Journal of Quality Measurement and Analysis *JQMA* 18(3) 2022, 111-130
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*

AN OVERVIEW OF HOMOGENEITY OF VARIANCE TESTS ON VARIOUS CONDITIONS BASED ON TYPE 1 ERROR RATE AND POWER OF A TEST

(Sorotan Ujian Kehomogenan Varians pada Pelbagai Keadaan Berdasarkan Kadar Ralat Jenis 1 dan Kuasa Ujian)

NUR FAZLIN ABDULLAH & NORA MUDA*

ABSTRACT

In most statistical analyses, the data variance used is assumed to be homogeneous, but not all cases follow the assumption. Therefore, the homogeneity of variance assumption testing should be carried out prior to performing the main analysis. There are various statistical tests of variance homogeneity that exist and to obtain the best statistical test in the testing of variance equality, this study makes a comparison of statistical tests against assumptions met, assumptions violated and the existence of outlier. The comparison is based on the Type 1 Error rate and the power of the statistical test. For normal distribution, the comparison is between parametric statistical tests such as the Fisher test, the Bartlett test, the Levene test, the Brown-Forsythe test, and the Cochran C test. While for chi-squared distribution and outlier data, the comparison is between parametric and nonparametric statistical tests. The nonparametric statistical tests used are the Mood test, the Ansari-Bradley test, and the Fligner-Killeen test. The data used is the result of a normal and Chi-squared Monte Carlo simulation. The results showed that almost all the parametric statistical tests can control Type 1 Errors well in almost all situations. For the Chi-squared distribution only the Brown-Forsythe parametric statistical test was found to be robust. But most of the robust tests on non-normal data are nonparametric statistical tests. While for normal data with heterogeneous variance, the power of the test of all parametric statistical tests is seen to increase and exceed 0.80 as the size effect increases. On non-normal distributions, the power of the test is smaller than normal, but the value will increase as the size effect increases. The case was different for the Fisher test, the Bartlett test, and the Cochran C test, which was tested against data with 10% outlier in one group. The power of the test for the 1:2 variance ratio is seen as large, but the value is decreasing as the size effect increases. Thus, it can be concluded that none of the statistical tests were found to be robust and suitable for use in all the conditions set.

Keywords: Monte Carlo simulation; nonparametric test; outlier; parametric test; robust test

ABSTRAK

Dalam kebanyakan analisis statistik, varians bagi data yang digunakan adalah bersifat homogen. Tetapi bukan semua kes yang andaian kehomogenan varians dicapai. Oleh sebab itu, pengujian andaian kehomogenan varians perlu dilakukan sebelum analisis utama dijalankan. Terdapat pelbagai ujian statistik kehomogenan varians yang wujud dan untuk mendapatkan ujian statistik yang terbaik dalam pengujian kesamaan varians, kajian ini membuat perbandingan ujian-ujian statistik terhadap andaian dipenuhi, andaian tidak dipenuhi dan kewujudan data terpencil. Perbandingan yang dilakukan adalah berdasarkan kadar Ralat Jenis 1 dan kuasa ujian statistik. Bagi taburan normal, perbandingan adalah antara ujian-ujian statistik berparameter seperti ujian Fisher, ujian Bartlett, ujian Levene, Ujian Brown-Forsythe dan ujian Cochran C. Manakala bagi taburan khi-kuasa dua dan data terpencil, perbandingan yang dibuat adalah antara ujian statistik berparameter dan tak berparameter. Ujian statistik tak berparameter yang digunakan adalah ujian Mood, ujian Ansari-Bradley dan ujian Fligner-Killeen. Manakala data yang digunakan adalah hasil simulasi Monte Carlo yang tertabur secara normal dan Khi-kuasa dua. Hasil kajian mendapati hampir kesemua ujian statistik berparameter dapat mengawal Ralat Jenis 1 dengan baik pada hampir kesemua keadaan. Bagi taburan Khi-kuasa dua hanya ujian statistik berparameter Brown-Forsythe yang didapati teguh. Tetapi kebanyakan ujian yang teguh pada keadaan data bukan normal adalah ujian statistik tak berparameter. Manakala bagi data normal dengan varians heterogen, kuasa ujian bagi kesemua ujian statistik berparameter dilihat meningkat dan melebihi 0.80 apabila kesan saiz semakin meningkat. Pada taburan bukan normal, kuasa ujian adalah lebih kecil berbanding normal tetapi nilai tersebut akan meningkat seiring dengan peningkatan kesan saiz. Kes ini berlainan pula bagi ujian Fisher, ujian Bartlett dan ujian Cochran C yang diuji terhadap data dengan kewujudan 10% nilai terpencil pada satu kumpulan. Kuasa ujian tersebut bagi nisbah varians 1:2 dilihat besar tetapi nilai tersebut semakin menurun apabila kesan saiz semakin meningkat. Maka, boleh disimpulkan bahawa tiada satu ujian statistik pun yang didapati teguh dan sesuai digunakan pada kesemua keadaan yang ditetapkan.

Kata kunci: simulasi Monte Carlo; ujian tak berparameter; terpencil; ujian berparameter; ujian teguh

1. Introduction

There are various types of parametric tests in hypothesis testing. Typically, the parametric hypothesis testing tests the mean, variance, and proportion parameters for cases of one population, two populations, or more than two populations. This study focuses on comparing homogeneity of variance hypothesis testing on two populations. Homogeneity of variance means the variance for each population is equal. Currently, there are various parametric and nonparametric statistical tests to test the homogeneity of variance. The comparison of the homogeneity of variance hypotheses testing is carried out based on the assumptions met, assumptions violated, and the presence of outliers. The assumptions tested in this study are normality, equality of variance, and sample size.

According to Johann Karl Gauss, most statistical methods for parametric tests, such as correlation, regression, t-test, and ANOVA, assume that the data used are based on a normal distribution, also known as a Gaussian distribution. A normal distribution is a bell-shaped distribution in which the high frequency is at the center of the graph shape while the low frequency is at the end of the graph (Gravetter & Wallnau 2000). According to Chambers *et al*. (1983), researchers can find out more information concerning the studies conducted by knowing and understanding the types of distribution used in their studies. When these assumptions of normality are skewed, the results and conclusions made based on a study can be inaccurate or not true.

Brown and Forsythe (1974) indicated that a robust measure of central tendency in statistical tests to test the homogeneity of variance hypothesis should be considered. Another study that considering the outlier and the violation of the assumption in the test were examined by Ercheg-Hurn and Mirosevich (2008). They examined an ANOVA and least square test with various range of modern robust and rank-based significance tests by using software such as SAS, SPSS and R and concluded the test of assumption by discussing with the robust effect size indices. On the other hand, Underwood (1997) indicated that ANOVA would be problematic if the variance between groups is heterogeneous. Thus, Legendre and Borcard (2000) compared Type 1 Error rate and power of a test for statistical tests of homogeneity of variance testing the assumption of ANOVA homogeneity of variance. The statistical tests compared in the study were the Bartlett, Box’s M, Scheffe Box log-anova, and Cochran’s C tests. Vorapongsathorn *et al*. (2004) carried out a comparative study of Type 1 Error rate and power of a test for Bartlett, Levene, and Cochran tests when the assumptions of data normality are not met. The study aimed to find out the most appropriate statistical test among the three tests for a given situation to test the equality of variance. Lee *et al*. (2010) and Mendes *et al*. (2006) conducted a study on the homogeneity of variance test using the Monte Carlo simulation. According to Oladejo and Adetunde (2009), and Mazahreh *et al*. (2009), to test the homogeneity of variance of a data prior to performing the t-test and ANOVA test, the Levene test (Gastwirth *et al*. 2009; Carroll and Schneider 1985; Tomarken and Serlin 1986) is often used in the SPSS software. Subsequently, the question concerning the Levene test is the best test to test the homogeneity of variance in the SPSS software was raised. Thus, Lee *et al*. (2010) compared the Levene test with six other tests of homogeneity of variance to determine the best homogeneity of variance test for a given data. Other tests of homogeneity of variance being compared were the Fmax test (Hartley 1950), Samiuddin cube root test (Samiuddin and Atiqullah 1976), Z-variance test, Z-variance modified test (Overall and Woodward 1976), O'Brien test (O'Brien 1978; 1981), and Levene modified test (Carroll and Scnheider 1985). The comparison in the study was based on the test robustness by obtaining the Type 1 Error rate and power of a test through the Monte Carlo simulation (Alabi *et al*. 2008; Rana *et al*. 2008; Agunbiade and Iyuniwura 2010) of 3000. Their study used four groups with a balanced design consisting of sample sizes of n=10 and n=30. The comparison was carried out on the normal distribution, chi-square distribution with 5 degrees of freedom, chi-square distribution with 6 degrees of freedom, leptokurtic distribution, and platykurtic distribution.

Lemeshko *et al*. (2010a), Lemeshko *et al*. (2010b), Lemeshko *et al*. (2010) and Gobunova & Lemeshko (2012) were studied on the statistical test to test the homogeneity of variance and Lemeshko *et al*. (2010) indicated that the best statistical test to test the homogeneity of variance when the assumption of normality is not met is the Cochran's C test. Their study aimed to obtain the best homogeneity of variance test for data not meeting the normality criteria. Thus, they compared the classical and nonparametric tests on the De, Laplace, and normal distributions for the groups, k=2, k=3, k=4, and k=5, by obtaining the power of statistical tests. The classical tests selected in their study were the Fisher, Bartlett, Cochran's C, Hartley, and Levene tests. Meanwhile, the Ansari-Bradley, Mood, and Siegel-Tukey tests were selected for the nonparametric tests. Upon testing and comparing, they found the Mood nonparametric test had the highest power of a test compared to other nonparametric tests. Nevertheless, when the classical tests were compared with the nonparametric test, the Cochran's C test still had the highest power of a test compared to others. Thus, it shows that Cochran's C test is the best test for distributions skewed from the normality.

The study by Nordstokke and Zumbo (2010) concerned the homogeneity of variance hypothesis testing using the Levene modified parametric and nonparametric statistical tests. The Levene modified parametric statistical test used the propensity of median measure of central tendency. This statistical test is also called the Brown-Forsythe test. In contrast, the Levene nonparametric test was based on the principle of rank transformation (Conover and Iman 1981). Various existing parametric and nonparametric tests for homogeneity of variances, and some variations of these tests were examined in Conover *et al*. (1981). These tests were conducted to do comparisons under the null hypothesis (for robustness) and under the alternative (for power) with Monte Carlo simulations of various symmetric and asymmetric distributions, for various sample sizes.

Ahad *et al*. (2011) conducted a comparative study of data normality testing using the Monte Carlo simulation based on four types of statistical test: Kolmogorov-Smirnov test (Smirnov 1935, Kolmogorov 1956), Anderson-Darling test (Anderson & Darling 1952), Cramer-von Mises test (Anderson 1962), and Shapiro-Wilk test (Shapiro & Wilk 1965). The study was carried out to determine the sample size that could detect non-normal data using non-normally distributed data. The comparison was based on the Type 2 Error rate. They found that the Shapiro-Wilk test could detect the non-normality for non-normal data on a small sample size compared to other normality tests. It was followed by the Anderson-Darling, Cramer-von Mises, and Kolmogorov-Smirnov tests. They concluded that the Shapiro-Wilk test was the most appropriate test to test the assumption of data normality. Sharma and Kibria (2013) have reviewed 25 test procedures that are widely reported in the literature for testing the hypothesis of homogeneity of variances under various experimental conditions but not considering the condition of outlier existence in the data. Conover *et al*. (2018) and Kim and Cribbie (2018) were updated on a comparative study of tests for homogeneity of variance with included the lognormal distribution and various conditions while Wang *et al*. (2017) and Yi *et al*. (2022) were compared of robust tests for homogeneity of variance in Factorial ANOVA. Patil and Kulkarni (2022) were proposed an uniformly superior exact multi-sample test procedure for homogeneity of variances under location-scale family of distributions.

In a study, not all cases meet the assumption of the equality of variance; there is also a possibility of cases not meeting the assumption. Thus, the equality of variance data testing should be carried out first not to skew the study's results. This study compares the statistical tests of homogeneity of variance on two samples for cases in which the assumptions are met, assumptions violated, and the presence of outliers to determine the study's best equality of variance test. The comparison is carried out to get the best statistical test to test a case and obtain solid results for a study.

Therefore, a comparison of the homogeneity of variance parametric test on the normally distributed data is carried out to test the equality of variance. It is followed by a comparison of parametric and nonparametric tests on the equality of variance for non-normal data and outliers and a determination of the best statistical test to overcome the assumptions of normality, equality of variance, and the presence of outlier skewness problem.

2. Monte Carlo Simulation

The analysis was carried out using the generated simulation data. The data used in this statistical test comparative study were the results of the R software simulation, based on two groups. Whereas the distributions used were the normal and chi-square distributions. The normality test was carried out on the simulated data to identify whether the simulated data is normal or non-normal. Then, the hypothesis testing on several statistical tests of the equality of variance was carried out using the simulated data. Each test carried out was evaluated based on the Type 1 Error rate and power of a test to determine the best statistical test.

The hypothesis testing in this study was carried out following the assumptions of normality, assumptions of equality of variance, and diversified sample sizes. Then, the test was carried out on non-normal data, i.e., chi-square distributed data with 3 degrees of freedom. The study's homogeneity of variance test was compared in two groups with sample sizes of $n = 10$, $n = 30$ and $n = 50$. The variance values for each group were determined based on ratios of 1:1, 1:2, 1:3, and 1:4. When coupled with heterogeneous variance, the inequality of the sample size in each group can affect Type 1 Error. Thus, this study also analyzed variance pairs and sample sizes, termed positive and negative pairs.

The positive pair here means that when the sample size is large, the variance is also large, and likewise, when the sample size is small, the variance is also small. On the other hand, the negative pair refers to a large sample size coupled with a small variance and a small sample size with a large variance.

For the study's Monte Carlo simulation for the statistical tests of homogeneity of variance, the steps for the simulation were shown to avoid confusion in carrying out the simulation. The Type 1 Error rate and power of a test calculation algorithms for the homogeneity of variance hypothesis testing are as follows:

Algorithm for the Type 1 Error rate

1. Generate a random sample from the Normal and Chi-square distributions. Assuming the null hypothesis, H_0 is correct.
2. Calculate the statistical test value.
3. Calculate the *p*-value.
4. Repeat step (1) to step (3) 10,000 times.
5. Count B = number (*p*-value < alpha)
6. Calculate the Type 1 Error = step (5) / 10,000

Algorithm for the power of a test

1. Generate a random sample from the normal and chi-square distributions. Assuming the null hypothesis, H_0 is correct.
2. Calculate the statistical test value.
3. Calculate the *p*-value.
4. Repeat step (1) to step (3) 10,000 times.
5. Count B = number (*p*-value < alpha)
6. Calculate the power of a test = step (5) / 10,000

2.1. *Method of testing data normality*

Data normality means the data is normally distributed and is around the mean. A normal distribution is a bell-shaped distribution with the highest frequency in the middle and the low frequency at the end. A normal distribution is also a continuous and symmetrical distribution with an area under the curve is equal to one.

The data should be normally distributed or close to normal for most statistical analyses because normality is important in the parametric statistical method. Thus, prior to carrying out the hypothesis testing on comparing the statistical tests of homogeneity of variance, the simulated data should be tested by data normality testing. The normality test is important to ensure that the assumption of normality is not skewed (Cochran & Cox 1957; Levene 1960; Conover *et al*. 1981; Weerahandi 1995; Zar 1999) so that the study's findings meet the hypotheses and the Type 1 Error rate for the test carried out is close to the significance level specified.

Several statistical tests are used in data normality testing (Ghasemi & Zahediasl 2012) and are readily available in the R computer software. Among them are the Kolmogorov-Smirnov test (Kolmogorov 1956; Smirnov 1935), Anderson-Darling test (Anderson & Darling 1952), Cramer-von Mises test (Anderson 1962), and Shapiro-Wilk test (Shapiro & Wilk 1965). The null hypothesis for these tests is that the data are normally distributed. When the *p*-value is greater than the specified significance level $\alpha = 0.05$, the null hypothesis is not rejected, meaning that the data is normally distributed. Similarly, the null hypothesis is rejected if the *p*-value is smaller than the significance level $\alpha = 0.05$, meaning the tested data is non-normally distributed.

One of the most important factors influencing the results of such data normality tests is the sample size. These tests are evaluated on various spectrums of non-normal data and different sample sizes. In this study, the Shapiro-Wilk test was selected to test the assumption of data normality because the test is the best normality test compared to other tests (Ahad *et al.* 2011). This test rejects the null hypothesis of the normality test on the smallest sample size for all levels of skewness and kurtosis for each distribution. The statistical test for Shapiro-Wilk is as follows:

$$W_n = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

where,

$$a' = (a_1, a_2, ..., a_n) = \frac{m'V^{-1}}{m'V^{-1}V^{-1}m},$$

$$m' = (m_1, m_2, ..., m_n)$$

is a vector of the expected value for a statistical standard normal distribution, $V = n \times n$ is a covariance matrix, and $X' = \left(X_1, X_2, ..., X_n\right)$ is a random sample.

2.2. *Outliers*

Outliers are data or observations that are far from other observations. The presence of outliers in the data can negatively impact the analysis results and provide skewed results. In addition, outliers can cause the assumption of data normality to be skewed and lower the value of the power of a test, in turn, causing the probability of making the right decision to decrease. Thus, prior to carrying out data analysis, a preliminary review of the data should be done to identify outliers and subsequently use the appropriate method on the analysis to obtain robust analysis results.

In this study the simulated normal distribution data were contaminated with 10% outliers by replacing 10% of the data with extreme or outlier values in one of the groups and both groups to see the Type 1 Error rates and power of a test. Then, this study selected several methods to identify outliers and ensure that the data used contain outliers. The methods used were the boxplot method and the comparison of data values with outliers determination intervals. The outliers in this study are refer to the data outside the interval (quantile-1-3 * *IQR* + quantile-3), with the *IQR* being the interquartile interval for the entire data.

3. Methodology

A comparative study of statistical tests of homogeneity of variance was carried out to obtain the best statistical test to test data. Thus, to realize the objective, this comparison was carried out based on the assumption is met, the assumption is violated, and the presence of outliers. This comparison was carried out by testing the hypotheses on the simulated data using statistical tests of the homogeneity of selected variance. The null and alternative hypotheses, H_0 and H_1, should be established first prior to testing the hypotheses on the statistical tests of homogeneity of variance. For this study, H_0 and H_1 are:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

*H*0 for testing this hypothesis is rejected if the statistical test value is greater than the critical value of the significance level specified at $\alpha =$ 0.05 when the *p*-value is smaller than the specified α value.

Next, several parametric and nonparametric tests of homogeneity of variance were selected to compare the statistical tests of homogeneity of variance in this study. The Fisher, Bartlett, Cochran's C, Levene, and Brown-Forsythe tests were the selected parametric tests. Meanwhile, the selected nonparametric tests were the Ansari-Bradley, Mood, and Fligner-Killeen modified tests. The comparison between parametric tests was carried out to see the best test for normally distributed data with the assumption that the equality of variance is met, violated, and diversified sample sizes. The study continued to compare parametric and nonparametric statistical tests on data not meeting the assumption of normality and data with outliers. The comparison was carried out by obtaining the Type 1 Error rate and the power of a statistical test.

The Fisher test was employed to test the equality of variance hypothesis for two groups with sample sizes of n_1 and n_2. The statistical test for the Fisher test is:

$$F = \frac{S_1^2}{S_2^2}$$

where S_1^2 and S_2^2 are non-bias samples of variance. If the sample is normally distributed and $H_0 : \sigma_1^2 = \sigma_2^2$ is true, then the statistics are F_{n_1-1,n_2-1} distributed.

The Bartlett test χ^2 was the outcome of Bartlett's (1937) proposal on special use of the χ^2 test to test the assumption of the equality of variance that the null hypothesis consisting of the same variance would be rejected if the Bartlett χ^2 statistical value is greater than the critical value χ^2 with its degree of freedom is *k*-1. On the other hand, when the Bartlett χ^2 statistical value is smaller than the critical value χ^2 with its degree of freedom is *k*-1, then *H*0 fails to be rejected. The statistical test for the Bartlett test is:

$$\chi^2 = \frac{\left(N-k\right)\log\left[\frac{\sum_{j=1}^{k}(n_j-1)s_j^2}{N-k}\right] - \sum_{j=1}^{k}(n_j-1)\log(s_j^2)}{1+\frac{\left(\sum_{j=1}^{k}\frac{1}{n_j-1}\right)\frac{1}{N-k}}{3(k-1)}}$$

where, *N* is the total sample size, $N_{\bullet j}$ is the sample size for the *j* group, *k* is the total group, and S_j^2 is the variance for the *j* group. When *H*0 is true, all $v_i > 3$ and the sample are normally distributed, the Bartlett statistical test is independent of the sample size, and the test is close to the distribution χ^2_{k-1}. However, if the sample used does not consist of a normal distribution, the Bartlett statistical test is dependent on the sample size, and its distribution is not a distribution χ^2_{k-1}.

The Cochran's C test was introduced by Cochran (1941) whereby this test is the ratio of the highest variance group and the total variance sample. The statistical test for the Cochran test is:

$$C = \frac{s^2_{\max}}{\sum s^2_j}$$

where, *n* is the number of observations in each group for a balanced design and $F_{\alpha/k}$ is the critical value of *F* at α/k with *n*-1,(*k*-1)(*n*-1) degrees of freedom. If the statistical value is greater than the critical value, then H_0 is rejected.

The Levene test uses a mean sample in the calculation and is an alternative to the Bartlett test introduced by Klotz and Johnson (1993). This test tests the assumption of variance normality for the independent *t*-test sample and ANOVA design. According to Levene (Gastwirth *et al.* 2009), the Levene test uses an absolute value of residuals or a square value of residuals transforming the variance test to a mean test because of its relatively robust nature on the assumption of normality compared to the variance test. The statistical test for the Levene test is:

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

where *N* is the total sample size, n_i is the sample size for the *i* group, $Z_{ij} = | Y_{ij} - \bar{Y}_i |$, $\bar{Y}_i$ is the mean of the *i* subgroup, $\bar{Z}_i$ is the mean of the group Z_{ij}, and $\bar{Z}$ is the overall mean of Z_{ij}. In addition, when $Z_{ij} = | Y_{ij} - \bar{Y}'_{i.} |$ where $\bar{Y}'_{i.}$ is the trimmed 10% mean for the *i* subgroup, and *W* is the statistical test for the Levene modified test. Then, H_0 is rejected if $W > F(\alpha, k-1, N-k)$, $F(\alpha, k-1, N-k)$ is the critical value of the *F* distribution with degrees of freedom of *k*-1 and *N*-*k*, and the significance level $\alpha = 0.05$.

The Brown-Forsythe test is a Levene test using a median sample in the calculation, and it is a modified Levene test to overcome the problem of inequality of variance. The statistical test for the Brown-Forsythe test is:

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

where *N* is the total sample size, n_i is the sample size for the *i* group, $Z_{ij} = |Y_{ij} - \tilde{Y}_i|, \tilde{Y}_i$ is the mean of the *i* subgroup, $\bar{Z}_i$ is the median of the group Z_{ij}, and $\bar{Z}$ is the overall median of Z_{ij}. In addition, when $Z_{ij} = |Y_{ij} - \tilde{Y}'_{i.}|$ where $\tilde{Y}'_{i\cdot}$ is the trimmed 10% mean for the *i*-th subgroup, and *W* is the statistical test for the Levene modified test. Then, H_0 is rejected if $W > F(\alpha, k-1, N-k)$, $F(\alpha, k-1, N-k)$ is the critical value of the *F* distribution with degrees of freedom of *k*-1 and *N*-*k*, and the significance level $\alpha = 0.05$.

The Ansary-Bradley test is a nonparametric statistical test to test the equality of variance. The statistical test for the Ansary-Bradley test is:

$$A = \sum_{i=1}^{m} (\frac{m+n+1}{2} - | R_i - \frac{m+n+1}{2} |),$$

where *m*, *n* are the sample size $(m \leq n)$, and R_i is the position of the *i*-th value in the sample.

In addition, the Mood test is also a nonparametric statistical test to test the equality of variance. The statistical test for the Mood test is:

$$M = \sum_{i=1}^{m} (R_i - \frac{n+m+1}{2})^2,$$

where *m*, *n* are the sample size $(m \leq n)$, and R_i is the position of the *i*-th value in the sample.

The Fligner-Killeen test is a nonparametric statistical test to test the equality of variance. This test is used on independent data and non-normally distributed data. The Fligner Killeen test was proposed by Fligner and Killeen (1976). Fligner and Killeen used a combination of absolute value ranks $|X_{ij}|$ and set the increasing scores, such as

$$a_{N,i} = i, a_{N,i} = i^2$$

or

$$a_{N,i} = \Phi^{-1} \left(\frac{\left(\frac{1+i}{N+1} \right)}{2} \right)$$

based on the ranks. However, Conover and Johnson *et al*. (1981) later modified the Fligner-Killeen test and proposed that this test uses ranking $|X_{ij} - \tilde{X}_j|$ where $\tilde{X}_j$ is the median sample for the *j*-th population. Thus, this Fligner-Killeen modified test is called the Fligner-Killeen test of median central tendency. Once the score $a_{N,i}$ is set to $|X_{ij} - \tilde{X}_j|$, the statistical test for the Fligner-Killeen modified test is:

$$X^2 = \frac{\sum_{j=1}^{k} n_j \left(\bar{A}_j - \bar{a}\right)^2}{V^2},$$

where, $\bar{A}_j$ is the mean score for the *j*-th sample, $\bar{a}$ is the overall mean score,

$$\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a_{N,j}$$

and

$$V^2 = \frac{1}{N-1}\sum_{i=1}^{N} (a_{N,i} - \bar{a})^2 .$$

For large sample sizes, the Fligner-Killeen statistical test, χ^2 asymptotically distributed in a chi-square distribution with degrees of freedom of *k*-1. This Fligner-Killeen modified test is more robust on data that the assumption of normality is skewed than the original Fligner-Killeen test.

3.1. *Type 1 Error and power of a test*

The tests were compared using the Type 1 Error value and power of a test obtained using the R software. Statistical tests with a Type 1 Error rate close to the nominal value and within Bradley's liberal criteria interval proposed by Bradley (1978) are robust tests. Meanwhile, for the power of a test, the higher the probability of detecting the inequality of variance correctly, the greater the probability of the power of a test.

3.1.1. *Type 1 Error*

Type 1 Error is the probability of rejecting a null hypothesis, whereas it is true and easy to calculate in the hypothesis testing. Once the Type 1 Error rate was obtained, this study selected Bradley's liberal criteria as the Type 1 Error rate interval. Based on these criteria, testing is robust when the empirical rate for Type 1 Error, α, is in the range $0.5\alpha \le \hat{\alpha} \le 1.5\alpha$. In this study, the significance level α was set at 0.05; a test is robust if α is in the interval of 0.025 and 0.075. However, according to Bradley (1978), a test that can control the Type 1 Error rate very well has the Type 1 Error rate within the stringent robust interval criteria, between 0.045 and 0.055. In addition, the power of the statistical test, $1-\beta$ was also obtained to compare the statistical tests in this study.

3.1.2. *Power of a test*

Power calculations are as important as significance calculations and power analysis should be regularly incorporated into experiments to find few dissenters. Therefore, to illustrate the importance of power analysis and the consequences of ignoring power, Thomas and Juanes (1996) had used an example from *Animal Behaviour* to determine whether the experiment had a good chance of producing a statistically significant result if a biologically significant difference existed in the population. In other words, whether the experiment had high power, given a biologically significant effect size.

The power of a statistical test is the probability of rejecting H_0 when H_0 is false. It means the power of a test is the probability of making the right decision. In addition, the power of a test power can also detect the effect size, i.e., σ_1^2 / σ_2^2 if such an effect is present. The effect size means the difference between the actual value and the value expressed in H_0. Thus, with the Type 1 Error rate and the power of a statistical test probability, the determination of the best statistical test of homogeneity of variance for a given data testing can be achieved.

4. Results and Discussion

The values of Type 1 Error rate and power of a test were obtained from the hypothesis testing carried out using selected tests of homogeneity of variance for the assumption that is met, the assumption that is violated, and the presence of outliers at the significance level, $\alpha = 0.05$. For a normal distribution, the comparison of homogeneity of variance carried out was only between parametric statistical tests, namely the Fisher, Bartlett, Levene, Brown-Forsythe, and Cochran's C tests. Meanwhile, for Chi-square distribution and outliers, the statistical tests compared were between parametric and nonparametric. It is because the assumption of nonparametric statistical tests is that data are not normally distributed.

4.1. *Type 1 Error*

The robustness of the testing method was assessed through its ability to control the Type 1 Error rate well. Type 1 Error is the probability value to reject the null hypothesis, H0, whereas it is true. The Type 1 Error rate was obtained when the variance for both groups are equal. A test was assumed to be robust in this study if the Type 1 Error was within Bradley's liberal criteria. At the significance level, $\alpha = 0.05$, Bradley proposed two robust criteria, i.e. the Type 1 Error rate is within the interval of 0.025 and 0.075 and the more stringent criteria, i.e. the interval of 0.045 and 0.055. Table 1 shows the complete results for the Type 1 Error rates obtained through the homogeneity of variance hypothesis testing using the Fisher, Bartlett, Cochran's C, Levene, Brown-Forsythe, Ansari-Bradley, Mood, and Fligner-Killeen modified tests. Type 1 Error rates within Bradley's liberal robust criteria interval are marked *, while Type 1 Error rates within the stringent criteria interval are marked **.

4.2. *Normal distribution*

The Fisher and the Bartlett tests showed a good level of control of Type 1 Error rates over the Normal distribution. Both tests' Type 1 Error rates were within a stringent robust criteria interval for all sample size conditions. Meanwhile, the Levene test showed good control of Type 1 Error rates on balanced designs for sample sizes of 30 and 50, respectively, and in groups with different sample sizes. However, the Levene test remained robust with the liberal criteria for a sample size of 10. In addition, the Brown-Forsythe test was able to control Type 1 Error rates well for large sample sizes, i.e. n=30 and n=50, but the test was still robust with the criteria liberal for a sample size group of 10 and different sample sizes. Next, the Cochran's C test showed good control of Type 1 Error rates only for the balanced design but not the imbalanced design.

Table 1: Type 1 Error Rates for hypothesis testing based on homogeneity of variance tests.

Sample Size	Variance pairing	Test	Normal	Chi-Square	Outlier -1 group	Outlier - 2 groups
10,10	1:1	*Fisher*	0.0526 **	0.1796	0.9999	0
		Bartlett	0.0526 **	0.1796	0.9999	0
		Levene	0.0617	0.1282	0.0038	0
		B.Forsythe	0.0382	0.0509	0	0
		Cochran	0.0526 **	0.1796	0.9999	0
		Mood	N	0.0496 **	0.0496 **	0.0168
		A.Bradley	NA	0.0393 *	0.0393 *	0.0194
		Fligner		0.0603 *	0.0378 *	0.0064
30,30	1:1	*Fisher*	0.0535 **	0.2137	1	0
		Bartlett	0.0534 **	0.2137	1	0
		Levene	0.0546 **	0.1141	0.9988	0
		B.Forsythe	0.0442 *	0.0463	0.1388	0
		Cochran	0.0535 **	0.2137	1	0
		Mood		0.0486 **	0.0961	0.0171
		A.Bradley	NA	0.0493 **	0.0832	0.0248
		Fligner		0.0703 *	0.1170	0.0043
50,50	1:1	*Fisher*	0.0498 **	0.2334	1	0
		Bartlett	0.0498 **	0.2334	1	0
		Levene	0.0520 **	0.1172	1	0
		B.Forsythe	0.0457 **	0.0491 **	0.9084	0
		Cochran	0.0498 **	0.2334	1	0
		Mood		0.0523 **	0.1422	0.0191
		A.Bradley	NA	0.0520 **	0.1089	0.0275 *
		Fligner		0.0750 *	0.2145	0.0058
10,50	1:1	*Fisher*	0.0476 **	0.1886	1	0
		Bartlett	0.0480 **	0.1901	1	0
		Levene	0.0475 **	0.1116	1	0
		B.Forsythe	0.0410 *	0.0433 *	0.5706	0
		Cochran	0.1659	0.3644	1	0
		Mood		0.0507 **	0.0789	0.0177
		A.Bradley	NA	0.0514 **	0.0607 *	0.023
		Fligner		0.0685 *	0.0627 *	0.0039
50,10	1:1	*Fisher*	0.0480 **	0.1791	1	0
		Bartlett	0.0491 **	0.1838	1	0
		Levene	0.0513 **	0.1039	0.00112	0
		B.Forsythe	0.0417 *	0.0408 *	0	0
		Cochran	0.1638	0.3555	1	0
		Mood		0.0498 **	0.0503 **	0.0162
		A.Bradley	NA	0.0505 **	0.0575 **	0.0250 *
		Fligner		0.0662 *	0.0726 *	0.0042

4.3. *Chi-square distribution,* χ^2

The Brown-Forsythe test showed good control of Type 1 Error rates for the balanced group on the chi-square distribution. However, for the imbalanced group, the test remained robust with liberal criteria. This statistical test was seen to be robust compared to other parametric statistical tests on the chi-square distribution probably because this test used a more robust measure of central tendency, i.e. median compared to mean. In addition, the Mood test was able to control Type 1 Error rates very well for all sample size conditions. As for the Ansari-Bradley test, only for a balanced group with a sample size of 10, the test was robust with liberal criteria. In other cases, the test was found to enter the interval of stringent robust criteria. Next, the Fligner-Killeen test was robust when Type 1 Error rates were within Bradley liberal criteria interval for all sample size conditions.

4.4. *Outliers in one group*

When normal data were contaminated with the presence of 10% outliers in one group, the Mood test was found to be able to control the Type 1 Error rate well for a balanced group with the small sample size, n=10, and for an imbalanced group with 10% outliers value in a large sample size. In addition, the Ansari-Bradley test also showed good control of Type 1 Error at different sample sizes with its outliers were within the group with a larger sample size. However, the test was still robust with liberal criteria of the same sample size, n=10, and the sample size differed with the outliers were within the group with the smaller sample size. Next, the Fligner-Killeen test was robust with liberal criteria for the balanced group with the small sample size, n=10 and the imbalanced group.

4.5. *Outliers in two groups*

For contaminated data with 10% outliers in both groups, only the Ansari-Bradley statistical test was found robust for the balanced group with large sample size, n=50, and different sample sizes with outliers were within the group with a larger sample size.

4.6. *Power of a test*

The power of a statistical test is the probability of rejecting H_0 when H_0 is false. It means the power of a statistical test, $1-\beta$ is the probability of making the right decision. In this study, the power of a test was obtained when the variance for both groups is heterogeneous, and the values of power of a test are equal to or more than 80%, indicating the test of homogeneity of variance is good and appropriate for use. Tables 2, 3, and 4 show the complete results for the power of statistical tests obtained through the homogeneity of variance hypothesis testing using the Fisher, Bartlett, Cochran's C, Levene, Brown-Forsythe, Ansari-Bradley, Mood, and Fligner-Killeen modified tests.

For normal distribution data with a variance ratio of 1:2, the values of power of a test for all equality of variance parametric statistical test were more than 0.80 when the sample sizes are the same, n=30 and n=50. In addition, only the Cochran's C test had a high power of a test compared to other parametric tests on the normal distribution for positive and negative variance pairs. Whereas, for a variance ratio of 1:3, only the power of a test for the Fisher, Bartlett, and Cochran's C tests had a rate of more than 0.80 for the balanced group with a sample size of 10. While for equally large sample sizes and positive and negative variance pairs, all parametric statistical tests showed a high power of a test.

For a variance ratio of 1:4, all parametric statistical tests showed the power of statistical tests exceeding 0.80 for all sample size conditions. It indicates that all the tests can detect differences in variance in both samples when the effect size is large. Thus, it can be concluded that the effect size can affect the power of a test.

For Chi-square distribution with a variance ratio of 1:2, all equality of variance tests showed a small or less than 0.80 power of statistical tests for all sample size conditions. Whereas, for data with a variance ratio of 1:3 and the same sample size n=50, all parametric statistical tests and the nonparametric Fligner-Killeen statistical test had a power of a test exceeding 0.80. However, for the positive variance pair, only the Mood and Ansari-Bradley tests showed a high power of a test, exceeding 0.80.

Next, for a variance ratio of 1:4 with the same sample sizes *n*=30 and *n*=50, all parametric tests of equality of variance and the Fligner-Killeen test showed the power of a test exceeding 0.80. On the positive and negative variance pairs, only the power of a test for the Cochran's C, Mood, and Ansari- Bradley tests were found greater than 0.80.

Table 2: Power of statistical tests for hypothesis testing based on homogeneity of variance test with a variance ratio of 1:2.

Sample Size	Test	Normal	Chi-Square	Outlier -1 group	Outlier – 2 groups
10,10	*Fisher*	0.4962	0.2655	0.6684	0
	Bartlett	0.4960	0.2655	0.6681	0
	Levene	0.4196	0.2148	0	0
	B.Forsythe	0.3092	0.1214	0	0
	Cochran	0.4962	0.2655	0.6684	0
	Mood	NA	0.0164	0.0653	0.1292
	A.Bradley		0.0158	0.0874	0.1335
	Fligner		0.1348	0.0239	0.0621
30,30	*Fisher*	0.9531	0.4932	0.9987	0.0002
	Bartlett	0.9531	0.4932	0.9987	0.0002
	Levene	0.9139	0.4219	0.0495	0
	B.Forsythe	0.8966	0.3401	0	0
	Cochran	0.9531	0.4932	0.9987	0.0002
	Mood	NA	0.0197	0.3666	0.5879
	A.Bradley		0.0266	0.4013	0.5763
	Fligner		0.3870	0.1332	0.4398
50,50	*Fisher*	0.9974	0.6398	1	0.0019
	Bartlett	0.9974	0.6398	1	0.0019
	Levene	0.9916	0.5930	0.31	0
	B.Forsythe	0.9903	0.5368	0.0002	0.0001
	Cochran	0.9974	0.6398	1	0.0019
	Mood	NA	0.0242	0.6368	0.8581
	A.Bradley		0.0339	0.6601	0.8354
	Fligner		0.6108	0.2915	0.7825
10,50	*Fisher*	0.6572	0.3263	0.9961	0
	Bartlett	0.7068	0.3593	0.9929	0
	Levene	0.5937	0.2446	0.0939	0
	B.Forsythe	0.5738	0.1736	0.0008	0
	Cochran	0.9279	0.5880	0.9997	0.0007
	Mood	NA	0.3409	0.057	0.1776
	A.Bradley		0.2794	0.1781	0.2917
	Fligner		0.2466	0.0157	0.1326
50,10	*Fisher*	0.7966	0.3389	0.7299	0.0034
	Bartlett	0.7685	0.3119	0.7834	0.0014
	Levene	0.7349	0.2930	0.0001	0
	B.Forsythe	0.6819	0.2113	0	0
	Cochran	0.8531	0.4387	0.9675	0.0218
	Mood	NA	0.1469	0.2775	0.3960
	A.Bradley		0.1250	0.2555	0.3529
	Fligner		0.2016	0.0666	0.1774

Table 3: Power of statistical tests for hypothesis testing based on homogeneity of variance test with a variance ratio of 1:3.

Sample Size	Test	Normal	Chi-Square	Outlier -1 group	Outlier – 2 groups
10,10	*Fisher*	0.8746	0.4232	0.1428	0.0012
	Bartlett	0.8744	0.4230	0.1425	0.0012
	Levene	0.7594	0.3567	0	0.0003
	B.Forsythe	0.6361	0.2360	0	0.0001
	Cochran	0.8746	0.4232	0.1428	0.0012
	Mood	NA	0.0016	0.2012	0.3132
	A.Bradley		0.0020	0.2442	0.3136
	Fligner		0.2453	0.0982	0.1801
30,30	*Fisher*	0.9999	0.7679	0.4223	0.0359
	Bartlett	0.9999	0.7679	0.4223	0.0359
	Levene	0.9988	0.7350	0	0.0243
	B.Forsythe	0.9985	0.6891	0.0005	0.0536
	Cochran	0.9999	0.7679	0.4223	0.0359
	Mood	NA	0.0011	0.8556	0.9196
	A.Bradley		0.0034	0.8551	0.9110
	Fligner		0.7215	0.6188	0.8876
50,50	*Fisher*	1	0.9078	0.6807	0.1212
	Bartlett	1	0.9078	0.6807	0.1212
	Levene	1	0.9028	0.0001	0.1502
	B.Forsythe	1	0.8929	0.0008	0.3441
	Cochran	1	0.9078	0.6807	0.1212
	Mood	NA	0.0014	0.9882	0.9951
	A.Bradley		0.0050	0.9844	0.9931
	Fligner		0.9203	0.9192	0.9959
10,50	*Fisher*	0.9866	0.5356	0.3009	0
	Bartlett	0.991	0.5711	0.2295	0
	Levene	0.9510	0.4573	0	0
	B.Forsythe	0.9449	0.3912	0	0.0011
	Cochran	0.9997	0.7771	0.4889	0.0223
	Mood	NA	0.9024	0.3045	0.5403
	A.Bradley		0.8416	0.5618	0.6701
	Fligner		0.4792	0.1222	0.4762
50,10	*Fisher*	0.9811	0.5880	0.1597	0.0795
	Bartlett	0.9773	0.5535	0.1938	0.0528
	Levene	0.9665	0.5410	0.0001	0.0176
	B.Forsythe	0.9517	0.4483	0.0001	0.0142
	Cochran	0.9870	0.6652	0.4813	0.166
	Mood	NA	0.7757	0.6336	0.7199
	A.Bradley		0.7328	0.5853	0.6679
	Fligner		0.4017	0.2866	0.4791

For data with 10% outliers in one group with a variance ratio of 1:2, the power of a test for the Fisher and Bartlett tests exceeded 0.80 for a balanced design consisting of sample sizes of 30 and 50 and for the positive variance pair. At the same time, the Cochran C parametric statistical test showed a high power of a test exceeding 80% for a balanced design with *n*=30 and positive and negative variance pairs. Whereas, for data with a variance ratio of 1:3 and a balanced group with large sample sizes of *n*=30 and *n*=50, only the Mood and Ansari-Bradley nonparametric statistical tests showed the power of a test exceeding 0.80. However, for a balanced group with a sample size of 50, the Fligner-Killeen test also had a large power of a test exceeding 0.80.

Table 4: Power of statistical tests for hypothesis testing based on homogeneity of variance test with a variance ratio of 1:4.

Sample Size	Test	Normal	Chi-Square	Outlier -1 group	Outlier – 2 groups
10,10	*Fisher*	0.9738	0.5446	0.0255	0.0143
	Bartlett	0.9738	0.5444	0.0253	0.0142
	Levene	0.8951	0.4577	0.0036	0.0163
	B.Forsythe	0.8124	0.3303	0.0041	0.0142
	Cochran	0.9738	0.5446	0.0255	0.0143
	Mood	NA	0.0001	0.3279	0.4495
	A.Bradley		0.0001	0.3813	0.4474
	Fligner		0.3365	0.1869	0.3062
30,30	*Fisher*	1	0.9015	0.0195	0.2801
	Bartlett	1	0.9015	0.0194	0.2801
	Levene	1	0.8915	0.0307	0.3835
	B.Forsythe	1	0.8732	0.0899	0.5391
	Cochran	1	0.9015	0.0195	0.2801
	Mood	NA	0	0.9689	0.9776
	A.Bradley		0	0.9662	0.9761
	Fligner		0.8823	0.8893	0.9814
50,50	*Fisher*	1	0.9803	0.0194	0.6008
	Bartlett	1	0.9803	0.0194	0.6008
	Levene	1	0.9810	0.0910	0.8239
	B.Forsythe	1	0.9810	0.3112	0.9487
	Cochran	1	0.9803	0.0194	0.6008
	Mood	NA	0	0.9994	0.9993
	A.Bradley		0.0001	0.9994	0.9995
	Fligner		0.9850	0.9956	0.9999
10,50	*Fisher*	0.9999	0.6757	0.0035	0.0012
	Bartlett	1	0.7108	0.0013	0.0031
	Levene	0.9960	0.6115	0.0016	0.0200
	B.Forsythe	0.9948	0.5644	0.0217	0.0741
	Cochran	1	0.8788	0.0122	0.1989
	Mood	NA	0.9975	0.6007	0.7565
	A.Bradley		0.9913	0.8153	0.8389
	Fligner		0.6417	0.3408	0.7659
50,10	*Fisher*	0.9975	0.7495	0.0357	0.2660
	Bartlett	0.9969	0.7235	0.0407	0.2195
	Levene	0.9944	0.7089	0.0195	0.1758
	B.Forsythe	0.9928	0.6381	0.0197	0.1611
	Cochran	0.9983	0.8051	0.1482	0.3918
	Mood	NA	0.9916	0.8036	0.8494
	A.Bradley		0.9878	0.7656	0.8100
	Fligner		0.5780	0.4925	0.6748

Furthermore, for data with a variance ratio of 1:4 and the balanced group with *n*=30 and *n*=50, all nonparametric statistical tests showed a high power of a test than parametric tests. However, only the power of a test for the Ansari-Bradley test for the positive variance pair exceeded 80%. For the negative variance pair, only the Mood nonparametric test had the power of a test exceeding 0.80.

For contaminated data with 10% outliers in both groups with a ratio of 1:2, the Mood and Ansari-Bradley tests showed the test of a power exceeding 0.80 for the same sample size *n*=50. For a variance ratio of 1:3, the power of a test for all nonparametric tests for the balanced group with sample sizes of 30 and 50 exceeded 0.80.

For a variance ratio of 1:4 with the same sample sizes of *n*=30 and *n*=50, the nonparametric statistical tests showed the power of a test exceeding 80%. At the same time, the power of a test for the Levene and Brown-Forsythe parametric statistical tests exceeded 0.80 with a sample size equal to *n*=50. For the positive variance pair, only the Ansari-Bradley test showed the power of a test exceeding 0.80. As for the negative variance pair, the Mood and Ansari-Bradley tests showed the power of a test exceeding 80% compared to other statistical tests.

5. Conclusion

The comparison can determine the best homogeneity of variance test results for a given case carried out on normally and chi-square distributed data and data contaminated with outliers. The equality of variance parametric statistical tests used in this study were the Fisher, Bartlett, Levene, Brown-Forsythe, and Cochran's C tests. While nonparametric tests used were the Mood, Ansari-Bradley, and Fligner-Killeen tests. For normal data, the comparison of tests carried out was between the parametric statistical tests themselves. While for Chi-square distributed data and data with outlier values, the comparison carried out was between parametric and nonparametric statistical tests. The comparison was carried out based on the Type 1 Error rates and the power of statistical tests. A test is robust and can control the Type 1 Error rate well when the Type 1 Error rate for the test is close to the nominal significance level of 0.05. Whereas for different variance values, a test is categorized as good and appropriate for use when the value of the power of a test exceeds the probability value of 0.80 (Peterman 1990; Cohen 1988).

A test is robust when the Type 1 Error is close to the significance level of 0.05. For normally distributed data with the same variance values, all parametric statistical tests of homogeneity of variance except the Cochran's C test are said to be robust for all sample size conditions. In comparison, the Cochran's C test is only robust for a balanced design condition. For chi-square distributed data and for all sample size conditions, all nonparametric statistical tests and the Brown-Forsythe parametric statistical test are said to be robust. It indicates that nonparametric statistical tests are more appropriate to test the homogeneity of variance for the Chi-square distributed data of two groups. Next, to test the equality of variance for contaminated data with 10% outliers values in one of the groups, the only robust and appropriate tests for all balanced group conditions *n*= 10, and in the imbalanced group are the Ansari-Bradley and Fligner-Killeen tests. In comparison, the Mood test is only found to be robust for the balanced group *n*=10. When the data used are contaminated in both groups, the Ansari-Bradley test is robust for the balanced group condition *n*=50 and the imbalanced group. It indicates that none of the tests is robust for all distribution conditions and sample sizes.

Moreover, for normal data for two groups with a variance ratio of 1:2 and the imbalanced group, the Cochran's C test is more appropriate for use than other parametric statistical tests. Whereas for large balanced sample sizes of *n*=30 and *n*=50, all power of a test for parametric statistical tests exceeds 80% but not for the balanced sample size of *n*=10. For the balanced sample size of *n*=10, with a variance ratio of 1:3, the power of a test for the Levene and

Brown-Forsythe tests are still found to be less than 0.80. Based on the study' results, it can be concluded that when normal data for two groups have different variance values in both groups, the power of a test for all parametric statistical tests is found to increase and exceed 0.80 to the extent some reaching a value of 1 when the variance for one group increases.

For Chi-square distribution, the power of a test for all homogeneity of variance tests is small for small effect sizes. As the variance ratio increases in one of the groups, the power of a test also increases. The power of parametric statistical tests exceeds 0.80 for the balanced sample size condition when the sample size and effect size are large. It is also similar to the Fligner-Killeen nonparametric statistical test. The Cochran's C test also detects the heterogeneity of variance for positive and negative variance pairs as the effect size increases. Meanwhile, the Mood and Ansari-Bradley are found to have the power of a test exceeding 80% for the positive variance pair with a variance ratio of 1:3 and positive and negative variance pairs with a variance ratio of 1:4.

For data with 10% outlier values in one of the groups for a variance ratio of 1:2, the power of a test for the Fisher and Bartlett tests exceeds 0.80 if the group is balanced with the large sample size and for the positive variance pair but the power of a test for those tests decreases as the effect size increases. Similar to the power of a test for the Cochran's C test, its value decreases as the effect size increases. For a variance ratio of 1:2, the power of a test for the Cochran's C test exceeds 80% for the balanced group with the large sample size and the positive and negative variance pairs. In another case with nonparametric statistical tests, the power of tests increases consistent with the effect size. For the variance ratios of 1:3 and 1:4, the power of nonparametric statistical tests exceeds 0.80 for the balanced group with large sample size. As for the positive and negative variance pairs and a variance ratio of 1:4, the power of a test for the Ansari-Bradley test is found to exceed 80% and higher than other statistical tests.

When the data are contaminated with 10% outlier values in both groups, the power of a test for all statistical tests consistently increases with the effect size. For a variance ratio of 1:2, the Mood and Ansari-Bradley tests have a power of a test exceeding 0.80 for the balanced group with n=50. However, when the variance for one of the groups increases, the power of a test for all nonparametric statistical tests exceeds 80% for the balanced group with large sample size. The power of a test for the Mood and Ansari-Bradley tests is also high, to some extent exceeding 0.80 for positive and negative variance pairs with a variance ratio of 1:4. In addition, for a balanced group with n=50 and a variance ratio of 1:4, the power of a test for the Levene and Brown-Forsythe parametric statistical tests exceeds 80%, apart from the nonparametric statistical tests.

In conclusion, none of the statistical tests is robust and appropriate for use for all conditions. Thus, to obtain the best homogeneity of variance test for a study, the conditions that need to be considered are data distribution, sample size, variance ratio, and the presence of outliers. Thus, with the comparison of homogeneity of variance hypothesis testing between parametric tests on normal data and the comparison between parametric and nonparametric tests on chi-square distribution data and outliers carried out, the determination of the best statistical test for all cases assumed in this study can be carried out. Thus, all three study's objectives have been carried out and answered.

Acknowledgments

This study is supported by the Research University Grant GGP-2020-023.

References

Agunbiade D.A. & Iyaniwura J.O. 2010. Estimation under multicollinearity: A comparative approach using Monte Carlo methods. *Journal of Mathematics and Statistics* **6**(2): 183-192.

Ahad N.A., Yin T.S., Othman A.R. & Yaacob C.R. 2011. Sensitivity of normality tests to non-normal data. *Sains Malaysiana* **40**(6): 637-641.

Alabi O.O., Ayinde K. & Olatayo T.O. 2008. Effect of multicollinearity on power rates of the ordinary least squares estimators. *Journal of Mathematics and Statistics* **4**(2): 75-80.

Anderson T.W. 1962. On the distribution of the two-sample Cramer-von Mises criterion. *Annals of Mathematical Statistics* **33**(3): 1148-1159.

Anderson T.W. & Darling D.A. 1952. Asymptotic theory of certain 'Goodness of Fit' criteria based on stochastic processes. *Annals of Mathematical Statistics* **23(2)**: 193–212.

Bartlett M.S. 1937. The statistical conception of mental factors. *British Journal of Psychology* **28**(1): 97-104.

Bradley J.V. 1978. Robustness?. *British Journal of Psychology* **31**(2): 144-152.

Brown M.B. & Forsythe A.B. 1974. Robust tests for the equality of variances. *Journal of the American Statistical Association* **69**(346): 364-367.

Carroll R.J. & Schneider H. 1985. A note on Levene's tests for equality of variances. *Statistics & Probability Letters* **3**(4): 191-194.

Chambers J.M., Cleveland W.S., Kleiner B. & Tukey P.A. 1983. *Graphical Methods for Data Analysis*. Boston: Duxbury Press.

Cochran W.G. 1941. The distribution of the largest of a set of estimated variances as a fraction of their total. *Annals of Eugenics* **11**(1): 47-52.

Cochran W.G. & Cox G.M. 1957. *Experimental Design*. NY: John Willey and Sons.

Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd Ed. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Conover W.J. & Iman R.L. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician* **35**(3): 124-129.

Conover W.J., Johnson M.E. & Johnson M.M. 1981. A comparative study for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics* **23**(4): 351-361.

Conover W.J., Guerrero-Serrano A.J. & Tercero-Gómez V.G. 2018. An update on 'A comparative study of tests for homogeneity of variance'. *Journal of Statistical Computation and Simulation* **88**(8): 1454-1469.

Ercheg-Hurn D.M. & Mirosevich V.M. 2008. Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist* **63**(7): 591-601.

Fligner M.A. & Killeen T.J. 1976. Distribution-free two-sample tests for scale. *Journal of the American Statistical Association* **71**(353): 210-213.

Gastwirth J.L., Gel Y.R. & Miao W. 2009. The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science* **24**(3): 343–360.

Ghasemi A. & Zahediasl S. 2012. Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology Metabolism* **10**(2): 486-489.

Gorbunova A.A. & Lemeshko B.Y. 2012. Application of parametric homogeneity of variances tests under violation of classical assumption. *Proceedings of the 2nd Stochastic Modeling Techniques and Data Analysis International Conference*, pp. 253–260.

Gravetter F.J. & Wallnau L.B. 2000. *Statistics for the Behavioral Sciences*. 5th Ed. Belmont: Wadsworth – Thomson Learning.

Hartley H.O. 1950. The use of range in analysis of variance. *Biometrika* **37**(3/4): 271-280.

Kim Y.J. & Cribbie R.A. 2018. ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. *British Journal of Mathematical and Statistical Psychology* **71**(1): 1-12.

Klotz S. & Johnson N.L. 1993. *Breakthroughs in Statistics: Volume 1: Foundations and Basic Theory*. NY: Springer.

Kolmogorov A.N. 1956. *Foundations of the Theory of Probability*. 2nd English Eds. NY: Chelsea Publishing Company.

Lee H.B., Katz G.S. & Restori A.F. 2010. A Monte Carlo study of seven homogeneity of variance tests. *Journal of Mathematics and Statistics* **6**(3): 359-366.

Legendre P. & Borcard D. 2000. Statistical comparison of univariate tests of homogeneity of variances. *Unpublished.*

Lemeshko B.Y., Lemeshko S.B. & Gorbunova A.A. 2010a. Application and power of criteria for testing the homogeneity of variances. Part I. Parametric criteria. *Measurement Techniques* **53**(3): 237-246.

Lemeshko B.Y., Lemeshko S.B. & Gorbunova A.A. 2010b. Application and power of criteria for testing the homogeneity of variances. Part II. Nonparametric criteria. *Measurement Techniques* **53**(5): 476-486.

Levene H. 1960. Robust tests for equality of variances. In Olkin I. (ed.). *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. CA: Stanford University Press.

Mazahreh A.S., Hammad H. & Abu-Jaber H. 2009. The attitudes of instructors and faculty members about the quality of technical education programs in community colleges in Jordan. *Journal of Social Sciences* **5**(4): 401-407.

Mendes M., Ozcaya Turhan N. & Gurbuz F. 2006. A new alternative in testing for homogeneity of variances. *Journal of Statistical Research* **40**(2): 65-83.

Nordstokke D.W. & Zumbo B.D. 2010. A new nonparametric Levene test for equal variances. *Psicológica* **31**(2): 401-430.

O'Brien R.G. 1978. Robust techniques for testing heterogeneity of variance effects in factorial designs. *Psychometrika* **43**(3): 327-342.

O'Brien R.G. 1981. A simple test for variance effects in experimental designs. *Psychological Bulletin* **89**(3): 570-574.

Oladejo N. & Adetunde I. 2009. A numerical test on the Riemann hypothesis with applications. *Journal of Mathematics and Statistics* **5**(1): 47-53.

Overall J.E. & Woodward J.A. 1976. A robust and powerful test for heterogeneity of variance. University of Texas Medical Branch Psychometric Laboratory.

Patil K.P. & Kulkarni H.V. 2022. An uniformly superior exact multi-sample test procedure for homogeneity of variances under location-scale family of distributions. *Journal of Statistical Computation and Simulation* **92(**18): 3931-3957.

Peterman R.M. 1990. The importance of reporting statistical power: The forest decline and acidic deposition example. *Ecology* **71**(5): 2024–2027.

Rana M.S., Midi H. & Imon A.R. 2008. A robust modification of the Goldfeld-Quandt test for the detection of heteroscedasticity in the presence of outliers. *Journal of Mathematics and Statistics* **4**(4): 277-283.

Samiuddin M. & Atiqullah M. 1976. A test for equality of variances. *Biometrika* **63**(1): 206-208.

Shapiro S. & Wilk M. 1965. An analysis of variance test for normality (complete samples). *Biometrika* **52**(3/4): 591-611.

Sharma D. & Kibria B.G. 2013. On some test statistics for testing homogeneity of variances: A comparative study. *Journal of Statistical Computation and Simulation* **83**(10): 1944-1963.

Smirnov S.S. 1935. Zur Copepodenfauna des Amur-Limans. K faune Copepoda Amurskogo limana. *Issledovaniya Morei SSSR, Explorations des Mers de l'URSS* **22**: 41-53.

Thomas L. & Juanes F. 1996. The importance of statistical power analysis. *Animal Behaviour* **52**(4): 856-859.

Tomarken A.J. & Serlin R.C. 1986. Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin* **99**(1): 90–99.

Underwood A.J. 1997. *Experiments in Ecology: Their Logical Design and Interpretation using Analysis of Variance*. Cambridge: Cambridge University Press.

Vorapongsathorn T., Taejaroenkul S. & Chukiat V. 2004. A comparison of type I error and power of Bartlett's test, Levene's test and Cochran's test under violation of assumptions. *Songklanakarin Journal of Science & Technology* **26**(4): 537-547.

Wang Y., Rodríguez de Gil P., Chen Y.H., Kromrey J.D., Kim E.S., Pham T., Nguyen D. & Romano J.L. 2017. Comparing the performance of approaches for testing the homogeneity of variance assumption in one-factor ANOVA models. *Educational and Psychological Measurement* **77**(2): 305-329.

Weerahandi S. 1995. ANOVA under unequal error variances. *Biometrics* **51**(2): 589– 599.

Yi Z., Chen Y.H., Yin Y., Cheng K., Wang Y., Nguyen D., Pham T. & Kim E.S. 2022. Brief research report: A comparison of robust tests for homogeneity of variance in factorial ANOVA. *The Journal of Experimental Education* **90**(2): 505-520.

Zar J.H. 1999. *Biostatistical Analysis*. 4th Ed. NJ: Prentice Hall.

Department of Mathematical Sciences
Faculty of Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi
Selangor DE, MALAYSIA
*E-mail: fazlinabdullah03@yahoo.com, noramuda@ukm.edu.my**

Received: 17 October 2022
Accepted: 2 November 2022

*Corresponding author