Revista Colombiana de Estadística - Applied Statistics
January 2026, volume 49, issue 1, pp. 275–291
http://doi.org/10.15446/rce.v49n1.118393

Performance and Agreement Between Some Normality Tests Under The Presence and Lack of Outliers

Desempeño y concordancia entre algunas pruebas de normalidad en presencia y ausencia de valores atípicos

Jerfson B. N. Honorio[1,a], Amanda S. Gomes[2,b], Jorge A. Sousa[2,c]

[1]Departament of de Statistics, CCEN, Federal University of Pernambuco, Recife, Brazil

[2]Departament of de Statistics, CCT, Federal University of Campina Grande, Campina Grande, Brazil

Abstract

This study evaluated the performance of various normality tests including Shapiro-Wilk, Shapiro-Francia, Anderson-Darling, Lilliefors, Cramer-von Mises, and Jarque-Bera under different conditions, both with and without the presence of outliers. Monte Carlo simulations were conducted to calculate the type I error rates, power, and the Kappa-Fleiss agreement coefficient, which measured the concordance among the tests. For normally distributed data without outliers, the Shapiro-Wilk and Shapiro-Francia tests showed the best control over the type I error rate. In contrast, with the introduction of outliers, the Lilliefors and Cramer-von Mises tests performed better. In terms of test power, the Shapiro-Wilk and Shapiro-Francia tests performed best for distributions without outliers, while the Jarque-Bera test was more robust in the presence of outliers. Overall, the results highlight the sensitivity of these tests to sample size and the presence of outliers, suggesting that Shapiro-Wilk and Shapiro-Francia are suitable for data without outliers, while Jarque-Bera may be preferred in contaminated samples. The tests showed higher concordance for exponential and lognormal distributions but lower concordance for beta, χ^2, and t-Student distributions, illustrating the complexity of normality identification across various contexts.

***Keywords*:** Concordance; Normality; Outliers; Performance; Simulation.

[a] Ph.D. E-mail: jerfson35@gmail.com
[b] Ph.D. E-mail: amanda.natalia.gomes@gmail.com
[c] Ph.D. E-mail: jorge.alves@professor.ufcg.edu.br

Resumen

Este estudio evaluó el desempeño de diversas pruebas de normalidad incluyendo Shapiro-Wilk, Shapiro-Francia, Anderson-Darling, Lilliefors, Cramer-von Mises y Jarque-Bera bajo diferentes condiciones, tanto con como sin la presencia de valores atípicos. Se realizaron simulaciones de Monte Carlo para calcular las tasas de error tipo I, la potencia y el coeficiente de concordancia Kappa-Fleiss, que mide la concordancia entre las pruebas.

Para datos distribuidos normalmente sin valores atípicos, las pruebas de Shapiro-Wilk y Shapiro-Francia mostraron el mejor control sobre la tasa de error tipo I. En contraste, con la introducción de valores atípicos, las pruebas de Lilliefors y Cramer-von Mises tuvieron un mejor desempeño. En términos de potencia, las pruebas de Shapiro-Wilk y Shapiro-Francia obtuvieron los mejores resultados para distribuciones sin valores atípicos, mientras que la prueba de Jarque-Bera fue más robusta en presencia de valores atípicos.

En general, los resultados destacan la sensibilidad de estas pruebas al tamaño de la muestra y a la presencia de valores atípicos, sugiriendo que Shapiro-Wilk y Shapiro-Francia son adecuadas para datos sin valores atípicos, mientras que Jarque-Bera puede ser preferida en muestras contaminadas. Las pruebas mostraron mayor concordancia para distribuciones exponenciales y lognormales, pero menor concordancia para distribuciones beta, χ^2 y t-Student, lo que ilustra la complejidad de identificar la normalidad en diversos contextos.

***Palabras clave*:** Concordancia; Desempeño; Pruebas de normalidad; Simulación; Valores atípicos.

1. Introduction

The assumption of data normality is common in the day-to-day work of professionals dealing with sample or experimental data, as many inferential methods rely on this assumption to ensure statistical validity when making inferences about population parameters. The normal distribution, a continuous probabilistic model, is one of the most widely used in statistics, forming the basis for robust statistical methods. Many random phenomena can be approximately described by this distribution, justifying its application across a wide range of contexts.

The probability density function for a continuous random variable X with a normal distribution is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \tag{1}$$

where μ is the mean and σ is the standard deviation (Morettin & de O. Bussab, 2010).

In statistical inference, various assumptions are essential, with normality, linearity, and homoscedasticity among the most important. This study focuses on the assumption of normality, as it is required by many statistical procedures, such as confidence interval construction, hypothesis testing, variance analysis, and statistical modeling. Verifying this assumption is thus crucial before proceeding with any analysis that depends on it (Anderson & Darling, 1952).

The (R Core Team, 2024) offers a variety of tools for verifying data normality, including graphical, numerical, and formal normality tests. A common graphical method is the `QQ-plot`, which visually inspects normality by positioning sample data along a reference line. Although useful, this technique is subjective, and for more objective conclusions, formal statistical tests are essential. All analyses in this study were conducted using (R Core Team, 2024).

This study was motivated by the interest in evaluating the performance of different normality tests including (Shapiro & Wilk, 1965), (Anderson & Darling, 1952), (Lilliefors, 1967), (Jarque & Bera, 1980), (Shapiro & Francia, 1972), and (Cramer, 1957) under two distinct conditions: data without outliers and data with outliers. Test performance was measured through type I error rates and empirical power. Additionally, the Kappa concordance coefficient was used to examine the level of agreement among tests in deciding whether to reject or retain the null hypothesis of normality (Cohen, 1960; Fleiss, 1971).

2. Theoretical Background

2.1. Literature Review

Numerous studies have evaluated the performance of normality tests under various conditions. Razali and Wah (2011) compared the Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling tests, concluding that the Shapiro-Wilk test generally provides the highest power for small sample sizes. Similarly, Yap and Sim (2011) found that while Shapiro-Wilk is superior for symmetric short-tailed distributions, the Anderson-Darling test performs well for various other distributions.

However, the performance of these tests in the presence of outliers remains a critical area of investigation. Assessing normality in contaminated samples is challenging because outliers can inflate variance or distort skewness and kurtosis, heavily impacting tests like Jarque-Bera (Thode, 2002). While previous literature has extensively covered Type I error and power (variables analyzed individually), there is a lack of studies focusing on the concordance between these tests. Understanding whether tests agree or disagree on the same contaminated sample is crucial for researchers who rely on a single test for decision-making. This study addresses this gap by employing the Kappa-Fleiss coefficient to measure agreement alongside traditional performance metrics.

2.2. Normality Tests

This section presents a theoretical foundation for the normality tests evaluated in this study. In each case, the hypotheses are as follows:

$$\begin{cases} H_0 : \text{The data follow a normal distribution;} \\ H_1 : \text{The data do not follow a normal distribution.} \end{cases} \tag{2}$$

For a sample $\{x_1, x_2, \dots, x_n\}$ of size n, the following notations are used:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad p_{(i)} = \Phi\left(\tfrac{x_{(i)} - \bar{x}}{s}\right), \tag{3}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, $p_{(i)}$ are the ordered percentiles of the standard normal distribution, and Φ is the cumulative distribution function of the standard normal distribution.

2.2.1. Jarque-Bera

The Jarque-Bera test evaluates data normality based on the skewness and kurtosis of the probability distribution of a statistical measure, based on a random sample, comparing them to those of a normal distribution. The test statistic is defined as:

$$JB = n\left(\frac{S^2}{6} + \frac{(C-3)^2}{24}\right), \tag{4}$$

where

$$S = \tfrac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 \quad \text{and} \quad C = \tfrac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 \tag{5}$$

represent the sample skewness and kurtosis, respectively. Under H_0, the JB statistic follows an asymptotically chi-squared distribution with 2 degrees of freedom (Jarque & Bera, 1980).

2.2.2. Cramer-von Mises

Proposed by Cramer (1957), this test assesses the fit of a cumulative distribution function F^* to an empirical cumulative distribution function F_n. The test statistic is given by:

$$W = \frac{1}{12n} + \sum_{i=1}^{n}\left(p_{(i)} - \frac{2i-1}{2n}\right)^2, \tag{6}$$

and detects significant differences between the distributions being compared. The p-value is calculated using the modified statistic $Z = W\left(1 + \frac{0.5}{n}\right)$, which follows a standard normal distribution.

2.2.3. Anderson-Darling

The Anderson-Darling test, similar to Cramer-von Mises, assesses the fit of a cumulative distribution function F^* to an empirical distribution F_n. This test is more sensitive to the tails of the distribution, making it ideal for cases where tail fit is critical. The test statistic is:

$$A = -n - \frac{1}{n}\sum_{i=1}^{n}[2i-1]\left[\ln(p_{(i)}) + \ln(1-p_{(n-i+1)})\right], \tag{7}$$

with the p-value calculated by $Z = A\left(1.0 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)$, following a standard normal distribution (Anderson & Darling, 1952).

2.2.4. Lilliefors

Developed by Lilliefors (1967), this test is an adaptation of the Kolmogorov-Smirnov test, used when theoretical distribution parameters are estimated from the data. The statistic D measures the maximum difference between the empirical and theoretical cumulative distribution functions:

$$D = \max\left|D^+, D^-\right|, \tag{8}$$

where

$$D^+ = \max\left\{\frac{i}{n} - p_{(i)}\right\}_{i=1,2,\ldots,n}, \quad D^- = \max\left\{p_{(i)} - \frac{i-1}{n}\right\}_{i=1,2,\ldots,n}. \tag{9}$$

The p-value is calculated using the statistic $Z = D\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}\right)$, which follows a standard normal distribution.

2.2.5. Shapiro-Wilk

Proposed by Shapiro & Wilk (1965), this test is widely used to evaluate normality in small samples. The test statistic is defined as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{10}$$

where the constants $a_1, a_2, \ldots, a_n$ are calculated as the solution of

$$(a_1, a_2, \ldots, a_n) = \frac{m^\top V^{-1}}{\left(m^\top V^{-1} V^{-1} m\right)^{1/2}}, \tag{11}$$

with $m = (m_1, m_2, \ldots, m_n)^\top$ representing the vector of expected values of the sample order statistics and V the covariance matrix of these statistics.

2.2.6. Shapiro-Francia

Shapiro & Francia (1972) proposed an alternative to the Shapiro-Wilk test, with the advantage of simpler implementation. In this test, the constants $a_1, a_2, \dots, a_n$ are determined by:

$$(a_1, a_2, \dots, a_n) = \frac{m^\top}{(m^\top m)^{1/2}}, \tag{12}$$

ignoring the covariance matrix to simplify the calculation of coefficients.

2.3. Kappa-Fleiss Concordance Coefficient

The Kappa coefficient, proposed by Cohen (1960), measures agreement between dependent sample proportions. Widely used in clinical studies, it evaluates the degree of agreement beyond chance, such as between diagnoses by different doctors or the same doctor at different times.

The Kappa coefficient is calculated as:

$$\kappa = \frac{P(O) - P(E)}{1 - P(E)}, \tag{13}$$

where $P(O)$ represents the observed proportion of agreements and $P(E)$ the expected proportion of agreements. Values of κ close to 1 indicate strong agreement, while values below 0 suggest a lack of agreement or even disagreement.

Fleiss (1971) proposed an extension of Kappa, known as Kappa-Fleiss, for more than two raters. It is calculated as:

$$\kappa = \frac{P(\bar{O}) - P(\bar{E})}{1 - P(\bar{E})}, \tag{14}$$

where $P(\bar{O})$ is the mean observed agreement proportion, and $P(\bar{E})$ is the mean expected agreement proportion. This coefficient is useful for evaluating concordance among multiple normality tests in rejecting or accepting the null hypothesis of normality.

3. Methodology

This study used specific strategies to assess type I error rates, power, and concordance among normality tests. Monte Carlo simulation random samples normal was applied to each test, with a pre-defined significance level of 5%, to verify whether the null hypothesis of normal distribution would be rejected.

If the null hypothesis was rejected for a sample generated from a normal distribution, this was considered a type I error. Conversely, if the null hypothesis was rejected for a sample obtained from a non-normal population, a correct decision was recorded. In each scenario, 10 000 repetitions were performed; the type I error rate corresponds to the proportion of incorrect decisions in the first case, and the empirical power corresponds to the proportion of correct decisions in the second. All simulations were conducted in R Core Team (2024).

3.1. Monte Carlo Simulation

Two Monte Carlo simulation experiments were performed. The first consisted of 10,000 simulations under H_0 and H_1 for different probability distributions, generating samples of varying sizes: $n = 10, 20, 30, 50, 100, 200, 300, 500$, and 1000.

The second experiment evaluated the robustness of the tests in the presence of outliers. To address the variability of contamination, we defined two distinct outlier scenarios:

- **Fixed Outliers:** Two outliers were added to each generated sample: one value equivalent to 10% of the minimum value and another equivalent to 100% of the maximum distribution value.
- **Percentage Contamination:** We replaced 5% of the observations in each sample with values generated from a Student's t-distribution with 2 degrees of freedom, simulating a heavy-tailed contamination.

The tests were implemented in R (R Core Team, 2024)using the `nortest`, `stats`, and `normtest` packages.

3.2. Type I Error

To evaluate type I error, 10 000 random samples of size n were generated from the normal distribution using the `rnorm()` function with mean 0 and standard deviation 1. When the null hypothesis of normality was rejected at a significance level of 5%, the distribution was considered erroneously classified as non-normal. The proportion of incorrect rejections for each test was calculated, representing the type I error rate.

3.3. Power

As described in Section 3, random samples from non-normal distributions were simulated to evaluate test power, i.e., the ability to correctly reject the null hypothesis when it is false.

In two experiments, 10 000 samples of size n from both symmetric and asymmetric non-normal distributions were generated. The first experiment excluded outliers, while the second incorporated outliers in the distributions.

The symmetric distributions simulated were Uniform$(0, 1)$, Beta$(2, 2)$, and t-Student(10). The asymmetric distributions included Gamma$(2, 1)$, $\chi^2(15)$, Lognormal$(0, 1)$, Exponential(1), and Weibull$(2, 1)$. The distributions were simulated using `random` functions in R Core Team (2024).

Depending on the chosen parameters for each distribution, the distance between them and the normal distribution can vary significantly, affecting test performance.

The normality tests were applied to each of the 10,000 samples generated from each distribution, and the proportion of correct null hypothesis rejections was computed. These values represent test power, which was then compared across tests.

3.4. Concordance

For each distribution and sample size, the results were computed, and test concordance was assessed using the Kappa-Fleiss coefficient (Fleiss, 1971). The `Kappam.fleiss` function in the `irr` package was used to verify the level of agreement among tests in deciding whether to reject the null hypothesis of normality. This concordance analysis was applied only to samples from non-normal distributions, allowing us to evaluate whether test power directly influences decision concordance.

4. Results and Discussions

This section presents the simulation results for assessing type I error rates, power, and concordance among normality tests.

4.1. Without Outliers

Six normality tests were evaluated: Anderson-Darling, Lilliefors, Shapiro-Francia, Cramer-von Mises, Shapiro-Wilk, and Jarque-Bera. Table 1 shows the type I error rates for different sample sizes, allowing us to observe the accuracy of the tests in correctly identifying normality when it truly exists.

The results indicate that the Shapiro-Wilk and Shapiro-Francia tests exhibited type I error rates closest to the nominal value of 5% for various sample sizes, demonstrating greater consistency across sample sizes. This behavior highlights the robustness of these tests in maintaining the significance level, especially in small samples. In contrast, the Jarque-Bera test showed greater fluctuations in type I error rate, indicating a tendency to reject the null hypothesis of normality more frequently, particularly in smaller samples.

The analysis suggests that for samples without outliers, the Shapiro-Wilk and Shapiro-Francia tests are preferable choices due to their control over type I error. Table 1 details error rate variations for each test as a function of sample size.

Table 1: Type I Error for samples without outliers across normality tests Anderson-Darling (AD), Lilliefors (LL), Shapiro-Francia (SF), Cramer-von Mises (CVM), Shapiro-Wilk (SW), and Jarque-Bera (JB).

	Tests					
Sample Size	**SW**	**AD**	**CVM**	**LL**	**SF**	**JB**
10	0.0466	0.0442	0.0424	0.0478	0.0506	0.0484
15	0.0482	0.0502	0.0496	0.0472	0.0528	0.0548
20	0.0496	0.0500	0.0500	0.0506	0.0496	0.0476
30	0.0544	0.0520	0.0514	0.0470	0.0514	0.0514
50	0.0460	0.0512	0.0542	0.0546	0.0482	0.0456
100	0.0512	0.0506	0.0498	0.0530	0.0520	0.0532
200	0.0528	0.0524	0.0530	0.0484	0.0524	0.0480
300	0.0478	0.0480	0.0434	0.0454	0.0488	0.0484
500	0.0514	0.0450	0.0482	0.0472	0.0504	0.0474
1000	0.0496	0.0468	0.0460	0.0470	0.0520	0.0500

The superior performance of the Shapiro-Wilk and Shapiro-Francia tests in controlling Type I error and maintaining high power for non-normal distributions aligns with the findings of Razali and Wah (2011). Our results corroborate that regression-and-correlation-based tests (like SW) are generally more sensitive to departures from normality than tests based on empirical distribution functions (like Lilliefors), especially in smaller samples.

Based on Figure 1 and Table 1, it was observed that the type I error rate of all tests fluctuated around the nominal level of 5% for small samples ($n \leq 30$).

As shown in Figure 1, as the sample size increases, the Shapiro-Francia and Shapiro-Wilk tests approach the nominal rate of 0.05, indicating their greater precision. Conversely, the Jarque-Bera test tends to reject the null hypothesis of normality more frequently, with fluctuations below the nominal level of 0.05. The Shapiro-Francia test, with little variation around the nominal level, proved to be one of the most accurate for controlling type I error. In contrast, the Anderson-Darling, Lilliefors, and Cramer-von Mises tests showed greater deviations from the nominal level, particularly in larger samples, tending to reject the hypothesis of normality more often. The Shapiro-Wilk test, standing out among them, showed low fluctuation around the nominal level, confirming its robustness in terms of power across different sample sizes.

Table 2 shows the variance of the tests relative to the nominal level across all samples.

Table 2: Variance of Type I Error for samples without outliers.

Test	All Samples	Small Samples ($n \leq 30$)	Large Samples ($n \geq 50$)
SW	6.4×10^{-7}	4.8×10^{-7}	2.048×10^{-6}
AD	1.024×10^{-5}	4.32×10^{-6}	3.2×10^{-6}
CVM	1.6×10^{-5}	1.45×10^{-5}	3.2×10^{-6}
LL	1.547×10^{-5}	1.825×10^{-6}	1.095×10^{-5}
SF	7.471×10^{-6}	6.453×10^{-6}	5.408×10^{-6}
JB	3.004×10^{-6}	1.613×10^{-6}	7.2×10^{-6}

In general, the Shapiro-Wilk and Shapiro-Francia tests were the most accurate. However, for small samples, the Shapiro-Francia test loses second place in precision to the Jarque-Bera test, which showed lower variance in this specific case.

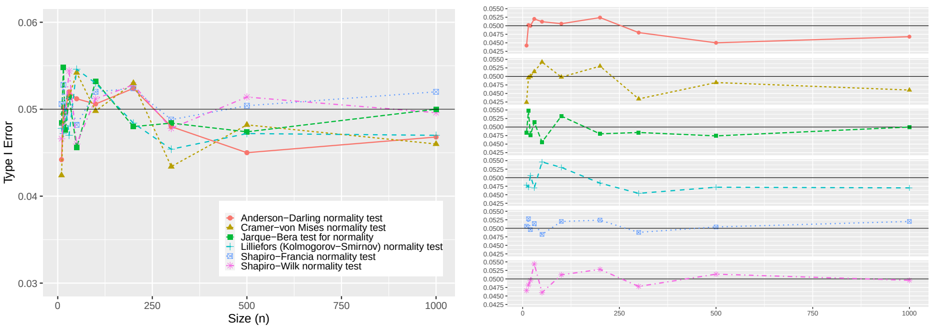


FIGURE 1: Type I Error: Normal Distribution (0,1) without Outliers.

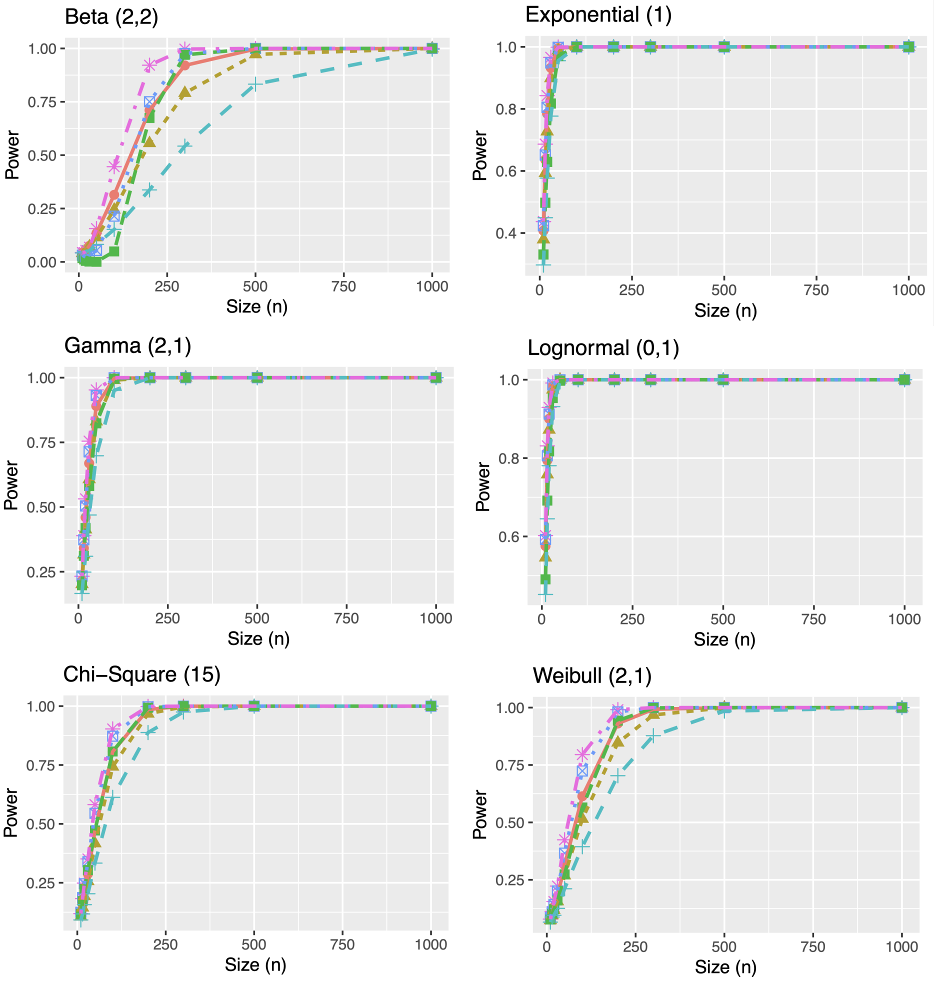
Figure 2 presents the power of each test for symmetric and asymmetric distributions with different sample sizes.

For samples from uniform and beta distributions, the Shapiro-Wilk test shows the best performance for small samples, followed by the Anderson-Darling test. Among the other tests, Jarque-Bera showed the lowest power for these distributions in small samples, while for samples larger than $n = 150$, the Lilliefors test had the lowest power.

For the t-Student distribution, the tests showed significant power differences even for larger sample sizes. This difficulty in distinguishing between the t-Student and normal distributions is due to their similar densities. Considering all sample sizes, the Jarque-Bera test performed best in identifying data from the t-Student distribution, followed by the Shapiro-Francia test, while the Lilliefors test had the lowest power.

For asymmetric distributions, the Shapiro-Wilk test was the most powerful, followed by the Shapiro-Francia test. For Weibull and χ^2 distributions, these tests stood out for sample sizes below 250. For larger samples ($n > 500$), all tests reached maximum power except the Lilliefors test, which continued to show the lowest power.

For the lognormal, exponential, and gamma distributions, all tests showed high power, with the Shapiro-Wilk and Shapiro-Francia tests maintaining superiority in all cases.



4.2. Kappa-Fleiss Concordance Coefficient for Samples without Outliers

Table 3 presents the Kappa-Fleiss concordance coefficient results for the various distributions analyzed. It is observed that, for most distributions, the tests showed a high level of concordance in deciding whether to reject the normality hypothesis. However, for beta, χ^2, and t-Student distributions, concordance among tests was lower.

This lower concordance can be seen in Figure 2, where the tests show reduced precision in identifying non-normality. This phenomenon occurs due to the similarity between these distributions and the normal distribution, making it more challenging to differentiate between them, especially for small and moderate samples.

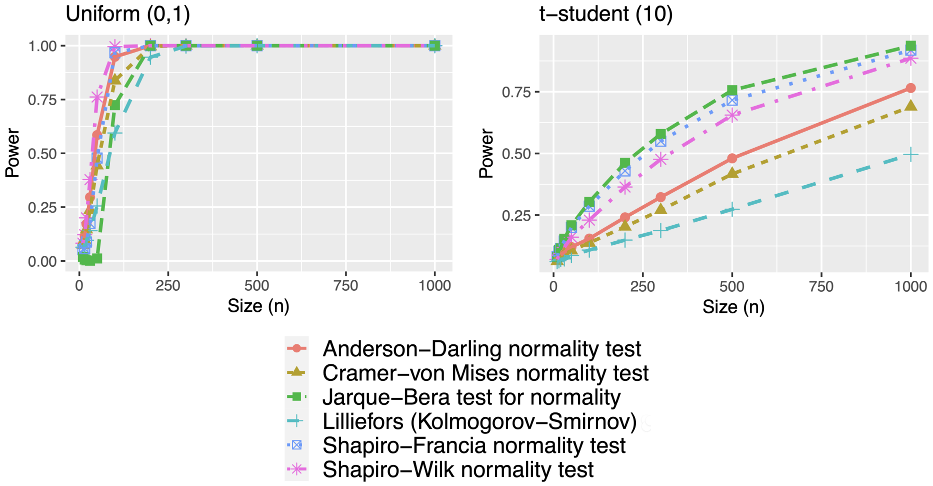


Figure 2: Power of tests for distributions under study without the presence of outliers.

Table 3: Kappa-Fleiss concordance coefficient.

Gamma	Beta	Exponential	Lognormal	χ^2	t-student	Uniform	Weibull
0.90	0.38	0.93	1.00	0.66	0.40	0.86	0.81

4.3. With Outliers

The following are the simulation results with the presence of outliers for assessing type I error rates, power, and concordance of normality tests. Two outliers were added to each sample: one value equivalent to 10% of the minimum value and another equivalent to 100% of the maximum distribution value.

Table 4: Type I Error for samples with *outliers* across normality tests Anderson-Darling (AD), Lilliefors (LL), Shapiro-Francia (SF), Cramer-von Mises (CVM), Shapiro-Wilk (SW), and Jarque-Bera (JB).

	Tests					
Sample Size	**SW**	**AD**	**CVM**	**LL**	**SF**	**JB**
10	0.16	0.15	0.15	0.14	0.20	0.27
15	0.27	0.23	0.22	0.19	0.35	0.45
20	0.38	0.28	0.26	0.20	0.47	0.57
30	0.49	0.32	0.28	0.22	0.60	0.71
50	0.64	0.35	0.29	0.21	0.76	0.83
100	0.84	0.34	0.27	0.20	0.93	0.94
200	0.96	0.31	0.23	0.16	0.99	0.97
300	0.99	0.25	0.18	0.13	1.00	0.98
500	1.00	0.20	0.15	0.10	1.00	0.99
1000	1.00	0.13	0.10	0.08	1.00	0.99

In general, all tests showed difficulty identifying data normality in the presence of outliers. As observed in Figure 3, the tests demonstrated sensitivity to outliers, deviating significantly from the nominal value of 0.05.

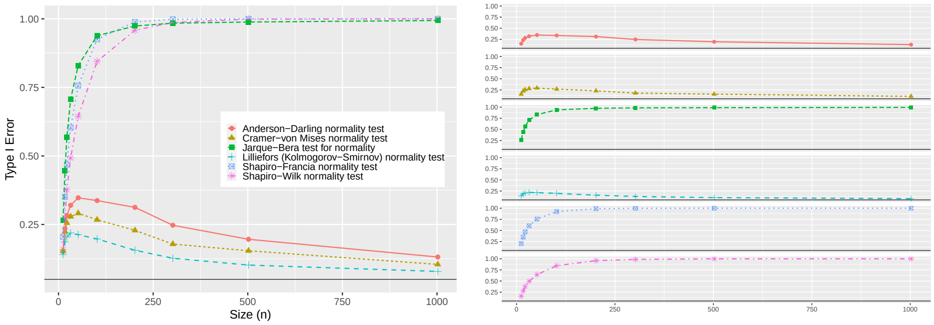


Figure 3: Type I Error: Normal Distribution $(0, 1)$ with Outliers. Source: Author

The Anderson-Darling, Lilliefors, and Cramer-von Mises tests were the least affected by outliers, showing better performance and closer proximity to the nominal value, as indicated in Table 4.

Table 5 presents the variance of the tests relative to the nominal level for all sample sizes.

Table 5: Variance of Type I Error for samples with outliers (Fixed and Percentage scenarios).

Test	All Samples	Small Samples ($n \leq 30$)	Large Samples ($n \geq 50$)
SW	4.313	0.4033	6.2050
AD	0.472	0.2028	0.4805
CVM	0.295	0.1680	0.2645
LL	0.142	0.1008	0.1125
SF	5.138	0.6721	7.0330
JB	5.760	1.0800	7.3447

Note: Values correspond to the fixed outlier scenario. Update with 5% results.

It is observed that, in general, the Anderson-Darling, Cramer-von Mises, and Lilliefors tests were the most precise in terms of type I error rates when outliers were present. The Lilliefors test exhibited the lowest variance for both small and large samples, followed closely by the Cramer-von Mises test. In contrast, the Jarque-Bera test was the most sensitive to outliers, showing a higher variability in its estimates.

In the presence of outliers, the Jarque-Bera test exhibited high power but also higher Type I error variance. This behavior is consistent with the literature (Thode, 2002), which notes that moment-based tests are highly sensitive to extreme values that distort skewness and kurtosis. Conversely, the Lilliefors test demonstrated greater resistance to outliers in terms of Type I error control, suggesting it may be a more conservative choice for contaminated datasets, despite its lower power.

The power of the tests for symmetric and asymmetric distributions in the presence of outliers is presented in Figure 4.

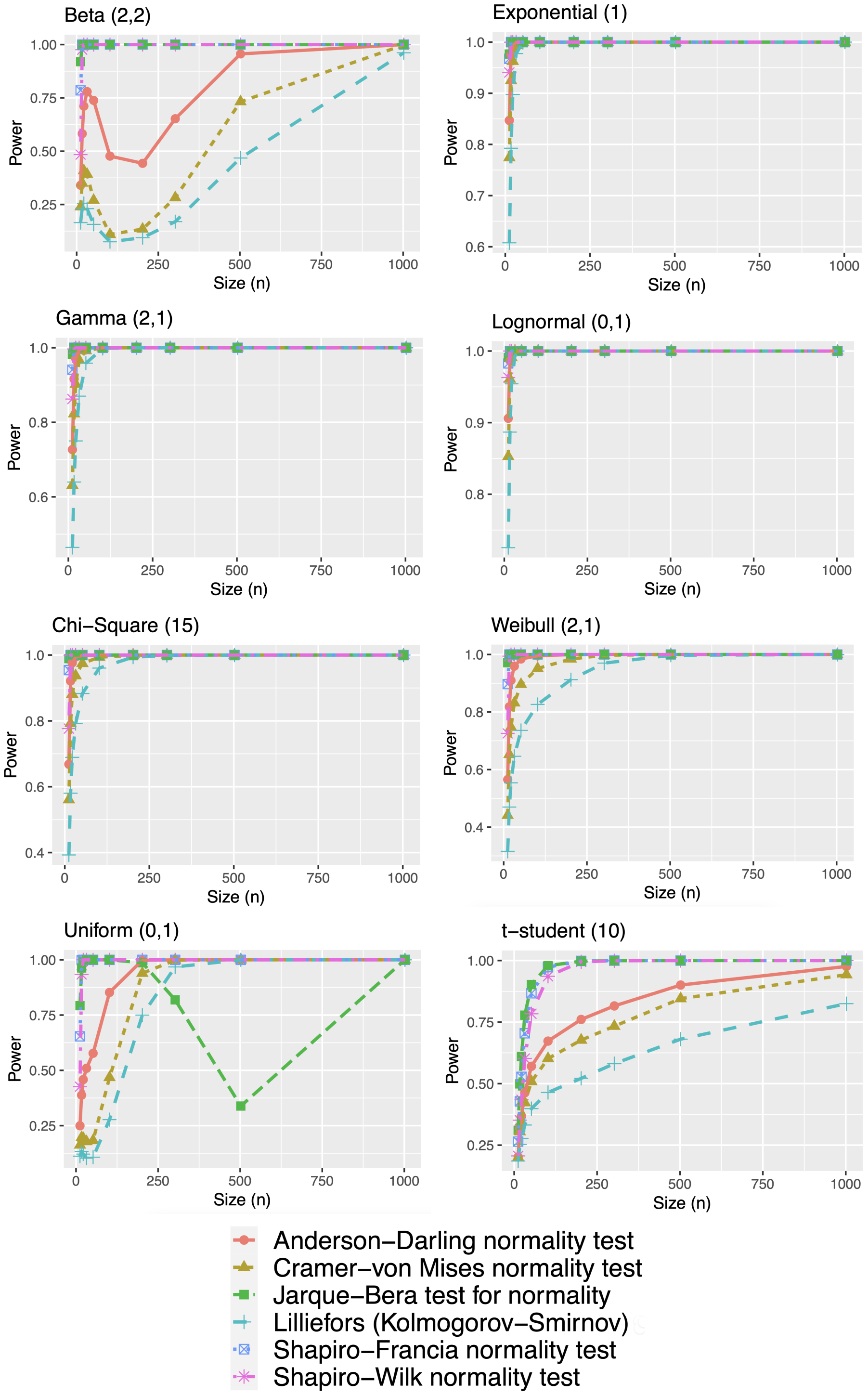


Figure 4: Power of tests for distributions under study with the presence of outliers.

For samples from a uniform distribution, the Jarque-Bera test exhibited the best performance in small samples, followed by the Shapiro-Francia test. For samples larger than 250, the Shapiro-Francia test became the most powerful, followed by the Shapiro-Wilk test, due to their stability compared to the oscillations observed in the Jarque-Bera test. Among other tests, the Lilliefors test demonstrated the lowest power.

In beta-distributed samples, the Jarque-Bera test again excelled in small samples, showing a rapid increase in power, followed by the Shapiro-Francia test. Although other tests showed oscillations as the sample size increased, these oscillations tended to grow, enhancing their power. The Lilliefors test consistently demonstrated the lowest power, even for large samples.

For samples with outliers from the t-Student distribution, the tests encountered similar difficulties as observed in the previous section. This difficulty arises from the similarity between the t-Student and normal distribution densities, making it challenging to distinguish between them. Across all sample sizes, the Jarque-Bera test performed best in the presence of outliers, followed by the Shapiro-Francia test, while the Lilliefors test demonstrated the lowest power.

For the asymmetric distributions gamma, exponential, lognormal, and χ^2, all tests exhibited high power, including the Lilliefors test. Nevertheless, the Shapiro-Wilk and Shapiro-Francia tests remained superior in all cases.

4.4. Kappa-Fleiss Concordance Coefficient for Samples with Outliers

Table 6 provides the Kappa-Fleiss concordance results for different distributions in the presence of *outliers*. It was observed that for most distributions, the tests displayed weak agreement on the decision to reject or retain the hypothesis of normality. However, maximum concordance was observed for the exponential and lognormal distributions. Lower concordance was found for other distributions, as illustrated in Figure 4, where the tests show reduced precision in identifying normality.

Table 6: Kappa-Fleiss Concordance Coefficient.

Gamma	Beta	Exponential	Lognormal	χ^2	t-student	Uniform	Weibull
0.43	0.06	1.00	1.00	0.37	0.52	0.15	0.19

The low concordance coefficients observed for distributions such as Beta and Weibull in the presence of outliers indicate a lack of consensus among methods. This suggests that contamination affects the test statistics heterogeneously, making the choice of normality test a determining factor in the conclusion of the analysis for contaminated data.

5. Conclusion

For normally distributed data without outliers, the evaluated tests demonstrated precision with low variation around the nominal 5% level. The Shapiro-Wilk and Shapiro-Francia tests provided the best performance for type I error rates. In contrast, for normally distributed data containing outliers, the Lilliefors and Cramer-von Mises tests were most effective in controlling type I error rates. It is worth noting that all tests exhibited increased variability in the presence of outliers, highlighting the difficulties that arise with extreme values.

In terms of power, the Shapiro-Wilk and Shapiro-Francia tests performed best for distributions generated without outliers, proving more suitable for this type of data. For distributions with outliers, the Jarque-Bera test achieved the highest power, followed closely by the Shapiro-Wilk test. These results underline the importance of test selection based on the data characteristics, particularly regarding sample size and the presence of outliers.

[Received: February 2025 — Accepted: December 2025]

References

Anderson, T. W. & Darling, D. A. (1952), 'Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes', *The Annals of Mathematical Statistics* **23**(2), 193–212.

Cohen, J. (1960), 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement* **20**, 37–46.

Cramer (1957), 'Mathematical theory of statistics', *The Annals of Mathematical Statistics* .

Fleiss, J. L. (1971), 'Measuring nominal scale agreement among many raters', *Psychological Bulletin* **76**(5), 378–382.

Jarque, C. M. & Bera, A. K. (1980), 'Efficient tests for normality, homoscedasticity and serial independence of regression residuals', *Economics Letters* **6**, 255–259.

Lilliefors, H. W. (1967), 'On the Kolmogorov-Smirnov test for normality with mean and variance unknown', *Journal of the American Statistical Association* **62**(318), 399–402.

Morettin, P. A. & de O. Bussab, W. (2010), *Estatística Básica*, 6ª edição, revista e atualizada edn, Saraiva.

R Core Team (2024), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Shapiro, S. S. & Francia, R. S. (1972), 'An approximate analysis of variance test for normality', *Journal of the American Statistical Association* **67**(337), 215–216. https://www.tandfonline.com/doi/abs/10.1080/01621459.1972.10481232

Shapiro, S. S. & Wilk, M. B. (1965), 'An analysis of variance test for normality (complete samples)', *Biometrika* **52**(3/4), 591–611.