Thailand Statistician
July 2018; 16(2): 106-128
http://statassoc.or.th
Contributed paper

Modifications of Levene's and O'Brien's Tests for Testing the Homogeneity of Variance Based on Median and Trimmed Mean

Kotchaporn Soikliew* [a] and Autcha Araveeporn [a]
[a] Department of Applied Statistics, Faculty of Science, King Monkgut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand.
*Corresponding author; e-mail: kotchaporn.fern@gmail.com

Received: 24 March 2017
Accepted: 11 December 2017

Abstract

The propose of this paper is to compare the efficiency of Levene's test with absolute deviation with the alternative tests consists of Brown-Forsythe test, modified Levene's test based on trimmed mean, three Levene's tests with squared deviation and three O'Brien's tests. In total, nine homogeneity of variance tests used in ANOVA were compared on probability of type I error (also robustness) and power of the test (also power). These tests are analyzed through four different shaped distributions contain normal distribution, Uniform distribution, Logistic distribution, and gamma distribution. The results from these studies showed that Levene's test with absolute deviation is neither the best nor worst in terms of robustness and power. It is the best choice for logistic distribution. There are better tests that could be used, and some were more preferable depending on the distributional shape of data. For normal distribution, modified Levene's test with squared deviation based on median is the best choice for equal sample size. For uniform distribution, modified O'Brien's test based on median is the best choice for small sample size both equal and unequal. For gamma distribution, Levene's test with absolute deviation based on trimmed mean is the best choice for both small and medium sample sizes whereas modified Levene's test with squared deviation based on median is the best choice same as Levene's test with absolute deviation based on trimmed mean and Levene's test with absolute deviation for large sample size.

Keywords: Homogeneity of variance, Levene's test, O'Brien's test, robustness, power of the test.

1. Introduction

The analysis of variance (ANOVA) is one of the most important and useful techniques for comparing different groups or treatments with respect to their means where there are more than two groups. A set of assumptions include normal distribution, homogeneity of variance and independence of observations. These assumptions are checked before analysis of data. It is now well established that the violation of the assumption of homogeneity of variance can have severe effects on the inference of the population mean, especially in the case of unequal sample sizes and non-normally distributed. These are explained in research papers such as Mendes (Mendes and Pala 2004).

In fact, the conventional ANOVA based on F test provides generally poor control over both type I and type II error rates under a wide variety of variance heterogeneity conditions (Bishop and Dudewicz 1978). Misleading assumption would impair the utility of the test, leading to wrong and invalid conclusions. The problem of homogeneity of variance has to be settled before performing an ANOVA (Cochran 1941). It is well known that the generally statistical packages use Levene's test to check the homogeneity of variance for conducting tests of the equality of means in ANOVA (Oladejo and Adetunde 2009). Unfortunately, Levene's test for comparing variance is extremely sensitive to the assumption of normality (Sharma and Kibria 2013). However, there are many others test statistics for checking homogeneity of variance under different situations O'Brien (1981), and the modification tests Brown and Forsythe (1974), Carroll and Schneider (1985), Tomarken and Serlin (1986), Gastwirth et al. (2010), that may have been a better choice than Levene's test.

Levene's test is the one-way analysis of variance F test on $\left|X_{ij}-\bar{X}_i\right|$, the absolute deviations of the X_{ij} from their group mean $\bar{X}_i$ various modifications of Levene's test have been proposed and investigated. Carroll and Schneider (1985) had described that Brown and Forsythe (1974) had considered the median and the 10% trimmed mean, which are more robust estimates of alternatives central tendency, as alternatives to mean in computing the absolute deviations. They found that the alternative formulations of Levene's tests are robust against non-normality (Gogoi and Gogoi 2015).

Lee et al. (2010) had compared Levene's test with modified Levene's test, O'Brien's test, Samiuddin Cobe Root test, Z-variance test, Overall-Woodward modified Z-variance test, and F-Max test. The results showed that Levene's test was neither the best nor worst in terms of robustness and power of a test, modified Levene's test showed very good robustness when compared with the other tests but lower power of a test than other tests.

Gorbunova and Lemeshko (2012) had compared F-test with Bartlett's test, O'Brien's test Neyman-Pearson test, Z-variance test, Overall-Woodward modified Z-variance test, Hartley test, Levene's test and modified Levene's test in term of power for sample were from any distribution. When the data distribution was mesokurtic or platykurtic, the best choice was O'Brien's test or Z-variance test, while the data distribution was leptokurtic or skewed, the modified Levene's test was chosen.

Parra (2012) had presented that Levene's test had higher power when distributions were leptokurtic and skewed. Power of Brown-Forsythe test was low for small sample sizes and decreased when coupled with unbalanced samples; on the other hand, statistical power increased with larger samples, both balanced and unbalanced. Although with normal or platykurtic distributions Levene had a lower power than those tests.

Wang et al. (2016) had suggested that O'Brien's, Brown-Forsythe, Bootstrap Brown-Forsythe, and Levene's with squared deviations test maintained adequate type I error control, performing better than the others across all the conditions.

Furthermore, Hanif et al. (2013) had presented that the mean was used for the homogeneous variance situations and the trimmed mean was used for the heterogeneous variance situations. They had suggested that using the 10% trimmed mean. The trimmed mean could be considered regardless distribution shapes and sample sizes. Likewise, Lu et al. (2014) had reported that the trimmed mean could be used efficiency for both non-normality (especially heavy tails) and heterogeneity of variance conditions. It is well known that the outliers have been a negative effect on arithmetic mean. The trimmed mean, one of the estimators was used alternative mean instead of arithmetic mean. It is not affected much from outliers and its computing was removing a proportion of the

largest and smallest observations then calculates the mean of remaining observations. Moreover, the median and mean were described in terms of trimming. The median was trimmed mean, trimmed at 50%. At the opposite extreme, the mean was a location measure with 0% of the scores trimmed prior to calculation (Othman et al. 2012).

The aim of this study is to compare the probability of type I error (also robustness) and power of Levene's tests with absolute and squared deviation, O'Brien's test and six modified test statistics based on median and trimmed mean for four distribution-types (normal distribution, uniform distribution, logistic distribution, and gamma distribution). These tests are simulated and analyzed by R program version 3.3.2. The straightforward modification of this idea is extracted from the modified Levene's test by Brown and Forsythe (1974). "Goodness" of each test is determined by examining the robustness and power of a test for each test. Should general statistic package consider other tests along with Levene's test?

2. Materials and Methods

In this section, the homogeneity of variance tests of k populations given sample X_{ij} $\left\{X_{ij} : i = 1,2,...,k, j = 1,2,...,n_i\right\}$ from the i^{th} population with mean μ_i, and variance σ_i^2. The null hypothesis is $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ and the alternative hypothesis is $H_0 : \sigma_i^2 \neq \sigma_j^2$ for at least one $i \neq j$,

where $\bar{X}_i = \dfrac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$ is the mean of the i^{th} group,

$S_i^2 = \dfrac{\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2}{n_i - 1}$ is the variance of the i^{th} group,

$N = \sum_{i=1}^{k} n_i$ is total sample size, and n_i is sample size of the i^{th} group.

2.1. Description of test statistics

Levene's test, Levene (1960) had proposed Levene's test is used to test if k samples have equal variances. Levene had proposed an alternative method to Bartlett's test for testing the assumption of homogeneity of variance for ANOVA designs. Bartlett's test works well for data that are normally or approximate normally distributed but it doesn't fare well for data that follow a leptokurtic or skewed distribution (Overall and Woodward 1974). According to Levene, that Levene's test is less sensitive to departures from normally. This says that Levene's test has fewer type I errors than Bartlett's test for non-normality distributions (Gastwirth et al. 2010).

Levene's test with absolute deviation is the one-way analysis of variance F-test based on $Z_{ij} = \left|X_{ij} - \bar{X}_i\right|$, where $\bar{X}_i$ is the mean of the i^{th} group, and it is denoted by L. The test statistic is

$$L = \frac{(N-k)\sum_{i=1}^{k} n_i (\bar{Z}_i - \bar{Z})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)^2}, \quad (1)$$

where $\bar{Z}_i = \dfrac{\sum_{j=1}^{n_i} Z_{ij}}{n_i}$ and $\bar{Z} = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}}{N}$.

The null hypothesis H_0 is rejected when $L > F_{(\alpha,k-1,N-k)}$ where $F_{(\alpha,k-1,N-k)}$ is the upper critical value of the F-distribution with $k-1$ and $N-k$ degrees of freedom at a significance level of α.

Levene's original paper only proposed using the mean. Brown and Forsythe (1974) extended Levene's test to use either the median or the trimmed mean in addition to the mean. Levene's test that using the trimmed mean performed the best when the underlying data followed heavy-tailed distribution and the median performed the best when the underlying data followed skewed distribution. Using the mean provided the best power for symmetric and moderate-tailed distributions.

Modified Levene's with absolute deviation based on median is usually called Brown-Forsythe's test. This the modified studies by Brown and Forsythe was referenced in Carroll and Schneider (1985). This test is the one-way analysis of variance F-test based on $Z_{Mij} = \left|X_{ij} - M_i\right|$, where M_i is median of the i^{th} group, which is used instead of the mean. The median version of Levene's test with absolute deviation is denoted by BF. The modified test statistic is

$$BF = \frac{(N-k)\sum_{i=1}^{k} n_i(\bar{Z}_{Mi} - \bar{Z}_M)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{Mij} - \bar{Z}_{Mi})^2}, \tag{2}$$

where $X_{i(j)} = X_{i(1)}, X_{i(2)}, ..., X_{i(n_j)}$ is sample order statistic of the i^{th} group,

$$M_i = X_{(n_i+1)/2}, \text{ for odd number of } n_i, \tag{3}$$

or $$M_i = \left(X_{n_i/2} + X_{(n_i/2)+1}\right)/2, \text{ for even number of } n_i, \tag{4}$$

where $\bar{Z}_{Mi} = \dfrac{\sum_{j=1}^{n_i} Z_{Mij}}{n_i}$, and $\bar{Z}_M = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{Mij}}{N}$.

The null hypothesis H_0 is rejected when $BF > F_{(\alpha,k-1,N-k)}$ where $F_{(\alpha,k-1,N-k)}$ is the upper critical value of the F-distribution with $k-1$ and $N-k$ degrees of freedom at a significance level of α.

Modified Levene's test based on trimmed mean is used 10% trimmed mean in addition to the mean. This the modified studies by Brown and Forsythe (1974). This test is the one-way analysis of variance F-test based on $Z_{Tij} = \left|X_{ij} - T_i\right|$, where T_i is 10% trimmed mean of the i^{th} group, which is used instead of the mean. The trimmed mean version of Levene's test with absolute deviation is denoted by $Ltrim$. The modified test statistic is

$$Ltrim = \frac{(N-k)\sum_{i=1}^{k} n_i(\bar{Z}_{Ti} - \bar{Z}_T)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{Tij} - \bar{Z}_{Ti})^2}, \tag{5}$$

where $X_{i(j)} = X_{i(1)}, X_{i(2)}, ..., X_{i(n_j)}$ is sample order statistic of the i^{th} group,

T_i is 10% trimmed mean of the i^{th} group,

where $\overline{Z}_{Ti} = \dfrac{\sum_{j=1}^{n_i} Z_{Tij}}{n_i}$ and $\overline{Z}_T = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{Tij}}{N}$.

According to Tukey (1948), the δ percent trimmed mean ($T_{\delta i}$) is defined as

$$T_{\delta i} = \frac{1}{n_i - 2\left\lceil n_i\left(\delta/100\right)\right\rceil}\sum_{i=g+1}^{n_i-g} X_{i(j)}, \tag{6}$$

where δ is percent of trimming the first and the last order observations and $g = \left\lceil n_i\left(\delta/100\right)\right\rceil$, $\left\lceil n_i\left(\delta/100\right)\right\rceil$ is the largest integer $\le n_i\left(\delta/100\right)$.

The null hypothesis H_0 is rejected when $Ltrim > F_{(\alpha,k-1,N-k)}$ where $F_{(\alpha,k-1,N-k)}$ is the upper critical value of the F-distribution with $k-1$ and $N-k$ degrees of freedom at a significance level of α.

Levene's test with squared deviation is the one-way analysis of variance F-test based on $Z_{ij}^* = (X_{ij} - \overline{X}_i)^2$ (Levene 1960), where $\overline{X}_i$ is the mean of the i^{th} group, and it is denoted by LS. The test statistic is

$$LS = \frac{(N-k)\sum_{i=1}^{k} n_i(\overline{Z}_i^* - \overline{Z}^*)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij}^* - \overline{Z}_i^*)^2}, \tag{7}$$

where $\overline{Z}_i^* = \dfrac{\sum_{j=1}^{n_i} Z_{ij}^*}{n_i}$ and $\overline{Z}^* = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}^*}{N}$.

The null hypothesis H_0 is rejected when $LS > F_{(\alpha,k-1,N-k)}$ where $F_{(\alpha,k-1,N-k)}$ is the upper critical value of the F-distribution with $k-1$ and $N-k$ degrees of freedom at a significance level of α.

Modified Levene's test with squared deviation based on median is the one-way analysis of variance F-test based on $Z_{Mij}^* = (X_{ij} - M_i)^2$, where M_i is median of the i^{th} group and M_i is given by (3) or (4), which is used instead of the mean. The median version of Levene's test with squared deviation is denoted by $LSmed$. The modified test statistic is

$$LSmed = \frac{(N-k)\sum_{i=1}^{k} n_i(\overline{Z}_{Mi}^* - \overline{Z}_M^*)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{Mij}^* - \overline{Z}_{Mi}^*)^2}, \tag{8}$$

where $\overline{Z}_{Mi}^* = \dfrac{\sum_{j=1}^{n_i} Z_{Mij}^*}{n_i}$ and $\overline{Z}_M^* = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{Mij}^*}{N}$.

The null hypothesis H_0 is rejected when $LSmed > F_{(\alpha,k-1,N-k)}$ where $F_{(\alpha,k-1,N-k)}$ is the upper critical value of the F-distribution with $k-1$ and $N-k$ degrees of freedom at a significance level of α.

Modified Levene's test with squared deviation based on trimmed mean is the one-way analysis of variance F-test based on $Z^*_{Tij} = (X_{ij} - T_i)^2$, where T_i is 10% trimmed mean of the i^{th} group and T_i is given by (6), which is used instead of the mean. The trimmed mean version of Levene's test with squared deviation is denoted by *LStrim*. The modified test statistic is

$$LStrim = \frac{(N-k)\sum_{i=1}^{k} n_i(\bar{Z}^*_{Ti} - \bar{Z}^*_T)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z^*_{Tij} - \bar{Z}^*_{Ti})^2}, \tag{9}$$

where $\bar{Z}^*_{Ti} = \frac{\sum_{j=1}^{n_i} Z^*_{Tij}}{n_i}$ and $\bar{Z}^*_T = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z^*_{Tij}}{N}$.

The null hypothesis H_0 is rejected when $LStrim > F_{(\alpha,k-1,N-k)}$ where $F_{(\alpha,k-1,N-k)}$ is the upper critical value of the F-distribution with $k-1$ and $N-k$ degree of freedom at a significance level of α.

O'Brien's test, O'Brien (1981) had claimed that this test is a general method and fairly well for behavioral science data. O'Brien had stated that the test is robust to data that departs from normality. It was also easy to program into statistical packages like SPSS, that was competitive with other tests in terms of power and it can be easily used in different ANOVA designs with equal or unequal sample sizes. O'Brien stated that not much research had been done on statistic.

The computational operations for this test are straightforward modification. Every raw score, X_{ij} in the study is transformed using formula which is defined by

$$V_{ij} = \frac{(n_i - 1.5)n_i(X_{ij} - \bar{X}_i)^2 - 0.5S_i^2(n_i - 1)}{(n_i - 1)(n_i - 2)}. \tag{10}$$

This test is the one-way analysis of variance F-test based on the V_{ij}, where $\bar{X}_i$ is the mean of the i^{th} group and it is denoted by *OB*. The test statistic is

$$OB = \frac{(N-k)\sum_{i=1}^{k} n_i(\bar{V}_i - \bar{V})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(V_{ij} - \bar{V}_i)^2}, \tag{11}$$

where $\bar{V}_i = \frac{\sum_{j=1}^{n_i} V_{ij}}{n_i}$ and $\bar{V} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} V_{ij}}{N}$.

After this transformation the mean of V values will be equal to the variance for original values, that is $\bar{V}_i = \frac{\sum_{j=1}^{n_i} V_{ij}}{n_i} = S_i^2$.

The null hypothesis H_0 is rejected when $OB > F_{(\alpha,k-1,N-k)}$ where $F_{(\alpha,k-1,N-k)}$ is the upper critical value of the F-distribution with $k-1$ and $N-k$ degrees of freedom at a significance level of α.

Modified O'Brien's test based on median is nearly identical to original O'Brien's test. The median is used instead of the mean. The differences are that median is used instead of the mean in computing V_{ij}.

The computational operations for this test are straightforward modification. Every raw score, X_{ij} in the study is transformed using formula which is defined by

$$V_{Mij} = \frac{(n_i - 1.5)n_i(X_{ij} - M_i)^2 - 0.5S_i^2(n_i - 1)}{(n_i - 1)(n_i - 2)}. \tag{12}$$

This test is the one-way analysis of variance F-test based on the V_{Mj}, where M_i is the median of the i^{th} group and it is denoted by $OBmed$. The modified test statistic is

$$OBmed = \frac{(N-k)\sum_{i=1}^{k} n_i(\bar{V}_{Mi} - \bar{V}_M)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(V_{Mij} - \bar{V}_{Mi})^2}, \tag{13}$$

where $X_{i(j)} = X_{i(1)}, X_{i(2)}, ..., X_{i(n_j)}$ is sample order statistic of the i^{th} group and M_i is given by (4) or (5) for odd or even number of n_i, respectively, where $\bar{V}_{Mi} = \frac{\sum_{j=1}^{n_i} V_{Mij}}{n_i}$ and $\bar{V}_M = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} V_{Mij}}{N}$.

The null hypothesis H_0 is rejected when $OBmed > F_{(\alpha,k-1,N-k)}$ where $F_{(\alpha,k-1,N-k)}$ is the upper critical value of the F-distribution with $k-1$ and $N-k$ degrees of freedom at a significance level of α.

Modified O'Brien's test based on trimmed mean is nearly identical to original O'Brien's test. The 10% trimmed mean is used instead of the mean. The differences are that 10% trimmed mean is used instead of the mean in computing V_{ij}.

The computational operations for this test are straightforward modification. Every raw score, X_{ij} in the study is transformed using formula which is defined by

$$V_{Tij} = \frac{(n_i - 1.5)n_i(X_{ij} - T_i)^2 - 0.5S_i^2(n_i - 1)}{(n_i - 1)(n_i - 2)}. \tag{14}$$

This test is the one-way analysis of variance F-test based on the V_{Tij}, where T_i is 10% trimmed mean of the i^{th} group and it is denoted by $OBtrim$. The modified test statistic is

$$OBtrim = \frac{(N-k)\sum_{i=1}^{k} n_i(\bar{V}_{Ti} - \bar{V}_T)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(V_{Tij} - \bar{V}_{Ti})^2}, \tag{15}$$

where $X_{i(j)} = X_{i(1)}, X_{i(2)}, ..., X_{i(n_j)}$ is sample order statistic of the i^{th} group, T_i is given by (6), where $\bar{V}_{Ti} = \dfrac{\sum_{j=1}^{n_i} V_{Tij}}{n_i}$ and $\bar{V}_M = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} V_{Tij}}{N}$.

The null hypothesis H_0 is rejected when $OBtrim > F_{(\alpha,k-1,N-k)}$ where $F_{(\alpha,k-1,N-k)}$ is the upper critical value of the F-distribution with $k-1$ and $N-k$ degrees of freedom at a significance level of α.

2.2. Research design

Let three population group cases contain equal sample sizes and unequal sample sizes. Configurations of sample size are determined under total sample sizes (45, 90, and 180), average sample sizes (15, 30, and 60) (Reiczigel 2003) and sample size patterns (equal and unequal). The configurations of sample size are given in Table 1.

Table 1 The three populations and sample size patterns

	Sample sizes	
	Equal	Unequal
Small	15:15:15	10:15:20
Medium	30:30:30	20:30:40
Large	60:60:60	40:60:80

Let the population distribution-types are four distributions contain (i) Normal distribution, (ii) Logistic distribution, (iii) Uniform distribution, and (iv) Gamma distribution.

Let the different population variance ratios by non-centrality parameter (ϕ), this ϕ is a measure of the different population variance (Game et al. 1972). The formula is defined by

$$\phi = \frac{\sqrt{\sum_{i=1}^{k}(\sigma_i^2 - \bar{\sigma}^2)^2 / k}}{\sigma_1^2}, \tag{16}$$

where σ_1^2 is the population variance with the lowest,

σ_i^2 is the population variance with the i^{th} group; $i = 1, 2, ..., k$,

$\bar{\sigma}^2$ is the mean of population variance with k groups.

The details are given in Table 2.

Table 2 The different population variance ratios by non-centrality parameter (ϕ)

Levels	Ratios	ϕ
Slightly ($0<\phi<1.5$)	1:2:4	1.25
Moderately ($1.5\leq\phi<3.0$)	1:4:8	2.87
Highly ($\phi\geq 3.0$)	1:8:16	6.13

To study the significance levels (α) and power of the tests, let $\alpha = 0.05$. The significance level must be chosen typically set to 0.05. That is usually chosen on the field of study (Craparo 2007). Let the criterion for determined robustness based on the Bradley limit (1978) for testing ability to control probability of type I error. The probability of type I error (α) has to fall between (0.025, 0.075) when the significance levels are 0.05.

2.3. Methodology

In this study, there are research methods as follows:

1. R program version 3.3.2 is written to carry out the analyses. The data are generated from normal distribution, uniform distribution, logistic distribution, and gamma distribution with equal variance. For all tests in this study, there are three groups are arranged as configurations of sample size. The details are given in Table 1.

2. Three group experimental data are simulated one thousand times and these experimental data are analyzed for each sample size. All homogeneity of variance tests are computed for each experimental data. The experimental data are counted the number of times the null hypothesis is rejected when it is true at the $\alpha = 0.05$ levels for 1,000 experimental data with R program version 3.3.2.

3. The probability of type I error is computed with the number of H_0 is rejected when H_0 is true divide by the number of replications 1,000 times for each test statistics, that is

$$\frac{\text{The number of } H_0 \text{ rejection when it is true}}{\text{The number of replications 1,000 times}}.$$

4. The probability of type I error is compared under the criterion of Bradley (1978). At significance level $(\alpha = 0.05)$, the test statistic is called robustness when the probability of type I error has to fall between (0.025, 0.075) in Bradley limit. If any the probability of type I error is over the limit, it shows that the test statistic isn't called robustness.

5. The analyses are repeated for the same both sample sizes and significance level. The data are generated from all distribution-types with unequal variances for only the test statistic is called robustness.

6. Power of a test or power is computed with the number of H_0 is rejected when H_0 is false divide by the number of replications 1,000 times for each test statistics, that is

$$\frac{\text{The number of } H_0 \text{ rejection when it is false}}{\text{The number of replications 1,000 times}}.$$

7. The maximum power of a test is 1.0 and the minimum is zero. The test statistic is called the best test when it shows the highest power or power of a test.

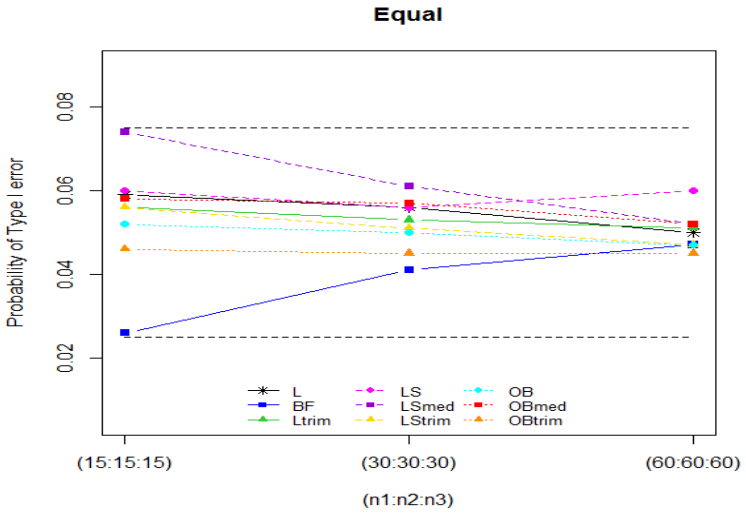
3. Results and Discussion

Here Table 3 and Figures 1-4 show the probability of type I error under equal variances at significance level $(\alpha = 0.05)$, whereas Table 4 and Figures 5-8 show that the power of the nine tests statistics. The data are generated from four distribution-types consists of normal distribution, uniform distribution, logistic distribution, and gamma distribution.

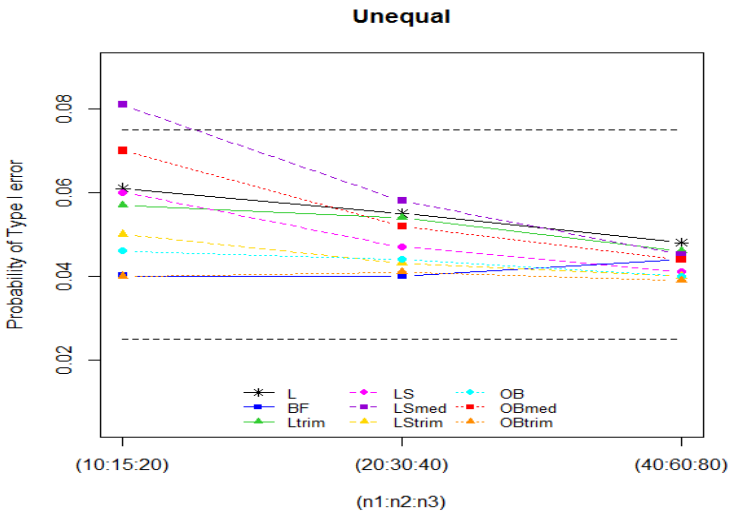
Table 3 Probability of type I error under equal variances of three Levene's tests with absolute deviation, three Levene's tests with squared deviation, and three O'Brien tests based on 1,000 simulations (robustness) when the data are generated from normal distribution, uniform distribution, logistic distribution, and gamma distribution

Distributions	Sample sizes	Test statistics								
		L	BF	Ltrim	LS	LSmed	LStrim	OB	OBmed	OBtrim
Normal	15:15:15	0.059*	0.026*	0.056*	0.060*	0.074*	0.056*	0.052*	0.058*	0.046*
	30:30:30	0.056*	0.041*	0.053*	0.056*	0.061*	0.051*	0.050*	0.057*	0.045*
	60:60:60	0.050*	0.047*	0.051*	0.049*	0.052*	0.047*	0.047*	0.052*	0.045*
	10:15:20	0.061*	0.040*	0.057*	0.060*	0.081	0.050*	0.046*	0.070*	0.040*
	20:30:40	0.055*	0.040*	0.054*	0.047*	0.058*	0.043*	0.044*	0.052*	0.041*
	40:60:80	0.048*	0.044*	0.046*	0.041*	0.045*	0.040*	0.040*	0.044*	0.039*
Uniform	15:15:15	0.042*	0.015	0.039*	0.039*	0.079	0.035*	0.034*	0.063*	0.028*
	30:30:30	0.065*	0.034*	0.064*	0.062*	0.081	0.060*	0.058*	0.074*	0.050*
	60:60:60	0.060*	0.043*	0.058*	0.055*	0.075*	0.052*	0.053*	0.070*	0.050*
	10:15:20	0.056*	0.024	0.050*	0.050*	0.078	0.044*	0.038*	0.071*	0.033*
	20:30:40	0.055*	0.031*	0.050*	0.055*	0.077	0.048*	0.049*	0.069*	0.043*
	40:60:80	0.040*	0.022	0.039*	0.040*	0.057*	0.036*	0.040*	0.053*	0.033*
Logistic	15:15:15	0.055*	0.026*	0.050*	0.043*	0.064*	0.039*	0.038*	0.052*	0.031*
	30:30:30	0.054*	0.040*	0.049*	0.047*	0.054*	0.041*	0.041*	0.049*	0.039*
	60:60:60	0.054*	0.046*	0.052*	0.040*	0.047*	0.038*	0.039*	0.045*	0.036*
	10:15:20	0.058*	0.031*	0.053*	0.053*	0.062*	0.042*	0.040*	0.048*	0.034*
	20:30:40	0.053*	0.038*	0.049*	0.048*	0.056*	0.041*	0.042*	0.050*	0.037*
	40:60:80	0.045*	0.038*	0.044*	0.034*	0.040*	0.033*	0.034*	0.039*	0.032*
Gamma	15:15:15	0.079	0.042*	0.072*	0.063*	0.091	0.050*	0.049*	0.075*	0.038*
	30:30:30	0.079	0.041*	0.066*	0.052*	0.076	0.039*	0.049*	0.068*	0.036*
	60:60:60	0.062*	0.038*	0.048*	0.043*	0.057*	0.033*	0.039*	0.051*	0.032*
	10:15:20	0.089	0.044*	0.074*	0.063*	0.090	0.052*	0.057*	0.080	0.046*
	20:30:40	0.080	0.048*	0.072*	0.046*	0.075*	0.039*	0.046*	0.068*	0.036*
	40:60:80	0.061*	0.039*	0.047*	0.048*	0.061*	0.036*	0.044*	0.058*	0.032*

*Test statistic is called robustness when the probability of type I error has to fall between (0.025, 0.075) at significance level (α =0.05).

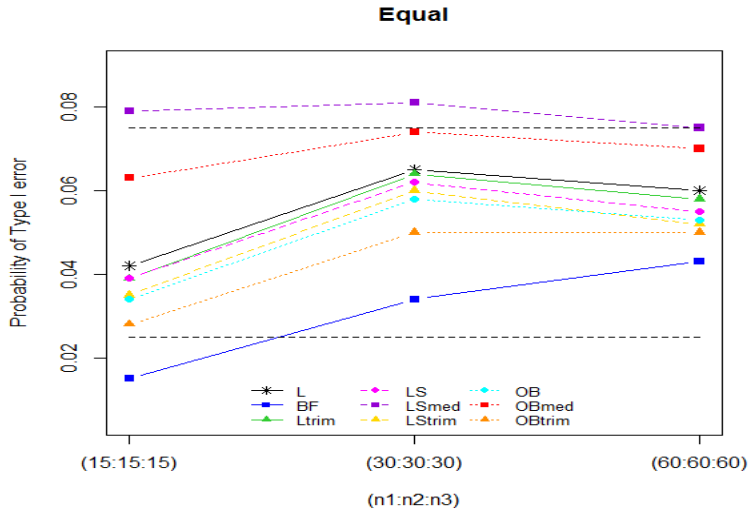


(a)

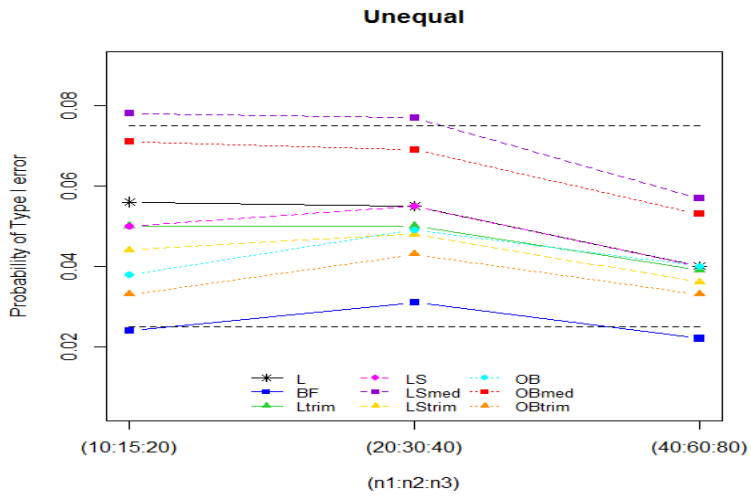


(b)

Figure 1 Probability of type I error under equal variances of three Levene's tests with absolute deviation, three Levene's tests with squared deviation, and three O'Brien tests when the data are generated from normal distribution based on (a) sample sizes are equal and (b) sample sizes are unequal

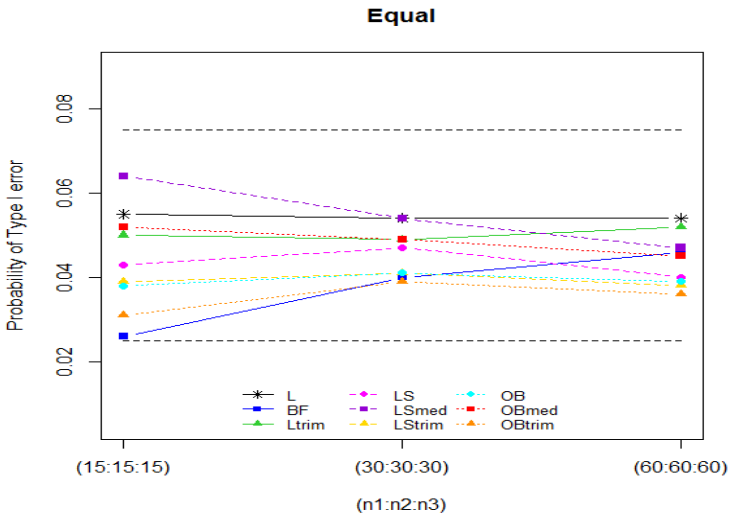


(a)

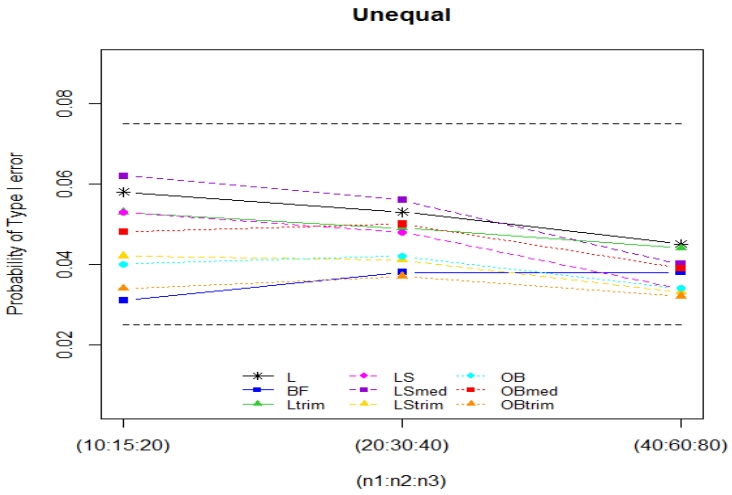


(b)

Figure 2 Probability of type I error under equal variances of three Levene's tests with absolute deviation, three Levene's tests with squared deviation, and three O'Brien tests when the data are generated from uniform distribution based on (a) sample sizes are equal and (b) sample sizes are unequal

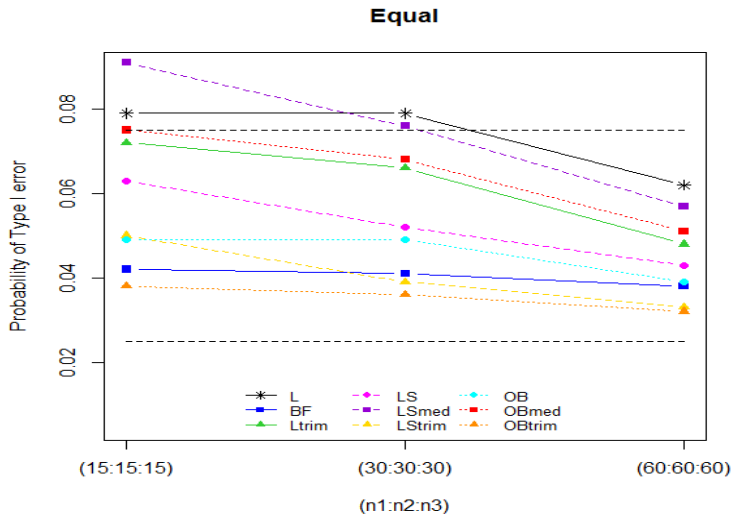


(a)

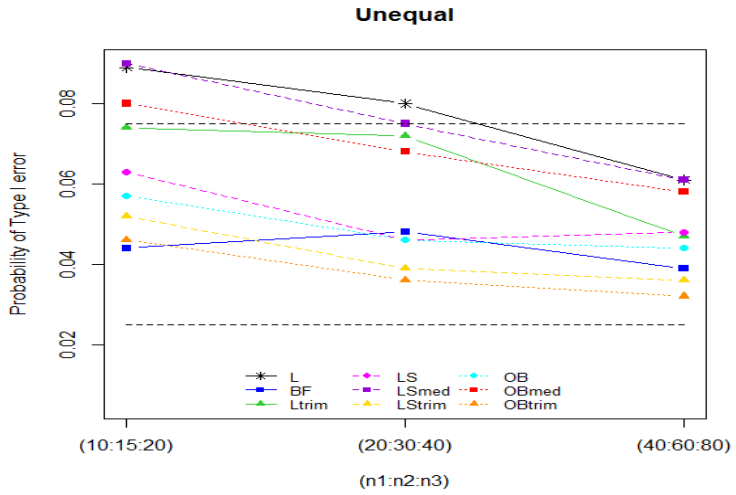


(b)

Figure 3 Probability of type I error under equal variances of three Levene's tests with absolute deviation, three Levene's tests with squared deviation, and three O'Brien tests when the data are generated from logistic distribution based on (a) sample sizes are equal and (b) sample sizes are unequal



(a)



(b)

Figure 4 Probability of type I error under equal variances of three Levene's tests with absolute deviation, three Levene's tests with squared deviation, and three O'Brien tests when the data are generated from gamma distribution based on (a) sample sizes are equal and (b) sample sizes are unequal

Table 4 Power of three Levene's tests with absolute deviation, three Levene's tests with squared deviation, and three O'Brien tests based on 1,000 simulations

Distributions	Sample sizes	Ratios of variance	Test statistics								
			L	BF	Ltrim	LS	LSmed	LStrim	OB	OBmed	OBtrim
Normal	15:15:15	1:2:4	0.517	0.408	0.501	0.481	0.522	0.453	0.440	0.479	0.408
		1:4:8	0.842	0.759	0.835	0.726	0.761	0.689	0.671	0.711	0.637
		1:8:16	0.980	0.950	0.974	0.846	0.875	0.818	0.806	0.841	0.771
	30:30:30	1:2:4	0.857	0.823	0.851	0.853	0.867	0.839	0.841	0.853	0.826
		1:4:8	0.995	0.991	0.995	0.986	0.986	0.983	0.983	0.986	0.977
		1:8:16	1.000	1.000	1.000	0.999	0.999	0.999	0.999	0.999	0.997
	60:60:60	1:2:4	0.989	0.989	0.990	0.993	0.993	0.992	0.993	0.992	0.992
		1:4:8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1:8:16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	10:15:20	1:2:4	0.452	0.344	0.439	0.290	-	0.264	0.259	0.279	0.229
		1:4:8	0.733	0.641	0.719	0.456	-	0.415	0.409	0.444	0.366
		1:8:16	0.913	0.844	0.902	0.572	-	0.520	0.521	0.559	0.467
	20:30:40	1:2:4	0.799	0.769	0.789	0.727	0.739	0.711	0.715	0.718	0.687
		1:4:8	0.984	0.980	0.983	0.945	0.950	0.937	0.938	0.945	0.922
		1:8:16	1.000	0.999	1.000	0.984	0.986	0.982	0.983	0.984	0.976
	40:60:80	1:2:4	0.990	0.989	0.990	0.990	0.990	0.989	0.989	0.989	0.988
		1:4:8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1:8:16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Uniform	15:15:15	1:2:4	0.716	-	0.705	0.797	-	0.775	0.767	0.803	0.737
		1:4:8	0.952	-	0.946	0.957	-	0.944	0.940	0.949	0.922
		1:8:16	0.998	-	0.997	0.986	-	0.979	0.980	0.982	0.967
	30:30:30	1:2:4	0.964	0.939	0.962	0.993	-	0.991	0.992	0.992	0.988
		1:4:8	1.000	1.000	1.000	1.000	-	1.000	1.000	1.000	1.000
		1:8:16	1.000	1.000	1.000	1.000	-	1.000	1.000	1.000	1.000
	60:60:60	1:2:4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1:4:8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1:8:16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	10:15:20	1:2:4	0.673	-	0.652	0.693	-	0.653	0.662	0.697	0.614
		1:4:8	0.921	-	0.902	0.896	-	0.864	0.869	0.879	0.827
		1:8:16	0.991	-	0.987	0.949	-	0.928	0.934	0.949	0.905
	20:30:40	1:2:4	0.958	0.934	0.957	0.992	-	0.988	0.990	0.989	0.985
		1:4:8	1.000	1.000	1.000	1.000	-	1.000	1.000	1.000	1.000
		1:8:16	1.000	1.000	1.000	1.000	-	1.000	1.000	1.000	1.000
	40:60:80	1.000	-	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1.000	-	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1.000	-	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4 (Continued)

Distributions	Sample sizes	Ratios of variance	Test statistics								
			L	BF	Ltrim	LS	LSmed	LStrim	OB	OBmed	OBtrim
Logistic	15:15:15	1:2:4	0.456	0.352	0.438	0.357	0.394	0.316	0.310	0.350	0.274
		1:4:8	0.757	0.648	0.732	0.539	0.574	0.500	0.483	0.536	0.443
		1:8:16	0.940	0.889	0.932	0.670	0.706	0.625	0.620	0.657	0.570
	30:30:30	1:2:4	0.777	0.741	0.768	0.685	0.707	0.657	0.668	0.692	0.629
		1:4:8	0.980	0.975	0.980	0.917	0.922	0.902	0.907	0.912	0.889
		1:8:16	1.000	1.000	1.000	0.968	0.971	0.960	0.961	0.967	0.949
	60:60:60	1:2:4	0.977	0.974	0.977	0.951	0.951	0.947	0.949	0.951	0.945
		1:4:8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
		1:8:16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	10:15:20	1:2:4	0.347	0.276	0.333	0.192	0.208	0.167	0.164	0.181	0.143
		1:4:8	0.605	0.510	0.579	0.280	0.303	0.255	0.244	0.265	0.221
		1:8:16	0.829	0.739	0.803	0.353	0.389	0.323	0.317	0.344	0.287
	20:30:40	1:2:4	0.706	0.668	0.699	0.524	0.532	0.495	0.509	0.515	0.480
		1:4:8	0.949	0.939	0.949	0.721	0.739	0.692	0.702	0.717	0.668
		1:8:16	0.996	0.996	0.996	0.850	0.862	0.820	0.823	0.840	0.791
	40:60:80	1:2:4	0.966	0.961	0.965	0.896	0.894	0.888	0.890	0.889	0.885
		1:4:8	1.000	1.000	1.000	0.989	0.990	0.984	0.989	0.990	0.984
		1:8:16	1.000	1.000	1.000	0.995	0.996	0.994	0.995	0.995	0.994
Gamma	15:15:15	1:2:4	-	0.367	0.479	0.431	-	0.397	0.381	0.432	0.348
		1:4:8	-	0.691	0.800	0.625	-	0.576	0.577	0.635	0.528
		1:8:16	-	0.880	0.940	0.694	-	0.627	0.637	0.710	0.570
	30:30:30	1:2:4	-	0.801	0.834	0.800	-	0.775	0.788	0.810	0.762
		1:4:8	-	0.987	0.991	0.939	-	0.923	0.930	0.945	0.909
		1:8:16	-	1.000	1.000	0.956	-	0.924	0.945	0.965	0.909
	60:60:60	1:2:4	0.993	0.991	0.993	0.990	0.993	0.990	0.990	0.992	0.990
		1:4:8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1:8:16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	10:15:20	1:2:4	-	0.332	0.408	0.285	-	0.248	0.248	-	0.220
		1:4:8	-	0.559	0.650	0.378	-	0.321	0.332	-	0.286
		1:8:16	-	0.741	0.840	0.414	-	0.349	0.371	-	0.302
	20:30:40	1:2:4	-	0.739	0.774	0.677	0.696	0.636	0.658	0.675	0.620
		1:4:8	-	0.955	0.964	0.811	0.840	0.768	0.800	0.817	0.748
		1:8:16	-	0.999	0.999	0.844	0.887	0.776	0.834	0.869	0.752
	40:60:80	1:2:4	0.990	0.981	0.985	0.976	0.979	0.965	0.975	0.978	0.964
		1:4:8	1.000	1.000	1.000	0.996	0.996	0.995	0.996	0.996	0.995
		1:8:16	1.000	1.000	1.000	0.985	0.988	0.982	0.984	0.988	0.982

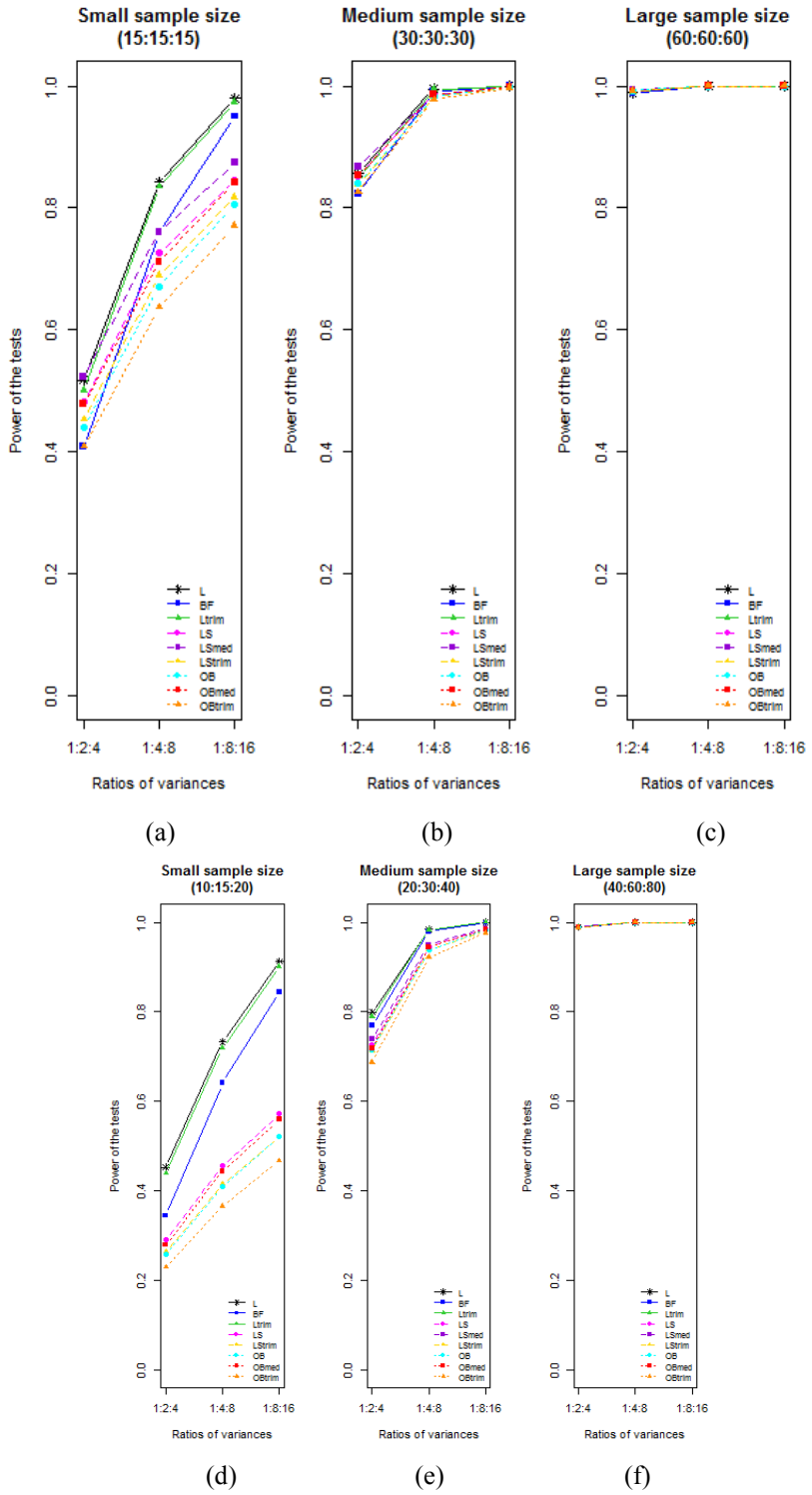


Figure 5 Power of three Levene's tests with absolute deviation, three Levene's tests with squared deviation, and three O'Brien tests when the data are generated from normal distribution

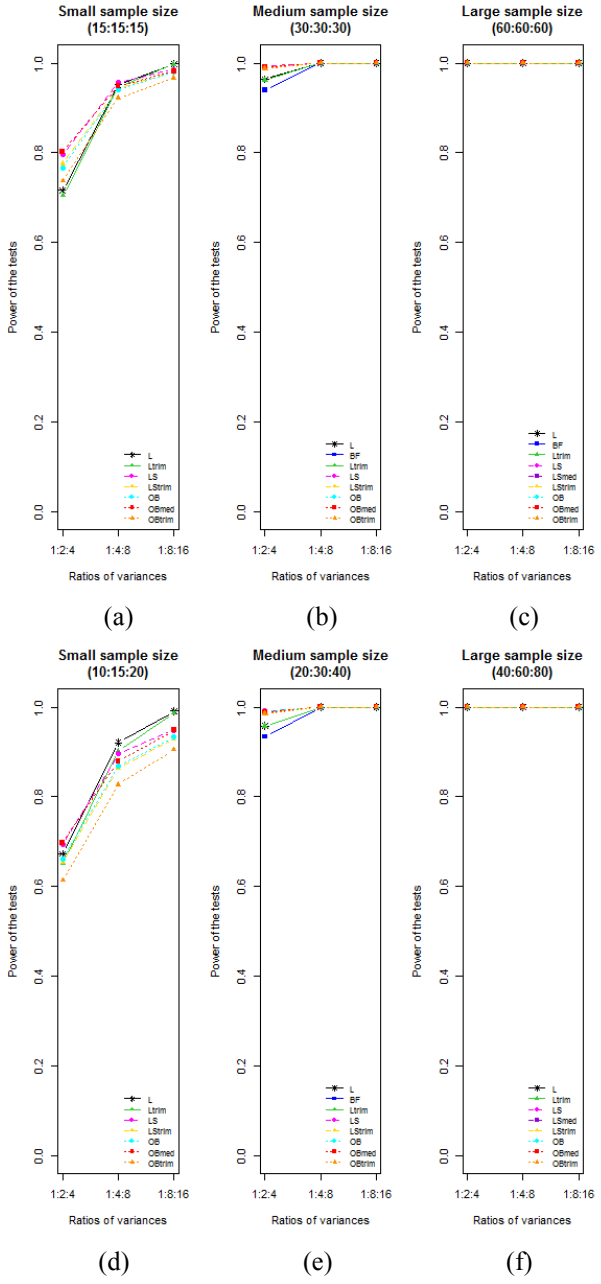


Figure 6 Power of three Levene's tests with absolute deviation, three Levene's tests with squared deviation, and three O'Brien tests when the data are generated from uniform distribution

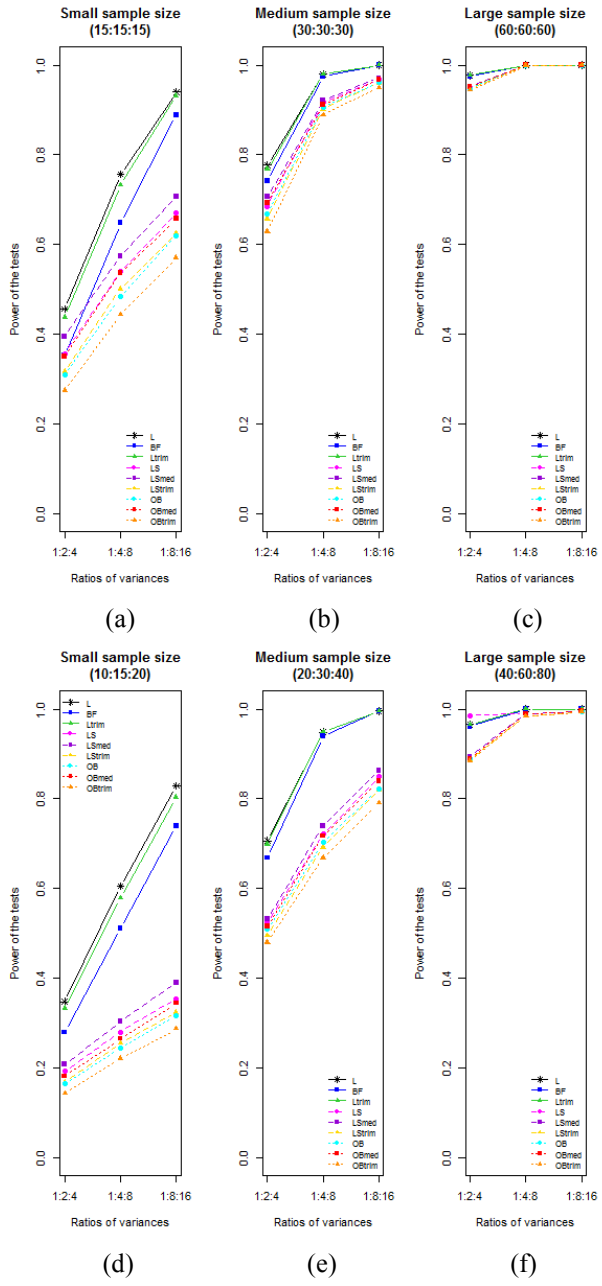


Figure 7 Power of three Levene's tests with absolute deviation, three Levene's tests with squared deviation, and three O'Brien tests when the data are generated from logistic distribution

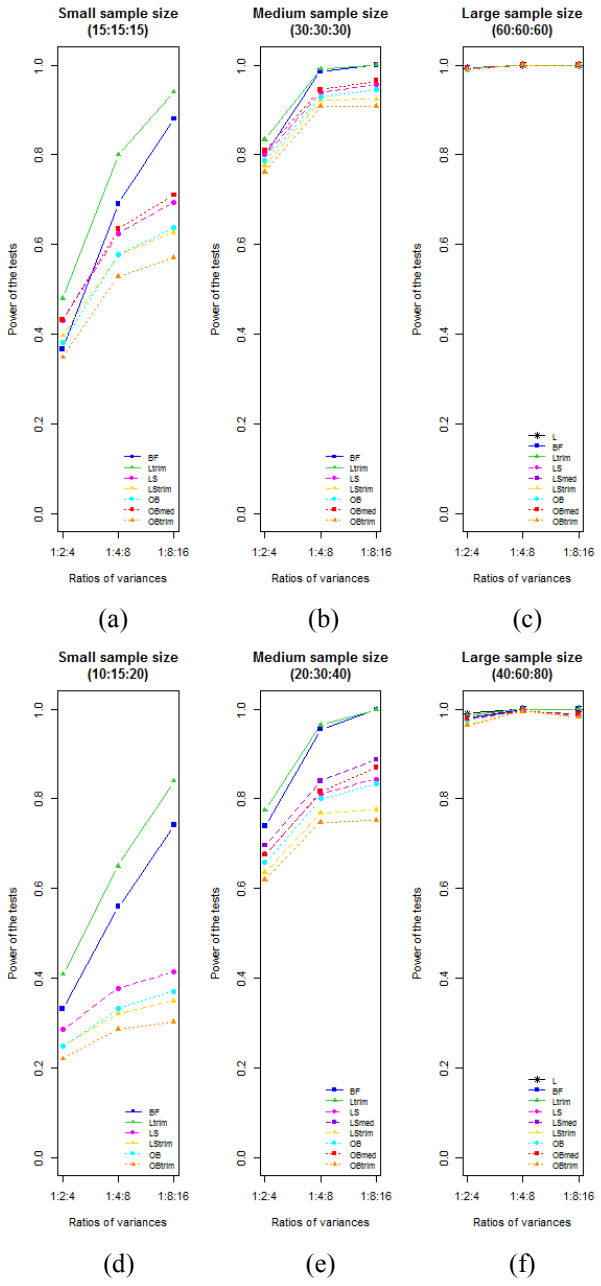


Figure 8 Power of three Levene's tests with absolute deviation, three Levene's tests with squared deviation, and three O'Brien tests when the data are generated from gamma distribution

The results of robustness: The results from these studies are found that the test statistics consists of Ltrim test and LStrim test are provide the probability of type I errors have to fall between (0.025, 0.075) in Bradley limit as well as OB test when the data are generated from all distributions, whereas L test provides the probability of type I error is over the Bradley limit when the data are generated from gamma distribution based on small and medium sample sizes. The

results show in Table 3 and Figures 1-4. When the results are considered each distribution, Figure 1 shows the all tests provide probability of type I error has to fall between (0.025, 0.075) in Bradley limit except LSmed test under normal distribution. Figure 2 shows LSmed test provide probability of Type I error has to fall in Bradley limit when the data are generated from uniform distribution based on large sample sizes both equal and unequal. On the other hand, BF test provides the probability of type I error is over the Bradley limit when large and unequal sample size under uniform distribution. Figure 3 shows all tests provide probability of type I error have to fall in Bradley limit for all conditions of sample sizes under logistic distribution. Figure 4 shows L test provides good probability of type I error for only large equal and unequal sample sizes same as LSmed test, while OBmed test provide probability of type error has to fall in Bradley limit for all conditions of sample sizes except small and unequal sample sizes under gamma distribution.

For all statistics tests, the probability of type I error decrease with sample sizes are large.

The results of power: The results from these studies are shown in Table 4 and Figures 5-8. The details are as follows:

Under normal distribution and equal sample sizes, Figure 5 shows that LSmed test shows the highest power, but L test shows the highest power when ratios of variances are increased when small sample sizes. For unequal sample sizes, L test shows the highest power for all conditions of sample sizes. Ltrim test is the highest power as well as L test when sample sizes and ratios of variances are increased. However, all tests are the highest power when large sample sizes and ratios of variances are increased both equal and unequal sample sizes.

Under uniform distribution, Figure 6 shows that OBmed test shows the highest power for small sample sizes both equal and unequal sample sizes. All tests show the highest power when sample sizes and ratios of variances are increased both equal and unequal sample sizes except BF test for unequal sample sizes.

Under Logistic distribution, Figure 7 shows that L test shows the highest power for all conditions of sample sizes. Whereas Ltrim test is the highest power as well as L test for medium and large sample sizes, and then sample sizes and ratios of variances are increased.

Under gamma distribution, Figure 8 shows that Ltrim test shows the highest power, while LSmed test has the highest power as well as Ltrim test for large and equal sample sizes.

Here these tests are observed that power increases with increase of sample sizes, for all the tests. The “best” values are printed in bold in Table 4. For all statistics tests, the power of all tests close to one when the data are generated based on ratios of variances and sample sizes are large.

4. Conclusions

From the above discussion, considering robustness and power of the tests concluded that Modified Levene’s test with squared deviation based on median (LSmed test) is the best choice for equal sample sizes when the data are established with normal distribution modified O’Brien (OBmed test) is the best choice for small sample sizes both equal and unequal sample sizes when the data are established with uniform distribution. Levene’s test with absolute deviation (Ltest) is still the best choice when the data are established with logistic distribution, and then modified Levene’s test with absolute deviation based on trimmed mean (Ltrim test) is the best choice for all conditions of sample sizes, while LSmed test is the best test same as Ltrim test and L test for large equal sample sizes when the data with logistic distribution, and then Ltrim test is the best choice for all conditions of sample sizes when the data are established from gamma distribution, whereas LSmed test is the best choice same as Ltrim test for large and equal sample sizes.

Therefore Levene's test with absolute deviation is used by generally statistical packages such as SPSS (now PASW) is not the best choice. There are better tests that can be used and some are more preferable depending on the distributional shape of the data. The modified these tests satisfied that one mentioned significance level $(\alpha = 0.05)$. But on the other hand the modified these tests less satisfies the significance level $(\alpha = 0.01)$. Since the mere substitution of the median and the trimmed mean for the mean in original tests contain Levene's test with absolute and squared deviations and O'Brien test computations will alleviate some of the problems.

References

Bishop TA, Dudewicz EJ. Exact analysis of variance with unequal variances: Test procedures and tables. Technometrics. 1978; 20: 419-430.

Bradley JV. Robustness?. Br J Math Stat Psychol. 1978; 31: 144-152.

Brown MB, Forsythe AB. Robust tests for equality of variances. J Am Stat Assoc. 1974; 69: 364-367.

Carroll RJ, Schneider H. A note on Levene's tests for equality of variances. Stat Probab Lett. 1985; 3: 191-194.

Cochran WG. The distribution of the largest of a set of estimated variances as a fraction of their total. Ann Hum Genet. 1941; 11: 47-52.

Craparo RM. Significance level. In: Salkind NJ, editor. Encyclopedia of Measurement and Statistics 3. Thousand Oaks CA: SAGE Publications; 2007, 889-891.

Game PA, Winkler H, Probert DA. Robust tests for homogeneity of variance. Educ Psychol Meas. 1972; 32: 887-910.

Gastwirth JL, Gel YR, Miao W. The impact of Levene's test of equality of variances on statistical theory and practice. Stat Sci. 2010; 24: 343-360.

Gogoi P, Gogoi B. Multi-sample scale tests for comparing scale parameters with equal and unequal location differences. Int J Adv Res. 2015; 2: 26-31.

Gorbunova AA, Lemeshko BY. Application of parametric homogeneity of variances tests under violation of classical assumption. Proceedings of 2nd Stochastic Modeling Techniques and Data Analysis International Conference; 2012 June 5-8, Chania Crete Greece. 2012.

Hanif HM, Norazan MR, Rasmani KA, Nor AMG. Enhancing hotelling T^2 control chart performance with decile and trimmed means. ICIOMS2013: Proceedings of the International Conference on Information, Operations Management and Statistics; 2013 Sep 1-3; Malaysia. Australia: IEOM Research Solutions Pty Ltd; 2013, 1-7.

Lee HB, Katz GS, Restori AF. A monte carlo study of seven homogeneity of variance tests. J Math Stat. 2010; 6: 359-366.

Levene H. Robust tests for equality of variances. In: Olkin I, editor. Contributions to Probability and Statistics. Palo Alto. CA: Stanford: University Press; 1960, 278-292.

Lu MH, Yiğit S, Mollaoğulla A, Genç S, Mendeş M. Influence of using alternative means on type-I error rate in the comparison of independent groups. JAPS. 2014; 24: 344-349.

Mendes M, Pala A. Evaluation of four tests when normality and homogeneity of variance assumptions are violated. J Appl Sci. 2004; 4: 38-42.

O'Brien RG. A simple test for variance effects in experimental designs. Psychol Bull. 1981; 89: 570-574.

Oladejo NK, Adetunde IA. Assessing computer knowledge in senior high school: A case study of the upper east region in Ghana. J Math Stat. 2009; 5: 287-297.

Othman AR, Yin TS, Keselman H, Wilcox RR, Algina J. Robust Modifications of the Levene and O'Brien Tests for spread. Journal of Modern applied Statistical Methods. 2012; 11: 54-68.

Overall JE, Woodward JA. A simple test for homogeneity of variance in complex factorial design. Psychometrika. 1974; 39: 311-318.

Parra FI. Testing homogeneity of variances with unequal sample sizes. Comput Stat. 2012;28: 1269-1297.

Reiczigel J. Confidence intervals for the binomial parameter: some new considerations. Statist Med. 2003; 22: 611-621.

Sharma D, Kibria BMG. On some test statistics for testing homogeneity of variances: A comparison study. J Stat Comput Simulat. 2013; 83:1944-1963.

Tormarkan AJ, Serlin RC. Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychol Bull. 1986; 99: 90-99.

Tukey JW. Some elementary problems of importance to small sample practice. Hum Biol. 1948; 20: 205-214.

Wang Y, Rodríguez PDG, Chen YH, et al. Comparing the Performance of approaches for testing the homogeneity of variance assumption in one-factor ANOVA models. Educ Psychol Meas. 2016; 77: 305-329.