International Journal of Advanced Statistics and Probability, 13 (1) (2026) 73-80

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP
https://doi.org/10.14419/j2tta864
Research paper

Sensitivity and Robustness of Bartlett Test, Levene Test and Randomization Test for The Analysis of Completely Randomized Design

Uchenna Valentine Ikebuife [1, 2] *, Kenechi Jane Ezema [3]

[1] African Institute for Mathematical Sciences, South Africa
[2] Department of Mathematics, Stellenbosch University, South Africa
[3] Department of Statistics, University of Nigeria, Nsukka
**Corresponding author E-mail: valentine@aims.ac.za*

Abstract

This Monte Carlo simulation study evaluates the robustness and sensitivity of the Bartlett test (B), the Levene test (L), and their randomization-based counterparts (RB and RL) for testing homogeneity of variances in a completely randomized design. The study took into account normal and non-normal data (uniform, beta, lognormal, and gamma distributions), three treatment groups (t = 2, 3, and 5), two significance levels (α = 0.01 and 0.05), two variance ratios (1 and 2), and seven sample sizes (n =30,60,90,120,150,300, and 600). Type-I-error rates were assessed using Bradley's criterion of robustness, with power comparisons restricted to tests satisfying this condition. The randomization framework improved the performance of both test statistics by stabilizing small-sample performance and enhancing power without compromising Bradley's criterion of robustness. Bartlett test fails to maintain nominal type-I-error rates under non-normal data, exhibiting strong sensitivity to non-normality. The Levene test had control of the type-I error rate, with the randomization-based Levene test consistently maintaining robustness across all settings. In terms of power, RL outperforms B, RB, and L across most conditions, while RB shows improved performance over B but with some instability at small sample sizes. Hence, RL should be used whenever data are suspected to deviate from normality in CRD.

Keywords: *Robustness; Sensitivity; Bartlett Test; Levene Test; Randomization Test; Monte Carlo Simulation.*

1. Introduction

The Bartlett test is a likelihood-based procedure for testing the equality of variances across two or more populations. It is well established that the Bartlett test attains high power when the assumptions of normality and independence are satisfied (Conover et al., 1981; Wang et al., 2022). However, this efficiency is achieved at the cost of strong dependence on the normality assumption. In practice, the Bartlett test is highly sensitive to even mild departures from normality, including skewed and heavy-tailed distributions, resulting in severely inflated type-I-error rates (Allingham & Rayner, 2012; Hossain et al., 2021). This sensitivity makes the test unsuitable for analyzing data generated by experimental designs (Wang et al., 2022). Due to this limitation, the Levene test is often resorted to when testing homogeneity of variance in experimental designs.

The Levene test is a robust, analysis of variance (ANOVA)-based procedure for testing the equality of variances across two or more populations. Unlike the Bartlett test, the Levene test reduces sensitivity to departures from normality (Esmailzadeh, 2018; Lim & Loh, 1996). In situations where the distribution of the data is approximately symmetric with mild departures from normality, the mean-based Levene test performs reasonably well and provides a good balance between power and robustness (Baydili and Sığırlı, 2017). However, when the data are moderately skewed or contain mild outliers, the median-based Brown–Forsythe modification is generally preferred as it maintains type-I-error rates closer to the nominal level (Brown & Forsythe, 1974; Conover et al., 1981). In cases of severe non-normality, heavy-tailed distributions, or substantial contamination by outliers, trimmed-mean-based versions of the Levene test have been shown to provide further improvement in robustness (Gastwirth et al., 2009; Wilcox, 2012).

Several works have been done to compare the type-I-error rate and power of the mean-based, median-based (Brown-Forsythe), and trimmed-mean-based Levene tests with those of bootstrap and permutation-based Levene tests (Parra-Frutos, 2009; Cahoy, 2010; Li et al., 2015; Baydili & Sığırlı, 2017; Esmailzadeh, 2018; Abdullah and Muda, 2022; Yi et al., 2022). The foregoing works showed the superiority of resampling-based Levene tests over their parametric counterparts in controlling type-I-error rates across varying sample sizes and variance ratios (Cahoy, 2010; Baydili & Sığırlı, 2017; Yi et al., 2022), under skewed distributions (Li et al., 2015; Esmailzadeh, 2018), heavy-tailed distributions, and in the presence of outliers (Parra-Frutos, 2009; Abdullah and Muda, 2022). However, it is important to note that resampling Levene tests are not entirely free from inflation of type-I-error when group sizes are extremely small, and distributions are heavily skewed, though they still markedly outperform the parametric Levene variants under those conditions (Parra-Frutos, 2009; Wang et al., 2022).

Copyright © Uchenna Valentine Ikebuife, Kenechi Jane Ezema. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

While these resampling efforts have been directed at the various Levene modifications, no study has yet investigated whether placing the Bartlett test statistic within a randomization framework can alleviate its sensitivity to non-normality. Moreover, the randomization test based on the mean-based Levene statistic has not been compared directly with its parametric counterpart or with the randomization-based Bartlett test.

Table 1: Summarizes the Main Variance-Homogeneity Procedures Considered in the Literature and Clarifies the Gap Addressed in This Study

Procedure	Main strength	Main limitation and gap
Bartlett	High power under normality	Highly sensitive to non-normality
Levene	More robust to non-normality	May lose power under heteroscedasticity
Brown-Forsythe	Robust under skewness/outliers	Not the focus of the present comparison
Bootstrap Levene	Improves finite-sample behaviour	Less direct comparison with the Bartlett statistic
Randomization- based Levene test	Strong robustness-power balance	Needs simulation evidence across CRD settings
Randomization- based Bartlett test	Tests whether randomization reduces Bartlett sensitivity	May remain unstable under non-normality

In this work, we construct randomization tests using only the Bartlett statistic and the mean-based Levene statistic, and compare their type-I-error rate and power with those of the classic parametric versions. By doing so, we aim to determine whether a randomization framework offers a meaningful advantage for each test under the same wide range of conditions, varying sample sizes, significance levels, variance ratios, distribution shapes, and number of treatments. A Monte Carlo simulation study is therefore conducted to evaluate the robustness and sensitivity of the Bartlett test and the Levene test and their randomization-based counterparts, each constructed using the corresponding test statistic, for testing homogeneity of variances in a completely randomized design (CRD).

2. Methodology

The Bartlett test (B), Levene test (L), with their randomization-based counterparts (RB and RL), are discussed in this section.

2.1. Bartlett test

The Bartlett test is a likelihood-based test for testing the hypothesis of variance homogeneity. Let X_{ij} denote the j- The observation in the group i, where $i = 1, \dots, k$, and $j = 1, \dots, n_i$. Let n_i be the sample size of the group i. The null hypothesis is formulated as H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$, against H_1: At least one group variance differs. Let the sample variance of the group i be denoted by s_i^2, and the pooled variance by $s_p^2 = \frac{\sum_{i=1}^{k}(n_i-1)s_i^2}{N-k}$. The Bartlett test statistic is given by

$$B = \frac{(N-k)ln\,(s_p^2) - \sum_{i=1}^{k}(n_i-1)ln\,(s_i^2)}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{N-k}\right)}, \quad (1)$$

Where $s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2$ is the sample variance of the ithe group, $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ is the group mean, and $N = \sum_{i=1}^{k} n_i$ is the total sample size. Under the null hypothesis, B is asymptotically distributed as a chi-square random variable with $k-1$ degrees of freedom. The null hypothesis is rejected when $B > \chi^2_{1-\alpha,k-1}$. Although the Bartlett test is optimal in terms of power under normality, it is highly sensitive to deviations from normality, as departures such as skewness or heavy tails may induce rejection even in the absence of true variance differences (Vorapongsathorn et al., 2004; Odoi et al., 2019).

2.2. Levene test

The Levene test was introduced as a more robust alternative to address the sensitivity of the Bartlett test to non-normality. The test transforms observations into absolute deviations from a measure of central tendency. For each observation X_{ij}, define $Z_{ij} = \mid X_{ij} - \tilde{X}_i \mid$, where $\tilde{X}_i$ represents the group means, median, or trimmed mean. In this study, the mean-based Levene statistic was used so that the classical statistic and its randomization version were compared under the same statistic. Let $\bar{Z}_i$ denote the mean of Z_{ij} within group i, and $\bar{Z}$ The overall mean of all transformed observations. The Levene test statistic is given by

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

In this expression, $Z_{ij} = \left|X_{ij} - \bar{X}_i\right|$ are the absolute deviations from the group mean, $\bar{Z}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}$, and $\bar{Z} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}$ is the overall mean of the transformed observations. Under the null hypothesis of equal variances, L follows approximately an F-distribution with $k-1$ and $N-k$ degrees of freedom. The null hypothesis is rejected when $L > F_{1-\alpha,k-1,N-k}$. Unlike the Bartlett test, the Levene test does not rely on normality assumptions and therefore maintains a more stable type-I-error rate under skewed or heavy-tailed distributions, although this robustness may be accompanied by a modest reduction in power when normality holds (Vorapongsathorn et al., 2004; Li et al., 2015).

2.3. Randomization test

Randomization test is a distribution-free test that repeatedly shuffles the data to construct an empirical reference distribution for any test statistic (Edgington and Onghena, 2007). The randomization test was implemented as follows for a dataset with *N* observations allocated to *t* balanced treatment groups:

- Compute the observed Bartlett and Levene statistics from the original data and denote them by B_{obs} and L_{obs}, respectively.

- Randomly permute the treatment labels while keeping the observed response values fixed and preserving the original group sizes.
- Compute the Bartlett and Levene statistics for the permuted data, denoted by B_i and L_i for the ith permutation.
- Repeat the procedure for B = 10,000 random permutations. The R scripts used set.seed(12) for the type-I-error randomization runs and set. seed(123) for the power runs.
- obtain the p-values of RB and RL as the proportion of the statistics that are greater than or equal to B_{obs} and L_{obs}, respectively. The null hypothesis is rejected if the p-value is less than α.

The number of random permutations was set to $B = 10{,}000$ for all randomization tests. This value balances computational feasibility with adequate precision for Monte Carlo estimation of p-values. Because the total number of possible permutations $\frac{(tr)!}{(r!)^t}$ is extremely large for the sample sizes considered, exact enumeration was infeasible; therefore, p-values were computed as Monte Carlo estimates (Oladugba and Ikebuife, 2026). For reproducibility, the random seed was fixed at a set value. seed(12) for the type-I-error simulations and set. seed(123) for power simulations. Computational complexity is approximately. $O(B \times N)$ per replication, which remained manageable for all scenarios studied.

3. Monte Carlo Simulation

The Monte Carlo simulation was conducted to evaluate the robustness and sensitivity of the Bartlett test, Levene test, and their randomization-based counterparts, each constructed using their corresponding test statistic. The study considered both normal and non-normal data-generating processes, including Uniform, Gamma, Beta, and Lognormal distributions. Two significance levels (1% and 5%), three treatment conditions ($t = 2, 3, 5$), and seven balanced sample sizes per treatment group ($n = 30, 60, 90, 120, 150, 300,$ and 600) were examined. The following exact parameter settings were used for data generation under the null hypothesis of equal variances (variance ratio = 1): Normal: $X{\sim}N(\mu = 0, \sigma^2 = 1)$, Uniform: $X{\sim}U(a = 3, b = 9)$, Beta: $X{\sim}Beta(\alpha = 3, \beta = 2)$, Lognormal: $X{\sim}LN(\mu_{log} = 0, \sigma_{log} = 1)$, and Gamma: $X{\sim}GAM(\alpha = 1, \beta = 2)$. Under the alternative hypothesis (variance ratio = 1:2 for two groups, and 1:1.5:2 for three or more groups), the parameters were adjusted to induce heteroscedasticity while preserving the shape of the distribution. Let the variances of the t groups be denoted by $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$. Under the null hypothesis of homogeneity of variances, the variances were specified as $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$, corresponding to a variance ratio of $1{:}\,1{:}\cdots{:}\,1$. Empirical rejection rates under this setting were used to estimate the type-I-error rate. Under the alternative hypothesis, the variances were specified as $\sigma_1^2{:}\,\sigma_2^2{:}\cdots{:}\,\sigma_k^2 = 1{:}\,1.5{:}\cdots{:}\,2$. These unequal variance settings were used to evaluate the statistical power of the tests. Specifically, for normal and lognormal distributions, the standard deviation of one group was multiplied by $\sqrt{2}$ relative to the others; for uniform, the range was widened; for beta, the shape parameters were altered to increase variance; and for gamma, the scale parameter was modified accordingly. All simulations assumed balanced designs with equal group sizes. In CRD, observations were allocated equally across treatment groups, such that the number of replications per treatment was given by $r = \frac{n}{t}$, ensuring a balanced design for all simulation scenarios.

The R statistical package was used to implement the Monte Carlo simulation, with 10,000 permutations sampled from the $\frac{(tr)!}{(r!)^t}$ possible permutations for the randomization-based tests. For each error distribution, the experiment was replicated 10,000 times. In each replication, the simulated data were analyzed using the B, RB, L, and RL to estimate both the type-I-error rate and power in CRD. For reproducibility, the data-generating parameters used in the R scripts were as follows: normal data were generated with rnorm(mean = 0, sd = 1) under H_0, and with one group assigned sd = 2 under H_1; uniform data used runif(min = 3, max = 9) under H_0 and unequal bounds in the power setting; beta data used rbeta(shape_1 = 3, shape_2 = 2) under H_0 and unequal beta-shape parameters in the power setting; lognormal data used rlnorm(meanlog = 0, sdlog = 1) under H_0 and one group assigned sdlog = 2 under H_1; and gamma data used rgamma(shape = 1, rate = 2) under H_0 and unequal gamma parameters under H_1.

The Bradley's criterion of robustness was adopted in order to assess the robustness of each test. That is, a test will be considered robust if the type-I-error rate lies within the interval $\left(\alpha_0 - \frac{\alpha_0}{2}, \alpha_0 + \frac{\alpha_0}{2}\right)$, for a nominal significance level $\alpha = \alpha_0$, where $\alpha_0 \in \{0.01, 0.05\}$. A test is classified as liberal if the type-I-error rate is above. $\alpha_0 + \frac{\alpha_0}{2}$ and conservative if the type-I-error rate is below $\alpha_0 - \frac{\alpha_0}{2}$. Hence, the test that has the closest type-I-error rate to the nominal α_0, would be considered to be more robust in terms of type-I-error, and the test that has the larger power would be taken to be more powerful than the other if and only if it is robust based on Bradley's criterion of robustness (Yi et al., 2022).

4. Results

The type-I-error rates and power of the Bartlett test (B), randomization-based Bartlett test (RB), Levene test (L), and randomization-based Levene test (RL) were evaluated via Monte Carlo simulation at nominal significance levels α = 0.01 and α = 0.05 for t = 2, 3, and 5 treatment groups under normal and skewed error distributions. Robustness was assessed using Bradley's criterion, and power comparisons were made only among tests that satisfied the robustness condition. In Tables 2 and 3, the values in bold denote the p-values that satisfied Bradley's criterion of robustness, while in Tables 4 and 5, the values in bold denote the p-values that exceeded the power value of 80%. Tests that are not robust (liberal or conservative) are excluded from power comparisons.

Table 2: Type-I-Error Rate of Four Statistical Tests for t = 2, 3, and 5 Under Normal and Skewed Distributions at α = 0.01

Distribution		t = 2				t = 3				t = 5			
	n	B	RB	L	RL	B	RB	L	RL	B	RB	L	RL
Normal	30	**0.0126**	0.0214	0.0044	**0.0061**	**0.0109**	0.0186	**0.0058**	**0.0060**	**0.0085**	0.0134	**0.0092**	**0.0087**
	60	**0.0105**	0.0176	**0.0090**	**0.0080**	**0.0106**	0.0156	0.0047	**0.0067**	**0.0132**	0.0116	**0.0051**	**0.0060**
	90	**0.0101**	**0.0076**	**0.0069**	**0.0085**	**0.0105**	**0.0100**	**0.0070**	**0.0073**	**0.0100**	0.0084	0.0042	**0.0077**
	120	**0.0094**	**0.0074**	**0.0075**	**0.0090**	**0.0104**	**0.0084**	**0.0082**	**0.0095**	**0.0098**	0.0054	**0.0078**	**0.0055**
	150	**0.0103**	0.0043	**0.0097**	**0.0091**	**0.0089**	0.0038	**0.0081**	**0.0077**	**0.0086**	0.0027	**0.0077**	**0.0059**
	300	**0.0091**	**0.0093**	**0.0088**	**0.0109**	**0.0090**	**0.0066**	**0.0105**	**0.0083**	**0.0099**	0.0085	**0.0080**	**0.0100**
	600	**0.0071**	**0.0126**	**0.0093**	**0.0107**	**0.0103**	**0.0119**	**0.0094**	**0.0104**	**0.0087**	0.0128	**0.0096**	**0.0093**
Uniform	30	0.0008	0.0002	0.0033	**0.0052**	0.0006	0.0003	**0.0055**	**0.0052**	0.0015	0.0007	**0.0064**	**0.0074**
	60	0.0005	0.0000	0.0049	**0.0066**	0.0001	0.0000	**0.0061**	**0.0056**	0.0002	0.0002	0.0035	0.0020

	90	0.0001	0.0000	**0.0068**	0.0047	0.0002	0.0000	**0.0062**	0.0046	0.0002	0.0001	0.0038	0.0037
	120	0.0000	0.0000	**0.0080**	**0.0055**	0.0001	0.0000	**0.0061**	**0.0058**	0.0002	0.0001	0.0037	0.0042
	150	0.0001	0.0001	**0.0070**	**0.0064**	0.0000	0.0002	**0.0060**	**0.0055**	0.0000	0.0000	**0.0052**	0.0029
	300	0.0000	0.0000	**0.0086**	**0.0071**	0.0000	0.0000	**0.0084**	**0.0075**	0.0000	0.0000	**0.0058**	0.0047
	600	0.0002	0.0000	**0.0098**	**0.0099**	0.0000	0.0000	**0.0093**	**0.0081**	0.0000	0.0000	**0.0076**	**0.0079**
	30	0.0045	0.0021	0.0046	**0.0077**	0.0046	**0.0144**	**0.0055**	**0.0078**	0.0036	0.0095	**0.0104**	**0.0145**
	60	0.0029	**0.0143**	**0.0069**	**0.0089**	0.0024	**0.0083**	**0.0064**	**0.0069**	0.0020	0.0049	**0.0062**	**0.0051**
	90	0.0028	**0.0060**	**0.0080**	**0.0088**	0.0018	0.0040	**0.0071**	**0.0071**	0.0013	0.0031	**0.0054**	**0.0070**
Beta	120	0.0016	**0.0050**	**0.0093**	**0.0106**	0.0014	**0.0051**	**0.0076**	**0.0093**	0.0007	0.0039	**0.0062**	**0.0087**
	150	0.0020	0.0031	**0.0093**	**0.0085**	0.0013	0.0021	**0.0078**	**0.0081**	0.0015	0.0016	**0.0062**	**0.0068**
	300	0.0015	0.0024	**0.0103**	**0.0105**	0.0008	0.0018	**0.0073**	**0.0084**	0.0011	0.0010	**0.0088**	**0.0081**
	600	0.0016	0.0046	**0.0089**	**0.0113**	0.0011	0.0028	**0.0093**	**0.0095**	0.0011	0.0021	**0.0087**	**0.0089**
	30	0.3125	0.4873	**0.0052**	0.0000	0.4245	0.6352	**0.0105**	0.0025	0.5129	0.9767	0.0227	0.0336
	60	0.3977	0.2399	**0.0059**	0.0036	0.5542	0.5931	**0.0069**	**0.0074**	0.7092	0.9316	**0.0106**	**0.0108**
	90	0.4437	0.2427	**0.0052**	**0.0072**	0.6106	0.5079	**0.0073**	**0.0091**	0.7701	0.8423	**0.0073**	**0.0119**
Lognormal	120	0.4641	0.2250	**0.0068**	**0.0086**	0.6387	0.3592	**0.0069**	**0.0117**	0.8175	0.6898	**0.0075**	**0.0094**
	150	0.4731	0.2095	**0.0065**	**0.0106**	0.6672	0.3445	**0.0068**	**0.0090**	0.8421	0.6209	**0.0076**	**0.0108**
	300	0.5226	0.1637	**0.0074**	**0.0108**	0.7210	0.2570	**0.0087**	**0.0077**	0.8935	0.4417	**0.0087**	**0.0116**
	600	0.5695	0.4084	**0.0064**	**0.0109**	0.7717	0.6219	**0.0096**	**0.0089**	0.9299	0.8716	**0.0085**	**0.0083**
	30	0.1339	0.0626	**0.0083**	**0.0056**	0.1808	0.0254	**0.0115**	**0.0103**	0.2146	0.0220	0.0228	0.0218
	60	0.1580	0.0353	**0.0093**	**0.0097**	0.2251	0.0502	**0.0089**	**0.0096**	0.3269	0.0448	**0.0081**	**0.0096**
	90	0.1653	0.0240	**0.0088**	**0.0114**	0.2552	0.0301	**0.0091**	**0.0104**	0.3675	0.0462	**0.0101**	**0.0111**
Gamma	120	0.1737	0.0272	**0.0109**	**0.0122**	0.2640	0.0323	**0.0092**	**0.0109**	0.3898	0.0403	**0.0086**	**0.0104**
	150	0.1766	0.0235	**0.0098**	**0.0102**	0.2655	0.0319	**0.0082**	**0.0116**	0.4017	0.0361	**0.0103**	**0.0082**
	300	0.1808	0.2036	**0.0101**	**0.0080**	0.2862	0.3135	**0.0093**	**0.0102**	0.4486	0.4856	**0.0085**	**0.0095**
	600	0.1929	0.1544	**0.0104**	**0.0098**	0.2959	0.2487	**0.0097**	**0.0121**	0.4642	0.4171	**0.0091**	**0.0099**

Table 3: Type-I-Error Rate of Four Statistical Tests for t = 2, 3, and 5 Under Normal and Skewed Distributions at α = 0.05

Distribution		t = 2				t = 3				t = 5			
	n	B	RB	L	RL	B	RB	L	RL	B	RB	L	RL
	30	**0.0529**	0.0849	**0.0302**	**0.0408**	**0.0521**	0.0984	**0.0349**	**0.0396**	**0.0492**	0.0832	**0.0355**	**0.0411**
	60	**0.0521**	**0.0693**	**0.0451**	**0.0464**	**0.0492**	**0.0749**	**0.0389**	**0.0395**	**0.0486**	**0.0735**	**0.0303**	**0.0358**
	90	**0.0512**	**0.0515**	**0.0419**	**0.0453**	**0.0501**	**0.0535**	**0.0384**	**0.0422**	**0.0520**	**0.0512**	**0.0347**	**0.0361**
Normal	120	**0.0499**	**0.0465**	**0.0443**	**0.0489**	**0.0470**	**0.0438**	**0.0433**	**0.0460**	**0.0491**	**0.0429**	**0.0355**	**0.0382**
	150	**0.0475**	**0.0321**	**0.0442**	**0.0457**	**0.0489**	**0.0256**	**0.0428**	**0.0429**	**0.0490**	0.0215	**0.0369**	**0.0382**
	300	**0.0484**	**0.0477**	**0.0485**	**0.0477**	**0.0470**	**0.0446**	**0.0475**	**0.0437**	**0.0489**	**0.0476**	**0.0441**	**0.0492**
	600	**0.0461**	**0.0551**	**0.0481**	**0.0509**	**0.0502**	**0.0530**	**0.0478**	**0.0489**	**0.0504**	**0.0564**	**0.0451**	**0.0450**
	30	0.0055	0.0103	**0.0257**	**0.0323**	0.0072	0.0070	**0.0273**	**0.0324**	0.0111	0.0063	**0.0284**	**0.0345**
	60	0.0039	0.0035	**0.0340**	**0.0385**	0.0019	0.0025	**0.0295**	**0.0309**	0.0024	0.0013	0.0219	0.0205
	90	0.0027	0.0017	**0.0390**	**0.0300**	0.0015	0.0005	**0.0325**	**0.0286**	0.0017	0.0002	0.0223	0.0190
Uniform	120	0.0028	0.0021	**0.0394**	**0.0363**	0.0017	0.0010	**0.0338**	**0.0293**	0.0006	0.0008	0.0240	0.0209
	150	0.0031	0.0022	**0.0422**	**0.0391**	0.0012	0.0013	**0.0360**	**0.0312**	0.0005	0.0002	**0.0259**	0.0246
	300	0.0021	0.0014	**0.0462**	**0.0428**	0.0009	0.0002	**0.0420**	**0.0377**	0.0002	0.0001	**0.0332**	**0.0279**
	600	0.0026	0.0019	**0.0504**	**0.0457**	0.0006	0.0004	**0.0492**	**0.0466**	0.0002	0.0001	**0.0332**	**0.0279**
	30	**0.0250**	0.0922	**0.0319**	**0.0457**	0.0240	0.0940	**0.0270**	**0.0444**	0.0246	**0.0716**	**0.0359**	**0.0480**
	60	0.0214	**0.0692**	**0.0401**	**0.0430**	0.0169	**0.0586**	**0.0390**	**0.0364**	0.0151	**0.0483**	**0.0303**	**0.0322**
	90	0.0212	**0.0415**	**0.0415**	**0.0456**	0.0143	**0.0364**	**0.0395**	**0.0407**	0.0129	**0.0303**	**0.0313**	**0.0388**
Beta	120	0.0179	**0.0348**	**0.0443**	**0.0453**	0.0142	**0.0333**	**0.0420**	**0.0480**	0.0114	**0.0270**	**0.0317**	**0.0433**
	150	0.0174	**0.0252**	**0.0455**	**0.0460**	0.0152	0.0199	**0.0433**	**0.0410**	0.0105	0.0159	**0.0380**	**0.0406**
	300	0.0197	0.0189	**0.0515**	**0.0473**	0.0129	0.0129	**0.0435**	**0.0443**	0.0091	0.0102	**0.0424**	**0.0457**
	600	0.0174	**0.0290**	**0.0495**	**0.0493**	0.0124	0.0241	**0.0444**	**0.0487**	0.0074	0.0195	**0.0481**	**0.0456**
	30	0.4504	0.5001	**0.0376**	**0.0323**	0.5829	0.8773	**0.0399**	**0.0264**	0.6839	1.0000	**0.0553**	**0.0486**
	60	0.5210	0.3069	**0.0401**	**0.0572**	0.6810	0.7552	**0.0372**	**0.0399**	0.8224	0.9842	**0.0384**	**0.0441**
	90	0.5617	0.4476	**0.0422**	**0.0486**	0.7262	0.5424	**0.0389**	**0.0455**	0.8641	0.9170	**0.0366**	**0.0474**
Lognormal	120	0.5748	0.3485	**0.0488**	**0.0504**	0.7469	0.5051	**0.0445**	**0.0490**	0.8939	0.7977	**0.0425**	**0.0451**
	150	0.5846	0.3364	**0.0450**	**0.0530**	0.7708	0.5004	**0.0389**	**0.0472**	0.9055	0.7490	**0.0422**	**0.0477**
	300	0.6257	0.2859	**0.0461**	**0.0529**	0.8069	0.4149	**0.0456**	**0.0448**	0.9401	0.6161	**0.0425**	**0.0530**
	600	0.6600	0.5293	**0.0479**	**0.0515**	0.8480	0.7351	**0.0451**	**0.0479**	0.9604	0.9276	**0.0429**	**0.0433**
	30	0.2631	0.1569	**0.0434**	**0.0506**	0.3334	0.1774	**0.0488**	**0.0402**	0.4042	0.1230	**0.0633**	**0.0583**
	60	0.2841	0.1208	**0.0480**	**0.0456**	0.3809	0.1635	**0.0443**	**0.0486**	0.5137	0.1876	**0.0414**	**0.0416**
	90	0.2905	0.0928	**0.0484**	**0.0548**	0.4123	0.1044	**0.0420**	**0.0506**	0.5536	0.1486	**0.0450**	**0.0475**
Gamma	120	0.3028	0.0933	**0.0488**	**0.0535**	0.4212	0.1089	**0.0459**	**0.0496**	0.5720	0.1448	**0.0424**	**0.0435**
	150	0.3042	0.0952	**0.0453**	**0.0486**	0.4235	0.1079	**0.0494**	**0.0516**	0.5794	0.1338	**0.0439**	**0.0418**
	300	0.2999	0.3516	**0.0474**	**0.0491**	0.4449	0.4986	**0.0467**	**0.0528**	0.6170	0.6998	**0.0484**	**0.0477**
	600	0.3194	0.2779	**0.0502**	**0.0508**	0.4495	0.4056	**0.0510**	**0.0527**	0.6354	0.5958	**0.0473**	**0.0465**

Table 4: Power of Four Statistical Tests for t = 2, 3, and 5 Under Normal and Skewed Distributions at α = 0.01

Distribution		$\sigma_1^2 : \sigma_2^2 = 1:2$ t = 2				$\sigma_1^2 : \sigma_2^2 : \sigma_3^2 = 1:1.5:2$ t = 3				$\sigma_1^2 : \sigma_2^2 : \ldots \sigma_5^2 = 1:1.5 \ldots :2$ t = 5			
	n	B	RB	L	RL	B	RB	L	RL	B	RB	L	RL
	30	**0.0826**	0.9857	0.0376	**0.9950**	0.0351	0.9858	0.0173	**0.9943**	0.0192	0.9879	0.0132	**0.9917**
	60	0.2206	0.9838	0.1446	**0.9931**	0.0859	0.9878	0.0464	**0.9941**	0.0420	0.9866	0.0181	**0.9942**
	90	0.3714	**0.9856**	0.2682	**0.9934**	0.1557	**0.9830**	0.0921	**0.9924**	0.0672	0.9853	0.0314	**0.9937**
Normal	120	0.5187	**0.9854**	0.4015	**0.9912**	0.2243	**0.9842**	0.1534	**0.9901**	0.1046	0.9865	0.0578	**0.9936**
	150	0.6390	0.9831	0.5336	**0.9913**	0.3193	0.9857	0.2265	**0.9928**	0.1477	0.9869	0.0810	**0.9937**
	300	**0.9453**	**0.9816**	**0.8960**	**0.9902**	0.7035	**0.9849**	0.5840	**0.9909**	0.4187	0.9823	0.2983	**0.9906**
	600	**0.9997**	**0.9835**	**0.9987**	**0.9902**	**0.9747**	**0.9848**	**0.9456**	**0.9907**	**0.8566**	0.9853	0.7481	**0.9915**
	30	0.0232	0.9989	0.0514	**0.9963**	0.0055	0.9992	0.0236	**0.9942**	0.0023	0.9980	0.0148	**0.9934**
Uniform	60	0.1180	0.9998	0.2441	**0.9929**	0.0137	0.9999	0.0696	**0.9945**	0.0020	0.9994	0.0202	0.9969

	90	0.3017	0.9991	0.4661	0.9921	0.0375	0.9999	0.1649	0.9943	0.0038	0.9998	0.0435	0.9963
	120	0.5322	0.9996	0.6664	**0.9922**	0.0983	0.9999	0.2883	**0.9950**	0.0098	0.9998	0.0817	0.9955
	150	0.7197	0.9994	**0.7909**	**0.9908**	0.1863	1	0.4224	**1**	0.0192	0.9999	0.1414	0.9962
	300	0.9957	0.9998	**0.9927**	**0.9913**	0.7592	0.9999	**0.8676**	**0.9903**	0.2237	1	0.5625	0.9938
	600	1	0.9997	**1**	**0.9915**	0.9988	0.9999	**0.9984**	**0.9908**	0.8927	1	**0.9624**	**0.9928**
	30	0.0112	0.9884	0.0123	**0.9857**	0.0397	**0.9582**	0.0251	**0.9757**	0.1601	0.8337	0.0322	**0.9659**
	60	0.0031	**0.9977**	0.0198	**0.9841**	0.0066	**0.9940**	0.0179	**0.9820**	0.0325	0.9713	0.0164	**0.9844**
	90	0.0023	**0.9986**	0.0161	**0.9813**	0.0035	0.9977	0.0175	**0.9854**	0.0088	0.9906	0.0148	**0.9857**
Beta	120	0.0020	**0.9996**	0.0192	**0.9850**	0.0012	**0.9986**	0.0155	**0.9848**	0.0031	0.9961	0.0146	**0.9864**
	150	0.0012	0.9995	0.0172	**0.9838**	0.0010	0.9995	0.0163	**0.9844**	0.0021	0.9977	0.0127	**0.9867**
	300	0.0021	1	0.0158	**0.9842**	0.0001	0.9998	0.0141	**0.9875**	0.0005	0.9995	0.0115	**0.9881**
	600	0.0059	0.9999	0.0167	**0.9893**	0.0004	**0.9999**	0.0135	**0.9882**	0.0002	0.9999	0.0138	**0.9875**
	30	0.3168	0.7049	0.0061	0.9938	0.3980	0.6060	0.0103	0.9908	0.5068	0.5065	0.0224	0.9765
	60	0.4074	0.6337	0.0052	0.9943	0.5341	0.4726	0.0077	**0.9932**	0.7022	0.3176	0.0110	**0.9905**
	90	0.4734	0.5802	0.0078	**0.9935**	0.6012	0.4205	0.0082	**0.9936**	0.7754	0.2480	0.0099	**0.9926**
Lognormal	120	0.5047	0.5491	0.0096	**0.9943**	0.6319	0.3844	0.0075	**0.9926**	0.8074	0.2016	0.0069	**0.9934**
	150	0.5426	0.5361	0.0117	**0.9941**	0.6737	0.3472	0.0087	**0.9924**	0.8329	0.1756	0.0105	**0.9925**
	300	0.6564	0.4808	0.0268	**0.9929**	0.7619	0.2910	0.0129	**0.9922**	0.8964	0.1223	0.0114	**0.9945**
	600	0.7720	0.4412	0.0675	**0.9914**	0.8410	0.2467	0.0280	**0.9923**	0.9402	0.0790	0.0193	**0.9927**
	30	0.7512	0.7821	0.2770	**0.9904**	0.6560	0.6677	0.1388	**0.9867**	0.6363	0.5261	0.0971	0.9741
	60	0.9291	0.7512	0.7342	**0.9920**	0.8648	0.6103	0.4146	**0.9920**	0.8559	0.3927	0.2368	**0.9896**
	90	0.9802	0.7383	**0.9360**	**0.9914**	0.9414	0.5791	0.6833	**0.9912**	0.9329	0.3493	0.4413	**0.9917**
Gamma	120	0.9937	0.7348	**0.9880**	**0.9920**	0.9758	0.5731	**0.8497**	**0.9927**	0.9676	0.3098	0.6464	**0.9925**
	150	0.9987	0.7317	**0.9987**	**0.9905**	0.9883	0.5502	**0.9395**	**0.9918**	0.9862	0.2941	0.7870	**0.9906**
	300	1	0.7147	**1**	**0.9910**	0.9996	0.5246	**0.9996**	**0.9921**	0.9995	0.2607	**0.9923**	**0.9923**
	600	1	0.7148	**1**	**0.9919**	1	0.5040	**1**	**0.9911**	1	0.2290	**1**	**0.9895**

Table 5: Power of Four Statistical Tests for t = 2, 3, and 5 Under Normal and Skewed Distributions at α = 0.05

Distribution		$\sigma_1^2 : \sigma_2^2 = 1:2$ $t = 2$				$\sigma_1^2 : \sigma_2^2 : \sigma_3^2 = 1:1:2$ $t = 3$				$\sigma_1^2 : \sigma_2^2 : \ldots \sigma_5^2 = 1:1 \ldots :2$ $t = 5$			
	n	B	RB	L	RL	B	RB	L	RL	B	RB	L	RL
	30	0.2259	0.9384	0.1524	**0.9691**	0.1226	0.9387	0.0748	**0.9683**	0.0830	0.9410	0.0530	**0.9624**
	60	0.4535	**0.9358**	0.3593	**0.9559**	0.2352	**0.9376**	0.1615	**0.9638**	0.1371	**0.9370**	0.0766	**0.9708**
	90	0.6297	**0.9340**	0.5320	**0.9566**	0.3548	**0.9327**	0.2645	**0.9586**	0.2007	**0.9365**	0.1236	**0.9631**
Normal	120	0.7560	**0.9321**	0.6756	**0.9551**	0.4717	**0.9389**	0.3764	**0.9616**	0.2714	**0.9371**	0.1831	**0.9645**
	150	**0.8467**	**0.9321**	0.7731	**0.9538**	0.5586	**0.9355**	0.4598	**0.9550**	0.3382	0.9327	0.2390	**0.9617**
	300	**0.9853**	**0.9284**	**0.9713**	**0.9525**	**0.8785**	**0.9339**	**0.8170**	**0.9547**	0.6774	**0.9335**	0.5637	**0.9574**
	600	**1**	**0.9281**	**0.9996**	**0.9489**	**0.9940**	**0.9340**	0.9861	**0.9524**	**0.9538**	**0.9328**	**0.9090**	**0.9535**
	30	0.1394	0.9879	0.1949	**0.9743**	0.0354	0.9914	0.0868	**0.9730**	0.0172	0.9890	0.0528	**0.9695**
	60	0.4280	0.9930	0.5170	**0.9649**	0.0930	0.9950	0.2209	**0.9685**	0.0194	0.9962	0.0823	0.9782
	90	0.6856	0.9935	0.7320	**0.9607**	0.2063	0.9962	0.3817	**0.9664**	0.0381	0.9982	0.1551	0.9776
Uniform	120	0.8561	0.9936	**0.8610**	**0.9587**	0.3623	0.9972	0.5460	**0.9632**	0.0663	0.9981	0.2442	0.9734
	150	0.9423	0.9926	**0.9327**	**0.9558**	0.5156	0.9974	0.6810	**0.9574**	0.1192	0.9987	0.3554	0.9735
	300	0.9996	0.9945	**0.9989**	**0.9561**	0.9489	0.9977	**0.9609**	**0.9546**	0.5788	0.9997	**0.8025**	**0.9638**
	600	1	0.9942	**1**	**0.9483**	0.9999	0.9979	**0.9995**	**0.9550**	0.9861	0.9995	**0.9912**	**0.9565**
	30	0.0266	0.9736	0.0471	**0.9540**	0.0636	0.9400	0.0632	**0.9418**	0.2161	0.7757	0.0775	**0.9226**
	60	0.0135	**0.9904**	0.0586	**0.9451**	0.0179	**0.9832**	0.0552	**0.9447**	0.0511	**0.9538**	0.0527	**0.9474**
	90	0.0108	**0.9926**	0.0623	**0.9463**	0.0110	**0.9909**	0.0621	**0.9444**	0.0188	**0.9827**	0.0508	**0.9521**
Beta	120	0.0121	**0.9951**	0.0621	**0.9433**	0.0076	**0.9939**	0.0615	**0.9452**	0.0100	**0.9884**	0.0532	**0.9482**
	150	0.0116	**0.9953**	0.0655	**0.9427**	0.0065	0.9952	0.0586	**0.9429**	0.0065	0.9936	0.0531	**0.9497**
	300	0.0208	0.9974	0.0623	**0.9459**	0.0042	0.9987	0.0569	**0.9496**	0.0031	0.9984	0.0532	**0.9447**
	600	0.0492	**0.9974**	0.0652	**0.9460**	0.0081	0.9992	0.0606	**0.9461**	0.0019	0.9995	0.0550	**0.9468**
	30	0.4534	0.5694	0.0356	**0.9626**	0.5586	0.4526	0.0408	**0.9642**	0.6722	0.3355	0.0602	**0.9409**
	60	0.5291	0.5085	0.0448	**0.9629**	0.6640	0.3387	0.0430	**0.9610**	0.8188	0.1972	0.0408	**0.9617**
	90	0.5811	0.4568	0.0531	**0.9583**	0.7188	0.2979	0.0457	**0.9620**	0.8639	0.1527	0.0440	**0.9636**
Lognormal	120	0.6137	0.4355	0.0620	**0.9557**	0.7389	0.2716	0.0505	**0.9597**	0.8855	0.1183	0.0435	**0.9608**
	150	0.6447	0.4240	0.0725	**0.9565**	0.7747	0.2439	0.0512	**0.9561**	0.9041	0.0998	0.0457	**0.9604**
	300	0.7354	0.3761	0.1195	**0.9559**	0.8385	0.1988	0.0745	**0.9564**	0.9417	0.0693	0.0537	**0.9619**
	600	0.8341	0.3410	0.2186	**0.9546**	0.8932	0.1696	0.1160	**0.9559**	0.9682	0.0451	0.0815	**0.9533**
	30	0.8470	0.6432	0.6073	**0.9511**	0.7866	0.4981	0.3296	**0.9549**	0.7862	0.3431	0.2151	**0.9317**
	60	0.9631	0.6070	**0.9202**	**0.9489**	0.9249	0.4577	0.6846	**0.9561**	0.9284	0.2469	0.4618	**0.9556**
	90	0.9906	0.6083	**0.9896**	**0.9502**	0.9684	0.4273	**0.8764**	**0.9526**	0.9691	0.2147	0.6840	**0.9572**
Gamma	120	0.9969	0.6025	**0.9986**	**0.9514**	0.9887	0.4217	**0.9564**	**0.9562**	0.9866	0.1918	**0.8348**	**0.9554**
	150	0.9994	0.6016	**1**	**0.9514**	0.9951	0.4027	**0.9872**	**0.9507**	0.9942	0.1810	**0.9210**	**0.9570**
	300	1	0.5885	**1**	**0.9491**	0.9999	0.3815	**0.9999**	**0.9497**	1	0.1555	**0.9991**	**0.9545**
	600	1	0.5898	**1**	**0.9513**	1	0.3700	**1**	**0.9512**	1	0.1413	**1**	**0.9516**

4.1. Normal distribution

From Table 2, it can be seen that, under normality, the Bartlett test maintained type-I-error rates within Bradley's robustness interval at both significance levels for all sample sizes and treatment numbers. The randomization-based Bartlett test was liberal at very small sample sizes but became robust once n reached approximately 90. In contrast, the Levene test and the randomization-based Levene test were predominantly robust, with type-I-error rates frequently falling within Bradley's robustness interval at both significance levels. Table 3 shows that B, RB, L, and RL were robust at $\alpha = 0.05$, although they approached the nominal level as the number of treatments increased. Among the robust tests, the power of RB and RL was substantially higher than that of B and L across all variance heterogeneity settings (Tables 4 and 5).

4.2. Uniform distribution

Under the uniform distribution, both B and RB were conservative across all sample sizes and significance levels (Tables 2 and 3), with type-I-error rates falling far below the lower bound of Bradley's robustness interval. Consequently, these tests are not robust for uniform data. L and RL attained robustness with moderate to large sample sizes. For $\alpha = 0.01$, L and RL lie within Bradley's robustness interval when $n \geq 90$ for $t = 2$ and $t = 3$, and for all n at $t = 5$. At $\alpha = 0.05$, L and RL were robust for $n \geq 60$, with RL exhibiting type-I-error rates closer to the nominal level. Since only L and RL can be regarded as robust, the power comparison is restricted to these two tests. In every scenario, RL was more powerful than L (Tables 4 and 5).

4.3. Beta distribution

For data generated from a beta distribution, B was generally conservative at both $\alpha = 0.01$ and $\alpha = 0.05$, particularly for $t = 3$ and 5, with type-I-error rates frequently below the lower bound of Bradley's robustness interval (Tables 2 and 3). RB exhibited unstable behaviour: it was liberal at small n but became robust or even conservative as sample sizes grew. In contrast, L and RL were robust across nearly all conditions: their type-I-error rates lie within Bradley's robustness interval for all t and for $n \geq 60$ at $\alpha = 0.01$ and for all n at $\alpha = 0.05$. Among the robust tests, RL was consistently the most powerful, outperforming the others by a considerable margin (Tables 4 and 5). The power of L remained close to the nominal α even for large samples, indicating very low sensitivity under the beta distribution, whereas RL maintained high and rapidly increasing power with sample size.

4.4. Lognormal distribution

Under the lognormal distribution, B and RB were grossly liberal at both significance levels, with type-I-error rates frequently above the upper bound of Bradley's robustness interval (Tables 2 and 3). Consequently, B and RB are not robust for lognormal data. L and RL maintained robust type-I-error control. At $\alpha = 0.01$, L was robust for $t = 2$ and $t = 3$ across all n, while for $t = 5$ it was slightly liberal at the smallest sample size ($n = 30$) but robust thereafter. RL was conservative at very small n but became robust as n increased. At $\alpha = 0.05$, both L and RL were robust for all t and $n \geq 60$. Because B and RB are not robust, power comparisons are restricted to L and RL. Table 4 shows that under $\alpha = 0.01$, the power of L was very low, barely exceeding the nominal level even at large sample sizes, whereas RL attained higher power for the same n. A similar pattern is observed at $\alpha = 0.05$ (Table 5), where L exhibited only modest power gains with increasing. n, while RL was more powerful. Overall, RL therefore demonstrated clear superiority.

4.5. Gamma distribution

With gamma-distributed errors, B and RB were consistently liberal at both significance levels (Tables 2 and 3). L was robust in most settings, but was liberal for $t = 5$ ($\alpha = 0.01$ and $n = 30$). RL remained robust across all conditions. Among the robust tests, RL was consistently more powerful than L (Tables 4 and 5). The power of RL approached 1 more rapidly than that of L, confirming its greater sensitivity in detecting unequal variances under gamma distributions.

Across all distributions, a consistent pattern emerges: the Bartlett test is optimal only under strict normality, RB provides partial correction but remains unstable in small samples, L ensures robustness at the cost of power loss, and RL achieves the best overall balance between robustness and sensitivity. The dominance of RL across skewed, heavy-tailed, and light-tailed distributions demonstrates that embedding Levene-type statistics within a randomization framework significantly improves inferential performance.

5. Discussion of Results

The evaluation of statistical tests for equality of variances under varying experimental conditions is essential for identifying methods that remain reliable when distributional assumptions are violated. Under normality, the Bartlett test consistently maintained type-I-error rates within Bradley's robustness interval, confirming its well-known optimality under Gaussian assumptions (Conover et al., 1981; Vorapongsathorn et al., 2004). However, the randomization-based Bartlett test showed instability at small sample sizes before stabilizing as the sample size increased, indicating that while the randomization test makes the errors at small sample sizes less severe, it does not eliminate them. In contrast, both L and RL tended to be conservative, particularly for small numbers of treatment groups. This aligns with Lee et al. (2010), who similarly reported that Levene-type tests often under-reject under normality. Despite this conservatism, RL consistently outperformed all other tests in power, showing that the randomization framework substantially enhances sensitivity even when classical assumptions already hold.

The theoretical reasons for the superior performance of the randomization-based Levene test merit further explanation. Randomization tests are distribution-free: they construct an empirical reference distribution by permuting treatment labels, thereby preserving the actual error distribution of the data. This property makes RL robust to skewness, heavy tails, and other non-normal features because the permutation distribution automatically adapts to the observed data structure. In contrast, classical Bartlett and Levene tests rely on asymptotic chi-square or F-approximations, which fail under non-normality. The Levene statistic, based on absolute deviations, is already more resistant to outliers than the squared deviations used in Bartlett; embedding it within a randomization framework further removes parametric assumptions, stabilizes small-sample behaviour, and enhances power without inflating type-I-error. The instability of the randomization-based Bartlett test under non-normality arises because the Bartlett statistic itself remains highly sensitive to skewed or heavy-tailed distributions; randomization cannot compensate for a fundamentally non-robust test statistic.

Under the uniform distribution, B and RB became conservative, failing to maintain nominal type-I-error rates. This behaviour is consistent with known limitations of variance-based likelihood methods under light-tailed distributions. In contrast, both L and RL achieved robustness with moderate to large sample sizes, confirming findings by Esmailzadeh (2018) and Baydili and Sığırlı (2017), who showed that Levene-type procedures are more stable under non-normal light-tailed settings. For beta-distributed data, both L and RL were robust across most scenarios, but RL consistently exhibited higher power. This agrees with findings by Li et al. (2015) and Yonar et al. (2024), who reported that classical Levene-type tests often become conservative under skewed distributions, while resampling methods offer greater power in detecting variance differences.

Across all non-normal distributions, RL uniquely maintained both robustness and high power, whereas B and RB failed to control type-I-error, and L, though robust, was substantially less powerful. This agrees with studies (Allingham & Rayner, 2012; Hossain et al., 2021;

Patrick & Ahmed, 2024), which documented severe inflation of type-I-error for Bartlett-type tests under non-normality. In contrast, L and RL maintained acceptable type-I-error control, with RL stabilizing more rapidly across increasing sample sizes. However, the most striking result was the pronounced power gap between RL and L: while L remained close to the nominal level even under substantial heteroscedasticity, RL achieved near-perfect detection as sample size increased. This pattern is consistent with Cahoy (2010) and Yi et al. (2022), who reported that resampling-based variance tests provide superior sensitivity while preserving robustness under complex distributional structures.

In summary, RL emerged as the most reliable and efficient method, consistently maintained type-I-error control within Bradley's robustness bounds while achieving the highest power across all conditions. These findings reinforce earlier conclusions in the literature that resampling-based methods improve robustness under model misspecification (Esmailzadeh, 2018; Cahoy, 2010; Yi et al., 2022), but extend them by demonstrating that the combination of Levene statistics with randomization provides a stronger and more stable alternative to both classical and randomization-based Bartlett approaches.

6. Conclusion

This study provides a systematic evaluation of classical and randomization-based procedures for testing homogeneity of variances under a wide range of distributional settings in a completely randomized design. By integrating Bradley's robustness criterion with power-based comparisons, the analysis establishes a clear distinction between methods that are theoretically optimal under restrictive assumptions and those that remain reliable in practice when such assumptions are violated.

The findings highlight that methodological performance is fundamentally driven by the interaction between distributional shape and sample size. In particular, tests derived from likelihood-based frameworks exhibit strong sensitivity to departures from normality, while rank- and deviation-based tests offer greater stability but may sacrifice efficiency. The introduction of randomization alters this trade-off by improving finite-sample behaviour and enhancing the ability to detect variance heterogeneity without relying on asymptotic approximations.

A key contribution of this work is the demonstration that embedding the Levene statistic within a randomization framework improves the robustness-power trade-off that characterizes variance testing. Rather than merely improving one aspect of performance, this approach delivers a balanced procedure that maintains valid inference while substantially increasing sensitivity across light-tailed, skewed, and heavy-tailed distributions. These results showed that the classical Bartlett test should only be used under clear evidence of normality and small sample sizes; otherwise, its type-I error is unreliable. Randomization-based Bartlett test is not recommended for routine use because its performance remains unstable under non-normal conditions. In this regard, the randomization-based Levene test remains valid but may be slightly conservative for small sample sizes.

The study is limited to balanced completely randomized designs, a finite set of distributional forms, and a single unequal-dispersion pattern in which one treatment group has larger dispersion. Future research should extend the comparison to unequal group sizes, additional variance ratios, heavy-tailed and contaminated distributions, randomized complete block designs, multivariate settings, missing observations, and explicit outlier mechanisms. Such extensions would provide a more comprehensive assessment of how these factors influence the robustness and sensitivity of the tests, particularly under conditions that deviate from ideal experimental settings.

References

[1] Abdullah, N. F., & Muda, N. (2022). An overview of homogeneity of variance tests on various conditions based on the Type I error rate and power of a test. *Journal of Quality Measurement and Analysis*, *18*(3),111–130.

[2] Allingham, D., & Rayner, J. C. W. (2012). Testing equality of variances for multiple univariate normal populations. *Journal of Statistical Theory and Practice*,*6*(3), 524–535. https://doi.org/10.1080/15598608.2012.695703.

[3] Baydili, İ., & Sığırlı, D. (2017). Comparison of bootstrap and permutation tests for equality of variances.

[4] *Turkiye Klinikleri Journal of Biostatistics*, *9*(2), 120–128.

[5] Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(346), 364–367. https://doi.org/10.1080/01621459.1974.10482955.

[6] Cahoy, D. O. (2010). A bootstrap test for equality of variances. *Computational Statistics & Data Analysis*,*54*(10), 2306–2316. https://doi.org/10.1016/j.csda.2010.04.012.

[7] Conover, W. J., Johnson, M. E., Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics.23*(4), 351-361. https://doi.org/10.1080/00401706.1981.10487680.

[8] Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9781420011814.

[9] Esmailzadeh, N. (2018). A comparison of five bootstrap and non-bootstrap Levene-type tests of homogeneity of variances. *Iranian Journal of Science and Technology Transactions Science,43*(3), 979–989. https://doi.org/10.1007/s40995-018-0485-0.

[10] 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. https://doi.org/10.1214/09-STS301.

[11] Hossain, K. S. M., Khan, M. A., & Sultana, N. (2021). The robustness of variance homogeneity tests under non-normality: A simulation study. *Journal of Statistical Theory and Practice*, *15*(4), 88.

[12] 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), 359–369. https://doi.org/10.3844/jmssp.2010.359.366.

[13] Li, X., Qiu, W., Morrow, J., DeMeo, D. L., Weiss, S. T., Fu, Y., & Wang, X. (2015). A comparative study of tests for homogeneity of variances with application to DNA methylation data. *PLoS ONE*, *10*(12), e0145295. https://doi.org/10.1371/journal.pone.0145295.

[14] Lim, T. S., & Loh, W.Y. (1996). A comparison of tests of equality of variances. *Computational Statistics & Data Analysis*, *22*(3), 287–301. https://doi.org/10.1016/0167-9473(95)00054-2.

[15] Odoi, B., Samita, S., Al-Hassan, S., & Twumasi-Ankrah, S. (2019). Efficiency of Bartlett and Levene's tests for testing homogeneity of variance under varying number of replicates and groups in one- way ANOVA. *International Journal of Innovative Technology and Exploring Engineering*, *8*(6S4), 2278-3075. https://doi.org/10.35940/ijitee.F1250.0486S419.

[16] Oladugba, A. V., & Ikebuife, U. V. (2026). A Monte Carlo simulation study of the type-I-error and power of ANOVA F-test, Friedman test and randomization test for randomized complete block design. *Qua lity & Quantity*,*60*(1),2781–2803. https://doi.org/10.1007/s11135-025-02374-6.

[17] Parra-Frutos, I. (2009). A bootstrap test for the equality of variances. *Computational Statistics & Data Analysis*, *53*(9), 3435–3446.

[18] Patrick, A. O., & Ahmed, I. (2024). Determine robust procedure for testing variance equality using Type I error rate and power. *Malaysian Journal of Applied Sciences*,*9*(2), 68–76.

[19] Vorapongsathorn, T., Taejaroenkul, S. & Viwatwongkasem, C. (2004). A comparison of type I error and power of Bartlett's test, Levene's test and Cochran's test under violation of assumptions. https://doi.org/10.37231/myjas.2024.9.2.397.

[20] *Songklanakarin Journal of Science and Technology,26*(4), 537–547.

[21] Wang, Y., Chen, X., & Li, H. (2022). A comprehensive comparison of tests for homogeneity of variances in one-way ANOVA. *Statistical Papers*, *63*(5), 1521–1541.
[22] Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Academic Press. https://doi.org/10.1016/B978-0-12-386983-8.00001-9.
[23] Yi, Z., Chen, Y. H., Yin, Y., Cheng, K., Wang, Y., Nguyen, D., Pham, T., & Kim, E. S. (2020). Brief research report: A comparison of robust tests for homogeneity of variance in factorial ANOVA. *The Journal of Experimental Education*.,*90*(2), 505–520. https://doi.org/10.1080/00220973.2020.1789833.
[24] Yonar, A., Yonar, H., Demirsöz, M., & Tekindal, M. A. (2024). A comparative analysis for homogene ity of variance tests. *Journal of Science and Arts*,*24*(2), 305–328. https://doi.org/10.46939/J.Sci.Arts-24.2-a06.