International Journal of Analysis and Applications

Efficiency Evaluation of Statistical Tests for Homogeneity of Variances under Normal, Beta, and Weibull Distributional Frameworks

Saowapa Chapitak[1], Jarukit Jaipetch[1], Kunita Wichayacheewin[1], Wasutida Suwanrach[1], Boonyarit Choopradit[2,*]

[1]Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand
[2]Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathumthani 12120, Thailand
**Corresponding author:* boonyarit@mathstat.sci.tu.ac.th

ABSTRACT. This research endeavor aims to evaluate six test statistics relevant to assessing of homogeneity of variance (HOV): Bartlett's (BL), Levene's (LV), modified Levene's (LVM), Klotz's (KL), Layard's (LY), and Samiuddin's (SMD). Simulated datasets were generated under the frameworks of normal, Beta, and Weibull distributions, encompassing both three and four groups, while incorporating variations in sample sizes that were both equal and unequal. Each experimental condition was replicated 5,000 times to ensure the precision of statistical outcomes. In the context of the normal distribution, the BL, LY, and SMD statistics exhibited strong control over Type I error rates, with the BL and LY statistics achieving the highest statistical power among the tests classified as acceptable. Whereas the LV and LVM statistics demonstrated competence in error control, they were characterized by reduced power; conversely, the SMD statistic exhibited significantly low power. In contrast, the KL statistic consistently yielded inflated error rates, rendering it inappropriate for practical application. In the realm of the Beta distribution, the KL, LVM, and LY statistics emerged as the most proficient performers, adeptly preserving Type I error rates. The KL statistic, notwithstanding its mediocre performance under normal distribution conditions, demonstrated the greatest resilience within this specific context. The LVM statistic maintained a conservative approach; the LY statistic exhibited precision yet was somewhat less robust when faced with skewed data, the LV statistic demonstrated moderate effectiveness, the BL statistic was excessively cautious, and the SMD statistic was classified as unreliable. In relation to the Weibull distribution, the LY, SMD, KL, and LVM statistics consistently controlled the Type I error rates. The BL statistic performed satisfactorily but exhibited a slight inclination towards inflation of Type I error rates, whereas the LV statistic was assessed as unreliable. The BL statistic attained the highest statistical power, albeit with correspondingly elevated Type I error rates. The LVM and LY statistics demonstrated considerable power across diverse scenarios, with the LY statistic being preferentially utilized for small to medium sample sizes and the LVM statistic for larger sample sizes. The SMD and KL statistics consistently ranked lowest in terms of empirical power.

Received Aug. 14, 2025

2020 Mathematics Subject Classification. 62E10, 62F03, 62G10.

Key words and phrases. normal; Beta; Weibull; homogeneity of variances; hypothesis testing.

https://doi.org/10.28924/2291-8639-23-2025-240

ISSN: 2291-8639

1. Introduction

In applying the F-test statistic for analysis of variance (ANOVA) to study the mean of more than two populations, the observed value is assumed to be a random sample from a population with a normal distribution, and it is assumed that treatment variances are equal (homogeneity of variance) [1]. The hypotheses of homogeneity of variances (HOV) are $H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ versus $H_1: \sigma_i^2 \neq \sigma_j^2$ for at least one pair (i, j). The null hypothesis assumes homogeneity of population variances across k groups. When this assumption is violated, the validity of the ANOVA results is compromised. In such cases, alternative statistical approaches—such as Welch's ANOVA, which adjusts for unequal variances, or non-parametric methods like the Kruskal-Wallis test—are considered more appropriate and statistically robust ([2], [3]). As a result, it is required to test for variance homogeneity before selecting the right methodologies for subsequent analysis. Scheffe [4] noted that the effect becomes greater when the sample sizes in each group differ. Violations of this assumption might lead to bias in the test's significance level (Type I error). For these reasons, various statistical tests can be used to examine variance homogeneity, including Levene's test and Bartlett's test, and Brown-Forsythe's test. Levene [5], Snedecor and Cochran [6], and Brown and Forsythe [7] are commonly used for checking the ANOVA assumptions. The choice of test depends on the nature of the data and the assumptions underlying each one. In general, Levene's test is commonly used and more robust to deviations from normality, whereas Bartlett's test is sensitive to such departures. The Brown-Forsythe test is preferred when the assumption of normality is violated.

Recently, numerous researchers have initiated studies to examine the effectiveness of various tests in various contexts ([8], [9]). Wang et al. [10] studied the performance of 14 tests by varying factors such as group size, variance ratios, and distribution shapes. The results indicated that many tests can control Type I error rates, providing practical guidelines for selecting appropriate variance tests. Conover et al. [11] update introduces tests for skewed and lognormal distributions. Three tests demonstrate superior power for skewed distributions and ease of application. Riansut [12] compared the performance of six test statistics —Bartlett, Levene, Brown-Forsythe, O'Brien, Klotz, and Mood —by simulating data under various conditions, including sample size and distributions. The results showed that Klotz's test and Mood's test are more robust to violations of normal distribution. Meanwhile, Bartlett's test is accurate when the data are normal but sensitive to skewness. Sinsomboonthong [13] investigated the performance of six tests for homogeneity of variance under distributions with high kurtosis and skewness. Results indicated that under a highly kurtotic normal distribution, Bartlett's, Levene's, Brown-Forsythe's, and O'Brien's tests adequately controlled Type I error rates. For the highly skewed gamma distribution, only Brown-Forsythe's and O'Brien's tests maintained control. Lehman's test exhibited the highest power under normality, while Bartlett's test was more effective with small

samples. Under the gamma distribution, Levene's and Bartlett's tests achieved the highest power, with Brown-Forsythe's test demonstrating the most consistent robustness. Soikliew and Araveeporn [14] compared the Type I error rates and power of Levene's, O'Brien's, and six modified tests across four distributions. The modified Levene's test using squared deviations based on the median performed best under normality. For logistic and gamma distributions, Levene's test using absolute deviations based on the trimmed mean showed the most consistent performance across all sample size conditions. Jiamwattanapong and Ingadapa [15] evaluated five tests through a simulation study, including Cochran's Q, Z-variance, O'Brien's F, Levene's test, and Modified Levene's test. Results revealed that Cochran's and Z-variance tests outperformed the other tests. Sritan and Phuenaree [16] compared type I error and the power of the test for five homogeneity of variance tests, which are Bartlett, Levene, Cochran, O'Brien, and Jackknife, under log-normal distributions. The results demonstrated that Levene's test works well for highly skewed distributions and Bartlett's test has good power. Zhou et al. [17] reviewed various techniques for determining the homogeneity of variation among groups and studied them under normal, heavy-tailed, and skewed normal data. The simulation demonstrated that the Jackknife and Cochran's tests are highly effective. These comparisons have been conducted under various distributional conditions to evaluate the efficiency and comparative performance of test statistics for assessing the equality of variances.

Although considerable research has been conducted on tests for variance homogeneity, there remains a notable gap in the comparative assessment of test power efficiency under normal, Beta and Weibull distributions when considering six specific test statistics: Bartlett's (BL), Levene's (LV), Modified Levene's (LVM), Klotz's (KL), Layard's (LY), and Samiuddin's (SMD). Consequently, our focus is on addressing this gap. The objective of this study is to evaluate and compare the performance of these six widely used tests, with a focus on their ability to maintain Type I error rates and maximize statistical power in assessing variance homogeneity.

The organization of this paper is as follows. Section 2 presents a detailed explanation of the six tests for homogeneity of variances. Section 3 presents the research methodology through the simulation. Section 4 reports the results regarding the empirical Type I error rate, robustness, and empirical power obtained from the tests. Finally, Section 5 provides the conclusion of the study.

2. Six Tests for Homogeneity of Variances

Let y_{ij} is the j-th observation from the group i-th sample, $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$, where k and n_i represent the number of groups and the sample size of the i-th group, respectively. Let the total sample size be $n = \sum_{i=1}^{k} n_i$. In this study, the six test statistics of the homogeneity of variances, including Bartlett's test (BL), Levene's test (LV), Modified Levene's test (LVM), Klotz's test (KL), Layard's test (LY), and Samiuddin's test (SMD), are detailed as follows.

2.1 Bartlett's Test

Bartlett's test statistic (BL) ([6], [18]) is calculated by Equation (1).

$$BL = \frac{(n-k)ln S_p^2 - \sum_{i=1}^{k} k(n_i-1)ln S_i^2}{1+\frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{n_i-1}-\frac{1}{n-k}\right]} \quad (1)$$

where S_i^2 is the i-th group sample variance calculated by $S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$, S_p^2 is the sample pooled variance calculated by $S_p^2 = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)\, S_i^2$, and $\bar{y}_{i.}$ is the i-th group sample mean calculated by $\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$. Under the assumption that each group's data is normally distributed, the BL statistic follows an approximate chi-square distribution with $k-1$ degrees of freedom.

2.2 Levene's Test

Levene's test statistic (LV) ([5], [19]) is expressed as Equation (2).

$$LV = \frac{\sum_{i=1}^{k} n_i(\bar{Z}_{i.}-\bar{Z}_{..})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Z_{ij}-\bar{Z}_{i.}\right)^2/(n-k)} \quad (2)$$

where Z_{ij} stands for the new variables resulting from the transformation; $Z_{ij} = \left|y_{ij} - \bar{y}_{i.}\right|$, $\bar{Z}_{i.}$ is the i-th group sample mean, and $\bar{Z}_{..}$ is the overall mean of the new variables Z_{ij}. Under H_0, the LV statistic follows an F-distribution with k - 1 and n – k degrees of freedom, respectively.

2.3 Modified Levene's Test

The Modified Levene's test statistic (LVM) was proposed by Brown and Forsythe [7], who extended Levene's test by incorporating the median or trimmed mean in addition to the mean. It is expressed in Equation (3).

$$LVM = \frac{\sum_{i=1}^{k} n_i(\bar{Z}_{i.}-\bar{Z}_{..})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Z_{ij}-\bar{Z}_{i.}\right)^2/(n-k)} \quad (3)$$

where Z_{ij} can have one of the following two definitions: (i) $Z_{ij} = \left|y_{ij} - \tilde{y}_{i.}\right|$, for $\tilde{y}_{i.}$ is the median of the i-th group and (ii) $Z_{ij} = \left|y_{ij} - \bar{y}'_{i.}\right|$, for $\bar{y}'_{i.}$ is the 10% trimmed mean of the i-th group. Under H_0, the LVM statistic follows an F-distribution with k - 1 and n– k degrees of freedom, respectively. In this study, the LVM test statistic was calculated using the median.

2.4 Klotz's Test

Klotz [20] proposed a rank-based test as a nonparametric alternative for detecting differences among several populations, and its test statistic (KL) has been widely applied in k-sample problems ([20], [21], [22]). It is given in Equation (4).

$$KL = \frac{1}{\sum_{i=1}^{n} A_i^4 - \frac{1}{n}\left(\sum_{i=1}^{n} A_i^2\right)^2}\sum_{j=1}^{k} n_j\left(\bar{A}_{j.}^2 - \bar{A}_{..}^2\right)^2 \quad (4)$$

where n_j is the sample size of the j-th group, $A_i = \Phi^{-1}(R_i/(n+1))$ is the normal score for the i-th observation with rank R_i in a combined sample of size n, Φ^{-1} is the quantile function (inverse CDF) of the standard normal distribution, $\bar{A}_{j.}^2 = \frac{1}{n_j}\sum_{i=1}^{n_j} A_{ij}^2$ is the mean of the squared normal

scores for the j-th group, $\bar{A}_{..}^2 = \frac{1}{n}\sum_{i=1}^{n} A_i^2$ is the overall mean of the squared normal scores for all observations. Under H_0, the test statistic KL approximately follows a chi-square distribution with $k-1$ degrees of freedom.

2.5 Layard's Test

Layard's test [23] for homogeneity of variances is a robust large-sample test that accounts for the kurtosis of the underlying distributions. It's designed to be less sensitive to departures from normality than, for example, Bartlett's test. It is expressed as Equation (5).

$$LY = \frac{1}{\hat{\tau}^2}\sum_{i=1}^{k}(n_i - 1)\left(lnS_i^2 - \frac{\sum_{i=1}^{k}(n_i-1)lnS_i^2}{\sum_{i=1}^{k}(n_i-1)}\right)^2 \quad (5)$$

where $\hat{\tau}^2 = 2 + \left(1 - \frac{1}{\bar{n}}\right)\hat{\gamma}$, when $\bar{n} = \frac{1}{k}\sum_{i=1}^{k} n_i$ or sometimes $\bar{n} = n$, and $\hat{\gamma}$ is an estimate of the common population kurtosis. The test statistic LY is asymptotically distributed as a chi-square distribution with $k-1$ degrees of freedom.

2.6 Samiuddin's Test

Samiuddin's test statistic (SMD), as proposed by [24], aims to provide a more robust alternative to classic tests, such as Bartlett's test, especially when the underlying data distributions are not perfectly normal. Its robustness stems from a transformation applied to the sample variances. It is given by Equation (6).

$$SMD = \sum_{i=1}^{k}\frac{(m_i - m)^2}{a_i^2} \quad (6)$$

where $m_i = \left(1 - \frac{2}{9(n_i-1)}\right)\left(S_i^2\right)^{-\frac{1}{3}}$, $a_i^2 = \frac{2}{9(n_i-1)\left(S_i^2\right)^{\frac{2}{3}}}$ and $m = \frac{\sum_{i=1}^{k}\frac{m_i}{a_i^2}}{\sum_{i=1}^{k}\frac{1}{a_i^2}}$.

The test statistic SMD follows a chi-square distribution with $k-1$ degrees of freedom.

3. Methodology

This study aims to identify the tests that best control the Type I error rate and their robustness while maintaining high statistical power, as discussed in [25], across various distributions and sample sizes. The research methodology is outlined as follows:

3.1 Data Simulation.

This section delineates the methodological framework employed to comparatively evaluate the efficiency of six statistical tests for assessing the homogeneity of variances under varying distributional assumptions. Specifically, the study focuses on normal, Beta, and Weibull distributions. Monte Carlo simulation techniques were utilized to generate data for each scenario, with 5,000 replications performed per condition to ensure statistical reliability and robustness. All simulations were performed using R version 4.1.2. The study populations were generated from three different distributions: normal, Beta, and Weibull. The number of population groups was set to three and four, respectively. This study adopted the non-centrality parameter (ϕ) as a

criterion for assessing the degree of variance differences among populations [26]. It is defined as: $\phi = \sqrt{\frac{\sum_{i=1}^{k}(\sigma_i^2 - \bar{\sigma}^2)^2/k}{\sigma_{(1)}^2}}$, where $\bar{\sigma}^2$ is the average of k population variances, and $\sigma_{(1)}^2$ is the minimum of k population variances. Four levels are defined as follows: $\phi = 0$, indicating homogeneity of variances (no difference); $0 < \phi < 1.5$ representing a low variance ratio; $1.5 \leq \phi < 3.0$ representing a moderate variance ratio, and $\phi \geq 3.0$ representing a high variance ratio. In this study, the population variance ratios were defined as shown in Table 1. A nominal significance level (α) of 0.05 was used for hypothesis testing across all scenarios.

3.2 Empirical Type I Error Rate and Robustness

The empirical Type I error rate is estimated based on the following steps:

Step 1. The populations are simulated to follow identical distributions, namely the normal, Beta, and Weibull distributions, under the assumption of the null hypothesis (H_0), which states that all population variances are equal. Under this null condition, the non-centrality parameter is set to zero, that is $\phi = 0$.

Table 1: Variance ratio

k	Variance Ratio	ϕ	Degree of Variance Differences
3	1.0: 1.0: 1.0	0	no difference
	1.0: 2.0: 2.0	0.4714	low
	1.0: 2.0: 3.0	0.8165	low
	1.0: 1.0: 5.0	1.8856	moderate
	1.0: 4.0: 7.0	2.4495	moderate
	1.0: 8.0: 8.0	3.2998	high
	1.0: 8.0: 15.0	5.7155	high
4	1.0: 1.0: 1.0: 1.0	0	no difference
	1.0: 1.5: 1.5: 2.0	0.3536	low
	1.0: 2.0: 3.0: 4.0	1.1180	low
	1.0: 3.0: 3.0: 6.0	1.7854	moderate
	1.0: 3.0: 5.0: 7.0	2.2361	moderate
	1.0: 4.0: 4.0: 10.0	3.2692	high
	1.0: 6.0: 11.0: 16.0	5.5902	high

The simulation settings are as follows: for the normal distribution, data are generated with zero mean and unit variance, that is $Y \sim N(0,1)$. For the Beta distribution, data are drawn from a Beta distribution with shape and scale parameters equal to 1, that is $Y \sim Beta(1,1)$. Finally, for the Weibull distribution, data are generated from a Weibull distribution with shape and scale parameters equal to 2, that is $Y \sim Weibull(2,2)$.

Step 2. Draw 5,000 random samples from populations with small, medium, and large sample sizes, both equal and unequal across groups, as shown in Table 2.

Step 3. All six test statistics are calculated for each replication and compared with their respective critical values, and the number of correct rejections of H_0 is recorded.

Step 4. Empirical Type I error rate is calculated as the proportion of replications in which the null hypothesis is correctly rejected when it is true, and the number of rejections is recorded.

Empirical Type I error rate = (number of times to reject $H_0 | H_0$ is true) / 5,000.

A test is considered to control the Type I error well if its empirical Type I error rate is close to the nominal level of 0.05, or falls within Bradley's acceptable range of [0.025–0.075], as proposed by [27]. The robustness of a test statistic is evaluated based on its ability to control the Type I error rate close to the nominal significance level of 0.05, specifically within Cochran's acceptable range of 0.04 to 0.06 [28]. To make more precise comparisons, the average relative error (ARE) values are calculated for each statistic. The ARE values indicate the deviation of the test statistic from the nominal significance level (α) and are calculated as given in Equation (7).

$$ARE = \frac{100}{M}\sum_{i=1}^{M}\frac{|\hat{\alpha}_i-\alpha|}{\alpha} \quad (7)$$

where M is the number of empirical Type I error rates calculated for each statistic in the table, $\hat{\alpha}_i$ is the i-th Type I error rate. It can be stated that the test statistic yielding the smallest ARE value demonstrates superior performance with respect to controlling the Type I error rate.

3.3 Empirical Power

The power of the test is the probability of rejecting false hypotheses. The test will be called the best method when it shows the highest power. However, we take into account only those tests that are able to control the Type I error rate. The empirical power of each test is estimated through the following steps:

Step 1. Simulate data under the alternative hypothesis (H_1), where the non-centrality parameter is set to a value greater than zero, that is $\phi > 0$, as given in Table 1.

Step 2. Random 5,000 samples from populations with equal and unequal sample sizes as given in Table 2.

Step 3. All six test statistics are calculated for each replication and compared with their respective critical values, and the number of correct rejections of H_0 is recorded.

Step 4. Empirical power is computed as the proportion of replications in which the null hypothesis is correctly rejected when it is false.

Empirical power = (number of times to reject $H_0 | H_1$ is true) / 5,000.

Table 2: Sample sizes

k	Equality of Sample Sizes	Sample Sizes		
		Small	Medium	Large
3	Equal	(10, 10, 10)	(30, 30, 30)	(90, 90, 90)
	Unequal	(2, 6, 10)	(20, 25, 30)	(70, 80, 90)
4	Equal	(10, 10, 10, 10)	(30, 30, 30, 30)	(90, 90, 90, 90)
	Unequal	(2, 6, 10, 14)	(20, 25, 30, 35)	(70, 80, 90, 100)

4. Results and Discussions

A simulation study was conducted to evaluate the performance of six test statistics—BL, LV, LVM, KL, LY, and SMD—concerning their ability to control the Type I error rate and to detect true differences (power) when data are generated from normal, Beta, and Weibull distributions. The assessment was carried out under various conditions, including both equal and unequal sample sizes, and across three and four populations. The empirical Type I error rates and power estimates obtained from the simulation are presented and discussed in the following sections.

4.1 Empirical Type I Error Rate and Robustness Results

The empirical Type I error rates for the six test statistics used to assess the homogeneity of population variances at the nominal significance level 0.05 under scenarios involving three and four populations with normal, Beta, and Weibull distributions are presented in Tables 3 to 5 and Figures 1 – 3, respectively.

The results presented in Table 3 and Figure 1 offer a comparative assessment of six test statistics—BL, LV, LVM, KL, LY, and SMD—regarding their ability to control the Type I error rate when the data followed a normal distribution with a common variance of one. Among these, the BL, LY, and SMD statistics demonstrated the most satisfactory performance, with average empirical Type I error rates of 0.0565, 0.0511, and 0.0486, respectively. These values were close to the nominal level of 0.05 and also consistently fell within Bradley's acceptable range (0.025–0.075) across various scenarios, including both equal and unequal sample sizes and differing numbers of populations. Together with the relatively low ARE values, which were 0.2167, 0.45, and 0.1983, respectively. While the LVM statistic produced the minimal average Type I error rate of 0.0416, it exhibited a relatively conservative stance across all scenarios, with an ARE value of 0.955. This guarantees rigorous control of the Type I error rate; however, it may potentially diminish statistical power in real-world applications. Consequently, notwithstanding its accuracy, the LVM might not be regarded as the most resilient overall, particularly in situations where the preservation of power is of paramount importance. The LV statistic demonstrated moderate efficacy, with mean empirical Type I error rates of 0.0649 and an ARE value of 0.8867; although the mean empirical Type I error rates generally adhered to Bradley's criterion under most conditions, they surpassed the acceptable limit in cases characterized by unequal and minimal sample sizes—such as (2,6,10) and (2,6,10,14)—which raises significant concerns regarding its dependability in such circumstances. The KL statistic persistently exhibited an inability to regulate the Type I error, manifesting an average empirical rate of 0.2076 and transgressions of Bradley's criterion in each scenario investigated, as denoted by asterisks in the corresponding table, accompanied by an ARE value of 22.4783, which is exceedingly elevated. This inflation indicates that KL is overly sensitive and prone to false positives, making it unsuitable for use under normality assumptions. Furthermore, an evaluation of robustness under the normal

distribution reveals that the BL, LY, and SMD test statistics exhibited strong robustness, as evidenced by their empirical Type I error rates falling within Cochran's acceptable range. Conversely, the LV, LVM, and KL statistics fell to meet this criterion, indicating a lack of robustness in controlling the Type I error rate under this distribution.

The results presented in Table 4 and Figure 2 offer a comparative assessment of six test statistics—BL, LV, LVM, KL, LY, and SMD—regarding their ability to control the Type I error rate when the data followed the Beta(1,1) distribution, equivalent to a uniform distribution over the interval [0,1]. Among these, the KL, LVM, and LY statistics demonstrated the most satisfactory performance, with average empirical Type I error rates of 0.0440, 0.0356, 0.0354, and 0.0440, respectively. These values were not only close to the nominal level of 0.05 but also consistently fell within Bradley's acceptable range across all scenarios, including both equal and unequal sample sizes and different numbers of populations. These empirical results indicate that the KL, LVM, and LY ensure a stable and precise control over the Type I error rate when the underlying data is characterized by the Beta distribution, exhibiting low ARE values of 0.5983, 1.4433, and 1.4617, respectively. Notably, the KL statistic, which displayed suboptimal performance under the normal distribution framework, emerged as one of the most resilient and dependable tests when applied to the Beta distribution context. These findings imply that the KL statistic is particularly well-adapted for symmetric and bounded distributional forms, exemplified by the uniform distribution denoted as Beta(1,1), and it surpasses a number of alternative statistical measures that otherwise demonstrate robust performance under the assumption of normality. Although the LVM statistic yielded slightly lower average empirical Type I error rates than the nominal level, it remained within acceptable bounds and showed no violations of Bradley's criterion. Its conservative nature may imply a trade-off with statistical power, yet its reliability in Type I error rate control under the Beta distribution is evident. The LY statistic maintained empirical Type I error rates that were consistently close to the nominal level of 0.05 and fell within Bradley's acceptable range in all conditions. Nevertheless, while its performance was stable and acceptable across all scenarios, its robustness in the strict sense—defined as the ability to maintain Type I error rate control under a wide range of non-normal distributions—was not as strong as that of the KL statistic. In particular, although LY performed well under the normal and Beta distributions, its error control slightly deteriorated under the Weibull distribution, where moderate skewness was present. Therefore, LY may be considered a reliable statistic with good Type I error rate control, but not the most robust when assessed under varying distributional assumptions. The LV statistic exhibited a moderate level of efficacy; although its overall mean empirical Type I error rate of 0.0631 (ARE value of 1.34) fell within an acceptable range, it surpassed the upper limit of Bradley's criterion under various conditions, particularly in scenarios characterized by unequal and minimal sample sizes, such as the configurations (2,6,10)

and (2,6,10,14). These findings suggest that the LV statistic is particularly sensitive to variations in sample size and may lack robustness when applied to such experimental designs. In contrast, the BL statistic exhibits an exceptionally low average Type I error rate of 0.0055, yet it demonstrates an excessively conservative tendency across nearly all conditions, infringing upon Bradley's lower bound in the preponderance of scenarios. This inherent conservativeness may result in an insufficient rejection of the null hypothesis, consequently leading to a diminishment in statistical power, thereby rendering it less advantageous in practical applications. Furthermore, it achieved a considerable ARE value of 4.45, indicative of suboptimal performance. The SMD statistic demonstrated the poorest performance in the overall analysis. It yielded the highest mean Type I error rate of 0.1289, with values surpassing the Bradley's criterion across all scenarios investigated, exhibiting an ARE value of 7.89, thereby indicating suboptimal performance relative to the other statistical tests. This persistent inflation signifies a profound deficiency in Type I error regulation, consequently rendering the SMD statistic inadequate when applied to Beta-distributed data.

Table 5 and Figure 3 present the empirical Type I error rates of six test statistics—BL, LV, LVM, KL, LY, and SMD—under the assumption that the data were generated from the Weibull distribution. The results demonstrated that four statistical metrics—LVM, KL, LY, and SMD—manifested the highest degree of reliability in regulating the Type I error rate, consistently adhering to Bradley's acceptable criteria across all experimental conditions. All of these metrics were in close alignment with the nominal threshold of 0.05, with average empirical Type I error rates documented at 0.0427, 0.0448, 0.0550, and 0.0463, respectively. Furthermore, the calculated ARE values were found to be 0.735, 0.525, 0.7383, and 0.4167, respectively. This indicates that they exhibited a pronounced superiority in comparison to the BL and LV statistics. Moreover, the SMD and KL statistics exhibited a notable degree of resilience, followed closely by the LVM and LY statistics, which also manifested robustness. The BL statistic demonstrated a slight inflation in Type I error rates under particular circumstances, especially within the unequal large-sample configuration of (70, 80, 90, 100), where the empirical Type I error rate reached 0.0774, marginally surpassing Bradley's upper threshold. Additionally, an observable trend was noted whereby empirical Type I error rates exhibited an increase with augmented sample sizes, suggesting a potential sensitivity of the BL statistic to variations in sample size. Although the overall average empirical Type I error rate for the BL statistic across all scenarios was recorded at 0.0647, which remains within an acceptable threshold, these localized anomalies imply that its reliability may be compromised in specific sample structures, particularly those characterized by larger sample sizes. Furthermore, the obtained ARE value of 1.54 serves as evidence of markedly inadequate performance and robustness. Ultimately, the LV statistic displayed inconsistent performance, with its average Type I error rate attaining 0.0758, slightly surpassing Bradley's upper limit and

exhibiting a considerable ARE value of 2.58. Various specific conditions, especially those involving unequal or diminutive sample sizes (e.g., (10,10,10), (2,6,10), (10,10,10,10), and (2,6,10,14)).

Table 3: Empirical Type I error rate for six tests when $Y \sim N(0,1)$

k	Sample Sizes		BL	LV	LVM	KL	LY	SMD
3	Equal	(10,10,10)	0.0522	0.0662	0.0340	0.0830*	0.0544	0.0494
		(30,30,30)	0.0490	0.0506	0.0378	0.1246*	0.0450	0.0514
		(90,90,90)	0.0498	0.0500	0.0462	0.1646*	0.0480	0.0490
	Unequal	(2,6,10)	0.0448	0.0766*	0.0436	0.1974*	0.0440	0.0480
		(20,25,30)	0.0518	0.0546	0.0390	0.2346*	0.0480	0.0510
		(70,80,90)	0.0488	0.0490	0.0446	0.2696*	0.0544	0.0538
4	Equal	(10,10,10,10)	0.0546	0.0658	0.0324	0.2968*	0.0474	0.0528
		(30,30,30,30)	0.0510	0.0516	0.0370	0.3272*	0.0458	0.0484
		(90,90,90,90)	0.0486	0.0516	0.0466	0.3588*	0.0498	0.0494
	Unequal	(2, 6, 10, 14)	0.0440	0.0778*	0.0444	0.3850*	0.0486	0.0508
		(20, 25, 30, 35)	0.0500	0.0554	0.0392	0.4130*	0.0564	0.0462
		(70, 80, 90, 100)	0.0486	0.0448	0.0406	0.4428*	0.0346	0.0544
Average			0.0565	0.0649	0.0416	0.2076*	0.0511	0.0486
ARE			0.2167	0.8867	0.955	22.4783	0.45	0.1983

**Denotes that the empirical Type I error rate exceeds Bradley's criterion.*

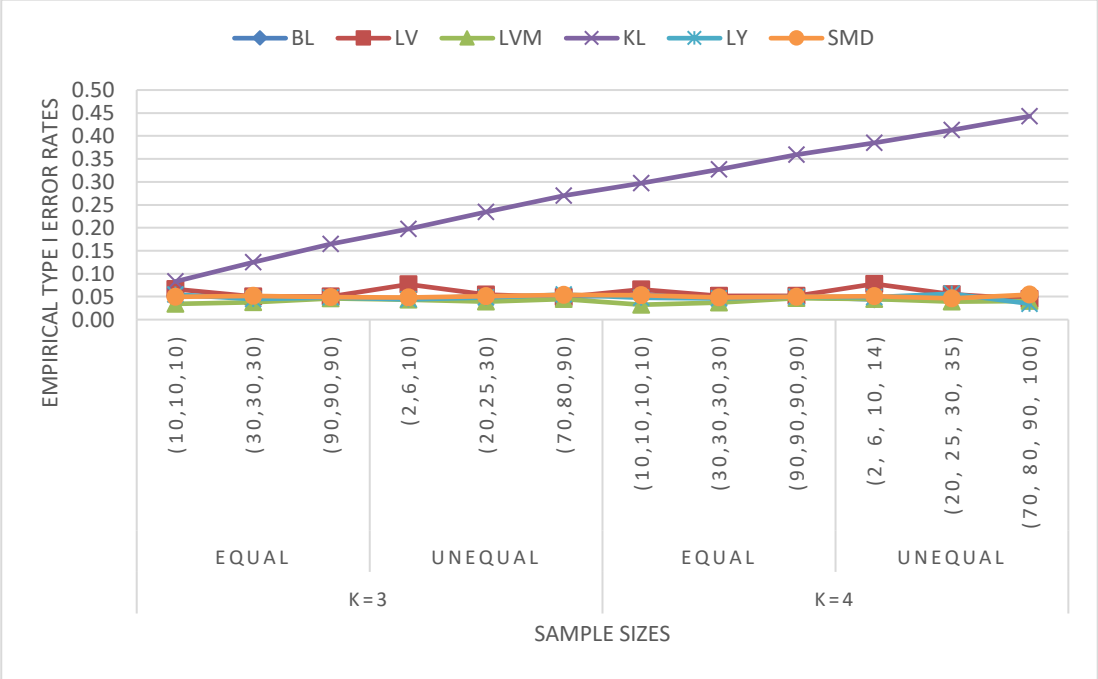

Figure 1: Plot of empirical Type I error rates for six tests when $Y \sim N(0,1)$

Table 4: Empirical Type I error rate for six tests when $Y \sim Beta(1,1)$

k	Sample Sizes		BL	LV	LVM	KL	LY	SMD
3	Equal	(10,10,10)	0.0078*	0.0714	0.0288	0.0388	0.0320	0.1116*
		(30,30,30)	0.0014*	0.0498	0.0308	0.0450	0.0292	0.1186*
		(90,90,90)	0.0008*	0.0520	0.0428	0.0478	0.0414	0.1122*
	Unequal	(2,6,10)	0.0254	0.0946*	0.0460	0.0396	0.0404	0.0690
		(20,25,30)	0.0016*	0.0550	0.0284	0.0466	0.0328	0.1088*
		(70,80,90)	0.0008*	0.0490	0.0392	0.0470	0.0420	0.1232*
4	Equal	(10,10,10,10)	0.0066*	0.0680	0.0272	0.0404	0.0288	0.1456*
		(30,30,30,30)	0.0014*	0.0510	0.0302	0.0432	0.0278	0.1616*
		(90,90,90,90)	0.0002*	0.0532	0.0408	0.0476	0.0386	0.1736*
	Unequal	(2, 6, 10, 14)	0.0184*	0.1042*	0.0432	0.0438	0.0402	0.0906*
		(20, 25, 30, 35)	0.0010*	0.0598	0.0316	0.0442	0.0328	0.1626*
		(70, 80, 90, 100)	0.0006*	0.0496	0.0378	0.0442	0.0386	0.1694*
Average			0.0055*	0.0631	0.0356	0.0440	0.0354	0.1289*
ARE			4.45	1.34	1.4433	0.5983	1.4617	7.89

**Denotes that the empirical Type I error rate exceeds Bradley's criterion.*

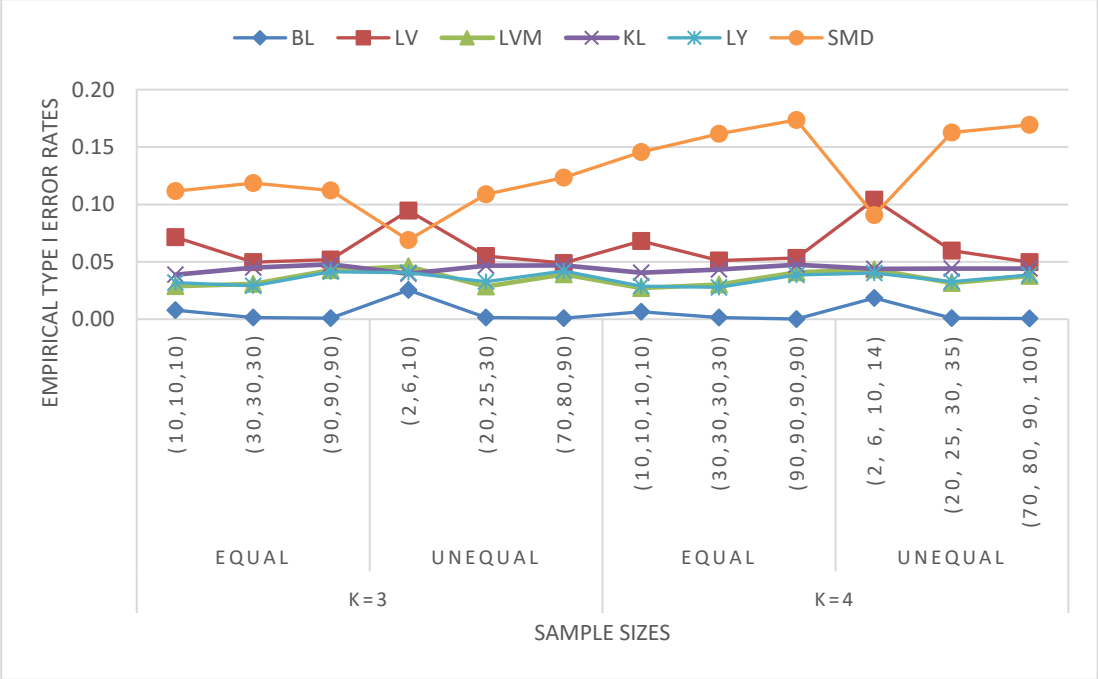

Figure 2: Plot of empirical Type I error rates for six tests when $Y \sim Beta(1,1)$

Table 5: Empirical Type I error rate for six tests when $Y \sim Weibull(2,2)$

k	Sample Sizes		BL	LV	LVM	KL	LY	SMD
3	Equal	(10,10,10)	0.0632	0.0778*	0.0382	0.0378	0.0572	0.0450
		(30,30,30)	0.0672	0.0696	0.0382	0.0474	0.0584	0.0474
		(90,90,90)	0.0670	0.0648	0.0462	0.0484	0.0576	0.0458
	Unequal	(2,6,10)	0.0482	0.0818*	0.0486	0.0420	0.0584	0.0510
		(20,25,30)	0.0646	0.0702	0.0404	0.0450	0.0424	0.0470
		(70,80,90)	0.0714	0.0678	0.0478	0.0468	0.0568	0.0450
4	Equal	(10,10,10,10)	0.0622	0.0926*	0.0374	0.0390	0.0590	0.0424
		(30,30,30,30)	0.0660	0.0700	0.0358	0.0444	0.0572	0.0520
		(90,90,90,90)	0.0720	0.0700	0.0456	0.0500	0.0570	0.0444
	Unequal	(2, 6, 10, 14)	0.0476	0.0938*	0.0454	0.0440	0.0582	0.0496
		(20, 25, 30, 35)	0.0696	0.0766*	0.0384	0.0464	0.0430	0.0436
		(70, 80, 90, 100)	0.0774*	0.0746	0.0498	0.0458	0.0542	0.0428
Average			0.0647	0.0758*	0.0427	0.0448	0.0550	0.0463
ARE			1.54	2.58	0.735	0.525	0.7383	0.4167

**Denotes that the empirical Type I error rate exceeds Bradley's criterion.*

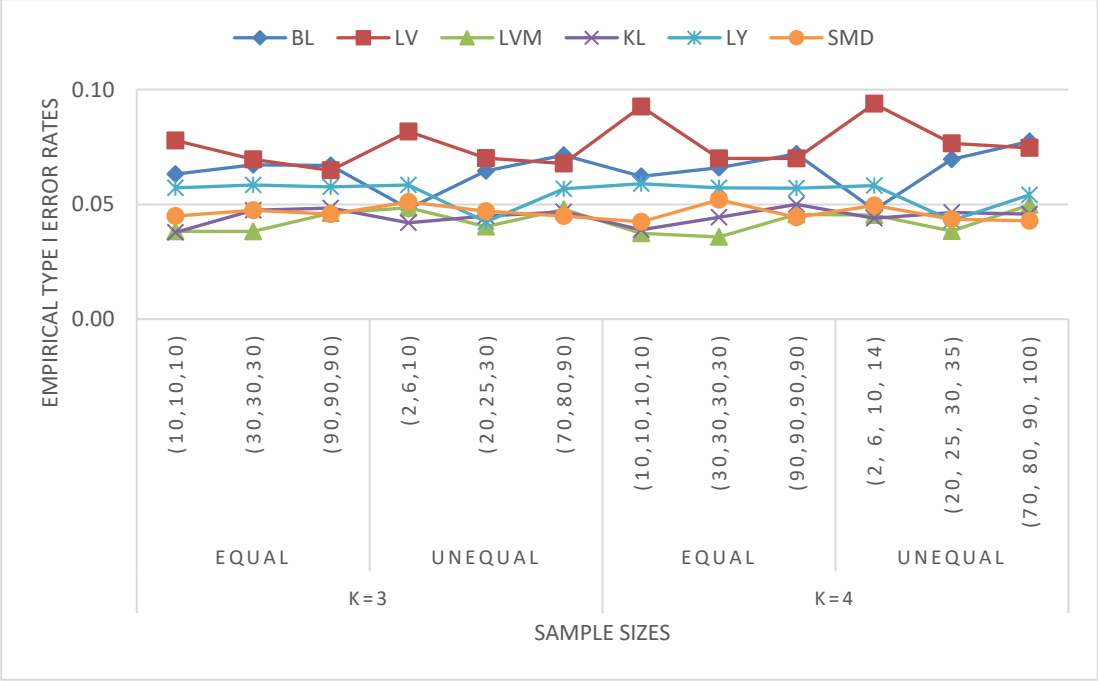

Figure 3: Plot of empirical Type I error rates for six tests when $Y \sim Weibull(2,2)$

4.2 Empirical Power Results

The empirical efficacy of six distinct test statistics was assessed utilizing datasets generated from three specific distributions encompassing three and four groups, with the resulting data systematically delineated in Tables 6 through 11.

As displayed in Tables 6 – 7, the comparative examination of test statistics was performed within the context of the normal distribution. The KL statistic did not satisfy the criterion for Type I error control and was consequently omitted from the power comparison. The results demonstrated that the BL and LY statistics displayed the highest empirical effectiveness among those that successfully regulated the Type I error rate. Both statistical metrics consistently manifested robust performance in contexts involving 3 and 4 groups. Conversely, the LV and LVM statistics, although capable of addressing the Type I error rate, yielded significantly reduced power. In contrast, the SMD statistic demonstrated exceedingly low power. It was determined that, in instances where the sample sizes were equivalent, the BL test statistic demonstrated the greatest statistical power. Conversely, in scenarios characterized by unequal sample sizes, the LY test statistic produced the maximum power. Furthermore, augmenting the sample size led to an enhancement in power across all test statistics.

As evidenced in Tables 8 - 9, the BL and SMD statistics on the Beta distribution did not fulfill the criteria for controlling the Type I error within the framework of this distribution and were thus excluded from the evaluation of statistical power. The LV statistic exhibited the highest statistical power across all experimental conditions and sample sizes. Nonetheless, the LV statistic is not advised for application with small and unequal sample sizes within this particular distribution. The LVM and LY statistics were positioned in the second tier, demonstrating comparable levels of statistical power. In scenarios involving smaller sample sizes, the KL statistic produced inferior power relative to the other statistics. Nevertheless, as sample sizes increased, the power of the KL statistic did improve, yet it consistently remained lower than that of the other statistics. Thus, the KL statistic is deemed inappropriate for relatively small sample sizes. As the sample size escalates, all statistics exhibit enhanced and increasingly comparable power.

As reported in Tables 10–11, the data distributed as the Weibull distribution revealed that the LV statistic exhibited the least reliability, often surpassing the acceptable threshold when confronted with unbalanced or small sample sizes. Consequently, it was omitted from the power evaluation. The findings demonstrated that BL achieved the highest levels of power, presumably attributable to its relatively elevated empirical Type I error rate. The LVM and LY statistics showcased substantial power across all group sizes and contexts, signifying considerable robustness in non-normal conditions. Notably, LY is particularly advantageous for small to medium sample sizes, as it consistently surpassed LVM in these instances. In contrast, LVM is more appropriate for larger sample sizes due to its superior power. The SMD statistic manifested the lowest power in all scenarios, while KL persistently yielded low power throughout all conditions.

Table 6: Empirical power of the tests under the normal distribution with $k = 3$

Sample Size	Variance Ratio	Test Statistic					
		BL	LV	LVM	KL	LY	SMD
(10, 10, 10)	1.0: 2.0: 2.0	0.1406	0.1488*	0.0828	-	0.1284	0.0004
	1.0: 2.0: 3.0	0.2620*	0.2442	0.1514	-	0.2082	0.0006
	1.0: 1.0: 5.0	0.6924*	0.6086	0.4788	-	0.5250	0.0026
	1.0: 4.0: 7.0	0.7016*	0.5626	0.4004	-	0.5868	0.0030
	1.0: 8.0: 8.0	0.8588*	0.6844	0.5012	-	0.7522	0.0032
	1.0: 8.0: 15.0	0.9540*	0.8268	0.6754	-	0.8776	0.0036
(30, 30, 30)	1.0: 2.0: 2.0	0.4446*	0.3846	0.3280	-	0.4284	0.0006
	1.0: 2.0: 3.0	0.7486*	0.6670	0.6202	-	0.7082	0.0008
	1.0: 1.0: 5.0	0.9946*	0.9854	0.9820	-	0.9876	0.0028
	1.0: 4.0: 7.0	0.9982*	0.9914	0.9878	-	0.9952	0.0032
	1.0: 8.0: 8.0	1.0000*	1.0000*	0.9994	-	0.9996	0.0046
	1.0: 8.0: 15.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0048
(90, 90, 90)	1.0: 2.0: 2.0	0.9262*	0.8768	0.8664	-	0.9102	0.0012
	1.0: 2.0: 3.0	0.9978*	0.9926	0.9922	-	0.9228	0.0018
	1.0: 1.0: 5.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0030
	1.0: 4.0: 7.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0032
	1.0: 8.0: 8.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0038
	1.0: 8.0: 15.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0042
(2, 6, 10)	1.0: 2.0: 2.0	0.0510	-	0.0310	-	0.0780*	0.0002
	1.0: 2.0: 3.0	0.0696	-	0.0370	-	0.1392*	0.0004
	1.0: 1.0: 5.0	0.3178	-	0.1590	-	0.4560*	0.0010
	1.0: 4.0: 7.0	0.1086	-	0.0522	-	0.4586*	0.0012
	1.0: 8.0: 8.0	0.0886	-	0.0384	-	0.6540*	0.0018
	1.0: 8.0: 15.0	0.1504	-	0.0660	-	0.6680*	0.0022
(20, 25, 30)	1.0: 2.0: 2.0	0.3144	0.2626	0.2132	-	0.3204*	0.0032
	1.0: 2.0: 3.0	0.6078	0.5100	0.4526	-	0.6486*	0.0038
	1.0: 1.0: 5.0	1.0000*	0.9712	0.9638	-	0.8760	0.0046
	1.0: 4.0: 7.0	0.9882*	0.9526	0.9346	-	0.9240	0.0048
	1.0: 8.0: 8.0	0.9990*	0.9896	0.9832	-	0.9822	0.0090
	1.0: 8.0: 15.0	1.0000*	0.9990	0.9980	-	1.0000*	0.0094
(70, 80, 90)	1.0: 2.0: 2.0	0.8634	0.7898	0.7766	-	0.9048*	0.0310
	1.0: 2.0: 3.0	0.9948*	0.9838	0.9822	-	0.9228	0.0370
	1.0: 1.0: 5.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.1590
	1.0: 4.0: 7.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0522
	1.0: 8.0: 8.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0384
	1.0: 8.0: 15.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0660

Note: **Denotes* the highest statistical power within each group among the tests that successfully control the Type I error rate according to Bradley's criterion; *- Indicates* that statistical power is not reported because the test statistic fails to control the Type I error rate.

Table 7: Empirical power of the tests under the normal distribution with $k = 4$

Sample Size	Variance Ratio	Test Statistic					
		BL	LV	LVM	KL	LY	SMD
(10, 10, 10, 10)	1.0: 1.5: 1.5: 2.0	0.1152	0.1218*	0.0692	-	0.0914	0.0004
	1.0: 2.0: 3.0: 4.0	0.3530*	0.3184	0.1914	-	0.2880	0.0006
	1.0: 3.0: 3.0: 6.0	0.5276*	0.4358	0.2910	-	0.4236	0.0010
	1.0: 3.0: 5.0: 7.0	0.6638*	0.5260	0.3628	-	0.5398	0.0014
	1.0: 4.0: 4.0: 10.0	0.7738*	0.6426	0.4706	-	0.6466	0.0042
	1.0: 6.0: 11.0: 16.0	0.9566*	0.8036	0.6234	-	0.8714	0.0312
(30, 30, 30, 30)	1.0: 1.5: 1.5: 2.0	0.2864*	0.2526	0.2090	-	0.2648	0.0004
	1.0: 2.0: 3.0: 4.0	0.9076*	0.8420	0.8008	-	0.8836	0.0008
	1.0: 3.0: 3.0: 6.0	0.9870*	0.9570	0.9386	-	0.9794	0.0056
	1.0: 3.0: 5.0: 7.0	0.9976*	0.9898	0.9828	-	0.9950	0.0076
	1.0: 4.0: 4.0: 10.0	1.0000*	0.9978	0.9956	-	0.9996	0.0086
	1.0: 6.0: 11.0: 16.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0298
(90, 90, 90, 90)	1.0: 1.5: 1.5: 2.0	0.7834*	0.7144	0.6990	-	0.7738	0.0006
	1.0: 2.0: 3.0: 4.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0016
	1.0: 3.0: 3.0: 6.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0068
	1.0: 3.0: 5.0: 7.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0084
	1.0: 4.0: 4.0: 10.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0134
	1.0: 6.0: 11.0: 16.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0346
(2, 6, 10, 14)	1.0: 1.5: 1.5: 2.0	0.0640	-	0.0416	-	0.0874*	0.0002
	1.0: 2.0: 3.0: 4.0	0.1016	-	0.0568	-	0.2564*	0.0010
	1.0: 3.0: 3.0: 6.0	0.1656	-	0.0970	-	0.3580*	0.0032
	1.0: 3.0: 5.0: 7.0	0.2086	-	0.1486	-	0.4308*	0.0084
	1.0: 4.0: 4.0: 10.0	0.2812	-	0.1612	-	0.6322*	0.0011
	1.0: 6.0: 11.0: 16.0	0.2150	-	0.1840	-	0.8672*	0.0124
(20, 25, 30, 35)	1.0: 1.5: 1.5: 2.0	0.2474*	0.2042	0.1806	-	0.2430	0.0102
	1.0: 2.0: 3.0: 4.0	0.8324	0.8104	0.6764	-	0.8450*	0.0124
	1.0: 3.0: 3.0: 6.0	0.9622*	0.9452	0.8688	-	0.9202	0.1890
	1.0: 3.0: 5.0: 7.0	0.9896	0.9610	0.9440	-	0.9910*	0.0192
	1.0: 4.0: 4.0: 10.0	0.9986	0.9890	0.9836	-	0.9994*	0.0438
	1.0: 6.0: 11.0: 16.0	1.0000*	0.9992	0.9978	-	1.0000*	0.0480
(70, 80, 90, 100)	1.0: 1.5: 1.5: 2.0	0.7360	0.6552	0.6404	-	0.7610*	0.0092
	1.0: 2.0: 3.0: 4.0	0.9996	0.9992	0.9992	-	1.0000*	0.0102
	1.0: 3.0: 3.0: 6.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0244
	1.0: 3.0: 5.0: 7.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0246
	1.0: 4.0: 4.0: 10.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0268
	1.0: 6.0: 11.0: 16.0	1.0000*	1.0000*	1.0000*	-	1.0000*	0.0520

Note: **Denotes* the highest statistical power within each group among the tests that successfully control the Type I error rate according to Bradley's criterion; *- Indicates* that statistical power is not reported because the test statistic fails to control the Type I error rate.

Table 8: Empirical power of the tests under the Beta distribution with $k = 3$

Sample Size	Variance Ratio	Test Statistic					
		BL	LV	LVM	KL	LY	SMD
(10, 10, 10)	1.0: 2.0: 2.0	-	0.2318*	0.1420	0.0490	0.1554	-
	1.0: 2.0: 3.0	-	0.3584*	0.2294	0.0528	0.2652	-
	1.0: 1.0: 5.0	-	0.6076*	0.4234	0.0124	0.5714	-
	1.0: 4.0: 7.0	-	0.7448*	0.6082	0.0644	0.5860	-
	1.0: 8.0: 8.0	-	0.8770*	0.7884	0.0364	0.7162	-
	1.0: 8.0: 15.0	-	0.9382*	0.8742	0.0698	0.7906	-
(30, 30, 30)	1.0: 2.0: 2.0	-	0.5688*	0.4986	0.2252	0.4746	-
	1.0: 2.0: 3.0	-	0.8180*	0.7794	0.2882	0.7322	-
	1.0: 1.0: 5.0	-	0.9954*	0.9940	0.0374	0.9748	-
	1.0: 4.0: 7.0	-	0.9974*	0.9968	0.5604	0.9784	-
	1.0: 8.0: 8.0	-	1.0000*	1.0000*	0.5310	0.9932	-
	1.0: 8.0: 15.0	-	1.0000*	1.0000*	0.7200	0.9966	-
(90, 90, 90)	1.0: 2.0: 2.0	-	0.9758*	0.9742	0.8782	0.9670	-
	1.0: 2.0: 3.0	-	1.0000*	0.9996	0.9576	0.9992	-
	1.0: 1.0: 5.0	-	1.0000*	1.0000*	0.9770	1.0000*	-
	1.0: 4.0: 7.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
	1.0: 8.0: 8.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
	1.0: 8.0: 15.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
(2, 6, 10)	1.0: 2.0: 2.0	-	-	0.1062	0.1378*	0.1044	-
	1.0: 2.0: 3.0	-	-	0.1610	0.1800	0.2280*	-
	1.0: 1.0: 5.0	-	-	0.3752	0.0390	0.5208*	-
	1.0: 4.0: 7.0	-	-	0.3360	0.3368	0.5420*	-
	1.0: 8.0: 8.0	-	-	0.3976	0.4202	0.6982*	-
	1.0: 8.0: 15.0	-	-	0.5080	0.4514	0.7804*	-
(20, 25, 30)	1.0: 2.0: 2.0	-	0.4590*	0.3734	0.3062	0.4082	-
	1.0: 2.0: 3.0	-	0.7210*	0.6442	0.4092	0.6720	-
	1.0: 1.0: 5.0	-	0.9870*	0.9788	0.0094	0.9540	-
	1.0: 4.0: 7.0	-	0.9906*	0.9828	0.9024	0.9678	-
	1.0: 8.0: 8.0	-	0.9980*	0.9970	0.9850	0.9892	-
	1.0: 8.0: 15.0	-	1.0000*	1.0000*	0.9956	0.9986	-
(70, 80, 90)	1.0: 2.0: 2.0	-	0.9384	0.9304	0.9184	0.9996*	-
	1.0: 2.0: 3.0	-	0.9976	0.9974	0.9822	0.9998*	-
	1.0: 1.0: 5.0	-	1.0000*	1.0000*	0.0462	1.0000*	-
	1.0: 4.0: 7.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
	1.0: 8.0: 8.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
	1.0: 8.0: 15.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-

Note: **Denotes* the highest statistical power within each group among the tests that successfully control the Type I error rate according to Bradley's criterion; *- Indicates* that statistical power is not reported because the test statistic fails to control the Type I error rate.

Table 9: Empirical power of the tests under the Beta distribution with $k = 4$

Sample Size	Variance Ratio	Test Statistic					
		BL	LV	LVM	KL	LY	SMD
(10, 10, 10, 10)	1.0: 1.5: 1.5: 2.0	-	0.1688*	0.0930	0.0644	0.0976	-
	1.0: 2.0: 3.0: 4.0	-	0.4940*	0.3202	0.1162	0.3350	-
	1.0: 3.0: 3.0: 6.0	-	0.6472*	0.4746	0.1868	0.4512	-
	1.0: 3.0: 5.0: 7.0	-	0.7618*	0.6094	0.1924	0.5506	-
	1.0: 4.0: 4.0: 10.0	-	0.8180*	0.6810	0.2782	0.6210	-
	1.0: 6.0: 11.0: 16.0	-	0.9534*	0.8936	0.3460	0.7766	-
(30, 30, 30, 30)	1.0: 1.5: 1.5: 2.0	-	0.4306*	0.3564	0.2864	0.3444	-
	1.0: 2.0: 3.0: 4.0	-	0.9468*	0.9260	0.7568	0.8638	-
	1.0: 3.0: 3.0: 6.0	-	0.9918*	0.9864	0.9496	0.9506	-
	1.0: 3.0: 5.0: 7.0	-	0.9994*	0.9994*	0.9838	0.9778	-
	1.0: 4.0: 4.0: 10.0	-	1.0000*	0.9996	0.9940	0.9848	-
	1.0: 6.0: 11.0: 16.0	-	1.0000*	1.0000*	1.0000*	0.9966	-
(90, 90, 90, 90)	1.0: 1.5: 1.5: 2.0	-	0.9044*	0.8922	0.8652	0.8922	-
	1.0: 2.0: 3.0: 4.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
	1.0: 3.0: 3.0: 6.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
	1.0: 3.0: 5.0: 7.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
	1.0: 4.0: 4.0: 10.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
	1.0: 6.0: 11.0: 16.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
(2, 6, 10, 14)	1.0: 1.5: 1.5: 2.0	-	-	0.0808	0.1376*	0.0920	-
	1.0: 2.0: 3.0: 4.0	-	-	0.2076	0.3226*	0.3108	-
	1.0: 3.0: 3.0: 6.0	-	-	0.2890	0.3926	0.4020*	-
	1.0: 3.0: 5.0: 7.0	-	-	0.3332	0.4882	0.5210*	-
	1.0: 4.0: 4.0: 10.0	-	-	0.3976	0.5016	0.6102*	-
	1.0: 6.0: 11.0: 16.0	-	-	0.5040	0.7184	0.7408*	-
(20, 25, 30, 35)	1.0: 1.5: 1.5: 2.0	-	0.3628*	0.2666	0.3454	0.3084	-
	1.0: 2.0: 3.0: 4.0	-	0.9002*	0.8516	0.8514	0.8208	-
	1.0: 3.0: 3.0: 6.0	-	0.9734*	0.9562	0.9664	0.9430	-
	1.0: 3.0: 5.0: 7.0	-	0.9916*	0.9838	0.9912	0.9810	-
	1.0: 4.0: 4.0: 10.0	-	0.9974*	0.9948	0.9964	0.9848	-
	1.0: 6.0: 11.0: 16.0	-	1.0000*	1.0000*	1.0000*	0.9966	-
(70, 80, 90, 100)	1.0: 1.5: 1.5: 2.0	-	0.8476	0.8274	0.8740	0.9402*	-
	1.0: 2.0: 3.0: 4.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
	1.0: 3.0: 3.0: 6.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
	1.0: 3.0: 5.0: 7.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
	1.0: 4.0: 4.0: 10.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-
	1.0: 6.0: 11.0: 16.0	-	1.0000*	1.0000*	1.0000*	1.0000*	-

Note: **Denotes* the highest statistical power within each group among the tests that successfully control the Type I error rate according to Bradley's criterion; - *Indicates* that empirical power is not reported because the test statistic fails to control the Type I error rate.

Table 10: Empirical power of the tests under the Weibull distribution with $k = 3$

Sample Size	Variance Ratio	Test Statistic

		BL	LV	LVM	KL	LY	SMD
(10, 10, 10)	1.0: 2.0: 2.0	0.3990*	-	0.0700	0.0366	0.1630	0.0154
	1.0: 2.0: 3.0	0.4700*	-	0.0998	0.0372	0.2080	0.0146
	1.0: 1.0: 5.0	0.6756*	-	0.2652	0.0518	0.3520	0.0038
	1.0: 4.0: 7.0	0.7076*	-	0.2090	0.0344	0.3706	0.0032
	1.0: 8.0: 8.0	0.7994*	-	0.2492	0.0236	0.4602	0.0028
	1.0: 8.0: 15.0	0.8790*	-	0.3518	0.0236	0.5576	0.0006
(30, 30, 30)	1.0: 2.0: 2.0	0.6150*	-	0.1642	0.0570	0.2318	0.0076
	1.0: 2.0: 3.0	0.7532*	-	0.3098	0.0720	0.3494	0.0066
	1.0: 1.0: 5.0	0.9498*	-	0.8004	0.1988	0.6694	0.0042
	1.0: 4.0: 7.0	0.9676*	-	0.7688	0.0812	0.7264	0.0030
	1.0: 8.0: 8.0	0.9892*	-	0.8864	0.0424	0.8388	0.0004
	1.0: 8.0: 15.0	0.9982*	-	0.9694	0.0644	0.9132	0.0002
(90, 90, 90)	1.0: 2.0: 2.0	0.8410*	-	0.4958	0.0900	0.4244	0.0074
	1.0: 2.0: 3.0	0.9542*	-	0.8112	0.1548	0.6626	0.0070
	1.0: 1.0: 5.0	0.9990	-	0.9992	0.6898	0.9626	0.0068
	1.0: 4.0: 7.0	1.0000*	-	0.9990	0.2156	0.9626	0.0060
	1.0: 8.0: 8.0	1.0000*	-	1.0000*	0.0646	0.9930	0.0004
	1.0: 8.0: 15.0	1.0000*	-	1.0000*	0.1882	0.9972	0.0002
(2, 6, 10)	1.0: 2.0: 2.0	0.1898*	-	0.0538	0.0342	0.1580	0.0142
	1.0: 2.0: 3.0	0.2174*	-	0.0500	0.0320	0.1980	0.0122
	1.0: 1.0: 5.0	0.4284*	-	0.0946	0.0270	0.3384	0.0036
	1.0: 4.0: 7.0	0.2690	-	0.0502	0.0320	0.3580*	0.0028
	1.0: 8.0: 8.0	0.2484	-	0.0460	0.0426	0.4560*	0.0018
	1.0: 8.0: 15.0	0.3222	-	0.0546	0.0476	0.5420*	0.0004
(20, 25, 30)	1.0: 2.0: 2.0	0.5678*	-	0.1114	0.0486	0.2202	0.0068
	1.0: 2.0: 3.0	0.7030*	-	0.2102	0.0528	0.3348	0.0064
	1.0: 1.0: 5.0	0.9350*	-	0.7246	0.1152	0.6580	0.0040
	1.0: 4.0: 7.0	0.9364*	-	0.5700	0.0500	0.7084	0.0028
	1.0: 8.0: 8.0	0.9752*	-	0.6652	0.0336	0.8204	0.0002
	1.0: 8.0: 15.0	0.9938*	-	0.8532	0.0444	0.9092	0.0002
(70, 80, 90)	1.0: 2.0: 2.0	0.8008*	-	0.3986	0.0794	0.4802	0.0068
	1.0: 2.0: 3.0	0.9380*	-	0.7280	0.1258	0.6504	0.0068
	1.0: 1.0: 5.0	0.9990	-	0.9980	0.5474	0.9508	0.0056
	1.0: 4.0: 7.0	1.0000*	-	0.9972	0.1536	0.9502	0.0058
	1.0: 8.0: 8.0	1.0000*	-	1.0000*	0.0422	0.9802	0.0002
	1.0: 8.0: 15.0	1.0000*	-	1.0000*	0.1404	0.9960	0.0002

Note: **Denotes* the highest statistical power within each group among the tests that successfully control the Type I error rate according to Bradley's criterion; *- Indicates* that empirical power is not reported because the test statistic fails to control the Type I error rate.

Table 11: Empirical power of the tests under the Weibull distribution with $k = 4$

Sample Size	Variance Ratio	Test Statistic					
		BL	LV	LVM	KL	LY	SMD
(10, 10, 10, 10)	1.0: 1.5: 1.5: 2.0	0.4586*	-	0.0616	0.2086	0.2734	0.0040
	1.0: 2.0: 3.0: 4.0	0.6190*	-	0.1204	0.2220	0.2444	0.0046
	1.0: 3.0: 3.0: 6.0	0.7044*	-	0.1680	0.2314	0.2946	0.0020
	1.0: 3.0: 5.0: 7.0	0.7576*	-	0.1930	0.2342	0.1500	0.0092
	1.0: 4.0: 4.0: 10.0	0.8110*	-	0.2502	0.0864	0.3986	0.0028
	1.0: 6.0: 11.0: 16.0	0.9094*	-	0.3194	0.1042	0.0018	0.0006
(30, 30, 30, 30)	1.0: 1.5: 1.5: 2.0	0.6270*	-	0.1188	0.2436	0.5300	0.0158
	1.0: 2.0: 3.0: 4.0	0.8818*	-	0.4134	0.2760	0.4466	0.0128
	1.0: 3.0: 3.0: 6.0	0.9460*	-	0.5982	0.2984	0.5918	0.0098
	1.0: 3.0: 5.0: 7.0	0.9750*	-	0.7154	0.3020	0.1746	0.0078
	1.0: 4.0: 4.0: 10.0	0.9904*	-	0.8328	0.3340	0.7702	0.0004
	1.0: 6.0: 11.0: 16.0	0.9994*	-	0.9596	0.3566	0.9480	0.0002
(90, 90, 90, 90)	1.0: 1.5: 1.5: 2.0	0.8244*	-	0.3496	0.6938	0.9110	0.0128
	1.0: 2.0: 3.0: 4.0	0.9952*	-	0.9584	0.7294	0.8422	0.0104
	1.0: 3.0: 3.0: 6.0	0.9994*	-	0.9960	0.7814	0.9418	0.0086
	1.0: 3.0: 5.0: 7.0	0.9996	-	1.0000*	0.7822	0.3094	0.0034
	1.0: 4.0: 4.0: 10.0	1.0000*	-	1.0000*	0.8480	0.9882	0.0010
	1.0: 6.0: 11.0: 16.0	1.0000*	-	1.0000*	0.8754	0.9996	0.0008
(2, 6, 10, 14)	1.0: 1.5: 1.5: 2.0	0.3098*	-	0.0488	0.2008	0.2456	0.0098
	1.0: 2.0: 3.0: 4.0	0.3674*	-	0.0446	0.2105	0.2012	0.0126
	1.0: 3.0: 3.0: 6.0	0.4124*	-	0.0586	0.2200	0.2780	0.0112
	1.0: 3.0: 5.0: 7.0	0.4148*	-	0.0472	0.2408	0.2340	0.0134
	1.0: 4.0: 4.0: 10.0	0.4874*	-	0.0800	0.0842	0.3870	0.0186
	1.0: 6.0: 11.0: 16.0	0.4806*	-	0.0548	0.0856	0.0009	0.0234
(20, 25, 30, 35)	1.0: 1.5: 1.5: 2.0	0.6188*	-	0.0914	0.2560	0.4502	0.0012
	1.0: 2.0: 3.0: 4.0	0.8374*	-	0.3050	0.2740	0.4080	0.0026
	1.0: 3.0: 3.0: 6.0	0.9148*	-	0.4710	0.2804	0.5200	0.0128
	1.0: 3.0: 5.0: 7.0	0.9484*	-	0.5374	0.2508	0.1602	0.0178
	1.0: 4.0: 4.0: 10.0	0.9740*	-	0.7096	0.3204	0.7204	0.0170
	1.0: 6.0: 11.0: 16.0	0.9970*	-	0.8406	0.3432	0.9082	0.0224
(70, 80, 90, 100)	1.0: 1.5: 1.5: 2.0	-	-	0.3114	0.6804	0.9008*	0.0008
	1.0: 2.0: 3.0: 4.0	-	-	0.9220	0.7032	0.8380	0.0018
	1.0: 3.0: 3.0: 6.0	-	-	0.9894	0.7678	0.9102	0.0065
	1.0: 3.0: 5.0: 7.0	-	-	0.9984	0.7764	0.2086	0.0078
	1.0: 4.0: 4.0: 10.0	-	-	0.9996*	0.8234	0.9884	0.0198
	1.0: 6.0: 11.0: 16.0	-	-	1.0000*	0.8890	0.9886	0.0212

Note: **Denotes* the highest statistical power within each group among the tests that successfully control the Type I error rate according to Bradley's criterion; *- Indicates* that empirical power is not reported because the test statistic fails to control the Type I error rate.

5. Conclusion

This study conducted a Monte Carlo simulation to systematically compare the performance of six test statistics for assessing the homogeneity of variance, including Bartlett's (BL), Levene's (LV), modified Levene's (LVM), Klotz's (KL), Layard's (LY), and Samiuddin's (SMD). The objective was to evaluate the capability of each test in accurately controlling the Type I error rate while maintaining adequate statistical power under varying conditions to identify the most effective tests. Simulated data were generated from three and four populations, assuming underlying distributions from the normal, Beta, and Weibull distributions. Each simulation scenario was repeated 5,000 times to ensure stability and precision of the empirical estimates. The design also incorporated both equal and unequal sample size configurations to examine the robustness and distributional sensitivity of each test.

In the framework of the normal distribution, the study revealed that the BL, LY, and SMD statistics demonstrated strong control over the Type I error rate and exhibited robustness under normal distribution conditions. The LV statistic should be used cautiously, particularly when sample sizes are unequal. Although the LVM statistic effectively controlled the Type I error rate, its conservative nature may reduce its effectiveness in scenarios where statistical power is critical. Conversely, the KL statistic should have been avoided due to its persistently inflated error rates. When considering the empirical power, the BL and LY statistics exhibited the highest empirical efficacy among those that adequately regulated the Type I error rate. Both statistical measures demonstrated consistent effectiveness in contexts involving three and four groups. The LV and LVM statistics, although proficient in controlling the Type I error rate, yielded significantly reduced statistical power. In contrast, the SMD statistic revealed exceedingly low statistical power. It was ascertained that, in cases where the sample sizes were equal, the BL test statistic exhibited the highest statistical power. Conversely, in contexts characterized by unequal sample sizes, the LY test statistic generated the maximum statistical power. Moreover, an increase in sample size resulted in an enhancement of power across all statistical tests.

Under the Beta distribution, the KL, LVM, and LY statistics emerged as the top-performing methods, demonstrating strong Type I error control across all scenarios, including various combinations of equal and unequal sample sizes and differing numbers of groups. Notably, the KL statistic, despite its poor performance under the normal distribution, proved to be the most robust and reliable in this setting. The LVM statistic maintained conservative behavior, with slightly deflated error rates but no violations of the acceptable range, indicating high reliability, albeit potentially at the expense of reduced statistical power. The LY statistic, while accurate and stable under both normal and Beta distributions, exhibited a minor decline in performance under skewed distributions, such as the Weibull, indicating that it is dependable but not the most robust across varying data conditions. In contrast, the LV statistic showed only moderate performance;

although its average error rate was within bounds, it exhibited sensitivity to small and unequal sample sizes. The BL statistic was overly conservative, with consistently low error rates that often fell below the lower bound of Bradley's criterion, potentially limiting its practical utility due to underpowered testing. Finally, the SMD statistic performed the worst, with significantly inflated Type I error rates across all scenarios, rendering it unsuitable for use under uniform distribution assumptions.

Under data generated from the Weibull distribution, LY, SMD, KL, and LVM demonstrated consistent control over the Type I error rate. Regarding robustness, the SMD and KL statistics were found to be particularly robust, followed by the LVM and LY statistics, which also exhibited robustness. Although BL performed reasonably well in most settings, it exhibited slight inflation under some conditions. The LV statistic exhibited the lowest degree of reliability, often surpassing the permissible threshold when confronted with unbalanced sample sizes or smaller group configurations. The LV statistic proved to be ineffective in the regulation of the Type I error rate and was therefore omitted from the ensuing power evaluation. In the analysis of empirical power, the BL statistic achieved the most elevated power levels, a phenomenon ascribed to its relatively heightened empirical Type I error rate. The LVM and LY statistics exhibited high power across all group sizes and scenarios, thereby demonstrating their robustness in the presence of non-normative conditions. Nevertheless, the LY statistic is deemed more appropriate for small to medium sample sizes, as it consistently yields superior power compared to the LVM statistic in such contexts. In contrast, for large sample sizes, the LVM statistic is recommended, as it offers enhanced power relative to the LY statistic, demonstrating strong efficacy in moderately skewed datasets. The SMD statistic was positioned at the lowest rank, displaying exceedingly low power across all examined scenarios. while the KL statistic consistently exhibited the lowest power across all conditions.

Future research endeavors might investigate the efficacy of the six test statistics across a more extensive spectrum of fundamental distributions, encompassing heavy-tailed or significantly skewed distributions [29], thereby enhancing the comprehension of their robustness in varied data conditions. Although the current investigation is predicated on simulation, the application of these tests to empirical datasets across diverse domains (e.g., education, medicine, environmental studies) would substantiate their practical applicability and elucidate context-specific constraints.

Acknowledgments: This work was supported by the International SciKU Branding (ISB), Faculty of Science, Kasetsart University, and the Department of Statistics, Faculty of Science, Kasetsart University, Thailand. We thank the referees for their valuable suggestions, which enhanced this paper's quality.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] R.W. Johnson, Alternate Forms of the One-Way ANOVA F and Kruskal–Wallis Test Statistics, J. Stat. Data Sci. Educ. 30 (2022), 82-85. https://doi.org/10.1080/26939169.2021.2025177.

[2] B.L. Welch, On the Comparison of Several Mean Values: An Alternative Approach, Biometrika 38 (1951), 330-336. https://doi.org/10.1093/biomet/38.3-4.330.

[3] W.H. Kruskal, W.A. Wallis, Use of Ranks in One-Criterion Variance Analysis, J. Am. Stat. Assoc. 47 (1952), 583-621. https://doi.org/10.1080/01621459.1952.10483441.

[4] H. Scheffé, The Analysis of Variance, John Wiley & Sons, New York, 1959.

[5] H. Levene, Robust Tests for Equality of Variance, in: Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, Stanford University Press, California, pp. 278–292, 1960.

[6] G. W. Snedecor, W. G. Cochran, Statistical Methods, 8th ed. Iowa State University Press, 1989.

[7] M.B. Brown, A.B. Forsythe, Robust Tests for the Equality of Variances, J. Am. Stat. Assoc. 69 (1974), 364-367. https://doi.org/10.2307/2285659.

[8] H. B. Lee, G. S. Katz, A. F. Restori, A Monte Carlo Study of Seven Homogeneity of Variance Tests, J. Math. Stat. 6 (2010), 359-366. http://doi.org/10.3844/jmssp.2010.359.366.

[9] D. Sharma, B.M.G. Kibria, On Some Test Statistics for Testing Homogeneity of Variances: A Comparative Study, J. Stat. Comput. Simul. 83 (2013), 1944-1963. https://doi.org/10.1080/00949655.2012.675336.

[10] Y. Wang, P. Rodríguez de Gil, Y. Chen, J.D. Kromrey, E.S. Kim, et al., Comparing the Performance of Approaches for Testing the Homogeneity of Variance Assumption in One-Factor ANOVA Models, Educ. Psychol. Meas. 77 (2016), 305-329. https://doi.org/10.1177/0013164416645162.

[11] W.J. Conover, A.J. Guerrero-Serrano, V.G. Tercero-Gómez, An Update on 'A Comparative Study of Tests for Homogeneity of Variance', J. Stat. Comput. Simul. 88 (2018), 1454-1469. https://doi.org/10.1080/00949655.2018.1438437.

[12] W. Riansut, A Comparison of the Efficiency of the Test Statistics for Testing Homogeneity of Variance, Naresuan Univ. J. Sci. Technol. 26 (2018), 170-180.

[13] S. Sinsomboonthong, An Efficiency Comparison of 3 Groups Homogeneity of Population Variance Tests under Highly Kurtosis and Skewness Distributions, Thai Sci. Tech. J. 26 (2018), 721–738.

[14] K. Soikliew, A. Araveeporn, Modifications of Levene's and O'Brien's Tests for Testing the Homogeneity of Variance Based on Median and Trimmed Mean, Thai. Stat. 16 (2018), 106–128.

[15] K. Jiamwattanapong, N. Ingadapa, Performance of Tests for Homogeneity of Variances for More than Two Samples, Int. J. Manag. Appl. Sci. 6 (2020), 57-62.

[16] K. Sritan, B. Phuenaree, A Comparison of Efficiency for Homogeneity of Variance Tests under Log-normal Distribution, Asian J. Appl. Sci. 9 (2021), 254-259. http://doi.org/10.24203/ajas.v9i4.6692.

[17] Y. Zhou, Y. Zhu, W.K. Wong, Statistical Tests for Homogeneity of Variance for Clinical Trials and Recommendations, Contemp. Clin. Trials Commun. 33 (2023), 101119. https://doi.org/10.1016/j.conctc.2023.101119.

[18] M. S. Bartlett, Properties of Sufficiency and Statistical Tests, Proc. R. Soc. Lond. A 160 (1937), 268–282. https://doi.org/10.1098/rspa.1937.0109.
[19] J.L. Gastwirth, Y.R. Gel, W. Miao, The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice, Stat. Sci. 24 (2009), 343-360. https://doi.org/10.1214/09-sts301.
[20] J. Klotz, Nonparametric Tests for Scale, Ann. Math. Stat. 33 (1962), 498-512. https://doi.org/10.1214/aoms/1177704576.
[21] F. Karaman, Generalization of Klotz's Test, J. Appl. Sci. 9 (2009), 2916-2924. https://doi.org/10.3923/jas.2009.2916.2924.
[22] SAS Institute Inc, SAS/STAT ® 13.1 User's Guide: The NPAR1WAY Procedure, Accessed: Feb. 11, 2025. https://support.sas.com/documentation/onlinedoc/stat/131/npar1way.pdf.
[23] M.W.J. Layard, Robust Large-Sample Tests for Homogeneity of Variances, J. Am. Stat. Assoc. 68 (1973), 195-198. https://doi.org/10.2307/2284168.
[24] M. Samiuddin, Bayesian Test of Homogeneity of Variance, J. Am. Stat. Assoc. 71 (1976), 515-517. https://doi.org/10.2307/2285344.
[25] S. Chaipitak, B. Choopradit, A New Test for Equality of Two Covariance Matrices in High-Dimensional Data, Math. Stat. 12 (2024), 455-464. https://doi.org/10.13189/ms.2024.120507.
[26] P.A. Games, H.B. Winkler, D.A. Probert, Robust Tests for Homogeneity of Variance, Educ. Psychol. Meas. 32 (1972), 887-909. https://doi.org/10.1177/001316447203200404.
[27] J.V. Bradley, Robustness?, Br. J. Math. Stat. Psychol. 31 (1978), 144-152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x.
[28] W.G. Cochran, The χ^2 Test of Goodness of Fit, Ann. Math. Stat. 23 (1952), 315-345. https://doi.org/10.1214/aoms/1177729380.
[29] B. Choopradit, S. Wasinrat, Zero-one Inflated Bell Distribution and Its Application to Insurance Data, Int. J. Math. Comput. Sci. (2025), 625-635. https://doi.org/10.69793/ijmcs/02.2025/sirithip.