Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 16(1), 127-140, 2026
Journal of the Institute of Science and Technology, 16(1), 127-140, 2026

ISSN: 2146-0574, eISSN: 2536-4618

Biyosistem Mühendisliği / *Biosystem Engineering* **DOI: 10.21597/jist.1846196**

Araştırma Makalesi / *Research Article*

Geliş tarihi / *Received*: 21.12.2025 **Kabul tarihi / *Accepted*: 15.01.2026**

Atıf İçin: Huyut, M. T. (2026). Normallik Testinde Örneklem Büyüklüğünün Yeterliliğine İlişkin Ampirik ve Asimptotik Perspektifler: Bir Monte Carlo Çalışması. *Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 16(1), 127-140.

To Cite*:* Huyut, M. T. (2026). Empirical and Asymptotic Perspectives on Sample Size Adequacy in Normality Testing: A Monte Carlo Study. *Journal of the Institute of Science and Technology*, 16(1), 127-140.

Normallik Testinde Örneklem Büyüklüğünün Yeterliliğine İlişkin Ampirik ve Asimptotik Perspektifler: Bir Monte Carlo Çalışması

Mehmet Tahir HUYUT[1*]

Öne Çıkanlar:

- Kritik değerlerdeki sonlu örneklem bozulmaları, normallik testlerinin ampirik gücünü önemli ölçüde etkiler.
- Örneklem büyüklüğüne bağlı güç artışı, normallikten sapmanın simetrik ya da asimetrik olmasına göre değişir.
- Simetrik sapmalarda orta büyüklükte örneklemler yeterli olabilirken, asimetrik alternatiflerde kararlı güç için daha büyük örneklemler gereklidir.

Anahtar Kelimeler:

- Normallik testleri,
- Sonlu örneklem etkileri,
- Ampirik ve asimptotik kritik değerler,
- Örneklem boyutu yeterliliği,
- Monte Carlo simülasyonu

ÖZET:

Normallik testleri istatistiksel uygulamalarda yaygın olarak kullanılmaktadır; ancak, örneklem büyüklüğü, kritik değer kalibrasyonu ve dağılım yapısı arasındaki etkileşimle şekillenen sonlu örneklem davranışları yeterince anlaşılmamıştır. Bu çalışma, örneklem büyüklüğünün ampirik ve asimptotik kritik değerlerin güvenilirliğini nasıl etkilediğini ve bunun da simetrik ve asimetrik normallikten sapmalar altında yaygın olarak kullanılan normallik testlerinin ampirik gücünü nasıl şekillendirdiğini araştırmaktadır. On altı yaygın olarak kullanılan normallik testi için büyük ölçekli bir Monte Carlo simülasyon çalışması yapılmıştır. Ampirik ve asimptotik kritik değerler, asimptotik referans değerle birlikte n=25, 50, 100 ve 500 örneklem büyüklüklerinde değerlendirilmiştir. Ampirik güç, α=0,05 ve α=0,10 anlamlılık seviyelerinde değerlendirilmiş ve sonuçlar, yapısal olarak benzer simetrik ve asimetrik alternatif dağılımlar üzerinden ortalama alınarak özetlenmiştir. Küçük ve orta örneklem büyüklüklerinde, çeşitli testler için ampirik ve asimptotik kritik değerler arasında önemli farklılıklar gözlemlenmiştir. Bu farklılıklar doğrudan heterojen güç davranışına dönüştü. Simetrik alternatiflerde, birçok test orta düzeydeki örneklem büyüklüklerine kadar hızlı güç kazanımları sergiledi, ardından belirgin bir doygunluk gözlendi. Buna karşılık, asimetrik alternatifler gecikmeli güç birikimi gösterdi ve anlamlı kazanımlar daha büyük örneklem büyüklüklerinde de devam etti. Anlamlılık düzeyinin artırılması gücü eşit şekilde artırdı ancak göreceli test sıralamalarını değiştirmedi. Normallik testinde örneklem büyüklüğü etkileri dağılıma büyük ölçüde bağlıdır ve yalnızca asimptotik teori ile yeterince yakalanamaz. Simetrik sapmaları tespit etmek için orta düzeydeki örneklemler yeterli olabilirken, asimetrik sapmalar güvenilir güç elde etmek için daha büyük örneklemler gerektirir. Bu bulgular, normallik testinde sonlu örneklem değerlendirmelerinin önemini vurgulamakta ve daha bilinçli test seçimi için mekanik bir temel sağlamaktadır.

Empirical and Asymptotic Perspectives on Sample Size Adequacy in Normality Testing: A Monte Carlo Study

Highlights:

- Finite-sample distortions in critical values substantially affect the empirical power of normality tests.
- Power gains with increasing sample size depend strongly on whether departures from normality are symmetric or asymmetric.
- While moderate sample sizes may suffice for symmetric alternatives, asymmetric deviations require larger samples to achieve stable power.

Keywords:

- Normality tests,
- Finite-sample effects,
- Empirical and asymptotic critical values,
- Sample size adequacy,
- Monte Carlo simulation

ABSTRACT:

Normality tests are widely used in statistical practice; however, their finite-sample behavior—shaped by the interaction between sample size, critical value calibration, and distributional structure—remains insufficiently understood. This study investigates how sample size governs the reliability of empirical and asymptotic critical values and, in turn, shapes the empirical power of widely used normality tests under symmetric and asymmetric departures from normality. A large-scale Monte Carlo simulation study was conducted for sixteen widely used normality tests. Empirical and asymptotic critical values were evaluated across sample sizes n=25, 50, 100 and 500, together with the asymptotic benchmark. Empirical power was assessed at significance levels α=0.05 and α=0.10, with results summarized by averaging across structurally similar symmetric and asymmetric alternative distributions. Substantial discrepancies between empirical and asymptotic critical values were observed for several tests at small and moderate sample sizes. These discrepancies translated directly into heterogeneous power behavior. Under symmetric alternatives, many tests exhibited rapid power gains up to moderate sample sizes, followed by clear saturation. In contrast, asymmetric alternatives showed delayed power accumulation, with meaningful gains persisting at larger sample sizes. Increasing the significance level increased power uniformly but did not alter relative test rankings. Sample size effects in normality testing are strongly distribution-dependent and cannot be adequately captured by asymptotic theory alone. Moderate samples may suffice for detecting symmetric deviations, whereas asymmetric departures require larger samples to achieve reliable power. These findings underscore the importance of finite-sample considerations in normality testing and provide a mechanistic basis for more informed test selection.

[1] Mehmet Tahir HUYUT (**Orcid ID:** 0000-0002-2564-991X), Department of Biostatistics and Medical Informatics, Faculty of Medicine, Erzincan Binali Yıldırım University, 24000 Erzincan, Türkiye
***Corresponding Author:** Mehmet Tahir HUYUT, e-mail: tahir.huyut@erzincan.edu.tr
Ethics Committee Approval: Ethics approval was not required for this study as it involves only simulation-based analyses using synthetic data generated under predefined statistical models.

INTRODUCTION

Assessing the assumption of normality plays a central role in statistical inference, as many classical procedures rely on normality either explicitly or asymptotically. Parametric hypothesis tests, confidence intervals, and regression-based methods often depend on this assumption, particularly in small and moderate samples. Consequently, a wide range of goodness-of-fit and normality tests has been proposed, including tests based on empirical distribution functions, moment-based measures, and correlation or regression principles (Anderson & Darling, 1952; Shapiro et al., 1965; Stephens, 1974; Stephens, 1976). Despite their widespread use, guidance on how these tests behave as a function of sample size remains incomplete.

Most normality tests are implemented using *asymptotic critical values***,** derived under the assumption of large samples. In practice, however, normality testing is frequently applied in finite-sample settings, where the asymptotic approximation may be inaccurate. In such cases, the null distribution of the test statistic can differ substantially from its limiting form, leading to distorted rejection thresholds and unreliable inference (D'Agostino & Stephens, 1986; Filliben, 1975). These finite-sample discrepancies may result in conservative or liberal testing behavior, directly affecting empirical power.

Although numerous studies have compared the empirical power of normality tests under selected alternative distributions, the *mechanisms driving power variation across sample sizes* are far less understood. Power is not determined solely by the sensitivity of a test statistic to deviations from normality; it is also shaped by the behavior of its critical values. In finite samples, mismatches between empirical and asymptotic critical values can suppress power even when a test statistic is theoretically well-designed (Jarque & Bera, 1987; Royston, 1982). As sample size increases, critical values may stabilize at different rates depending on the test, leading to heterogeneous convergence patterns.

From a theoretical standpoint, increasing the sample size should improve both the approximation of the null distribution and the ability of a test to detect departures from normality. Empirically, however, this improvement is often *nonlinear***.** Some tests exhibit rapid gains in power with modest increases in sample size, while others require substantially larger samples before asymptotic behavior becomes reliable (Hosking, 1990). Moreover, beyond a certain point, additional increases in sample size may yield only marginal improvements in power, raising the practical question of *sample size adequacy* in normality testing.

The concept of sample size adequacy has received limited attention in the context of normality tests. Existing guidelines often rely on heuristic thresholds or general rules of thumb, without accounting for test-specific behavior or finite-sample distortions (Thode, 2002). As a result, practitioners may either overestimate the reliability of asymptotic critical values in small samples or unnecessarily increase sample sizes beyond the point where meaningful power gains are achieved.

Monte Carlo simulation provides a natural framework for addressing these issues. By explicitly estimating empirical critical values and power across a range of sample sizes, simulation studies allow the finite-sample behavior of normality tests to be examined in detail (Razali & Wah, 2011; Romao et al., 2010; Yap & Sim, 2011). Such an approach makes it possible to quantify the divergence between empirical and asymptotic critical values, to assess the rate at which convergence occurs, and to evaluate how these factors influence empirical power. Importantly, this perspective shifts the focus from ranking tests under fixed conditions to understanding *how and why* test performance evolves with sample size.

The present study adopts this mechanistic perspective to investigate *sample-size–dependent behavior in normality testing***.** Rather than emphasizing distribution-specific power comparisons, the

analysis focuses on the interaction between critical value behavior and empirical power. Using an extensive Monte Carlo design, we examine how empirical and asymptotic critical values differ across sample sizes, how quickly asymptotic approximations become reliable, and how marginal power gains change as the sample size increases.

The primary contributions of this work are threefold. First, it provides a systematic assessment of empirical versus asymptotic critical values for a broad set of normality tests across increasing sample sizes. Second, it links finite-sample critical value distortion to observed power behavior, offering a clear explanation for why some tests perform poorly in small samples despite favorable asymptotic properties. Third, it introduces an operational view of *sample size adequacy,* identifying conditions under which further increases in sample size yield only marginal improvements in power.

The remainder of the article is structured as follows: First, it presents the conceptual framework underlying empirical and asymptotic critical values and their role in finite sample inference. It then describes the Monte Carlo simulation design. The next section examines the behavior of critical values across sample sizes, followed by an analysis of the effects of sample size on empirical power and marginal gains. Following this, a mechanistic interpretation of the findings is presented, and practical implications are discussed. Finally, it concludes with recommendations for applied normality testing and guidance for future research.

MATERIALS AND METHODS

Normality Tests and Testing Framework

Normality testing is formulated as a hypothesis testing problem in which the null hypothesis states that the observed data arise from a normal distribution. Let $X_1,\ldots,X_n$ denote an independent and identically distributed sample of size n. For a given normality test, a test statistic T_n is computed from the sample and compared with a critical value c_α, corresponding to a nominal significance level α. The null hypothesis of normality is rejected when $T_n > c_\alpha$.

In practical applications, critical values are most often obtained from the *asymptotic distribution* of the test statistic. As the sample size increases, many normality test statistics converge in distribution to a known limiting form, allowing rejection thresholds to be tabulated independently of n. These asymptotic critical values are widely used in statistical software and routine data analysis. However, their validity relies on sufficiently large sample sizes.

In finite samples, the null distribution of T_n may deviate from its asymptotic approximation. As a result, *empirical (finite-sample) critical values*, defined as quantiles of the exact or simulated null distribution for a fixed sample size, may differ from asymptotic thresholds. This discrepancy can directly influence both type-I error control and empirical power. The present study explicitly distinguishes between empirical and asymptotic critical values and examines how their divergence evolves with increasing sample size. In this study, normality tests used are categorized into three different groups based on their structures, and the results are reported from this perspective.

The EDF-based tests:

1) Kolmogorov–Smirnov test with Lilliefors correction (K-S),

2) Anderson–Darling test (AD),

3) Cramér–von Mises test (C_VM),

4) Watson's U^2 test (U^2),

5) Zhang–Wu Z_C test (Z_C),

6) Zhang–Wu ZA test (Z_A), and

7) Glen–Leemis–Barr test (P_S).

Moment-based tests:

8) D'Agostino–Pearson K^2 test (K^2),

9) Jarque–Bera test (JB),

10) Doornik–Hansen omnibus test (DH) , and

11) a robust Jarque–Bera–type test (RJB) based on the Gel–Gastwirth formulation.

12) Hosking test based on trimmed L-moments ($T_{Lmom}^{(3)}$)

Regression- and correlation-based tests:

13) Shapiro–Wilk test (W),

14) Shapiro–Francia test (W_{SF}),

15) Chen–Shapiro regression-based normality test (CS) and

16) Filliben probability plot correlation coefficient test (Fr).

Empirical and Asymptotic Critical Values

Empirical critical values were obtained through Monte Carlo simulation under the null hypothesis of normality. For each test and sample size, repeated samples were generated from the standard normal distribution, and the corresponding test statistics were computed. The empirical critical value at level α\alphaα was defined as the *(1−α)-quantile* of the simulated null distribution of the test statistic.

Asymptotic critical values were taken from the limiting distributions associated with each test, as commonly reported in the literature or implemented in standard statistical software. These values serve as reference thresholds representing the large-sample behavior of the tests. By comparing empirical and asymptotic critical values across sample sizes, it becomes possible to assess the rate at which asymptotic approximations become reliable.

The difference between empirical and asymptotic critical values reflects *finite-sample distortion*. If the asymptotic critical value is smaller than the empirical one, the test becomes liberal in finite samples; if larger, the test becomes conservative. Such distortions are expected to diminish as sample size increases, but the speed of convergence is test-specific. This study investigates these convergence patterns in detail.

Sample Size Grid and Monte Carlo Design

The behavior of normality tests was examined across a grid of sample sizes representing small, moderate, and large samples. . For each test, sample size, and significance level, 1.000.000 Monte Carlo replications were performed to ensure stable estimation of empirical critical values and power. Specifically, sample sizes n = 25, 50, 100, 500 were considered, along with the asymptotic benchmark corresponding to $n\to\infty$. This range allows both early finite-sample effects and later stabilization toward asymptotic behavior to be observed.

For each combination of test, sample size, and significance level *($\alpha=0.05$ and $\alpha=0.10$)*, Monte Carlo simulations were conducted to estimate empirical critical values and empirical power. A large number of simulation replications was used to ensure stable estimation of quantiles and rejection probabilities. All simulations were performed independently for each test and sample size.

Monte Carlo simulation provides a flexible and transparent framework for studying finite-sample behavior, as it allows the exact sampling distribution of test statistics to be approximated without relying on asymptotic assumptions. This approach is particularly well suited for investigating how critical values and power evolve as functions of sample size.

The significance levels $\alpha = 0.05$ and $\alpha = 0.10$ were selected as they represent the most commonly used thresholds in applied statistical practice. More stringent levels such as $\alpha = 0.01$ were not considered, as they tend to induce highly conservative behavior in finite samples and may obscure the assessment of sample size adequacy, which constitutes the primary focus of this study.

Empirical Power Evaluation

Empirical power was defined as the probability of rejecting the null hypothesis of normality when the data were generated from non-normal alternative distributions. For each test and sample size, power was estimated as the proportion of Monte Carlo replications in which the test statistic exceeded the corresponding critical value.

To focus on sample size effects rather than distribution-specific idiosyncrasies, empirical power results were summarized by averaging across sets of alternative distributions. Alternatives were grouped according to their distributional structure, allowing power behavior to be examined separately under symmetric and asymmetric departures from normality. This aggregation provides a concise representation of test performance while retaining sensitivity to broad distributional features.

To facilitate interpretation of sample size effects, empirical power results were summarized by averaging across sets of alternative distributions sharing similar structural properties. Accordingly, non-normal alternatives were grouped into symmetric and asymmetric classes, and for each test and sample size, the reported average power corresponds to the mean rejection probability across all distributions within the respective group. These alternatives were designed to represent a broad range of departures from normality, capturing variations in tail behavior, skewness, and kurtosis rather than emphasizing specific distributional families. For completeness, the exact distributional specifications used in the simulation design follow the same framework as a related power-focused study by the authors and are therefore not reproduced here.

Marginal Power Gain and Sample Size Adequacy

To quantify the effect of increasing sample size, *marginal power gain* was defined as the increase in empirical power obtained when moving from one sample size to the next larger one. Marginal gains provide a direct measure of how much additional information is gained by increasing the sample size.

A test was considered to approach *sample size adequacy* when further increases in sample size resulted in only negligible marginal power gains. This concept emphasizes that larger samples do not always translate into meaningful improvements in performance and that the point of diminishing returns may differ substantially across tests.

By jointly analyzing empirical and asymptotic critical values, empirical power, and marginal power gains, the present study provides a comprehensive assessment of sample size effects in normality testing.

RESULTS AND DISCUSSION

Empirical and Asymptotic Critical Value Behavior Across Sample Sizes

Figures 1,2 and 3 present a comprehensive comparison of empirical and asymptotic critical values across increasing sample sizes for all normality tests considered in this study. Results are displayed for

four nominal percentiles (90%, 95%, 97.5%, and 99%) and sample sizes n=25,50,100,500, together with the asymptotic reference corresponding to n→∞. Figure 1 focuses on tests based on the empirical distribution function (EDF), whereas Figure 2 and Figure 3 summarizes moment-based tests and regression- or correlation-based normality tests.

EDF-based tests

As shown in Figure 1, EDF-based tests exhibit heterogeneous finite-sample behavior with respect to their critical values. The Lilliefors test displays the most pronounced deviation between empirical and asymptotic critical values at small sample sizes. For n=25, empirical critical values are substantially larger than their asymptotic counterparts across all percentiles, indicating a strong finite-sample distortion. This discrepancy decreases monotonically as the sample size increases, but noticeable differences persist even at n=100, particularly at the upper percentiles. Convergence toward asymptotic critical values becomes clearly apparent only at n=500, suggesting that asymptotic approximations for the Lilliefors test may be unreliable in small to moderate samples.

In contrast, Anderson–Darling, Cramér–von Mises, and Watson U^2 tests demonstrate markedly faster convergence to their asymptotic critical values (Figure 1). For these tests, empirical and asymptotic thresholds are already closely aligned at n=50 with only minor residual discrepancies observed at the 99% percentile. This behavior indicates a comparatively stable finite-sample distribution and supports the practical use of asymptotic critical values even at moderate sample sizes.

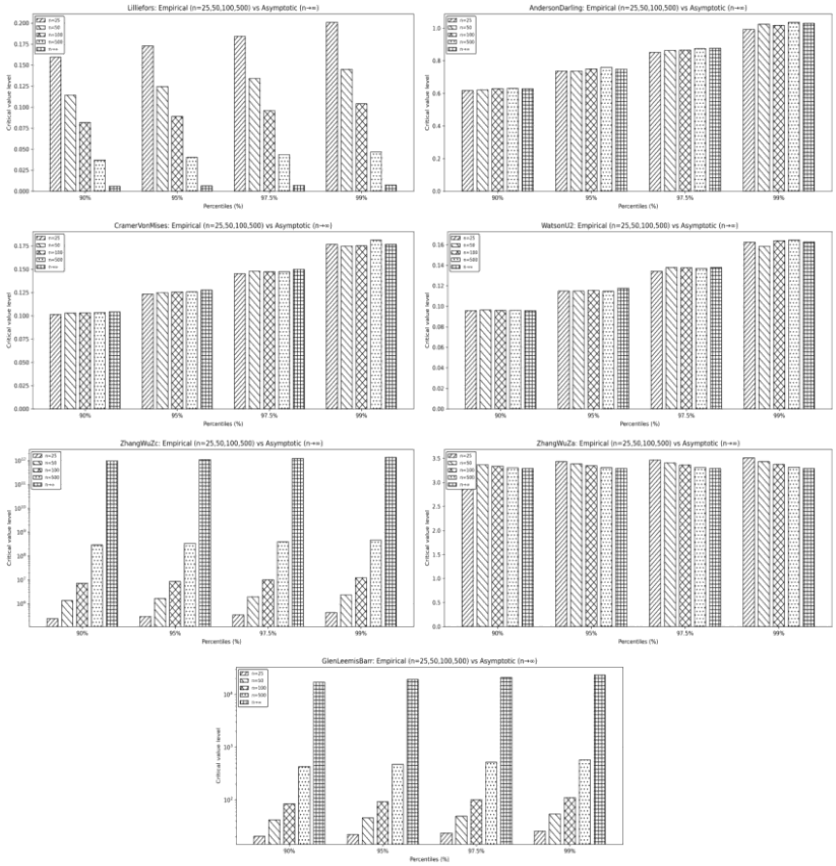

Figure 1. Empirical and asymptotic critical values for EDF-based normality tests.

The Zhang–Wu statistics exhibit a more complex pattern. While one variant shows gradual stabilization with increasing sample size, the other displays extremely large asymptotic critical values relative to the empirical ones, reflecting the heavy-tailed nature of the corresponding null distribution. This highlights that, for certain EDF-based statistics, asymptotic critical values may not provide a meaningful approximation to finite-sample behavior, even at large nnn.

Finally, the Glen–Leemis–Barr test shows systematic inflation of asymptotic critical values relative to empirical estimates across all percentiles (Figure 1). Although convergence improves with increasing sample size, the magnitude of the asymptotic values suggests that finite-sample calibration remains important for this test family.

Moment-based and regression-based tests

Figure 2 summarizes the behavior of moment-based tests (D'Agostino–Pearson K^2, Jarque–Bera, Doornik–Hansen, and Robust Jarque–Bera) as well as regression- and correlation-based tests (Shapiro–Wilk, Shapiro–Francia, Chen–Shapiro, and Filliben).

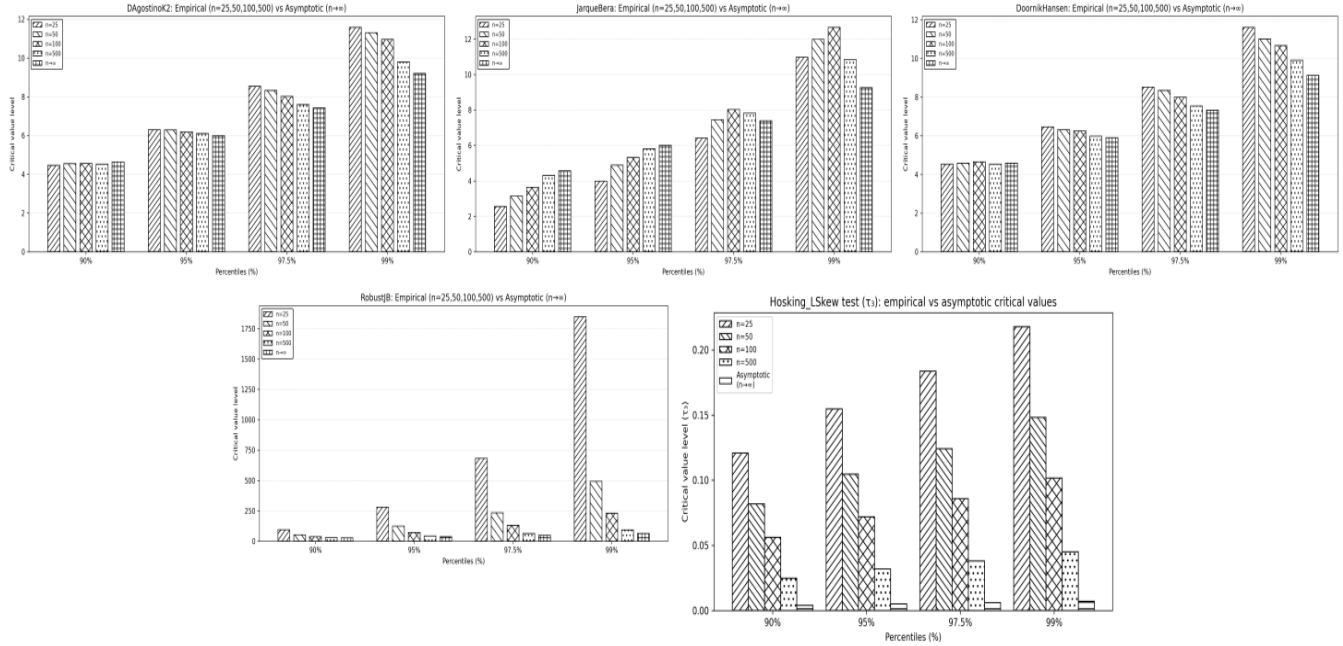

Figure 2. Empirical and asymptotic critical values for moment-based normality tests.

Moment-based tests exhibit substantial finite-sample sensitivity, particularly at high percentiles. The Jarque–Bera test shows pronounced inflation of empirical critical values at small sample sizes, especially at the 99% percentile, where empirical thresholds exceed asymptotic values by a wide margin (Figure 2). Although this discrepancy decreases with increasing nnn, convergence remains slow relative to EDF-based tests, indicating that asymptotic approximations for Jarque–Bera may substantially underestimate rejection thresholds in small samples.

The D'Agostino–Pearson and Doornik–Hansen tests demonstrate smoother convergence patterns. Empirical and asymptotic critical values begin to align at n=100, particularly at lower percentiles, although residual discrepancies persist at extreme quantiles. Robust Jarque–Bera exhibits the largest finite-sample distortion among moment-based tests, with asymptotic critical values substantially exceeding empirical estimates across all sample sizes, suggesting overconservatism when asymptotic thresholds are applied.

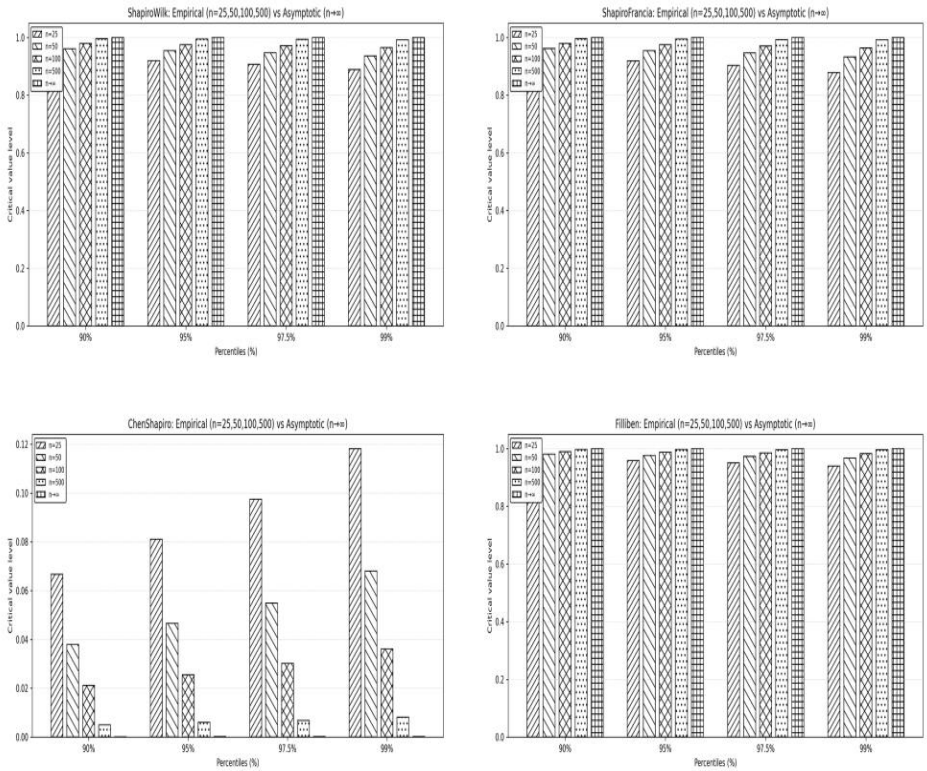

Figure 3. Empirical and asymptotic critical values omnibus normality tests.

Regression- and correlation-based tests display markedly different behavior. Shapiro–Wilk, Shapiro–Francia, Filliben, and Chen–Shapiro show minimal deviation between empirical and asymptotic critical values even at n=25 (Figure 3). Across all percentiles, empirical thresholds rapidly stabilize and closely match asymptotic values, indicating strong finite-sample robustness. This rapid convergence suggests that the null distributions of these statistics are well-approximated by their asymptotic forms at relatively small sample sizes.

Implications for finite-sample testing

Taken together, Figures 1,2 and 3 demonstrate that convergence of empirical to asymptotic critical values is highly test-specific. While some tests achieve near-asymptotic behavior at moderate sample sizes, others retain substantial finite-sample distortion even at n=100 or beyond. These differences provide a mechanistic explanation for the heterogeneous power behavior observed in subsequent analyses: tests with inflated or deflated critical values in finite samples inherently exhibit altered rejection probabilities under alternatives.

Importantly, the results indicate that reliance on asymptotic critical values may lead to misleading inference for certain normality tests in small to moderate samples, whereas other tests appear largely insensitive to finite-sample effects. This distinction motivates the power comparisons presented in the following sections, where the interaction between sample size, critical value behavior, and empirical power is examined in detail.

Sample size effects on empirical power under symmetric and asymmetric alternatives

Following the examination of empirical versus asymptotic critical values, we next investigated how sample size influences empirical power across different classes of distributional departures from normality. To facilitate interpretation, power results were summarized by averaging rejection probabilities across symmetric and asymmetric alternatives separately (Table 1).

Table 1. Average empirical power (mean rejection probability) across symmetric and asymmetric alternative distributions for all sample sizes.

Distribution	Sample size	K-S	AD	C_VM	U^2	Z_C	Z_A	P_S	K^2	JB	DH	RJB	$T^{(3)}_{Lmom}$	W	W_{SF}	CS	F_r
Simetrik (α=0.05)	25	41.7	47.2	45.5	45.9	38.9	15.3	41.7	44.2	37.9	44.2	35.5	27.9	48.2	45.9	45.9	45.9
	50	52.8	60.9	58.4	59.5	46.4	21.4	52.8	59.7	46.2	59.7	43.4	29.9	62.0	59.3	59.3	59.3
	100	63.8	72.4	69.7	70.9	53.4	24.4	63.8	72.7	68.9	72.7	52.8	31.9	74.5	73.2	73.2	73.2
	500	85.1	89.3	88.2	88.5	54.6	25.3	85.1	84.9	85.6	84.9	83.3	34.7	90.9	91.5	91.5	91.5
Asimetrik (α=0.05)	25	43.4	50.9	48.4	46.2	35.1	5.2	43.4	43.9	46.0	43.9	49.9	51.9	54.9	52.3	52.3	52.3
	50	58.6	67.5	64.2	61.4	43.7	4.8	58.6	59.0	60.1	59.0	65.9	68.2	71.7	68.9	68.9	68.9
	100	72.2	81.4	77.6	74.1	54.1	5.1	72.2	79.6	81.5	79.6	81.7	83.4	87.2	85.2	85.2	85.2
	500	94.9	96.9	96.4	95.2	66.4	7.2	94.9	97.8	97.8	97.8	97.7	96.8	97.9	97.8	97.8	97.8
Simetrik (α=0.10)	25	49.2	54.7	53.1	53.7	44.1	18.9	49.2	52.2	45.5	52.2	41.1	33.6	55.3	53.3	53.3	53.3
	50	59.2	66.7	64.2	64.7	50.9	23.7	59.2	66.9	62.4	66.9	49.7	36.3	68.6	66.4	66.4	66.4
	100	70.1	77.2	75.1	75.9	55.9	25.6	70.1	76.3	74.9	76.3	61.3	37.2	78.9	77.7	77.7	77.7
	500	88.2	90.9	90.1	90.3	55.5	25.7	88.2	86.2	86.5	86.2	85.4	39.5	92.0	92.5	92.5	92.5
Asimetrik (α=0.10)	25	52.15	59.4	56.7	54.9	41.3	8.7	52.2	52.9	56.7	52.9	57.9	59.7	62.7	60.1	60.1	60.1
	50	65.5	73.6	70.3	67.1	50.7	7.9	65.5	70.1	74.4	70.1	73.7	75.9	78.8	76.2	76.2	76.2
	100	78.5	86.1	83.2	79.9	58.9	8.2	78.5	86.3	88.1	86.3	86.3	87.8	90.5	89.0	89.0	89.0
	500	96.4	97.8	97.3	96.4	69.9	8.7	96.4	98.5	98.6	98.5	98.5	97.7	98.6	98.4	98.4	98.4

Across all tests, empirical power increased monotonically with sample size under both symmetric and asymmetric alternatives; however, the rate and magnitude of this increase differed substantially between distribution classes. Under symmetric alternatives, power gains were already substantial when moving from n=25 to n=50, with a further pronounced improvement observed at n=100. Beyond this point, additional increases in sample size resulted in comparatively smaller gains, indicating an early saturation of power for many tests in symmetric settings (Table 1).

In contrast, asymmetric alternatives exhibited a slower and more gradual power accumulation as sample size increased. While the transition from n=25 to n=50 improved detection ability, many tests required at least n=100 observations to achieve power levels comparable to those observed under symmetric departures. Even at n=500, several tests continued to show non-negligible power gains, suggesting that asymmetry introduces a stronger finite-sample burden on normality testing (Table 1).

Importantly, the observed differences between symmetric and asymmetric scenarios cannot be attributed solely to sample size. Rather, they reflect an interaction between distributional structure and the finite-sample behavior of the test statistics. Tests whose critical values stabilize rapidly tended to exhibit earlier power saturation under symmetric alternatives, whereas tests sensitive to skewness and tail asymmetry benefited more strongly from increasing sample size.

Overall, these results demonstrate that sample size requirements for reliable normality testing are intrinsically linked to the type of distributional deviation considered. While moderate sample sizes may be sufficient for detecting symmetric departures, substantially larger samples are often necessary to achieve comparable power under asymmetric alternatives. This finding provides a mechanistic explanation for the heterogeneous power patterns observed across tests and underscores the importance of accounting for distributional structure when assessing sample size adequacy.

Comparative empirical power profiles across sample sizes and significance levels

Figure 4 presents a consolidated comparison of the empirical power of all sixteen normality tests across sample sizes (n=25, 50, 100, 500), significance levels (α=0.05 and α=0.10), and distributional

classes (symmetric vs. asymmetric alternatives). For each test, the reported power represents the average rejection probability across the corresponding group of alternative distributions.

Across all panels, a consistent ordering of tests is observed as sample size increases. Under symmetric alternatives at α=0.05 (Figure 4), tests based on the Shapiro–Wilk family (Shapiro–Wilk, Shapiro–Francia, and Filliben) exhibit relatively strong performance already at moderate sample sizes, with pronounced gains between n=25 and n=100. EDF-based tests such as Anderson–Darling and Cramér–von Mises display a more gradual increase in power, reflecting their stronger reliance on asymptotic critical values.

Increasing the significance level to α=0.10 (Figure 4) leads to a systematic upward shift in power for all tests, while largely preserving their relative ranking. The magnitude of this shift is more pronounced for tests with weaker small-sample performance, indicating that liberal significance levels partially compensate for finite-sample limitations but do not eliminate structural differences among tests.

Under asymmetric alternatives (Figure 4), power accumulation is markedly slower for most tests, particularly at smaller sample sizes. Moment-based tests, including Jarque–Bera and D'Agostino–Pearson–type statistics, show clear improvements as nnn increases but remain comparatively less competitive at n=25. In contrast, tests explicitly sensitive to skewness-related departures, such as Hosking's L-skewness–based test, benefit more strongly from increasing sample size and achieve substantially higher power at $n \geq 100$.

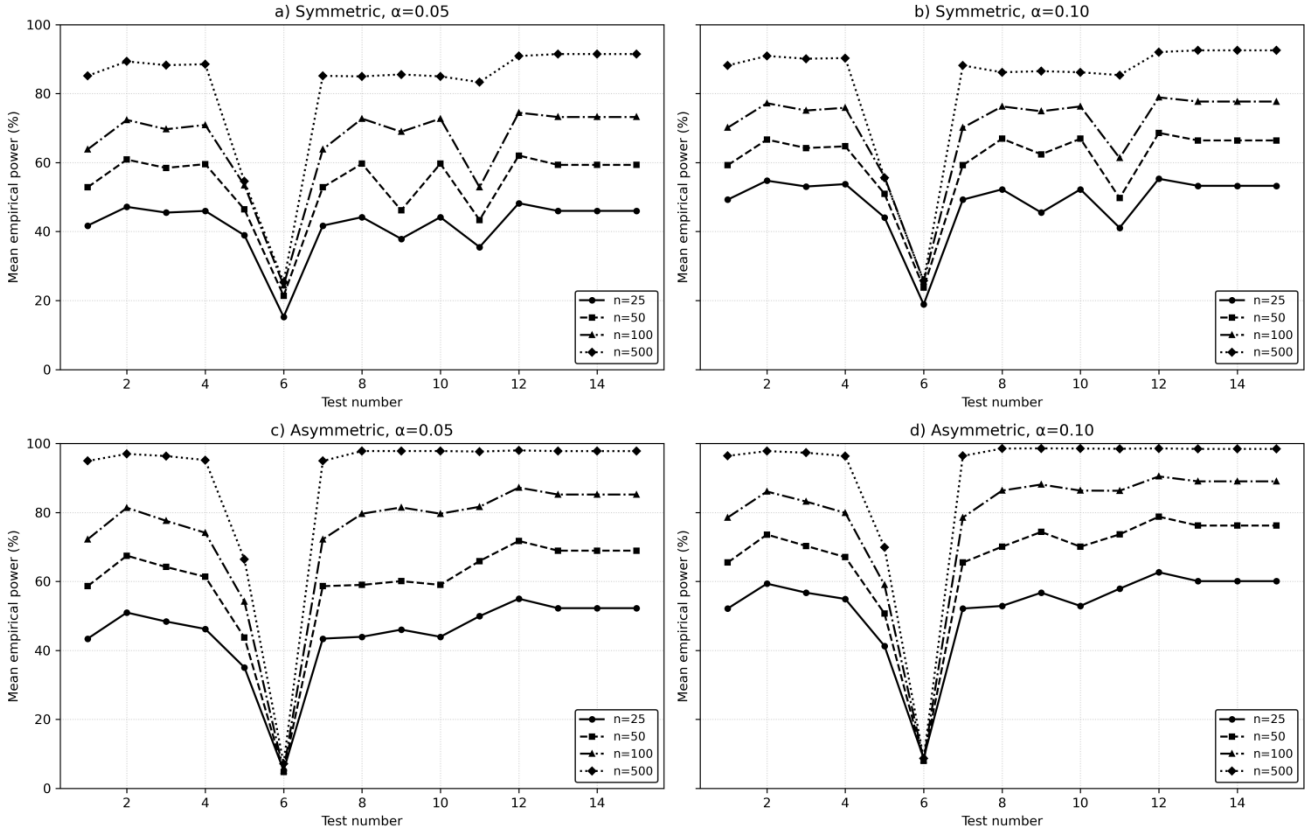

Figure 4. Comparative empirical power profiles of normality tests across sample sizes, significance levels, and distributional structures

Notably, the transition from n=100 to n=500 yields diminishing power gains for many tests under symmetric alternatives, whereas under asymmetric alternatives several tests continue to exhibit non-negligible improvements. This pattern highlights an interaction between distributional structure and

sample size: while moderate samples may suffice for detecting symmetric deviations, asymmetric departures impose stronger finite-sample constraints that delay power saturation.

Overall, Figure 4 demonstrates that no single test uniformly dominates across all scenarios. Instead, empirical power is shaped jointly by sample size, significance level, and the nature of the distributional deviation. These results reinforce the need to consider both the expected form of non-normality and the available sample size when selecting normality tests in practice.

DISCUSSION

This study examined how sample size influences the empirical power of univariate normality tests through its interaction with finite-sample critical value behavior under symmetric and asymmetric departures from normality. Unlike prior comparative studies that primarily focus on ranking tests by power, the present work emphasizes the underlying mechanisms that govern power gains, saturation, and instability as the sample size increases.

A key finding is that discrepancies between empirical and asymptotic critical values play a central role in shaping power performance, particularly at small and moderate sample sizes. For several tests, substantial deviations from asymptotic thresholds persist even at n=50, leading to either conservative or liberal rejection behavior. Such finite-sample distortions have long been recognized in the theoretical literature (Conover, 1999; Darling, 1957; Lehmann & Romano, 2005; Wilcox, 2017), yet they are frequently overlooked in applied practice, where asymptotic critical values are often used by default. Our results demonstrate that these distortions translate directly into reduced or inflated empirical power, independent of the intrinsic sensitivity of the test statistic.

The impact of sample size on power is strongly modulated by the structural form of the alternative distribution. Under symmetric deviations from normality—such as heavy-tailed or kurtotic alternatives—power increases rapidly with sample size and typically stabilizes around n≈100. Beyond this point, additional observations yield only marginal gains. This saturation behavior is consistent with classical asymptotic arguments suggesting that variance reduction dominates once higher-order moments are sufficiently well estimated (Epps & Pulley, 1983; Lilliefors, 1967; Stephens, 1987; Zhang & Wu, 2005). From a practical perspective, this finding implies that for symmetric departures, moderate sample sizes may already provide near-optimal detection performance for many widely used tests.

In contrast, asymmetric alternatives exhibit a markedly different pattern. Power gains under skewed distributions are more gradual, and meaningful improvements often persist beyond n=100. This behavior reflects the increased difficulty of reliably estimating skewness-related features in finite samples, as such features are more sensitive to sampling variability and extreme observations (Cox & Hinkley, 1974; Cramér, 1946; Glen & Leemis, 2004; Serfling, 1980). Consequently, a sample size that is adequate for detecting symmetric deviations may remain insufficient when the underlying departure from normality is asymmetric. These findings reinforce the notion that “sufficient sample size” is inherently context-dependent and cannot be defined independently of the alternative structure.

Another important observation concerns the relative insensitivity of power rankings to the nominal significance level. While increasing α\alphaα from 0.05 to 0.10 uniformly raises rejection probabilities, it does not fundamentally alter the ordering of tests across sample sizes. This suggests that sample size effects dominate significance-level adjustments in determining empirical power, a conclusion that aligns with earlier Monte Carlo investigations of goodness-of-fit procedures (Cramér, 1946; Davison & Hinkley, 1997; Kendall & Stuart, 1977; Lawless, 2003). In practical terms, relaxing the significance level cannot compensate for poor finite-sample calibration or inadequate sample size.

Taken together, the results highlight that the empirical power of normality tests emerges from the combined action of three interrelated components: (i) finite-sample critical value distortion, (ii) sample size–driven variance reduction, and (iii) the structural characteristics of the alternative distribution. Ignoring any of these components risks oversimplified interpretations and potentially suboptimal test selection in applied data analysis. Rather than advocating a single universally optimal test, the findings underscore the importance of aligning test choice and sample size planning with the expected form of distributional deviation.

Limitations

Several limitations of this study should be acknowledged. First, empirical power was summarized by averaging across groups of symmetric and asymmetric alternatives. While this approach facilitates interpretation of sample size effects, it may obscure distribution-specific behaviors that could be relevant in specialized applications. Second, the analysis was restricted to univariate normality tests; extensions to multivariate settings or dependent data structures may yield different convergence and power characteristics. Finally, as with all simulation-based studies, the results depend on the chosen set of alternatives and parameter ranges, although care was taken to represent a broad spectrum of realistic deviations.

Future Perspectives

Future research may build on these findings by conducting distribution-specific power analyses that complement the group-averaged results presented here. In particular, adaptive testing strategies that incorporate sample size–dependent critical value adjustments warrant further investigation. A related line of work, planned as a companion study, will focus on detailed empirical power comparisons across individual distributional scenarios, with an emphasis on practical guidance for test selection under finite-sample constraints.

Taken together, these findings highlight that sample size adequacy in normality testing cannot be assessed independently of finite-sample calibration and distributional structure. The following section summarizes the practical implications of these results and provides concluding recommendations.

CONCLUSION

This study investigated how sample size governs the empirical power of univariate normality tests through its interaction with finite-sample critical value behavior under symmetric and asymmetric departures from normality. Using extensive Monte Carlo simulations, we demonstrated that asymptotic critical values may provide a poor approximation to finite-sample behavior, particularly for small and moderate sample sizes.

Our results show that power gains are not linear in sample size and that the point at which additional observations yield only marginal improvements depends strongly on the structural form of the alternative distribution. While symmetric deviations often exhibit early power stabilization around moderate sample sizes, asymmetric departures require substantially larger samples to achieve comparable detection performance. These findings highlight that sample size adequacy cannot be defined independently of distributional structure.

Importantly, no single normality test uniformly dominates across all scenarios. Instead, empirical power reflects a complex interplay between test construction, critical value calibration, and sample size. From a practical standpoint, the results caution against routine reliance on asymptotic thresholds and emphasize the need for finite-sample awareness when interpreting normality test outcomes.

Overall, this work provides a mechanistic framework for understanding sample size effects in normality testing and offers guidance for more informed test selection and sample size planning in applied statistical analysis.

Conflict of Interest

The authors declare no conflict of interest

Author's Contributions

Conceptualization, M.T.H.; methodology, M.T.H.; software, M.T.H.; validation, M.T.H.; formal analysis, M.T.H.; investigation, M.T.H.; resources, M.T.H.; data curation, M.T.H.; writing—original draft preparation, M.T.H.; writing—review and editing, M.T.H.; visualization, M.T.H.; supervision, M.T.H. All authors have read and agreed to the published version of the manuscript.

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.

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York, NY: Wiley.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London, England: Chapman & Hall.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

D'Agostino, R. B., & Stephens, M. A. (1986). *Goodness-of-fit techniques*. New York, NY: Marcel Dekker.

Darling, D. A. (1957). The Kolmogorov–Smirnov, Cramér–von Mises tests. *The Annals of Mathematical Statistics, 28*(4), 823–838.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, England: Cambridge University Press.

Epps, T. W., & Pulley, L. B. (1983). A test for normality based on the empirical characteristic function. *Biometrika, 70*(3), 723–726.

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics, 17*(1), 111–117.

Glen, A. G., & Leemis, L. M. (2004). Computational and graphical tools for analyzing probability distributions. *Computational Statistics & Data Analysis, 46*(2), 295–312.

Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society: Series B, 52*(1), 105–124.

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

Kendall, M. G., & Stuart, A. (1977). *The advanced theory of statistics* (Vol. 1). London, England: Griffin.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data* (2nd ed.). Hoboken, NJ: Wiley.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York, NY: Springer.

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.

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.

Romão, X., Delgado, R., & Costa, A. (2010). An empirical power comparison of univariate goodness-of-fit tests for normality. *Journal of Statistical Computation and Simulation, 80*(5), 545–591.

Royston, P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics, 31*(2), 115–124.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York, NY: Wiley.

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

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

Stephens, M. A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *The Annals of Statistics, 4*(2), 357–369.

Stephens, M. A. (1987). *Goodness-of-fit techniques*. New York, NY: Marcel Dekker.

Thode, H. C. (2002). *Testing for normality*. New York, NY: Marcel Dekker.

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). San Diego, CA: Academic Press.

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

Zhang, J., & Wu, Y. (2005). Likelihood-ratio tests for normality. *Computational Statistics & Data Analysis, 49*(3), 709–721.