Evaluation of empirical type I error rates of F and normality tests under different variance and mean conditions in multi-treatment CRDs

Homero Ribeiro Neto[1]*, Marciel Lelis Duarte[1] and Nerilson Terra Santos[1]

[1] *Department of Statistics, Federal University of Viçosa, Viçosa, Brazil.*
** Corresponding author. E-mail: homero.neto@ufv.br*

ABSTRACT. Hypothesis tests, such as normality tests, are extensively employed in Agricultural Sciences to evaluate the normality assumption of the F test in the Analysis of Variance (ANOVA) when large sample sizes are unavailable. Nonetheless, researchers conducting these tests are exposed to the risk of committing type I or type II errors, with probabilities that are influenced by different experimental conditions. This study assesses the empirical type I error rate of hypothesis tests by considering the equality (inequality) of treatment means, the homogeneity (heterogeneity) of variances, and different numbers of repetitions per treatment. Applying Completely Randomized Designs (CRD), sub-scenarios were simulated for each experimental scenario, with 10,000 iterations performed for each sub-scenario. Response variable values and experimental residuals were generated and subjected to appropriate tests. The results demonstrate that when the assumption of homogeneity of variances is violated, both the F and normality tests (excluding the Kolmogorov-Smirnov test) exhibit higher empirical type I error rates. Additionally, for normality tests, these error rates increase with the number of repetitions. Conversely, without such violations, the error rates remain stable and closely approximate the theoretical significance level for all analyzed hypothesis tests.

Keywords: Level of Significance; ANOVA; Completely Randomized Design; Normal Distribution; Experimental Errors.

DOI: https://dx.doi.org/10.33837/msj.v8i1.1719

Received: 07/02/2025 *Online published: 22/04/2025*
Associate editor: Dr. Anderson Rodrigo da Silva

INTRODUCTION

In Statistics, hypothesis tests involve the decision between the null hypothesis (H0) and the alternative hypothesis (Ha). According to Bussab and Morettin (2010), these tests rely on the distributions of the estimator (test statistic), assuming the truthfulness of H0. The applicability of hypothesis tests arises from the fact that decisions regarding the hypotheses are based on samples, thus conserving resources compared to assessing an entire population. However, these tests entail the possibility of two types of errors: type I error (rejecting H0 when it is true) and type II error (not rejecting H0 when it is false), with inversely complementary probabilities α and β, respectively, which are present in all hypothesis tests. In agricultural research, Piepho and Edmondson (2018) emphasize that understanding these errors is crucial for interpreting results and making informed decisions about crop varieties or farming practices, highlighting the practical implications of statistical theory in applied settings.

Within the domain of Agricultural Sciences, hypothesis tests such as the F test of Analysis of Variance (ANOVA) and normality tests are frequently employed to verify assumptions and facilitate decision-making. Acutis et al. (2012) underscored the prevalence of these tests and multiple comparison tests to evaluate differences among means. As exemplified by Henrique and Laca-Buendía (2010), who assessed a novel genotype and five cotton cultivars across various production response variables, these comparisons among means play a vital role in selecting the most suitable treatments.

Nevertheless, it is imperative to acknowledge the risk of errors, particularly type I errors, as an erroneous selection of the perceived optimal treatment may lead to significant economic losses. Taking into account the results obtained by Laca-Buendía (2010), Rodrigues et al. 2010) conducted a comparative analysis of type I error rates across various mean tests, utilizing significant outcomes derived from the ANOVA F test. These precautions are indispensable in minimizing additional costs and potential adverse impacts on farmers.

Further emphasizing the importance of statistical accuracy in agricultural research, we must consider the broader implications in field trials. Researchers and farmers face challenges in accurately assessing the efficacy of different treatments, such as fertilizers, pesticides, or irrigation methods. As Piepho and Edmondson (2018) highlight, the challenge lies in drawing accurate conclusions from field experiments. Decision-making based on statistical models employed in these trials can result in type I error (falsely concluding a treatment effect when there is none) or type II error (failing to detect a genuine treatment effect). Hence, minimizing and comprehending the occurrence of these errors is of utmost importance. Factors such as small sample sizes, non-normal distributions, or variance heterogeneity can inflate Type I errors, potentially exacerbating the challenges in this context.

Within this context, the F-test, one of the most widely utilized tests for decision-making (Mood, 1974; Casella and Berger, 2002; Searle and Gruber, 2016), requires that the model error follows a normal distribution. Moreover, an additional assumption is that the variances within each treatment are identical. Consequently, it is expected that if the assumptions of normality and homogeneity are not met, there might be some repercussions on the type I and type II error rates. Violations of these assumptions can distort p-values, leading to unreliable conclusions. This distortion occurs because the p-value is calculated based on the theoretical distribution of the test statistic under the null hypothesis. When the underlying assumptions are violated, the actual distribution of the test statistic may differ from the theoretical one, resulting in p-values that no longer accurately represent the probability of obtaining the observed results under the null hypothesis, compromising the integrity of the decision-making process in agricultural insurance claims.

Considering the importance of the classic ANOVA F-test and its assumptions, the main goal of this study was to assess the type I error rate under different experimental conditions, considering smaller sample sizes that are more representative of applied research, in contrast to what has been done in other studies. Additionally, the study aims were to evaluate the impact of homogeneity (or heterogeneity) of treatment variances, equality (or inequality) of their means, and the number of repetitions per treatment on type I error rates of the normality tests Shapiro-Wilk (SW), Anderson-Darling (AD), Cramér-von Mises (CVM), Kolmogorov-Smirnov (KS), and Lilliefors (LI).

MATERIAL AND METHODS

The study investigated the impact of violating key assumptions of the F-test and normality tests. To this end, four scenarios were simulated, representing possible values of a response variable (y_{ij}) in an experiment conducted based on a completely randomized design (CRD) with five treatments and k replications per treatment, as shown in Figure 1. The four scenarios were derived from the factorial combination of two primary factors, each with two levels: the equality or inequality of treatment means and the homogeneity or heterogeneity of variances across treatments.

In the simulations, it was established that each treatment should follow a specific normal distribution. To achieve this, the four scenarios were distinguished based on the equality (inequality) of treatment means and the homogeneity (heterogeneity) of variances within treatments. The parameters for these distinct normal distributions used in the simulation are presented in Table 1.

For each of the four scenarios (Table 1), sub-scenarios were simulated by varying the number k of repetitions per treatment, such that k = 2, 4, 6, 8, and 10. Within each of these sub-scenarios, 10,000 iterations were simulated. For each iteration w such that $w = 1; 2; ...; 10{,}000$, a set w of $5k$ experimental residual values were obtained using Equation 1, where, $\hat{\varepsilon}_{ijw}$ represents the residual for the observed value y_{ijw} of the response variable in iteration w for replication j of treatment i; $\hat{\mu}_{iw}$ is the mean of the values of the response variable for treatment i in iteration w such that $w = 1, 2, \cdots, 10.000$; $i = 1, \cdots, 5$; $j = 1, \cdots, k$ and $k = 2, 4, 6, 8, 10$.

$$\hat{\varepsilon}_{ijw} = y_{ijw} - \hat{\mu}_{iw} \quad (1)$$

Subsequently, each set of residuals was evaluated by each one of the normality tests: Anderson-Darling (AD), Cramér-von Mises (CVM), Kolmogorov-Smirnov (KS), Lilliefors (LI), and Shapiro-Wilk (SW). These five tests were selected because they are

commonly used to assess normality across different sample sizes and are widely recognized in statistical literature. The p-value for each normality test in each iteration was recorded to calculate the empirical type I error rate ($\hat{\alpha}$) of each normality test in each sub-scenario, using Equation 2.

$$\hat{\alpha} = \frac{\boldsymbol{number\ of\ p-values \leq 0.05}}{\boldsymbol{10.000}} \quad (2)$$

Each set w of values of the response variable (not the residuals) for scenarios C1 and C2 sub-scenarios was also subjected to the F test. Notably, only these two scenarios were subjected to the F test as they represent conditions where the null hypothesis (H0) of equality of means is true. This selection is crucial because type I error occurs under a true H0, which is the fundamental premise in this context. The F test evaluates the null hypothesis that all treatment means are equal against the alternative hypothesis that at least one mean differs. The p-value of the F test in each iteration of the sub-scenarios was recorded to calculate the empirical rate of type I error ($\hat{\alpha}$), also using Equation 2.

Based on the results, graphs were plotted to illustrate the relationship between the number of repetitions (k) per treatment and the empirical rates of type I error ($\hat{\alpha}$) for each scenario.

To assess the empirical rates of type I error for the normality tests and the F test, the level of empirical significance for each of these tests was classified as follows: If $\hat{\alpha} > 0.05$, the test exhibited a high probability of type I error. If $\hat{\alpha} \leq 0.05$, the test exhibited a low probability of type I error.

The chi-squared test was employed to evaluate the independence between the homogeneity (heterogeneity) of treatment variances and the empirical rates of type I error ($\hat{\alpha}$). In cases where the chi-squared test was significant, and the sample size was less than 40, with at least one class having an expected frequency of less than 5, the Yates correction was applied as per Fukunaga et al. (2018).

This correction was deemed appropriate in these instances as it helps to reduce the upward bias in the chi-squared statistic that can occur when dealing with small sample sizes or low expected frequencies, thereby providing a more conservative and accurate estimate of statistical significance.

For the simulation iterations and the application of the normality and F tests, the software R, version 4.0.2 (R Core Team, 2020) was used. The simulations and statistical analyses were conducted using a suite of packages, including "tidyverse", "xlsx", "car", "GAD", "PMCMRplus", "DescTools", "outliers", "stats", "coin", "dplyr", and "nortest and "onewaytests". Furthermore, the graphical representations were generated using the "lattice" and "dplyr" packages.

Figure 1 shows the organization chart of each simulated scenario pattern and sub-scenario.

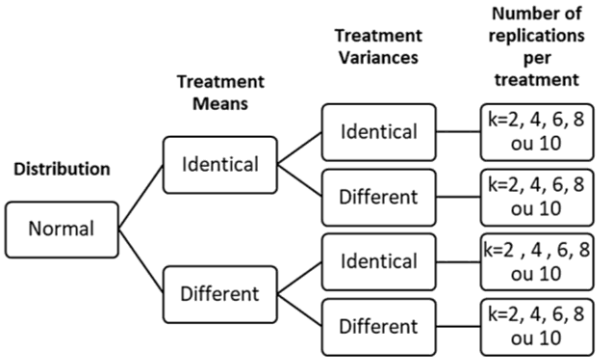

Figure 1. Simulated scenarios and sub-scenarios defining normal distributions and considering equality (inequality) of treatment means, homogeneity (heterogeneity) of treatment variances, and the number of replications per treatment (k=2, 4, 6, 8, 10).

Table 1 shows the values of means and standard deviations for each scenario.

Table 1. Values of means and standard deviations were established to simulate scenarios considering different normal distributions.

Distribution	Mean	Standard Deviation	Scenario
Normal	$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = 100$	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 1$	C1
		$\sigma_1 = 1; \sigma_2 = 2; \sigma_3 = 3; \sigma_4 = 4;\ \sigma_5 = 5$	C2
	$\mu_1 = 100; \mu_2 = 200; \mu_3 = 300; \mu_4 = 400; \mu_5 = 500$	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 1$	C3
		$\sigma_1 = 1; \sigma_2 = 2; \sigma_3 = 3; \sigma_4 = 4;\ \sigma_5 = 5$	C4

RESULTS AND DISCUSSION

Figure 2 presents the results for the empirical type I error rates in the C1 (homogeneous variances) and C2 (heterogeneous variances) scenarios and their sub-scenarios.

All normality tests exhibited satisfactory empirical type I error rates, that is $\hat{\alpha} \leq 0.05$ when the

means and standard deviations of the treatments were identical (Figure 2 (a)). For the AD, CVM, LI, and SW tests, the $\hat{\alpha}$ values were very close to the theoretical significance level (α = 0.05) for almost all numbers of replications (k) per treatment. The $\hat{\alpha}$ values for the KS test were visibly lower than the threshold of 0.05 and approached zero. These results confirm the conservative nature of the KS test, which was also reported by Torman et al. (2012), meaning that the KS test tends to reject the normality hypothesis less frequently than expected, resulting in fewer false positives. Conversely, Arnastauskaitė et al. (2021), Razali and Yap (2011), and Ogunleye et al. (2018) demonstrated that the Shapiro-Wilk (SW), Anderson-Darling (AD), and Lilliefors (LI) tests tend to be more powerful and less conservative than the Kolmogorov-Smirnov (KS) test in detecting deviations from normality. The higher power of SW, AD, and LI tests may be associated with their slightly higher empirical Type I error rates observed in the present study's C1 scenario, albeit still within acceptable limits. The F test also exhibited (Figure 2 (a)) satisfactory results for the empirical type I error rate for all evaluated numbers of repetitions.

However, these satisfactory $\hat{\alpha}$ values were not observed in the simulations of scenario C2 (Figure 2). The empirical type I error rates ($\hat{\alpha}$) increased and exceeded the theoretical threshold of $\hat{\alpha}$ = 0.05 in most sub-scenarios with different standard deviations (Figure 2 (b)). Except for F and KS, for all tests the significance level increased as the number of replications per treatment (k) increased. This outcome suggested a possible dependence between $\hat{\alpha}$ and the homogeneity (or heterogeneity) of variances within treatments. Except for the KS test, this dependence was confirmed by a chi-squared test (Table 2) for all tests.

Therefore, the empirical rates of type I error of the normality tests and the F test tend to be higher when the assumption of variances' homogeneity is false. It is worth noting that the KS test differed from the other normality tests as it did not exhibit a level of empirical significance higher than the theoretical threshold of 0.05 in any of the sub-scenarios. Consequently, it was impossible to use the chi-squared test for the KS test, as some cells of contingency chi-squared table were null. However, such test was not needed because the KS test pattern was essentially the same in both C1 and C2 (Figure 2), that is, $\hat{\alpha}$ values close to zero. This result indicates that the heterogeneity of variances did not influence the empirical error rate of the KS normality test, which might be explained by the fact that the parameters of the theoretical distribution (μ, σ^2) are supposed to be known and completely specified while performing a KS test, making it behave like an exact test.

Regarding the F test, the chi-squared test of independence (Table 2) revealed that when the within treatment variances were different, the empirical type I error rate increased. Therefore, there was a higher probability of rejecting the null hypothesis of equal treatment means when it should not be rejected. This result clearly shows the importance of the evaluation of the assumption of homogeneity of treatment variances before performing an Analysis of Variance (ANOVA). However, unlike the observed results for the normality tests, the increases in $\hat{\alpha}$ values for the F test in sub-scenarios of C2 (Figure 2) were relatively minor.

It was also evaluated the effect of inequality of treatment means on the normality test $\hat{\alpha}$ values (Figure 3). Considering the scenario C3 (unequal means and homogeneous within treatment variance) the empirical type I error rates remained practically stable for all numbers of repetitions evaluated in all sub-scenarios of C3, almost all $\hat{\alpha}$ values for the normality tests were either very close to 0.05 or lower. The results of the C3 sub-scenarios were very similar to those observed for sub-scenarios with equal means and homogeneous variances (C1 in Figure 2).

However, when comparing the simulation results of the C3 sub-scenarios (distinct means and homogeneous variances) with those obtained for C4 (distinct means and heterogeneous variances) in Figures 3 (a) and (b), respectively, the $\hat{\alpha}$ values increase, and this increase is proportional to the number of repetitions per treatment. To investigate the relationship between the empirical type I error rate and the homogeneity (or heterogeneity) of variances within treatments when the treatment means are different, the chi-squared test of independence was applied to each of the normality tests, considering the database of C3 and C4, which differ only in terms of the homogeneity or heterogeneity of treatment variances. The results of these chi-squared tests are presented in Table 3.

The results in Table 3 show that, for all evaluated normality tests, there is a significant dependence relationship between the empirical type I error rates and the condition of homogeneity (or heterogeneity) of variances within treatments in the scenarios. These results allow us to conclude that when the variances are heterogeneous, the incidence of type I error significantly increases.

It is noteworthy that the equality or inequality of means did not affect the simulation results (empirical type I error rates of normality tests), as evidenced by identical results observed for scenarios C1 and C3, and for C2 and C4, which differ only in treatment mean equality. This consistency explains the identical chi-squared test outcomes presented in Tables 2 and 3. Conversely, the alteration of variances (heterogeneous or homogeneous) had a significant

impact on the outcomes. For the simulation, a random seed was utilized to ensure reproducibility of results under the same conditions. Nevertheless, this approach did not prevent differences in type I error rates between scenarios C1 and C2 or between C3 and C4, where the variance conditions differed, further emphasizing the impact of variance heterogeneity on the results.

Figure 2 shows the results of scenarios C1 and C2. These scenarios differ only in terms of the residual variances of the treatments (homogeneity in C1 and heterogeneity in C2).

Figure 3 shows the results for scenarios C3 and C4. Although these two scenarios were simulated under inequality of means, their sub-scenarios differ regarding the residual variances of the treatments (homogeneity in C3 and heterogeneity in C4).

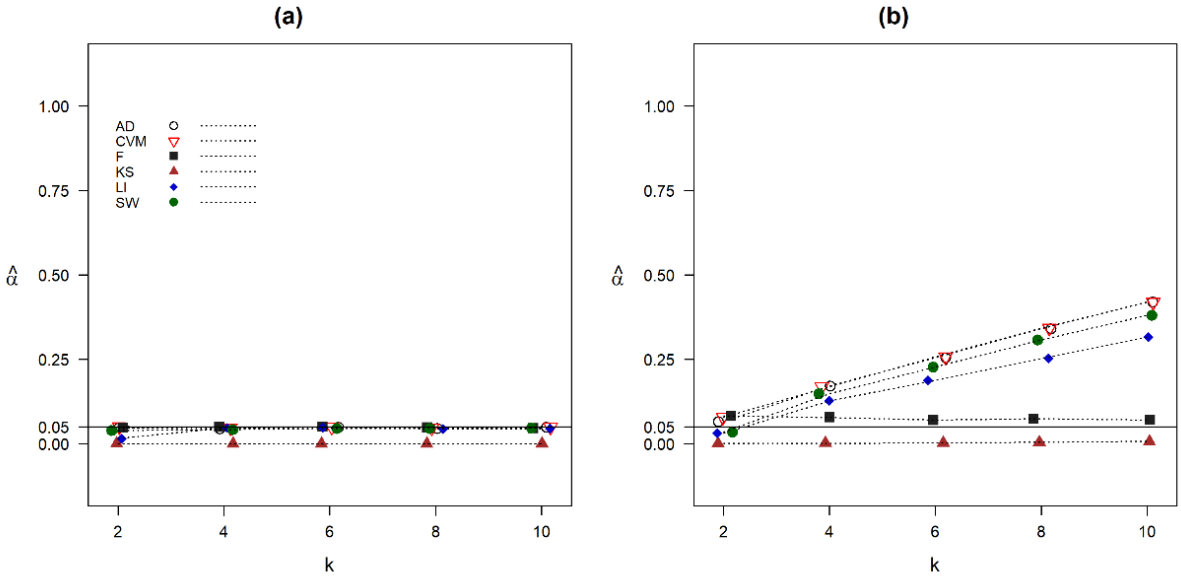

Figure 2. Empirical type I error rate ($\hat{\alpha}$) of the AD, CVM, KS, LI, SW, F tests as a function of the number of repetitions (k) per treatment in the sub-scenarios of C1 (a) and C2 (b) and the theoretical significance level (α=0.05) adopted in all tests.

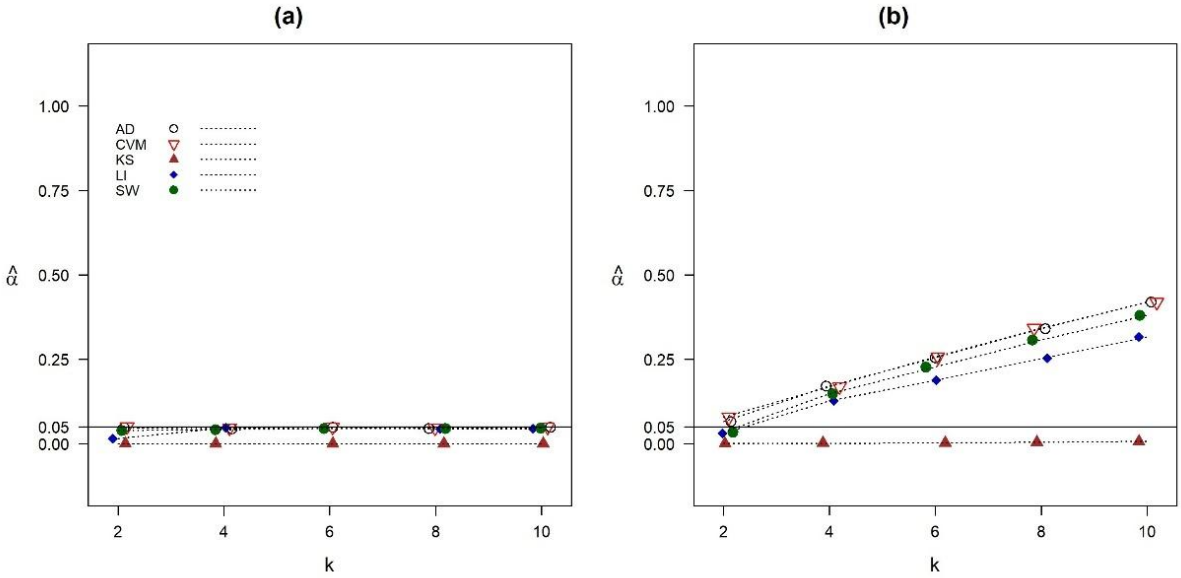

Figure 3. Empirical type I error rate ($\hat{\alpha}$) of the normality tests (AD, CVM, KS, LI, and SW) as a function of the number of repetitions (k) per treatment in the sub-scenarios of scenarios C3 (a) and C4 (b) and the theoretical significance level (α=0.05) adopted in all tests.

Table 2 shows the results for the chi-squared independence tests between empirical type I error rates (high if $\hat{\alpha}$>0.05 and low if $\hat{\alpha}$≤0.05) observed in the F, AD, KS, CVM, LI, and SW tests under the conditions of homogeneity (or heterogeneity) of treatment variances, using the database from scenarios C1 and C2.

Table 2. Chi-squared tests of independence between empirical type I error rates (high if $\hat{\alpha}$>0.05 and low if $\hat{\alpha}$≤0.05) observed in the F, AD, KS, CVM, LI, and SW tests under the conditions of homogeneity (or heterogeneity) of treatment variances, using the database from scenarios C1 and C2. Values in bold indicate significance (p-value≤ 0.05).

Test	χ^2	p-value
Anderson-Darling (AD)	6.40	0.01141
Kolmogorov-Smirnov (KS)	-	-
Cramér-von Mises (CVM)	3.75	0.05281
Lilliefors (LI)	3.75	0.05281
Shapiro-Wilk (SW)	3.75	0.05281
F	3.75	0.05281

Table 3 shows the results for the chi-squared independence tests between the empirical type I error rates (high if $\hat{\alpha}$>0.05 and low if $\hat{\alpha}$≤0.05) observed in the AD, KS, CVM, LI, and SW tests under the conditions of homogeneity (or heterogeneity) of treatment variances, using the database from scenarios C3 and C4.

Table 3. Chi-squared tests of independence between the empirical type I error rates (high if $\hat{\alpha}$>0.05 and low if $\hat{\alpha}$≤0.05) observed in the AD, KS, CVM, LI, and SW tests under the conditions of homogeneity (or heterogeneity) of treatment variances, using the database from scenarios C3 and C4. Values in bold indicate significance (p-value≤ 0.05).

Test	χ^2	p-value
Anderson-Darling (AD)	6.40	0.01141
Kolmogorov-Smirnov (KS)	-	-
Cramér-von Mises (CVM)	3.75	0.05281
Lilliefors (LI)	3.75	0.05281
Shapiro-Wilk (SW)	3.75	0.05281

The comparison of the results obtained in this work with those obtained in previous studies is crucial. The present study employs a novel approach by incorporating a defined experimental design, whereas most previous papers simulated data without such specification. This methodological distinction is critical as it allows for evaluating the effect of equality (inequality) of treatment means and homogeneity (heterogeneity) within treatment variances on empirical type I error rates of normality and F tests. Such evaluation was not feasible in previous studies due to their design limitations. Incorporating a defined experimental design enables a more comprehensive assessment of these tests under conditions that more closely approximate real experimental scenarios, thereby enhancing the applicability of the findings. Nevertheless, the results obtained in scenarios with treatment variance homogeneity (C1 and C3) are consistent with those observed by Ogunleye et al. (2018), Öztuna et al. (2006), Keskin (2006), and Torman et al. (2012).

Ogunleye et al. (2018) found that, in general, the normality tests that exhibited empirical type I error rates closest to the theoretical level of 5% significance were the SW test, followed by the KS test and the AD test. However, the differences between their empirical type I error rates were not significant. Furthermore, they observed specific stability of empirical significance levels across sample sizes ranging from 10 to 100, with 5,000 iterations for each. The present study employed a different methodology, utilizing a completely randomized design (CRD) experimental layout with 5 treatments and 2, 4, 6, 8, or 10 repetitions per treatment, resulting in total sample sizes of 10, 20, 30, 40, or 50 observations. Additionally, 10,000 iterations were used for each sample size to calculate the empirical type I error rate. Despite these methodological differences, similar conclusions to Ogunleye et al. (2018) were obtained in the present study for scenarios C1 and C3, representing conditions with homogeneous variances across treatments. This suggests that the findings regarding the performance of normality tests may be robust across different sample size configurations and iteration counts, specifically in scenarios with homogeneous variances. However, it is important to note that this conclusion cannot be extended to scenarios with heterogeneous variances among treatments.

Öztuna et al. (2006) also concluded that there was not a significant difference between the type I

error rates of the LI and SW tests. They varied the sample size from 5 to 200 and found relatively stable rates, a behavior pattern very similar to what was observed in the present study for scenarios with identical treatment variances (C1 and C3). However, as their experiment was not linked to any specific experimental design, they could not evaluate the influence of homogeneity (or heterogeneity) of treatment variances since the simulations were conducted for only one level of a factor (i.e., only one treatment). The present study extends this line of inquiry by employing a multi-treatment approach within a completely randomized design. This design allows for examining normality test performance across different treatments and, crucially, enables assessing how variance homogeneity or heterogeneity among treatments affects these tests. This approach provides insights into the robustness of normality tests under more complex experimental conditions. It offers a more comprehensive understanding of their behavior in practical research scenarios where multiple treatments with varying degrees of variance homogeneity are common.

Keskin (2006) reached similar conclusions to those mentioned above, as the type I error rate of the SW test was around 0.05 for all sample sizes ranging from 10 to 150. They conducted a total of 100,000 iterations for each sample size. However, even with this large number of iterations, it was still impossible to assess the influence of different experimental conditions on the results.

In contrast, this study was able to conclude that the heterogeneity of treatment variances has a significant effect on the significance level of almost all evaluated normality tests (AD, CVM, LI, and SW), both when the treatment means are equal (C2) and when the means are different (C4), as indicated by the results of the chi-squared tests for independence presented in Tables 2 and 3, respectively. These results confirm that when treatment variances are heterogeneous, type I error rates increase significantly, exceeding the threshold of 0.05, either if the treatment means are identical (C2) or different (C4). Under these conditions, the values of $\hat{\alpha}$ become increasingly higher as the number of repetitions per treatment increases. This inflation of type I error rates in heterogeneous variance scenarios might be attributed to the sensitivity of these tests' statistics to variance differences among treatments. For the Anderson-Darling (AD) and Cramér-von Mises (CVM) tests, which are based on the empirical distribution function, heterogeneous variances lead to more extreme values in the tails of the distribution, affecting the cumulative distribution function F(Yi) used in their test statistics. This results in larger deviations from the expected values under normality, inflating the test statistics. Similarly, for the Shapiro-Wilk (SW) test, variance heterogeneity distorts the relationship between the numerator and denominator of the test statistic, typically resulting in smaller W values that indicate greater deviation from normality. Consequently, these tests become more likely to reject the null hypothesis of normality when variances are heterogeneous, even if the underlying distribution is normal within each treatment. This sensitivity underscores the importance of considering the overall distribution shape and the variance structure when assessing normality in multi-treatment experimental designs.

Regarding the F test, the results of the present study were similar to those obtained by Nguyen et al. (2019), Kulkarni and Patil (2021), and Kelter (2021). Overall, it was found that when treatment variances are homogeneous (C1), the F test is highly effective, exhibiting empirical type I error rates ($\hat{\alpha}$) very close to the adopted theoretical significance level of 0.05. In contrast, when the treatment variances are heterogeneous (C2), the $\hat{\alpha}$ values exceeded 0.05 for all numbers of replications, indicating lower effectiveness of the F test. The influence of variance heterogeneity on the increase in type I error rates was further confirmed by significant results from the chi-squared tests for independence (Table 2).

Consistent with these findings, Nguyen et al. (2019) concluded that under conditions of variance homogeneity, the F test yielded satisfactory results regarding type I error rates, with the majority of values below or equal to 0.05. However, under variance heterogeneity conditions, the F test's performance was unsatisfactory, as the calculated type I error rates exceeded 0.05, particularly as the disparity between treatment variances increased. In such situations, non-parametric or semi-parametric methods, such as the Wilcox test proposed by Wilcox (1988) and Wilcox (1989), as well as the Welch test proposed by Welch (1951), exhibited better control over type I error rates compared to the F test. While the present study did not explicitly explore these alternative methods, the results suggest that their application might be beneficial in scenarios with high variance heterogeneity. However, it is important to note that the effectiveness of these non-parametric or semi-parametric approaches in the context of normality testing within multi-treatment experimental designs requires further investigation. Future research should focus on evaluating the performance of these alternative methods across various scenarios of variance heterogeneity and sample sizes to provide more comprehensive recommendations for practitioners dealing with heteroscedastic data in experimental settings.

Kulkarni and Patil (2021) arrived at a more general conclusion that many conventional hypothesis tests, including the F test, exhibit high type I error rates under specific parametric conditions, particularly when comparing multiple groups with small sample

sizes and when there is variance heterogeneity between the groups. To address this issue, Kulkarni and Patil (2021) proposed alternative tests based on integrated ratios between likelihood functions with respect to problematic parameters, aiming to reduce type I error rates under the most critical conditions. This approach involves integrating the ratio of likelihood functions over the parameter space of nuisance parameters (such as variances) rather than using point estimates. By doing so, the method accounts for the uncertainty in these parameters, particularly in small sample sizes, and provides a more robust test statistic. This integration-based approach differs from the standard F test by incorporating the full range of possible parameter values, potentially offering improved performance in scenarios where traditional methods struggle due to variance heterogeneity or small sample sizes. While promising, the applicability and effectiveness of this method in the context of normality testing within multi-treatment experimental designs would require further investigation to determine its potential benefits over conventional approaches.

Unlike the present study, Kelter (2021) investigated the behavior of hypothesis testing for two samples. He concluded that both frequentist hypothesis tests for two samples and their Bayesian counterparts have reduced type II errors and, consequently, increased type I errors as the sample size increased. Kelter (2021) also highlighted that frequentist tests exhibit higher type I error rates than their Bayesian counterparts. He also suggested that adopting a theoretical significance level lower than 0.05 could be considered to achieve lower type I error rates. However, this would inevitably lead to increased type II error rates.

The results from Kelter's (2021) study and the present work demonstrate that as sample sizes or numbers of replications per treatment increase, the samples from each treatment become more representative, and the variability within treatments increases. However, the experimental error variance estimator does not reflect this increase, as it is a weighted average that assigns the same weights to the variance within each treatment due to the assumption of variance homogeneity within treatments. Consequently, as this estimator is underestimated and appears in the denominator of the F test statistic, the values of F are higher than they should be, resulting in the incorrect rejection of the null hypothesis of equal means and an increased type I error rate. To mitigate this issue, several approaches could be considered: (1) using Welch's ANOVA, which does not assume equal variances; (2) applying variance-stabilizing transformations to the data before analysis; or (3) employing robust statistical methods that are less sensitive to violations of homogeneity assumptions. Additionally, mixed-effects models could account for both fixed and random effects in cases where the design allows, potentially providing a more nuanced analysis of the variance structure. Future research could focus on evaluating the effectiveness of these mitigation strategies in maintaining appropriate type I error rates across various scenarios of variance heterogeneity and sample sizes in the context of normality testing within multi-treatment experimental designs.

CONCLUSIONS

Kolmogorov-Smirnov test presents the lowest empirical rates $\hat{\alpha}$ of type I error in all scenarios. This superior performance is due to its conservative nature and lower sensitivity to variance heterogeneity compared to AD, CVM, and SW tests. While other tests are more affected by distortions in distribution tails and overall variance structure, the KS test's focus on maximum distribution differences provides greater robustness across various experimental conditions.

When there is the homogeneity of residual variances of the treatments, all normality tests studied and the F test present empirical rates of type I error ($\hat{\alpha}$) close to or lower than 0.05 and, therefore, satisfactory, unlike when these variances are heterogeneous.

There is a significant dependence relationship between the empirical significance levels ($\hat{\alpha}$) and the homogeneity condition (heterogeneity) of variances. Thus, when the standard deviations of the treatments are different, all studied hypothesis tests, except the KS, have a higher incidence of type I error.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

To Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – Código de Financiamento 001) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), for financing, in part, this study and to Federal University of Viçosa, for providing knowledge, human and computational resources.

REFERENCES

Acutis, M., Scaglia, B., & Confalonieri, R. (2012). Perfunctory analysis of variance in agronomy, and its consequences in experimental results interpretation. *European Journal of Agronomy* 43, 129-135.

Anderson, T.W. & Darling, D.A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics* 23, 193-212.

Arnastauskaitė, J., Ruzgas, T. & Bražėnas, M. (2021). An Exhaustive Power Comparison of Normality Tests. Mathematics 9, 788.

Barbetta, P. A., Reis, M. M., Bornia, A. C. (2004). *Estatística: para cursos de engenharia e informática.* (São Paulo, Atlas).

Bussab, W. O., Morettin, P. A. (2010). *Estatística Básica. 6ª ed.* (Editora Saraiva).

Casella, G. & Berger, R. L. (2022). *Statistical Inference. 2nd. ed* (Duxbury/Thomson Learning).

Campos, H. (1976). *Estatística Experimental Não-Paramétrica. 2ª ed.* (Piracicaba, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo).

Fukunaga, E. T., Guibu, I. A., Moraes, J. C. (2018). *Bases de Estatística para Profissionais de Saúde*. (Memnon: CEALAG- Centro de Estudos Augusto Leopoldo Ayrosa Galvão).

Henrique, F.H., Laca-Buendía, J.P. (2010). Comportamento morfológico e agronômico de genótipos de algodoeiro no município de Uberaba-MG. *FAZU em Revista* **7**, 32-36.

Kelter, R. (2021). Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality. *Computational Statistics* 36, 1263–1288.

Keskin, S. (2006). Comparison of Several Univariate Normality Tests Regarding Type I Error Rate and Power of the Test in Simulation based Small Samples. *Journal of Applied Science Research* 2, 296–300.

Kulkarni, H. V., Patil, S. M. (2021). Uniformly implementable small sample integrated likelihood ratio test for one-way and two-way ANOVA under heteroscedasticity and normality. *AStA Advances in Statistical Analysis* 105, 273–305.

Mood, A. M. (1974). *Introduction to the theory of statistics. 3. ed.* (McGraw-Hill, Inc).

Nguyen, D., Kim, E., Wang, Y., Pham, T. V., Chen, Y.H. & Kromrey, J. D. (2019). Empirical comparison of tests for one-factor ANOVA under heterogeneity and non-normality: A Monte Carlo study. *Journal of Modern Applied Statistical Methods* 18.

Ogunleye, L. I., Oyejola, B. A., Obisesan, K. O. (2018). Comparison of Some Common Tests for Normality. *International Journal of Probability and Statistics* 7, 130–137.

Öztuna, D.; Elhan, A. H., Tüccar, E. (2006). Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions. *Turkish Journal of Medical Sciences* 36, 171–176.

Piepho, H. P., & Edmondson, R. N. (2018). A tutorial on the statistical analysis of factorial experiments with qualitative and quantitative treatment factor levels. *Journal of Agronomy and Crop Science* 204, 429-455.

Razali, N.M. & Wah, Y.B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. Stat. Model Anal.* 2, 21-33.

Rodrigues, J., Piedade, S. M. E; Lara, I.A.R; Henrique, F.H. (2021). Type I error in multiple comparison tests in analysis of variance. *Acta Scientiarum* 45.

Searle, S. R., Gruber, M. H. J. (2016). *Linear Models. 2nd. ed. [S.l.]* (Wiley Series in Probability and Statistics).

Shapiro, S.S. & Wilk, M.B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52, 591-611.

Thadewald, T., Büning, H. (2007). Jarque-Bera Test and its Competitors for Testing Normality - A Power Comparison. *Journal of Applied Statistics* 34, 87–105.

Torman, V.B.L., Coster, R., Riboldi, J. (2012). Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. *Revista Clinical & Biomedical Research* 32.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika* 38, 330–336.

Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results on James's second-order method. *British Journal of Mathematical and Statistical Psychology* 41, 109–117.

Wilcox, R. R. (1989). Adjusting for Unequal Variances When Comparing Means in One-Way and Two-Way Fixed Effects ANOVA Models. *Journal of Educational and Behavioral Statistics* 14, 269–278.

To cite this paper, use:

Ribeiro Neto, H., Duarte, M.L. & Santos, N.T. (2025). Evaluation of empirical type I error rates of F and normality tests under different variance and mean conditions in multi-treatment CRDs. *Multi-Science Journal*, 8(1): 1-9. DOI: https://dx.doi.org/10.33837/msj.v8i1.1719