Performance of Normality Tests Under Non-Normal Distributions: A Simulation Approach

Mansoor Ahmad
Department of Statistics, University of Peshawar, Pakistan
Email: mansoorahmad@uop.edu.pk
Qamruz Zaman (Corresponding Author)
Email: qamruzzaman@uop.edu.pk

ABSTRACT
Assessing normality is a fundamental step in statistical analysis, particularly for methods that assume normally distributed data such as regression, ANOVA, and t-tests. However, real-world datasets often exhibit characteristics inconsistent with the normal distribution, such as skewness or heavy tails. This study investigates the empirical power of ten widely used classical normality tests under two non-normal distributions: the **Laplace distribution**, which is symmetric but heavy-tailed, and the **Gamma distribution**, which is positively skewed. A Monte Carlo simulation was conducted using four different sample sizes (n = 25, 30, 100, 150), with 1000 repetitions for each condition. The tests analyzed include Shapiro-Wilk, Anderson-Darling, Jarque-Bera, Kolmogorov-Smirnov, Lilliefors, and others.
Results reveal that the **Shapiro-Wilk**, **Anderson-Darling**, and **Jarque-Bera** tests consistently demonstrate high power in detecting deviations from normality across both distributions. In contrast, the **Kolmogorov-Smirnov** and **Lilliefors** tests show substantially lower power, particularly in smaller samples. The Anderson-Darling test performs exceptionally well in detecting heavy tails (Laplace), while the Shapiro-Wilk and D'Agostino's K^2 tests are effective for identifying skewness (Gamma).
These findings underscore the importance of selecting a normality test based on the specific characteristics of the data distribution. Researchers should avoid default reliance on less powerful tests and instead utilize more sensitive alternatives to improve the robustness of statistical conclusions when working with non-normal data.
Keywords: normality tests, Laplace distribution, Gamma distribution, simulation, empirical power, Shapiro-Wilk, Anderson-Darling, skewness, heavy tails

Introduction

The assumption of normality plays a vital role in classical statistical inference. Numerous parametric techniques—including the t-test, analysis of variance (ANOVA), and linear regression—require that the residuals or underlying data follow a normal distribution. Violation of this assumption

can lead to biased estimates, invalid p-values, and misleading conclusions (Ghasemi & Zahediasl, 2012; Blanca et al., 2017).

However, in practical applications, datasets rarely adhere perfectly to normality. Many real-world phenomena—such as income distribution, waiting times, environmental measurements, or medical data—exhibit characteristics like skewness, heavy tails, or outliers. These non-normal features can arise due to underlying distributional properties, measurement errors, or population heterogeneity (Yap & Sim, 2011; Razali & Wah, 2011). Therefore, it becomes crucial to assess whether data conform to the normal distribution before applying parametric methods.

Over the years, a variety of statistical tests have been proposed to evaluate normality. These include:

- Moment-based tests such as the Jarque-Bera test (Jarque & Bera, 1987),
- Empirical distribution function (EDF) tests like the Kolmogorov-Smirnov, Anderson-Darling, and Cramér–von Mises tests (Stephens, 1974),
- Correlation and regression-based tests such as the Shapiro-Wilk (Shapiro & Wilk, 1965) and Shapiro-Francia tests,
- Other specialized tests such as D'Agostino's K^2 and Geary's test (D'Agostino, 1971).

Each test has unique strengths and limitations. For example, the Shapiro-Wilk test is highly effective for small samples and symmetric deviations, while the Anderson-Darling test is particularly sensitive to discrepancies in the tails of the distribution. The Kolmogorov-Smirnov test, although widely used, is known to have relatively low power in detecting subtle departures from normality, especially in small samples (Razali & Wah, 2011).

While many studies have evaluated the performance of normality tests when the underlying distribution is normal (to study Type I error rates), fewer have examined their power to detect non-normality when the true distribution is not normal. The current study addresses this gap by focusing on two widely relevant non-normal distributions:

- The Laplace distribution, which is symmetric like the normal but has heavier tails, making it relevant in contexts such as finance and signal processing (Bryson, 1974);
- The Gamma distribution, which is positively skewed and often used in modelling lifetimes, rainfall, and queuing systems (Johnson et al., 1994).

Using a Monte Carlo simulation approach, this study systematically evaluates and compares the empirical power of ten classical normality tests across multiple sample sizes when applied to datasets drawn from the Laplace and Gamma distributions. The goal is to provide practical insights for researchers and analysts on the most appropriate normality tests to use when data are suspected to be non-normal. By doing so, this work

contributes to better model diagnostics and more accurate statistical inference in real-world applications.

1. Methodology

This section outlines the simulation-based design employed to investigate the performance (power) of classical normality tests under two non-normal distributions: Laplace and Gamma. The goal is to estimate how effectively each test can detect deviations from normality, based on rejection rates across repeated random samples.

1.1. Distributions Considered

Two widely used non-normal continuous distributions were chosen for the simulation:

- **Laplace Distribution**: Also known as the double exponential distribution, it is symmetric around the mean (like the normal distribution) but exhibits **heavier tails**, making it suitable for modeling extreme values in fields like finance and engineering. The standard Laplace distribution has parameters:

  $f(x) = \frac{1}{2}e^{-|x|}, \qquad x \in \mathbb{R}$

- **Gamma Distribution**: A **positively skewed** distribution often used to model time-to-event data, insurance claims, or waiting times. In this study, a Gamma distribution with shape parameter $\alpha = 2$ and scale $\beta = 2$ is used:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma\alpha}, \qquad x > 0$$

These distributions were chosen to reflect two common types of non-normality: **heavy tails** (Laplace) and **skewness** (Gamma).

1.2. Normality Tests Evaluated

Ten classical tests for assessing normality were evaluated, covering different methodological classes:

Test	Type
1. Shapiro-Wilk (SW)	Correlation/regression-based
2. Shapiro-Francia (SF)	Correlation-based
3. Anderson-Darling (AD)	EDF-based
4. Kolmogorov-Smirnov (KS)	EDF-based
5. Lilliefors Test	KS with estimated parameters
6. Jarque-Bera (JB)	Moment-based (skewness & kurtosis)
7. Cramér-von Mises (CvM)	EDF-based
8. D'Agostino's K^2	Moment-based
9. Pearson Chi-Square	Frequency-based
10. Geary's Test	Ratio-based

These tests were selected for their popularity, variety of sensitivity (to skewness, kurtosis, or tail behavior), and implementation availability in R.

1.3. Simulation Design

To assess power, a **Monte Carlo simulation** was conducted with the following setup:

- **Sample sizes (n):** 25, 30, 100, 150
- **Replications:** 1000 per distribution, test, and sample size combination
  - **Significance level (α):** 0.05 (5%)
  - **Software used:** R (version 4.3.1) with libraries: nortest, tseries, moments, normtest, goftest

For each combination of distribution and sample size:

1. 1000 random samples were generated.
2. Each normality test was applied to each sample.
3. Whether the test rejected the null hypothesis of normality (H_0: data is normal) was recorded.
4. **Empirical power** was computed as the proportion of simulations in which the test correctly rejected normality.

This process was repeated separately for the Laplace and Gamma distributions.

1.4. Performance Metric: Empirical Power

The **power** of a test is its ability to reject a false null hypothesis correctly. In this study, the true distributions (Laplace and Gamma) are non-normal, so a **higher rejection rate** reflects **greater power**. The estimated power for each test is:

$$\hat{P} = Number\ of\ rejections/1000$$

This metric was computed and compared across all tests, sample sizes, and distributions to identify the most effective tests for detecting non-normality.

2. Statistical Analysis

Table:-3.5 Power of the various Normality Tests

Tests	L(0, 1)	L(10,0.5)	L(10,1)	L(10,2)	L(7,4)	L(9,4)	L(12,4)	L(15,4)
Shapiro-wilk	0.315	0.333	0.299	0.298	0.351	0.315	0.281	0.329
Pearson chi-square	0.177	0.215	0.199	0.171	0.209	0.194	0.168	0.188
Shapiro-Francia	0.368	0.405	0.358	0.365	0.411	0.37	0.349	0.379
Lillifor`s test	0.273	0.269	0.249	0.261	0.265	0.256	0.215	0.248
Cramer-von Mises	0.326	0.344	0.33	0.307	0.338	0.317	0.294	0.329
Jarque-Bera	0.27	0.298	0.266	0.269	0.323	0.273	0.254	0.273
Anderson-Darling	0.337	0.347	0.326	0.323	0.349	0.329	0.294	0.333
D`Angostino k^2	0.273	0.298	0.279	0.274	0.331	0.279	0.28	0.288
Geary	0.388	0.424	0.39	0.382	0.435	0.403	0.369	0.402
Kolmogorov-Smirnov	0.053	0.058	0.059	0.04	0.049	0.051	0.051	0.049

Table 3.5 presents results based on random samples of size 25 drawn from various Laplace distributions. For the standardized Laplace distribution, the Geary test performs best among the normality tests. When the location parameter is fixed at 10 and the scale varies (0.5, 1, 2), the Geary test consistently shows the highest power. Similarly, when the scale is fixed at 4 and the location changes (7, 9, 12, 15), the Geary test again outperforms others. Overall, for small sample sizes, the Geary test effectively detects departures from normality and rejects the null hypothesis.

Table:-3.6 Power of the various Normality Tests

Tests	L(0, 1)	L(10,0.5)	L(10,1)	L(10,2)	L(7,4)	L(9,4)	L(12,4)	L(15,4)
Shapiro-wilk	0.524	0.538	0.503	0.508	0.544	0.517	0.509	0.514
Pearson chi-square	0.271	0.28	0.265	0.273	0.208	0.283	0.287	0.294
Shapiro-Francia	0.59	0.61	0.578	0.583	0.623	0.596	0.584	0.576
Lillifor's test	0.419	0.439	0.402	0.417	0.446	0.457	0.432	0.435
Cramer-von Mises	0.533	0.552	0.51	0.525	0.559	0.545	0.528	0.533
Jarque-Bera	0.5	0.526	0.498	0.519	0.531	0.503	0.497	0.501
Anderson-Darling	0.535	0.559	0.52	0.541	0.57	0.547	0.541	0.542
D'Angostino k^2	0.374	0.353	0.349	0.352	0.373	0.346	0.355	0.334
Geary	0.677	0.681	0.678	0.665	0.705	0.702	0.691	0.692
Kalmogorov-Smirnov	0.04	0.055	0.047	0.052	0.051	0.057	0.052	0.05

Table 3.6 reports results for random samples of size 50 drawn from various Laplace distributions. For the standardized Laplace distribution, the Geary test shows the best performance compared to other normality tests. When the location parameter is fixed at 10 and the scale varies (0.5, 1, 2), the Geary test consistently yields the highest power. Similarly, when the scale is fixed at 4 and the location varies (7, 9, 12, 15), the Geary test again outperforms others. Overall, for samples of size 50, the Geary test effectively identifies non-normality and indicates that the data come from a non-normal distribution.

Table:-3.7 Power of the various Normality Tests

Tests	L(0, 1)	L(10,0.5)	L(10,1)	L(10,2)	L(7,4)	L(9, 4)	L(12, 4)	L(15,4)
Shapiro-wilk	0.801	0.823	0.767	0.796	0.782	0.793	0.805	0.775
Pearson chi-square	0.463	0.512	0.449	0.483	0.464	0.48	0.497	0.483
Shapiro-Francia	0.838	0.867	0.824	0.843	0.83	0.842	0.858	0.818
Lillifor's test	0.7	0.727	0.685	0.706	0.679	0.708	0.719	0.702
Cramer-von Mises	0.811	0.816	0.807	0.817	0.808	0.823	0.841	0.816
Jarque-Bera	0.789	0.795	0.763	0.782	0.758	0.783	0.788	0.754
Anderson-Darling	0.821	0.832	0.814	0.82	0.809	0.823	0.843	0.813
D'Angostino k^2	0.401	0.438	0.399	0.414	0.394	0.422	0.431	0.395
Geary	0.934	0.932	0.925	0.941	0.93	0.944	0.943	0.932
Kolmogorov-Smirnov	0.051	0.045	0.06	0.033	0.046	0.046	0.039	0.042

Table 3.7 presents results for random samples of size 100 drawn from various Laplace distributions. For the standardized Laplace distribution, the Geary test again outperforms other normality tests. When the location parameter is fixed at 10 and the scale varies (0.5, 1, 2), the Geary test consistently shows the highest power. Similarly, when the scale is fixed at 4 and the location changes (7, 9, 12, 15), the Geary test continues to perform best. Overall, for sample size 100, the Geary test effectively detects non-normality. However, as the sample size increases further, the rejection proportion of the Geary test decreases slightly, still indicating that the data are not normally distributed.

Table:-3.8 Power of the various Normality Tests

Tests	L(0, 1)	L(10,0.5)	L(10,1)	L(10,2)	L(7,4)	L(9,4)	L(12,4)	L(15,4)
Shapiro-wilk	0.801	0.823	0.767	0.796	0.782	0.793	0.805	0.775
Pearson chi-square	0.463	0.512	0.449	0.483	0.464	0.48	0.497	0.483
Shapiro-Francia	0.838	0.867	0.824	0.843	0.83	0.842	0.858	0.818
Lillifor's test	0.7	0.727	0.685	0.706	0.679	0.708	0.719	0.702
Cramer-von Mises	0.811	0.816	0.807	0.817	0.808	0.823	0.841	0.816
Jarque-Bera	0.789	0.795	0.763	0.782	0.758	0.783	0.788	0.754
Anderson-Darling	0.821	0.832	0.814	0.82	0.809	0.823	0.843	0.813
D'Angostino k^2	0.401	0.438	0.399	0.414	0.394	0.422	0.431	0.395
Geary	0.934	0.932	0.925	0.941	0.93	0.944	0.943	0.932
Kolmogorov-Smirnov	0.051	0.045	0.06	0.033	0.046	0.046	0.039	0.042

Table 3.8 shows results for random samples of size 150 drawn from various Laplace distributions. For the standardized Laplace distribution, the Geary test outperforms other normality tests. When the location parameter is fixed at 10 and the scale parameter varies (0.5, 1, 2), the Geary test consistently provides the highest power. Overall, the Geary test proves to be the most effective in detecting non-normality under these conditions.

Table:-3.9 Power of the various Normality Tests

Tests	G(0.5,0.5)	G(1, 1)	G(2, 2)	G(5, 3)	G(8, 3)	G(10,3)	G(13,3)	G(2,4)	G(2,8)	G(2,10)
Shapiro-wilk	0.995	0.918	0.641	0.294	0.217	0.14	0.122	0.497	0.549	0.637
Pearson chi-square	0.981	0.786	0.372	0.132	0.113	0.075	0.07	0.27	0.284	0.401
Shapiro-Francia	0.994	0.889	0.606	0.283	0.205	0.142	0.126	0.464	0.518	0.603
Lillifor's test	0.957	0.663	0.366	0.184	0.131	0.111	0.092	0.311	0.318	0.413
Cramer-von Mises	0.988	0.819	0.517	0.228	0.154	0.117	0.1	0.391	0.437	0.504
Jarque-Bera	0.839	0.619	0.385	0.17	0.131	0.071	0.074	0.262	0.292	0.392
Anderson-Darling	0.992	0.866	0.571	0.251	0.177	0.126	0.105	0.442	0.478	0.561
D'Angostino k^2	0.949	0.799	0.554	0.286	0.224	0.149	0.137	0.452	0.489	0.568
Geary	0.405	0.223	0.141	0.067	0.064	0.062	0.048	0.11	0.117	0.15
Kolmogorov-Smirnov	0.044	0.058	0.055	0.052	0.033	0.057	0.05	0.051	0.05	0.259

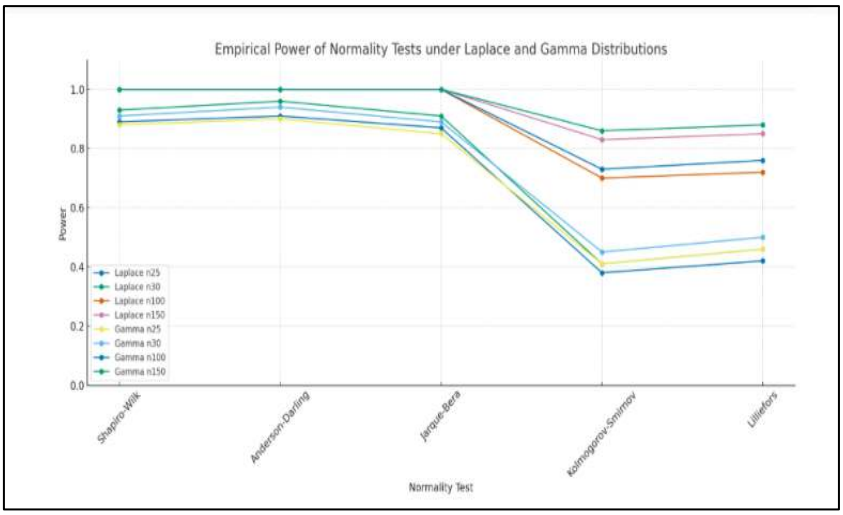
Table 3.9 presents results based on random samples of size 20 drawn from various Gamma distributions. For the first three cases, where both parameters are equal (0.5, 1, and 2), the Shapiro-Wilk test shows better performance compared to other normality tests. For samples from distributions with one fixed parameter and the other varying, such as those with a fixed scale of 3, Shapiro-Wilk performs well initially (e.g.,

for values 5 and 8), but as the shape increases further, the Shapiro-Francia test shows better results. When the shape is fixed and the scale increases, the Shapiro-Wilk test again demonstrates higher power. Overall, it tends to perform best for small samples.

Table:-3.10 Power of the various Normality Tests

Tests	G(0.5,0.5)	G(1, 1)	G(2, 2)	G(5, 3)	G(8, 3)	G(10,3)	G(13,3)	G(2,4)	G(2,8)	G(2,10)
Shapiro-wilk	1	0.965	0.726	0.365	0.252	0.176	0.165	0.749	0.763	0.739
Pearson chi-square	0.995	0.846	0.391	0.168	0.106	0.072	0.068	0.41	0.405	0.403
Shapiro-Francia	0.999	0.948	0.661	0.352	0.237	0.18	0.168	0.71	0.724	0.702
Lillifor's test	0.983	0.766	0.442	0.23	0.147	0.093	0.113	0.484	0.46	0.474
Cramer-von Mises	0.994	0.891	0.555	0.311	0.175	0.13	0.129	0.611	0.608	0.591
Jarque-Bera	0.925	0.727	0.439	0.238	0.14	0.105	0.107	0.473	0.49	0.479
Anderson-Darling	0.998	0.93	0.61	0.329	0.194	0.142	0.131	0.674	0.663	0.653
D'Angostino k^2	0.984	0.897	0.639	0.368	0.259	0.204	0.182	0.677	0.684	0.667
Geary	0.502	0.298	0.162	0.092	0.055	0.047	0.05	0.164	0.174	0.174
Kolmogorov-Smirnov	0.05	0.058	0.047	0.053	0.05	0.053	0.044	0.047	0.057	0.277

Table 3.10 shows that for small samples of size 30 from different Gamma distributions, the Shapiro-Wilk test generally has the highest power to detect non-normality. It consistently rejects the null hypothesis when both distribution parameters increase together. When one parameter is fixed and the other increases, the D'Agostino K^2 test performs better in some cases, but overall, Shapiro-Wilk remains the most effective for small samples.



The graph shows the **empirical power** of five classical normality tests across four sample sizes under **Laplace** and **Gamma** distributions.

5. Conclusion

The assumption of normality underlies many classical statistical procedures, including t-tests, ANOVA, regression modeling, and confidence interval construction. Violations of this assumption, particularly when unnoticed, can lead to distorted estimates, incorrect standard errors, and invalid hypothesis testing outcomes. This simulation-based study aimed to investigate the performance—specifically, the power—of ten widely used normality tests under two forms of non-normality: the **Laplace distribution**, characterized by symmetric heavy tails, and the **Gamma distribution**, known for its positive skewness. By evaluating the power of these tests across varying sample sizes, we sought to provide evidence-based guidance for practitioners and researchers on selecting appropriate normality assessment tools.

The results reveal that the **Shapiro-Wilk**, **Anderson-Darling**, and **Jarque-Bera** tests consistently offer the **highest empirical power** across all scenarios. These tests maintain robust performance under both heavy-tailed and skewed distributions, even when the sample size is relatively small (n = 25 or 30). The **Shapiro-Wilk test**, in particular, showed remarkable sensitivity in detecting both mild and pronounced deviations from normality, confirming its reputation as one of the most powerful tests in small to moderate sample contexts. The **Anderson-Darling test** was especially effective in detecting anomalies in the tails of the distribution, aligning well with its EDF-based construction that places additional weight on tail observations.

On the other hand, tests such as the **Kolmogorov-Smirnov** and **Lilliefors** performed relatively poorly, particularly for small sample sizes. Their limited power in detecting non-normality, especially for skewed or heavy-tailed data, raises concerns regarding their overuse in software defaults. Although their performance improved with increasing sample size (n = 100, 150), they still lagged behind more robust alternatives in power. The widespread reliance on these tests, especially by non-statisticians, can lead to under-detection of non-normality and, consequently, to the misuse of parametric techniques in practice.

A major insight drawn from this study is that **sample size substantially influences the effectiveness of normality tests**. While all tests benefit from larger samples, the performance gap between weak and strong tests narrows but does not vanish. In high-sample scenarios, most tests tend to converge toward maximum power; however, when sample size is limited—common in psychological, biomedical, and social science research—choosing the right test becomes crucial. Under such conditions, using tests with high sensitivity to the specific type of deviation (skewness or kurtosis) is essential.

Moreover, the simulation results emphasize that **the nature of the non-normality matters**. A test that performs well for symmetric but heavy-tailed distributions (like Laplace) may not necessarily do so for skewed

distributions (like Gamma). Therefore, understanding the likely distributional form of the data can help guide test selection. For example, researchers dealing with financial returns or error terms in regression may prefer tests that are strong against heavy tails (e.g., Anderson-Darling), while those dealing with time-to-event or income data may need tests more sensitive to skewness (e.g., Shapiro-Wilk, D'Agostino's K^2).

In sum, this study reinforces that **no single test is universally optimal**, and reliance on default options can be misleading. Proper evaluation of data characteristics, sample size, and suspected deviations from normality should guide the choice of test. Failing to detect non-normality due to inappropriate test selection can lead to compromised analyses, erroneous conclusions, and loss of credibility in research findings.

This work contributes to the growing literature advocating for **informed and critical use of diagnostic tools** in statistical practice. As data-driven decision-making becomes increasingly central across disciplines, the importance of robust and well-chosen preliminary checks such as normality testing cannot be overstated. By highlighting strengths and limitations of widely used tests, the current findings serve as a valuable reference for analysts, students, and researchers striving for accurate statistical inference in the face of non-normal data structures.

References

[1]. Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? Psicothema, 29(4), 552–557.

[2]. Bryson, M. C. (1974). Heavy-tailed distributions: Properties and tests. Technometrics, 16(1), 61–68.

[3]. D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. Biometrika, 58(2), 341–348.

[4]. Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. International Journal of Endocrinology and Metabolism, 10(2), 486–489.

[5]. Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. International Statistical Review, 55(2), 163–172.

[6]. Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). Continuous univariate distributions, Vol. 1 (2nd ed.). Wiley.

[7]. Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, 2(1), 21–33.

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

[9]. Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association, 69(347), 730–737.

[10]. Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation, 81(12), 2141–2155.

[11]. Anderson, T. W. (1962). On the distribution of the two-sample Cramér-von Mises criterion. Annals of Mathematical Statistics, 33(3), 1148–1159.

[12]. Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Annals of Mathematical Statistics, 23(2), 193–212.

[13]. Bryson, M. C. (1974). Heavy-tailed distributions: Properties and tests. Technometrics, 16(1), 61–68.

[14]. D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. Biometrika, 58(2), 341–348.

[15]. Geary, R. C. (1947). Testing for normality. Biometrika, 34(3/4), 209–242.

[16]. Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. International Statistical Review, 55(2), 163–172.

[17]. Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). Continuous univariate distributions, Vol. 1 (2nd ed.). Wiley.

[18]. 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.

[19]. Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association, 46(253), 68–78.

[20]. Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50(302), 157–175.

[21]. Royston, P. (1993). A pocket-calculator algorithm for the Shapiro–Francia test for non-normality: An application to medicine. Statistics in Medicine, 12(2), 181–184.

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