Advances in Methodology and Statistics / Metodološki zvezki, Vol. 18, No. 2, 2021, 53–71
https://doi.org/10.51936/ltgt2135

Power comparison of ANOVA and Kruskal–Wallis tests when error assumptions are violated

Felix N. Nwobi*, Felix C. Akanno

Imo State University, Department of Statistics, Owerri, Nigeria

Abstract

The effects of the violations of normality and homogeneity of variances assumptions on the power of the one-way ANOVA F-test is studied in this paper. Simulation experiments were conducted to compare the power of the parametric F-test with the non-parametric Kruskal–Wallis (KW) test in normal/non-normal, equal/unequal variances scenarios and equal/unequal sample group means. Each of these 184 simulation experiments was replicated $N = 1000$ times and power obtained for both F and KW−tests. The Shapiro–Wilk's test for normality and Bartlett's/Levene's tests for homogeneity of variances was conducted in each experiment. Results show that the power of the KW tests outperformed those of the F-tests in the 92 (85/92) non-normal cases. Although the power of the F-tests is higher than those of the KW tests in 85 out of the 92 experiments under normality assumptions, these differences, in all cases in this study are not significant ($p > 0.05$) using both t and sign tests. Based on these results, this study favours the KW test as a more robust test and safer to use rather than the F-test especially when the distributional assumptions of data sets are in doubt.

Keywords: ANOVA, Homogeneity test, Kruskal–Wallis test, Normality test, Power comparison

1. Introduction

The one-way analysis of variance (ANOVA) is one of the most popular statistical methods for the comparison of treatment means from completely randomized designed (CRD) experiments since its introduction into the statistics literature from the 1920s by R. A. Fisher. As a classical statistical method, two of the major requirements of the method to produce optimal results are: (i) data set be normally distributed, and (ii) group variances be homogeneous. In real-life situations, data sets are not often normally distributed and group variances unequal making these assumptions always unattainable. However, many applied researchers such as those in the fields of business, economics, the social sciences, etc., go ahead to apply the method because they are either not aware of these restrictions or ignorant of the seriousness of their violations.

*Corresponding author

Email addresses: fnnwobi@imsu.edu.ng (Felix N. Nwobi), felixtemple@yahoo.com (Felix C. Akanno)

To circumvent the effects of these assumptions, Kruskal and Wallis (1952) introduced a non-parametric version of the analysis of means by ranks as an alternative to ANOVA's F-test. The remarkable thing about Kruskal–Wallis (KW) test is that the assumptions are milder than those of the parametric F-test. Effects of violations of homogeneity of variances, non-normality of group means or both on Type I error rate are available (see, for example, Kutner et al., 2005; Legendre & Borcard, 2008; Marcinko, 2014; Moder, 2007, 2010). A good review of the effect of non-normality on the robustness of the F-test can be found in Blanca Mena et al. (2017), especially when degrees of skewness and kurtosis ranging from -1 to 1 are considered.

To compare these methods, Hecke (2012) whose simulation experiments were based on permutation to determine the power of both tests, observed a higher power in KW as compared to the classical F-test in the case of non-symmetrical distributions. Lachenbruch and Clements (1991) had demonstrated that the KW test may have greater power than the F-test when the population distributions are not normal. They further argued that in comparison with F-test, the KW test is more robust against the departures from assumptions of equality of variance. The research carried out by Glass et al. (1972) focused on the powers of the F-test and KW test when the population of interest is skewed. They observed that non-normality has some effect on the Type I error, but the minimal effect when the variances are equal. For a completely randomized fixed-effect model of data with binomial errors, the F-test behaved in general, better than the KW test, controlling the nominal level of significance and presenting higher power (Ferreira et al., 2012).

From the foregoing, opinions of researchers are divided on the robustness of the classical F-test in the analysis of data sets from completely randomized experiments. In this study, we use Monte Carlo simulations to investigate the effects of violations of these assumptions on the power of the F-test and KW test under various scenarios, e.g., unequal means and sample sizes.

2. Two competing tests

2.1. The ANOVA F-test

The ANOVA test is a powerful statistical tool for tests of equality of a group means. By using Fisher notation, a one-way ANOVA model may be represented mathematically as

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, Y_{ij} is the yield from the j-th observation at i-th treatment, μ is the general mean effect given by

$$\mu = \sum_{i=1}^{k} \mu_i \alpha_i / n$$

and α_i is the fixed/random effect due to the i-th treatment. This means that if there were no treatment differences and no chance causes, then the yield of each observation will be μ. The α_i which is the effect of the i-th treatment is given by

$$\alpha_i = \mu_i - \mu.$$

Therefore the i-th treatment increases or decreases the yield by an amount α_i. The two basic assumptions of this model are (i) that the data set is normally distributed, $Y \sim N(\mu, \sigma^2)$ and $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and (ii) group variances are equal. The test hypothesis is therefore stated as

$$H_0 \colon \mu_1 = \mu_2 = \cdots = \mu_k,$$

for

$$F = \frac{MS_{tr}}{MS_e} \tag{2.1}$$

where F is the test statistic, MS_{tr} and MS_e are treatment mean squares and error mean squares respectively. Based on Equation (2.1), H_0 is rejected for a given α if $F > F_{k-1,k(n-1);\alpha}$.

2.2. The Kruskal–Wallis test

The KW test as a non-parametric alternative to the one-way ANOVA assumes that observations in each group come from a population with the same shape of the distribution. It becomes a problem when truly the observations are not coming from the population with the same shape. The null hypothesis associated with this test is given by

$$H_0 \colon \eta_1 = \eta_2 = \cdots = \eta_k$$

where η_i is the median of the i-th group. This is equivalent to H_0: *Samples are from identical populations*. Let n denote the total number of observations $n = \sum_{i=1}^{k} n_i$ where n_i is the size of the i-th sample, $i = 1,2,\ldots,k$ and k is the number of groups. Rank the n observations in either ascending or descending order of magnitude and use average ranks when there are ties. Let $R(X_{ij})$ represent the rank assigned to the j-th observation from the i-th group, X_{ij} and R_i represent the sum of the ranks assigned to the i-th group $R_i = \sum_{i=1}^{n_i} R(X_{ij}), i = 1,2,\ldots,k$. Define the test statistic T as

$$T = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{{R_i}^2}{n_i} - \frac{n(n+2)^2}{4}\right)$$

where

$$S^2 = \frac{1}{n-1}\left(\sum_{allrank} R(X_{ij})^2 - n\frac{(n-1)^2}{4}\right).$$

If there are no ties, S^2 simplifies to $n(n+1)/12$ and the test statistic reduces to

$$T = \frac{12}{n(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1).$$

Under the null hypothesis, H_0 (Lehmann, 2006), T is asymptotically chi-square distributed with $k-1$ degrees of freedom, i.e., $T \sim \chi^2_{k-1}$.

3. Methodology

3.1. The power of a test

In testing the equality of group means of a data set by whatever method, the researcher will be interested in the correctness or otherwise of the outcome of the test hypothesis. The outcome is interpreted using a p-value, which is the probability of observing the result with a specified level of significance, α, given that H_0 is true. When an experimenter erroneously accepts a H_0 when H_1 is true, then that experimenter has committed a Type II error stated as P(Accept $H_0|H_1$ is true) = β. Then define a statistical power of a test as the probability that the test correctly rejects H_0 when a specified alternative is true, i.e., Power = P(Reject $H_0|H_1$ is true) = $1-\beta$.

Power is influenced mainly by the chosen significant level of the test and the sample size. Since power is stated in terms of probability, its value is within the range $0 \leq P \geq 1$, therefore with two methods testing a given dataset, the method with higher P will be considered a better method. Therefore, a method with a lower power has a higher risk of committing Type II errors (of accepting a null hypothesis when indeed the alternative is true).

3.2. Tests on assumptions

Tests on assumptions will be conducted on each simulation experiment. For the normality assumption, the Shapiro–Wilks W test will be implemented. The hypothesis of interest is H_0: *Data are normally distributed*, versus the alternative, H_1: *Data are not normally distributed*.

For the homogeneity of group variances assumption, the Bartlett's K^2 test will be applied when data are assumed normal or near-normal while the Levene's L test will be used when data set is either skewed or for non-normal data. The Levene's test, unlike the Bartlett's, is known to be less sensitive to departures from normality. In these two cases, the hypothesis is given by

$$H_0 \colon \sigma_i^2 = \sigma_j^2$$

$$H_0 \colon \sigma_i^2 \neq \sigma_j^2.$$

For these three tests in this subsection, H_0 will be rejected if $p < \alpha$; if $p > \alpha$, then the H_0 cannot be rejected and conclude that the group variances are equal. The reader is referred to Sahai and Ageel (2000, pp. 93–107) for Shapiro–Wilk's, Bartlett's and Levene's tests respectively.

3.3. Tests for differences

3.3.1. The t-test

Each data set generated will be analyzed using two methods, the F-test and the KW test each of them independently reporting the power of the test. To compare the powers from these methods, a paired t-test is considered to determine whether the difference between their powers is significantly different from zero.

Let A_i and B_i, $i = 1, 2, \ldots, n$ denote the power of the F-test and the KW test on the i-th experiment respectively. Further, let $d_i = A_i - B_i$, be assumed to be identically distributed, all with the same expected population mean values μ_d and variances σ_d^2. The hypotheses for this test are given as follows

$$H_0 \colon \mu_d = 0 \quad \text{vs.} \quad H_1 \colon \mu_d \neq 0.$$

The t-statistic for this test is

$$T_n = \frac{\bar{d}}{s/\sqrt{n}} \sim t_{(n-1;\alpha)}$$

where $\bar{d}$ and s are the mean and standard deviation respectively of the d_i. Rejection of this null hypothesis at α level of significance will lead to the conclusion that the power of the F-test is significantly different from the power of the KW test. T-tests will be performed for all the experiments.

3.3.2. The Sign test S

A nonparametric sign test will then be used to determine if the number of F-test with positive power differences is significantly greater than or equal to the number of KW tests with negative power differences.

Taking S_+ to be the number of positive differences (+) in favour of F-tests out of m pairs, then the null hypothesis of interest is

$$H_0 \colon P(S_+) = P(S_-) = 0.5$$

$$H_1 \colon P(S_+) < P(S_-)$$

$$P(S \leq q | p = 0.5) = \alpha$$

where S is the number ($S = \frac{S_+}{m^*}$, where $m^* = S_+ + S_-$) of F-test with positive differences, q is the critical value and α the level of significance. If $P(S_+) \leq q$, then the F-test, then H_0 will be rejected and conclude that F-test with positive differences is significantly less than the number of KW test power with negative differences.

4. Simulation studies

4.1. The criteria

A Monte Carlo simulation is implemented to access the performances of both the one-way ANOVA and KW methods for the following scenarios are listed in Table 1. The following setup and conditions have been defined for purposes of clarity (eight scenarios for normal and eight scenarios for non-normal distributions) and reproducibility. All simulations were carried out using R software and plots with MATLAB. Each of the 194 experiments were replicated $N = 1000$ times.

Table 1: The 16 scenarios for both normal and non-normal distributions

S/N	n	μ	σ
1	Equal	Equal	Equal
2	Equal	Equal	Unequal
3	Equal	Unequal	Equal
4	Equal	Unequal	Unequal
5	Unequal	Equal	Equal
6	Unequal	Equal	Unequal
7	Unequal	Unequal	Equal
8	Unequal	Unequal	Unequal

Note: S/N = simulation scenario, n = sample size, μ = mean, σ = standard deviation

4.1.1. Balanced/unbalanced design

In this study n_i ($i = 1,2,3$) was decided upon for convenience without loss of generality. By definition, a completely randomized design is said to be balanced if the group sizes are equal, i.e., $n = n_1 = n_2 = n_3$. In this work balanced data is taken to mean $n = n_1 = n_2 = n_3 = 5,10,15,\ldots,60$. In the unbalanced cases, the total sample size is given by the sum of all group sizes, i.e., $n = n_1 + n_2 + n_3$. For example, $n_1 = 3$, $n_2 = 3$, $n_3 = 4$ so $n = 10$. Or, if $n_1 = 13$, $n_2 = 23$, $n_3 = 24$ then $n = 60$.

4.1.2. Equal/unequal group means

In the simulated data sets, equal means are taken to be $\mu = \mu_1 = \mu_2 = \mu_3 = 8$ and unequal means understood to mean the vector $\mu = (8,9,11) = (\mu_1, \mu_2, \mu_3)$.

4.1.3. Homogeneity/heterogeneity of variances

For homogeneity of group variances, each group standard deviation (SD) σ is taken to be $\sigma = \sigma_1 = \sigma_2 = \sigma_3 = 5$. However in the heterogeneity scenario, the vector $\sigma = (\sigma_1, \sigma_2, \sigma_3) = (5.3, 8.5, 11.3)$.

4.1.4. Normal/non-normal

In all simulation experiments, equal or unequal sample sizes, equal or unequal group means and equal or unequal group variances were considered. The eight simulation studies carried out were based on the assumption of normality of data sets so generated. For the non-normal simulations, two scenarios were considered:

1. For equal n the multivariate non-normal distributed data sets were generated with the `mvnonnorm()` function in the `semTools` and `MASS` packages of the R software based on Vale and Maurelli (1983) method. The parameters of this function include, among others, the `variance-covariance matrix` $V = (v_{11}, v_{22}, v_{33}, v_{21}, v_{31}, v_{32})$, `skewness` $Sk = (s_1 = s_2 = s_3 = 1.5)$, and `kurtosis` $K = (k_1 = k_2 = k_3 = 3.5)$ vectors. While the skewness and kurtosis parameters were kept constant, V varied according as equal or unequal variances; $V = (v_{11} = 5.3, v_{22} = 8.5, v_{33} = 11.3, v_{21} = 2, v_{31} = 1, v_{32} = 2)$ for unequal and $\text{diag}(V) = (v_{11} = v_{22} = v_{33} = 8)$ for equal variances.
2. For unequal n, the log-normal distribution data sets were also generated for equal or unequal means; `meanlog` = $\mu_1 = \mu_2 = \mu_3 = 8$ or `meanlog` = $(8, 9, 11)$, and standard deviation, `sdlog` = $(5, 5, 5)$ or $(5.3, 8.5, 11.3)$ respectively.

4.2. The algorithm

The algorithm of the simulation experiment can be depicted in the following steps:

1. Generate three random samples in accordance with Sections 4.1.1–4.1.4.
2. Run both the ANOVA and KW tests on the independent groups simulated in step 1 at $\alpha = 0.05$ level of significance.
3. Calculate the p-values from the tests.
4. Repeat steps 1–3 1000 times.
5. Calculate the probability of rejecting the null hypothesis when it is true (i.e., Type I error).
6. Compute power by obtaining the proportion of simulation runs that rejected H_0.

5. Results and discussions

Results of the simulation studies carried out in Section 4 above are presented in Tables 2–17. Each of these tables consists of nine columns, the first of which is the sample size, n, an integer for equal sample sizes and a vector with three elements in the case of unbalanced design. The computed power values corresponding to the respective sample sizes for the F-test and KW test are given in columns two and three. The test statistic W for the Shapiro–Wilk's test for normality and its corresponding p-values are in columns 4–6. Similarly, values displayed in columns 7–9 are results for the K^2 statistic from the Bartlett's K^2 test for homogeneity of variances when the data set under consideration is assumed normal, otherwise the Levene L-test was used. The hypotheses were true (T) or false (F) according to the test results are reflected in the p-value.

From Table 2, 12 experiments were performed and each replicated 1000 times. At the end of each experiment, the F-test and KW-test were conducted and their respective powers obtained. Similarly, the Shapiro–Wilk's W statistic for normality test and Bartlett's K^2 statistic and Levene's L statistic for tests for equality of variances were extracted. Following the Shapiro–Wilk's test on data set when $n = 5, \mu = 8, \sigma = 5$ in Table 2, we assume that the data set is normally distributed ($W = 0.978, p > 0.05(0.955)$). Similarly, the Bartlett's test on the same data set confirm the homogeneity ($K^2 = 0.307, p > 0.05(0.858)$) of group variances of the data set.

For equal means, variances and sample sizes, the normality and equality of variances assumptions were maintained (or nearly so) in all 12 experiments displayed in Table 2. It was observed that the power of the F-test was higher than that of the KW test in eight out of 12 experiments (Table 2 and Figure 1(a)). Similarly, results of power analyses displayed in Table 3 show that F-test was slightly better than the KW test only in five experiments, while the KW test performed better in five experiments and both tests tied in two of the 12 experiments. This is displayed in Figure 1(b). This poor performance of the F-test may have been due to the violation of the equal variance assumption where $\sigma_i = (\sigma_1, \sigma_2, \sigma_3) = (5.3, 8.5, 11.3)$. F-test performed better in all 12 experiments in Table 4 as the group sample sizes were equal and equality of variances assumptions were respected. In situations of unequal group means, the power of both tests showed a positive trend with increasing sample sizes (panels (c) and (d) of Figures 1–4). Increasing sample size, however, did not influence the power of F-test in the log-normal situations. Regardless of group sizes and whether or not the variances were equal, the two tests under comparison demonstrated nearly identical power except in the log-normal scenarios when the KW maintained dominance (Figure 4). Similar observations could be rendered to results presented in Tables 6–9, and displayed graphically in Figures 1 and 2 where normal distributions were assumed.

The situation in non-normal scenarios is of interest in this work. Except for results in Table 10, KW test performed better than the F-test (Tables 11–14), but in multivariate non-normal cases, both tests showed indications of asymptotic convergence when sample sizes are equal and the means are unequal (Tables 12 and 13, and Figure 3 in panels (c) and (d)). The superior performance of the power of the KW test over the F-test in a non-normal scenario is demonstrated under the lognormal distribution (Tables 14–17, and Figure 4).

These results are in tandem with the trends in research involving parametric and non-parametric statistics (Blanca Mena et al., 2017; Sawilowsky et al., 1989) where non-parametric tests are less powerful than parametric tests but such power gap is small. On the other hand, these authors observed that the power advantage of the non-parametric tests under conditions of non-normality can be dramatic.

Further analysis of the values of power of the F-test and the KW test were performed using data in Tables 2–17 and displayed in Table 18. The negative values in the t-statistic column indicate scenarios where the power of the KW is higher than the F-test, and positive otherwise. The corresponding p-values show the significance of the differences ($p < 0.05$); the hypothesis of no difference in power are in 11 out of the 16 tests especially when data sets were assumed to be normally distributed.

The result of the t-test that was carried out to see if the power of the F-test is indeed higher than that of the KW test using values in columns 3 and 4 of Table 18 showed that with 11 degrees of freedom for Tables 2–5 and Tables 10–13 and 10 degrees of freedom for those with 11 experiments, the t-tests rejected the hypothesis of no difference ($p < 0.05$) only in 5 out of 16 tests. The KW test performed about three times better than the conventional F-test. Since the p-values are very small ($p < 0.001$) in the five cases, there are very small probability of these results occurring by chance.

Similarly, the non-parametric sign test results as displayed (Table 18) where the statistic S (shown in bold) is the number of positive differences where F-test performed better than KW test. For instance, **8**/12 is understood to mean F-test is higher in power than KW in 8 out of 12 experiments in Table 2 and that the difference is statistically different ($p > 0.05$). The values of S in this table show also that KW test outperformed the F-test, ($p < 0.05$). The overall result shows that the F-test is better in only 74 out of 184 experiments.

Table 2: Normal: Equal n, $\mu = 8$, $\sigma = 5$, 12 experiments

	Power		Normality			Homogeneity		
n	F	KW	W	p	H_0	K^2	p	H_0
5	0.045	0.042	0.978	0.955	T	0.307	0.858	T
10	0.058	0.054	0.970	0.532	T	2.403	0.297	T
15	0.054	0.032	0.977	0.519	T	1.807	0.405	T
20	0.048	0.047	0.981	0.477	T	1.519	0.468	T
25	0.049	0.054	0.978	0.229	T	0.366	0.833	T
30	0.049	0.046	0.990	0.697	T	2.191	0.334	T
35	0.056	0.062	0.983	0.192	T	1.310	0.520	T
40	0.038	0.042	0.982	0.109	T	5.325	0.070	T
45	0.048	0.046	0.988	0.312	T	4.169	0.124	T
50	0.053	0.057	0.996	0.941	T	0.537	0.765	T
55	0.047	0.041	0.997	0.969	T	0.018	0.991	T
60	0.041	0.036	0.987	0.101	T	0.221	0.895	T

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 3: Normal: Equal n, $\mu = 8$, $\sigma = (5.3, 8.5, 11.3)$, 12 experiments

	Power		Normality			Homogeneity		
n	F	KW	W	p	H_0	K^2	p	H_0
5	0.048	0.041	0.959	0.678	T	2.499	0.287	T
10	0.051	0.052	0.982	0.865	T	0.118	0.943	T
15	0.044	0.046	0.978	0.540	T	2.436	0.296	T
20	0.058	0.044	0.962	0.059	T	12.998	0.002	F
25	0.057	0.058	0.964	0.032	F	8.874	0.012	F
30	0.047	0.043	0.979	0.154	T	15.743	0.000	F
35	0.046	0.046	0.991	0.704	T	3.976	0.137	T
40	0.051	0.058	0.961	0.002	F	18.580	0.000	F
45	0.058	0.052	0.979	0.034	F	26.686	0.000	F
50	0.045	0.050	0.991	0.414	T	28.913	0.000	F
55	0.065	0.060	0.987	0.123	T	24.962	0.000	F
60	0.058	0.061	0.988	0.114	T	24.053	0.000	F

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 4: Normal: Equal n, $\mu = (8,9,11)$, $\sigma = 5$, 12 experiments

	Power		Normality			Homogeneity		
n	F	KW	W	p	H_0	K^2	p	H_0
5	0.103	0.089	0.958	0.658	T	0.444	0.801	T
10	0.189	0.165	0.948	0.150	T	6.057	0.048	F
15	0.283	0.271	0.952	0.063	T	2.753	0.252	T
20	0.382	0.360	0.986	0.696	T	0.204	0.903	T
25	0.474	0.435	0.994	0.982	T	3.725	0.155	T
30	0.559	0.535	0.979	0.165	T	2.832	0.243	T
35	0.615	0.585	0.994	0.919	T	2.434	0.296	T
40	0.680	0.665	0.988	0.368	T	1.203	0.548	T
45	0.704	0.688	0.991	0.495	T	0.486	0.784	T
50	0.767	0.750	0.997	0.981	T	1.577	0.455	T
55	0.818	0.796	0.988	0.187	T	1.529	0.466	T
60	0.861	0.846	0.991	0.323	T	0.091	0.956	T

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 5: Normal: Equal n, $\mu = (8,9,11)$, $\sigma = (5.3,8.5,11.3)$, 12 experiments

	Power		Normality			Homogeneity		
n	F	KW	W	p	H_0	K^2	p	H_0
5	0.072	0.067	0.941	0.390	T	1.262	0.532	T
10	0.096	0.082	0.985	0.934	T	2.672	0.263	T
15	0.128	0.111	0.982	0.710	T	4.990	0.082	T
20	0.145	0.130	0.983	0.551	T	8.523	0.014	F
25	0.169	0.162	0.971	0.081	T	28.763	0.000	F
30	0.200	0.196	0.963	0.012	F	12.461	0.002	F
35	0.239	0.212	0.988	0.448	T	3.879	0.144	T
40	0.283	0.253	0.967	0.035	F	24.372	0.000	F
45	0.269	0.257	0.991	0.557	T	10.809	0.004	F
50	0.324	0.306	0.990	0.354	T	20.851	0.000	F
55	0.350	0.328	0.991	0.367	T	30.040	0.000	F
60	0.344	0.323	0.989	0.161	T	28.190	0.000	F

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 6: Normal: Unequal n, $\mu = 8$, $\sigma = 5$, 11 experiments

	Power		Normality			Homogeneity		
(n_1, n_2, n_3)	F	KW	W	p	H_0	K^2	p	H_0
3,3,4	0.052	0.026	0.896	0.198	T	2.198	0.333	T
4,5,6	0.043	0.037	0.923	0.211	T	6.214	0.045	F
5,7,8	0.043	0.043	0.954	0.432	T	1.940	0.379	T
6,9,10	0.035	0.029	0.964	0.509	T	0.529	0.767	T
7,11,12	0.060	0.056	0.940	0.092	T	1.295	0.523	T
8,13,14	0.053	0.045	0.954	0.152	T	0.021	0.990	T
9,15,16	0.054	0.046	0.961	0.177	T	0.015	0.992	T
10,17,18	0.054	0.051	0.968	0.249	T	1.105	0.576	T
11,19,20	0.042	0.048	0.977	0.421	T	1.199	0.549	T
12,21,22	0.053	0.051	0.985	0.743	T	3.158	0.206	T
13,23,24	0.050	0.035	0.974	0.233	T	2.884	0.237	T

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 7: Normal: Unequal n, $\mu = 8$, $\sigma = (5.3, 8.5, 11.3)$, 11 experiments

	Power		Normality			Homogeneity		
(n_1, n_2, n_3)	F	KW	W	p	H_0	K^2	p	H_0
(3,3,4)	0.042	0.040	0.966	0.854	T	3.712	0.156	T
(4,5,6)	0.044	0.035	0.880	0.047	F	5.232	0.073	T
(5,7,8)	0.037	0.038	0.976	0.874	T	3.958	0.138	T
(6,9,10)	0.027	0.026	0.995	0.761	T	5.334	0.069	T
(7,11,1)	0.035	0.039	0.951	0.185	T	8.036	0.018	F
(8,13,14)	0.042	0.040	0.968	0.379	T	5.198	0.074	T
(9,15,16)	0.028	0.030	0.934	0.022	T	5.910	0.052	T
(10,17,18)	0.041	0.043	0.977	0.515	T	8.023	0.018	F
(11,19,20)	0.042	0.049	0.985	0.765	T	13.503	0.001	F
(12,21,22)	0.028	0.030	0.971	0.208	T	5.704	0.058	T
(13,23,24)	0.037	0.037	0.958	0.036	F	9.948	0.007	F

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 8: Normal: Unequal n, $\mu = (8,9,11)$, $\sigma = 5$, 11 experiments

	Power		Normality			Homogeneity		
(n_1,n_2,n_3)	F	KW	W	p	H_0	K^2	p	H_0
(3,3,4)	0.076	0.053	0.949	0.655	T	0.739	0.691	T
(4,5,6)	0.105	0.095	0.978	0.953	T	2.133	0.344	T
(5,7,8)	0.149	0.138	0.950	0.373	T	0.468	0.791	T
(6,9,10)	0.157	0.141	0.969	0.613	T	0.819	0.664	T
(7,11,1)	0.175	0.157	0.980	0.814	T	1.066	0.587	T
(8,13,14)	0.188	0.186	0.972	0.497	T	2.821	0.244	T
(9,15,16)	0.241	0.226	0.963	0.218	T	3.338	0.188	T
(10,17,18)	0.294	0.267	0.983	0.749	T	3.265	0.196	T
(11,19,20)	0.286	0.273	0.996	0.999	T	0.715	0.699	T
(12,21,22)	0.344	0.318	0.968	0.145	T	0.275	0.871	T
(13,23,24)	0.350	0.335	0.960	0.047	F	0.017	0.991	T

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 9: Normal: Unequal n, $\mu = (8,9,11)$, $\sigma = (5.3,8.5,11.3)$, 11 experiments

	Power		Normality			Homogeneity		
(n_1,n_2,n_3)	F	KW	W	p	H_0	K^2	p	H_0
(3,3,4)	0.058	0.041	0.878	0.125	T	3.490	0.175	T
(4,5,6)	0.057	0.053	0.957	0.648	T	3.268	0.195	T
(5,7,8)	0.055	0.048	0.955	0.448	T	3.311	0.191	T
(6,9,10)	0.060	0.064	0.882	0.008	F	8.301	0.016	F
(7,11,1)	0.068	0.076	0.944	0.117	T	2.419	0.298	T
(8,13,14)	0.073	0.073	0.962	0.260	T	6.517	0.038	F
(9,15,16)	0.088	0.090	0.987	0.931	T	3.860	0.145	T
(10,17,18)	0.085	0.078	0.991	0.972	T	5.523	0.063	T
(11,19,20)	0.096	0.099	0.970	0.223	T	18.740	0.000	F
(12,21,22)	0.095	0.094	0.980	0.492	T	9.512	0.009	F
(13,23,24)	0.112	0.101	0.963	0.068	T	18.660	0.000	F

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 10: Multivariate Non-normal: Equal n, $\mu = 8$, $\sigma = 5$, 12 experiments

	Power		Normality			Homogeneity		
n	F	KW	W	p	H_0	L	p	H_0
5	0.025	0.023	0.900	0.095	T	1.904	0.386	T
10	0.019	0.017	0.800	0.000	F	3.544	0.170	T
15	0.023	0.020	0.850	0.000	F	1.443	0.486	T
20	0.016	0.022	0.848	0.000	F	7.634	0.022	F
25	0.015	0.019	0.797	0.000	F	0.852	0.653	T
30	0.014	0.013	0.793	0.000	F	5.715	0.057	T
35	0.021	0.021	0.879	0.000	F	1.980	0.371	T
40	0.015	0.014	0.874	0.000	F	6.144	0.046	F
45	0.019	0.022	0.940	0.000	F	2.355	0.308	T
50	0.011	0.014	0.925	0.000	F	3.346	0.188	T
55	0.014	0.017	0.873	0.000	F	7.189	0.028	F
60	0.017	0.019	0.911	0.000	F	3.670	0.160	T

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 11: Multivariate Non-normal: Equal n, $\mu = 8$, $\sigma = (5.3, 8.5, 11.3)$, 12 experiments

	Power		Normality			Homogeneity		
n	F	KW	W	p	H_0	L	p	H_0
5	0.030	0.035	0.864	0.028	F	4.093	0.129	T
10	0.023	0.031	0.968	0.494	T	1.377	0.502	T
15	0.022	0.029	0.768	0.000	F	8.910	0.012	F
20	0.035	0.044	0.811	0.000	F	16.417	0.000	F
25	0.023	0.042	0.910	0.000	F	10.169	0.006	F
30	0.024	0.043	0.891	0.000	F	2.911	0.233	T
35	0.023	0.049	0.879	0.000	F	0.562	0.755	T
40	0.022	0.053	0.874	0.000	F	18.222	0.000	F
45	0.028	0.046	0.923	0.000	F	1.142	0.565	T
50	0.023	0.066	0.888	0.000	F	18.459	0.000	F
55	0.021	0.063	0.901	0.000	F	7.396	0.025	F
60	0.024	0.081	0.850	0.000	F	18.287	0.000	F

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 12: Multivariate Non-normal: Equal n, $\mu = (8,9,11)$, $\sigma = 5$, 12 experiments

	Power		Normality			Homogeneity		
n	F	KW	W	p	H_0	L	p	H_0
5	0.434	0.425	0.913	0.148	T	2.730	0.255	T
10	0.734	0.824	0.944	0.114	T	1.230	0.541	T
15	0.907	0.962	0.947	0.040	F	3.762	0.153	T
20	0.973	0.993	0.976	0.275	T	0.437	0.804	T
25	0.994	0.999	0.943	0.002	F	3.370	0.186	T
30	0.998	1.000	0.933	0.000	F	3.578	0.167	T
35	1.000	1.000	0.920	0.000	F	5.186	0.075	T
40	1.000	1.000	0.936	0.000	F	0.199	0.905	T
45	1.000	1.000	0.952	0.000	F	2.228	0.328	T
50	1.000	1.000	0.948	0.000	F	1.752	0.417	T
55	1.000	1.000	0.976	0.005	F	3.326	0.190	T
60	1.000	1.000	0.980	0.012	F	5.754	0.056	T

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 13: Multivariate Non-normal: Equal n, $\mu = (8,9,11)$, $\sigma = (5.3,8.5,11.3)$, 12 experiments

	Power		Normality			Homogeneity		
n	F	KW	W	p	H_0	L	p	H_0
5	0.501	0.538	0.890	0.005	F	5.792	0.055	T
10	0.743	0.813	0.840	0.000	F	4.099	0.129	T
15	0.225	0.240	0.787	0.003	F	5.704	0.058	T
20	0.865	0.914	0.923	0.001	F	2.016	0.365	T
25	0.936	0.974	0.945	0.003	F	0.757	0.685	T
30	0.976	0.996	0.833	0.000	F	15.273	0.001	F
35	0.988	0.996	0.928	0.000	F	1.549	0.461	T
40	0.998	1.000	0.874	0.000	F	26.109	0.000	F
45	0.998	1.000	0.957	0.000	F	0.475	3.789	T
50	0.999	1.000	0.890	0.000	F	9.985	0.007	F
55	1.000	1.000	0.870	0.000	F	1.805	0.406	T
60	1.000	1.000	0.913	0.000	F	5.562	0.062	T

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 14: Lognormal: Unequal n, $\mu = 8$, $\sigma = 5$, 11 experiments

	Power		Normality			Homogeneity		
(n_1, n_2, n_3)	F	KW	W	p	H_0	K^2	p	H_0
(3,3,4)	0.042	0.029	0.962	0.811	T	0.505	0.624	T
(4,5,6)	0.051	0.041	0.909	0.131	T	0.148	0.864	T
(5,7,8)	0.053	0.040	0.945	0.297	T	1.806	0.194	T
(6,9,10)	0.043	0.043	0.956	0.333	T	0.044	0.957	T
(7,11,12)	0.059	0.047	0.938	0.080	T	2.604	0.092	T
(8,13,14)	0.052	0.050	0.912	0.008	F	1.657	0.207	T
(9,15,16)	0.052	0.056	0.969	0.340	T	0.000	1.000	T
(10,17,18)	0.054	0.052	0.923	0.005	F	1.508	0.233	T
(11,19,20)	0.054	0.055	0.923	0.005	F	1.508	0.233	T
(12,21,22)	0.056	0.048	0.976	0.323	T	1.192	0.312	T
(13,23,24)	0.053	0.050	0.980	0.437	T	0.229	0.796	T

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 15: Lognormal: Unequal n, $\mu = 8$, $\sigma = (5.3, 8.5, 11.3)$, 11 experiments

	Power		Normality			Homogeneity		
(n_1, n_2, n_3)	F	KW	W	p	H_0	K^2	p	H_0
(3,3,4)	0.042	0.029	0.962	0.811	T	0.505	0.624	T
(4,5,6)	0.051	0.041	0.909	0.131	T	0.148	0.864	T
(5,7,8)	0.053	0.040	0.945	0.297	T	1.806	0.194	T
(6,9,10)	0.043	0.043	0.956	0.333	T	0.044	0.957	T
(7,11,12)	0.059	0.047	0.938	0.080	T	2.604	0.092	T
(8,13,14)	0.052	0.050	0.912	0.008	F	1.657	0.207	T
(9,15,16)	0.052	0.056	0.969	0.340	T	0.000	1.000	T
(10,17,18)	0.054	0.052	0.923	0.005	F	1.508	0.233	T
(11,19,20)	0.054	0.055	0.923	0.005	F	1.508	0.233	T
(12,21,22)	0.056	0.048	0.976	0.323	T	1.192	0.312	T
(13,23,24)	0.053	0.050	0.980	0.437	T	0.229	0.796	T

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 16: Lognormal: Unequal n, $\mu = (8,9,11)$, $\sigma = 5$, 11 experiments

	Power		Normality			Homogeneity		
(n_1,n_2,n_3)	F	KW	W	p	H_0	K^2	p	H_0
(3,3,4)	0.006	0.046	0.560	0.000	F	0.995	0.417	T
(4,5,6)	0.002	0.030	0.285	0.000	F	0.723	0.506	T
(5,7,8)	0.002	0.035	0.289	0.000	F	0.997	0.390	T
(6,9,10)	0.003	0.044	0.203	0.000	F	0.733	0.492	T
(7,11,1)	0.001	0.041	0.232	0.000	F	0.988	0.355	T
(8,13,14)	0.001	0.033	0.232	0.000	F	0.777	0.468	T
(9,15,16)	0.001	0.042	0.147	0.000	F	0.740	0.484	T
(10,17,18)	0.001	0.036	0.136	0.000	F	0.819	0.448	T
(11,19,20)	0.001	0.044	0.125	0.000	F	0.742	0.482	T
(12,21,22)	0.000	0.026	0.128	0.000	F	0.828	0.443	T
(13,23,24)	0.000	0.030	0.208	0.000	F	0.000	0.145	T

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 17: Lognormal: Unequal n, $\mu = (8,9,11)$, $\sigma = (5.3,8.5,11.3)$, 11 experiments

	Power		Normality			Homogeneity		
(n_1,n_2,n_3)	F	KW	W	p	H_0	K^2	p	H_0
(3,3,4)	0.008	0.059	0.433	0.000	F	0.673	0.540	T
(4,5,6)	0.004	0.089	0.348	0.000	F	0.817	0.465	T
(5,7,8)	0.007	0.118	0.282	0.000	F	0.830	0.453	T
(6,9,10)	0.007	0.159	0.346	0.000	F	0.946	0.404	T
(7,11,1)	0.004	0.161	0.187	0.000	F	0.840	0.443	T
(8,13,14)	0.006	0.206	0.399	0.000	F	2.279	0.119	T
(9,15,16)	0.009	0.235	0.286	0.000	F	1.913	0.162	T
(10,17,18)	0.008	0.244	0.197	0.000	F	0.511	0.604	T
(11,19,20)	0.009	0.304	0.270	0.000	F	0.725	0.490	T
(12,21,22)	0.004	0.298	0.120	0.000	F	0.735	0.485	T
(13,23,24)	0.011	0.332	0.116	0.000	F	1.828	0.170	T

Note: n = sample size, F = F-test, KW = Kruskal–Wallis test, p = p-value

Table 18: Summary of performances and tests for differences

		t-test		Sign test	
Distribution	Table #	*t*	*p*	*S*	*p*
Normal	2	0.5941	0.559	**8**/12	0.927
Normal	3	0.5053	0.618	**5**/12	0.500
Normal	4	0.2031	0.841	**12**/12	1.000
Normal	5	0.4121	0.684	**12**/12	1.000
Normal	6	1.8049	0.087	**9**/11	0.999
Normal	7	–0.1323	0.896	**4**/11	0.377
Normal	8	0.4035	0.691	**11**/11	1.000
Normal	9	0.3186	0.753	**6**/11	0.828
Non-normal	10	–0.6485	0.524	**5**/12	0.500
Non-normal	11	–5.1883	0.000	**0**/12	0.000
Non-normal	12	–0.1960	0.846	**1**/12	0.109
Non-normal	13	–0.4068	0.688	**0**/12	0.000
Non-normal	14	–13.6000	0.000	**0**/11	0.000
Non-normal	15	–8.2373	0.000	**1**/11	0.005
Non-normal	16	–7.0466	0.000	**0**/11	0.000
Non-normal	17	–11.6210	0.000	**0**/11	0.000

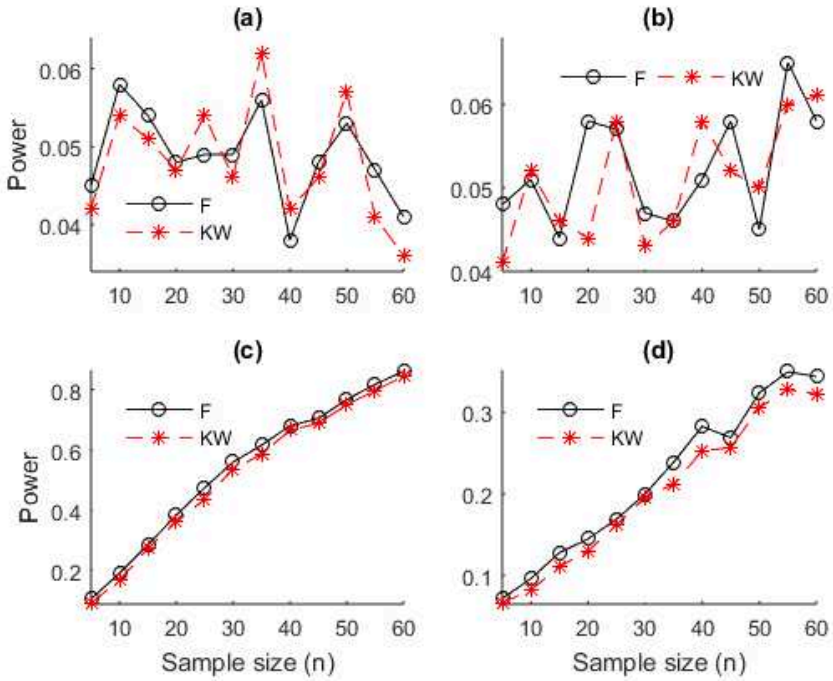

Figure 1: Normal: (a) Equal n, $\mu = 8$, $\sigma = 5$; (b) Equal n, $\mu = 8$, $\sigma = (5.3, 8.5, 11.3)$; (c) Equal n, $\mu = (8, 9, 11)$, $\sigma = 5$; (d) Equal n, $\mu = (8, 9, 11)$, $\sigma = (5.3, 8.5, 11.3)$

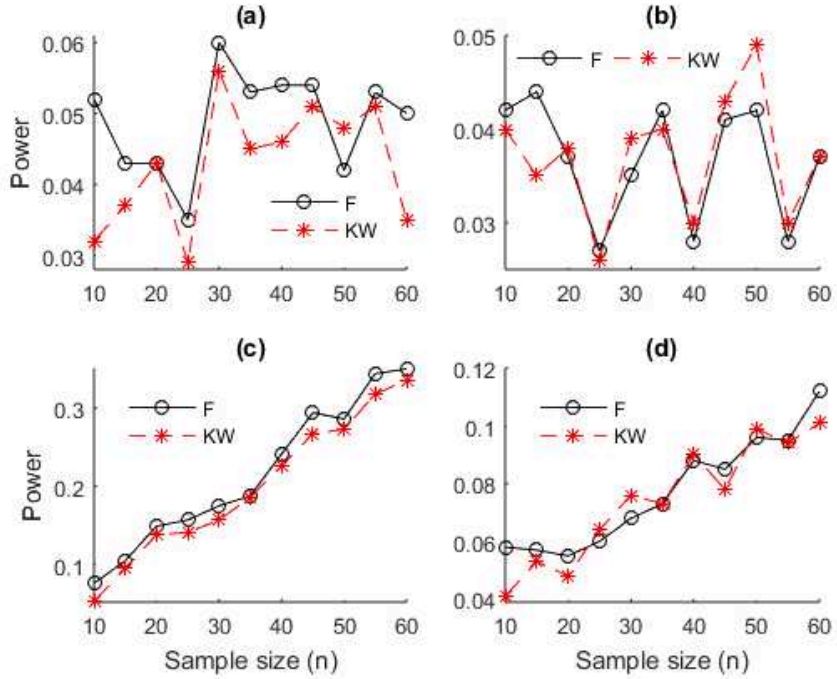

Figure 2: Normal: (a) Unequal n, $\mu = 8$, $\sigma = 5$; (b) Unequal n, $\mu = 8$, $\sigma = (5.3, 8.5, 11.3)$; (c) Unequal n, $\mu = (8, 9, 11)$, $\sigma = 5$; (d) Unequal n, $\mu = (8, 9, 11)$, $\sigma = (5.3, 8.5, 11.3)$

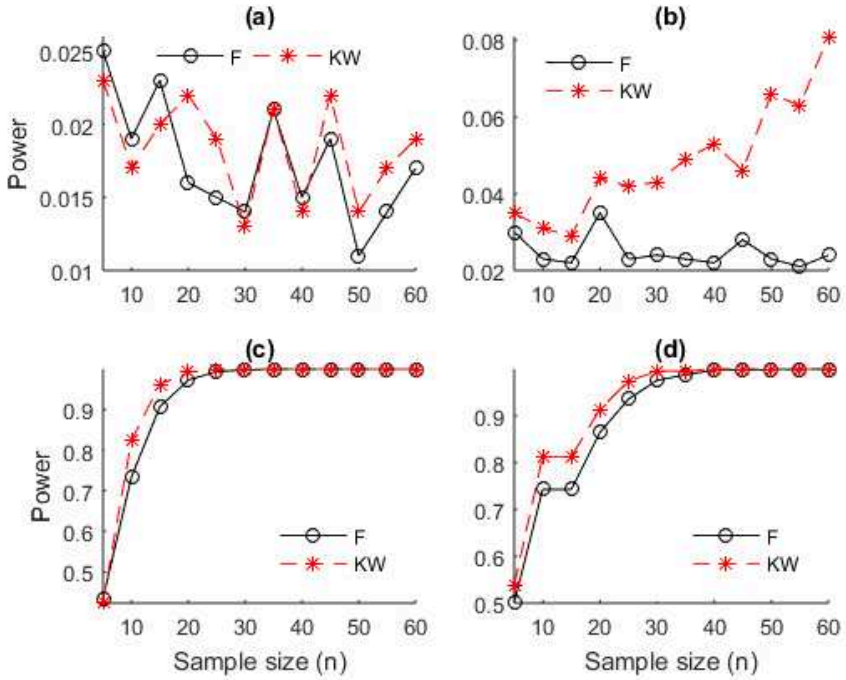

Figure 3: Multivariate Non-normal: (a) Equal n, $\mu = 8$, $\sigma = 5$; (b) Equal n, $\mu = 8$, $\sigma = (5.3, 8.5, 11.3)$; (c) Equal n, $\mu = (8, 9, 11)$, $\sigma = 5$; (d) Equal n, $\mu = (8, 9, 11)$, $\sigma = (5.3, 8.5, 11.3)$

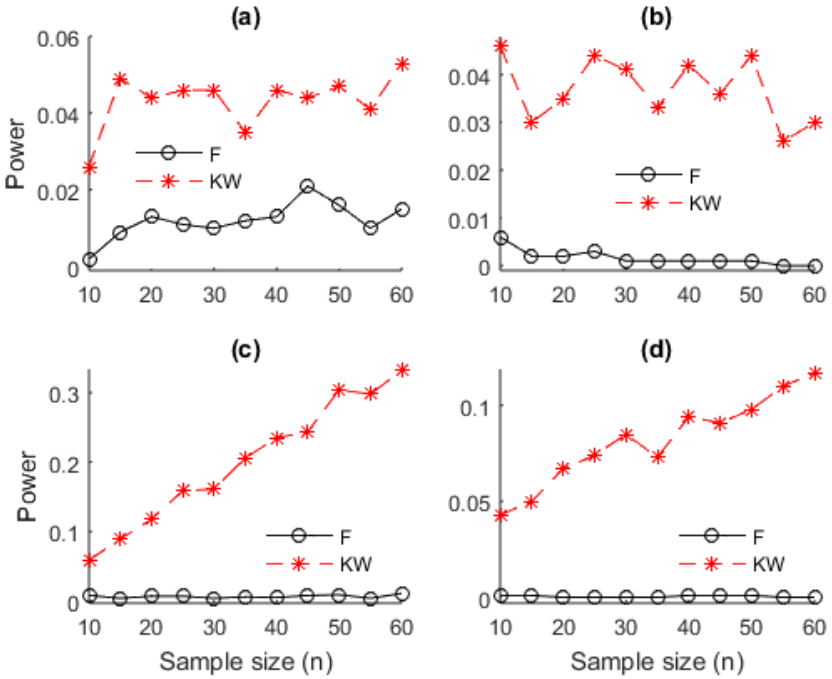

Figure 4: Lognormal: (a) Unequal n, $\mu = 8$, $\sigma = 5$; (b) Unequal n, $\mu = 8$, $\sigma = (5.3, 8.5, 11.3)$; (c) Unequal n, $\mu = (8, 9, 11)$, $\sigma = 5$; (d) Unequal n, $\mu = (8, 9, 11)$, $\sigma = (5.3, 8.5, 11.3)$

6. Conclusion

The purpose of this study was to compare the power of the parametric ANOVA F-test and its alternative, the non-parametric Kruskal–Wallis KW test where the assumptions of normality and homogeneity of variances are violated. The power of both tests showed a particular pattern in the case of equal means for normal and non-normal situations. In unequal group mean scenarios, they showed positive trends with increasing sample sizes for balanced or unbalanced designs, the distribution of the data set notwithstanding.

This study has shown that the instances when the F-test was more powerful than the KW test, it is often very difficult to distinguish. However, when the KW test was demonstrated to be more powerful, especially in non-normal scenarios, it came with a significant difference ($p < 0.05$). These results in general imply that the F-test has a higher risk of accepting the hypothesis of equality of group means when, indeed, they are not so. Specifically, the risk of using the F-test in the analysis of non-normal data is very high. Since it is rare to have perfect normality if ever, this study has provided more evidence that there is quite literally little to lose in using the Kruskal–Walis test as a non-parametric alternative to the parametric analysis of variance F-test.

Acknowledgements

The authors gratefully acknowledge the anonymous reviewers and the Editors for their time, constructive comments, and suggestions that led to the significant improvement of this paper.

References

Blanca Mena, M. J., Alarcón Postigo, R., Arnau Gras, J., Bono Cabré, R., Bendayan, R., et al. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, *29*(4), 552–557. https://doi.org/10.7334/psicothema2016.383

Ferreira, E. B., Rocha, M. C., & Mequelino, D. B. (2012). Monte Carlo evaluation of the ANOVA's F and Kruskal–Wallis tests under binomial distribution. *Sigmae*, *1*(1), 126–139.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. https://doi.org/10.3102/00346543042003237

Hecke, T. V. (2012). Power study of ANOVA versus Kruskal–Wallis test. *Journal of Statistics and Management Systems*, *15*(2-3), 241–247. https://doi.org/10.1080/09720510.2012.10701623

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill.

Lachenbruch, P. A., & Clements, P. J. (1991). ANOVA, Kruskal–Wallis, normal scores and unequal variance. *Communications in Statistics - Theory and Methods*, *20*(1), 107–126.

Legendre, P., & Borcard, D. (2008). *Statistical comparison of univariate tests of homogeneity of variances* [Unpublished manuscript]. Département de sciences biologiques, Université de Montréal.

Lehmann, E. L. (2006). *Nonparametrics: Statistical methods based on ranks*. Springer.

Marcinko, T. (2014). Consequences of assumption violations regarding one-way ANOVA. *Proceedings of The 8th International Days of Statistics and Economics*, *116*(47), 974–985.

Moder, K. (2007). How to keep the Type I error rate in ANOVA if variances are heteroscedastic. *Austrian Journal of Statistics*, *36*(3), 179–188. https://doi.org/10.17713/ajs.v36i3.329

Moder, K. (2010). Alternatives to F-test in one way ANOVA in case of heterogeneity of variances (a simulation study). *Psychological Test and Assessment Modeling*, *52*(4), 343–353.

Sahai, H., & Ageel, M. I. (2000). *The analysis of variance: Fixed, random and mixed models*. Springer.

Sawilowsky, S. S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the Type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational and Behavioral Statistics*, *14*(3), 255–267. https://doi.org/10.3102/10769986014003255

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*(3), 465–471. https://doi.org/10.1007/BF02293687