Brazilian Journal of Biometrics **43**, 1-13 (2025)
doi:10.28951/bjb.v43i4.760

ARTICLE

Empirical power of F and normality tests under different experimental conditions

Homero Ribeiro Neto*[1], Marciel Lelis Duarte[1] and Nerilson Terra Santos[1]
[1]Department of Statistics, Federal University of Viçosa, Viçosa-MG, Brazil
*Corresponding author. Email: homero.neto@ufv.br; ribeironetohomero@gmail.com

(Received: August 15,2024; Revised: June 10, 2025; Accepted: July 24, 2025; Published: September 10, 2025)

Abstract

The assumption of normality holds significant importance in inferential methods, such as the F-test of Analysis of Variance (ANOVA), which finds extensive application across various fields such as agricultural and clinical trials. Consequently, normality tests serve the purpose of evaluating the distribution of experimental errors for normality. However, prior studies aiming to compare the power of these tests should have considered the experimental design employed for simulated studies and assessed the impact of varying experimental conditions on the test powers. This study, therefore, focuses on assessing the effects of symmetry (or asymmetry) in the empirical distributions of the response variable for each treatment, the equality (or inequality) of their means, and the homogeneity (or heterogeneity) of their variances on the empirical power of both normality tests and the F-test, considering a Completely Randomized Design (CRD). To achieve this objective, normality tests were applied to 10,000 simulated sets of experimental residuals, while the F-test was applied to 10,000 simulated sets of response variable values. The findings of this study indicate that, in the majority of scenarios, power increases with an increasing number of replications per treatment. Furthermore, it was observed that the presence of symmetry tends to diminish the power of normality tests, while the F-test exhibits remarkable robustness to violations of normality assumptions. However, the power of the F-test can be influenced when the homogeneity of variances is compromised in conjunction with the asymmetry of non-normally distributed empirical data.

Keywords: Analysis of Variance (ANOVA); Completely Randomized Design (CRD); Normal Distribution; Experimental Errors.

1. Introduction

In the field of agricultural sciences, parametric inference plays a crucial role as a statistical tool, particularly in studies involving plants, such as analyzing grain yield (Souza *et al.*, 2023; Mwiinga *et al.*, 2020) or measuring plant cover (Wright *et al.*, 2017). Similarly, in clinical research, these inferential methods are essential for analyzing medical data and making evidence-based decisions (Kwak & Park, 2019). Some inferential procedures, such as the Analysis of Variance (ANOVA), require a normal distribution for experimental errors. Acutis *et al.* (2012)

highlight that ANOVA has been a standard method in agricultural sciences since the late 1930s, providing robust, probabilistic-based statistical analysis to support agronomists in reaching reliable information about scientific findings for technicians and farmers.

In ANOVA, the assumption of normality is primarily related to the F statistic, which is based on the Fisher-Snedecor F distribution (Fisher, 1925). The F statistic represents the ratio of two estimators of variance, that is, the estimator of variance between treatment means divided by the estimator of variance within treatments. The estimator of variance within treatments is known as residual variance, which estimates the variance of the experimental errors. Each one of these estimators follows a specific chi-square distribution, but both chi-squares require that the experimental errors follow a normal distribution. However, it is important to acknowledge that normality is not often evaluated in parametric inferences (Souza *et al.*, 2023; Knief and Forstmeier, 2021).

If the assumption of normality is not valid, then it is not possible to guarantee that the ratio between variance estimators will result in a statistic that follows an F distribution (Mood,1974; Casella and Berger, 2002). Consequently, the conclusion regarding the null hypothesis may be erroneous.

Hence, evaluating the assumption of normality of errors is a crucial step and can be accomplished through various methods. These methods include graphical techniques (e.g., histograms and boxplots), numerical approaches that utilize measures of skewness and kurtosis, and formal normality tests (Souza *et al.*, 2023; Razali & Wah, 2011). According to Shapiro and Wilk (1965) and Razali and Wah (2011), numerical methods and normality tests offer greater precision compared to graphical techniques.

One class of normality tests employed to verify the assumption of normality is the adherence test, such as Kolmogorov-Smirnov (KS), Lilliefors (LI), Cramér-von Mises (CVM), and Anderson-Darling (AD) tests. These tests compare the empirical distribution of the data to a theoretical normal distribution. A larger discrepancy observed between these cumulative distributions indicates a lower likelihood that the data are normally distributed, as noted by Anderson and Darling (1952).

On the other hand, the other class of normality tests compares two variance estimators: one specific to symmetric distributions (such as the normal distribution) and another for general distributions. One such test is the Shapiro-Wilk (SW) test (Shapiro and Wilk, 1965). This test computes the W statistic, which is a ratio between the variance estimator for the symmetric distribution (numerator) and the general variance estimator. The discrepancy between these estimators allows inference regarding the typical distribution of the data. A W statistic close to one indicates a high probability that the experimental error follows a normal distribution.

According to Souza *et al.* (2023), Patrício *et al.* (2016), González-Estrada and Cosmes (2019), the Shapiro-Wilk test (SW) is widely recommended because it can be used for small sample sizes (< 50 samples) and have greater power than adherence tests (Razali and Wah, 2011; Pino, 2014; Patrício *et al.*, 2016, Arnastauskaitė *et al.*, 2021; Uyanto, 2022). Many important agronomic studies have small sample sizes, such as corn (Singh *et al.*, 2023; Ullah *et al.*; 2023), rice (Li *et al.*, 2023), soybean (Karges *et al.*, 2022; Pierozan Junior *et al.*, 2023), sugarcane (Kölln *et al.*, 2022) and wheat (Basso *et al.*, 2013; Mizuta *et al.*, 2023).

Nevertheless, apart from the sample size, other experimental conditions may influence the power of normality tests and the F-test itself. Previous studies have primarily focused on sample size and consistently shown that test power increases as the sample size increases (Confalonieri *et al.*, 2007; Anderson *et al.*, 2017; Doulah, 2019; Arnastauskaitė *et al.*, 2021; Islam, 2021; Uyanto, 2022). Nonetheless, these studies were conducted using simulated sample sizes generally much larger than those typically encountered in agricultural experiments and did not explore the effects

of conditions such as symmetry (or asymmetry) of empirical data distributions and equality (or inequality) of means across treatment groups, as well as the homogeneity (or heterogeneity) of variances, on the power of normality tests and the F-test.

Given this context, studies like Nguyen *et al.* (2019) emphasize that the F-test is significantly sensitive to violations of the homogeneity of variances assumption. However, the impact of such violations on the test's power, when combined with other experimental conditions, remains unclear.

Therefore, the present study aims to evaluate the performance of normality tests and the F-test when applied to sample sizes representative of real agricultural, medical, and other applied sciences experiments. Additionally, evaluates the influence of symmetry (or asymmetry) of non-normal empirical distributions, equality (or inequality) of means, and homogeneity (or heterogeneity) of variances on the empirical power of normality tests and the F-test.

2. Materials and Methods

Data sets of a Completely Randomized Design (CRD) with five treatments and k replications were simulated under different scenarios such that the response variable follows different inverse gamma probability distributions. The chosen parameters, α and β (Gelman *et al.*, 2013), of these distributions yielded the scenarios C1, C2, C3, C4, C5, C6, C7, and C8, as presented in Table 1. The differences among these scenarios are due to distribution symmetry (asymmetry), equality (inequality) of treatment means, and equality (inequality) of treatment variances. For this study, approximately symmetric inverse gamma distributions were considered symmetric, as they do not differ significantly from a symmetric one. Furthermore, for the simulation of scenarios with different treatment means, their ratio was 1:2:3:4:5. Similarly, for scenarios with different treatment variances, their standard deviation ratio was also 1:2:3:4:5.

Five sub-scenarios were simulated for each of the eight scenarios, with each sub-scenario characterized by a number of k = 2, 4, 6, 8, or 10 replications per treatment (Figure 1). Therefore, 40 sub-scenarios were evaluated in this study. For each of these sub-scenarios, 10,000 iterations were simulated. For each iteration $\boldsymbol{w}$, such that $\boldsymbol{w}$=1,2,...,10,000, a $\boldsymbol{w}$ set of 5k residuals was obtained, according to Equation 1, where $\hat{\boldsymbol{\varepsilon}}_{ijw}$, considering the CRD statistical model, is the residual obtained for the observed value $\boldsymbol{y}_{ijw}$ of the response variable on iteration $\boldsymbol{w}$ for replication $\boldsymbol{j}$ of treatment $\boldsymbol{i}$:

$$\hat{\boldsymbol{\varepsilon}}_{ijw}=\boldsymbol{y}_{ijw}-\hat{\boldsymbol{\mu}}_{iw} \tag{1}$$

Therefore, for each one of the 10,000 iterations, a residual data set, composed of 5k residuals, was obtained. Each normality test (KS, LI, CVM, AD, and SW), which has as null hypothesis that experimental errors follow a normal distribution, was applied to each residual data set.

Under α=0.05, an empirical power $\widehat{\boldsymbol{P}}$ for each normality test was computed by Equation 2:

$$\widehat{\boldsymbol{P}} = \frac{\boldsymbol{number\ of\ p-values \leq 0.05}}{\boldsymbol{10.000}} \tag{2}$$

For a given normality test, the number of $\boldsymbol{p-values}$ was the number of $\boldsymbol{w}$ residuals data sets that yielded rejection of the null hypothesis of normality (α=0.05).

Unlike the normality tests, the F-test was applied to the response variable values, only in scenarios where the treatment means were unequal, specifically for C3, C4, C7, and C8 (Table 1). However, their empirical powers ($\widehat{\boldsymbol{P}}$) were also calculated using Equation 2, considering as null hypothesis, the equality of the five treatments means and a 5% significance level.

For each sub-scenario, each normality test and F-test had its empirical power classified as:

- Not powerful, if $\widehat{\boldsymbol{P}} < 0.75$;
- Powerful, if $\widehat{\boldsymbol{P}} \geq 0.75$.

The independence between the classification of empirical power and experimental conditions (symmetry of distribution, equality of means, and equality of variance) was evaluated by a chi-square test, according to Siegel and Castellan Jr. (2006).

Software R, version 4.0.2 (R Core Team, 2020), was used to obtain all simulated data and evaluate all normality and F tests. The simulations and statistical analyses were conducted using a suite of packages, including “tidyverse”, “xlsx”, “car”, “GAD”, “PMCMRplus”, “DescTools”, “outliers”, “stats”, “coin”, “dplyr”, “nortest”, “onewaytests”, “invgamma”, “nimble” and “moments”. The graphical representations were generated using the "lattice" and "dplyr" packages.

2.1 Figures and tables

Figure 1 illustrates the organizational chart for each simulated scenario pattern and sub-scenario.

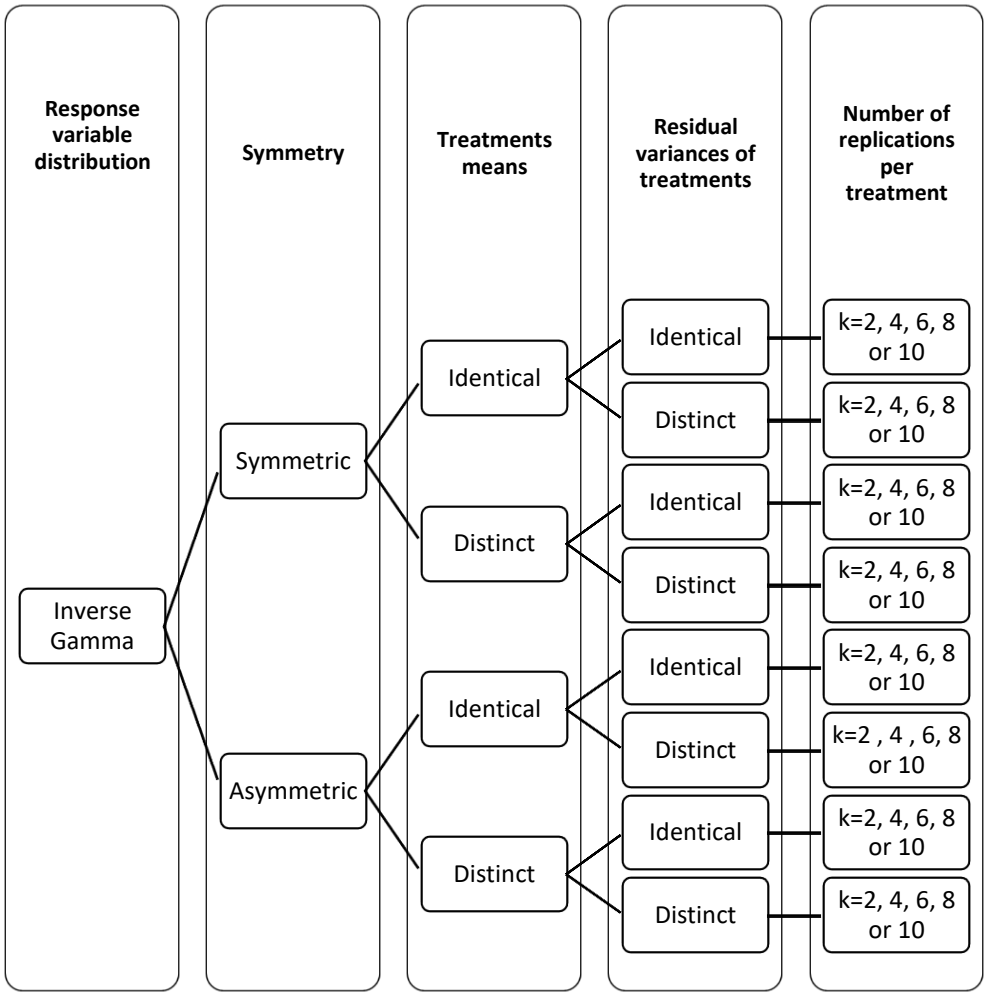

Figure 1. Organization chart of each scenario pattern and sub-scenario, under the CRD, defined by an inverse gamma probability distribution of the response variable, considered symmetric or asymmetric distribution, treatment means equal or unequal, treatment variances equal or unequal, and k replications per treatment.

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

Table 1. Values of means and standard deviations for the treatments were established to simulate scenarios considering different inverse gamma distributions

Inverse Gamma	Mean	Standard Deviation	Scenario
Asymmetric	$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = 0{,}25$	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 0{,}144$	C1
		$\sigma_1 = 0{,}144$; $\sigma_2 = 0{,}288$; $\sigma_3 = 0{,}432$; $\sigma_4 = 0{,}576$; $\sigma_5 = 0{,}720$	C2
	$\mu_1 = 0{,}25$; $\mu_2 = 0{,}5$; $\mu_3 = 0{,}75$; $\mu_4 = 1{,}00$; $\mu_5 = 1{,}25$	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 0{,}144$	C3
		$\sigma_1 = 0{,}144$; $\sigma_2 = 0{,}288$; $\sigma_3 = 0{,}432$; $\sigma_4 = 0{,}576$; $\sigma_5 = 0{,}720$	C4
Symmetric	$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = 100$	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 1$	C5
		$\sigma_1 = 1$; $\sigma_2 = 2$; $\sigma_3 = 3$; $\sigma_4 = 4$; $\sigma_5 = 5$	C6
	$\mu_1 = 100$; $\mu_2 = 200$; $\mu_3 = 300$; $\mu_4 = 400$; $\mu_5 = 500$	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 = 1$	C7
		$\sigma_1 = 1$; $\sigma_2 = 2$; $\sigma_3 = 3$; $\sigma_4 = 4$; $\sigma_5 = 5$	C8

3. Results

Empirical power of each normality and F-test (the latter, in scenarios with distinct treatment means), as a function of the number of replications per treatment, obtained from simulated iterations of asymmetric distributions scenarios C1, C2, C3, and C4 (Table 1), is presented in Figure 2. A threshold line at $\widehat{\boldsymbol{P}} = 0.75$ is plotted in each figure to distinguish between powerful and not powerful tests at each replication number k.

Under asymmetric distribution with identical means and identical standard deviations (C1), the empirical power of all normality tests linearly increased as the number of replications (k) per treatment increased (Figure 2 (a)). It should also be noted that the KS test showed the lowest empirical power at all k. This result suggests a highly conservative behavior of this test. In most iterations, the KS test did not reject the hypothesis of normality when it should have because the simulated data follows an inverse gamma distribution. For all normality tests under scenario C1 (Figure 2 (a)), the greatest empirical power was observed for the highest number of replications (k=10), representing a total of 50 observations for the experiment. Other than KS, all normality tests were classified as powerful for such replications. For k=8, the LI was not powerful. For k=2 (10 observations), k=4 (20 observations), and k=6 (30 observations), no test was classified as powerful according to the empirical powerfulness criteria.

As in scenario C1, an increase in the empirical power of all normality tests, except for the KS test, was observed in C3 as k increases (Figure 2(b)). However, their empirical power values were considerably smaller than those observed for C1 (Figure 2 (a)), and no normality test was classified as powerful at any number of replications, not even k=10. The only difference between scenarios C1 and C3 is that the treatment means were equal in C1 and not in C3. However, the independence chi-squared test between equality (not equality) of treatment means and powerfulness (not powerfulness) for each normality test was not significant (Table 2).

Under scenario C3, the F-test was powerful ($\widehat{\boldsymbol{P}} \geq 0.75$) even for a small number of replicates per treatment (k=2) (Figure 2 (b)). This result indicates the robustness of the F-test when the assumption of normality is not valid. However, it is worth mentioning that in C3, although the treatment means were not equal, the within-treatment variances were equal.

From the results of simulations for scenario C2 (Figure 2 (c)), when treatment means are identical but within treatment variances are different and k≥6, practically, all normality tests, except KS test, were powerful ($\widehat{\boldsymbol{P}} \geq 0.75$). As observed for scenarios C1 and C3, the empirical power of the normality tests under C2 increases as k increases.

In general, C2 showed higher empirical powers than C1 for all k. However, according to the chi-squared test (Table 3), there is no significant dependence ($\alpha = 0.05$) between power and within-treatment homogeneity of variance under scenarios simulated for asymmetric distributions

with equal treatment means.

Finally, for asymmetric inverse gamma distributions, Figure 2 (d) also shows the empirical power of scenario C4 (not equal treatment means and not equal within treatment variances). An increase in empirical power for all normality tests (AD, CVM, LI, SW, and KS) can be observed as the number of replications increases. For K≥8, all normality tests were powerful, except KS, with a noticeable tie between the most powerful ones (SW, AD, and CVM). On the other hand, the F-test was powerful if k>4, regardless of the homogeneity of variances, for scenarios with asymmetric distributions (C3 and C4).

Furthermore, under asymmetric distributions with different treatment means, F-test empirical power was lower in scenario C4 (with heterogeneity of variances) than in scenario C3 (with homogeneity of variances). This may be due to the violation, in C4, of the variance homogeneity assumption required for using the F-test.

When symmetric inverse gamma distributions with equal variances were selected for simulations, for either case of treatment means, equal (C5) or not equal (C7), no normality test was classified as powerful, showing empirical power near zero for every number of replications per treatment (Figure 3 (a) and (b)). However, under scenario C7, the F-test was classified as powerful for every number of replications per treatment (Figure 3 (b)).

The chi-square independence test between the powerfulness class (powerful or not powerful) and symmetry (asymmetry) of the simulated distributions was significant for most of the normality tests (Table 4). Such results explain the decreased empirical power observed compared to results obtained from simulations of asymmetric (Figure 2) and symmetric (Figure 3) inverse gamma distributions.

Moreover, the pattern observed for the empirical power of the F-test (Figures 2 and 3) indicates relative robustness, as the empirical power of the F-test was very close to 1 under sub-scenarios C3, C7, and C8.

The empirical power of most normality tests increased as the number of replications per treatment also increased for scenarios simulated using a symmetric inverse gamma distribution with unequal within-treatment variances, for both equal (C6) and unequal (C8) treatment means (Figure 3 (c) and (d)). Besides this increase, none of the normality tests were classified as powerful. It is worth mentioning that the empirical power of KS test remained remarkably stable and close to zero.

Although there is a clear pattern difference, in Figure 3, between scenarios simulated with equal (C5 and C7) and unequal variances (C6 and C8), it was not possible to test the significance of the independence between powerfulness class and equality of variance because there were some cells with zero frequency on the contingency table needed for chi-square independent test.

Furthermore, under symmetric inverse gamma distributions with different means and standard deviations (Figure 3 (d)), no normality test was powerful ($\widehat{\boldsymbol{P}}$<0.75), once again demonstrating how these tests were sensitive to the symmetry condition.

3.1 Figures and tables

Figure 2 shows the results of scenarios C1, C2, C3, and C4.

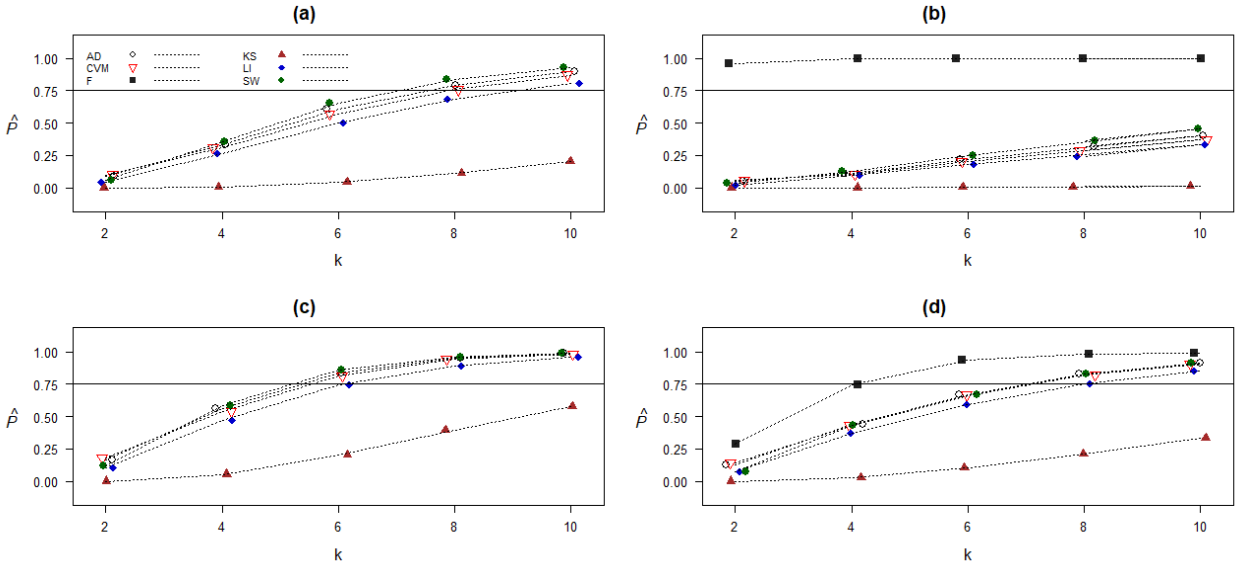

Figure 2. Empirical power ($\hat{P}$) of F-test for scenarios C3 (b) and C4 (d) and Anderson-Darling (AD), Kolmogorov-Smirnov (KS), Crámer-von Mises (CVM), Lilliefors (LI) and Shapiro-Wilk (SW) tests for scenarios C1 (a), C2 (c), C3 (b), and C4 (d) as a function of number of replications (k) per treatment obtained from simulated asymmetric inverse gamma distributions and empirical power threshold value at $\hat{P}$=0.75.

Figure 3 shows the results of scenarios C5, C6, C7, and C8.

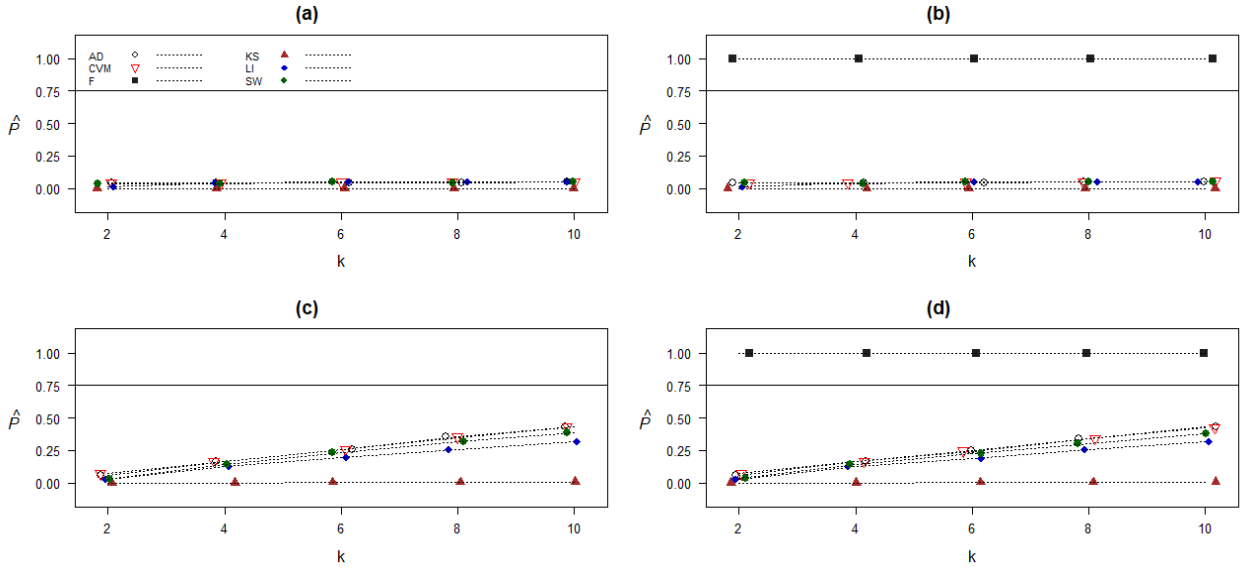

Figure 3. Empirical power ($\hat{P}$) of F-test for scenarios C7 (b) and C8 (d) and Anderson-Darling (AD), Kolmogorov-Smirnov (KS), Crámer-von Mises (CVM), Lilliefors (LI) and Shapiro-Wilk (SW) tests for scenarios C5 (a), C6 (c), C7 (b), and C8 (d) as a function of number of replications (k) per treatment obtained from simulated asymmetric inverse gamma distributions and empirical power threshold value at $\hat{P}$=0.75.

Tables 2, 3, and 4 show chi-square independence test results.

Table 2. Results of chi-square independence tests between classes of empirical power (powerful for $\hat{P}\geq 0.75$ and not powerful for $\hat{P}<0.75$) and equality (inequality) of treatment means, for Anderson-Darling (AD), Kolmogorov-Smirnov (KS), Crámer-von Mises (CVM), Lilliefors (LI) and Shapiro-Wilk (SW) normality tests, according to simulations of scenarios C1 and C3. Values in bold indicate significance (p-value≤0.05)

Test	P-value
AD	0.1138
KS	-
CVM	0.1138
LI	0.2918
SW	0.1138

Table 3. Results of chi-square independence tests between classes of empirical power (powerful for $\hat{P}\geq 0.75$ and not powerful for $\hat{P}<0.75$) and within treatment homogeneity (heterogeneity) of variance, for Anderson-Darling (AD), Kolmogorov-Smirnov (KS), Crámer-von Mises (CVM), Lilliefors (LI) and Shapiro-Wilk (SW) tests, according to simulations of scenarios C1 and C2. Values in bold indicate significance (p-value≤ 0.05)

Test	P-value
AD	0.5271
KS	-
CVM	0.5271
LI	0.4902
SW	0.5271

Table 4. Results of chi-square independence tests between classes of empirical power (powerful for $\hat{P}\geq 0.75$ and not powerful for $\hat{P}<0.75$) and symmetry (asymmetry), for Anderson-Darling (AD), Kolmogorov-Smirnov (KS), Crámer-von Mises (CVM), Lilliefors (LI) and Shapiro-Wilk (SW) tests, according to simulations of scenarios C1, C2, C3, C4, C5, C6, C7, and C8. Values in bold indicate significance (p-value≤ 0.05)

Test	P-value
AD	**0.01253**
KS	-
CVM	**0.01253**
LI	0.05583
SW	**0.01253**

4. Discussion

The reported results were obtained by adopting a methodology that differs significantly from previous studies that evaluated normality tests in two key aspects. This study used sample sizes much closer to practical, real-world sample sizes (<50 observations), in contrast to the thousands of observations employed in previous research. Furthermore, in this study, all normality tests were evaluated based on experimental errors rather than response values, as was done in prior studies. The experimental errors were computed based on the experimental design used to simulate the response values. While this study focused on Completely Randomized Design (CRD), future research should extend this methodological approach to other experimental designs, such as Randomized Block Design (RBD) and Latin Square Design (LSD), evaluating how treatment means equality (inequality), within treatment variances homogeneity (heterogeneity), and symmetry (asymmetry) of response variable distributions behave under these alternative experimental arrangements.

Unlike previous studies that did not employ experimental designs, this study assessed the effects of the equality or inequality of treatment means, as well as the homogeneity or heterogeneity of treatment variances, which are required for the Analysis of Variance (ANOVA). Despite these differences, the results of this study align with those of Arnastauskaitė *et al.* (2021), Razali and Wah (2011), Ogunleye *et al.* (2018), Doulah (2019), Torman *et al.* (2012), and Uyanto (2022), which showed a similar pattern of increasing empirical power as the number of observations increased, particularly when the data originated from asymmetric probability distributions for AD, CVM, KS, LI, and SW normality tests (see Figure 2). However, this pattern was not observed in this study when the residuals were obtained from scenarios simulated with symmetric inverse gamma distributions, either with equal treatment means and within-treatment variances (C5) or unequal treatment means and equal within-treatment variances (C7), as depicted in Figure 3. All normality tests exhibited stable and extremely low empirical power in these scenarios. Nevertheless, scenarios with symmetric inverse gammas and heterogeneity of variances (C6 and C8, Figure 3) demonstrated slightly higher power compared to C5 and C7 but still lower than the scenarios with asymmetric inverse gammas (Figure 2). These findings concur with those of Farrell and Stewart (2006), who stated that normality tests have lower power when empirical distributions are symmetric with short tails (C5 and C7) and higher power only when the curves have longer tails (C6 and C8).

It is worth mentioning that Farrell and Stewart (2006) concluded that the modification proposed in the SW test by Rahman and Govindarajulu (1997) increases the power of the SW test when the empirical distribution of the data is symmetric with a short tail. However, these results were not associated with any experimental design, as in the present study. Future studies could investigate whether the modification proposed by Rahman and Govindarajulu (1997) effectively increases empirical power when experimental errors are used instead of response values under symmetric scenarios. Additionally, it could be evaluated whether similar modifications enhance the power of other normality tests (AD, CVM, LI, and KS) in these cases, given their previous ineffectiveness.

Among all the normality tests evaluated in this study, the KS test exhibited the lowest power across all scenarios and replication numbers per treatment. These findings are consistent with those of Ogunleye *et al.* (2018), Torman *et al.* (2012), and Uyanto (2022), who observed a conservative pattern for the KS test.

This study also concludes that when the assumption of homogeneity of treatment variances holds true, the F-test is powerful for all sample sizes and effectively rejects the hypothesis of equality of treatment means. This finding aligns with the results of Nguyen *et al.* (2019), who demonstrated that, under homogeneity of treatment variances, the F-test outperforms non-parametric methods in assessing equality of means.

Like the present study, Nguyen *et al.* (2019) observed reduced power of the F-test when the homogeneity of variance is not met and the number of replications per treatment is low.

However, the robustness of the F-test is evident in this study, as it exhibited considerable power even when normality assumptions were violated. These results support the conclusions of Knief and Forstmeier (2021) that parametric tests, such as the F-test, are robust to violations of the normality assumption if there is no substantial asymmetry in the distribution of experimental data.

In this study, the F-test displayed robustness and stability across all replication numbers per treatment in C7 and C8. Nevertheless, the only scenario in which the F-test lacked power was C4 ($k < 4$), characterized by a combination of asymmetry in inverse gamma distributions and heterogeneity of within-treatment variances, indicating an interaction between these two experimental conditions that resulted in reduced F-test power. None of the scenarios with symmetry exhibited a lack of power for the F-test. In this regard, considering only symmetric inverse gamma distributions, although a chi-squared test was not employed, it is evident that there is no dependence between the power classification and the homogeneity or heterogeneity of variances for the F-test, as the empirical powers of the F-test were consistently equal to 1 for every

k replications per treatment for both scenarios C7 and C8.

Based on our findings, we propose the following practical guidelines for researchers applying these statistical methods in real-world settings. When dealing with experimental data analysis, (1) Researchers should be particularly cautious with normality test results under two specific conditions: when the experimental error distribution is symmetric or the number of replications per treatment is less than 8 (or the total experimental errors are fewer than 40). Under these conditions, the empirical power of normality tests is substantially reduced. (2) For ANOVA applications, the F-test power is compromised only when there is a simultaneous occurrence of heterogeneous within-treatment variances, asymmetric distribution of experimental errors, and fewer than 4 replications per treatment. Outside of this specific combination, the F-test demonstrates considerable robustness. (3) The F-test performs particularly well with symmetric error distributions, maintaining high power regardless of whether variances are homogeneous or heterogeneous. (4) For experimental design planning, we recommend a minimum of 8 replications per treatment to ensure adequate power for normality testing, although the F-test may perform reliably with fewer replications under symmetric distributions. When these minimum requirements cannot be met, complementary diagnostic tools, such as graphical methods (e.g., Q-Q plots) and descriptive statistics (skewness and kurtosis measures), may provide valuable additional information for data analysis decisions.

5. Conclusions

It can be concluded that, in general, the empirical power of the normality tests is considerably lower when:

- the distribution of the errors is symmetric;
- the within-treatment variances are homogeneous (comparing scenarios under the same symmetry and treatment means conditions);
- the treatment means are not equal (considering the asymmetric distribution of the errors and comparing scenarios under the same variance condition);
- the number of replications per treatment is smaller than 8, or the number of experimental errors is smaller than 40;

Also, it can be concluded that the empirical power of the F-test:

- is lower when the within-treatment variances are heterogeneous, the distribution of experimental errors is asymmetric, and the number of replications per treatment is smaller than 4;
- is higher under symmetric error distribution for either homogeneous or heterogeneous treatment variances.

Acknowledgments

The work was mainly funded by the Brazilian Federal Agency “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (CAPES – Código de Financiamento 001), Brazil. We would also like to thank the “Fundação de Amparo à Pesquisa do Estado de Minas Gerais” (FAPEMIG) for funding part of the research, too.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Author Contributions

Conceptualization: RIBEIRO NETO, H.; SANTOS, N. T. **Data curation:** RIBEIRO NETO, H.; SANTOS, N. T.. **Formal analysis:** RIBEIRO NETO, H.; SANTOS, N. T.; DUARTE, M. L. **Funding acquisition:** SANTOS, N. T.. **Investigation:** RIBEIRO NETO, H.; SANTOS, N. T.. **Methodology:** RIBEIRO NETO, H.; SANTOS, N. T.. **Project administration:** SANTOS, N. T.. **Resources:** SANTOS, N. T.. **Supervision:** SANTOS, N. T.. **Visualization:** RIBEIRO NETO, H.; SANTOS, N. T.; DUARTE, M. L.. **Writing – original draft:** RIBEIRO NETO, H.; SANTOS, N. T.; DUARTE, M. L.. **Writing – review & editing:** RIBEIRO NETO, H.; SANTOS, N. T.; DUARTE, M. L..

References

1. Acutis, M., Scaglia, B., & Confalonieri, R. Perfunctory analysis of variance in agronomy, and its consequences in experimental results interpretation. *European Journal of Agronomy* **43,** 129-135 (2012). https://doi.org/10.1016/j.eja.2012.06.006
2. Anderson, T.W. & Darling, D.A. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics* **23,** 193-212 (1952). https://doi.org/10.1214/aoms/1177729437
3. Anderson, S.F., Kelley, K. & Maxwell, S.E. Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological Science* **28,** 1547–1562 (2017). https://doi.org/10.1177/0956797617723724
4. Arnastauskaitė, J., Ruzgas, T. & Bražėnas, M. An Exhaustive Power Comparison of Normality Tests. Mathematics **9,** 788 (2021). https://doi.org/10.3390/math9070788
5. Basso, B., Cammarano, D., Fiorentino, C., & Ritchie, J.T. Wheat yield response to spatially variable nitrogen fertilizer in Mediterranean environment. *European Journal of Agronomy* **51,** 65-70 (2013). https://doi.org/10.1016/j.eja.2013.06.007
6. Casella, G. & Berger, R. L. *Statistical Inference. 2nd. ed* (Duxbury/Thomson Learning, 2022).
7. Confalonieri, R., Acutis, M., Bellocchi, G. & Genovese, G. Resampling-based software for estimating optimal sample size. *Environmental Modelling & Software* **22,** 1796-1800 (2007). https://doi.org/10.1016/j.envsoft.2007.02.006
8. Doulah, M.S.U. A Comparison among Twenty-Seven Normality Tests. *Research & Reviews: Journal of Statistics* **8,** 41–59 (2019).
9. Farrell, P. & Stewart, K.R. Comprehensive study of tests for normality and symmetry: Extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation* **76,** 803–816 (2006). https://doi.org/10.1080/10629360500109023
10. Fisher, R.A. *Statistical Methods for Research Workers.* In: Kotz, S., Johnson, N.L. (eds) *Breakthroughs in Statistics* (Springer, 1992). https://doi.org/10.1007/978-1-4612-4380-9_6
11. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D., Vehtari, A., & Rubin, D. *Bayesian Data Analysis* (CRC Press, 2013).
12. Gonzalez-Estrada, E. & Cosmes, W. Shapiro-Wilk test for skew normal distributions based on data transformations. *J. Stat. Comput. Simul* **89,** 3258–3272 (2019). https://doi.org/10.1080/00949655.2019.1658763
13. Islam, T.U. Min-max approach for comparison of univariate normality tests. *PLOS ONE* **16,** (2021). https://doi.org/10.1371/journal.pone.0255024
14. Karges, K., Bellingrath-Kimura, S. D., Watson, C. A., Stoddard, F. L., Halwani, M., & Reckling, M. Agro-economic prospects for expanding soybean production beyond its current northerly limit in Europe. *European Journal of Agronomy* **133,** (2022). https://doi.org/10.1016/j.eja.2021.126415

15. Knief, U. & Forstmeier, W. Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods* **53,** 2576–2590 (2021). https://doi.org/10.3758/s13428-021-01587-5
16. Kölln, O., Boschiero, B., Franco, H., Soldi, M., Sanches, G., Castro, S. & Trivelin, P. Preferential mineral N form uptake by sugarcane genotypes contrasting in nitrogen use efficiency. *Experimental Agriculture* **58,** (2022). http://dx.doi.org/10.1017/S0014479722000229
17. Kwak, S.G. & Park, S.H. Normality Test in Clinical Research. Journal of Rheumatic Diseases 26, 5-11 (2019). https://doi.org/10.4078/jrd.2019.26.1.5
18. Li, L., He, L., Li, Y., Wang, Y., Ashraf, U., Hamoud, Y. A., Hu, X., Wu, T., Tang, X. & Pan, S. Deep fertilization combined with straw incorporation improved rice lodging resistance and soil properties of paddy fields. *European Journal of Agronomy* **142,** (2023). https://doi.org/10.1016/j.eja.2022.126659
19. Mizuta, K., Araki, H. & Takahashi, T. Relationship between canopy coverage at the initiation of stem elongation and lodging in wheat. *European Journal of Agronomy* **148,** (2023). https://doi.org/10.1016/j.eja.2023.126855
20. Mood, A. M. *Introduction to the theory of statistics. 3. ed.* (McGraw-Hill, Inc, 1974).
21. Mwiinga, B., Sibiya, J., Kondwakwenda, A., Musvosvi, C. & Chigeza, G. Genotype x environment interaction analysis of soybean (Glycine max (L.) Merrill) grain yield across production environments in Southern Africa. *Field Crops Research* **256,** (2020). https://doi.org/10.1016/j.fcr.2020.107922
22. Nguyen, D., Kim, E., Wang, Y., Pham, T. V. & Chen, Y.H. Empirical comparison of tests for one-factor ANOVA under heterogeneity and non-normality: A Monte Carlo study. *Journal of Modern Applied Statistical Methods* **18,** (2019). https://dx.doi.org/10.22237/jmasm/1604190000
23. Ogunleye, L. I., Oyejola, B. A., Obisesan, K. O. Comparison of Some Common Tests for Normality. *International Journal of Probability and Statistics* **7,** 130–137 (2018). https://doi.org/10.5923/j.ijps.20180705.02
24. Patrício, M., Ferreira, F., Oliveiros, B. & Caramelo, F. Comparing the performance of normality tests with ROC analysis and confidence intervals. *Commun. Stat. Simul. Comput.* **46,** 7535–7551 (2016). http://dx.doi.org/10.1080/03610918.2016.1241410
25. Pierozan Junior, C., Favarin, J., Baptistella, J., De Almeida, R., Maciel de Oliveira, S., Lago, B., & Tezotto, T. Controlled release urea increases soybean yield without compromising symbiotic nitrogen fixation. *Experimental Agriculture* **59,** (2023). http://dx.doi.org/10.1017/S0014479722000540
26. Pino, F.A. The question of non-normality: a review. *Rev. De. Econ. Agríc.* **61,** 17-33 (2014).
27. R CORE TEAM. R: A language and environment for statistical computing. R, -Foundation for Statistical Computing, Vienna, Austria, 2020.
28. Rahman, M.M. & Govindarajulu, Z. A modification of the test of Shapiro and Wilk for normality. *Journal of Applied Statistics* **24,** 219–235 (1997). https://doi.org/10.1080/02664769723828
29. Razali, N.M. & Wah, Y.B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. Stat. Model Anal.* **2,** 21-33 (2011).
30. Shapiro, S.S. & Wilk, M.B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **52,** 591-611 (1965). https://doi.org/10.2307/2333709
31. Siegel, S. & Castellan JR., N.J. *ESTATÍSTICA NÃO-PARAMÉTRICA PARA CIENCIAS DO COMPORTAMENTO. 2.ed. ed.* (ARTMED EDITORA S.A, 2006).
32. Singh, R., Sawatzky, S.K., Thomas, M., Akin, S., Zhang, H., Raun, W. & Arnall, D.B. Nitrogen, Phosphorus, and Potassium Uptake in Rain-Fed Corn as Affected by NPK Fertilization. *Agronomy* **13,** (2023). https://doi.org/10.3390/agronomy13071913

33. Souza, R.R., Toebe, M., Mello, A.C., & Bittencourt, K.C. Sample size and Shapiro-Wilk test: An analysis for soybean grain yield. *European Journal of Agronomy* **142,** (2023). https://doi.org/10.1016/j.eja.2022.126666
34. Souza, R.R., Toebe, M., Marchioro, V.S., Filho, A.C., Mello, A.C., Manfio, G.L., Soldateli, F.J., Soares, S., Martins, V. & Junges, D.L. Soybean grain yield in highland and lowland cultivation systems: A genotype by environment interaction approach. *Annals of Applied Biology* **179,** 302-318 (2021). https://doi.org/10.1111/aab.12709
35. Torman, V.B.L., Coster, R., Riboldi, J. Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. *Revista Clinical & Biomedical Research* **32,** (2012). https://seer.ufrgs.br/index.php/hcpa/article/view/29874
36. Ullah, J., Chen, S., Ruan, Y., Ali, A., Khan, N.M., Rehman, M.N.U., & Fan, P. Combined Di-Ammonium Phosphate and Straw Return Increase Yield in Sweet Corn. *Agronomy* **13,** 1885 (2023). https://doi.org/10.3390/agronomy13071885
37. Uyanto, S.S. An Extensive Comparisons of 50 Univariate Goodness-of-fit Tests for Normality. *Austrian Journal of Statistics* **51,** 45–97 (2022). https://doi.org/10.17713/ajs.v51i3.1279
38. Wright, W.J., Irvine, K.M., Warren, J.M. & Barnett, J.K. Statistical design and analysis for plant cover studies with multiple sources of observation errors. *Methods Ecol. Evol.* **8,** 1832–1841 (2017). https://doi.org/10.1111/2041-210X.12825