DOI: 10.18621/eurj.1030038 *Biostatistics*

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

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

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

ABSTRACT

Objectives: The performances of the Welch test, the Alexander-Govern test, the Brown-Forsythe test and the James Second-Order test, which are among the 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 and heterogeneous, the sample sizes are balanced and unbalanced, the distribution of the data is in accordance with the normal distribution and 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 Welch test, the Alexander-Govern test and the James Second-Order test were not affected by the distribution and performed well in situations where variances were heterogeneous. The Brown-Forsythe test was not affected by the distribution, it performed well when the variance was homogeneous and the sample size in the groups to be compared was not equal.

Conclusions: The Welch test, the Alexander-Govern test and the James Second-Order test are the tests that can be recommended as an alternative to the F test.

Keywords: Analysis of variance, conformity of normal distribution, 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, 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].

Pearson [3], Glass *et al*. [4], and Wilcox [5] exam-

Received: ***November 29, 2021;*** ***Accepted:*** ***August 8, 2022;*** ***Published Online:*** ***August 30, 2022***

e-ISSN: 2149-3189

How to cite this article: *Ocakoğlu G, Macunluoğlu AC. Comparison of the performances of parametric k-sample test procedures as an alternative to one-way analysis of variance. Eur Res J 2023;9(1):239-48. DOI: 10.18621/eurj.1030038*

Address for correspondence: *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*

ined the effect of the normality assumption violation on the Type-I error. Wilcox [5] concluded that samples that do not conform to normal distribution have some impact on the Type-I error rate, but the effect is minimal if the variances are homogeneous. Glass *et al*. [4] reported similar results to Wilcox [5] in their studies if the variances were homogeneous. In his study, Buning [6] 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 Welch test, the Alexander-Govern test, the Brown-Forsythe test, the James Second-Order test, which are among the parametric alternatives of the F test, to protect Type-I error under various simulation scenarios.

Table 1. Sample sizes of the groups

	Balanced Sample	Non-balanced Sample		
Number of groups		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

METHODS

In our study, the Welch test, the Alexander-Govern test, the Brown-Forsythe test, the James Second-Order test in terms of maintaining the probability of Type-I error determined at the beginning of the experiment were compared with the F test. Simulation scenarios run under the R program [7].

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 units 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 H0 hypotheses were determined, which were rejected at the end of 50000 repetitions. In our study, the evaluation criterion proposed by Peterson [8] 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 μi (i= 1, 2, . . ., k) is different. The F test statistic,

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

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

The Welch Test

The Welch test is a robust test against the violation of the assumption of variance homogeneity, which is considered as an alternative to the F test [9]. 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 formula for the Welch test is

$$F_W=\frac{\sum_{i=1}^{k} w_i(\bar{x}_i-\widehat{\mu})^2/(k-1)}{1+[2(k-2)/k^2-1]\sum h_i} \quad (2)$$

where

$$(w_i)=n_i/s_i^2;\ \hat{\mu}=\sum_{i=1}^{k} w_i x_i/W;\ W=\sum_{i=1}^{k} w_i;\ \ h_i=(1-w_i/W)^2/(n_i-1)$$

The null hypothesis should then be rejected at the α level of significance when $F_W > F_{\alpha;k-1,\ f.}$

The Alexander Govern Test

The Alexander-Govern (AG) test is another alternative to the F test developed by Alexander and Govern [10]. This test is used when the sample sizes in the groups are not equal. It is a parametric test that can be used instead of the F test if the data conform to normal distribution. To calculate the test statistic, the t statistic is first calculated for each group.

$$t_i=\frac{\bar{X}_i-X^+}{S_{\bar{X}_i}} \quad (3)$$

where

$$S_{\bar{X}_i}=\sqrt{\frac{\sum_{j=1}^{n_i}\left(X_j-\bar{X}_i\right)^2}{n_i(n_i-1)}};\ W_i=\frac{1/S_i^2}{\sum_{i=1}^{k}(1/S_i^2)};\ X^+=\sum_{i=1}^{k} W_i\bar{X}_i.$$

Calculated t values are converted to the standard normal distribution Z using the normalization ap-

Table 2. Variance rates of groups (k)

	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	0.1:0.1:0.1:0.1:0.1:0.1:0.1:0.2 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 0.1:0.1:0.1:0.3:0.3:0.3:0.5:0.5 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

proach [11].

$$Z_i = c + \frac{(c^3+3c)}{b} - \frac{(4c^7+33c^5+240c^3+855)}{(10b^2+8bc^4+1000b)} \tag{4}$$

where

$$a_i = v_i - 0.5,\ b = 48a^2,\ c = \left[a * ln\left(1 + \frac{t_i^2}{v_i}\right)\right]^{\frac{1}{2}} \text{ and } v_i = n_i - 1.$$

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 test statistic is calculated as [10],

$$AG = \sum_{i=1}^{k} Z_i^2 \tag{5}$$

The null hypothesis should then be rejected at the α level of significance when AG > X_{k-1}^2.

The James Second-Order Test

The James Second-Order (JSO) test, developed by James [12] as an alternative to the F test, is a robust test against violating the assumption of variance homogeneity. To calculate the test statistic, the t statistic is first calculated for each group.

$$t_i = \frac{\bar{X}_i - \bar{Y}}{S_i^2} \tag{6}$$

where

$$\bar{Y} = \sum_{i=1}^{k} a_i \bar{X}_i,\ a_i = \frac{1/S_i^2}{\sum_{i=1}^{k}(1/S_i^2)}.$$

The null hypothesis H_0: $\mu_1=\mu_2=\ldots=\mu_k$ versus alternative H1: at least one μ_i (i= 1, 2, . . ., k) is different. The test statistic (J) is calculated as [9],

$$J = \sum_{i=1}^{k} t_i^2 \tag{7}$$

The null hypothesis should then be rejected at the α level of significance when J > CV_α. The test statistic, J, is compared to a critical value, CV_α, where

$$CV_\alpha = C + \frac{1}{2}(3\chi_4+\chi_2)T + \frac{1}{16}(3\chi_4+\chi_2)^2\left(1-\frac{J-3}{c}\right)T^2 + \frac{1}{2}(3\chi_4+\chi_2)\Big[(8R_{23} - 10R_{22} + 4R_{21} - 6R_{12}^2 + 8R_{12}R_{11} - 4R_{11}^2) + (2R_{23} - 4R_{22} + 2R_{21} - 2R_{12}^2 + 4R_{12}R_{11} - 2R_{11}^2)(\chi_2 - 1) + \frac{1}{4}(-R_{12}^2 + 4R_{12}R_{11} - 2R_{12}R_{10} - 4R_{11}^2 + 4R_{11}R_{10} - R_{10}^2)(3\chi_4 - 2\chi_2 - 1)\Big] + (R_{23} - 3R_{22} + 3R_{21} - R_{20})(5\chi_6 + 2\chi_4 + \chi_2) + \frac{3}{16}(R_{12}^2 - 4R_{23} + 6R_{22} - 4R_{21} + R_{20})(35\chi_8 + 15\chi_6 + 9\chi_4 + 5\chi_2) + \frac{1}{16}(-R2_{22} + 4R_{21} - R_{20} + 2R_{12}R_{10} - 4R_{11}R_{10} - 4R_{11}R_{10} + R_{10}^2) \times (9\chi_8 - 3\chi_6 - 5\chi_4 - \chi_2) + \frac{1}{4}(-R_{22} + R_{11}^2)(27\chi_8 + 3\chi_6 + \chi_4 + \chi_2) + \frac{1}{4}(R_{23} - R_{12}R_{11})(45\chi_8 + 9\chi_6 + 7\chi_4 + 3\chi_2) \tag{8}$$

with C denoting the 1 − α quantile of a X_{k-1}^2 distribution and with

$$T = \sum_{i=1}^{k} \frac{(1-a_i)^2}{n_i - 1},\ \chi_{2s} = \frac{\chi_{k-1,\alpha}^{2s}}{[(k-1)(k+1)\ldots(k+2s-3)]},\ R_{st} = \sum_{i=1}^{k} \frac{a_i^t}{(n_i-1)^2}.$$

The JSO test was accepted as the best option for both data with normal distribution, heterogeneous variance [10], and situations that do show the non-normal distribution and heterogeneous variance [13]. The disadvantage of this method is the complexity of the computation of critical values [14].

The Brown-Forsythe Test

One of the parametric alternatives to the F test is the Brown-Forsythe test. It is a robust test if the sample size is small, the population heterogeneous variance, and the normality assumption is provided. The null hypothesis H_0: $\mu_1=\mu_2=\ldots=\mu_k$ versus alternative H1: at least one μ_i (i= 1, 2, . . ., k) is different. The test statistic is calculated as [15],

$$BF = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{\sum_{i=1}^{k}(1 - n_i/N)(S_i)^2} \tag{9}$$

The null hypothesis, H_0: $\mu_1=\mu_2=\ldots=\mu_k$, should then be rejected at the α level of significance when F > $F_{\alpha;k\text{-}1,\ f.}$

BF statistic has an approximately F distribution with *k-1* and *f* degrees of freedom, where *f* is obtained with

$$f = \left(\sum_{i=1}^{k} \frac{C_i^2}{n_i - 1}\right)^{-1} \tag{10}$$

C_i used in calculating degrees of freedom f is calculated with the Satterthwaite [16] approach.

$$C_i = \frac{\left(1-\frac{n_i}{N}\right){S_i}^2}{\left[\sum_{i=1}^{k}\left(1-\frac{n_i}{N}\right){S_i}^2\right]} \qquad (11)$$

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 [7].

Comparisons in which sample size is balanced, the group variances are homogeneous, and the data follow to the normal distribution

Considering all simulation scenarios given in Table 3, it was observed that the F test and the JSO test were able to maintain the Type-I error level (α = 0.05) determined at the beginning. When *Supplementary* Table 1 is examined, it has been observed that the AG test can maintain the Type-I error level initially determined. The F test is the test that shows the most successful performance in estimating the Type-I error level determined at the beginning, according to the alternative parametric tests (*Supplementary* Table 2).

Comparisons in which the sample size is not balanced, the group variances are homogeneous, and the data follow to the normal distribution

The F test and the BF test are the tests that show the most successful performance in estimating the Type-I error level determined at the beginning. The F test tended to maintain the Type-I error in all simulation scenarios given in the tables. The BF test estimated the Type-I error level as deviant only in a simulation scenario (*Supplementary* Table 3, *Supplementary* Table 4, and *upplementary* Table 5). The F test was also not affected by excessive differences of sample size in groups and tended to maintain the Type-I error level initially determined in all simulation scenarios according to the Peterson criterion (*Supplementary* Table 6, *Supplementary* Table 7, and *Supplementary* Table 8).

Comparisons in which the sample size is balanced, group variances are heterogeneous, but the data follow to the normal distribution

As expected, when the simulation scenarios were examined according to the Peterson criterion, the F test was highly affected by distortion in-group variance and failed to maintain the Type-I error at a nominal level and gave highly deviant results. When the simulation scenarios given in *Supplementary* Table 9, *Supplementary* Table 10, and *Supplementary* Table 11 are examined, it is seen that the AG test is the best alternative to the F test. Among the other tests included in the study, the alternatives of the F test after the AG test in this trial can be seen as the Welch test and the JSO test.

Comparisons in which the sample size is not balanced, group variances are heterogeneous, but the data follow to the normal distribution

When the combinations of observations in which the sample size in the groups are not equal are examined (*Supplementary* Table 12, *Supplementary* Table 13, and *Supplementary* Table 14), it is seen that the AG test and Welch tests are the best alternative of the F test respectively. Although the performance of the JSO test is negatively affected by the increase in the number of groups compared, it can be seen as an alternative test after the AG test and the Welch test. When the simulation scenarios (*Supplementary* Table 15, *Supplementary* Table 16, and *Supplementary* Table 17) are examined, 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. When the simulation scenarios in which the assumption of homogeneity of variances were not met, a lower number of observations was assigned to the group with high variance, and a higher number of observations was assigned to the group with a low variance (*Supplementary* Table 18, *Supplementary* Table 19, and *Supplementary* Table 20), it was seen that the Welch, the AG test and the JSO test were alternatives to the F test.

Comparisons in which the sample size is balanced, group variances are homogeneous, and the data follow to log-normal distribution

As expected, the F test is the test that shows the most successful performance to estimate the level of Type-I error determined at the beginning when considering the parametric alternatives available. The JSO test tends to preserve the Type-I error in all simulation scenarios in three group comparisons. The AG test

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

σ^2	n	F	Welch	AG	JSO	BF
1	3	4.70%	3.44%	3.91%	4.99%	3.10%
	5	4.98%	4.47%	4.39%	4.94%	4.19%
	10	5.03%	4.92%	4.79%	5.08%	4.83%
	15	5.09%	5.11%	4.99%	5.20%	5.01%
	20	5.07%	5.04%	4.95%	5.11%	5.01%
	25	5.08%	5.09%	4.99%	5.12%	5.04%
	30	4.93%	4.90%	4.83%	4.92%	4.91%
	50	5.11%	4.94%	4.90%	4.95%	5.09%
	80	5.07%	5.06%	5.03%	5.07%	5.06%
	100	5.04%	5.06%	5.05%	5.06%	5.04%
2	3	4.82%	3.64%	3.65%	4.85%	2.92%
	5	4.93%	4.51%	4.45%	4.99%	4.12%
	10	4.88%	4.80%	4.66%	4.96%	4.68%
	15	5.03%	5.07%	4.97%	5.16%	4.95%
	20	4.80%	4.79%	4.69%	4.85%	4.74%
	25	5.15%	5.17%	5.09%	5.19%	5.12%
	30	5.03%	5.15%	5.04%	5.16%	5.01%
	50	4.93%	4.93%	4.89%	4.94%	4.93%
	80	4.99%	5.01%	4.99%	5.02%	4.99%
	100	5.00%	5.02%	5.00%	5.03%	4.99%
4	3	5.08%	3.90%	3.75%	4.84%	2.92%
	5	4.96%	4.36%	4.40%	4.93%	4.17%
	10	5.06%	4.99%	4.79%	5.07%	4.83%
	15	4.77%	4.70%	4.85%	5.02%	4.91%
	20	5.10%	5.13%	4.95%	5.11%	5.01%
	25	4.95%	4.91%	4.94%	5.07%	5.04%
	30	5.21%	5.20%	4.69%	4.78%	4.74%
	50	5.10%	5.06%	4.80%	4.86%	4.90%
	80	4.89%	4.88%	4.94%	4.97%	5.02%
	100	4.86%	4.83%	4.87%	4.92%	4.88%
8	3	4.97%	3.71%	3.78%	4.84%	3.00%
	5	4.78%	4.33%	4.32%	4.77%	3.98%
	10	5.07%	4.98%	4.85%	5.17%	4.87%
	15	5.15%	5.12%	4.99%	5.22%	5.06%
	20	4.90%	4.81%	4.73%	4.86%	4.86%
	25	5.01%	4.94%	4.87%	4.98%	4.98%
	30	5.00%	4.94%	4.89%	4.98%	4.97%
	50	5.08%	5.06%	5.02%	5.07%	5.07%
	80	5.21%	5.27%	5.25%	5.27%	5.20%
	100	5.09%	5.09%	5.07%	5.09%	5.09%
10	3	5.10%	3.80%	3.81%	5.03%	3.17%
	5	4.99%	4.49%	4.48%	4.92%	4.09%
	10	4.88%	4.73%	4.61%	4.88%	4.67%
	15	5.01%	4.88%	4.76%	4.99%	4.89%
	20	5.02%	4.92%	4.81%	4.99%	4.97%
	25	5.08%	5.00%	4.92%	5.05%	5.05%
	30	4.90%	4.90%	4.85%	4.93%	4.88%
	50	5.06%	5.00%	4.90%	4.96%	5.01%
	80	5.08%	5.09%	4.97%	5.02%	5.05%
	100	5.09%	5.12%	5.10%	5.12%	5.09%

tends to preserve the Type-I error in all simulation scenarios in five group comparisons (*Supplementary* Table 21, *Supplementary* Table 22, and *Supplementary* Table 23).

Comparisons in which the sample size is not balanced, group variances are homogeneous, and the data follow to log-normal distribution

When the simulation scenarios are examined, the F test and the BF test are the tests that show the most successful performance in estimating the Type-I error level determined at the beginning (*Supplementary* Table 24, *Supplementary* Table 25, and *Supplementary* Table 26). The F test tended to maintain the Type-I error in all simulation scenarios given in the tables. The BF test has given biased estimates in only two simulation scenarios in three group comparisons, in only one simulation scenario in five group comparisons. It tends to preserve Type-I error in all simulation scenarios for eight groups. It was observed that the other tests included in the study were negatively affected by the imbalance of the number of units in the groups, and their performance in maintaining the Type-I error level determined at the beginning was not considered sufficient. When the simulation scenarios involving observation combinations in which the sample size in the groups differ excessively (*Supplementary* Table 27, *Supplementary* Table 28, and *Supplementary* Table 29), it was observed that the Welch test, the AG test, the BF test, the JSO test were affected by the extreme differences in the sample size.

Comparisons in which the sample size is balanced, group variances are heterogeneous, and the data follow to log-normal distribution

The F test was highly affected by the deterioration of group variances and failed to maintain the Type-I error at the nominal level. Considering the performances determined according to Peterson criteria, it was seen that the AG test is the best alternative of the F test. Among the other tests included in the study, the alternatives of the F test in these simulation scenarios after the AG test can be accepted as the JSO test and the Welch test (*Supplementary* Table 30, *Supplementary* Table 31, and *Supplementary* Table 32).

Comparisons in which the sample size is not balanced, group variances are heterogeneous, and the data follow to log-normal distribution

When the simulation scenarios (*Supplementary* Table 33, *Supplementary* Table 34, and *Supplementary* Table 35) are examined, as expected, the F test was highly affected by the deterioration in group variances and failed to protect the Type-I error at the nominal level. Among the other tests included in the study, the alternatives of the F test in these simulation scenarios after the AG test can be accepted as the JSO test and the Welch test. When the simulation scenarios (*Supplementary* Table 36, *Supplementary* Table 37, and *Supplementary* Table 38) are examined, 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. When the simulation scenarios (*Supplementary* Table 39, *Supplementary* Table 40, and *Supplementary* Table 41) are examined, the alternatives of the F test in these simulation scenarios after the AG test can be accepted as the JSO test and the Welch test respectively.

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 [6] 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.* [17], Clinch and Keselman [18], Gamage and Weerahandi [19], Lantz [20] and Schmider *et al.* [21] 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 [22], Blanca *et al.* [17], Brown and Forsythe [23], Buning [6], Debeuckelaer [24], Lee and Ahn [25], Li *et al*. [26], Lu and Mathew [27], Markowski [28], Keselman *et al*. [29], Tomarken and Serlin [30] concluded that the F test is highly affected by the deterioration in group variances.

In this study, the Welch test, the AG test and the JSO test were not affected by the distribution of the data, and in cases where the variances were not homogeneous, they tend to protect the Type-I error. Penfield [31], Lix *et al*. [32] and Hartung *et al*. [33] found that the Welch test is not affected by the distribution of data and performs better in simulation scenarios where variances are heterogeneous. Bishop and Dudewicz [22], Brown and Forsythe [23], Buning [6], DeBeuckelaer [24], Keselman *et al*. [29], Markowski [28], Rafinetti [34], Tomarken and Serlin [30], Wilcox *et al*. [35] similar results in their work; They found that the Welch test performed better in cases where both assumptions were not provided. In their studies, Alexander and Govern [10], Myers [36], Oshima and Algina [13] concluded that the performance of the AG test was sufficient in terms of protecting Type-I error in cases where the data conformed to normal distribution, but the variance was not homogeneous. They stated that in cases where the assumption of conformity to normal distribution and homogeneity of variances is not realized, the sample size should be considered in order to use the AG test. Alexander and Govern [10] and Myers [36] stated that the JSO test is a good alternative to the F test when the distribution of the data is symmetrical and the assumption of homogeneity of variances is not met. Oshima and Algina [13] and Wilcox [37] found that the JSO test performed better in cases where both assumptions were not provided.

It has been concluded that the BF test shows an adequate performance in cases where the data show normal and log-normal distribution, the assumption of homogeneity of the variances is met, and the sample size in the groups to be compared are not equal. DeBeuckelaer [24], found that the BF test gives better results than the F test when one or both of the assumptions of normality and variance homogeneity cannot be achieved. Gamage and Weerahandi [19], Roth [38], Steel *et al*. [39], when group variances were not homogeneous, Wilcox *et al*. [35], stated that in cases where groups with large variances have small sample sizes, Oshima and Algina [13] stated that in cases where the homogeneity and normality assumption of variances cannot be achieved, the BF test can be used to make comparisons between groups.

CONCLUSION

As a result 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. The Welch, the AG test and the JSO test are tests that can be recommended as an alternative to the F test because they are less affected by the sample size in the groups, the distribution of the data or the number of groups to be compared, if the homogeneity of the variances is neglected.

Authors' Contribution

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

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 2017;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. Pearson ES. The analysis of variance in cases of non-normal variation. Biometrika 1931;23:114-33.
4. Glass GV, Peckham PD, Sanders JR. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev Educ Res 1972;42:237-88.
5. Wilcox RR. ANOVA: a paradigm for low power and misleading measures of effect size? Rev Educ Res 1995;65:51-77.

6. Buning H. Robust analysis of variance. J Appl Stat 1997;24:319-32.

7. 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/

8. Peterson K. Six modifications of the aligned rank transform test for interaction. J Modern Appl Stat Methods 2002;1:100-9.

9. Cribbie RA, Fiksenbaum L, Keselman HJ, Wilcox RR. Effect of non-normality on test statistics for one-way independent groups designs. Br J Math Stat Psychol 2012;65:56-73.

10. Alexander R, Govern D. A new and simpler approximation for anova under variance heterogeneity. J Educ Stat 1994;19:91-101.

11. Hill G. Algorithm 395. Student's t-distribution. Commun ACM 1970;13:617-9.

12. James GS. The comparison of several groups of observations when the ratios of the population variances are unknown. Biometrika 1951;38:324-9.

13. Oshima T, Algina J. Type-I error rates for James's second-order test and Wilcox's Hm test under heteroscedasticity and non-normality. Br J Math Stat Psychol 1992;45:255-63.

14. Dag O, Dolgun A, Konar N. One-way tests: an R package for one-way tests in independent groups designs. R J 2018;10:175-99.

15. Brown M, Forsythe A. The small sample behavior of some statistics which test the equality of several means. Technometrics 1994:16:129-32.

16. Satterhwaite FE. Synthesis of variance. Psychometrika 1941;6:309-16.

17. Blanca M, Alarcón R, Arnau J, Bono R, Bendayan R. Non-normal data: Is ANOVA still a valid option? Psicothema 2017;29:552-7.

18. Clinch J, Kesselman H. Parametric alternatives to the analysis of variance. J Educ Behav Stat 1982;7:207-14.

19. Gamage J, Weerahandi S. Size performance of some tests in one-way ANOVA. Commun Stat Simul Comput 1998;27:625-40.

20. Lantz B. The impact of sample non-normality on ANOVA and alternative methods. Br J Math Stat Psychol 2013;66:224-44.

21. Schmider E, Ziegler M, Danay E, et al. Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. Methodology 2010;6:147-51.

22. Bishop TA, Dudewicz EJ. Exact analysis of variance with unequal variances: test procedures and tables. Technometrics 1978;20:419-30.

23. Brown MB, Forsythe AB. The small sample behavior of some statistics which test the equality of several means. Technometrics 1974;16:129-32.

24. De Beuckelaer A. A closer examination on some parametric alternatives to the ANOVA F-test. Stat Pap (Berl) 1996;37:291-305.

25. Lee S, Ahn C. Modified ANOVA for unequal variances. Commun Stat Simul Comput 2003;32:987-1004.

26. Li X, Wang J, Liang H. Comparison of several means: a fiducial based approach. Comput Stat Data Anal 2011;55:1993-2002.

27. Lu F, Mathew T. A parametric bootstrap approach for ANOVA with unequal variances: fixed and random models. Comput Stat Data Anal 2007;51:5731-42.

28. Markowski CA. Conditions for the effectiveness of a preliminary test of variance. Am Stat 1990;44:322-6.

29. 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.

30. Tomarken A, Serlin RC. Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychol Bull 1986;99:90-9.

31. Penfield D. Choosing a two- sample location test. J Exp Educ 1994;62:343-60.

32. Lix L, Keselman J, Keselman H. Concequences of assumption violations revisited: a quantitative review of alternatives to the one-way analysis of variance F test. Rev Educ Res 1996;66:579-619.

33. Hartung J, Argaç D, Makambi K. Small sample properties of tests on homogeneity in one-way anova and meta-analysis. Stat Pap (Berl) 2002;43:197-235.

34. Rafinetti R. Demonstrating the concequences of violations of assumptions in between-subjects analysis of variance. Teach Psychol 1996;23:51-4.

35. Wilcox RR, Charlin V, Thompson KL. NewMonte Carlo results on the robustness of the ANOVA F, W, and F statistics. Commun Stat Simul Comput 1986;15:933-44.

36. Myers L. Comparability of the James's second-order aproximantion tets and the Alexander and Govern a statistic for non-normal heterosdecatic data. J Stat Comput Simul 1998;60:207-23.

37. Wilcox RR. A new alternative to the ANOVA F and new results on James's second-order method. Br J Math Stat Psychol 2011;41:109-17.

38. Roth AJ. Robust trend tests derived and simulated: analogs of the Welch and Brown-Forsythe tests. J Am Stat Assoc 1983;78:972-80.

39. Steel R, Torrie J, Dickey D. Principles and procedures of Statistics: A Biometrical Approach. 3rd ed. New York, NY: McGraw-Hill; 1997.