TecnoLógicas

Simulation Study on the Power and Sensitivity of Sixteen Normality Tests Under Different Non-Normality Scenarios

Estudio de simulación sobre la potencia y sensibilidad de dieciséis pruebas de normalidad en distintos escenarios de no normalidad

Cristian David Correa-Álvarez[1]; Jessica María Rojas-Mora[2]; Antonio Elías Zumaqué-Ballesteros[3]; Osnamir Elias Bru-Cordero[4]

[1]Universidad Nacional de Colombia, Sede Manizales, Facultad de Ciencias Exactas y Naturales, Manizales-Colombia
[2]Instituto Tecnológico Metropolitano, Medellín-Colombia
[3]Universidad de Córdoba, Montería-Colombia
[4]Universidad Nacional de Colombia, Sede De La Paz, Dirección Académica, La Paz-Colombia

Correspondence: crdcorreaal@unal.edu.co

Received: 18 October 2024 **Accepted:** 05 March 2025 **Available:** 31 March 2025

How to cite / Cómo citar

C. D. Correa-Álvarez, J. M. Rojas-Mora, A. E. Zumaqué-Ballesteros, and O. E. Bru-Cordero, "Simulation Study on the Power and Sensitivity of Sixteen Normality Tests Under Different Non-Normality Scenarios," *TecnoLógicas*, vol. 28, no. 62, e3293, 2025. https://doi.org/10.22430/22565337.3293

Abstract

In data analysis, validating the normality assumption is crucial for determining the suitability of applying parametric methods. The objective of this research was to compare the power and sensitivity of sixteen normality tests, classified according to various aspects. The methodology involved simulating data using the Fleishman contamination system. This approach allowed us to evaluate the tests under non-normality conditions across ten distributions with varying degrees of deviation from normality. The results obtained showed that tests based on correlation and regression, such as Shapiro-Wilk and Shapiro-Francia, outperform the others in power, especially for large samples and substantial deviations from normality. For moderate deviations, the D'Agostino-Pearson and skewness tests performed well, while for low deviations, the Robust Jarque-Bera and Jarque-Bera tests were the most effective. Additionally, some tests exhibited high power across multiple distribution types, such as Snedecor-Cochran and Chen-Ye, which performed well for both symmetric platykurtic and asymmetric leptokurtic distributions. These findings offer valuable insights for selecting appropriate normality tests based on sample characteristics, which improves the reliability of statistical inference. Finally, it is concluded that this research demonstrates scenarios in which the most commonly used statistical tests are not always the most effective.

Keywords

Distribution classification method, Fleishman's method, Monte Carlo simulation, normality tests, power comparison.

Resumen

En el análisis de datos, la validación del supuesto de normalidad es crucial para determinar si es correcto aplicar métodos paramétricos. El objetivo de esta investigación fue comparar la potencia y sensibilidad de dieciséis pruebas de normalidad, clasificadas según diversos aspectos. La metodología utilizada consistió en simular datos a partir del sistema de contaminación Fleishman para evaluar las pruebas en situaciones de no normalidad y diez distribuciones con distintos grados de desviación de la normalidad. Los resultados obtenidos fueron que las pruebas basadas en la correlación y la regresión, como Shapiro-Wilk y Shapiro-Francia, superaron a las demás en potencia, especialmente, para muestras grandes y desviaciones sustanciales de la normalidad. Para desviaciones moderadas se observó que las pruebas de D'Agostino-Pearson y de sesgo se desempeñaron bien, mientras que, para desviaciones bajas, sobresalieron la prueba robusta de Jarque-Bera y la prueba de Jarque-Bera. Además, algunas pruebas mostraron una elevada potencia en distintos tipos de distribuciones, como Snedecor-Cochran y Chen-Ye para distribuciones platicurticas simétricas, y Snedecor-Cochran y Chen-Ye para distribuciones leptocurticas asimétricas. Estos resultados aportaron información valiosa sobre la selección de pruebas de normalidad adecuadas en función de las características de la muestra, lo que ayuda a los investigadores a mejorar la fiabilidad de la inferencia estadística. En conclusión, este artículo muestra escenarios donde las pruebas estadísticas más conocidas no siempre son las más efectivas.

Palabras clave

Método de clasificación de distribuciones, método de Fleishman, simulación Monte Carlo, pruebas de normalidad, comparación de potencias.

1. INTRODUCTION

Normality tests play a crucial role in maintaining the accuracy of statistical results and aiding in the decision-making process during data analysis. Their primary purpose is to determine whether a dataset follows a normal distribution or comes from a population with a normal distribution pattern. This requirement is essential for many statistical analyses, highlighting its inherent significance [1]-[3].

The validity of population inferences drawn from sample data often depends on the normality assumption, which is especially relevant when using certain parametric statistical techniques. This assumption is characterized by a symmetric bell-shaped curve with a defined mean and standard deviation. When data follows a normal distribution, we can confidently use classical and parametric statistical methods since all necessary assumptions are met [4]-[6]. Recent comparative studies have provided insights into the efficacy of various normality tests. [7] conducted an exhaustive power comparison of 40 normality tests, identifying the Hosking 1 test as the most powerful for smaller sample sizes and the newly proposed N-metric test for larger sample sizes. Additionally, the empirical power and p-value distributions of normality tests were assessed by [1], who found that the 2nd Zhang-Wu test demonstrated robustness for small sample sizes, while the 1st Zhang-Wu test exhibited robustness for moderate sample sizes. Both tests were observed to be effective when data exhibited symmetric or asymmetric distributions.

The most widely used tests to determine data normality include the [8]-[10] tests. The choice of test depends on the sample size, data distribution, and other contextual factors specific to the data analysis. It should be noted that no normality test is infallible. Visual interpretation and subject matter expertise may occasionally assist in evaluating data normality [11], [12]. Once the normality test is conducted and it is established that the data follows a normal distribution, further tests can be performed, such as variance homogeneity tests, hypothesis testing, and other parametric analyses [13]-[15]. Nevertheless, [16] used Monte Carlo simulations to assess the power of 50 univariate goodness-of-fit tests for normality against different theoretical distributions, with sample sizes ranging from 10 to 100. The study showed that the Robust Jarque-Bera and Gel-Miao-Gastwirth tests generally had the highest performance for symmetric distributions. On the other hand, the Cabana-Cabana and Zhang-Wu tests showed superior performance for asymmetric distributions. In particular, the second Zhang-Wu test showed consistently robust performance across different distributions, highlighting its versatility and reliability.

Power and sensitivity are critical features of a normality test, vital for data analysis. Statistical tests can identify natural effects or differences if present. Higher power increases the likelihood of detecting deviations from normality in our data for normality tests, which is crucial in interpreting the results [15]. The sensitivity of a test refers to its ability to correctly identify cases that do not adhere to normality. A test with high sensitivity reduces analysis errors by accurately detecting when the data does not follow a normal distribution [2], [5]. This article employed Monte Carlo procedures to assess the power of sixteen normality test statistics. Random sample generation was performed using two methods: Fleishman's method, also known as the power method for generating non-normal data [17], and the distribution classification method.

Different statistical functions are used to categorize normality tests, which assess the coherence of data with a normal distribution. For instance, the Shapiro-Wilk test is based on regression coefficients and compares observed quantities with anticipated quantities under the assumption of normality. The Shapiro-Wilk test is commonly used to test the hypothesis of normality in an experimental design [9]. On the other hand, the Anderson-Darling test uses a statistical model built based on differences between empirical and theoretical cumulative distribution functions [12], [18].

Statistical tests such as the Kolmogorov-Smirnov test measure the difference between theoretical and empirical cumulative distribution functions [8]. The Jarque-Bera test, based on statistical moments of symmetrization and kurtosis, determines whether the data is suitable for a normal distribution [19]. Other less popular tests, such as the D'Agostino-Pearson tests, use statistical moments to evaluate normality [20]. The Cramer-von Mises test, similar to the Anderson-Darling test, evaluates differences within the cumulative distribution function, focusing on deviations across the entire range rather than emphasizing the tails [1], [21]. This article aims to evaluate the efficacy of several normality tests under different simulation scenarios.

This article is structured into several sections, including the following: Section 2 presents the methodology employed, which comprises four groups of normality tests: moment-based tests, empirical distribution, correlation, and regression. Additionally, two data generation methods are considered: Fleishman's method and the distribution classification method. Section 3 provides a comprehensive analysis of the simulation study results, along with a discussion of the findings in relation to recent research. Finally, Section 4 presents the conclusions of the study.

2. METHODOLOGY

This section details the methodology, organized into four subsections. The first subsection introduces the selected normality tests and provides the rationale for its application. The second subsection introduces the Fleishman's method, which is utilized to generate non-normal data with varying levels of contamination. The third subsection describes the distribution classification and specifies the probability distributions employed in the study. Finally, the fourth subsection outlines the structure of the simulation study, detailing the data generation process and the application of normality tests, with particular emphasis on assessing their power and sensitivity.

2.1 Normality tests

The motivation behind conducting sixteen tests of normality is based on the work conducted by [22] and the method for generating non-normal data proposed by [17]. Most of these tests were utilized by [16], where they selected a group of distributions with important properties such as skewness and kurtosis, following the categorization proposed by [23]. The normality tests employed in the development of this article were classified into 4 groups, where

group 1 corresponds to moment-based normality tests; group 2 encompasses tests based on empirical distribution; group 3 corresponds to normality tests based on correlation and regression, and finally, group4 contains tests with specific case specification [22].

In Table 1, reference is made to the tests that comprise each of the groups of interest. Expanded information on the distributions used can be found in [1], [16]. The group of moment-based normality tests is a set of analytical tools that assess the discrepancy between empirical distributions and the cumulative normal distribution function. These tools employ the analysis of the first statistical moments of the samples to determine the extent of the discrepancy. The objective of this study is to compare the values obtained from the data with the expected values when the distribution is normal, as previously discussed in [6].

Table 1. Classification of normality tests by coding, authors, and R functions.
Source: own elaboration.

Test coding	Author	R Functions
Moment-based normality tests (group 1)		
DK	D'Agostino-Pearson [20]	agostino.test(x)
JB	Jarque-Bera [19]	ajb.norm.test(x)
RJB	Robust Jarque-Bera [24]	rjb.test(x)
BS	Bonett-Seier [25]	bonett.test(x)
BM	Bontemps-Meddahi [26]	statcompute(stat.index=14, data=x)
SK	Bai-Ng [27]	skewness.norm.test(x)
KU	Bai-Ng [27]	kurtosis.norm.test(x)
Empirical distribution-based normality test (group 2)		
LL	Kolmogorov-Smirnov [8]	lillie.test(x)
CS	Snedecor-Cochran [28]	chisq.test(x)
G	Chen-Ye [29]	G.test(x)
AD	Anderson-Darling [10]	ad.test(x)
BH	Brys-Hubert-Struy [30]	statcompute(stat.index=16, data=x)
Correlation and regression-based normality tests (group 3)		
SW	Shapiro-Wilk [9]	shapiro.test(x)
SF	Shapiro-Francia [31]	sf.test(x)
Tests with specific case specifications (group 4)		
DH	Doornik-Hansen [32]	statcompute(stat.index=8, data=x)
BHBS	Brys-Hubert-Struy [30]	statcompute(stat.index=18, data=x)

Conversely, the empirical distribution-based normality test group employs a common strategy involving the analysis of the correlation between the theoretical and empirical distributions. This strategy relies on the relationship of two estimations by weighted least squares using a scale obtained from order statistics, as outlined in [33]. The empirical distribution function obtained from the sample data is compared to the theoretical cumulative distribution function. This method allows the degree of concordance or discordance between the two functions to be assessed, providing an indication of how closely the observed data matches the theoretical distribution.

The Correlation and Regression-Based Normality tests group encompasses tests that rely on the idea that deviations from normality can be detected by two sample moments— the skewness and kurtosis. These tests are based on the ratio of two scale estimations derived from the least squares method of order statistics. Specifically, the numerator employs a weighted least squares estimation, while the denominator uses the variance estimated from a sample

drawn from a different population [7]. Finally, the group designated as tests with specific case specifications comprises approaches that cannot be classified into the previous sets.

The methods employed to generate the data are outlined in subsections 2.2 and 2.3, with each of the methodologies explained in detail. The generated data are then tested for normality using the tests listed in Table 1.

2.2 Fleishman's method

In this paper, we employed the Fleishman power method [17] to generate non-normal data. This method uses a transformation of a polynomial function, as shown in (1):

$$Z = a + bX + cX^2 + dX^3 \tag{1}$$

Where Z is a variable with unknown distribution and parameters ($\mu = 0;\ \sigma^2 = 1;\ g_1;\ g_2$) X is a normally distributed random variable with mean 0 and variance 1. This procedure calculates the coefficients a, b, c and d using a polynomial transformation according to the third and fourth moments; skewness (g_1) and kurtosis (g_2). The values of skewness, kurtosis, and Fleishman's coefficients, calculated with their respective levels of deviation from normality, are shown in Table 2 [34]-[36]. The different scenarios of deviation from the normal distribution are generated from these values.

Table 2. Skewness, kurtosis, and Fleishman's coefficients used to generate non-normal simples. Source: own elaboration.

Levels of contaminated	Skewness (g_1)	Kurtosis (g_2)	Coefficients Fleishman (a, b, c, d)
None	0	0	$(0, 1, 0, 0)$
Under	0.25	0.75	$(-0.037, 0.933, 0.037, 0.021)$
Moderated	0.75	1	$(-0.119, 0.956, 0.119, 0.009)$
High	1.3	2	$(-0.249, 0.984, 0.249, -0.016)$
Severe	2	6	$(-0.314, 0.826, 0.314, 0.023)$

2.3 Methods for generating data by distribution classification

To compare the performance and power of normality tests using distribution classification, samples from ten non-normal distributions were classified according to [3], in terms of the levels of skewness and kurtosis, as shown in Table 3.

Table 3. Classify alternative distributions according to their skewness and kurtosis. Source: own elaboration.

Case	g_1 (Skewness) g_2 (Kurtosis)	Classification	Distributions Alternatives
1	$g_1 = 0$ $2.5 \leq g_2 \leq 4.5$	Symmetric mesocurtic	Weibull $(4, 5)$ Logistic $(9, 3)$
2	$g_1 = 0$ $g_2 > 4.5$	Symmetric leptokurtic	t (4) t (1) Cauchy $(0, 0.5)$
3	$g_1 = 0$ $g_2 < 2.5$	Symmetric platicurtic	Beta $(2, 2)$
4	$g_1 = 0$ $g_2 > 4.5$	Asymmetric leptokurtic	Beta $(1, 6)$ Gamma $(2, 9)$ Gamma $(6.5, 2.8)$ Weibull $(1, 2)$

2.4 Methods evaluation of normality tests using Monte Carlo simulations and data generation methods

Monte Carlo methods were employed to assess the performance and power of sixteen normality tests: D'Agostino-Pearson, Jarque-Bera, Robust Jarque-Bera, Bonett-Seier, Bontemps-Meddahi, Skewness, Kurtosis, Lilliefors, Anderson-Darling, Snedecor-Cochran, Chen-Ye, Brys-Hubert-Struyf, Shapiro-Wilk, Shapiro-Francia and Doornik-Hansen. A theoretical comparison was deemed impractical, making simulation necessary [16].

Two methods were used to generate samples: Fleishman's method and distribution classification. Fleishman's method generates non-normal data by introducing controlled deviations in skewness and kurtosis. This study considers five contamination levels: uncontaminated, low, moderate, high, and severe. The distribution classification method, on the other hand, uses different probability distributions, categorizing them into symmetric distributions (mesokurtic, leptokurtic, and platykurtic symmetry) and asymmetric distributions (leptokurtic asymmetry).

To compare the power of the normality tests, samples of sizes n=10, 20, 30, 50, 100, 200 and 500 were generated using both Fleishman's method and the distribution classification method. The groups corresponded to those proposed by [22]. A total of 100,000 simulations were conducted at significance levels of α=1 %,5 % and 1 %. In each simulation, the power was measured as the proportion of times the null hypothesis was correctly rejected in 100,000 replications, using several R statistical software packages [37]. The following is a simulation study designed to evaluate the sensitivity and effectiveness of certain normality tests.

3. RESULTS AND DISCUSSION

This section presents the results of the simulation study, which involves two methods for generating samples: the Fleishman's method and distribution classification. These methods are described in their respective cases in Section 2 (Tables 2 and 3). In each case, the normality tests listed in Table 1 are evaluated. Furthermore, a discussion of the results of the article in the context of recent work related to normality tests is provided.

3.1 Results of Fleishman's method

In this subsection, a comparative analysis of power was conducted according to Fleishman's method. The methodology proposed in subsection 2.4 was applied to obtain the statistical power and sensitivity of the normality tests. The results of the study, conducted across a range of contamination levels, are presented below.

3.1.1 Simulated power for the low contamination level

Figure 1 illustrates the power of various normality tests in the presence of low contamination levels. It is evident that, at a significance level of 10 %, the tests in group 1, with sample sizes 10≤n≤20, underestimate the significance level. However, as the sample size increases, there is an overestimation of the significance level in the BM test, while the DK, JB, RJB, BS, SK, and KU tests present estimated power values close to the significance level. Additionally, in this scenario, two main groups can be distinguished, considering the overestimation and underestimation of the significance level.

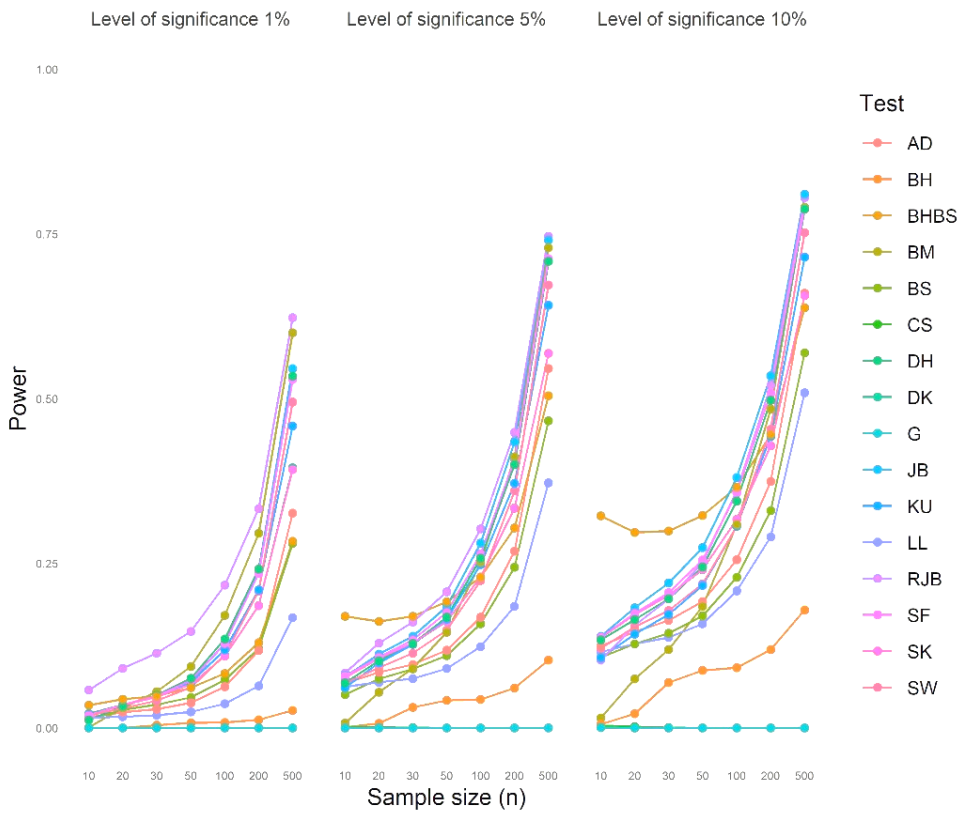

Figure 1. Simulated power for the low contamination level across three significance levels. Source: own elaboration.

On the other hand, the CS and G tests exhibit considerable variability compared to the other tests, as they underestimate the significance level. This indicates that the null hypothesis (H_0), which states that the sample comes from a population with a normal distribution, is rejected more often than it should be, despite being true in all scenarios. Figure 1 shows that, when the significance level is set at 10 %, tests based on the empirical distribution exhibit distinct behaviors. For instance, the BH test underestimates the significance level compared to the other tests, while the LL and AD tests demonstrate a good fit for all sample sizes.

Similarly, groups 3 and 4 show a good fit for all sample sizes, exhibiting an overestimation of the significance levels at 1 % and 5 %, respectively. Moreover, it is notable that the estimated power displays a trend as the significance level increases for all analyzed sample sizes.

3.1.2 Simulated power for the low contamination level

Figure 2 illustrates the power for the moderate contamination level. It can be observed that, when using a significance level of 1 %, the tests in group 1, with a sample size of $n = 10$, underestimate the significance level, with a value of 0.00227 for the BM test. However, as the sample size increases, this test overestimates the significance level. Conversely, the DK, JB, RJB, BS, SK, and KU tests demonstrate a satisfactory fit for all sample sizes, as their respective estimated significance levels approach 1 %. It is evident that the DK and SK tests are the most powerful in this group, followed by the BM test, and then the RJB, JB, KS, and BS tests, respectively.

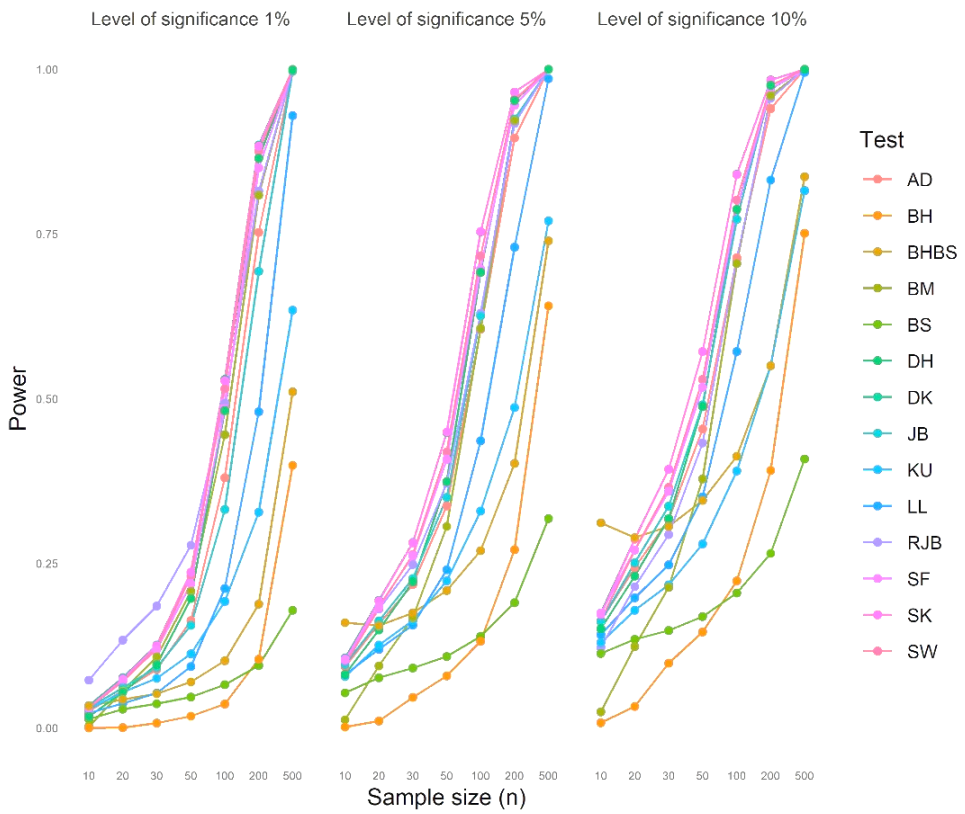

Figure 2. Simulated power for the moderate contamination level across three significance levels. Source: own elaboration.

Conversely, the BH test exhibits a degree of underestimation, albeit to a lesser extent than that observed for the CS and G tests, for sample sizes 10≤n≤30. Conversely, the LL and AD tests demonstrate a satisfactory fit for all sample sizes. However, as the sample size increases for the five tests in group 2, only three overestimate the significance level.

For a significance level of 5 %, the tests in group 3, SW and SF, exhibit a good fit across all sample sizes. This is evidenced by their tendency to overestimate the 5 % value (see Figure 2). A similar behavior is observed with the tests in group 4. Likewise, for a significance level of 10%, the tests in groups 3 and 4 similarly overestimate across all sample sizes (see Figure 2). It should be noted that the CS and G tests are not depicted in Figures 2, 3, and 4 because the power values reported for the various significance levels and sample sizes approach zero. This occurs because they underestimate the significance level, leading to more frequent rejection of H_0, even when it may be true.

3.1.3 Simulated power for the high contamination level

Figure 3 illustrates that the simulated power increases as the sample size and significance level grow, reaching a maximum in some normality tests. For a significance level of 1 %, the tests in group 1 show a good fit for all sample sizes, becoming more sensitive as the degree of contamination increases. This is evidenced by the fact that their respective estimated significance levels approach 1 %, with the DK and SK tests exhibiting higher power. In group 2, it is noteworthy that at a significance level of 10 %, the LL and AD tests maintain a good fit for all sample sizes. However, as the sample size increases for the five tests in group 2, only three overestimate the significance level of 1 %.

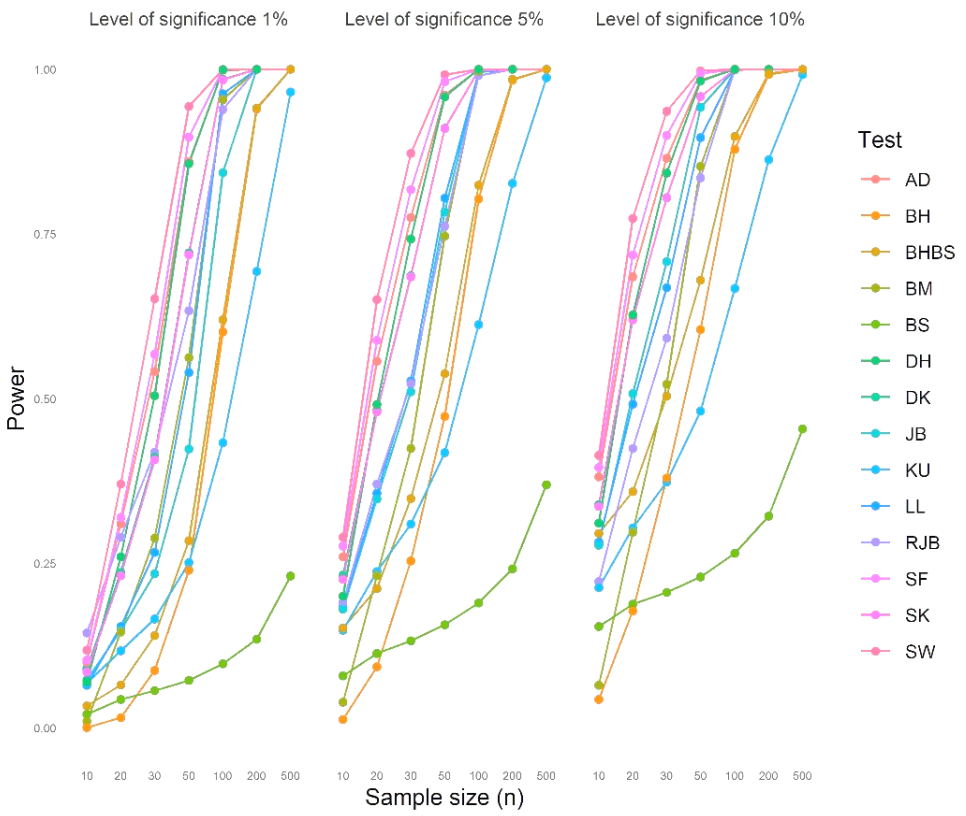

Figure 3. Simulated power for the high contamination level across three significance levels. Source: own elaboration.

For a significance level of 5 %, the tests in group 3, as shown in Figure 3, exhibit a good fit. This is evidenced by the fact that all sample sizes overestimate their 5 % value, reaching maximum power between sample sizes 100≤n≤500. A similar pattern is observed in group 4, where the DH test reaches its maximum power between sample sizes 200≤n≤500, and the BHB-S test reaches its maximum power at a sample size of n=500 with a significance level of 5 %. These tests similarly overestimate the significance level across all sample sizes, making them highly powerful.

3.1.4 Simulated power for the severe contamination level

Figure 4 illustrates the impact of severe contamination on test power, showing a clear trend where power increases with sample size and significance level. For instance, in the most powerful tests of group 1, DK and SK reach their maximum power from a sample size of $n = 50$, indicating that these moment-based normality tests have a good fit. In group 2, the most powerful test is AD, reaching its maximum power from a sample size of $n = 50$. For the tests in group 3, the most powerful test is SW, exhibiting a good fit across all sample sizes, as it overestimates the significance level and reaches its maximum power between sample sizes $100 \leq n \leq 500$. In contrast, in group 4, the most powerful test is DH, which reaches its maximum power between sample sizes $200 \leq n \leq 500$. Similarly, the BHBS test reaches its maximum power at a sample size of $n = 500$.

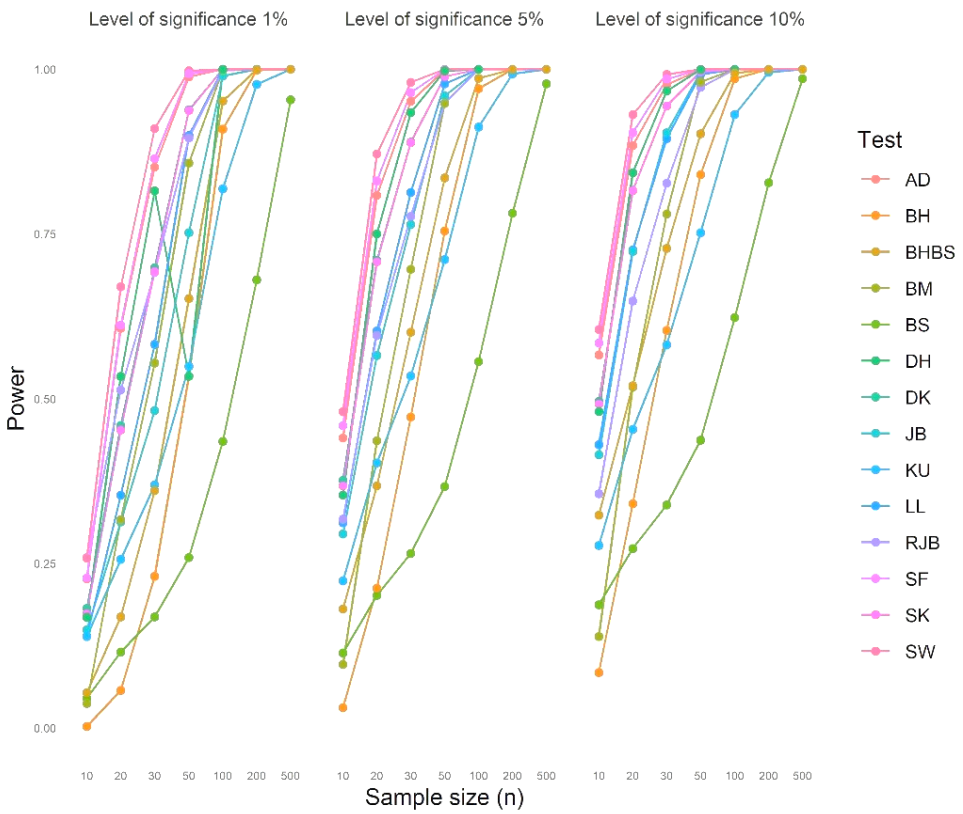

Figure 4. Simulated power for the severe contamination level across three significance levels. Source: own elaboration.

Notably, Figure 4 provides evidence in favor of using the Shapiro-Wilk test (SW), which exhibits higher power compared to the other tests analyzed. However, it is crucial to note that although the test shows relatively high power across all sample sizes, it never exceeds a value that could be considered acceptable (minimum of 0.6) for small samples, within the interval (0,1) where it takes values. From a high contamination level onwards, high power is observed only for sample sizes n≥30. It is noteworthy that most of the tests analyzed become highly powerful under severe contamination. In this regard, it can be concluded that normality tests are only effective when the departure from theoretical distribution is significant.

3.2 Results of the distribution classification method

This section presents a comparison of the power according to the classification of distributions (symmetric mesokurtic distributions, symmetric leptokurtic distributions, symmetric platykurtic distributions, and asymmetric leptokurtic distributions) using the simulation methodology outlined in Section 2. The statistical power of the normality tests is then obtained.

3.2.1 Symmetric mesokurtic distributions

In this initial classification, the distributions utilized exhibit characteristics consistent with a normal distribution, as evidenced by their proximity to zero for both skewness and kurtosis. The Logistic (9,3) distribution and the Weibull (4,5) distribution were employed, as illustrated in Figures 5 and 6. Figure 5 shows that the BH normality test is the most powerful, while the G test is the least sensitive and powerful for the Logistic (9,3) distribution. However, for sample sizes ($n \geq 30$), the lowest reported power is observed with the SF test. The BH test performs optimally for all sample sizes and significance levels. Conversely, the tests with the lowest performance at a significant level of 1 % are G, CS, DK, and SF. At a significant level of 5 %, the tests with lower performance are G and CS. Finally, at a significant level of 10 %, the tests G, CS, and SF show the lowest performance.

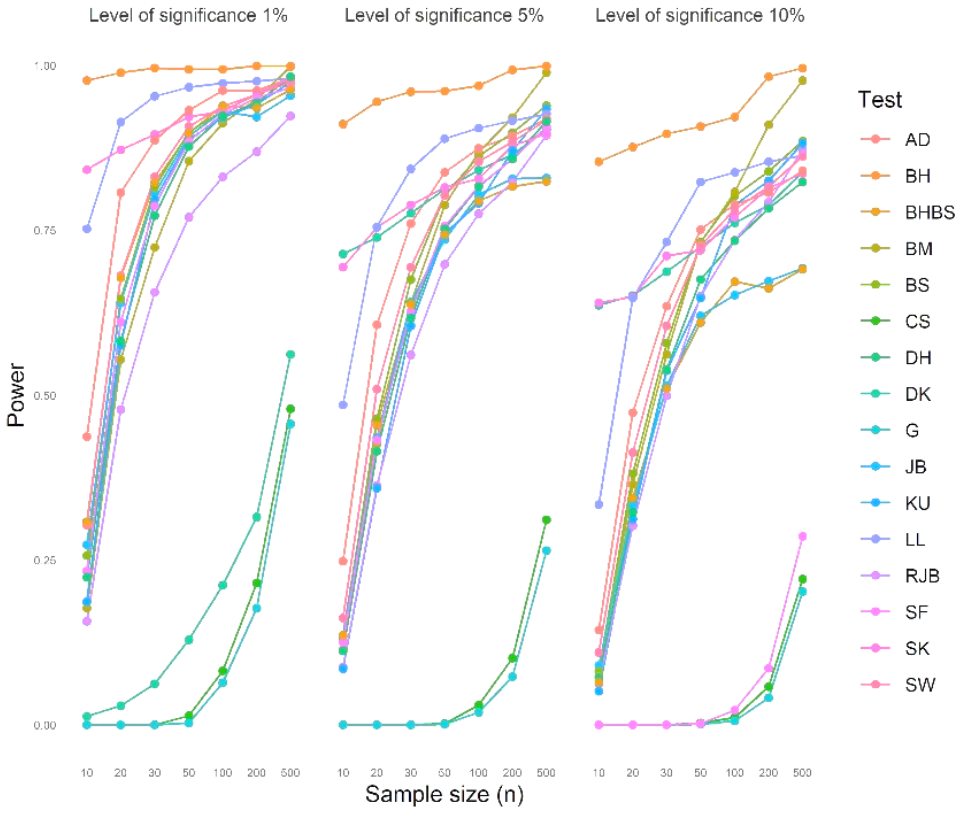

Figure 5. Simulated power under the Logistic (9,3) distribution at different significance levels. Source: own elaboration.

On the other hand, it is noteworthy that the G and CS tests exhibit good performance for all sample sizes in the case of the Weibull (4,5) distribution (see Figure 6). Conversely, at a significance level of 10 % and larger sample sizes (n≥30), the SF normality test shows low power compared to the others. Similarly, the JB and BHBS tests show low performance, particularly at a significance level of 5 %.

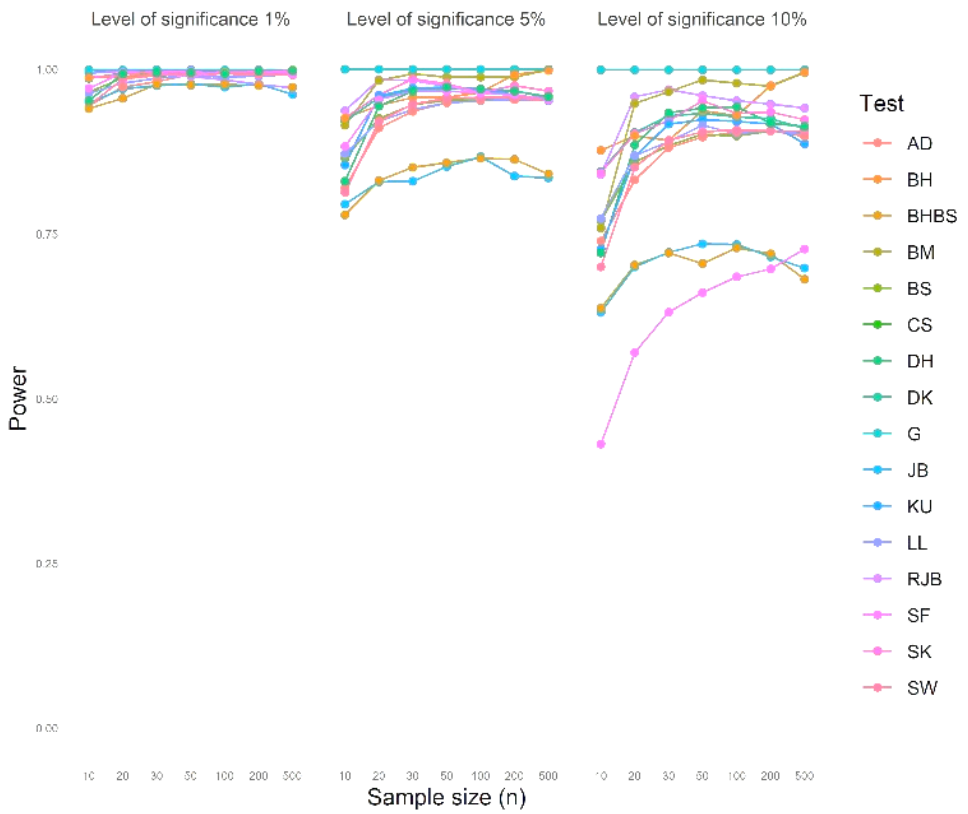

Figure 6. Simulated power under the Weibull (4,5) distribution at different significance levels. Source: own elaboration.

The results obtained from the two mesokurtic symmetric distributions (see Figures 5 and 6) indicate that for sample sizes less than or equal to 100 (n≤100), both the CS and G tests exhibit low power and sensitivity with the Logistic (9,3) distribution. In most cases, it is accepted that the data originate from a normal distribution. In contrast, when using the Weibull (4,5) distribution, the opposite is true. It is important to note that, at a significance level of 1 %, the LL test performs well for all samples and is the second most powerful test when the data originate from the Logistic (9,3) distribution. Furthermore, it was observed that the remaining tests exhibited a notable decline in sensitivity and an increase in power as the sample size increased.

3.2.2 Symmetric leptokurtic distributions

In this classification, the test that demonstrates the most robust performance in terms of power is the BH test with the Cauchy (0,0.5) and t (1) distributions, followed by the G test with the t (4) distribution. Conversely, the test with the lowest power for all sample sizes is the RJB test, followed by the CS test for sample sizes smaller than thirty (see Figures 7, 8, and 9).

As illustrated in Figure 7, it can be observed that as the significance level increases, the normality tests lose power. This indicates that the BH test is the most powerful and least sensitive compared to the others at significance levels of 1 % and 5 %. However, at a 10 % confidence level, the SF test outperforms the BH test, maintaining its power above 0.883. In contrast, the RJB test is the least powerful for the Cauchy (0,0.5) distribution. For sample sizes of at least 50 (n≥50), the tests with the worst power performance are RJB, SF, AD, SW, BS, DH, BHBS, KU, JB, LL, and BM at a 1 % significance level.

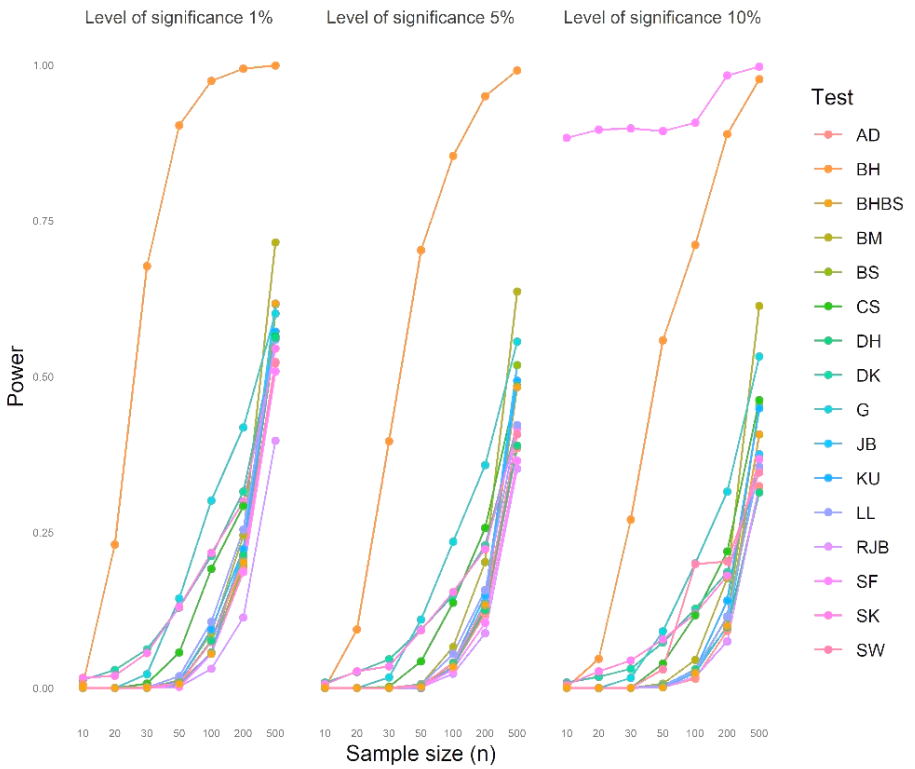

Figure 7. Simulated power under the Cauchy (0,0.5) distribution with different significance levels. Source: own elaboration.

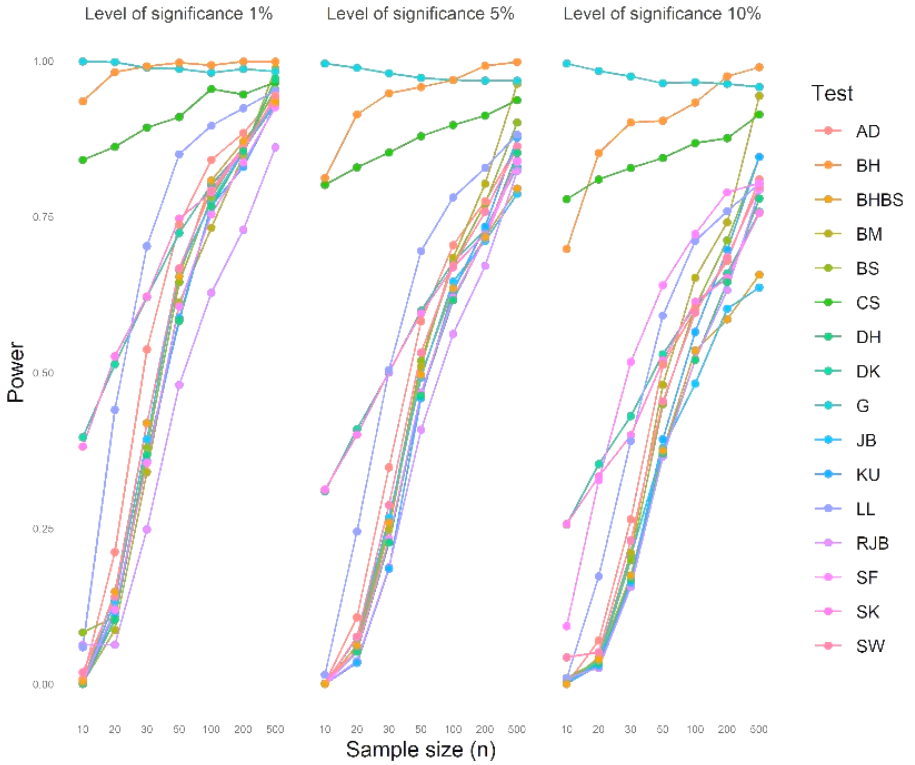

Figure 8. Simulated power under the *t* (4) distribution with different significance levels. Source: own elaboration.

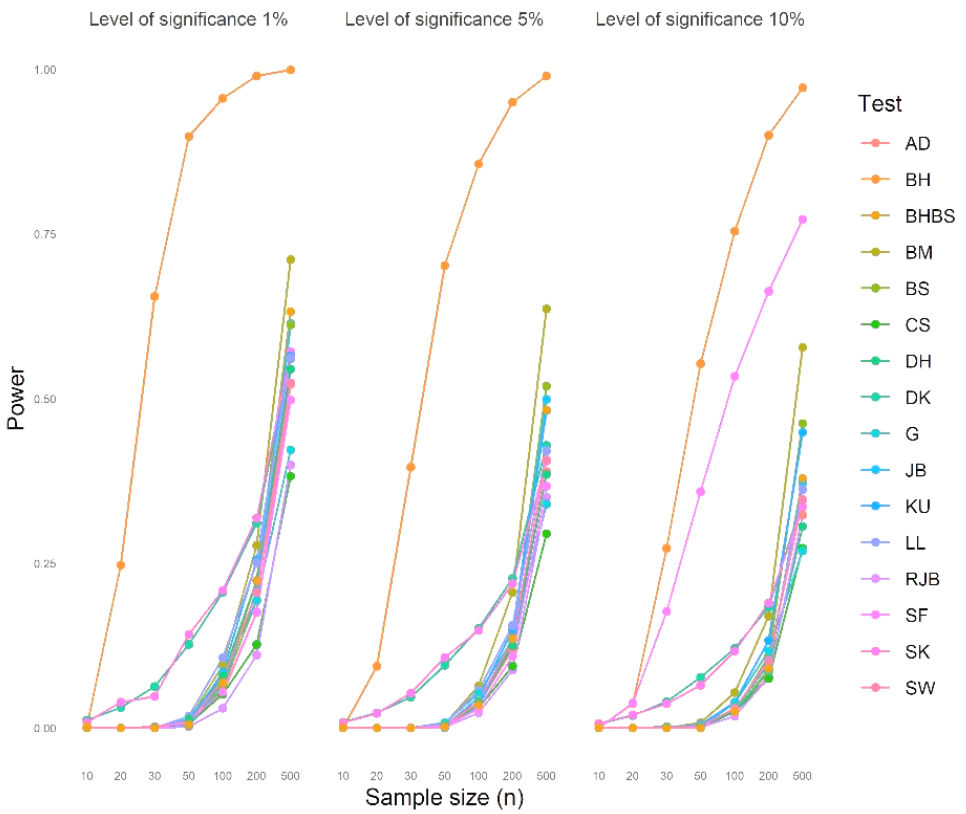

Figure 9. Simulated power under the t (1) distribution with different significance levels. Source: own elaboration.

In the case of the t (4) distribution (Figure 8), the excellent performance of the G, BH, and CS tests at all sample sizes is particularly noteworthy. Additionally, it is observed that the normality tests G, BH, and CS exhibit high power for the three significance levels and all sample sizes, in contrast to the rest of the tests. Conversely, the RJB and KU tests demonstrate the opposite pattern for significance levels of 1 % and 5 %, while for a significance level of 10 %, the JB and BHBS tests exhibit low power and higher sensitivity for sample sizes of (n≥100).

Figure 9 illustrates that the BH test maintains its high power, followed by the SK and DK tests, for significance levels of 1 % and 5 %. At a significance level of 10 %, however, the SF and BM tests exhibit the greatest power under this distribution, while the opposite is true for the RJB, CS, and DH tests (less powerful). Finally, the results of the power and sensitivity normality tests for a leptokurtic symmetric distribution (Figures 8 and 9) indicate that the distribution with the greatest power is the t (4), with the normality tests G, BH, and CS performing best at all significance levels.

3.2.3 Symmetric platicurtic distribution

In the case of a symmetric platykurtic distribution, the Beta (2,2) distribution was considered. It can be observed that half of the normality tests exhibit low statistical power for sample sizes ranging from small to medium ($10 \leq n \leq 20$) at a significance level of 10 %. This is illustrated in Figure 10. It can also be seen that all normality tests perform well for sample sizes ($n \geq 200$), with the exception of the SF test with α = 10 %.

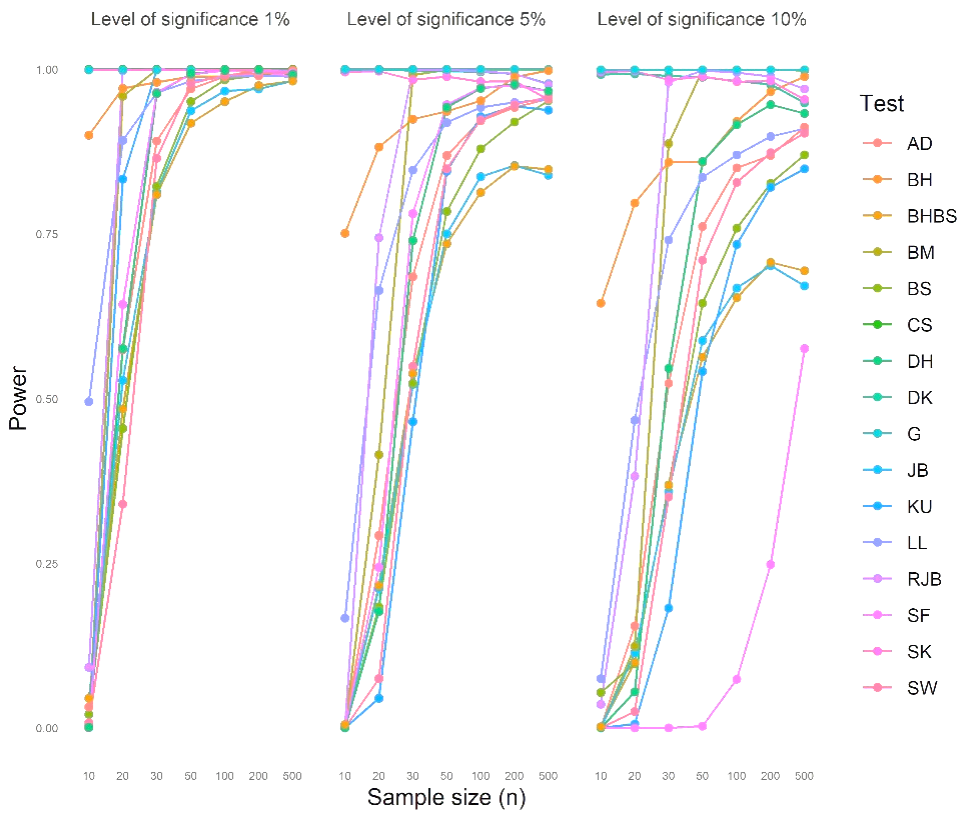

Figure 10. Simulated power under the Beta (2,2) distribution with different significance levels. Source: own elaboration.

Figure 10 illustrates that the CS, G, and DK tests exhibit high power, followed by the SK and BH tests at significance levels of 1 % and 5 %. At a significance level of 10 %, the SK and DK tests are among the four most powerful tests, while the BHBS test is the least powerful.

3.2.4 Asymmetric leptokurtic distributions

In this classification, the analyzed distributions exhibit a degree of asymmetry with long tails. It can be observed that in all the tests analyzed, the statistical power and sensitivity increase as the sample size increases. That is, in most cases, it is rejected that the data come from a normal distribution when this is false. The tests with the greatest statistical power are CS and G in the case of Beta (1,6), Gamma (2,9), and Gamma (6.5, 2.8), while in the case of Weibull (1,2), the most effective test is BS (see Figures 11-14).

Figure 11 illustrates that the power of tests CS, G, BS, and KU is high, followed by tests BH, BHBS, and JB for all significance levels. The latter stand out in their high power among the seven tests under this distribution. Conversely, the SW test becomes the least powerful for samples with a minimum of 30 observations (n≥30).

The CS, G, and DK tests exhibit high power, followed by the BH, BHBS, and JB tests for significance levels of 1 % and 5 % (see Figure 12). However, at the 10 % level, it is the BH and KU tests that are the most powerful among the seven tests. Conversely, the tests in group 3, based on correlation and regression, become the least powerful under this distribution.

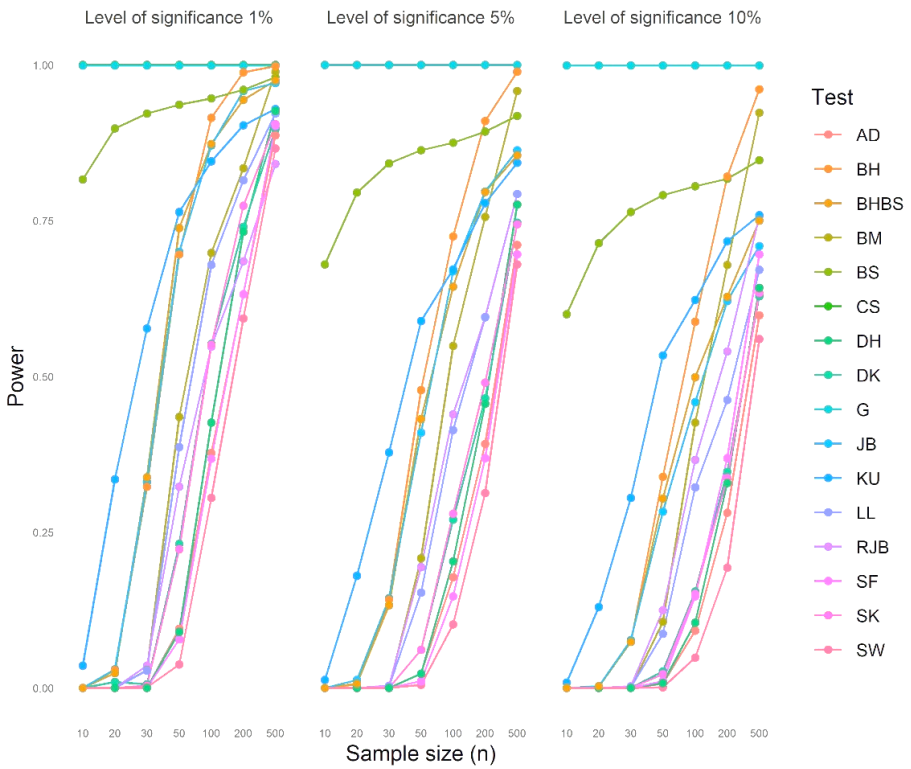

Figure 11. Simulated power under the Beta (1,6) distribution with different significance levels. Source: own elaboration.

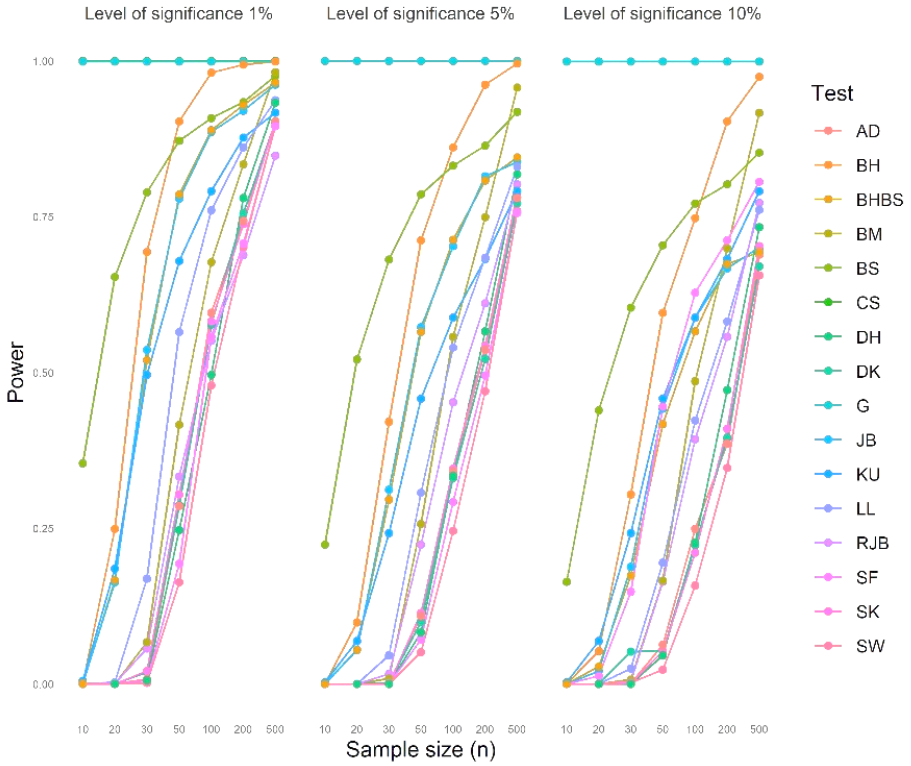

Figure 12. Simulated power under the Gamma (2,9) distribution with different significance levels. Source: own elaboration.

Figure 13 illustrates that the CS, G, and BS tests maintain their high power for samples with at least 20 observations (n≥20), followed by the BH, BS, and JB tests for significance levels of 1 % and 5 %. However, at the 10 % level, the power varies, with the BH test remaining the most powerful, followed by the KU and JB tests, respectively. Therefore, it can be concluded that these tests are the most powerful under the specified distribution. Conversely, the SF test exhibits the lowest power for a significance level of 10 %.

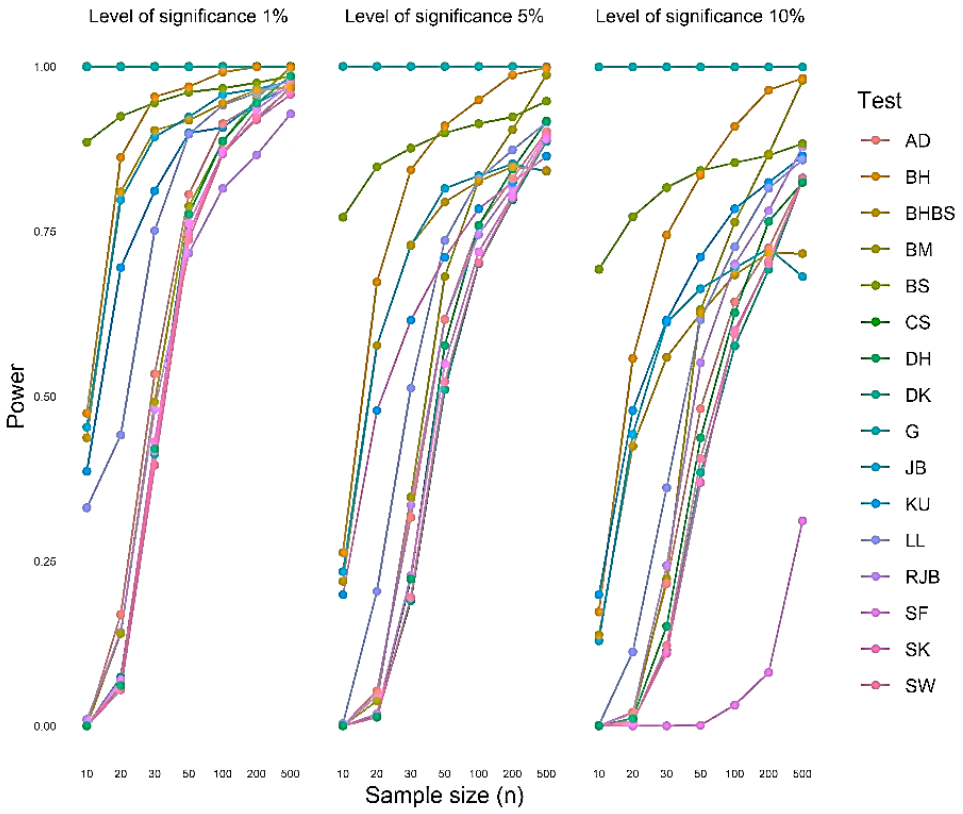

Figure 13. Simulated power under the Gamma (6.5,2.8) distribution with different significance levels. Source: own elaboration.

The BS, BH, JB, and BHBS tests exhibit high power, followed by the KU, G, and BM tests for all significance levels (see Figure 14). Conversely, lower power is observed in the SF test with α = 1 % and samples (n≤50), the SK test with α = 5 % and samples (n≤50), and the SW test with α = 10 % and samples (n≤200).

In summary, the comparative analysis of normality tests with regard to power and sensitivity for leptokurtic asymmetric distributions indicates that the distribution with the highest power is the Gamma (6.5, 2.8) under the CS, G, and BS normality tests for all significance levels and sample sizes.

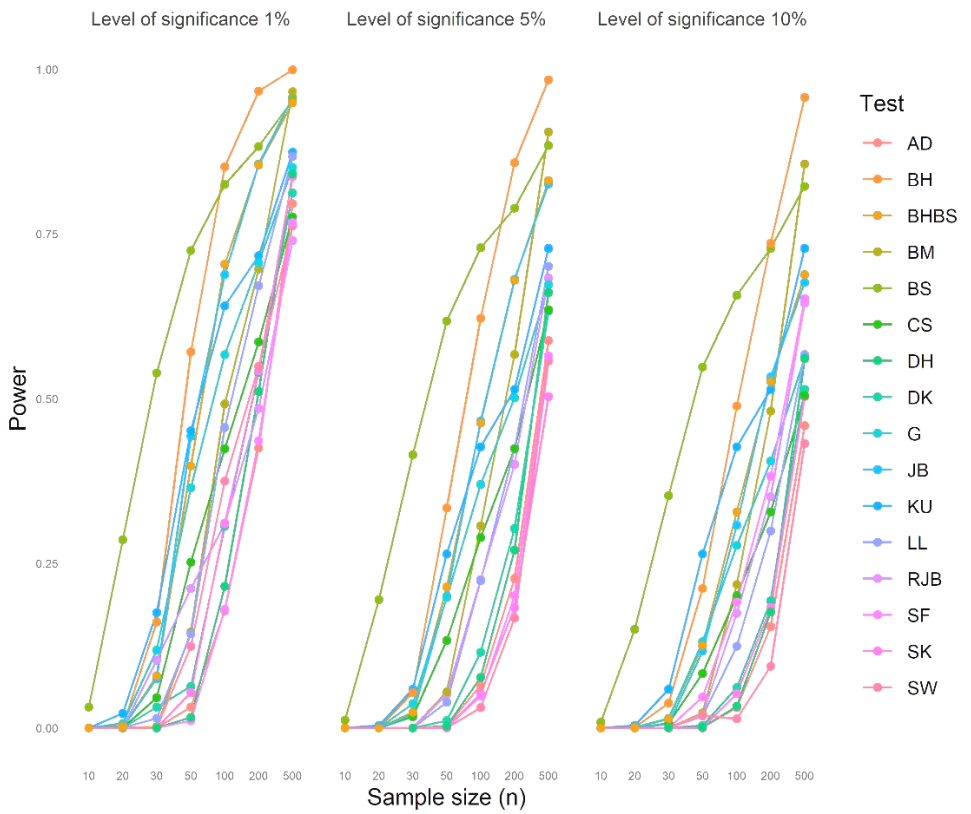

Figure 14. Simulated power under the Weibull (1,2) distribution with different significance levels. Source: own elaboration.

3.3 Normality assessment and testing procedures

This subsection presents several normality tests used to determine whether the generated data follow a normal distribution. The null hypothesis (H_0) states that the data originate from a normal distribution, while the alternative hypothesis (H_1) suggests that the data deviate from normality. To generate controlled non-normal data, Fleishman's method is employed, allowing for systematic manipulation of skewness and kurtosis. A variety of tests—each based on distinct statistical principles—have been applied to ensure a robust evaluation. These tests include:

- Moment-based tests: D'Agostino-Pearson, Jarque-Bera, and Bonett-Seier.
- Empirical distribution-based tests: Kolmogorov-Smirnov.
- Correlation and regression-based tests: Shapiro-Wilk.

A test rejects (H_0) when the p-value falls below a predefined significance level (e.g., $\alpha = 0.05$), indicating that the data likely exhibit non-normal behavior. Below, we describe the key normality tests and their implementation in R.

3.3.1 D'Agostino-Pearson test (group 1: moment-based tests)

The D'Agostino-Pearson test evaluates normality by transforming skewness and kurtosis into a joint test statistic [20]. It is implemented in R as follows:

```
# Install and load the 'moments' package if not already installed
install.packages("moments")
```

```
library(moments)

# Define the D'Agostino-Pearson test function
agostino_test <- function(x) {
  skewness_value <- skewness(x)
  kurtosis_value <- kurtosis(x)
  n <- length(x)

  # Calculate the test statistic and corresponding p-value
  stat <- ((skewness_value^2) / 6) + ((kurtosis_value^2) / 24)
  p_value <- 1 - pchisq(stat, df = 2)
  return(list(statistic = stat, p_value = p_value))
}

# Fleishman function to generate non-normal data
fleishman <- function(n, skewness, kurtosis) {
  X <- rnorm(n, mean = 0, sd = 1) # Generate standard normal variables

  # Fleishman coefficients based on the desired skewness and kurtosis
  if (skewness == 1.3 && kurtosis == 2) {
    a <- -0.249; b <- 0.984; c <- 0.249; d <- -0.016
  } else if (skewness == 0.75 && kurtosis == 1) {
    a <- -0.119; b <- 0.956; c <- 0.119; d <- 0.009
  } else {
    a <- 0; b <- 1; c <- 0; d <- 0
  }
    Z <- a + b * X + c * X^2 + d * X^3 # Apply Fleishman transformation
  return(Z)
}

# Generate non-normal data with high contamination (skewness = 1.3, kurtosis = 2)
set.seed(123)
fleishman_data <- fleishman(1000, skewness = 1.3, kurtosis = 2)

# Visualize the generated data
hist(fleishman_data, col = 4, main = "Data Generated with Fleishman Method")
grid()

# Apply the D'Agostino-Pearson test
result_dk <- agostino_test(fleishman_data)
print(result_dk)
```

3.3.2 Kolmogorov-Smirnov test (group 2: empirical distribution-based tests)

The Kolmogorov-Smirnov test compares the empirical cumulative distribution function of a sample with the theoretical normal distribution [8]. It is implemented as follows:

```
# Perform the Kolmogorov-Smirnov test
ks_test_skewed <- ks.test(fleishman_data, "pnorm", mean = mean(fleishman_data), sd = sd(fleishman_data))
print(ks_test_skewed)
```

3.3.3 Shapiro-Wilk test (group 3: correlation and regression-based tests)

The Shapiro-Wilk test is widely used for assessing normality, particularly in small sample sizes [9]. Its implementation is shown below:

```
# Apply the Shapiro-Wilk test
shapiro_test_skewed <- shapiro.test(fleishman_data)
print(shapiro_test_skewed)
```

3.3.4 Jarque-Bera test (group 1: moment-based tests)

The Jarque-Bera test evaluates normality based on skewness and kurtosis [19]. To implement this test, the following code is used:

```
# Install and load the 'tseries' package if not already installed
install.packages("tseries")
library(tseries)

# Apply the Jarque-Bera test
result_jarque_bera <- jarque.bera.test(fleishman_data)
print(result_jarque_bera)
```

3.3.5 Bonett-Seier test (group 1: moment-based tests)

The Bonett-Seier test provides a robust method for assessing normality based on variance estimation [26]. The following R code implements this test:

```
# Install and load the BSDA package if not already installed
install.packages("BSDA")
library(BSDA)

# Apply the Bonett-Seier test
result_bonett <- bonett.test(fleishman_data)
print(result_bonett)
```

The results from the applied normality tests reveal that the generated data deviate from a normal distribution. Specifically, low p-values ($\alpha < 0.05$) obtained consistently across tests—namely, the D'Agostino-Pearson, Kolmogorov-Smirnov, Shapiro-Wilk, Jarque-Bera, and Bonett-Seier tests—indicate that the Fleishman-transformed data exhibit controlled non-normal characteristics such as skewness and kurtosis. These findings underscore the importance of using multiple normality tests, as each offers a distinct perspective on data distribution, thereby ensuring a comprehensive evaluation. Researchers are encouraged to select the appropriate test based on sample size and distribution shape to enhance the reliability of statistical inferences. All analyses were conducted using the R statistical software [37], with packages such as moments, tseries, and BSDA facilitating the computations.

3.4 Discussion

The statistical power and sensitivity of normality tests were evaluated using two methodologies for data generation: Fleishman's method and the distribution classification method, each under multiple contamination scenarios. These methodologies allowed for a

comprehensive comparison across symmetric and asymmetric distributions, offering insights into the behavior of various normality tests.

In low contamination scenarios (Fleishman's method: skewness g_1=0.25 and kurtosis g_2=0.70), tests such as DK, JB, RJB, BS, SK, and KU demonstrated a good fit for all sample sizes at a 10 % significance level. This finding is consistent with those of previous studies, particularly those of [16], who observed that the RJB test exhibited high power in symmetric distributions but demonstrated variability in its performance with respect to sample size and contamination level. Notably, at the significance level (5 %), tests SW and SF (group 3) demonstrate a good fit across all sample sizes and tests DH and BHBS (group 4). However, at lower significance levels (1 %), these tests tend to overestimate the significance value.

In moderate contamination scenarios (skewness g_1=0.75, and kurtosis g_2=1), RJB (Robust Jarque-Bera) demonstrated the highest power at a 1 % significance level, maintaining robustness across sample sizes. This result aligns with findings by [1], who also identified RJB's superior performance under moderate contamination. However, this power diminishes or becomes sensitive as the significance level increases to 5 % and (10≤n≤50), suggesting that while effective at stricter significance levels, it may become overly sensitive as the criteria are relaxed. This is consistent with earlier findings that moment-based tests like RJB can lose sensitivity under lenient significance thresholds [7]. It was also observed that most tests overestimate the significance level with increasing sample size, except the BH test with sample size (10≤n≤100) and α = 1 %. Furthermore, the estimated significance value is underestimated for sample sizes between 10 and 30 with significance levels of 5 % and 10 %. This is also true for the CS and Gtests, indicating that the associated null hypothesis is rarely rejected.

In high contamination scenarios (skewness g_1=1.3, kurtosis g_2=2), moment-based tests such as DK and SK began to lose sensitivity, particularly at a 10 % significance level, where tests like CS and G began to underestimate the significance level. This underestimation was exacerbated in small samples, particularly for the BM test. Under symmetric leptokurtic distributions, group 2 tests (G, BH, CS) remained the most powerful across all significance levels. This underperformance is especially critical under asymmetric distributions like Gamma (2,9) or Beta (1,6), as reported in [34] comprehensive comparison study, where these tests consistently failed to detect departures from normality in smaller samples. Similar behavior was observed in other distribution tests, such as BH and AD, which, while performing well for large sample sizes, failed to maintain power in smaller datasets.

In scenarios involving symmetric platykurtic distributions like Beta (2,2), the tests with the highest power were DK, SK, and BM at both 1 % and 5 % significance levels, as illustrated by their consistency across all sample sizes (n≥200) [34]. These tests performed well, consistently identifying deviations from normality, particularly for distributions with low kurtosis. This supports the conclusions from [7], who highlighted the efficacy of these tests in detecting non-normality in symmetric, platykurtic distributions. A particularly interesting observation comes from tests under Weibull (1,2) and other asymmetric distributions. BS proved to be the most effective test, outperforming other tests like RJB and JB in these cases. This is in line with findings by [1], who pointed out the particular suitability of BS for heavy-tailed distributions. For larger sample sizes (n≥500) under severe contamination scenarios confirmed the superiority of CS, G, BS, and BH across all significance levels. These tests, particularly BS and BH, showed remarkable consistency in rejecting the null hypothesis, supporting their recommended use in large datasets, as highlighted by [16].

Finally, in this article, the problem discussed by [7] and [16] is addressed by examining the sensitivity and power of various tests. The contribution of this research lies in the method used to generate non-normal samples through the Fleishman transformation—a marked departure from the approach adopted by [7] and [16], who generated data based on different probability distributions. Moreover, rather than restricting the selection of normality tests to the most common ones, this study offers a broader range of alternatives by highlighting lesser-known tests that may prove beneficial in scenarios where traditional methods exhibit limitations.

Details of the data-generation process and test implementation, including R syntax, are provided in subsection 3.3.

4. CONCLUSIONS

This study demonstrates that normality tests exhibit significant variation in power and sensitivity depending on the distribution type and contamination level. By employing Fleishman's method and distribution classification, we established a robust framework for evaluating their performance under diverse conditions. In low contamination scenarios, tests such as DK, JB, RJB, BS, SK, and KU are highly effective at a 10 % significance level for all sample sizes, though their performance diminishes under stricter significance levels. Conversely, SW and SF tests consistently perform well across all scenarios, maintaining a high degree of power even at lower significance levels.

Moderate contamination scenarios highlight the robustness of the RJB test at a 1 % significance level, but its power reduces as the sample size increases and the significance level reaches 5 % or 10 %. Additionally, empirical distribution-based tests, such as CS and G, demonstrate reduced sensitivity under moderate contamination, reflecting their conservative nature in rejecting the null hypothesis. High contamination scenarios show that moment-based tests (e.g., DK and SK) struggle with sensitivity as contamination increases, particularly at higher significance levels. This is particularly pronounced in small samples where tests such as BM tend to underestimate significance, rendering them less reliable in these cases.

For symmetric distributions, particularly leptokurtic and platykurtic types, tests such as CS, G, DK, SK, and BM remain among the most powerful. This trend is consistent across distributions like Beta (2,2) and Gamma (6.5, 2.8), emphasizing the importance of sample size and distribution shape in determining test efficacy. For larger sample sizes (n≥500), the most powerful tests across asymmetric distributions, particularly Beta (1,6) and Weibull (1,2), are the CS, G, BS, and BH tests, which maintain strong performance even as significance levels rise. This suggests that these tests are particularly well-suited for detecting non-normality in heavily skewed distributions.

Future studies should further explore how sample size influences normality test performance, particularly in small-sample scenarios with high contamination levels. Investigating the use of alternative methods for calculating p-values, especially under varying degrees of skewness and kurtosis, could further refine the understanding of normality test performance.

5. REFERENCES

[1] T. Uhm, and Y. Seongbaek, "A Comparison of Normality Testing Methods by Empirical Power and Distribution of P-Values," *Communications in Statistics Simulation and Computation*, vol. 52, no. 9, pp. 4445-4458, Aug. 2021. ĥţţŗ ˈ ://đôï.ôŗĝ/10.1080/03610918.2021.1963450

[2] K. Rani Das, and A. H. M. Rahmatullah Imon "A Brief Review of Tests for Normality," *American Journal of Theoretical and Applied Statistics*, vol. 5, no. 1, pp. 5-12, Jan. 2016. https://doi.org/10.11648/j.ajtas.20160501.12

[3] S. S. Shapiro, M. B. Wilk, and H. J. Chen, "A Comparative Study of Various Tests for Normality," *Journal of the American Statistical Association*, vol. 63, no. 324, pp. 1343-1372, Apr. 1968. ĥţţŗ ˈ ://đôï.ôŗĝ/10.1080/01621459.1968.10480932

[4] H. C. Thode, "Plots, probability plots and regression tests," in *Testing for Normality*, 1st ed: Boca Raton: CRC Press, 2002. https://doi.org/10.1201/9780203910894

[5] X. Romão, R. Delgado, and A. Costa, "An Empirical Power Comparison of Univariate Goodness-of-Fit Tests for Normality," *Journal of Statistical Computation and Simulation*, vol. 80, no. 5, pp. 545-591, Mar. 2009. https://doi.org/10.1080/00949650902740824

[6] H. A. Noughabi, and N. R. Arghami, "Monte Carlo Comparison of Seven Normality Tests," *Journal of Statistical Computation and Simulation*, vol. 81, no. 8, pp. 965-972, Dec. 2010. https://doi.org/10.1080/00949650903580047
[7] J. Arnastauskaitė, T. Ruzgas, and M. Bražėnas, "An Exhaustive Power Comparison of Normality Tests," *Mathematics*, vol. 9, no. 7, p. 788, Apr. 2021. https://doi.org/10.3390/math9070788
[8] A. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *Giornale dell'Istituto Italiano degli Attuari,* vol. 4, pp. 89-91, 1933. https://www.sciepub.com/reference/1552
[9] S. S. Shapiro, and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591-611, Dec. 1965. https://doi.org/10.2307/2333709
[10] T. W. Anderson, and D.A. Darling, "Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes," *Annals of Mathematical Statistics*, vol. 23, no. 2, pp. 193-212, Jun. 1952. https://doi.org/10.1214/aoms/1177729437
[11] K. V. Mardia, "9 Tests of Univariate and Multivariate Normality," in *Handbook of Statistics,* Amsterdam: Elsevier, 1980, pp. 279–320. https://doi.org/10.1016/S0169-7161(80)01011-5
[12] B. W. Yap, and C.H. Sim, "Comparisons of Various Types of Normality Tests," *Journal of Statistical Computation and Simulation*, vol. 81, no. 12, pp. 2141-2155, May. 2011. https://doi.org/10.1080/00949655.2010.520163
[13] S. S. Uyanto, "Monte Carlo Power Comparison of Seven Most Commonly Used Heteroscedasticity Tests," *Communications in Statistics - Simulation and Computation,* vol. 51, no. 4, pp. 2065-2082, Nov. 2019. https://doi.org/10.1080/03610918.2019.1692031
[14] H. Hernandez, "Testing for Normality: What is the Best Method?," ForsChem Research Reports, Medellín, Colombia, Technical Report, 2021. Accessed: Jan. 11, 2024. [Online]. Available: https://www.researchgate.net/publication/351128739_Testing_for_Normality_What_is_the_Best_Method
[15] E. M. Gandica de Roa, "Potencia y Robustez en Pruebas de Normalidad con Simulación Montecarlo," *Rev. Sci.,* vol. 5, no. 18, pp. 108–119, Jan. 2020. https://doi.org/10.29394/Scientific.issn.2542-2987.2020.5.18.5.108-119
[16] S. S. Uyanto, "An Extensive Comparisons of 50 Univariate Goodness-of-fit Tests for Normality," *Austrian Journal of Statistics*, vol. 51, no. 3, pp. 45-97, Aug. 2022. https://doi.org/10.17713/ajs.v51i3.1279
[17] A. I. Fleishman, "A Method for Simulating Non-Normal Distributions," *Psychometrika*, vol. 43, no. 4, pp. 521-532, Dec. 1978. https://doi.org/10.1007/BF02293811
[18] L. Baringhaus, R. Danschke, and N. Henze, "Recent and Classical Tests for Normality - A Comparative Study," *Communications in Statistics - Simulation and Computation*, vol. 18, no. 1, pp. 363-379, Jul. 2007. https://doi.org/10.1080/03610918908812764
[19] C. M. Jarque, and A. K. Bera, "A Test for Normality of Observations and Regression Residuals," *International Statistical Review*, vol. 55, no. 2, pp. 163-172, Aug. 1987. https://doi.org/10.2307/1403192
[20] R. D'Agostino, and E.S. Pearson, "Tests for Departure from Normality. Empirical Results for the Distributions of b2 and √b1," *Biometrika*, vol. 60, no. 3, pp. 613-622, Dec. 1973. https://doi.org/10.1093/biomet/60.3.613
[21] H. Cramér, "On the Composition of Elementary Errors: First Paper: Mathematical Deductions," *Scandinavian Actuarial Journal*, vol. 1928, no. 1, pp. 13-74, Dec. 2011. https://doi.org/10.1080/03461238.1928.10416862
[22] M. Arshad, M. Rasool, and M. Ahmad, "Anderson-Darling and Modified Anderson-Darling Tests for Generalized Pareto Distribution," *Pakistan Journal of Applied Sciences*, vol. 3, no. 2, pp. 85-88, 2003. https://doi.org/10.3923/jas.2003.85.88
[23] M. D. Esteban, M. E. Castellanos, D. Morales, and I. Vajda, "Monte Carlo Comparison of Four Normality Tests Using Different Entropy Estimates," *Communications in Statistics - Simulation and Computation*, vol. 30, no. 4, pp. 761-785, Aug. 2006. https://www.tandfonline.com/doi/full/10.1081/SAC-100107780
[24] Y. R. Gel, and J. L. Gastwirth, "A Robust Modification of the Jarque–Bera Test of Normality," *Economics Letters*, vol. 99, no. 1, pp. 30-32, Apr. 2008. https://doi.org/10.1016/j.econlet.2007.05.022
[25] D. G. Bonett, and E. Seier, "A Test of Normality with High Uniform Power," *Computational Statistics & Data Analysis*, vol. 40, no. 3, pp. 435-445, Sep. 2002. https://doi.org/10.1016/S0167-9473(02)00074-9
[26] C. Bontemps, and N. Meddahi, "Testing Normality: A GMM Approach," *Journal of Econometrics*, vol. 124, no. 1, pp. 149-186, Jan. 2005. https://doi.org/10.1016/j.jeconom.2004.02.014
[27] J. Bai, and S. Ng, "Tests for Skewness, Kurtosis, and Normality for Time Series Data," *Journal of Business & Economic Statistics*, vol. 23, no. 1, pp. 49-60, 2012. https://doi.org/10.1198/073500104000000271

[28] G. W. Snedecor, and W. G. Cochran, "The mean and standard deviation," *Statistical Methods*, 8th ed. New York, NY, USA: Wiley-Blackwell, 1989. https://www.wiley.com/en-kr/Statistical+Methods%2C+8th+Edition-p-9780813815619#description-section

[29] Z. Chen, and C. Ye, "An Alternative Test for Uniformity," *International Journal of Reliability, Quality and Safety Engineering*, vol. 16, no. 4, pp. 343-356, Jul. 2009. https://doi.org/10.1142/S0218539309003435

[30] G. Brys, M. Hubert, and A. Struyf, "A Robust Measure of Skewness," *Journal of Computational and Graphical Statistics*, vol. 13, no. 4, pp. 996-1017, Jan. 2012. http://www.jstor.org/stable/27594089

[31] S. S. Shapiro, and R. S. Francia, "An Approximate Analysis of Variance Test for Normality," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 215-216, Jun. 1972. https://doi.org/10.1080/01621459.1972.10481232

[32] J. A. Doornik, and H. Hansen, "An omnibus test for univariate and multivariate normality," *Oxf. Bull. Econ. Stat.,* vol. 70, no. s1, pp. 927–939, Dec. 2008. https://onlinelibrary.wiley.com/doi/10.1111/j.1468-0084.2008.00537.x

[33] J. M. Dufour, A. Farhat, L. Gardiol, and L. Khalaf, "Simulation-based Finite Sample Normality Tests in Linear Regressions," *The Econometrics Journal*, vol. 1, no. 1, pp. C154-C173, Jun. 2008. https://doi.org/10.1111/1368-423X.11009

[34] M. J. Blanca, J. Arnau, D. López-Montiel, R. Bono, and R. Bendayan, "Skewness and Kurtosis in Real Data Samples," *Methodology*, vol. 9, no. 2, pp. 78-84, Jan. 2013. https://doi.org/10.1027/1614-2241/a000057

[35] R. Bendayan, J. Arnau, M. J. Blanca, and R. Bono, "Comparación de los procedimientos de Fleishman y Ramberg et al. para generar datos no normales en estudios de simulación," *Anales de Psicología / Annals of Psychology,* vol. 30, no. 1, Dec. 2013. https://doi.org/10.6018/analesps.30.1.135911

[36] P. Flores Muñoz, L. Muñoz Escobar, and T. Sánchez Acalo, "Estudio de potencia de pruebas de normalidad usando distribuciones desconocidas con distintos niveles de no normalidad," *Perfiles*, vol. 1, no. 21, pp. 4-11, Jun. 2019. https://doi.org/10.47187/perf.v1i21.42

[37] R Core Team, *R: A Language and Environment for Statistical Computing*, (2024), R Foundation for Statistical Computing, Vienna, Austria, 2024. Accessed: Jan. 11, 2024. [Online]. Available: https://www.R-project.org/

CONFLICT OF INTEREST

The authors have no conflicts of interest pertaining to the findings presented in this article.

AUTHOR CONTRIBUTIONS

Cristian David Correa-Álvarez: Conceptualization, Research, Organization, and Writing of the manuscript.

Jessica María Rojas-Mora: Design and development of the research, as well as the review of the manuscript.

Antonio Elías Zumaqué-Ballesteros: Simulation aspects and the development of the manuscript.

Osnamir Elias Bru-Cordero: Organization of the code, Preparation of figures, and Review of the manuscript.