The European Research Journal 2023;9(4):687-696 Original Article

DOI: 10.18621/eurj.1037546 *Biostatistics*

Comparison of the performances of non-parametric k-sample test procedures as an alternative to one-way analysis of variance

Aslı Ceren Macunluoğlu[1], Gökhan Ocakoğlu[2]

[1]Department of Biostatistics, Bursa Uludağ University, Institute of Health Sciences, Bursa, Turkey; [2]Department of Biostatistics, Bursa Uludağ University, Faculty of Medicine, Bursa, Turkey

ABSTRACT

Objectives: The performances of the Kruskal-Wallis test, the van der Waerden test, the modified version of Kruskal-Wallis test based on permutation test, the Mood's Median test and the Savage test, which are among the non-parametric alternatives of one-way analysis of variance and included in the literature, to protect the Type-I error probability determined at the beginning of the trial at a nominal level, were compared with the F test.

Methods: Performance of the tests to protect Type-I error; in cases where the variances are homogeneous/heterogeneous, the sample sizes are balanced/unbalanced, the distribution of the data is in accordance with the normal distribution/the log-normal distribution, how it is affected by the change in the number of groups to be compared has been examined on simulation scenarios.

Results: The Kruskal-Wallis test, the van der Waerden test, the modified version of the Kruskal-Wallis test based on the permutation test were not affected by the distribution of the data, but by the violation of the homogeneity of the variances. The performance of the Mood's Median test and the Savage test were not found to be sufficient in terms of protection of theType-I error compared to other tests.

Conclusions: It was determined that the Kruskal-Wallis test, the van der Waerden test, the modified version of Kruskal-Wallis test based on permutation test were not affected by the distribution of the data and tended to preserve the Type-I error when the variances were homogeneous.

Keywords: Analysis of variance, conformity of normal distribution, non-parametric k-sample tests

Data analysis methods that will be allied to the data obtained from research with at least interval scale; variance varies according to sample size, distribution of data, and the number of groups to be compared. One of the most critical steps of statistical data analysis is to decide whether the test procedure to be used to analyze the data will be a parametric or non-parametric test. Parametric tests are statistical methods that require data to be measured on an interval or ratio scale, which can be applied due to certain assumptions. Non-parametric test procedures are alternatively preferred when the necessary assumptions are not met for performing parametric tests.

One-way analysis of variance (ANOVA) or F-test,

Received: ***December 16, 2021;*** *Accepted:* ***August 10, 2022;*** *Published Online:* ***October 24, 2022***

How to cite this article: *Macunluoğlu AC, Ocakoğlu G. Comparison of the performances of non-parametric k-sample test procedures as an alternative to one-way analysis of variance. Eur Res J 2023;9(4):687-696. DOI: 10.18621/eurj.1037546*

Address for correspondence: *Gökhan Ocakoğlu, PhD., Professor, Bursa Uludağ University, Faculty of Medicine , Department of Biostatistics, Bursa, Turkey, E-mail: gocakoglu@gmail.com, Phone: +90 224 295 38 71*

e-ISSN: 2149-3189

which is a parametric test, is used to compare the mean of more than two populations and is one of the most frequently used and most important statistical methods for this purpose [1]. The assumptions for the F test include that the data is normally distributed, the sample variances are equal, and the samples are independent [2]. If the assumptions of conformity to normal distribution or homogeneity of variance are violated, the probability of Type-I error obtained at the end of the trials and the power of the test are adversely affected. This adversely affect becomes even more evident if the sample sizes in the groups compared are not balanced [3]. Non-parametric tests are statistical procedures that are preferred as an alternative to parametric tests when assumptions are not met. Non-parametric tests have less assumptions than parametric tests [4]. The data need not conform to a normal distribution. Non-parametric tests can be applied to data measured with a classifier or ordinal scale.

Pearson [5], Glass *et al*. [6], and Wilcox [7] examined the effect of the normality assumption violation on the Type-I error. Wilcox [7] concluded that samples that do not conform to normal distribution have some impact on the Type-I error rate, but the effect is mini-

Table 1. Sample sizes of the groups

Number of groups	Balanced Sample	Non-balanced Sample		
		Observation combinations where the number of sample sizes are not equal	Observation combinations where the number of sample sizes differs excessively	Observation combinations with inverse matching between variance and number of sample sizes
3	3:3:3 5:5:5 10:10:10 15:15:15 20:20:20 25:25:25 30:30:30 50:50:50 80:80:80 100:100:100	3:5:7 5:10:15 20:25:30 50:60:70 65:75:85 70:90:100	3:25:30 3:80:80 5:20:100	7:5:3 15:10:5 30:25:20 70:60:50 85:75:65 100:90:70
5	3:3:3:3:3 5:5:5:5:5 10:10:10:10:10 15:15:15:15:15 20:20:20:20:20 25:25:25:25:25 30:30:30:30:30 50:50:50:50:50 80:80:80:80:80 100:100:100:100:100	3:5:7:9:11 5:7:9:12:15 20:22:24:28:30 50:55:60:65:70 55:65:75:85:95 60:70:80:90:100	3:20:25:80:100 3:5:30:80:100 5:10:20:25:80 3:5:10:15:100	7:5:3 15:10:5 30:25:20 70:60:50 85:75:65 100:90:70
8	3:3:3:3:3:3:3:3 5:5:5:5:5:5:5:5 10:10:10:10:10:10:10:10 15:15:15:15:15:15:15:15 20:20:20:20:20:20:20:20 25:25:25:25:25:25:25:25 30:30:30:30:30:30:30:30 50:50:50:50:50:50:50:50 80:80:80:80:80:80:80:80 100:100:100:100:100:100:100:100	3:5:7:9:11:12:14:15 20:22:24:25:26:28:29:30 50:55:60:65:70:75:80:85 60:65:75:80:85:90:95:100	3:5:10:20:25:30:80:100 5:10:20:20:25:80:90:100 3:5:10:80:80:90:100:100 20:25:30:80:90:90:100:100	15:14:12:11:9:7:5:3 30:29:28:26:25:24:22:20 85:80:75:70:65:60:55:50 100:95:90:85:80:75:65:60

mal if the variances are homogeneous. Glass *et al*. [6] reported similar results to Wilcox [7] in their studies if the variances were homogeneous. In his study, Bunning [8] examined the performances of the Kruskal-Wallis test, the normal score test and the Welch test, which he included as an alternative to the F test and the F test, in terms of Type-I error and power. He evaluated the performances of the tests under various simulation scenarios in terms of whether the variances are homogeneous or not in equal and unequal sample sizes if the data show normal distribution or not. In his study, Moder [2] stated that the location parameters of the groups should be investigated in detail when there are unbalanced sample sizes.

In our study, we compared the performances of the Kruskal-Wallis test, the Mood's Median test, the van der Waerden test, the modified version of Kruskal-Wallis test based on permutation test and the Savage test, which are among the non-parametric alternatives of the F test, to protect the Type-I error under various simulation scenarios.

METHODS

In our study, the Kruskal-Wallis test, the modified version of Kruskal-Wallis test based on permutation test, the Mood's Median test, the van der Waerden test and the Savage test in terms of maintaining the probability of the Type-I error determined at the beginning of the experiment was compared with the F test. Simulation scenarios were run under the R program [9].

The performance of the tests was evaluated as a result of comparisons between three, five, and eight groups for simulation scenarios involving balanced/non-balanced sample sizes (Table 1), normal distribution or log-normal distribution, homogenous or heterogeneous variances (Table 2). In addition to the specified simulation conditions, observation combinations are also included, where the number of sample size varies excessively among the group with higher variance is assigned a lower number of observations, and the group with a lower variance is assigned a higher number of observations and inverse matching between variance and sample size.

In comparisons made to determine Type-I error, group means were taken equally. The Type- I error probabilities for each of the simulation scenarios were obtained after the numbers of H_0 hypotheses were determined, which were rejected at the end of 50000 repetitions. In our study, the evaluation criterion proposed by Peterson [10] was adopted and it was concluded that the performance of the tests with a probability of the Type-I error between 4.49% and 5.49% was sufficient to maintain Type-I error.

Table 2 shows the variance rates of the groups that are suitable for normal distribution and the scale parameter values of the groups that are suitable for log-normal distribution.

The F Test

One-way analysis of variance (ANOVA) or F test is used to compare the mean of more than two populations. It is one of the most important and frequently used methods of applied statistics [1]. The null hypothesis H_0: $\mu_1=\mu_2=\ldots=\mu_k$ versus alternative H_1: at least one μ_i (i= 1, 2, . . ., k) is different. The F test statistic,

$$F = \frac{\sum_{i=1}^{k} n_i(X_{i.}-\bar{X}_{..})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{x}_{i.})^2/(N-k)} \sim F_{1-\alpha;k-1,N-k} \qquad (1)$$

In Equation, k is the number of groups, N is the total number of observations, Xij is the jth observation (j = 1, 2, . . ., *ni*) in the ith group (i = 1, 2, . . . , *k*), $N=\sum n_i$, $\bar{X}_{..}$ is the overall mean, $\bar{X}_{i.}$ is the sample mean for the *i*th group. The F test is more powerful if the assumptions of normality and variance homogeneity hold. The null hypothesis, H_0: $\mu_1=\mu_2=\ldots=\mu_k$, should then be rejected at the α level of significance when $F \geq F_{1-\alpha;k-1,N-k}$.

The Kruskal-Wallis Test

One of the non-parametric alternatives to the F test is the Kruskal-Wallis (KW) test. The KW test is a non-parametric test procedure used to compare three or more groups independently [11]. It is carried out using ranks given to observation values instead of actual observation values. To calculate the test statistics, the data are sorted from small to large, and each is assigned a rank. $R_i=\sum_{j=1}^{n_i} R_{ij}$ is the sum of ranks assigned to the observations in the ith group. The null hypothesis H_0: $\theta_1=\theta_2=\ldots=\theta_k$ versus alternative H_1: at least one θ_i (i= 1, 2, . . ., k) is different. The test statistic is calculated as,

Table 2. Variance rates of groups

	Normal distribution		Log- normal distribution	
Number of groups	Homogeneous Variance	Heterogeneous Variance	Homogeneous Scale parameter (b)	Heterogeneous scale parameter (b)
3	1:1:1 2:2:2 4:4:4 8:8:8 10:10:10	1:1:2 1:2:2 1:1:4 1:4:4 1:1:8 1:8:8 1:1:10 1:10:10 1:4:8 2:1:1 2:2:1 4:1:1 4:4:1 8:1:1 8:8:1 10:1:1 10:10:1 8:4:1	0.1:0.1:0.1 0.2:0.2:0.2 0.3:0.3:0.3 0.4:0.4:0.4 0.5:0.5:0.5 0.6:0.6:0.6 0.7:0.7:0.7 0.8:0.8:0.8	0.10:0.10:0.20 0.10:0.20:0.20 0.10:0.30:0.50 0.10:0.40:0.50 0.10:0.10:0.50 0.10:0.50:0.60 0.10:0.60:0.80 0.20:0.10:0.10 0.20:0.20:0.10 0.50:0.30:0.10 0.50:0.40:0.10 0.50:0.10:0.10 0.60:0.50:0.10 0.80:0.60:0.10
5	1:1:1:1:1 2:2:2:2:2 4:4:4:4:4 8:8:8:8:8 10:10:10:10:10	1:1:2:2:2 1:1:4:4:4 1:1:8:8:8 1:1:10:10:10 1:2:4:8:10 2:2:2:1:1 4:4:4:1:1 8:8:8:1:1 10:10:10:1:1 10:8:4:2:1	0.1:0.1:0.1:0.1:0.1 0.2:0.2:0.2:0.2:0.2 0.3:0.3:0.3:0.3:0.3 0.4:0.4:0.4:0.4:0.4 0.5:0.5:0.5:0.5:0.5 0.6:0.6:0.6:0.6:0.6 0.7:0.7:0.7:0.7:0.7 0.8:0.8:0.8:0.8:0.8	0.1:0.1:0.2:0.2:0.2 0.1:0.1:0.4:0.4:0.4 0.1:0.1:0.5:0.5:0.5 0.1:0.1:0.6:0.7:0.8 0.1:0.3:0.5:0.7:0.8 0.2:0.2:0.2:0.1:0.1 0.4:0.4:0.4:0.1:0.1 0.5:0.5:0.5:0.1:0.1 0.8:0.7:0.6:0.1:0.1 0.8:0.7:0.5:0.3:0.1
8	1:1:1:1:1:1:1:1 2:2:2:2:2:2:2:2 4:4:4:4:4:4:4:4 8:8:8:8:8:8:8:8 10:10:10:10:10:10:10:10	1:1:1:1:1:1:1:2 1:1:1:1:1:1:1:4 1:1:1:1:1:1:1:8 1:1:1:1:1:1:1:10 1:1:1:2:2:2:4:4 1:1:1:1:4:4:4:4 1:1:1:1:8:8:10:10 2:1:1:1:1:1:1:1 4:1:1:1:1:1:1:1 8:1:1:1:1:1:1:1 10:1:1:1:1:1:1:1 4:4:2:2:2:1:1:1 4:4:4:4:1:1:1:1 10:10:8:8:1:1:1:1 10:10:8:8:4:4:2:1	0.1:0.1:0.1:0.1:0.1:0.1:0.1:0.1 0.2:0.2:0.2:0.2:0.2:0.2:0.2:0.2 0.3:0.3:0.3:0.3:0.3:0.3:0.3:0.3 0.4:0.4:0.4:0.4:0.4:0.4:0.4:0.4 0.5:0.5:0.5:0.5:0.5:0.5:0.5:0.5 0.6:0.6:0.6:0.6:0.6:0.6:0.6:0.6 0.7:0.7:0.7:0.7:0.7:0.7:0.7:0.7 0.8:0.8:0.8:0.8:0.8:0.8:0.8:0.8	0.1:0.1:0.1:0.1:0.1:0.1:0.1:0.1:0.2 0.1:0.1:0.1:0.1:0.1:0.1:0.1:0.1:0.3 0.1:0.1:0.1:0.1:0.1:0.1:0.1:0.1:0.5 0.1:0.1:0.1:0.1:0.1:0.1:0.1:0.1:0.7 0.1:0.1:0.1:0.3:0.3:0.3:0.5:0.5 0.1:0.1:0.1:0.1:0.6:0.6:0.8:0.8 0.2:0.3:0.4:0.5:0.6:0.7:0.7:0.8 0.2:0.2:0.2:0.4:0.4:0.8:0.8:0.8 0.2:0.1:0.1:0.1:0.1:0.1:0.1:0.1 0.3:0.1:0.1:0.1:0.1:0.1:0.1:0.1 0.5:0.1:0.1:0.1:0.1:0.1:0.1:0.1 0.7:0.1:0.1:0.1:0.1:0.1:0.1:0.1 0.5:0.5:0.3:0.3:0.3:0.1:0.1:0.1 0.8:0.8:0.6:0.6:0.1:0.1:0.1:0.1 0.8:0.7:0.7:0.6:0.5:0.4:0.3:0.2 0.8:0.8:0.8:0.4:0.4:0.2:0.2:0.2

$$KW = \frac{1}{s^2}\left(\sum_{i=1}^{K}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right) \quad (2)$$

where

$$S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} r_{ij}^2 - \frac{N(N+1)^2}{4}\right) \quad (3)$$

Note that, when there are no ties, S^2 simplifies to $N(N+1)/12$.

The null hypothesis H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$,should then be rejected at the α level of significance when $\chi^2_{KW} \geq \chi^2_{k-1;\alpha}$.

The Modified Version of Kruskal-Wallis Test Based on Permutation Test

Permutation test is the test procedure which is presented by Fisher [12] and the probability values obtained are exact probabilities, and it is also stated by Hecke [13] as a simulation method used to determine the strength of the test.

There are two methods for calculating the KW test: permutation and rank transformations. The modified version of the KW test based on the permutation test is obtained by combining the permutation method based on the F statistic with the rank method [11]. The process of obtaining the permutations starts by choosing the test statistic T and the acceptable significance level α. $\pi_1, \pi_2, \ldots, \pi_n$ be a set of all distinct permutations of the ranks of the data set in the experiment. For permutation testing, the data are sorted from small to large, each is given a rank and the KW test statistic is calculated ($H_1=t_0$). Different permutation (π_i) values are obtained for each data sorted from small to large. The KW test statistic is calculated for the obtained permutation (n_i) values ($H_i=H(\pi_i)$) and this process i is repeated (i= 2, 3, ..., M).

The null hypothesis H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$ versus alternative H_1: at least one θ_i (i= 1, 2, …, k) is different. The test statistic is calculated as [13],

$$p_0 = p(H \leq H_i) = \frac{1}{M}\sum_{i=1}^{M}\psi(t_0 - H_i) \quad (4)$$

where

$$\psi(.) = \begin{cases} 1, if\ t_0 \geq H_i \\ 0, if\ t_0 < H_i \end{cases}$$

Under the empirical distribution, if $p_0 \leq \alpha$, reject the null hypothesis.

The Mood's Median Test

The Mood's Median (MM) test is the generalized version of the median test used to test data from two independent groups, used for three and more sample comparisons [15]. The null hypothesis H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$ versus alternative H_1: at least one θ_i (i= 1, 2, …, k) is different.

To obtain the test statistics of the MM test, the common median value of all data is first calculated. As a second step, for each sample, it is determined how many observations are greater than the calculated median value and how many are equal to or less than it. As a result, a 2xk frequency table is obtained. The test statistic is calculated as,

$$\chi^2 = \sum_{i=1}^{k}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \quad (5)$$

The null hypothesis H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$,should then be rejected at the α level of significance when $\chi^2 \geq \chi^2_{\alpha;(i-1)*(j-1)}$

The van der Waerden Test

The advantage of the van der Waerden test is that it provides the high efficiency of the standard ANOVA analysis when the normality assumptions are in fact satisfied, but it also provides the robustness of the KW test when the normality assumptions are not satisfied [16]. The KW test is based on the ranks of the data. The van der Waerden test converts the ranks to quantiles of the standard normal distribution. These are called normal scores and the test is computed from these normal scores [17]. The null hypothesis H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$ versus alternative H_1: at least one θ_i (i= 1, 2, …, k) is different. The formula for the van der Waerden test is

$$V = \frac{1}{S^2}\sum_{i=1}^{k} n_i \bar{A}_i^2 \quad (6)$$

where

$$A_{ij} = \phi^{-1}\left(\frac{R_{ij}}{N+1}\right), \bar{A}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} A_{ij}, S^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i} A_{ij}^2, \phi^{-1}$$

is the normal quantile of x. The null hypothesis should then be rejected at the α level of significance when $V \geq \chi^2_{\alpha;k-1}$

The Savage Test

The Savage test is among the non-parametric alternatives to the F test used to test the differences between location parameters. The Savage test is powerful to compare scale differences or position differences in the extreme value distribution, which are compatible with exponential distribution [18].

The Savage test statistic is calculated by Savage scores. The null hypothesis H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$ versus alternative H_1: at least one θ_i (i= 1, 2, …, k) is different. The formula for the Savage test is

$$T_E = \frac{1}{S_E^2}\sum_{i=1}^{k} n_i \bar{S}_i^2 \qquad (7)$$

where

$$S_{ij} = \sum_{i=1}^{k}\frac{1}{N-i+1} - 1 \; ; \; \bar{S}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} S_{ij} \; ; \; S_E^2 = \frac{1}{(N-1)}\sum_{i=1}^{k}\sum_{j=1}^{n_i} S_{ij}^2$$

The null hypothesis should then be rejected at the α level of significance when $T_E \geq \chi^2_{\alpha;k-1}$.

RESULTS

In this study, the tests were compared with the help of simulation scenarios in terms of the Type-I error protection. Simulation scenarios were performed under the R program [9]. The obtained Type-I errors are given in tables.

Comparisons in which sample size is balanced, the group variances are homogeneous, and the data follow to the normal distribution (Table 3, Supplementary Table 1, 2)

The F test is the test that shows the most successful performance when the non-parametric alternatives are taken into consideration and the predetermined Type-I error level was determined.

In addition to the F test, the KW test also tends to maintain the Type-I error level in terms of observation combinations, and the increase in the number of groups to be compared (especially in the case of eight groups) has a positive effect on its performance.

Comparisons in which the sample size is not balanced, the group variances are homogeneous, and the data follow to the normal distribution (Supplementary Table 3-8)

The F test and the KW test based on permutation test are the most successful tests for estimating the Type-I error level initially determined. The F test and the modified version of KW test based on permutation test are followed by the KW test with deflection estimates shown only in a single simulation scenario.

The other tests included in the study were found to be adversely affected by the imbalance of the number of sample sizes in the groups, and their performance to protect the Type-I error determined at the beginning was not sufficient.

When simulation scenarios involving observation combinations in which the number of sample sizes in groups differ excessively, it was observed that the F test and modified version of KW test based on permutation test were not affected by the extreme differences in the number of sample sizes in groups and tended to maintain the Type-I error level initially determined in all simulation scenarios according to the Peterson criterion.

On the other hand, in cases where the number of sample size in the groups varies in a balanced manner, the KW test, which performs at a level that can accompany these two tests, was observed to have affected its performance and gave deviated results if the difference in the number of sample size was excessive.

Comparisons in which the sample size is balanced, group variances are heterogeneous, but the data follow to the normal distribution (Supplementary Table 9-11)

It has been seen that the tests included in the study generally give deviated results in terms of protecting the Type-I error and their performance was not found sufficient.

Comparisons in which the sample size is not balanced, group variances are heterogeneous, but the data follow to the normal distribution (Supplementary Table 12-20)

It has been seen that the tests included in the study generally give deviated results in terms of protecting

Table 3. Type-I error rates (%) for k=3 groups where $\sigma_1^2:\sigma_2^2:\sigma_3^2$= 1:…:1~10:…:10, $\mu_1=\mu_2=\mu_3=0$, sample size is balanced ($n_1=n_2=n_3$)

σ^2	n	F	KW	permKW	MM	VW	Savage
1	3	4.70%	8.51%	5.03%	0%	1.11%	0.71%
	5	4.98%	5.49%	4.90%	6.67%	4.85%	2.74%
	10	5.03%	5.11%	4.42%	3.48%	4.47%	3.97%
	15	5.09%	5.21%	4.74%	4.69%	4.78%	4.29%
	20	5.07%	5.15%	4.78%	4.74%	4.87%	4.43%
	25	5.08%	5.14%	4.83%	4.77%	4.78%	4.48%
	30	4.93%	4.92%	4.75%	4.15%	4.68%	4.63%
	50	5.11%	5.13%	4.99%	4.88%	4.99%	4.68%
	80	5.07%	5.08%	4.98%	5.54%	5.04%	4.89%
	100	5.04%	5.05%	5.00%	5.12%	5.00%	4.84%
2	3	4.82%	8.31%	4.60%	0%	1.02%	0.71%
	5	4.93%	5.51%	4.82%	6.80%	3.86%	2.64%
	10	4.88%	5.15%	4.29%	3.37%	4.42%	3.71%
	15	5.03%	5.20%	4.64%	4.67%	4.77%	4.19%
	20	4.80%	4.99%	4.51%	4.61%	4.62%	4.24%
	25	5.15%	5.26%	4.95%	4.88%	5.01%	4.67%
	30	5.03%	5.09%	4.86%	4.40%	4.96%	4.62%
	50	4.93%	5.01%	4.84%	4.79%	4.93%	4.79%
	80	4.99%	5.06%	4.92%	5.48%	4.95%	4.78%
	100	5.00%	4.97%	4.95%	5.10%	4.98%	4.76%
4	3	5.08%	8.57%	4.70%	0%	1.03%	0.67%
	5	4.96%	5.46%	4.79%	6.59%	3.83%	2.64%
	10	5.06%	5.29%	4.46%	3.31%	4.55%	3.82%
	15	4.77%	4.90%	4.60%	4.49%	4.73%	4.19%
	20	5.10%	5.10%	4.78%	4.74%	4.87%	4.43%
	25	4.95%	4.94%	4.84%	4.81%	4.85%	4.61%
	30	5.21%	5.13%	4.60%	4.14%	4.62%	4.52%
	50	5.10%	5.01%	4.80%	4.84%	4.78%	4.87%
	80	4.89%	4.85%	4.95%	5.42%	4.97%	4.83%
	100	4.86%	4.92%	4.85%	5.09%	4.88%	4.80%
8	3	4.97%	8.59%	4.84%	0%	1.13%	0.72%
	5	4.78%	5.29%	4.67%	6.66%	4.84%	2.73%
	10	5.07%	5.20%	4.46%	3.43%	4.55%	3.85%
	15	5.15%	5.35%	4.76%	4.72%	4.84%	4.42%
	20	4.90%	4.93%	4.62%	4.79%	4.56%	4.37%
	25	5.01%	5.12%	4.79%	4.70%	4.85%	4.55%
	30	5.00%	5.04%	4.82%	4.36%	4.76%	4.53%
	50	5.08%	5.03%	4.97%	4.82%	4.95%	4.84%
	80	5.21%	5.22%	5.13%	5.51%	5.12%	4.99%
	100	5.09%	4.93%	5.05%	5.14%	5.07%	5.01%
10	3	5.10%	8.66%	4.90%	0%	1.16%	0.72%
	5	4.99%	5.47%	4.91%	6.74%	3.90%	2.71%
	10	4.88%	5.08%	4.29%	3.46%	4.40%	3.76%
	15	5.01%	5.05%	4.61%	4.47%	4.61%	4.23%
	20	5.02%	5.09%	4.73%	4.85%	4.77%	4.28%
	25	5.08%	5.08%	4.85%	4.82%	4.83%	4.55%
	30	4.90%	4.99%	4.73%	4.38%	4.82%	4.52%
	50	5.06%	4.99%	4.84%	4.85%	4.86%	4.60%
	80	5.08%	5.10%	4.95%	4.88%	4.92%	4.68%
	100	5.09%	5.06%	5.04%	4.95%	5.00%	4.80%

F: F test; KW: Kruskal-Wallis test; permKW: the modified version of Kruskal-Wallis test based on permutation test; MM: Mood's Median test; VW: van der Waerden test

the Type-I error and their performance is not sufficient.

Comparisons in which the sample size is balanced, group variances are homogeneous, and the data follow to log-normal distribution (Supplementary Table 21-23)

As expected, the F test is the test that shows the most successful performance in order to estimate the level of Type-I error determined at the beginning when considering the non-parametric alternatives available.

In addition to the F test, the KW test also tends to maintain the Type-I error level in terms of observation combinations, and the increase in the number of groups to be compared (especially in the case of eight groups) has a positive effect on its performance.

The performance of the MM test was also positively affected by the increase in the number of groups. Although its performance in protecting the Type-I error is lower than that of the KW test, its performance in the case of eight groups has increased significantly compared to the number of groups to be compared with three and five.

Comparisons in which the sample size is not balanced, group variances are homogeneous, and the data follow to log-normal distribution (Supplementary Table 24-29)

When simulation scenarios involving observation combinations in which the number of sample size in the groups are not equal are examined, the F test and the KW test are the tests that show the most successful performance in order to estimate the Type-I error level determined at the beginning. These tests are followed by modified version of KW test based on permutation test.

When the simulation scenarios involving observation combinations in which the number of sample size in the groups differ excessively, it was observed that the permutation version of the F test and the KW test was not affected by the extreme differences in the number of sample size in the groups.

T he other tests included in the study were found to be adversely affected by the imbalance of the number of sample size in the groups, and their performance in maintaining the Type-I error level determined at the beginning was not sufficient.

Comparisons in which the sample size is balanced, group variances are heterogeneous, and the data follow to log-normal distribution Supplementary (Table 30-32)

It has been seen that the tests included in the study generally give deviated results in terms of protecting the Type-I error and their performance was not found sufficient.

Comparisons in which the sample size is not balanced, group variances are heterogeneous, and the data follow to log-normal distribution (Supplementary Table 33-41)

It has been seen that the tests included in the study generally give deviated results in terms of protecting the Type-I error and their performance was not found sufficient.

DISCUSSION

The F test is the test that shows the most successful performance as expected in cases where the conformity to the normal distribution and the homogeneity of the variances are provided. When the simulation scenarios where the assumption of homogeneity of variances are not met, as expected, the F test was highly affected by the deterioration in group variances and failed to maintain the Type-I error at the nominal level ($\alpha = 0.05$). The results of our study reach similar results to the studies conducted by Buning [8] and Moder [2]. It is the test that shows the most successful performance compared to other alternative tests in cases where the data conform to the log-normal distribution, and the variances are homogeneous. Blanca *et al*. [19] Clinch and Keselman [20], Gamage and Weerahandi [21], Lantz [22] and Schmider *et al*. [23] reported that the F test tends to protect the Type-I error in cases where the assumption of conformity to the normal distribution is violated. It was observed that the effect of violation of the homogeneity of variances on the performance of the F test was more than the violation of the assumption of conformity to normal distribution. Bishop and Dudewicz [3], Blanca *et al*. [19], Brown and Forsythe [24], Buning [8], Debeuckelaer [25], Lee and Ahn [26], Li *et al*. [27], Lu and Mathew [28], Markowski [29], Keselman *et al*. [30], Tomarken and Serlin [31] concluded that the F test is highly affected by the deterioration in group variances.

In this study, the KW test was not affected by the distribution of the data. It was concluded that the violation of the homogeneity of variances and the number of sample sizes (equal and unequal) in the groups were effective on the performance of the KW test to protect the Type-I error. In their studies, Hoeffding [32] ve Terry [33] concluded that the performance of the KW test was not sufficient in terms of protecting Type-I error in cases where the variance was not homogeneous. Lantz [22], Luh and Guo [34], Jett and Speer [35] found in their studies that the KW test was not affected by the distribution of the data, and in cases where the variances were homogeneous, they tend to protect the Type-I error.

The modified version of the KW test is not affected by the distribution of the data; It is highly affected by the homogeneity violation of variances such as the KW test. It can be suggested as an alternative for the F test for observation combinations where the number of sample sizes in the groups are not equal and excessively different. Odiase and Ogbonmwan [14] reported in their study that the permutation test does not require assumptions for the distribution of the data, and that it performs well on data that are normally distributed and not normally distributed.

The van der Wearden test was not affected by the distribution of the data and showed successful performance in protecting the Type-I error in observation combinations where the number of sample sizes in the groups where the group variances were homogeneous differed significantly. The van der Wearden test was greatly affected by the breakdown in group variance. Luepsen [1] stated that the van der Wearden test was the most successful test after the F test in estimating the Type-I error level in cases where there is no relationship between group variances and the number of observations belonging to the groups.

Although the MM test performed well as the number of groups compared increased, it did not show a successful performance in protecting the Type-I error in general. Jett and Speer [35] stated in their simulation studies that the performance of the MM test was not sufficient to protect the Type-I error and reported our study with supporting findings.

The Savage test could not perform adequately to protect the Type-I error at nominal level and gave biased results. There is no study in the literature regarding the Savage test. Our study aims to contribute to the literature by reporting that the Savage test's performance in protecting the Type-I error compared to other tests gives very poor and biased results.

CONCLUSION

In conclusion as stated in the literature, it was determined that the F test tends to maintain its robustness in case of violation of the normal distribution, however, it is more affected by the violation of the homogeneity assumption of variances. It was concluded that the distribution of the data was not effective on the KW test's performance in protecting Type-I error, the violation of homogeneity of variances and the sample size in the groups were effective. The modified version of KW test based on permutation test is not affected by the distribution of the data; like the KW test, it is highly affected by the violation of homogeneity of variances. It can be suggested as an alternative to the F test for combinations of observations where the sample sizes in the groups are not equal and vary excessively. The van der Wearden test was not affected by the distribution of the data and showed successful performance in protecting the Type-I error in observation combinations where the number of sample sizes in the groups where the group variances were homogeneous differed significantly. In general, the MM test did not show a successful performance in protecting the Type-I error. It has been found that the Savage test's performance in protecting the Type-I error compared to other tests gives very poor and biased results.

Authors' Contribution

Study Conception: GO; Study Design: GO; Supervision: GO; Funding: N/A; Materials: N/A ; Data Collection and/or Processing: ACM; Analysis and/or Data Interpretation: ACM, GO; Literature Review: ACM; Manuscript Preparation: ACM, GO and Critical Review: ACM, GO.

Conflict of interest

The authors disclosed no conflict of interest during the preparation or publication of this manuscript.

Financing

The authors disclosed that they did not receive any grant during conduction or writing of this study.

Supplementary Tables 1 to 41

REFERENCES

1. Luepsen H. Comparison of nonparametric analysis of variance methods: a vote for van der Waerden. Commun Stat Simul Comput 2018;47:2547-76.
2. Moder K. Alternatives to F-test in one way ANOVA in case of heterogeneity of variances (a simulation study). Psychol Test Assess Model 2010;52:343-53.
3. Bishop TA, Dudewicz EJ. Exact analysis of variance with unequal variances: test procedures and tables. Technometrics 1978;20:419-30.
4. McSeeney M, Katz B. Nonparametric statistics: use and nonuse. Percept Mot Skills 1978;46(3_suppl):1023-32.
5. Pearson ES. The analysis of variance in cases of non-normal variation. Biometrika 1931;23:114-33.
6. Glass G, Peckham P, Sande J. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev Educ Res 1972;42:237-88.
7. Wilcox RR. ANOVA: a paradigm for low power and misleading measures of effect size? Rev Educ Res 1995;65:51-77.
8. Buning H. Robust analysis of variance. J Appl Stat 1997;24:319-32.
9. R Development Core Team. R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria:. [cited 2018] Available from http://www.Rproject.org/
10. Peterson K. Six modifications of the aligned rank transform test for interaction. J Modern Appl Stat Methods 2002;1:100-9.
11. Kruskal WH, Wallis A. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583-621.
12. Fisher RA. The Design of Experiments. Edinburgh: Oliver and Boyd; 1935.
13. Hecke TV. Power Study of Anova versus Kruskal-Wallis Test, 2010.
14. Odiase JI, Ogbonmwan SM. JMASM20: exact permutation critical values for the Kruskal-Wallis One-way ANOVA. J Modern Appl Stat Methods 2005;4:609-20.
15. Brown GW, Mood AM. On Median Tests for Linear Hypotheses. University of California Press, 1951: pp. 159-66.
16. Conover WJ. Practical Nonparameteric Statistics. 3rd ed. Wiley; 1999: p. 396-406.
17. van der Waerden B. Order Tests for The Two-Sample Problem II, III, Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen. Serie A 1953;564:303-10 and 311-6.
18. Hajek J. A Course in Nonparametric Statistics. San Francisco: Holden-Day, 1969: p.83.
19. Blanca M, Alarcón R, Arnau J, Bono R, Bendayan R. Non-normal data: Is ANOVA still a valid option? Psicothema 2017;29:552-7.
20. Clinch J, Kesselman H. Parametric alternatives to the analysis of variance. J Educ Stat 1982;7:207-14.
21. Gamage J, Weerahandi S. Size performance of some tests in one-way ANOVA. Commun Stat Simul Comput 1998;27:625-40.
22. Lantz B. The impact of sample non-normality on ANOVA and alternative methods. Br J Math Stat Psychol 2013;66:224-44.
23. Schmider E, Ziegler M, Danay E, Beyer L, Bühner M. Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. Methodology (Gott) 2010;6:147-51.
24. Brown MB, Forsythe AB. The small sample behavior of some statistics which test the equality of several means. Technometrics 1974;16:129-32.
25. De Beuckelaer A. A closer examination on some parametric alternatives to the ANOVA F-test. Stat Papers 1996;37:291-305.
26. Lee S, Ahn C. Modified ANOVA for unequal variances. Commun Stat Simul Comput 2003;32:987-1004.
27. Li X, Wang J, Liang H. Comparison of several means: a fiducial based approach. Comput Stat Data Analysis 2011;55;1993-2002.
28. Lu F, Mathew T. A parametric bootstrap approach for ANOVA with unequal variances: fixed and random models. Comput Stat Data Analysis, 2007;51:5731-42.
29. Markowski CA. Conditions for the effectiveness of a preliminary test of variance. Am Stat 1990;44:322-6.
30. Keselman HJ, Rogan JC, Fier-Walsh BJ. An evaluation of some non-parametric and parametric tests for location equality. Br J Math Stat Psychol 1977;30:213-21.
31. Tomarken A, Serlin RC. Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychol Bull 1986;99:90-9.
32. Hoeffding W. Optimum" nonparametric tests. Berkeley Symposium on Mathematical Statistics and Probability. Universy of California 2nd ed. 1951: pp.83-92.
33. Terry MH. Some rank order test which are most powerful aganist specific parametric alternatives. Ann Math Stat 1952;23:346-66.
34. Luh W, Guo J. Approximate transformation trimmed mean methods to the test of simple linear regression slope equality. J Appl Stat 2000;27:843-57.
35. Jett D, Speer J. Comparison of parametric and nonparametric tests for differences in distribution. Proceedings of The National Conference On Undergraduate Research (NCUR) 2016 University of North Carolina-Asheville Asheville, North Carolina April 7-9, 2016: 1765-70.