International Journal of Analysis and Applications

Comparative Robustness of Rank-Based Multiple Comparison Procedures under Non-Normality and Heteroscedasticity

Pupe Sudsila, Ampai Thongteeraparp*

Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand

**Corresponding author:* fsciamu@ku.ac.th

ABSTRACT. This study compares the performance of three rank-based nonparametric tests—Brunner–Munzel (BM), Bootstrap Rank Welch (BRW), and Fitted (FT) tests for multiple comparisons across three independent groups under conditions of non-normality and heteroscedasticity. Simulation data were generated using SAS 9.4 under normal, *t*, and lognormal distributions with three balanced sample sizes (10, 30 and 50 per group) and variance ratios (1:1:1, 1:2:3, and 1:4:7), based on 1,000 replications. The tests were evaluated according to Bradley's robustness criterion for Type I error control and statistical power. The results show that the BRW test achieves the highest power across *t* and lognormal distributions while maintaining Type I error control for all sample sizes. Under normality, the FT test demonstrates the strongest power but fails to control the error rate for large samples. The BM test remains stable in most conditions, providing a balance between robustness and efficiency. A practical decision rule is proposed to guide the selection of appropriate nonparametric methods for analyzing multiple comparisons under unequal variances and non-normal distributions.

1. Introduction

In applied research, investigators often deal with experiments or studies involving more than two treatment or comparison groups. The classical analysis of variance (ANOVA) provides a global test for the equality of group means; however, once the null hypothesis is rejected, it offers no information about which group means differ. To address this limitation, researchers typically apply multiple comparison procedures (MCPs) to identify specific pairwise differences among the means. Traditional MCPs—such as Tukey's, Scheffé's, Bonferroni's, and Duncan's tests—are derived under strict assumptions of normality and homogeneity of variances. In practice, these assumptions are rarely satisfied. Real-world data frequently exhibit skewness,

Received Oct. 28, 2025

2020 Mathematics Subject Classification. 62G10, 62G35.

Key words and phrases. nonparametric test; multiple comparisons; rank-based procedures; bootstrap rank Welch test; Brunner–Munzel test; fitted test.

https://doi.org/10.28924/2291-8639-24-2026-132

ISSN: 2291-8639

unequal variances, or small sample sizes, all of which may distort Type I error rates and reduce the statistical power of classical tests. When such conditions arise, nonparametric statistics offer a practical and robust alternative. Nonparametric approaches rely on ranks rather than the raw data values, allowing inference without strong distributional assumptions.

Several two-sample rank-based tests have been proposed to handle heteroscedastic or non-normal data. For example, the Fitted test (FT) is a modification of Welch's t-test designed to improve power under heteroscedasticity. Brunner and Munzel [1] developed a rank-based test for the nonparametric Behrens–Fisher problem, which performs well with small samples and heterogeneous variances. Reiczigel, Zakarias, and Rozsa [2] proposed the Bootstrap Rank Welch (BRW) test, which combines rank transformation with resampling to enhance robustness under heavy-tailed distributions. Neuhäuser [3] further examined the Brunner–Munzel test and confirmed its suitability for non-normal, skewed, and heteroscedastic data, even with small samples. Fagerland and Dandvik [4] compared several methods—including the *t*-test, Welch–U, Yuen–Welch, Wilcoxon–Mann–Whitney, and Brunner–Munzel tests—and concluded that each performs optimally under different data conditions.

Despite extensive investigations of pairwise comparisons between two groups, limited research has systematically compared these three rank-based tests—BM, BRW, and FT—in the context of multiple groups (≥3) under heteroscedastic and non-normal conditions. This study addresses that gap through a comprehensive Monte Carlo simulation comparing the Brunner–Munzel, Bootstrap Rank Welch, and Fitted tests in terms of their ability to control Type I error and maintain satisfactory statistical power. The objectives of this study are threefold:

1. To evaluate the robustness of the BM, BRW, and FT tests under varying distributions and variance ratios.
2. To compare their power performance conditional on acceptable Type I error control.
3. To propose a practical guideline for selecting an appropriate test according to data characteristics.

The remainder of this paper is organized as follows. Section 2 provides a brief theoretical overview of the three rank-based tests. Section 3 describes the simulation design and evaluation criteria. Section 4 presents the empirical results, while Section 5 discusses their theoretical and practical implications. Finally, Section 6 outlines the study's limitations and future research directions, and Section 7 concludes the paper.

2. Theoretical Framework of Rank-Based Multiple Comparison Procedures

This section summarizes the theoretical background of the three nonparametric test statistics considered in this study: the Brunner–Munzel (BM), Bootstrap Rank Welch (BRW), and Fitted

(FT) tests. All three belong to the family of rank-based approaches that assess stochastic dominance between groups without assuming normality or equal variances.

2.1 Brunner–Munzel Test (BM)

The Brunner–Munzel test [1] is a nonparametric alternative to the two-sample *t*-test, designed to handle unequal variances and non-normal data. It is based on the relative effect size, defined as the probability that a randomly selected observation from one group exceeds an observation from another.

The test statistic is calculated as in Equation (1):

$$BM = \frac{n_1 n_2 \left(\bar{r}_2 - \bar{r}_1\right)}{\left(n_1 + n_2\right)\sqrt{n_1 s_1^2 + n_2 s_2^2}} \tag{1}$$

Under the null hypothesis, the test statistic BM follows t distribution with degree of freedom

$$df_{BM} = \frac{\left(n_1 s_1^2 + n_2 s_2^2\right)^2}{\frac{\left(n_1 s_1^2\right)^2}{n_1 - 1} + \frac{\left(n_2 s_2^2\right)^2}{n_2 - 1}}$$

where

$$s_i^2 = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(r_{ik} - w_{ik} - \bar{r}_i + \frac{n_i + 1}{2}\right)^2$$

when $\bar{r}_i$ is the mean rank of sample of group *i*.

n_i is the sample size of group *i*.

r_{ik} is the rank of observation *k* within group *i*, and

w_{ik} is the rank of k sample of all group *i*.

The BM test is robust to variance heterogeneity, performs well with small to moderate sample sizes, and remains asymptotically consistent under the Behrens–Fisher framework.

2.2 Bootstrap Rank Welch Test (BRW)

The BRW test [2] extends the Welch-type approach by combining rank transformation with bootstrap resampling to approximate the sampling distribution of the test statistic. Through repeated resampling, it stabilizes inference under both non-normality and heteroscedasticity. The BRW statistic is computed as in Equation (2):

$$BRW = \frac{\left(\bar{r}_2 - \bar{r}_1\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \tag{2}$$

Under the null hypothesis, the BRW statistic approximately follows a *t*-distribution with degree of freedom

$$df_{BRW} = \frac{\left(n_2 s_1^2 + n_1 s_2^2\right)^2}{\frac{\left(n_2 s_1^2\right)^2}{n_1 - 1} + \frac{\left(n_1 s_2^2\right)^2}{n_2 - 1}}$$

where $$s_i^2 = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}(r_{ik} - \bar{r}_i)^2$$

When $\bar{r}_i$ is the mean rank of sample of group *i*.

n_i is the sample size of group *i*.

r_{ik} is the rank of observation *k* within group *i*.

The bootstrap procedure generates *B* resampled datasets from the pooled data, and the empirical distribution of the resampled BRW statistics is then used to obtain the *p*-value or confidence interval.

2.3 Fitted Test (FT)

The Fitted test (FT) modifies the Welch *t*-test by estimating fitted coefficients that reflect the variance structure of the data. This adjustment improves power under moderate heteroscedasticity and near-normal conditions. However, as sample size increases and variance disparity becomes more pronounced, the test may over-reject the null hypothesis, resulting in inflated Type I error rates. The FT statistic is expressed as in Equation (3):

$$FT = \frac{\left(\bar{r}_2 - \bar{r}_1\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(3)

where $$s_i^2 = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}(r_{ik} - \bar{r}_i)^2$$

When $\bar{r}_i$ is the mean rank of sample of group *i*.

n_i is the sample size of group *i*.

r_{ik} is the rank of observation *k* within group *i*.

Calculate the critical value as

$$V\left(\hat{C}\right) = \frac{a_1 + a_2\hat{C} + a_3\hat{C}^2}{1 + a_4\hat{C} + a_5\hat{C}^2}$$

when $$\hat{C} = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

Which $a_1, \ldots, a_5$ is the Fitted coefficient.

From $a_1 = t_{n_2-1,\alpha}$, $a_i = b_{i0} + \frac{b_{i1}}{n_1 - 1} + \frac{b_{i2}}{n_2 - 1} + \frac{b_{i3}}{(n_1 - 1)(n_2 - 1)} + \frac{b_{i4}}{(n_1 - 1)^2} + \frac{b_{i5}}{(n_2 - 1)^2}$; $i = 2,3,4$

and $a_5 = \frac{a_1 + a_2 + a_3}{t_{n_1-1,\alpha}} - 1 - a_4$, when $b_{i1}, ..., b_{i5}$; i = 2, 3, 4 are constant of the Fitted coefficient table at α level of significance.

The null hypothesis will be rejected if the computed *FT* exceeds the critical value corresponding to the α-level significance.

3. Simulation Design and Evaluation Criteria

The study employed Monte Carlo simulations to evaluate the robustness and efficiency of three rank-based tests—Brunner–Munzel (BM), Bootstrap Rank Welch (BRW), and Fitted (FT) for multiple comparisons under various non-normal and heteroscedastic conditions.

3.1 Data Generation Process

Data were generated for three independent groups. Three balanced sample sizes were considered: (10, 10, 10), (30, 30, 30), and (50, 50, 50). Three probability distributions were selected to represent distinct shapes and tail behaviors:

(1) The Normal distribution, representing a symmetric and light-tailed case.

(2) The t-distribution, representing moderately heavy-tailed data.

(3) The Lognormal distribution, representing positively skewed and heavy-tailed data.

Variance structures were set to reflect increasing degrees of heterogeneity: equal variances (1:1:1), moderate heterogeneity (1:2:3), and severe heterogeneity (1:4:7).

For each combination of distribution, variance structure, and sample size, 1,000 Monte Carlo replications were performed. All simulated data were generated independently and identically distributed within groups.

3.2 Statistical Tests and Implementation

All simulations were performed using SAS 9.4. For each replication, pairwise comparisons among the three groups were performed using the following tests:

(1) The Brunner–Munzel test (BM),

(2) The Bootstrap Rank Welch test (BRW), using 1,000 bootstrap resamples per comparison, and

(3) The Fitted test (FT).

Each test was conducted at a significance level of 0.05. The procedures were implemented identically across all simulation conditions to ensure comparability of results.

3.3 Evaluation Criteria

The performance of each method was evaluated using three key indices: Type I error control, statistical power, and error–performance balance.

1. **Type I Error Control**

For simulated datasets generated under $H_0: \mu_i = \mu_j$ versus $H_1: \mu_i \neq \mu_j$, for $i \neq j$, $i, j = 1, 2, 3$. The proportion of rejected null hypotheses was compared with the nominal level 0.05. Two types of error rates were evaluated:

1.1 Comparisonwise Error Rate (CER): the probability of committing a Type I error for a single comparison.

1.2 Experimentwise Error Rate (EER): the probability of committing at least one Type I error across all pairwise comparisons within an experiment.

$$CER = \frac{\text{Number of rejections of } H_0 \text{ when } H_0 \text{ is true}}{\text{Total number of pairwise comparisons}}$$

$$EER = \frac{\text{Number of experiments rejecting any } H_0 \text{ when } H_0 \text{ is true}}{\text{Total number of experiments}}$$

Following Bradley's robustness criterion [5], the empirical Type I error rate was considered acceptable if it fell within the range [0.025, 0.075]. This rule was used to determine whether each procedure maintained satisfactory error control under various combinations of non-normality and variance heterogeneity.

2. Statistical Power

For data generated under the alternative hypothesis H_1, the empirical power was defined as the proportion of correct rejections of H_0.

The magnitude of mean differences was induced while keeping variance structures constant across replications.

3. Error-Performance Balance

When multiple procedures met Bradley's criterion for robustness, the test achieving the highest empirical power was considered the most efficient.

All results were summarized in three-dimensional tables that present Type I error and power across different distributions, sample sizes, and variance ratios.

4. Results

A Monte Carlo simulation was conducted to evaluate the performance of the three test statistics—BM, BRW, and FT—under various distributional and variance conditions. The empirical Type I error rates and statistical power obtained from the simulation are summarized below.

4.1 Type I Error Control

The evaluation focuses on two primary indicators: the Comparisonwise Error Rate (CER) and the Experimentwise Error Rate (EER), which assess robustness at the pairwise and experimentwise levels, respectively.

Tables 1 and 2 report the CER and EER values across sample sizes (10, 30, and 50), variance ratios (1:1:1, 1:2:3, and 1:4:7), and underlying distributions (Normal, *t*, and Lognormal). These results provide a comprehensive view of each procedure's ability to control Type I error rates under both mild and severe violations of normality and homoscedasticity.

Table 1 Comparisonwise error rate (CER) at the 0.05 level of significance

Distributions	Variances ratios	Sample sizes								
		(10, 10, 10)			(30, 30, 30)			(50, 50, 50)		
		BM	BRW	FT	BM	BRW	FT	BM	BRW	FT
Normal	1:1:1	0.0560	0.0547	0.0403	0.0533	0.0500	0.0683	0.0487	0.0477	**0.0810**
	1:2:3	0.0540	0.0513	0.0383	0.0517	0.0523	0.0723	0.0497	0.0507	**0.0847**
	1:4:7	0.0460	0.0497	0.0383	0.0490	0.0560	0.0740	0.0540	0.0600	**0.0903**
t	1:1:1	0.0567	0.0550	0.0360	0.0490	0.0470	0.0560	0.0550	0.0547	**0.0827**
	1:2:3	0.0627	0.0600	0.0343	0.0490	0.0483	0.0627	0.0513	0.0503	**0.0833**
	1:4:7	0.0230	0.0507	0.0270	0.0510	0.0503	0.0640	0.0517	0.0523	**0.0783**
Lognormal	1:1:1	0.0523	0.0490	0.0250	0.0497	0.0467	0.0390	0.0457	0.0450	**0.0843**
	1:2:3	0.0497	0.0493	0.0283	0.0513	0.0513	0.0560	0.0483	0.0503	**0.0873**
	1:4:7	0.0473	0.0533	0.0277	0.0573	0.0660	0.0553	0.0517	0.0583	**0.0833**

Note bold underline means the statistical test cannot control probability of type I error based on Bradley's robustness criterion

As shown in Table 1, both the BM and BRW tests maintained the CER within Bradley's acceptable range (0.025–0.075) under most distributional and variance conditions. In contrast, the FT test occasionally exceeded the upper limit, particularly under strong heteroscedasticity or with larger sample sizes, indicating greater sensitivity to variance inequality.

To assess error control at the experimentwise level, Table 2 summarizes the corresponding EER values. The EER reflects the probability of committing at least one Type I error across all pairwise comparisons within an experiment, thereby providing insight into global error control.

Table 2 Experimentwise error rate (EER) at the 0.05 significance level

Distributions	Variance ratios	Sample size								
		(10, 10, 10)			(30, 30, 30)			(50, 50, 50)		
		BM	BRW	FT	BM	BRW	FT	BM	BRW	FT
Normal	1:1:1	**0.1340**	**0.1340**	**0.1010**	**0.1290**	**0.1210**	**0.1640**	**0.1230**	**0.1200**	**0.1940**
	1:2:3	**0.1290**	**0.1230**	**0.0930**	**0.1300**	**0.1320**	**0.1800**	**0.1230**	**0.1240**	**0.2010**
	1:4:7	**0.1110**	**0.1180**	**0.0920**	**0.1170**	**0.1330**	**0.1700**	**0.1170**	**0.1300**	**0.1910**
t	1:1:1	**0.1420**	**0.1380**	**0.0890**	**0.1230**	**0.1180**	**0.1770**	**0.1320**	**0.1300**	**0.1950**
	1:2:3	**0.1530**	**0.1470**	**0.0820**	**0.1220**	**0.1210**	**0.1660**	**0.1220**	**0.1200**	**0.1900**
	1:4:7	0.0680	**0.1270**	0.0710	**0.1230**	**0.1210**	**0.1480**	**0.1290**	**0.1300**	**0.1920**
Lognormal	1:1:1	**0.1280**	**0.1210**	0.0610	**0.1260**	**0.1180**	**0.1000**	**0.1200**	**0.1180**	**0.1080**
	1:2:3	**0.1280**	**0.1260**	0.0470	**0.1280**	**0.1280**	**0.1450**	**0.1190**	**0.1250**	**0.2330**
	1:4:7	**0.1120**	**0.1220**	0.0470	**0.1310**	**0.1480**	**0.1420**	**0.1320**	**0.1460**	**0.2160**

Note bold underline means the statistical test cannot control probability of type I error based on Bradley's robustness criterion

As summarized in Table 2, all three procedures exhibited inflated EER values under the Normal and *t*-distributed conditions. However, the FT test achieved acceptable EER control only for small samples under the Lognormal distribution, whereas both BM and BRW tended to become conservative in those scenarios.

Overall, these results confirm that controlling the experimentwise error rate is more challenging than the comparisonwise rate, especially when assumptions of normality and variance homogeneity are violated. Among the three methods, the BRW procedure provides the most stable Type I error performance across heterogeneous data conditions, while the FT test remains highly sensitive to heteroscedasticity and distributional skewness.

4.2 Statistical Power

To complement the Type I error analysis, this section examines the statistical power of the three test procedures. Table 3 presents the estimated power values based on 1,000 replications under each simulation setting, highlighting each method's ability to correctly reject false null hypotheses under different combinations of non-normality and heteroscedasticity.

Table 3 Power of Test for BM, BRW, and FT at 0.05 Significance Level

Distributions	Variance ratios	Sample size								
		(10,10,10)			(30,30,30)			(50,50,50)		
		BM	BRW	FT	BM	BRW	FT	BM	BRW	FT
Normal	1:1:1	0.4307	0.4220	**0.4770**	0.7607	0.7553	**0.8120**	**0.9140**	0.9117	-
	1:2:3	0.2500	0.2420	**0.3103**	0.5670	0.5707	**0.6390**	0.7183	**0.7193**	-
	1:4:7	0.1520	0.1443	**0.2513**	0.3573	0.3740	**0.4530**	0.5077	**0.5270**	-
t	1:1:1	0.7323	**0.9337**	0.8783	0.9970	**1.0000**	0.9983	**1.0000**	**1.0000**	-
	1:2:3	0.7253	**0.8980**	0.8170	0.9887	**0.9997**	0.9830	**1.0000**	**1.0000**	-
	1:4:7	0.4373	**0.8803**	0.7670	0.9830	**0.9997**	0.9600	0.9993	**1.0000**	-
Lognormal	1:1:1	**0.4690**	0.4360	0.4027	**0.7777**	0.7627	0.7627	**0.9103**	0.9093	-
	1:2:3	**0.2927**	0.2763	0.2437	**0.5777**	0.5703	0.5140	**0.7487**	0.7310	-
	1:4:7	**0.1703**	0.1697	0.1230	0.3747	**0.3887**	0.3413	0.5113	**0.5303**	-

Note '–' indicates that the procedure failed to satisfy Bradley's robustness criterion for Type I error control.

When the data followed a Normal distribution, the FT test produced the highest power for small and moderate sample sizes (10 and 30) across all variance structures, while the BM test performed best under equal variances. The BRW test surpassed both methods in unequal variance conditions, particularly for larger samples. Under the *t*-distribution, BRW consistently achieved the highest power across all settings, confirming its superior robustness for heavy-tailed data. When the data followed a Lognormal distribution, the BM test performed best for small samples (n = 10), but BRW outperformed as variance inequality or sample size increased. Overall, power increased with larger sample sizes and decreased under stronger heteroscedasticity. These results demonstrate that while the FT test remains optimal under near-normality, the BRW test provides the most reliable performance across non-normal and unequal variance conditions, combining robustness and efficiency.

4.3 Comparative Summary

To synthesize the findings, the results presented in Tables 1–3 were consolidated into a comparative summary (Table 4). This integration allows for a direct evaluation of the three test procedures—BM, BRW, and FT—in terms of both Type I error control and power performance.

By aggregating results across all heteroscedastic conditions, this summary highlights the consistency, robustness, and efficiency of each method under increasing variance disparity.

Table 4 Comparative Summary of Power and Type I Error Across Methods.

Distributions	variance ratios	Sample size		
		(10,10,10)	(30,30,30)	(50,50,50)
Normal	1:1:1	FT	FT	BM
	1:2:3	FT	FT	BRW
	1:4:7	FT	FT	BRW
t	1:1:1	BRW	BRW	BM, BRW
	1:2:3	BRW	BRW	BM, BRW
	1:4:7	BRW	BRW	BRW
Lognormal	1:1:1	BM	BM	BM
	1:2:3	BM	BM	BM
	1:4:7	BM	BRW	BRW

A consistent trend emerges: the BRW test provides the most balanced performance, maintaining empirical Type I error rates close to the nominal 0.05 while achieving high power across all variance conditions. The BM test performs reliably under moderate heterogeneity but tends to be conservative as variance ratios increase. The FT test, despite its computational simplicity, shows inflated error rates and lower power under severe heteroscedasticity.

Analytical Interpretation: The evidence in Table 4 establishes a clear performance hierarchy among the three procedures. By combining rank transformation with a resampling-based Welch framework, the BRW test achieves a strong balance between robustness and efficiency. These comparative findings confirm the method's reliability under unequal variances and serve as a conceptual bridge to the subsequent discussion of theoretical implications and potential applications.

5. Discussion

Taken together, the findings demonstrate that the BRW procedure achieves a statistically reliable balance between robustness and efficiency across all heteroscedastic conditions examined. It consistently preserved the nominal Type I error rate while attaining higher statistical power than both the BM and FT tests. This advantage stems from its dual mechanism of rank transformation and bootstrap resampling, which jointly stabilize variance estimation and correct bias arising from unequal group variances. By incorporating the Welch-type adjustment within a bootstrap framework, the BRW test effectively adapts to heterogeneous data without depending on the assumptions of normality or homoscedasticity.

The contrast between BM and FT further clarifies their methodological differences. Although the BM test provides a theoretically justified nonparametric alternative, its conservative tendency under large variance ratios limits sensitivity in detecting genuine group effects.

Conversely, the FT test—while computationally straightforward—relies on fitted variance structures that may not adequately capture true heterogeneity in small or unbalanced samples, resulting in inflated Type I errors and reduced power. These comparative outcomes reaffirm that the BRW procedure provides a more flexible and empirically robust solution, balancing Type I error control with power efficiency. The implications extend beyond simulation. In applied contexts—particularly biomedical and social-science research where variance heterogeneity and non-normality are common—the BRW test serves as a dependable tool for post-hoc comparisons or independent-sample inference. Its resampling-based design enables more accurate *p*-values and confidence intervals without rigid parametric constraints. Moreover, the conceptual foundation of the BRW test suggests promising extensions to rank-based multiple-comparison procedures or bootstrap-enhanced nonparametric ANOVA frameworks, paving the way for broader methodological integration.

Overall, the BRW approach not only improves empirical performance but also provides theoretical coherence and interpretability, making it a valuable contribution to the development of robust nonparametric inference.

6. Limitations and Future Work

While this study offers valuable insights into the comparative performance of the BRW, BM, and FT procedures under heteroscedasticity, several limitations should be acknowledged. First, the simulation framework was limited to balanced group sizes and a fixed number of replications. Although this design allowed controlled comparison, it may not fully capture the variability of practical data where group sizes often differ. Second, the range of heteroscedasticity considered was moderate; more extreme variance ratios or nonstandard error structures (e.g., heavy-tailed or highly skewed distributions) could yield different inferential patterns. Third, the analysis focused solely on continuous outcomes; extending the BRW framework to ordinal, censored, or mixed-type responses would enhance its applicability.

From a methodological standpoint, future research should consider developing generalized BRW-type procedures that incorporate adaptive weighting or robust resampling strategies. Investigating the asymptotic properties of the BRW statistic under various dependence structures may also deepen understanding of its robustness. Finally, integrating bootstrap-based rank statistics with machine-learning or Bayesian frameworks could represent a promising direction, enabling robust nonparametric inference for complex data settings. Such developments would strengthen both the theoretical foundation and the practical relevance of BRW-based methodology across disciplines.

7. Conclusion

Building on the comparative evidence and methodological implications discussed above, this study re-emphasizes its essential contribution and practical significance. The simulation results show that the BRW test provides the most balanced performance—maintaining both comparisonwise and experimentwise error rates within acceptable limits while achieving high power across diverse scenarios. The BM test behaves slightly more conservatively yet remains reliable, particularly when stringent Type I error control is required. In contrast, the FT test achieves the highest power only under normality and equal variances, reflecting its limited robustness for non-normal or heteroscedastic data. Accordingly, the BRW test is recommended for general use in applications involving unequal variances or skewed distributions, whereas BM may be preferred for confirmatory analyses emphasizing error control.

These findings reinforce the importance of rank- and resampling-based inference as powerful, resilient alternatives to classical parametric methods in multiple-comparison analysis.

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

References

[1] E. Brunner, U. Munzel, The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation, Biom. J. 42 (2000), 17-25. https://doi.org/10.1002/(sici)1521-4036(200001)42:1<17::aid-bimj17>3.0.co;2-u.

[2] J. Reiczigel, I. Zakariás, L. Rózsa, A Bootstrap Test of Stochastic Equality of Two Populations, Am. Stat. 59 (2005), 156-161. https://doi.org/10.1198/000313005x23526.

[3] M. Neuhäuser, A Nonparametric Two-Sample Comparison for Skewed Data with Unequal Variances, J. Clin. Epidemiol. 63 (2010), 691-693. https://doi.org/10.1016/j.jclinepi.2009.08.026.

[4] M.W. Fagerland, L. Sandvik, Performance of Five Two-Sample Location Tests for Skewed Distributions with Unequal Variances, Contemp. Clin. Trials 30 (2009), 490-496. https://doi.org/10.1016/j.cct.2009.06.007.

[5] J.V. Bradley, Robustness?, Br. J. Math. Stat. Psychol. 31 (1978), 144-152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x.