International Journal of Assessment Tools in Education
2025, Vol. 12, No. 4, 1170–1185
https://doi.org/10.21449/ijate.1606406
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Investigating homogeneity of variance in normal, skewed-normal, and gamma distributions: A simulation study

Serpil Çelikten-Demirel[1*], Ayşenur Erdemir[2], Esra Oyar[3], Tuba Gündüz[4]

[1]Dicle University, Ziya Gokalp Faculty of Education, Department of Educational Sciences, Diyarbakır, Türkiye
[2]Turkish National Police Academy, Institute of Forensic Sciences, Department of Forensic Psychology, Ankara, Türkiye
[3]Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[4]Mugla Sitki Kocman University, Faculty of Education, Department of Educational Sciences, Muğla, Türkiye

ARTICLE HISTORY

Received: Dec. 31, 2024

Accepted: Aug. 28, 2025

Keywords:

Homogeneity of variances,
Bartlett's test,
Levene's test,
Brown-Forsythe test,
Fligner-Killeen test.

Abstract: It is an important point to test the homogeneity of variances in statistical methods such as the *t*-test or *F*-test used to make comparisons between groups. An erroneous decision regarding the homogeneity of variances will affect the test to be selected and thus lead to different results. For this reason, there are many tests for homogeneity of variance in the literature. This study aims to examine the type I error and power ratios of Levene, Bartlett, Brown-Forsythe, and Fligner-Killeen tests under different conditions. In this study, conducted within the scope of basic research, analyses were performed using simulated data. The simulation conditions included variance ratio (1:1, 1:2, 1:3, 2:1, 3:1), distributions (normal, skewed-normal, gamma), sample sizes (60, 120, and 240), and ratio of group sizes (1/1, 1/2, 1/4, 1/9). According to the study results, when controlling for type I error is a primary concern, the Brown–Forsythe and Fligner–Killeen tests are recommended, particularly under non-normal distributions. If the power is a major concern for research, the Bartlett's test and the Levene's test should be used in general.

1. INTRODUCTION

In many studies conducted in social sciences, various demographic variables are discussed, and inferences are made by comparing group averages according to these variables. To draw these inferences, statistical tests are employed. The choice of test depends on the characteristics of the data and is generally classified as either parametric or nonparametric, based on whether the relevant assumptions are met. Non-parametric tests are easier to calculate than parametric tests. However, they are less powerful, being less likely to reject a false null hypothesis (Woodbury, 2002, p. 591). Consequently, when the assumptions for parametric tests are satisfied, their use is recommended. These assumptions primarily involve normality and homogeneity of variances, which are fundamental when comparing group means in parametric statistics (Tabachnick & Fidell, 2007, p. 201).

*CONTACT: Serpil ÇELİKTEN-DEMİREL ✉ serpil.celikten@dicle.edu.tr Dicle University, Faculty of Education, Department of Educational Sciences, Diyarbakır, Türkiye

e-ISSN: 2148-7456

Because many variables in science and nature follow a normal distribution, it is the most widely applied distribution in statistics (Kirk, 2008, p.230). Most statistical methods, including correlation, regression, and group comparisons, including the *t*-test and ANOVA *F*-test, are grounded in the normal distribution and rely on the assumption of normality. In this case, it is necessary to examine whether the data have a normal distribution before using the relevant tests (Orcan, 2020; Sedgwick, 2015; Woodbury, 2002). At this point, the tests used to test the normality assumption under different simulation conditions are compared, and evaluations are made on the strength of the test. Another key assumption is homogeneity of variance, which means that the variance of a variable remains constant across the levels of another variable (Howell, 2010, p.213). Box (1954) stated that the *F*-test is robust to violations of the homogeneity of variances assumption provided that (1) there are equal numbers of observations at each variable level, (2) the population distributions are normal, and (3) the ratio of the largest variance to the smallest variance does not exceed 3. However, studies have shown that even when sample sizes are equal, the *F*-test is not robust against heterogeneity of variances, which is frequently encountered in social and educational sciences research. In this case, it is important that researchers should not ignore violations of the homogeneity of variances assumption (Kirk, 2008, pp.411-412). In addition, if the variances of the groups are not homogeneous, the inferences to be obtained as a result of the use of parametric tests may not be valid. In addition, although the variances are homogeneous, if this cannot be detected, invalid inferences may be obtained as a result of the use of nonparametric tests due to lower statistical power.

Various tests have been developed to examine the homogeneity of variances. The null hypothesis for all tests considered in this study is that variances are equal between groups and is shown as follows:

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2$$

On the other hand, the case that the variances of at least two populations are not equal to each other is established as the alternative hypothesis and is shown as follows:

$$H_1: \sigma_i^2 = \sigma_j^2 \quad \text{for some } 1 \le i \ne j \le k$$

One of the frequently used tests for testing homogeneity of variance is Levene's Test for Equality of Variance (Field, 2018, p.346; Gamst *et al.*, 2008, p.58). Since Levene's test is not calculated depending on the sample variance, it is not affected by outliers (Zhou *et al.*, 2023). However, it is stated that the power of Levene's test to detect variance differences between levels of a variable depends on the amount of data collected. In large samples, small differences in variances lead to a significant Levene's test, while conversely, in small samples, relatively large differences in variances may not be detected. In other words, in large samples, small differences in variance may increase the significance of Levene's test. However, in small sample conditions, even quite large differences may not be detected (Field, 2018, p.359). In this case, sample size is a point to be considered when using Levene's test.

Another method used to test the homogeneity of variances is Bartlett's test (Bartlett, 1937). Similar to Levene's test, it tests the null hypothesis that the variances in the population are equal. Since Bartlett's test is derived from the likelihood ratio test under normal distribution, it is affected by the normality assumption (Arsham & Lovric, 2011). Therefore, it is stated that this test is powerful when populations are normally distributed (Arsham & Lovric, 2011; Glass, 1966). In cases where populations are not normally distributed, Bartlett's test can be used if the sample is large enough. On the other hand, Levene's test is stated as a powerful test for situations where normal distribution is not provided (Gastwirth *et al.*, 2009).

Another method used to test the homogeneity of variances is the Fligner-Killeen test (Fligner & Killeen, 1976). It is stated that it is a nonparametric test for testing the homogeneity of variance between groups. Also, it can be used in cases where the data do not show a normal distribution. A robust statistical method for determining if variances in groups are equal, the

Brown-Forsythe test was created specifically to overcome Levene's test's shortcomings when dealing with non-normal distributions. It is a modified form of Levene's test where the absolute deviations of observations within each group are calculated using the median (or other trimmed means) in place of the mean (Brown & Forsythe, 1974).

Many studies are comparing the performance of the tests used to test the homogeneity of variances under different conditions (Abdullah & Muda, 2022; Conover *et al.*, 1981; Katsileros *et al.*, 2024; Keskin, 2002; Kim & Cribbie, 2017; Öztürk, 2020; Park, 2018; Wang *et al.*, 2017; Yonar *et al.*, 2024). While some of these studies analyzed only simulation data, others analyzed both simulation and real data. Within the scope of simulation studies, percent of correct rejection (power) and percent of false rejection (type I error rate) values were compared for sample size (Abdullah & Muda, 2022; Gökpınar, 2020; Kim & Cribbee, 2017; Wang *et al.*, 2017; Yonar, 2024), number of groups (Gökpınar, 2020; Kim & Cribbee, 2017; Zhou *et al.*, 2023), variance ratios between groups (Kim & Cribbee, 2017; Wang *et al.*, 2017; Zhou *et al.*, 2023) and different distribution types (such as normal distribution, Laplace distribution, Chi-square distribution; Wang *et al.*, 2017; Yonar *et al.*, 2024; Zhou *et al.*, 2023).

This study aims to examine various tests for homogeneity of variances under different distributions, different variance ratios between groups, and different sample sizes, and sample ratios, and to propose the most appropriate test for various conditions likely to be encountered in practice by crossing possible conditions. Gamma, skewed-normal, and normal distributions were utilized for the distribution condition. Total sample sizes of 60, 120, and 240 were considered, along with group size ratios of 1:1, 1:2, 1:4, and 1:9, and variance ratios of 1:1, 1:2, 1:3, 2:1, and 3:1. The performance of Bartlett's test, Levene's test, Fligner-Killeen test, and the Brown-Forsythe test was evaluated under these conditions.

In this direction, the research questions of the study are as follows:

1. What are the error rates (false rejections) of the tests for homogeneity of variances for varying sample sizes and sample ratios under different distribution conditions when the variance condition is 1:1?
2. What are the power ratios (correct rejections) of the tests for homogeneity of variances under different sample sizes, sample ratios, and variance ratios for normal and skewed-normal distributions?

What are the power ratios (correct rejections) of the tests for homogeneity of variances under different sample sizes, sample ratios, and variance ratios for gamma distributions?

2. METHOD

2.1. Study Design

Within the scope of the study, the power and type I error rates of the tests developed to test the homogeneity of variances under different conditions were evaluated. Basic research includes studies aimed at revealing the processes underlying a theoretical hypothesis (Fraenkel & Wallen, 2009, p.7). In this context, the study is basic research because there are hypotheses about the tests discussed in the study, and the validity of these hypotheses is tested under different conditions.

2.2. Data

Within the scope of the study, data were generated over the beta distribution, skewed-normal, and gamma distribution. The data were generated in a Python environment. Descriptive statistics were analyzed to check the validity of the generated data. The data were generated in such a way that there are two groups within each data set. The variables simulated in the study are as follows:

2.2.1. *Ratio of variance between groups*

In group comparison, the homogeneity of variances is expressed as the variances of the groups being equal to each other (Kirk, 2008, p.326). Since it is not very common for the variance ratios of the groups to be greater than 3 (Kirk, 2008), the variance ratios between the pairs were set at 1:1, 1:2, and 1:3. In addition, 2:1 and 3:1 ratios were included to evaluate conditions of small sample-large variance and large sample-small variance. Thus, five levels were considered for the variance ratio condition.

2.2.2. *Data distribution*

The type of distribution affects the significance of homogeneity tests (Brown & Forsythe, 1974; Gastwirth *et al.*, 2009). Given its frequent use in the social sciences, normal distribution plays a critical role in statistical tests (Shavelson, 1996, p. 115). In addition, skewed-normal distribution, which is close to normal distribution, has also been used in studies on statistical tests (Arnold *et al.*, 2014; Sarısoy *et al.*, 2013; Zhou *et al.*, 2023). In addition to these distributions, a systematic review study conducted by Bono and colleagues (2017) examined the types of distributions used in studies in the fields of health, education, and social sciences. Among the 262 studies reviewed in the Web of Science database between 2010 and 2015, the gamma distribution emerged as the most frequently used non-normal distribution, appearing in 57 articles. As a result, three different distribution types were used in the study: normal, skewed-normal, and gamma distribution. For normal distribution, N ~ (0, 1, 0), for skewed-normal distribution, slightly skewed (shape=2) and highly skewed (shape=10), and for gamma distribution, slightly skewed (shape=5, scale=1) and highly skewed (shape=2, scale=1) (Ahsanullah, 2017, p. 24; Azzalini, 1985, p. 174).

2.2.3. *Sample sizes and ratio of group sizes*

In studies conducted in the field of social sciences, sample size is a point to be considered in order to avoid errors in interpreting the results. In studies with different subgroups and comparisons between groups (e.g., gender, school, marital status, etc.), it is recommended to have at least 30 units from each subgroup; in experimental studies, it is recommended to have a sample size between 10-20 (Roscoe, 1975). Four different sample sizes ($N = (n_1 + n_2)$): 60, 120, and 240, and four levels of ratio of group sizes (n_1/n_2= 1/1, 1/2, 1/4, 1/9) were investigated for each distribution.

As a result, a total of 300 different conditions were obtained, 5 (variance ratios) x 5 (distributions) x 3 (sample sizes) x 4 (ratio of group sizes). For each condition, 10000 replications were generated.

2.3. Data Analysis

Three different variance homogeneity tests, which are Bartlett's test, Levene's test, Fligner-Killeen test, and the Brown-Forsythe test, were considered in the analysis of the data. All the analysis was done in a Python environment. Then, the type I error (false rejection) and power (correct rejection) rates were compared over the test results. When the variances were equal, the performance of the tests was evaluated according to the false rejection (type I error). On the other hand, when the variances were not equal, the performance of the tests was evaluated according to the correct rejection (power).

3. RESULTS

The simulation study was conducted to compare the empirical Type I error rates and the statistical power of the four different homogeneity tests of variances, manipulating the type of distributions, sample sizes, ratio of group sizes, and ratio of group variances. Results were examined based on the crossed conditions mentioned in the methods section.

Three different sample sizes (N=(n_1+n_2)): 60, 120, 240, and four levels of ratio of group sizes (n_1/n_2=1/1; 1/2; 1/5; 1/9) were investigated under the normal, the skewed-normal distributions (slightly-skewed [alpha=2], highly-skewed [alpha=10]), and the gamma distributions (high degree of skewness-GA [2,1]; slight degree of skewness-GA [5,1]), respectively.

3.1. Results of the Type I Error Rates under Each Distribution

In this section, the false rejection rates (Type I error) of the Bartlett, Levene, Brown-Forsythe, and Fligner-Killeen tests were evaluated under normal, skewed-normal, and gamma distributions when group variance ratios were equal. These examinations were conducted under a range of conditions where the ratios of group sizes were 1/1, 1/2, 1/5, 1/9, and the total sample size was 60, 120, and 240 in each condition, respectively. Table 1 summarizes the error rates (false rejections/type I error) for the results of the tests under each distribution.

Table 1. *The empirical type I error rates under normal, skewed-normal, and gamma distributions in terms of different ratios of group sizes and sample sizes.*

			Normal Distribution	Skewed-Normal Distribution		Gamma Distribution	
N	RoGS (n_1,n_2)			Slightly-skewed	Highly-skewed	Slightly-skewed	Highly-skewed
60	1/1 (30, 30)	BRT	0.051	0.070	0.098	0.104	0.180
		LEV	0.052	0.066	0.093	0.071	0.102
		BF	0.041	0.048	0.049	**0.041**	0.049
		FK	**0.040**	0.047	0.055	0.045	0.061
	1/2 (20,40)	BRT	0.049	0.068	0.093	0.103	0.177
		LEV	0.051	0.059	0.091	0.072	0.104
		BF	0.042	0.044	0.047	0.043	0.047
		FK	0.042	0.045	0.057	0.047	0.056
	1/5 (10,50)	BRT	0.048	0.065	0.082	0.094	0.160
		LEV	0.053	0.061	0.080	0.070	0.101
		BF	0.045	0.046	0.044	0.042	0.044
		FK	0.044	0.046	0.055	0.048	0.062
	1/9 (6,54)	BRT	0.049	0.059	0.076	0.081	0.127
		LEV	**0.054**	0.055	0.077	0.068	0.089
		BF	0.047	**0.040**	**0.040**	**0.041**	**0.039**
		FK	0.048	0.046	0.055	0.049	0.062
120	1/1 (60,60)	BRT	0.053	0.065	0.096	0.112	0.195
		LEV	0.052	0.058	0.088	0.072	0.099
		BF	0.047	0.045	0.049	0.045	0.049
		FK	0.045	0.046	0.060	0.050	0.062
	1/2 (40,80)	BRT	0.050	0.070	0.100	0.105	0.189
		LEV	0.052	0.060	0.088	0.071	0.097
		BF	0.046	0.048	0.051	0.047	0.046
		FK	0.044	0.048	0.062	0.049	0.062
120	1/5 (20, 100)	BRT	0.050	0.069	0.093	0.099	0.177
		LEV	0.050	0.060	0.088	0.068	0.096
		BF	0.046	0.048	0.052	0.044	0.046
		FK	0.044	0.049	0.062	0.050	0.062
	1/9 (12, 108)	BRT	0.049	0.061	0.085	0.093	0.163
		LEV	0.051	0.056	0.082	0.063	0.095
		BF	0.045	0.046	0.045	0.042	0.043
		FK	0.046	0.046	0.056	0.050	0.059

Table 1. *Continued.*

			Normal Distribution	Skewed-Normal Distribution		Gamma Distribution	
N	RoGS (n_1,n_2)			Slightly-skewed	Highly-skewed	Slightly-skewed	Highly-skewed
240	1/1 (120, 120)	BRT	0.049	**0.072**	**0.101**	**0.117**	**0.201**
		LEV	0.049	0.063	0.088	0.072	0.096
		BF	0.046	0.054	0.050	0.050	0.044
		FK	0.046	0.054	0.064	0.054	0.063
	1/2 (80, 160)	BRT	0.047	0.067	0.097	0.116	0.195
		LEV	0.048	0.057	0.082	0.071	0.096
		BF	0.045	0.049	0.047	0.050	0.050
		FK	0.045	0.050	0.060	0.054	0.069
	1/5 (40, 200)	BRT	0.052	0.067	0.095	0.110	0.191
		LEV	0.051	0.058	0.086	0.073	0.099
		BF	0.049	0.048	0.051	0.048	0.047
		FK	0.048	0.049	0.063	0.052	0.066
	1/9 (24, 216)	BRT	0.050	0.067	0.094	0.105	0.184
		LEV	0.050	0.057	0.091	0.069	0.097
		BF	0.048	0.046	0.054	0.044	0.045
		FK	0.048	0.048	0.068	0.050	0.066

N: Total sample size; RoGS: Ratio of group size; BRT: Bartlett's test; LEV: Levene's test; BF: Brown-Forsythe test; FK: Fligner-Killeen test

Error ratios in Table 1 show that the Fligner-Killeen test and the Brown-Forsythe test produced the lowest type I error rates with values close to each other for each crossed condition in terms of distribution, sample size, and group size (e.g., under the normal distribution with *N*=60 and RoGS: 1/1); however, the results varied, at times favoring the Fligner-Killeen and, alternatively, the Brown-Forsythe test. Detailed examinations of the lowest and highest error rates for tests of homogeneity of variances are presented for each of the distributions, considering the group ratio sizes.

For normal distribution, the Fligner-Killeen test yielded the best result with the lowest type I error rate for the balanced group sizes of 1/1 (30, 30). The highest type I error rate was observed in Levene's test with the group ratio of 1/9 (6, 54). For the slightly-skewed normal distribution, the Brown-Forsythe test showed the lowest type I error rate with the group ratio of 1/9 (6, 54). The highest error rates were produced by the BRT test for the balanced group sizes of 1/1 (120, 120). For the highly-skewed normal distribution, the Brown-Forsythe test again yielded the lowest type I error rate with the group ratio of 1/9 (6, 54), and the highest error rates came from Bartlett's test for the group sizes of 1/1 (120, 120), as in the slightly-skewed normal distribution.

For the slightly-skewed gamma distribution, the Brown-Forsythe test yielded the lowest error rate for the group ratio sizes of 1/1 (30, 30) and 1/9 (6, 54), respectively. The highest type I error rate was observed in the Bartlett's test with the group ratio sizes of 1/1 (120, 120). For the highly-skewed gamma distribution, the Brown-Forsythe test again produced the lowest error rate with the group ratio sizes of 1/9 (6, 54), while Bartlett's test produced the highest error rate for the group size ratio of 1/1 (120, 120).

When considering the varying degrees of skewness in both the skewed-normal and gamma distributions simultaneously, it was observed that methods producing the highest and lowest error rates were consistent across distributions. The Brown-Forsythe test resulted in the lowest error rates with the ratio of group sizes of 1/9 (6, 54), while Bartlett's test showed the highest error rates for the balanced group sizes of 1/1 (120, 120). An exception occurred in the slightly skewed gamma distribution, where the Brown-Forsythe test also produced the lowest error rate for the group ratio of 1/1 (30, 30), equal to the lowest error observed for the group ratio of 1/9 (6, 54). However, the cases yielding the highest and lowest error rates under the normal distribution occurred under different conditions than those observed in the skewed-normal and gamma distributions.

In order to achieve a more profound comprehension of the outcomes, it is crucial to undertake a holistic analysis of the conditions under the normal, skewed-normal, and gamma distributions. Analyses revealed that the error values observed under the gamma distribution generally tended to be higher than those observed under the normal and skewed-normal distributions.

When analyzing each distribution separately, it was found that under conditions where all sample sizes and group proportions were crossed, the highest type I error rates occurred in Bartlett's test when both distributions were highly skewed, except for one case arising from Levene's test. In some instances, Levene's test resulted in higher error rates than Bartlett's test, while in others, Bartlett's test produced higher errors than Levene's. A closer examination of these discrepancies showed that the results of the two tests were highly similar.

A comparison of the less skewed and highly skewed distributions, under both the skewed-normal and gamma conditions, showed that the highly skewed distribution produced higher error rates. Methodologically, the lowest error values were obtained with the Brown-Forsythe and Fligner-Killeen tests, though the difference between them was relatively small.

An evaluation of the results in terms of group size ratios, with each sample size held constant, revealed that differences in group sizes generally led to a decline in type I error, though a few cases contradicted this pattern. Lastly, when examining the impact of sample size on each method while maintaining constant group sizes across distributions, it was found that the observed fluctuations were not systematic.

3.2. Results of the Power for the Normal and Skewed-Normal Distributions

The powers of tests for homogeneity of variance were computed for skewed-normal distributions, including the normal, slightly skewed, and highly skewed normal distributions, separately. These computations were conducted under four different group variance ratio conditions (1:2, 1:3, 2:1, 3:1) and across different sample sizes with varying group size ratios (1/1, 1/2, 1/5, 1/9). Results are presented in Table 2, with the highest and lowest values highlighted in bold for each distribution crossed by the variance ratio conditions.

Table 1. *Correct rejection rate under normal and skewed-normal distribution with varying conditions in terms of sample size, ratio of group size, and ratio of group variance.*

			Normal distribution				Slightly-skewed-normal distribution				Highly-skewed-normal distribution			
N	RoGS (n_1,n_2)	VR / HT	1:2	1:3	2:1	3:1	1:2	1:3	2:1	3:1	1:2	1:3	2:1	3:1
60	1/1 (30, 30)	BRT	0.441	0.826	0.459	0.831	0.450	0.812	0.453	0.811	0.460	0.792	0.457	0.787
		LEV	0.390	0.758	0.398	0.763	0.387	0.737	0.396	0.741	0.412	0.738	0.416	0.743
		BF	0.352	0.722	0.363	0.728	0.340	0.688	0.346	0.699	0.308	0.644	0.314	0.649
		FK	0.327	0.684	0.340	0.690	0.319	0.654	0.331	0.665	0.314	0.632	0.313	0.632
	1/2 (20,40)	BRT	0.379	0.770	0.422	0.788	0.399	0.753	0.419	0.764	0.409	0.737	0.421	0.743
		LEV	0.330	0.678	0.383	0.731	0.338	0.657	0.376	0.706	0.357	0.673	0.395	0.708
		BF	0.302	0.653	0.334	0.692	0.300	0.617	0.320	0.653	0.262	0.565	0.295	0.615
		FK	0.295	0.633	0.298	0.629	0.300	0.612	0.286	0.588	0.290	0.593	0.272	0.561
	1/5 (10,50)	BRT	0.221	0.486	0.283	0.584	0.240	0.500	0.282	0.569	0.266	0.518	0.291	0.547
		LEV	0.188	0.407	0.278	0.554	0.185	0.394	0.274	0.540	0.230	0.437	0.294	0.540
		BF	0.178	0.393	0.227	0.489	0.163	0.360	0.221	0.472	0.147	0.320	0.205	0.437
		FK	0.188	0.407	0.188	0.401	0.181	0.393	0.173	0.384	0.194	0.407	0.168	0.354
	1/9 (6,54)	BRT	0.139	0.273	0.195	0.408	0.155	0.293	0.200	0.393	0.174	0.317	0.203	0.389
		LEV	0.114	0.219	0.210	0.419	0.111	0.212	0.211	0.401	0.151	0.261	0.226	0.407
		BF	**0.110**	**0.212**	0.167	0.353	**0.098**	**0.187**	0.161	0.332	**0.086**	**0.166**	0.149	0.308
		FK	0.120	0.232	**0.134**	**0.266**	0.122	0.228	**0.120**	**0.247**	0.140	0.247	**0.118**	**0.229**
120	1/1 (60,60)	BRT	0.749	0.987	0.752	0.985	0.734	0.980	0.741	0.982	0.720	0.969	0.725	0.969
		LEV	0.690	0.973	0.686	0.970	0.664	0.964	0.674	0.964	0.669	0.957	0.684	0.956
		BF	0.672	0.969	0.669	0.967	0.634	0.957	0.646	0.959	0.587	0.934	0.600	0.934
		FK	0.647	0.960	0.644	0.955	0.612	0.944	0.618	0.944	0.582	0.924	0.598	0.927
	1/2 (40,80)	BRT	0.683	0.978	0.705	0.973	0.677	0.967	0.698	0.964	0.668	0.952	0.670	0.949
		LEV	0.619	0.949	0.649	0.952	0.595	0.937	0.637	0.943	0.611	0.931	0.637	0.929
		BF	0.606	0.946	0.624	0.943	0.573	0.930	0.598	0.932	0.519	0.903	0.555	0.905
		FK	0.592	0.936	0.590	0.922	0.561	0.921	0.561	0.907	0.540	0.903	0.534	0.878
	1/5 (20, 100)	BRT	0.458	0.856	0.507	0.866	0.465	0.842	0.500	0.845	0.475	0.829	0.498	0.820
		LEV	0.387	0.781	0.472	0.835	0.389	0.753	0.468	0.806	0.414	0.763	0.481	0.800
		BF	0.388	0.781	0.430	0.809	0.370	0.738	0.419	0.774	0.325	0.682	0.393	0.741
		FK	0.388	0.778	0.389	0.747	0.378	0.745	0.369	0.709	0.374	0.727	0.349	0.675
	1/9 (12, 108)	BRT	0.288	0.623	0.356	0.693	0.301	0.621	0.361	0.681	0.325	0.627	0.357	0.658
		LEV	0.241	0.532	0.342	0.668	0.236	0.511	0.350	0.653	0.272	0.551	0.355	0.651
		BF	0.248	0.542	0.297	0.626	0.224	0.497	0.297	0.604	0.196	0.439	0.271	0.565
		FK	0.257	0.548	0.254	0.547	0.246	0.526	0.249	0.517	0.257	0.533	0.228	0.471
240	1/1 (120, 120)	BRT	**0.963**	**1.000**	**0.967**	**1.000**	**0.952**	**1.000**	**0.950**	**1.000**	**0.937**	**1.000**	**0.933**	**1.000**
		LEV	0.934	**1.000**	0.939	**1.000**	0.927	0.999	0.920	**1.000**	0.918	0.999	0.915	0.999
		BF	0.932	**1.000**	0.936	**1.000**	0.919	**1.000**	0.912	**1.000**	0.887	0.999	0.885	0.999
		FK	0.921	**1.000**	0.927	**1.000**	0.905	0.999	0.898	**1.000**	0.883	0.998	0.881	0.998
	1/2 (80, 160)	BRT	0.944	**1.000**	0.946	**1.000**	0.934	**1.000**	0.927	0.999	0.914	**1.000**	0.905	0.999
		LEV	0.907	**1.000**	0.915	0.999	0.891	0.999	0.897	0.999	0.886	0.999	0.883	0.998
		BF	0.905	**1.000**	0.907	0.999	0.882	0.999	0.886	0.998	0.847	0.998	0.845	0.997
		FK	0.898	0.999	0.894	0.998	0.872	0.998	0.863	0.997	0.853	0.998	0.832	0.995

N: Total sample size; RoGS: Ratio of group size; HT: Homogeneity tests; VR: Variance ratio; BRT: Bartlett's test; LEV: Levene's test; BF: Brown-Forsythe test; FK: Fligner-Killeen test

Table 2. *Continued.*

			Normal distribution				Slightly-skewed-normal distribution				Highly-skewed-normal distribution			
N	RoGS (n_1,n_2)	VR / HT	1:2	1:3	2:1	3:1	1:2	1:3	2:1	3:1	1:2	1:3	2:1	3:1
240	1/5 (40, 200)	BRT	0.776	0.996	0.797	0.990	0.774	0.990	0.782	0.988	0.758	0.985	0.748	0.976
		LEV	0.713	0.986	0.760	0.984	0.692	0.979	0.739	0.976	0.698	0.973	0.722	0.968
		BF	0.716	0.986	0.740	0.981	0.682	0.979	0.713	0.972	0.625	0.959	0.661	0.957
		FK	0.711	0.984	0.705	0.970	0.682	0.978	0.670	0.959	0.660	0.965	0.628	0.936
	1/9 (24, 216)	BRT	0.564	0.940	0.624	0.935	0.572	0.926	0.604	0.919	0.568	0.909	0.587	0.895
		LEV	0.492	0.894	0.593	0.913	0.476	0.874	0.569	0.898	0.503	0.869	0.575	0.879
		BF	0.504	0.900	0.560	0.900	0.471	0.871	0.528	0.878	0.419	0.817	0.498	0.843
		FK	0.505	0.898	0.519	0.871	0.481	0.875	0.480	0.840	0.479	0.852	0.453	0.795

N: Total sample size; RoGS: Ratio of group size; HT: Homogeneity tests; VR: Variance ratio; BRT: Bartlett's test; LEV: Levene's test; BF: Brown-Forsythe test; FK: Fligner-Killeen test

Results in Table 2 indicated that Bartlett's test tended to provide the best power among the homogeneity of variances tests considered in this study, apart from a few cases where Levene's test showed slightly higher power, though with only a small increase.

Under the normal distribution, all homogeneity of variance tests achieved perfect power with variance ratios of 1:3 and 3:1 for the balanced group sizes of 1/1 (120, 120). For the group size ratio of 1/2 (80, 160), Bartlett's, Levene's, and Brown-Forsythe tests yielded perfect power with a variance ratio of 1:3, while for the variance ratio of 3:1, only Bartlett's test achieved perfect power. For the variance ratios of 1:2 and 2:1, again with the balanced group sizes of 1/1 (120, 120), Bartlett's test provided the highest percentage of correct rejections. Results under the slightly skewed normal distribution indicated that, for the balanced group size of 1/1 (120, 120) with a variance ratio of 3:1, all methods produced perfect power. For the group size ratio of 1/2 (80, 160), Bartlett's and the Brown-Forsythe test achieved perfect power under the variance ratio of 1:3. Moreover, for the 1:2 and 2:1 variance ratios with the balanced group sizes of 1/1 (120, 120), Bartlett's test again provided the highest power, consistent with the normal distribution results. Under the highly skewed normal distribution, results showed that for the balanced group size of 1/1 (120, 120) with a variance ratio of 3:1, and for the group size ratio of 1/2 (80, 160), Bartlett's test achieved perfect power in each case. For the variance ratios of 1:2 and 2:1, Bartlett's test again showed the highest power for the balanced group size of 1/1 (120, 120). Examining the lowest true rejection rates, the Brown-Forsythe test produced the lowest power for variance ratios of 1:2 and 1:3 with the group size ratio of 1/9 (6, 54) under the normal, slightly skewed, and highly skewed normal distributions, respectively. For the variance ratios of 2:1 and 3:1 with the group size ratio of 1/9 (6, 54), the Fligner-Killeen test resulted in the lowest power under the normal, slightly skewed, and highly skewed normal distributions in the same manner.

When group sizes remained constant, it was observed that an increase in the overall sample size systematically led to higher power of the tests across all distribution types. Moreover, when the total sample size remained constant, increasing the disparity between group sizes reduced statistical power. In summary, larger sample sizes consistently increased the power of the tests, whereas greater differences in group size ratios decreased it.

In general, Bartlett's test tended to provide the highest power rates compared to the other methods under all distributions crossed by variance ratio conditions. Only under the group size ratio of 1/9 (6, 54) did Levene's test produce higher power values than Bartlett's test in some cases (e.g., under the normal distribution with group variance ratios of 2:1 and 3:1, and under the slightly skewed distribution with group variance ratios of 2:1 and 3:1). These differences were small, with a maximum of 0.018, suggesting that they may have been random rather than systematic. When the lowest correct rejection rates were examined, the Brown-Forsythe and Fligner-Killeen tests tended to produce the lowest values under all conditions, although in some cases their values were the same or very similar.

Increasing the group variance ratios led to a higher correct rejection rate of the homogeneity of variance tests (the expected result). However, changing the order of the group variances (e.g., 2:1 instead of 1:2) did not significantly impact the correct rejection rates. The small changes observed between the group variance conditions of 1:2 and 2:1, and between 1:3 and 3:1 under each distribution, were not systematic. When the differences between the 1:2–2:1 and 1:3–3:1 under each distribution condition (for example, the difference of 1:2 and 2:1 under normal distribution) were analyzed, it was observed that these difference values were generally low. When these differences were analyzed quantitatively, it was found that only 6% of the values were above 0.1, with the remaining values being 0.1 or below. However, this variation was not systematic.

In circumstances where the total sample size is minimal and the rate of differentiation between group sizes is substantial, it is imperative to meticulously select the most suitable method. In such cases, although Bartlett's test generally stood out, both Bartlett's and Levene's tests gained particular importance based on the quantitative comparison of correct rejection rates. It was established that the efficacy of the methods is directly proportional to the total sample size, with power values approaching 1. However, it was also demonstrated that an increase in the discrepancy between group sizes adversely affected the observed power values.

3.3. Results of the Power for the Gamma Distributions

Table 3 summarizes the correct rejection rates under the gamma distributions across varying conditions of sample size, ratio of group size, and ratio of group variance. The highest and lowest results are highlighted in bold for each distribution crossed by variance ratio conditions.

Table 2. *Correct rejection rate under gamma distribution across varying conditions in terms of sample size, ratio of group size, and ratio of group variance.*

			Slightly-skewed- gamma distribution				Highly-skewed-gamma distribution			
N	RoGS (n_1,n_2)	VR / HT	1:2	1:3	2:1	3:1	1:2	1:3	2:1	3:1
60	1/1 (30, 30)	BRT	0.463	0.792	0.456	0.786	0.468	0.740	0.468	0.744
		LEV	0.382	0.720	0.383	0.724	0.366	0.654	0.374	0.666
		BF	0.307	0.651	0.308	0.651	0.248	0.538	0.254	0.542
		FK	0.305	0.629	0.306	0.638	0.279	0.562	0.285	0.570
	1/2 (20, 40)	BRT	0.418	0.741	0.421	0.734	0.440	0.705	0.436	0.689
		LEV	0.326	0.638	0.372	0.692	0.316	0.583	0.364	0.640
		BF	0.256	0.562	0.295	0.620	0.204	0.443	0.256	0.530
		FK	0.272	0.584	0.269	0.568	0.262	0.525	0.250	0.505
60	1/5 (10,50)	BRT	0.280	0.523	0.287	0.541	0.341	0.550	0.311	0.509
		LEV	0.198	0.390	0.274	0.520	0.202	0.361	0.278	0.488
		BF	0.144	0.310	0.209	0.436	0.105	0.227	0.188	0.380
		FK	0.186	0.385	0.169	0.355	0.187	0.362	0.161	0.313
	1/9 (6,54)	BRT	0.179	0.330	0.204	0.380	0.239	0.387	0.225	0.371
		LEV	0.115	0.212	0.213	0.389	0.120	0.206	0.220	0.374
		BF	**0.079**	**0.152**	0.153	0.313	**0.051**	**0.096**	0.142	0.274
		FK	0.124	0.234	**0.116**	**0.230**	0.133	0.235	**0.118**	**0.213**
120	1/1 (60,60)	BRT	0.709	0.963	0.711	0.965	0.676	0.931	0.667	0.926
		LEV	0.643	0.948	0.647	0.948	0.594	0.908	0.584	0.903

Table 3. *Continued.*

N	RoGS (n_1,n_2)	HT \ VR	Slightly-skewed-gamma distribution 1:2	1:3	2:1	3:1	Highly-skewed-gamma distribution 1:2	1:3	2:1	3:1
		BF	0.586	0.935	0.595	0.936	0.492	0.866	0.481	0.861
		FK	0.580	0.923	0.588	0.927	0.531	0.878	0.519	0.876
	1/2 (40,80)	BRT	0.667	0.948	0.664	0.947	0.654	0.915	0.624	0.898
		LEV	0.579	0.917	0.613	0.927	0.534	0.873	0.556	0.880
		BF	0.521	0.897	0.555	0.910	0.420	0.816	0.458	0.834
		FK	0.534	0.900	0.524	0.886	0.488	0.858	0.471	0.825
	1/5 (20, 100)	BRT	0.482	0.819	0.489	0.818	0.508	0.785	0.477	0.753
		LEV	0.368	0.726	0.458	0.793	0.350	0.657	0.427	0.731
		BF	0.316	0.681	0.390	0.750	0.233	0.534	0.327	0.656
		FK	0.358	0.726	0.340	0.685	0.332	0.657	0.305	0.607
	1/9 (12, 108)	BRT	0.339	0.636	0.348	0.645	0.402	0.639	0.357	0.589
		LEV	0.239	0.494	0.338	0.626	0.234	0.451	0.322	0.574
		BF	0.192	0.425	0.274	0.565	0.137	0.303	0.235	0.488
		FK	0.246	0.507	0.228	0.479	0.235	0.470	0.198	0.416
240	1/1 (120, 120)	BRT	**0.929**	**0.999**	**0.923**	**0.999**	**0.882**	**0.996**	**0.881**	**0.995**
		LEV	0.906	0.998	0.899	**0.999**	0.852	**0.996**	0.851	**0.995**
		BF	0.889	0.998	0.881	**0.999**	0.802	0.994	0.798	0.993
		FK	0.878	0.997	0.875	0.998	0.824	**0.996**	0.823	0.994
	1/2 (80, 160)	BRT	0.901	0.999	0.892	**0.999**	0.856	0.993	0.845	0.991
		LEV	0.865	0.997	0.870	0.997	0.804	0.991	0.814	0.989
		BF	0.843	0.997	0.849	0.996	0.736	0.985	0.757	0.986
		FK	0.846	0.997	0.834	0.995	0.784	0.990	0.767	0.985
	1/5 (40, 200)	BRT	0.752	0.978	0.747	0.974	0.729	0.956	0.687	0.941
		LEV	0.660	0.961	0.716	0.966	0.604	0.927	0.650	0.932
		BF	0.617	0.956	0.667	0.959	0.495	0.887	0.570	0.912
		FK	0.645	0.961	0.627	0.943	0.601	0.931	0.561	0.899
	1/9 (24, 216)	BRT	0.574	0.902	0.583	0.890	0.591	0.866	0.538	0.828
		LEV	0.460	0.840	0.552	0.875	0.428	0.772	0.505	0.816
		BF	0.410	0.813	0.494	0.851	0.310	0.673	0.415	0.769
		FK	0.459	0.848	0.448	0.807	0.430	0.789	0.387	0.732

N: Total sample size; RoGS: Ratio of group size; HT: Homogeneity tests; VR: Variance ratio; BRT: Bartlett's test; LEV: Levene's test; BF: Brown-Forsythe test; FK: Fligner-Killeen test

Results in Table 3 indicated that under the gamma distribution, Bartlett's test tended to provide the best power among the homogeneity tests under all crossed conditions, followed by Levene's test. However, the Brown-Forsythe and Fligner-Killeen tests tended to yield the lowest power compared to the other homogeneity of variance tests considered in this study.

For the gamma distribution characterized by a slightly skewed shape, the maximum power values were generally attained under the condition of balanced group sizes 1/1 (120, 120), and in one case under the group size ratio of 1/2 (80, 160). Bartlett's test provided the highest power for the variance conditions of 1:2, 1:3, and 2:1 with the group size ratio of 1/1 (120, 120). For the variance condition of 3:1, Bartlett's, Levene's, and Brown-Forsythe tests all produced the highest power, with a value of 0.999, again for the balanced group size of 1/1 (120, 120). For the group size ratio of 1/2 (80, 160), Bartlett's test also yielded the highest power, with a value of 0.999 as in the previous cases.

For the gamma distribution characterized by a highly skewed shape, the maximum power values were attained under the balanced group sizes of 1/1 (120, 120), as in the slightly-skewed gamma distribution. Specifically, Bartlett's test yielded the highest power for the variance conditions of 1:2 and 2:1. For the 1:3 variance condition, Bartlett's, Levene's, and Fligner-Killeen tests yielded the same value of 0.996, while the Brown-Forsythe test produced a value of 0.994. For the variance condition of 3:1, Bartlett's and Levene's tests both yielded the same value of 0.995.

In both the slightly-skewed and highly-skewed distributions, the lowest power was observed under the group size ratio of 1/9 (6, 54). An analysis of the group variance ratios was conducted to ascertain the impact of the different methods under this condition. It was observed that the Brown-Forsythe test produced the lowest power in the 1:2 and 1:3 variance ratio conditions. In contrast, for the 3:1 and 2:1 variance ratio conditions, the Fligner-Killeen test yielded the lowest power values. Overall, the Brown-Forsythe and Fligner-Killeen tests provided the lowest power values when the sample size was minimal and the group size disparity was maximal (1/9, 6 vs. 54). These outcomes were consistent under both the slightly-skewed and highly-skewed gamma distributions. Furthermore, the power values under each variance condition demonstrated that the slightly-skewed gamma distribution yielded higher power than the highly-skewed distribution.

A comparative analysis of the methods revealed that Bartlett's test produced optimal results under all crossed conditions when the sample sizes were 60 and 120. Levene's test aligned closely with Bartlett's test. For the sample size of 240, Bartlett's test generally produced the highest values; however, the differences between methods were small, and there was a substantial increase in true rejection rates overall. Notably, in the highly-skewed distribution with group sizes 1/1 (120, 120) under the 1:3 variance condition, Bartlett's and Levene's tests provided the same values. In the slightly-skewed distribution under the 3:1 variance condition, Bartlett's, Levene's, and Brown-Forsythe tests all produced identical results, while the Fligner-Killeen test produced an almost identical value of 0.998.

As the difference between group sizes widened, the differences among methods also became more pronounced. When analyzed in terms of the lowest power under each crossed condition, the Brown-Forsythe and Fligner-Killeen tests generally produced the lowest values. Finally, as the total sample size increased, correct rejection rates also increased for all methods, resulting in higher power values. Altering the order of the variance ratios (1:2 versus 2:1; 1:3 versus 3:1) had only a minor impact on the results.

A general evaluation of the results regarding power values indicates that the correct rejection rates increased as the sample size increased when the group size remained constant. However, an increase in group size ratios led to a decrease in power values when the total sample size remained constant. It was established that an increase in the discrepancy between group variance ratios corresponded to an increase in the correct rejection rate. By contrast, no systematic change was observed when the order of group variance ratios was altered (1:2 versus 2:1; 1:3 versus 3:1).

The analysis revealed that Bartlett's test tended to yield the highest power values, while the Brown-Forsythe and Fligner-Killeen tests generally produced the lowest values. When the sample size was large (240) and the group size ratio was balanced (1/1), all tests provided close to, or even near-perfect, values. In contrast, when the sample size was modest (60) and the group size ratio was unbalanced (e.g., 1/9), a decline in power values was observed across all methods, with the disparities between methods becoming more pronounced. In such cases, Bartlett's test stood out for its consistently high power.

4. DISCUSSION and CONCLUSION

This study compared the performance of Bartlett, Levene, Brown-Forsythe, and Fligner-Killeen tests in terms of error rates and power for two groups when the variances were equal (1:1) and unequal (1:2, 1:3, 2:1, 3:1), across different distribution types (normal, skewed-normal, gamma), and varying sample sizes (60, 120, and 240) with different sample ratios (1/1, 1/2, 1/4, 1/9). In the first stage, the false rejection rates of the homogeneity tests were examined under normal, skewed-normal, and gamma distributions. When the variances for the two groups were equal under the normal distribution, the Fligner-Killeen test had the lowest false rejection rate when group sizes were balanced. In unbalanced cases, the Brown-Forsythe and Fligner-Killeen tests alternately produced the lowest false rejection rates. In other words, the type I error rates for Brown-Forsythe and Fligner-Killeen were lower than those of the other tests under the normal distribution. Similarly, Yi et al. (2020) stated that the Brown-Forsythe test was adequate for most population distribution shapes.

It should be noted that under the normal distribution, type I error rates across all simulation conditions were approximately 0.05 for all tests. However, under the skewed-normal and gamma distributions, the type I error rates for Bartlett's test and Levene's test increased more than those for the Brown-Forsythe and Fligner-Killeen tests. In particular, Bartlett's test showed the largest type I error rates under non-normal distributions. In other words, Bartlett's test was extremely sensitive to non-normal distributions and only performed well under the normal distribution. In parallel with this result, Chang et al. (2017) suggested using Bartlett's test only when the normality assumption is nearly certain. As in Yonar's (2024) research, Levene's test showed lower performance and larger type I error rates for normal distributions. In general, in some cases Levene's produced higher error rates than Bartlett's, while in other cases Bartlett's produced higher error rates than Levene's. However, a closer examination revealed that the results of the two tests were highly similar. A comparison of the slightly-skewed and highly-skewed distributions under both the skewed-normal and gamma distributions revealed that the highly-skewed distributions resulted in higher error rates. In general, greater skewness led to higher error levels as the sample size increased for all homogeneity of variance tests.

In the next stage, the results of the correct rejection rates of the homogeneity tests under normal, skewed-normal, and gamma distributions were examined. When the variances for the two groups were different under the normal distribution, Bartlett's test was generally the most powerful, followed by Levene's test. As Gastwirth et al. (2009) pointed out, Levene's test is a powerful method for situations where the normal distribution is not satisfied. The power of the Brown-Forsythe and Fligner-Killeen tests generally tended to decrease as the total sample size became smaller. Across all sample size conditions, the correct rejection rates of all tests were closer to each other when the group sizes were balanced. As the total sample size increased, the correct rejection rates increased for all methods. In parallel with the study of Wang et al. (2017), as the sample size increased, the correct rejection rates of all tests increased overall. Although the progression of the correct rejection rates of the methods varied under different sample conditions, it was evident that increasing the sample size had a positive effect on the performance of all methods.

For the normal, skewed-normal, and gamma distributions, the highest power values were observed when the sample size was large and the group sizes were balanced. Conversely, the lowest power values were observed when the sample size was smallest and the disparity between group sizes was greatest.

Depending on the conclusions of the current study, it is recommended that researchers select the most appropriate homogeneity of variance test based on the distribution, total sample size, group sample sizes, and variance ratios of the data. For example, if the sample size is 120, the data are normally distributed with a group variance ratio of 1:3, and the group size ratio is 1/5 (20, 100), Bartlett's test is a good choice when the focus is on power. The following recommendations are made:

- In general, the homogeneity of variance tests achieve perfect or near-perfect power as the total sample size increases and the disparity between group sizes decreases. Therefore, it is recommended that researchers use balanced and large samples whenever feasible.
- If type I error is a major concern in research, the Brown-Forsythe test should be preferred, particularly for non-normal distributions (skewed-normal and gamma). If the Brown-Forsythe test is not applicable, the Fligner-Killeen test may be considered a viable alternative.
- The Fligner-Killeen test is recommended for detecting small differences between group variances under a skewed-normal distribution.
- If there is an assumption or suspicion of heterogeneity of variance between groups, Bartlett's test may be preferred. If Bartlett's test is not applicable, Levene's test generally yields comparable results and may be considered a viable alternative without substantially reducing statistical power.
- If the group sizes are unbalanced and the sample size is small, Bartlett's test is recommended, as it tends to provide higher power values than the other methods.

As with all simulation studies, a limitation is that the results are only applicable to the conditions investigated in this research. To broaden the applicability of the current study, future researchers should conduct comparative analyses using real data.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Contribution of Authors

Serpil Çelikten Demirel: Investigation, Resources, Visualization, and Writing-original draft. **Ayşenur Erdemir:** Methodology, Software, Formal Analysis, and Writing-original draft. **Esra Oyar:** Methodology, Formal Analysis, Validation, and Writing-original draft. **Tuba Gündüz:** Validation, and Writing-original draft.

Orcid

Serpil Çelikten Demirel https://orcid.org/0000-0003-3868-3807
Ayşenur Erdemir https://orcid.org/0000-0001-9656-0878
Esra Oyar https://orcid.org/0000-0002-4337-7815
Tuba Gündüz https://orcid.org/0000-0002-0921-9290

REFERENCES

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

Ahsanullah, M. (2017). *Characterizations of univariate continuous distributions* (Vol. 1). Amsterdam: Atlantis Press.

Arsham, H., & Lovric, M. (2011). Bartlett's Test. In M. Lovric (Ed.), *International encyclopedia of statistical science (pp.* 87–88). Springer. https://doi.org/10.1007/978-3-642-04898-2_132

Arnold, B.C., Gómez, H.W., & Salinas, H.S. (2014). A doubly skewed normal distribution. *Statistics*, *49*(4), 842-858. https://doi.org/10.1080/02331888.2014.918618

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*(2), 171-178. http://www.jstor.org/stable/4615982

Bartlett, M.S. (1939, April). A note on tests of significance in multivariate analysis. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 35, No. 2, pp. 180-185). Cambridge University Press.

Bono, R., Blanca, M.J., Arnau, J., & Gómez-Benito, J. (2017). Non-normal distributions commonly used in health, education, and social sciences: A systematic review. *Frontiers in Psychology*, *8*, 1602. https://doi.org/10.3389/fpsyg.2017.01602

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/10.1080/01621459.1974.104829557

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance. I: Effect of inequality of variance in one-way classification. *Annals of Mathematical Statistics, 25*, 290–302. http://www.jstor.org/stable/2236731

Chang, C.H., Pal, N., & Lin, J.J. (2017). A revisit to test the equality of variances of several populations. *Communications in Statistics-Simulation and Computation, 46*(8), 6360-6384. https://doi.org/10.1080/03610918.2016.1202277

Conover, W.J., Johnson, M.E., & Johnson, M.M. (1981). A comparative study of test for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, *23*, 351–361. https://doi.org/10.2307/1268225

Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. Sage publications limited.

Fligner, M.A., & Killeen, T.J. (1976). Distribution-free two-sample tests for scale. *Journal of the American Statistical Association 71*(353), 210-213. https://doi.org/10.1080/01621459.1976.10481517

Fraenkel, J., & Wallen, N. (2009). *How to Design and Evaluate Research in Education* (7th ed.) McGraw-Hill Education.

Gamst, G., Meyers, L.S., & Guarino, A.J. (2008). *Analysis of variance designs: A conceptual and computational approach with SPSS and SAS*. Cambridge University. https://doi.org/10.1017/CBO9780511801648

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. http://dx.doi.org/10.1214/09-STS301

Glass, G.V. (1966). Testing homogeneity of variances. *American Educational Research Journal*, *3*(3), 187-190. https://doi.org/10.3102/00028312003003187

Gökpınar, E. (2022). Standardized likelihood ratio test for homogeneity of variance of several normal populations. *Communications in Statistics-Simulation and Computation, 51*(11), 6309-6319. https://doi.org/10.1080/03610918.2020.1800037

Howell, D.C. (2010). *Statistical methods for psychology*. PWS-Kent Publishing Co.

Katsileros, A., Antonetsis, N., Mouzaidis, P., Tani, E., Bebeli, P.J., & Karagrigoriou, A. (2024). A comparison of tests for homoscedasticity using simulation and empirical data. *Communications for Statistical Applications and Methods*, *31*(1), 1-35. https://doi.org/10.29220/CSAM.2024.31.1.001

Keppel, G., & Wickens, T.D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, Pearson Prentice Hall.

Keskin, S. (2002). *Varyansların homojenliğini test etmede kullanılan bazı yöntemlerin I. tip hata ve testin gücü bakımından irdelenmesi* [An examination of some methods used in testing the homogeneity of variances in terms of type I error and test power] [Unpublished doctoral dissertation]. Ankara University.

Kim, Y.J., & Cribbie, R.A. (2018). ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. *British Journal of Mathematical and Statistical Psychology*, *71*(1), 1-12. https://doi.org/10.1111/bmsp.12103

Kirk, R.E. (2008). *Statistics an Introduction* (5th ed.). Thomson Wadsworth

Moitra, S.D. (1990). Skewness and the beta distribution. *Journal of the Operational Research Society*, *41*(10), 953-961. https://doi.org/10.1057/jors.1990.147

Orcan, F. (2020). Parametric or non-parametric: skewness to test normality for mean comparison. *International Journal of Assessment Tools in Education, 7*(2), 255-265. https://doi.org/10.21449/ijate.656077

Öztürk, N.N. (2020). *Varyansların homojenliği için bazı testler ve karşılaştırmaları* [Some tests for the homogeneity of variances and comparisons] [Unpublished master dissertation]. Gazi University.

Park, H.I. (2018). Tests of equality of several variances with the likelihood ratio principle. *Communications for Statistical Applications & Methods*, *25*(4), 329-339. https://doi.org/10.29220/CSAM.2018.25.4.329

Roscoe, J.T. (1975). *Fundamental Research Statistics for the Behavioural Sciences* (2nd ed.). Holt Rinehart & Winston.

Sedgwick, P. (2015). A comparison of parametric and non-parametric statistical tests. *BMJ*, *350*. https://doi.org/10.1136/bmj.h2053

Shavelson, R.J. (1996). *Statistical reasoning for the behavioral sciences* (3rd ed.). Pearson Education.

Sarısoy, E.E., Potas, N., & Kara, M. (2013). A simulation study goodness-of-fit tests for the skewed normal distribution. In *Chaos, Complexity and Leadership 2012* (pp. 277-283). Springer Netherlands.

Sarısoy, E.E., Potas, N., Kara, M. (2014). A simulation study goodness-of-fit tests for the skewed normal distribution. In S. Banerjee and Ş. Erçetin (eds). *Chaos, Complexity and Leadership 2012* (pp. 277-283). Springer Proceedings in Complexity. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-7362-2_36

Tabachnick, B.G., & Fidell, L.S. (2007). *Using Multivariate Statistics* (5th ed.). Pearson.

Wang, Y., Rodríguez de Gil, P., Chen, Y.H., Kromrey, J.D., Kim, E.S., Pham, T., Nguyen, D., & Romano, J.L. (2017). Comparing the Performance of Approaches for Testing the Homogeneity of Variance Assumption in One-Factor ANOVA Models. *Educ. Psychol. Meas.*, *77*, 305–329. https://doi.org/10.1177/0013164416645162

Woodbury, G. (2002). *An Introduction to Statistics: Improving Your Grade*. Belmont, CA: Brooks/Cole.

Yi, Z., Chen, Y.H., Yin, Y., Cheng, K., Wang, Y., Nguyen, D., … Kim, E. (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

Yonar, A., Yonar, H., Demirsöz, M., & Tekindal, M.A. (2024). A comparative analysis for homogeneity of variance tests. *Journal of Science and Arts*, *24*(2), 305-328. https://doi.org/10.46939/J.Sci.Arts-24.2-a06

Zhou, Y., Zhu, Y., & Wong, K.Y. (2023). Statistical tests for homogeneity of variance for clinical trials and recommendations. *Contemporary Clinical Trials Communications, 33,* 101119. https://doi.org/10.1016/j.conctc.2023.101119