	$A3_5$	$A3_{20}$
Equal sample sizes	Failure rate	25.4	28.9	10.2	10.2	4.3	4.7
	Min $\hat{\alpha}$	4.2	4.3	3.1	3.4	2.1 (2.7)	1.9 (4.7)
	Max $\hat{\alpha}$	61.2 (25.4)	61.2 (28.9)	13.8 (10.2)	15.2 (10.2)	9 (1.6)	7.4
Uneq. sample sizes	Failure rate	37.9	38.6	14.5	14.1	5.0	3.6
	Min $\hat{\alpha}$	3.3	3	2.9	3.4	1.8 (2.7)	1.8 (2.9)
	Max $\hat{\alpha}$	74.7 (37.9)	74.7 (38.6)	16.6 (14.5)	16.6 (14.1)	9.4 (2.3)	8 (0.7)
Small samples	Failure rate	32.7	34.1	14.4	14.1	6.1	5.0
	Min $\hat{\alpha}$	3.3	3	2.9	3.4	1.8 (3.4)	1.8 (4.4)
	Max $\hat{\alpha}$	52.7 (32.7)	52.7 (34.1)	16.6 (14.4)	16.6 (14.1)	9.4 (2.7)	8 (0.6)
Large samples	Failure rate	38.6	40.9	8.5	8.5	0.0	0.0
	Min $\hat{\alpha}$	3.7	3.7	3.7	3.9	2.7	2.7
	Max $\hat{\alpha}$	74.7 (38.6)	74.7 (40.9)	12.1 (8.5)	12.2 (8.5)	6.6	6.5
Equal	Failure rate	29.7	34.4	9.4	9.4	0.0	0.0
	Min $\hat{\alpha}$	4.2	4.3	3.9	3.9	2.7	2.7
	Max $\hat{\alpha}$	61.2 (29.7)	61.2 (34.4)	12.1 (9.4)	12.2 (9.4)	6.6	6.5
Unequal	Failure rate	43.8	44.6	8.0	8.0	0.0	0.0
	Min $\hat{\alpha}$	3.7	3.7	3.7	4.2	3	3
	Max $\hat{\alpha}$	74.7 (43.8)	74.7 (44.6)	10.7 (8)	10.7 (8)	6.5	6.4
Positive pairing	Failure rate	26.7	25.8	11.7	12.1	3.3	3.3
	Min $\hat{\alpha}$	3.3	3	2.9	3.4	1.8 (3.3)	1.8 (3.3)
	Max $\hat{\alpha}$	68.1 (26.7)	68.1 (25.8)	11.7 (11.7)	14.4 (12.1)	7.2	6.4
Negative pairing	Failure rate	59.2	61.7	22.1	20.8	6.3	2.5
	Min $\hat{\alpha}$	5.2	4.9	3.3	3.7	2.1 (0.8)	2 (0.8)
	Max $\hat{\alpha}$	74.7 (59.2)	74.7 (61.7)	16.6 (22.1)	16.6 (20.8)	9.4 (5.4)	8 (1.7)

Table 9 Failure rates, minimum and maximum estimated significance level for the A1, A2 and A3 tests for combinations of sample sizes and variances

			$A1_5$	$A1_{20}$	$A2_5$	$A2_{20}$	$A3_5$	$A3_{20}$
Equal sample size	Homos.	Failure rate	6.3	6.3	0.0	0.0	7.8	10.9
		Min $\hat{\alpha}$	4.4	4.3	3.1	3.4	2.3 (2.1)	1.9 (2.9)
		Max $\hat{\alpha}$	9.8 (1.7)	9.7 (1.7)	6.2	6.3	5.7	5.1
	Low heteros.	Failure rate	21.9	21.9	0.0	0.0	1.6	3.9
		Min $\hat{\alpha}$	4.2	4.3	3.3	3.6	2.1 (1.6)	2 (3.9)
		Max $\hat{\alpha}$	20.9 (21.9)	20.9 (21.9)	7	7.2	6.2	5.7
	High heteros.	Failure rate	51.6	65.6	40.6	40.6	6.3	0.0
		Min $\hat{\alpha}$	5.4	6.1	3.9	3.8	3.4	3.3
		Max $\hat{\alpha}$	61.2 (51.6)	61.2 (65.6)	13.8 (40.6)	15.2 (40.6)	9 (6.3)	7.4
Positive pairing	Low heteros.	Failure rate	18.1	16.9	0.0	0.0	5.0	5.0
		Min $\hat{\alpha}$	3.5	3	2.9	3.4	1.8 (5)	1.8 (5)
		Max $\hat{\alpha}$	24.1 (18.1)	24.1 (16.9)	6.3	6.3	5.6	5.9
	High heteros.	Failure rate	43.8	43.8	35.0	36.3	0.0	0.0
		Min $\hat{\alpha}$	3.3	3	3.5	3.7	3.2	2.9
		Max $\hat{\alpha}$	68.1 (43.8)	68.1 (43.8)	11.7 (35)	14.4 (36.3)	7.2	6.4
Negative pairing	Low heteros.	Failure rate	46.9	43.1	9.4	7.5	3.8	1.3
		Min $\hat{\alpha}$	5.2	4.9	3.3	3.7	2.1 (1.3)	2 (1.3)
		Max $\hat{\alpha}$	32.4 (46.9)	32.4 (43.1)	10.2 (9.4)	10 (7.5)	8.1 (2.5)	7
	High heteros.	Failure rate	83.8	98.8	47.5	47.5	11.3	5.0
		Min $\hat{\alpha}$	5.8	7.1	3.8	3.8	3.4	3.3
		Max $\hat{\alpha}$	74.7 (83.8)	74.7 (98.8)	16.6 (47.5)	16.6 (47.5)	9.4 (11.3)	8 (5)

Table 10 Results according to sample sizes for the A4, A5, A6 and BA tests

		$A4_5$	$A4_{20}$	$A4_m$	$A5_{2.5}$	$A5_{10}$	**A6**	**BA**
Equal sample sizes	Failure rate	0.0	2.3	0.0	5.1	5.1	5.1	5.1
	Min $\hat{\alpha}$	2.5	2.1 (2.3)	2.5	1.8 (5.1)	1.8 (5.1)	1.8 (5.1)	1.8 (5.1)
	Max $\hat{\alpha}$	6.7	6.5	6.8	6.4	6.5	6.5	6.5
Uneq. sample sizes	Failure rate	0.0	1.3	0.0	2.5	2.9	2.9	2.9
	Min $\hat{\alpha}$	2.6	2.2 (1.3)	2.6	1.9 (2.5)	1.9 (2.7)	1.7 (2.9)	1.8 (2.9)
	Max $\hat{\alpha}$	7.4	7.4	7.4	7.4	7.7 (0.2)	7.4	7.4
Small samples	Failure rate	0.0	2.0	0.0	4.2	4.5	4.5	4.5
	Min $\hat{\alpha}$	2.5	2.1 (2)	2.5	1.8 (4.2)	1.8 (4.4)	1.7 (4.5)	1.8 (4.5)
	Max $\hat{\alpha}$	7.4	7.4	7.4	7.4	7.7 (0.2)	7.4	7.4
Large samples	Failure rate	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Min $\hat{\alpha}$	3.3	2.8	3.3	2.8	2.8	2.7	2.7
	Max $\hat{\alpha}$	6.4	6.4	6.4	6.3	6.4	6.4	6.4
Positive pairing	Failure rate	0.0	1.7	0.0	2.9	2.9	3.3	3.3
	Min $\hat{\alpha}$	2.6	2.2 (1.7)	2.6	1.9 (2.9)	1.9 (2.9)	1.7 (3.3)	1.8 (3.3)
	Max $\hat{\alpha}$	6.1	5.8	6.1	5.8	5.7	5.9	5.9
Negative pairing	Failure rate	0.0	0.4	0.0	0.8	1.3	0.8	0.8
	Min $\hat{\alpha}$	3	2.4 (0.4)	3	2 (0.8)	1.9 (0.8)	2 (0.8)	2 (0.8)
	Max $\hat{\alpha}$	7.4	7.4	7.4	7.4	7.7 (0.4)	7.4	7.4

Table 11 Failure rates, minimum and maximum estimated significance level for the A4, A5, A6 and BA tests for combinations of sample sizes and variances

	Heterosc.		$A4_5$	$A4_{20}$	$A4_m$	$A5_{2.5}$	$A5_{10}$	A6	BA
Equal sample size	Homos.	Failure rate	0.0	3.1	0.0	10.9	10.9	10.9	10.9
		Min $\hat{\alpha}$	2.5	2.1 (0.8)	2.5	1.8 (2.9)	1.8 (2.9)	1.8 (2.9)	1.8 (2.9)
		Max $\hat{\alpha}$	5.4	5.2	5.4	5.1	5.1	5	5
	Low heteros.	Failure rate	0.0	3.1	0.0	4.7	4.7	4.7	4.7
		Min $\hat{\alpha}$	2.5	2.2 (3.1)	2.7	1.9 (4.7)	1.8 (4.7)	1.9 (4.7)	1.9 (4.7)
		Max $\hat{\alpha}$	6	5.9	6	5.7	5.7	5.7	5.7
	High heteros.	Failure rate	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Min $\hat{\alpha}$	3.2	3.1	3.2	3.1	3.1	3.1	3.1
		Max $\hat{\alpha}$	6.7	6.5	6.8	6.4	6.5	6.5	6.5
Positive pairing	Low heteros.	Failure rate	0.0	2.5	0.0	4.4	4.4	5.0	5.0
		Min $\hat{\alpha}$	2.6	2.2 (2.5)	2.6	1.9 (4.4)	1.9 (4.4)	1.7 (5)	1.8 (5)
		Max $\hat{\alpha}$	5.7	5.7	5.7	5.8	5.7	5.9	5.9
	High heteros.	Failure rate	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Min $\hat{\alpha}$	3.1	3	3.1	3.1	3.1	3	3
		Max $\hat{\alpha}$	6.1	5.8	6.1	5.8	5.7	5.8	5.8
Negative pairing	Low heteros.	Failure rate	0.0	0.6	0.0	1.3	1.3	1.3	1.3
		Min $\hat{\alpha}$	3	2.4 (0.6)	3	2 (1.3)	1.9 (1.3)	2 (1.3)	2 (1.3)
		Max $\hat{\alpha}$	6.2	6.3	6.2	6.4	6.1	6.5	6.5
	High heteros.	Failure rate	0.0	0.0	0.0	0.0	1.3	0.0	0.0
		Min $\hat{\alpha}$	3.1	3	3.1	3	3	3	3
		Max $\hat{\alpha}$	7.4	7.4	7.4	7.4	7.7 (1.3)	7.4	7.4

Table 12 Results for the A1, A2 and A3 tests when distributions are symmetric

		$A1_5$	$A1_{20}$	$A2_5$	$A2_{20}$	$A3_5$	$A3_{20}$
Homosced.	Failure rate	0.0	0.0	0.0	0.0	0.0	0.0
	Min $\hat{\alpha}$	4.3	4	4.1	3.8	3.6	3.3
	Max $\hat{\alpha}$	6.2	5.9	5.8	6.3	5.9	6.2
Low heteros.	Failure rate	4.1	2.0	3.6	1.5	1.0	0.0
	Min $\hat{\alpha}$	3.5	3	2.9	3.4	3	2.8
	Max $\hat{\alpha}$	9.8 (4.1)	9.2 (2)	9.7 (3.6)	8.1 (1.5)	8.1 (1)	6.1
High heteros.	Failure rate	22.4	41.8	0.0	0.0	0.0	0.0
	Min $\hat{\alpha}$	3.3	3	3.5	3.7	3.2	2.9
	Max $\hat{\alpha}$	11.4 (22.4)	15.4 (41.8)	6	6	5.7	5.7
Low kurtosis	Failure rate	7.8	13.1	2.6	1.3	1.3	0.0
	Min $\hat{\alpha}$	3.6	3.1	3.2	3.7	3.3	3.3
	Max $\hat{\alpha}$	11.4 (7.8)	15.4 (13.1)	9.7 (2.6)	8.1 (1.3)	8.1 (1.3)	6.2
High kurtosis	Failure rate	8.8	12.3	1.5	0.5	0.0	0.0
	Min $\hat{\alpha}$	3.3	3	2.9	3.4	3	2.8
	Max $\hat{\alpha}$	10.6 (8.8)	13.2 (12.3)	9.3 (1.5)	7.9 (0.5)	7.2	5.9
Equal sample size	Failure rate	0.0	6.3	0.0	0.0	0.0	0.0
	Min $\hat{\alpha}$	4.2	4.3	3.9	3.8	3.4	3.2
	Max $\hat{\alpha}$	7.4	7.8 (6.3)	6	6.1	5.7	5.7
Equal sample size and low heteros.	Failure rate	0.0	0.0	0.0	0.0	0.0	0.0
	Min $\hat{\alpha}$	4.2	4.3	4	3.9	3.4	3.2
	Max $\hat{\alpha}$	6	6	6	6.1	5.7	5.6
Unequal sample size	Failure rate	12.2	15.5	2.9	1.2	0.8	0.0
	Min $\hat{\alpha}$	3.3	3	2.9	3.4	3	2.8
	Max $\hat{\alpha}$	11.4 (12.2)	15.4 (15.5)	9.7 (2.9)	8.1 (1.2)	8.1 (0.8)	6.2

Table 12 continued

		$A1_5$	$A1_{20}$	$A2_5$	$A2_{20}$	$A3_5$	$A3_{20}$
Small	Failure rate	12.2	15.8	3.6	1.5	1.0	0.0
	Min $\hat{\alpha}$	3.3	3	2.9	3.4	3	2.8
	Max $\hat{\alpha}$	11.4 (12.2)	15.4 (15.8)	9.7 (3.6)	8.1 (1.5)	8.1 (1)	6.2
Large	Failure rate	12.2	14.3	0.0	0.0	0.0	0.0
	Min $\hat{\alpha}$	3.7	3.7	3.9	4.2	3.9	3.9
	Max $\hat{\alpha}$	11 (12.2)	11.1 (14.3)	6	5.6	5.5	5.6
Small samples	Failure rate	3.8	5.5	1.1	0.5	0.3	0.0
	Min $\hat{\alpha}$	3.3	3	2.9	3.4	3	2.8
	Max $\hat{\alpha}$	11.4 (3.8)	15.4 (5.5)	9.7 (1.1)	8.1 (0.5)	8.1 (0.3)	6.2
Large samples	Failure rate	3.4	5.7	0.0	0.0	0.0	0.0
	Min $\hat{\alpha}$	3.7	3.7	3.9	3.9	3.7	3.7
	Max $\hat{\alpha}$	11 (3.4)	11.1 (5.7)	6	5.7	5.7	5.7
Positive pairing	Failure rate	0.0	0.0	0.0	0.0	0.0	0.0
	Min $\hat{\alpha}$	3.3	3	2.9	3.4	3	2.8
	Max $\hat{\alpha}$	6.2	6	6	6	5.7	5.9
Negative pairing	Failure rate	28.6	36.2	6.7	2.9	1.9	0.0
	Min $\hat{\alpha}$	5.2	4.9	3.8	3.8	3.4	3.1
	Max $\hat{\alpha}$	11.4 (28.6)	15.4 (36.2)	9.7 (6.7)	8.1 (2.9)	8.1 (1.9)	6.1
Negative pairing and low heteros.	Failure rate	11.4	5.7	10.0	4.3	2.9	0.0
	Min $\hat{\alpha}$	5.2	4.9	4.8	4.5	3.5	3.1
	Max $\hat{\alpha}$	9.8 (11.4)	9.2 (5.7)	9.7 (10)	8.1 (4.3)	8.1 (2.9)	6.1
Negative pairing and high heteros.	Failure rate	62.9	97.1	0.0	0.0	0.0	0.0
	Min $\hat{\alpha}$	5.8	7.1	3.8	3.8	3.4	3.3
	Max $\hat{\alpha}$	11.4 (62.9)	15.4 (97.1)	5.9	5.9	5.5	5.7

homoscedastic. On the other hand, the A2 test seems to control the estimated significance level when distributions are homoscedastic. The $A3_{20}$ test seems to be robust if distributions are normal or symmetric or if there is one asymmetric distribution. The rest of tests may control the Type I error rate in many more situations (Table 7). The A5, A6 and BA tests show a null failure rate if distributions are normal or symmetric or some of them symmetric or highly heteroscedastic.

In Tables 8, 9, 10 and 11 we classify the simulation results according to sample sizes. If sample sizes are large, we find that all tests except A1 and A2 control the Type I error rate (see Tables 8 and 10). In small samples only the A4 test seems to be robust. The A2 test seems to control the estimated significance level when sample sizes are equal if there is homoscedasticity or low heteroscedasticity and when there is positive pairing if there is low heteroscedasticity. If there is high heteroscedasticity, the A3 test shows control of the Type I error rate if sample sizes are equal ($A3_{20}$ test) or there is positive pairing ($A3_5$ and $A3_{20}$ tests). Equal or unequal sample sizes or the type of pairing between sample sizes and variances do not affect the robustness of the A4, A5, A6 and BA tests.

If we focus only on symmetric distributions we find that the $A3_{20}$, A4, A5, A6 and BA tests were robust (Tables 6 and 7). However, for some specific situations we also find some more robust tests (Table 12). In particular, all tests studied may be robust at the significant level if, additionally to symmetric distributions, they are homoscedastic, or sample sizes are equal or there is positive pairing.

6 Conclusions

We have compared the estimated significance level of twelve adaptive tests, including the classical ones used by applied researchers, and the bootstrap ANOVA in an ample variety of settings. Our simulation results reveal that increasing the nominal significance level in preliminary testing does not seem, in general, to improve the performance of an adaptive test.

The A1 and A2 tests, commonly used by applied researchers, show bad behaviours, especially the A1 test, that may have an estimated significance level of up to 74.7 %, when the nominal level is 5 %. Even with large sample sizes its performance does not improve. The A2 test is a better strategy than the A1 test and its performance improves in large samples.

The unconditional application of the bootstrap ANOVA test seems to perform better than the adaptive tests proposed in this paper, except for the A4 test. So, in some cases the accuracy of the final inference can be improved by using preliminary tests. The A4 test (in particular, $A4_5$ and $A4_m$) controls the Type I error rate in all the situations simulated, so it can be recommended for general use. The A4 test bases its procedure on improving the detection of heteroscedasticity by pre-testing the normality in the first stage in order to choose the appropriate homoscedasticity test in the second stage. As a result, in the third stage a better selection of the test of mean equality is made between the Brown–Forsythe and the bootstrap ANOVA tests. The decision on normality is not used in the third stage. When we focus on normal or symmetric distributions,

more tests are robust at the significance level, like the $A3_{20}$, A4, A5, A6 and BA tests.

References

Anderson TW, Darling DA (1954) A test of goodness of fit. J Am Stat Assoc 49:765–769

Albers W, Boon PC, Kallenberg WCM (2000) The asymptotic behavior of tests for normal means based on a variance pre-test. J Stat Plan Inference 88:47–57

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

Bancroft TA (1964) Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance. Biometrics 20:427–442

Bartlett MS (1937) Properties of sufficiency and statistical tests. Proc R Soc Ser A A160:268–282

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

Cabilio P, Masaro J (1996) A simple test of symmetry about an unknown median. Can J Stat 24:349–361

Cahoy DO (2010) A bootstrap test for equality of variances. Comput Stat Data Anal 54:2306–2316

Charway H, Bailer AJ (2007) Testing multiple-group variance equality with randomization procedures. J Stat Comput Simul 77:797–803

Cribbie RA, Fiksenbaum L, Keselman HJ, Wilcox RR (2012) Effect of non-normality on test statistics for one-way independent group designs. Br J Math Stat Psychol 65:56–73

D'Agostino RB, Stephens MA (1986) Goodness-of-fit techniques. Marcel Dekker, New York

D'Agostino RB, Belanger A, D'Agostino RB Jr (1990) A suggestion for using powerful and informative tests of normality. Am Stat 44:316–321

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York

Farrel PJ, Rogers-Stewart K (2006) Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test. J Stat Comput Simul 76:803–816

Freidlin B, Miao W, Gastwirth JL (2003) On the Use of the Shapiro–Wilk test in two-stage adaptive inference for paired data from moderate to very heavy tailed distributions. Biom J 45:887–900

Gastwirth JL, Gel YR, Miao W (2009) The impact of Levene's test of equality of variances on statistical theory and practice. Stat Sci 24:343–360

Gnandesikan R (1977) Methods for statistical analysis of multivariate observations. Wiley, New York

Hall P, Padmanabhan AR (1997) Adaptive inference for the two-sample scale problem. Technometrics 39:412–422

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

Hoaglin DC (1985) Summarizing shape numerically: The g- and h-distributions. In: Hoaglin, Mosteller, Tukey (eds) Exploring data tables, trends, and shapes. Wiley, New York, pp 461–513

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

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

Keskin S (2006) Comparison of several univariate normality tests regarding type I error rate and power of the test in simulation based small samples. J Appl Sci Res 2:296–300

Keyes TK, Levy MS (1997) Analysis of Levene's test under design imbalance. J Educ Behav Stat 22:227–236

Kolmogorov A (1933) Sulla Determinazione Empirica di una Legge di Distribuzione. Giornalle dell'Instituto Italiano degli Attuari 4:83–91

Levene H (1960) Robust tests for equality of variances. In: Olkin I et al (eds) Contributions to probability and statistics: essays in honor of harold hotelling. Stanford University Press, Palo Alto, pp 278–292

Lilliefors HW (1967) On the Kolmogorov–Smirnov test for normality with mean and variance unknown. J Am Stat Assoc 62:399–402

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
Mendes M, Pala A (2003) Type I error rate and power of three normality tests. Pak J Inf Technol 2:135–139
Miao W, Gastwirth J (2009) A new two stage adaptive nonparametric test for paired differences. Stat Interface 2:213–221
Neuhäuser M (2007) A comparative study of nonparametric two-sample tests after Levene's transformation. J Stat Comput Simul 77:517–526
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'Gorman T (1997) A comparison of an adaptive two-sample test to the t-test and the rank sum test. Commun Stat Simul Comput 26:1393–1411
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
Parra-Frutos I (2013) Testing homogeneity of variances with unequal sample sizes. Comput Stat 28:1269–1297
Parra-Frutos I (2014) Controlling the Type I error rate by using the nonparametric bootstrap when comparing means. Br J Math Stat Psychol 67:117–132
Paull AE (1950) On a preliminary test for pooling mean squares in the analysis of variance. Ann Math Stat 21:539–556
Rasch D, Kubinger KD, Moder K (2011) The two-sample t test: pre-testing its assumptions does not pay off. Stat Pap 52:219–231
Razali NM, Wah YB (2010) Power comparisons of some selected normality tests. In: Proceedings of the regional conference on statistical sciences (RCSS'10), pp 126–138
Rochon J, Kieser M (2011) A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample t-test. Br J Math Stat Psychol 64:410–426
Schoder V, Himmelmann A, Wilhelm KP (2006) Preliminary testing for normality: some statistical aspects of a common concept. Clin Exp Dermatol 31:757–761
Schucany WR, Ng HKT (2006) Preliminary goodness-of-fit tests for normality do not validate the one-sample Student t. Commun Stat Theory Methods 35:2275–2286
Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52:591–611
Shapiro SS, Wilk MB, Chen HJ (1968) A comparative study of various tests for normality. J Am Stat Assoc 63:1343–1372
Smirnov NV (1939) On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bull Mosc Univ 2:3–16 (**Russian**)
Welch BL (1951) On the comparison of several mean values: an alternative approach. Biometrika 38:330–336
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
Zimmerman DW (2004) A note on preliminary tests of equality of variances. Br J Math Stat Psychol 57:173–181