mathematics

MDPI

Article

An Exhaustive Power Comparison of Normality Tests

Jurgita Arnastauskaitė [1,2,*], Tomas Ruzgas [2] and Mindaugas Bražėnas [3]

[1] Department of Applied Mathematics, Kaunas University of Technology, 51368 Kaunas, Lithuania
[2] Department of Computer Sciences, Kaunas University of Technology, 51368 Kaunas, Lithuania; tomas.ruzgas@ktu.lt
[3] Department of Mathematical modelling, Kaunas University of Technology, 51368 Kaunas, Lithuania; mindaugas.brazenas@ktu.lt
* Correspondence: jurgita.arnastauskaite@ktu.lt

Abstract: A goodness-of-fit test is a frequently used modern statistics tool. However, it is still unclear what the most reliable approach is to check assumptions about data set normality. A particular data set (especially with a small number of observations) only partly describes the process, which leaves many options for the interpretation of its true distribution. As a consequence, many goodness-of-fit statistical tests have been developed, the power of which depends on particular circumstances (i.e., sample size, outlets, etc.). With the aim of developing a more universal goodness-of-fit test, we propose an approach based on an N-metric with our chosen kernel function. To compare the power of 40 normality tests, the goodness-of-fit hypothesis was tested for 15 data distributions with 6 different sample sizes. Based on exhaustive comparative research results, we recommend the use of our test for samples of size $n \geq 118$.

Keywords: goodness of fit test; normal distribution; power comparison

Citation: Arnastauskaitė, J.; Ruzgas, T.; Bražėnas, M. An Exhaustive Power Comparison of Normality Tests. *Mathematics* **2021**, *9*, 788. https://doi.org/10.3390/math9070788

Academic Editor: Vasile Preda

Received: 12 February 2021
Accepted: 31 March 2021
Published: 6 April 2021

1. Introduction

A priori information about data distribution is not always known. In those cases, hypothesis testing can help to find a reasonable assumption about the distribution of data. Based on assumed data distribution, one can choose appropriate methods for further research. The information about data distribution can be useful in a number of ways, for example:

- it can provide insights about the observed process;
- parameters of model can be inferred from the characteristics of data distributions; and
- it can help in choosing more specific and computationally efficient methods.

Statistical methods often require data to be normally distributed. If the assumption of normality is not satisfied, the results of these methods will be inappropriate. Therefore, the presumption of normality is strictly required before starting the statistical analysis. Many tests have been developed to check this assumption. However, tests are defined in various ways and thus react to abnormalities, present in a data set, differently. Therefore, the choice of goodness-of-fit test remains an important problem.

For these reasons, this study examines the issue of testing the goodness-of-fit hypotheses. The goodness-of-fit null and alternative hypotheses are defined as:

$$\begin{aligned} H_0 &: \textit{The distribution is normal}, \\ H_A &: \textit{The distribution is not normal}. \end{aligned} \tag{1}$$

A total of 40 tests were applied to analyze the problem of testing the goodness-of-fit hypothesis. The tests used in this study were developed between 1900 and 2016. In the early 19th century, Karl Pearson published an article defining the chi-square test [1]. This test is considered as the basis of modern statistics. Pearson was the first to examine

the goodness-of-fit assumption that the observations x_i can be distributed according to the normal distribution, and concluded that, in the limit as n becomes large, X^2 follows the chi-square distribution with $k-1$ degrees of freedom. The statistics for this test are defined in Section 2.1. Another popular test for testing the goodness-of-fit hypothesis is the Kolmogorov and Smirnov test [2]. This test statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution [3]. The Anderson and Darling test is also often used in practice [4]. This test assesses whether a sample comes from a specified distribution [3]. The end of 19th century and the beginning of 20th century was a successful period for the development of goodness-of-fit hypothesis test criteria and their comparison studies [5–19].

In 2010, Xavier Romão et al. conducted a comprehensive study comparing the power of the goodness-of-fit hypothesis tests [20]. In the study, 33 normality tests were applied to samples of different sizes, taking into account the significance level α and many symmetric, asymmetric, and modified normal distributions. The researchers found that the most powerful of the selected normality tests for the symmetric group of distributions were Coin β_3^2, Chen–Shapiro, Bonett–Seier, and Gel–Miao–Gastwirth tests; for the asymmetric group of distributions, Zhang–Wu Z_C and Z_A, and Chen–Shapiro; while the Chen–Shapiro, Barrio–Cuesta-Albertos–Matrán–Rodríguez-Rodríguez, and Shapiro–Wilk tests were the most powerful for the group of modified normal distributions.

In 2015, Adefisoye et al. compared 18 normality tests for different sample sizes for symmetric and asymmetric distribution groups [3]. The results of the study showed that the Kurtosis test was the most powerful for a group of symmetric data distributions and the Shapiro–Wilk test was the most powerful for a group of asymmetric data distributions.

The main objective of this study is to perform a comparative analysis of the power of the most commonly used tests for testing the goodness-of-fit hypothesis. The procedure described in Section 3 was used to calculate the power of the tests.

Scientific novelty—the comparative analysis of test power was carried out using different methods for goodness-of-fit in the case of many different types of challenges to curve tests. The goodness-of-fit tests have been selected as representatives of popular techniques, which have been analyzed by other researchers experimentally. We have proposed a new kernel function and its usage in an N-metric-based test. The uniqueness of the kernel function is that its shape is chosen in such a way that the shift arising in the formation of the test is eliminated by using sample values.

The rest of the paper is organized as follows. Section 2 provides descriptions of the 40 goodness-of-fit hypothesis tests and the procedure for calculating the power of the tests. The samples generated from 15 distributions are given in Section 4. Section 5 presents and discusses the results of a simulation modeling study. Finally, Section 6 concludes the results.

2. Statistical Methods

In this section, the most popular tests for normality are overviewed.

2.1. Chi-Square Test (CHI2)

In 1900, Karl Pearson introduced the chi-square test [1]. The statistic of the test is defined as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{1}$$

where O_i is the observed frequency and E_i is the expected frequency.

2.2. Kolmogorov–Smirnov (KS)

In 1933, Kolmogorov and Smirnov proposed the KS test [2]. The statistic of the test is defined as:

$$\begin{gathered} \chi^2 = D^+ = \max\left\{\left(\frac{i}{n}\right) - z_i\right\}, \quad 1 \le i \le n; \\ D^- = \max\left\{z_i - \frac{i-1}{n}\right\}, \quad 1 \le i \le n; \\ D = \max(D^+, D^-), \end{gathered} \tag{2}$$

where z_i is the cumulative probability of standard normal distribution and D is the difference between observed and expected values.

2.3. Anderson–Darling (AD)

In 1952, Anderson and Darling developed a variety of the Kolmogorov and Smirnov tests [4]. This test is more powerful than the Kolmogorov and Smirnov test. The statistic of the test is defined as:

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} (\ln(F(x_i)) + \ln(1 - F(x_{n+1-i}))), \tag{3}$$

where $F(x_i)$ is the value of the distribution function at point x_i and n is the empirical sample size.

2.4. Cramer–Von Mises (CVM)

In 1962, Cramer proposed the Cramer–von Mises test. This test is an alternative to the Kolmogorov and Smirnov test [21]. The statistic of the test is defined as:

$$CM = \frac{1}{12n} + \sum_{i=1}^{n} \left(Z_i - \frac{2i-1}{2n}\right)^2, \tag{4}$$

where Z_i is the cumulative distribution function of the specified distribution $Z_i = X_{(i)} - \overline{X}/S$, $\overline{\text{and } X}$ and S are the sample mean and sample standard deviation.

2.5. Shapiro–Wilk (SW)

In 1965, Shapiro and Wilk formed the original test [22]. The statistic of the test is defined as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\left(\sum_{i=1}^{n} x_i - \overline{x}\right)^2}, \tag{5}$$

where $x_{(i)}$ is the i^{th} order statistic, $\overline{x}$ is the sample mean, and a_i constants obtained:

$$a_i = (a_1, \ldots, a_n) = \frac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{1/2}},$$

where $m = (m_1, \ldots, m_n)^T$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and V is the covariance matrix of those order statistics.

2.6. Lilliefors (LF)

In 1967, Lilliefors modified the Kolmogorov and Smirnov test [23]. The statistic of the test is defined as:

$$T = \sup_x |F^*(x) - S(x)|, \tag{6}$$

where $F^*(x)$ is the standard normal distribution function and $S(x)$ is the empirical distribution function of the z_i values.

2.7. D'Agostino (DA)

In 1971, D'Agostino introduced the test for testing the goodness-of-fit hypothesis, which is an extension of the Shapiro–Wilk test [8]. The test proposed by D'Agostino does not need to define a weight vector. The statistic of the test is defined as:

$$D = \frac{\sum_{i=1}^{n}(i-(n+1)/2)\cdot x_{(i)}}{n^2\cdot\sqrt{m_2}}, \tag{7}$$

where m_2 is the second central moment that is defined as:

$$m_2 = n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

2.8. Shapiro–Francia (SF)

In 1972, Shapiro and Francia simplified the Shapiro and Wilk test and developed the Shapiro and Francia test, which is computationally more efficient [24]. The statistic of the test is defined as:

$$W_{SF} = \frac{\left(\sum_{i=1}^{n} m_i x_i\right)^2}{\left(\sum_{i=1}^{n} x_i-\overline{x}\right)^2\sum_{i=1}^{n} m_i^2}, \tag{8}$$

where m_i is the expected values of the standard normal order statistics.

2.9. D'Agostino–Pearson (DAP)

In 1973–1974, D'Agostino and Pearson proposed the D'Agostino and Pearson test [25]. The statistic of the test is defined as:

$$DP = \frac{\sum_{i=1}^{n}(i-(n+1)/2)x_{(i)}}{n^2\sqrt{m_2}}, \tag{9}$$

where n is the size of sample and m_2 is the sample variance of order statistics.

2.10. Filliben (Filli)

In 1975, Filliben defined the probabilistic correlation coefficient r as a test for the goodness-of-fit hypothesis [26]. This test statistic is defined as:

$$r = \frac{\sum_{i=1}^{n} x_{(i)}\cdot M_{(i)}}{\sqrt{\sum_{i=1}^{n} M_{(i)}^2}\cdot\sqrt{(n-1)\cdot\sigma^2}}, \tag{10}$$

where σ^2 is the variance, $M_{(i)} = \Phi^{-1}\left(m_{(i)}\right)$, when $m_{(i)}$ is the estimated median values of the order statistics, each $m_{(i)}$ is obtained by:

$$m_{(i)} = \begin{cases} 1-0.5^{\left(\frac{1}{n}\right)} & i=1, \\ \frac{(i-0.3175)}{(n+0.365)} & 1<i<n, \\ 0.5^{\left(\frac{1}{n}\right)} & i=n. \end{cases}$$

2.11. Martinez–Iglewicz (MI)

In 1981, Martinez and Iglewicz proposed a normality test based on the ratio of two estimators of variance, where one of the estimators is the robust biweight scale estimator S_b^2 [27]:

$$S_b^2 = \frac{n\cdot\sum_{|\widetilde{z}_i|<1}(x_i-M)^2\left(1-\widetilde{z}_i^2\right)^4}{\left[\sum_{|\widetilde{z}_i|<1}\left(1-\widetilde{z}_i^2\right)\left(1-5\widetilde{z}_i^2\right)\right]^2},$$

where M is the sample median, $\widetilde{z}_i = (x_i-M)/(9A)$, with A being the median of $|x_i-M|$.

This test statistic is then given by:

$$I_n = \frac{\sum_{i=1}^{n}(x_i - M)^2}{(n-1)\cdot S_b^2}. \tag{11}$$

2.12. Epps–Pulley (EP)

In 1983, Epps and Pulley proposed a test statistic based on the following weighted integral [28]:

$$T_{EP} = \int_{-\infty}^{\infty} |\varphi_n(t) - \hat{\varphi}_0(t)|^2 dG(t),$$

where $\varphi_n(t)$ is the empirical characteristic function and $G(t)$ is an adequate function chosen according to several considerations. By setting $dG(t) = g(t)dt$ and selecting:

$$g(t) = \sqrt{m_2/2\pi}\cdot exp\left(-0.5m_2t^2\right)$$

the following statistic is obtained:

$$T_{EP} = 1 + \frac{n}{\sqrt{3}} + \frac{2}{n}\sum_{k=2}^{n}\sum_{j=1}^{k-1}\exp\left(\left(-\left(x_j - x_k\right)^2\right)/2m_2\right) - \sqrt{2}\sum_{j=1}^{n}\exp\left(\left(-\left(x_j - \bar{x}\right)^2\right)/4m_2\right), \tag{12}$$

where m_2 is the second central moment.

2.13. Jarque–Bera (JB)

In 1987, Jarque and Bera proposed a test [29] with statistic defined as:

$$JB = \frac{n}{6}\left(s + \frac{(k-3)^2}{4}\right), \tag{13}$$

where $s = \frac{m_3^2}{m_2^3}$ and $k = \frac{m_4}{m_2^3}$ are the sample skewness and kurtosis.

2.14. Hosking ($H_1 - H_3$)

In 1990, Hosking and Wallis proposed the first Hosking test [5]. This test statistic is defined as:

$$H_i = \frac{V_i - \mu_V}{\sigma_V}, \tag{14}$$

where μ_V and σ_V are the mean and standard deviation of number of simulation data values of V. V_i is calculated as:

$$V_1 = \sqrt{\frac{\sum_{i=1}^{N} n_i\left(t^{(i)} - t^R\right)^2}{\sum_{i=1}^{N} n_i}},\ V_2 = \sum_{i=1}^{N} n_i \frac{\sqrt{\left(t^{(i)} - t^R\right)^2 + \left(t_3 - t_3^R\right)^2}}{\sum_{i=1}^{N} n_i},$$

$$V_3 = \sum_{i=1}^{N} n_i \frac{\sqrt{\left(t_3^{(i)} - t_3^R\right)^2 + \left(t_4^{(i)} - t_4^R\right)^2}}{\sum_{i=1}^{N} n_i},\ t^R = \frac{\sum_{i=1}^{N} n_i t^{(i)}}{\sum_{i=1}^{N} n_i},$$

where $t^{(i)}$ is the coefficient of variation of the L-moment ratio, $t_3^{(i)}$ is the coefficient of skewness of the L- moment, and $t_4^{(i)}$ is the coefficient of kurtosis of the L- moment.

2.15. Cabaña–Cabaña (CC1-CC2)

In 1994, Cabaña and Cabaña proposed the CC1 and CC2 tests [6]. The CC1 ($T_{S,l}$) and CC2 ($T_{K,l}$), respectively, are defined as:

$$T_{S,l} = \max\left|\, w_{S,l}(x)\right|, \quad T_{K,l} = \max\left|\, w_{K,l}(x)\right|, \tag{15}$$

where $w_{S,l}(x)$ and $w_{K,l}(x)$ approximate transformed estimated empirical processes sensitive to changes in skewness and kurtosis and are defined as:

$$w_{S,l} = \Phi(x)\cdot\overline{H}_3 - \phi(x)\cdot\sum_{j=1}^{l}\frac{1}{\sqrt{j}}H_{j-1}(x)\cdot\overline{H}_{j+3},$$
$$w_{K,l} = -\phi(x)\cdot\overline{H}_3 + [\Phi(x) - x\cdot\phi(x)]\cdot\overline{H}_4 - \phi(x)\cdot\sum_{j=2}^{l}\left(\sqrt{\frac{j}{j-1}}H_{j-2}(x)\cdot H_j(x)\right)\cdot\overline{H}_{j+3},$$

where l is a dimensionality parameter, $\Phi(x)$ is the probability density function of the standard normal distribution, $H_j(\cdot)$ is the jth order normalized Hermite polynomial, and $\overline{H}_j$ is the jth order normalized mean of the Hermite polynomial defined as: $\overline{H}_j = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}H_j(x_i)$.

2.16. The Chen–Shapiro Test (ChenS)

In 1995, Chen and Shapiro introduced an alternative test statistic based on normalized spacings and defined as [9]:

$$CS = \frac{1}{(n-1)\cdot s}\sum_{i=1}^{n-1}\frac{x_{(i+1)} - x_{(i)}}{M_{i+1} - M_i}, \tag{16}$$

where M_i is the ith quantile of a standard normal distribution.

2.17. Modified Shapiro-Wilk (SWRG)

In 1997, Rahman and Govindarajulu proposed a modification to the Shapiro–Wilk test [8]. This test statistic is simpler to compute and relies on a new definition of the weights using the approximations to m and V. Each element a_i of the weight vector is given as:

$$a_i = -(n+1)(n+2)\phi(m_i)[m_{i-1}\phi(m_{i-1}) - 2m_i\phi(m_i) + m_{i+1}\phi(m_{i+1})], \tag{17}$$

where it is assumed that $m_0\phi(m_0) = m_{n+1}\phi(m_{n+1}) = 0$. Therefore, the modified test statistic assigns larger weights to the extreme order statistics than the original test.

2.18. Doornik–Hansen (DH)

In 1977, Bowman and Shenton introduced the Doornik–Hansen goodness-of-fit test [9]. This test statistic is obtained using transformations of skewness and kurtosis:

$$s = \frac{m_3}{\sqrt{m_2}^3} \quad and \quad k = \frac{m_4}{m_2^2} \tag{18}$$

where $m_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^i\overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$ and n is sample size.

The DH test statistics have a chi-square distribution with two degrees of freedom. It is defined as:

$$DH = z_1^2 + z_2^2 \sim \chi^2(2),$$

where $z_1 = \delta\log\left(y + \sqrt{y^2-1}\right)$, $\delta = \frac{1}{\sqrt{log(w^2)}}$, $w^2 = -1 + \sqrt{2(\beta-1)}$,

$\beta = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$, $y = s\sqrt{\frac{(w^2-1)(n+1)(n+3)}{12(n-2)}}$,
$s = z_2 = \sqrt{2\alpha}\left(\frac{1}{9\alpha} - 1 + \sqrt[3]{\frac{\chi}{2\alpha}}\right)$, $\alpha = a + c\times s^2$, $a = \frac{(n-2)(n+5)(n+7)(n^2+27n-70)}{6\delta}$,
$c = \frac{(n-7)(n+5)(n+7)(n^2+2n-5)}{6\delta}$, $\delta = (n-3)(n+1)(n^2+15n-4)$, $\chi = 2l(k-1-s^2)$,
$l = \frac{(n+5)(n+7)(n^3+37n^2+11n-313)}{12\delta}$.

2.19. Zhang Q (ZQ), Q(ZQstar), Q − Q* (ZQQstar)*

In 1999, Zhang introduced the *Qtest* statistic based on the ratio of two unbiased estimators of standard deviation, q_1 and q_2, given by $Q = ln(q_1/q_2)$ [10]. The estimators q_1 and q_2 are calculated by $q_1 = \sum_{i=1}^{n} a_i x_{(i)}$ and $q_2 = \sum_{i=1}^{n} b_i x_{(i)}$, where the ith order linear coefficients a_i and b_i are:

$$b_i = \begin{cases} a_i = [(u_i - u_1)(n-1)]^{-1}, \text{given } i \neq 1,\ a_1 = \sum_{i=2}^{n} a_i, \\ -b_{n-i+1} = [(u_i - u_{i+4})(n-4)]^{-1} \qquad i = 1, \ldots, 4, \\ (n-4)^{-1} \cdot \left[(u_i - u_{i+4})^{-1} - (u_{i-4} - u_i)^{-1} \right] \ i = 5, \ldots, n-4, \end{cases} \tag{19}$$

where u_i is the ith expected value of the order statistics of a standard normal distribution, $u_i = \Phi^{-1}[(i - 0.375)/(n + 0.25)]$.

Zhang also proposed the alternative statistic Q^* by switching the ith order statistics $x_{(i)}$ in q_1 and q_2 by $x^*_{(i)} = -x_{(n-i+1)}$.

In addition to those already discussed, Zhang proposed joint test $Q - Q^*$, based on the fact that Q and Q^* are approximately independent.

2.20. Barrio–Cuesta-Albertos–Matran–Rodriguez-Rodriguez (BCMR)

In 1999, Barrio, Cuesta-Albertos, Matrán, and Rodríguez-Rodríguez proposed a new BCMR goodness-of-fit test [11]. This test is based on L_2-Wasserstein distance and is defined as:

$$BCMR = \frac{m_2 - \left[\sum_{i=1}^{n} x_{(i)} \cdot \int_{(i-1)/n}^{i/n} \Phi^{-1}(t) dt \right]^2}{m_2}, \tag{20}$$

where the numerator represents the squared L_2-Wasserstein distance.

2.21. Glen–Leemis–Barr (GLB)

In 2001, Glen, Leemis, and Barr extended the Kolmogorov–Smirnov and Anderson–Darling test to form the GLB test [12]. This test statistic is defined as:

$$P_S = -n - \frac{1}{n} \cdot \sum_{i=1}^{n} [(2n + 1 - 2i) \cdot \ln\left(p_{(i)}\right) + (2i - 1) \cdot \ln\left(1 - p_{(i)}\right)], \tag{21}$$

where $p_{(i)}$ is the elements of the vector p containing the quantiles of the order statistics sorted in ascending order.

2.22. Bonett–Seier T_w (BS)

In 2002, Bonett and Seier introduced the BS test [13]. The statistic for this test is defined as:

$$T_w = \frac{\sqrt{n+2} \cdot (\hat{\omega} - 3)}{3.54}, \tag{22}$$

where $\hat{\omega} = 13.29 \left[ln\sqrt{m_2} - \ln\left(n^{-1} \sum_{i=1}^{n} |x_i - \overline{x}| \right) \right]$, $m_2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$.

2.23. Bontemps–Meddahi (BM1–BM_{3-4}, BM2–BM_{3-6})

In 2005, Bontemps and Meddahi proposed a family of normality tests based on moment conditions known as Stein equations and their relation with Hermite polynomials [24]. The statistic of the test is defined as:

$$BM_{3-p} = \sum_{k=3}^{p} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} H_k(z_i) \right)^2, \tag{23}$$

where $z_i = (x_i - \overline{x})/s$ and $H_k(\cdot)$ is the kth order normalized Hermite polynomial having the general expression given by:

$$\forall i > 1: \; H_i(u) = \frac{1}{\sqrt{i}}\Big[u \cdot H_{i-1}(u) - \sqrt{i-1} \cdot H_{i-2}(u)\Big], \quad H_0(u) = 1, \quad H_1(u) = u.$$

2.24. Zhang–Wu (ZW1–Z_C, ZW2–Z_A)

In 2005, Zhang and Wu presented the ZW1 and ZW2 goodness-of-fit tests [15]. The Z_C and Z_A statistics are similar to the Cramér–von Mises and Anderson–Darling tests statistics based on the empirical distribution function. The statistic of the test is defined as:

$$\begin{aligned} Z_C &= \sum_{i=1}^{n}\left[ln\frac{\Phi\left(z_{(i)}\right)^{-1}-1}{\frac{n-0.5}{i-0.75}-1}\right]^2, \\ Z_A &= -\sum_{i=1}^{n}\left[\frac{ln\,\Phi\left(z_{(i)}\right)}{n-i+0.5} + \frac{\ln\left[1-\Phi\left(z_{(i)}\right)\right]}{i-0.5}\right], \end{aligned} \tag{24}$$

where $\Phi\left(z_{(i)}\right) = (i - 0.5)/n$.

2.25. Gel–Miao–Gastwirth (GMG)

In 2007, Gel, Miao, and Gastwirth proposed the GMG test [16]. The statistic of the test is defined as:

$$R_{sJ} = \frac{s}{J_n}, \tag{25}$$

where J_n is the ratio of the standard deviation and the robust measure of dispersion is defined as:

$$J_n = \frac{\sqrt{\pi/2}}{n}\sum_{i=1}^{n}|x_i - M|,$$

where M is the median of the sample.

2.26. Robust Jarque–Bera (RJB)

In 2007, Gel and Gastwirth modified the Jarque–Bera test and got a more powerful Jarque–Bera test [16]. RJB test statistic is defined as:

$$RJB = \frac{n}{6}\left(\frac{m_3}{J_n^3}\right)^2 + \frac{n}{64}\left(\frac{m_4}{J_n^4} - 3\right)^2, \tag{26}$$

where m_3, m_4 are the third and fourth moments, respectively, and J_n is the ratio of the standard deviation.

2.27. Coin β_3^2

In 2008, Coin proposed a test based on polynomial regression to determine the group distributions of symmetric distributions [17]. The type of model for this test is:

$$z_{(i)} = \beta_1 \cdot \alpha_i + \beta_3 \cdot \alpha_i^3, \tag{27}$$

where β_1 and β_3 are fitting parameters and α_i is the expected values of standard normal order statistics.

2.28. Brys–Hubert–Struyf T_{MC-LR} (BHS)

In 2008, Brys, Hubert, and Struyf introduced the BHS tests [3]. This test is based on skewness and long tails. The statistics for this test T_{MC-LR} is defined as:

$$T_{MC-LR} = n(w - \omega)^T V^{-1}(w - \omega), \tag{28}$$

where w is set as $[MC,\ LMC,\ RMC]^T$, MC is medcouple, LMC is left medcouple, RMC is right medcouple, and ω and V are obtained based on the influence function of the estimators in ω. In the case of a normal distribution:

$$\omega = [0,\ 0.199,\ 0.199]^T, \quad V = \begin{bmatrix} 1.25 & 0.323 & -0.323 \\ 0.323 & 2.62 & -0.0123 \\ -0.323 & -0.0123 & 2.62 \end{bmatrix}.$$

2.29. Brys–Hubert–Struyf–Bonett–Seier T_{MC-LR} & T_w (BHSBS)

In 2008, Brys, Hubert, Struyf, Bonett, and Seier introduced the combined BHSBS test [3]. This test statistic is defined as:

$$T_{MC-LR}\ \&\ T_w = n(w-\omega)^T V^{-1}(w-\omega) \& \frac{\sqrt{n+2}\cdot(\hat{\omega}-3)}{3.54}, \tag{29}$$

where ω is asymptotic mean and V is covariance matrix.

2.30. Desgagné–Lafaye de Micheaux–Leblanc R_n (DLDMLRn), X^a_{APD} (DLDMXAPD), Z^a_{EPD} (DLDMZEPD)

In 2009, Desgagné, Lafaye de Micheaux, and Leblanc introduced the R_n and X^a_{APD} tests [18]. The statistic $R_n(\mu,\sigma)$ for this test is defined as:

$$R_n(\mu,\sigma) = \frac{1}{n}\sum_{i=1}^{n} d_\theta(Y_i) = \begin{pmatrix} -2\left[\frac{1}{n}\sum\limits_{i=1}^{n} Y_i^2\ sign(Y_i)\right] \\ -2^{-1}\left[\frac{1}{n}\sum\limits_{i=1}^{n} Y_i^2 log|Y_i| - (2-log2-\gamma)/2\right] \end{pmatrix}, \tag{30}$$

where $Y_i = \sigma^{-1}(X_i-\mu)$. When μ and σ are unknown, the following maximum-likelihood estimators can be used:

$$\hat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\sigma}_n = S_n = \left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2\right]^{1/2}.$$

The DLDMXAPD test is based on skewness and kurtosis which are defined as:

$$s = \frac{1}{n}\sum_{i=1}^{n} Z_i^2\ sign(Z_i), \qquad k = \frac{1}{n}\sum_{i=1}^{n} Z_i^2\ log|(Z_i)|, \tag{31}$$

where $Z_i = S_n^{-1}(X_i-\overline{X}_n)$, $\overline{X}_n$, S_n are defined above.

The DLDMXAPD test is suitable for use when the sample size is greater than 10. The statistic X^a_{APD} for this test is defined as:

$$X^a_{APD} = \frac{ns^2}{3-8/\pi} + \frac{n(k-(2-log2-\gamma)/2)^2}{(3\pi^2-28)/8}, \quad X_{APD} = Z^2(s) + Z^2\left(k-s^2\right), \tag{32}$$

where $\gamma = 0.577215665$ is the Euler–Mascheroni constant and s, k are skewness and kurtosis, respectively.

In 2016, Desgagné, Lafaye de Micheaux, and Leblanc presented the DLDMZEPD test based on the skewness [18]. The statistic Z^a_{EPD} for this test is defined as:

$$Z^a_{EPD} = \frac{n^{1/2}(k-(2-log2-\gamma)/2)}{[(3\pi^2-28)/8]^{1/2}}, \quad Z_{EPD} = Z_{EPD}(k). \tag{33}$$

2.31. N-Metric

We improved the Bakshaev [30] goodness-of-fit hypothesis test based on N-metrics. This test is defined in the following way.

Under the null hypothesis statistic, $T_n = -n \int_0^1 \int_0^1 K(x) d(F_n^*(x) - x)$ has the same asymptotic distribution as quadratic form:

$$T_n = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \frac{a_{kj}}{\pi^2 kj} \xi_k \xi_j, \tag{34}$$

where ξ_k are independent random variables from the standard normal distribution and:

$$a_{kj} = -2 \int_0^1 \int_0^1 K(x) d sin(\pi k x).$$

In this case, Bakshaev applied the kernel function $K(x) = |x - y|$, and we propose to apply another kernel function (Figure 1):

$$K(x) = \varphi(\overline{g}(x)) \overline{g}'(x), \tag{35}$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

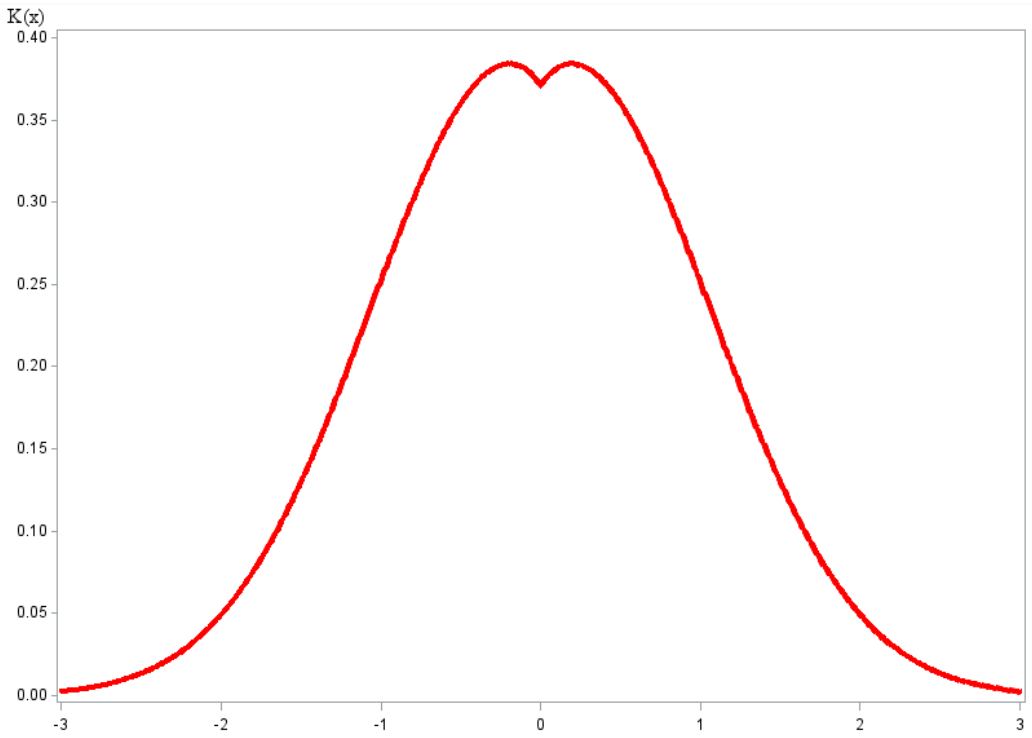

Figure 1. Plot of out kernel function $K(x)$ with experimentally chosen optimal parameters $a = 0.95$, $b = 0.25$, and $c = 1$.

An additional bias is introduced when the kernel function is calculated at the sample values (i.e., for $x = X(t)$). Therefore, to eliminate this bias, the shape of the kernel function is chosen so that the influence in the environment of the sample values is as small as possible.

Let X be the standard normal random variable, Φ and φ be its distribution and density functions, respectively, and $g : R \to R$ is an odd strictly monotonically increasing function. Then the distribution function F_Y of the random variable $Y = g(X)$ is $\Phi(\overline{g}(x))$, where $\overline{g}$ is the inverse of the function g. The distribution density f_Y of a random variable

Y is $\varphi(\overline{g}(x))\overline{g}'(x)$. Let us consider the parametric class of functions $\overline{g}$, which depends on three parameters:

$$\overline{g} = x\left(c + |x|^b\right)^a, \quad a, b, c > 0,$$
$$\overline{g}' = \left(c + |x|^b\right)^a + a|x|\left(c + |x|^b\right)^{a-1} b|x|^{b-1}$$

where a is variance, b is trough, and c is peak shape parameter.

3. The Power of Test

The power of the test is defined as the probability of rejecting a false H_0 hypothesis. Power is the opposite of type II error. Decreasing the probability of type I error α increases the probability of type II error and decreases the power of the test. The smaller the error is, the more powerful test is. In practice, the tests are designed to minimize the type II error for a fixed type I error. The most commonly chosen value for α is 0.05. The probability of the opposite event is calculated as $1 - \beta$, i.e., the power of the test (see in Figure 2) β is the probability of rejecting hypothesis H_0 when it is false. The power of the test makes it possible to compare two tests significance level and sample sizes. A more powerful test has a higher value of $1 - \beta$. Increasing the sample size usually increases the power of the test [31,32].

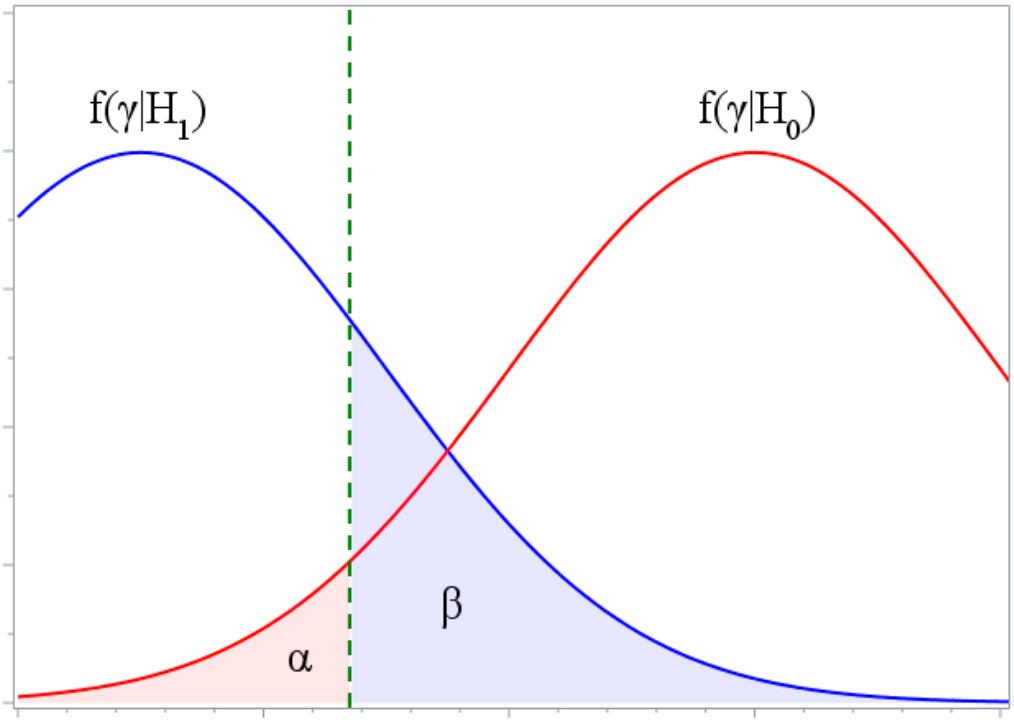

Figure 2. Illustration of the power.

When exact null distribution of a goodness-of-fit test statistic is a step function created by the summation of the exact probabilities for each possible value of the test statistic, it is possible to obtain the same critical value for a number of different adjacent significance levels α. Linear interpolation of the power of the test statistic using the power for a significance levels (see in Figure 3) less than (denoted α_1) and greater than (denoted α_2) the desired significance level (denoted as α) is preferred by many authors to overcome this problem (see, for example, [33]). Linear interpolation gives a weighting to the power based

on how close α_1 and α_2 are to α. In this case, the power of the test is calculated according to the formula [19]:

$$Power = \frac{(\alpha - \alpha_1)P(T \geq \gamma_2(\alpha)|H_1) + (\alpha_2 - \alpha)P(T \geq \gamma_1(\alpha)|H_1)}{\alpha_2 - \alpha_1}, \quad (36)$$

where $\gamma_1(\alpha)$ and $\gamma_2(\alpha)$ are the critical values immediately below and above the significance level α. α_1=$P(T \geq \gamma_1(\alpha)|H_0)$ and α_2=$P(T \geq \gamma_2(\alpha)|H_0)$ are the significance levels for $\gamma_1(\alpha)$ and $\gamma_2(\alpha)$, respectively.

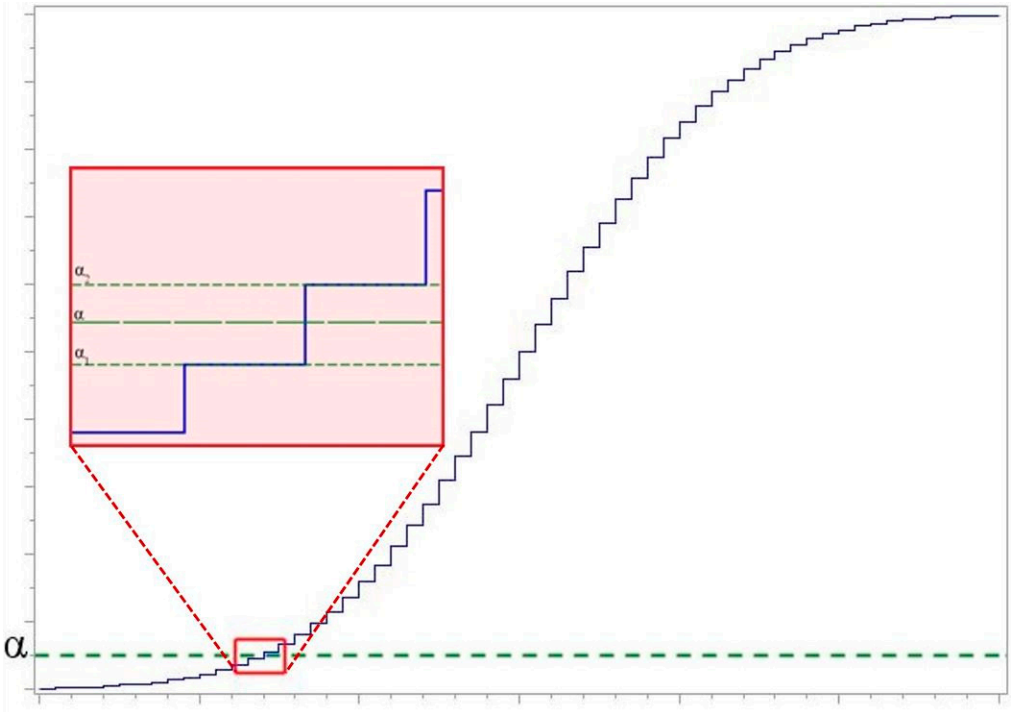

Figure 3. Significance levels of the statistic step function.

The power of test statistics is determinate by the following steps [19]:

1. The distribution of the analyzed data $x_1, x_2, \ldots, x_n$ is formed.
2. Statistics of the compatibility hypothesis test criteria are calculated. If the obtain value of statistic is greater than the corresponding critical value ($\alpha = 0.05$ is used), then hypothesis H_0 is rejected.
3. Steps 1 and 2 are repeated for k (in our experiments, $k = 1,000,000$) times.
4. The power of a test is calculated as $count/k$, where $count$ is the number of false hypotheses rejections.

4. Statistical Distributions

The simulation study considers fifteen statistical distributions for which the performance of the presented normality tests are assessed. Statistical distributions are grouped into three groups: symmetric, asymmetric, and modified normal distributions. A description of these distribution groups is presented in the following.

4.1. Symmetric Distributions

Symmetric distributions considered in this research are [20]:

- three cases of the $Beta(a, b)$ distribution$Beta(0.5; 0.5)$, $Beta(1; 1)$, and $Beta(2; 2)$, where a and b are the shape parameters;
- three cases of the $Cauchy(t, s)$ distribution—$Cauchy(0; 0.5)$, $Cauchy(0; 1)$, and $Cauchy(0; 2)$, where t and s are the location and scale parameters;

- one case of the $Laplace(t,\ s)$ distribution$Laplace(0;1)$, where t and s are the location and scale parameters;
- one case of the $Logistic(t,\ s)$ distribution$Logistic(2;2)$, where t and s are the location and scale parameters;
- four cases of the $t-Student(\nu)$ distribution$t(1),\ t(2),\ t(4)$, and $t(10)$, where ν is the number of degrees of freedom;
- five cases of the $Tukey(\lambda)$ distribution$Tukey(0.14),\ Tukey(0.5),\ Tukey(2),\ Tukey(5)$, and $Tukey(10)$, where λ is the shape parameter; and
- one case of the standard normal $N(0;1)$ distribution.

4.2. Asymmetric Distributions

Asymmetric distributions considered in this research are [20]:

- four cases of the $Beta(a,b)$ distribution$Beta(2;1),\ Beta(2;5),\ Beta(4;0.5)$, and $Beta(5;1)$;
- four cases of the $Chi\text{-}squared(\nu)$ distribution$\chi^2(1),\ \chi^2(2),\ \chi^2(4)$, and $\chi^2(10)$, where ν is the number of degrees of freedom;
- six cases of the $Gamma(a,\ b)$ distribution—$Gamma(2;2),\ Gamma(3;2),\ Gamma(5;1),\ Gamma(9;1),\ Gamma(15;1)$, and $Gamma(100;1)$, where a and b are the shape and scale parameters;
- one case of the $Gumbel(t,\ s)$ distribution$Gumbel(1;2)$, where t and s are the location and scale parameters;
- one case of the $Lognormal(t,\ s)$ distribution$LN(0;1)$, where t and s are the location and scale parameters; and
- four cases of the $Weibull(a,\ b)$ distribution$Weibull(0.5;1),\ Weibull(1;2),\ Weibull(2;3.4)$, and $Weibull(3;4)$, where a and b are the shape and scale parameters.

4.3. Modified Normal Distributions

Modified normal distributions considered in this research are [20]:

- six cases of the standard normal distribution truncated at a and b $Trunc(a;\ b)Trunc(-1;1),\ Trunc(-2;2),\ Trunc(-3;3),\ Trunc(-2;1),\ Trunc(-3;1)$, and $Trunc(-3;2)$, which are referred to as NORMAL1;
- nine cases of a location-contaminated standard normal distribution, hereon termed $LoConN(p;\ a)LoConN(0.3;1),\ LoConN(0.4;1),\ LoConN(0.5;1),\ LoConN(0.3;3),\ LoConN(0.4;3),\ LoConN(0.5;3),\ LoConN(0.3;5),\ LoConN(0.4;5)$, and $LoConN(0.5;5)$, which are referred to as NORMAL2;
- nine cases of a scale-contaminated standard normal distribution, hereon termed $ScConN(p;\ b)ScConN(0.05;0.25),\ ScConN(0.10;0.25),\ ScConN(0.20;0.25),\ ScConN(0.05;2),\ ScConN(0.10;2),\ ScConN(0.20;2),\ ScConN(0.05;4),\ ScConN(0.10;4)$, and $ScConN(0.20;4)$, which are referred to as NORMAL3; and
- twelve cases of a mixture of normal distributions, hereon termed $MixN(p;\ a;\ b)$ $MixN(0.3;1;0.25),\ MixN(0.4;1;0.25),\ MixN(0.5;1;0.25),\ MixN(0.3;3;0.25),\ MixN(0.4;3;0.25),\ MixN(0.5;3;0.25),\ MixN(0.3;1;4),\ MixN(0.4;1;4),\ MixN(0.5;1;4),\ MixN(0.3;3;4),\ MixN(0.4;3;4)$, and $MixN(0.5;3;4)$, which are referred to as NORMAL4.

5. Simulation Study and Discussion

This section provides a comprehensive modeling study that is designed to evaluate the power of selected normality tests. This modeling study takes into account the effects of sample size, the level of significance ($\alpha = 0.05$) chosen, and the alternative type of distribution (Beta, Cauchy, Laplace, Logistic, Student, Chi-Square, Gamma, Gumbel, Log-normal, Weibull, and modified standard normal). The study was performed by applying 40 normality tests (including our proposed normality test) for the generated 1,000,000 standardized samples of size 32, 64, 128, 256, 512, and 1024.

The best set of parameters $(a,\ b,\ c)$ was selected experimentally: the value of a was examined from 0.001 to 0.99 by step 0.01, the value of b was examined from 0.01 to 10 by step 0.01, and the value of c was examined from 0.5 to 50 by step 0.25. The *N-metric* test gave the most powerful results with the parameters: $a = 0.95,\ b = 0.25,\ c = 1$. In those cases, a test has several modifications, we present results only for the best variant. The Tables 1–3 present average power obtained for the symmetric, asymmetric, and modified normal distribution sets, for samples sizes of 32, 64, 128, 256, 512, and 1024. By comparing Tables 1–3, it can be seen that the most powerful test for small samples was *Hosking1* (*H1*), the most powerful test for large sample sizes was our presented test (*N-metric*). According to Tables 1–3, it is observed that for large sample sizes, most tests' power is approaching 1 except for the D'Agostino (DA) test, the power of which is significantly lower.

Table 1. Average empirical power obtained for a group of symmetric distributions.

		Sample Size					
		32	**64**	**128**	**256**	**512**	**1024**
Tests	**AD**	0.714	0.799	0.863	0.909	0.939	0.955
	BCMR	0.718	0.809	0.875	0.920	0.947	0.947
	BHS	0.431	0.551	0.663	0.752	0.818	0.868
	BHSBS	0.680	0.778	0.783	0.903	0.938	0.959
	BM2	0.726	0.835	0.905	0.945	0.965	0.974
	BS	0.717	0.810	0.877	0.920	0.947	0.961
	CC2	0.712	0.805	0.873	0.920	0.949	0.936
	CHI2	0.663	0.778	0.842	0.884	0.941	0.945
	CVM	0.591	0.733	0.805	0.855	0.919	0.949
	ChenS	0.729	0.806	0.871	0.915	0.943	0.960
	Coin	0.735	0.830	0.891	0.930	0.952	0.963
	DA	0.266	0.295	0.314	0.319	0.315	0.311
	DAP	0.723	0.820	0.883	0.924	0.948	0.962
	DH	0.709	0.805	0.877	0.925	0.950	0.963
	DLDMZEPD	0.730	0.826	0.889	0.929	0.952	0.963
	EP	0.706	0.828	0.974	0.910	0.946	0.959
	Filli	0.712	0.805	0.875	0.922	0.949	0.962
	GG	0.658	0.760	0.850	0.915	0.949	0.962
	GLB	0.712	0.798	0.863	0.909	0.943	0.918
	GMG	0.787	0.862	0.914	0.946	0.965	0.975
	H1	0.799	0.862	0.852	0.999	0.999	0.999
	JB	0.643	0.762	0.856	0.918	0.949	0.963
	KS	0.585	0.723	0.789	0.836	0.905	0.939
	Lillie	0.669	0.758	0.828	0.883	0.921	0.947
	MI	0.632	0.676	0.705	0.724	0.736	0.745
	N-metric	0.245	0.585	0.971	0.999	0.999	0.999
	SF	0.715	0.807	0.876	0.923	0.949	0.962
	SW	0.718	0.808	0.874	0.919	0.946	0.962
	SWRG	0.694	0.775	0.834	0.882	0.916	0.946
	ZQstar	0.513	0.576	0.630	0.669	0.697	0.718
	ZW2	0.715	0.806	0.869	0.912	0.939	0.957

Table 2. Average empirical power obtained for a group of asymmetric distributions.

		Sample Size					
		32	**64**	**128**	**256**	**512**	**1024**
Tests	**AD**	0.729	0.835	0.908	0.949	0.969	0.984
	BCMR	0.749	0.856	0.924	0.971	0.995	0.991
	BHS	0.529	0.664	0.769	0.855	0.915	0.950
	BHSBS	0.538	0.652	0.747	0.914	0.902	0.944
	BM2	0.737	0.859	0.931	0.965	0.981	0.993
	BS	0.506	0.588	0.665	0.738	0.805	0.859

Table 2. *Cont.*

		Sample Size					
		32	**64**	**128**	**256**	**512**	**1024**
Tests	**CC2**	0.579	0.682	0.777	0.853	0.938	0.956
	CHI2	0.645	0.799	0.881	0.934	0.965	0.980
	CVM	0.594	0.755	0.836	0.887	0.935	0.957
	ChenS	0.756	0.862	0.928	0.961	0.978	0.991
	Coin	0.480	0.556	0.630	0.700	0.769	0.916
	DA	0.237	0.223	0.209	0.198	0.191	0.192
	DAP	0.705	0.826	0.910	0.955	0.977	0.990
	DH	0.724	0.845	0.921	0.957	0.977	0.991
	DLDMXAPD	0.726	0.843	0.918	0.955	0.975	0.989
	EP	0.753	0.846	0.913	0.967	0.975	0.993
	Filli	0.732	0.842	0.915	0.953	0.974	0.991
	GG	0.672	0.805	0.898	0.949	0.973	0.988
	GLB	0.725	0.831	0.905	0.987	0.970	0.984
	GMG	0.683	0.751	0.809	0.859	0.901	0.932
	H1	0.816	0.896	0.896	0.999	0.999	0.999
	JB	0.662	0.808	0.904	0.953	0.975	0.989
	KS	0.582	0.736	0.810	0.863	0.921	0.945
	Lillie	0.671	0.786	0.872	0.929	0.959	0.976
	MI	0.644	0.731	0.798	0.843	0.872	0.913
	N-metric	0.464	0.761	0.990	0.999	0.999	0.999
	SF	0.736	0.846	0.918	0.955	0.975	0.989
	SW	0.753	0.859	0.925	0.959	0.977	0.991
	SWRG	0.758	0.861	0.927	0.960	0.977	0.999
	ZQstar	0.570	0.639	0.693	0.732	0.761	0.748
	ZW2	0.764	0.870	0.932	0.962	0.980	0.997

Table 3. Average empirical power obtained for a group of modified normal distributions.

		Sample Size					
		32	**64**	**128**	**256**	**512**	**1024**
Tests	**AD**	0.662	0.756	0.825	0.872	0.905	0.931
	BCMR	0.652	0.756	0.831	0.880	0.913	0.935
	BHS	0.463	0.585	0.676	0.744	0.796	0.834
	BHSBS	0.568	0.701	0.787	0.847	0.890	0.918
	BM2	0.641	0.770	0.854	0.904	0.934	0.953
	BS	0.587	0.688	0.770	0.833	0.881	0.916
	CC2	0.576	0.675	0.763	0.833	0.887	0.923
	CHI2	0.566	0.728	0.808	0.866	0.914	0.939
	CVM	0.557	0.708	0.779	0.833	0.897	0.930
	ChenS	0.656	0.759	0.833	0.882	0.915	0.937
	Coin	0.579	0.691	0.781	0.846	0.889	0.918
	DA	0.314	0.342	0.367	0.388	0.405	0.418
	DAP	0.617	0.733	0.818	0.872	0.906	0.930
	DH	0.617	0.727	0.815	0.872	0.907	0.930
	DLDMXAPD	0.651	0.754	0.831	0.879	0.912	0.935
	EP	0.640	0.748	0.819	0.865	0.906	0.931
	Filli	0.637	0.743	0.823	0.877	0.911	0.933
	GG	0.529	0.657	0.775	0.860	0.906	0.932
	GLB	0.659	0.755	0.823	0.870	0.903	0.930
	GMG	0.688	0.771	0.836	0.883	0.917	0.942
	H1	0.743	0.816	0.799	0.999	0.999	0.999
	JB	0.515	0.662	0.783	0.861	0.904	0.930
	KS	0.564	0.710	0.772	0.825	0.893	0.924
	Lillie	0.626	0.724	0.796	0.850	0.889	0.917
	MI	0.494	0.536	0.563	0.578	0.585	0.590
	N-metric	0.243	0.582	0.972	0.999	0.999	0.999

Table 3. *Cont.*

		Sample Size					
		32	**64**	**128**	**256**	**512**	**1024**
Tests	**SF**	0.642	0.747	0.826	0.879	0.912	0.934
	SW	0.654	0.758	0.832	0.882	0.915	0.937
	SWRG	0.643	0.746	0.818	0.864	0.901	0.931
	ZQstar	0.394	0.423	0.450	0.472	0.487	0.498
	ZW2	0.640	0.749	0.826	0.876	0.907	0.931

An additional study was conducted to determine the exact minimal sample size at which the *N-metric* test (statistic (34) with kernel function (35)) is the most powerful for groups of symmetric, asymmetric, and modified normal distributions. *Hosking1* and *N-metric* tests were applied for data sets of sizes: 80, 90, 100, 105, 110, and 115. The obtained results showed that the *N-metric* test was the most powerful for sample size ≥ 112 for the symmetric distributions, for sample size ≥ 118 for the asymmetric distributions, and for sample size ≥ 88 for a group of modified normal distributions (see in Table 4). The *N-metric* test is the most powerful for the Gamma distribution for sample size ≥ 32. It has been observed that in the case of Cauchy and Lognormal distributions, the *N-metric* test is the most powerful when the sample size is ≥ 255, which can be influenced by the long tail of these distributions.

Table 4. The minimal sample size at which the *N-metric* test is most powerful.

Nr.	**Distribution**	**Groups of Distributions**	**Minimal Sample Size (*n*)**
1.	Standard normal	Symmetric	46
2.	Beta	Symmetric	88
3.	Cauchy	Symmetric	257
4.	Laplace	Symmetric	117
5.	Logistic	Symmetric	71
6.	Student	Symmetric	96
7.	Beta	Asymmetric	108
8.	Chi-square	Asymmetric	123
9.	Gamma	Asymmetric	<32
10.	Gumbel	Asymmetric	125
11.	Lognormal	Asymmetric	255
12.	Weibull	Asymmetric	65
13.	Normal1	Modified normal	70
14.	Normal2	Modified normal	93
15.	Normal3	Modified normal	72
16.	Normal4	Modified normal	117

To complement the results given in Tables 1–3, Figure 4 (and Figures A1–A3 in Appendix A) presents the average power results of the most powerful goodness-of-fit tests. Figure 4 presents two distributions from each group of symmetric (Standard normal and Student), asymmetric (Gamma and Gumbel), and modified normal (standard normal distribution truncated at *a* and *b* and location-contaminated standard normal distribution) distributions. Figures of all other distributions are given in Appendix A. In Figure 4, it can be seen that for the standard normal distribution, our proposed test (*N-metric*) is the most powerful when the sample size is 64 or larger. Figure 4 shows that our proposed test (*N-metric*) is the most powerful in the case of Gamma data distribution for all sample sizes examined. In general, it can be summarized that the power of the Chen–Shapiro (*ChenS*), Gel–Miao–Gastwirth (*GMG*), *Hosking1* (*H1*), and Modified Shapiro–Wilk (*SWRG*) tests increases gradually with increasing sample size. The power of our proposed test (*N-metric*) increases abruptly when the sample size is 128 and its power value remains close to 1 for larger sample sizes.

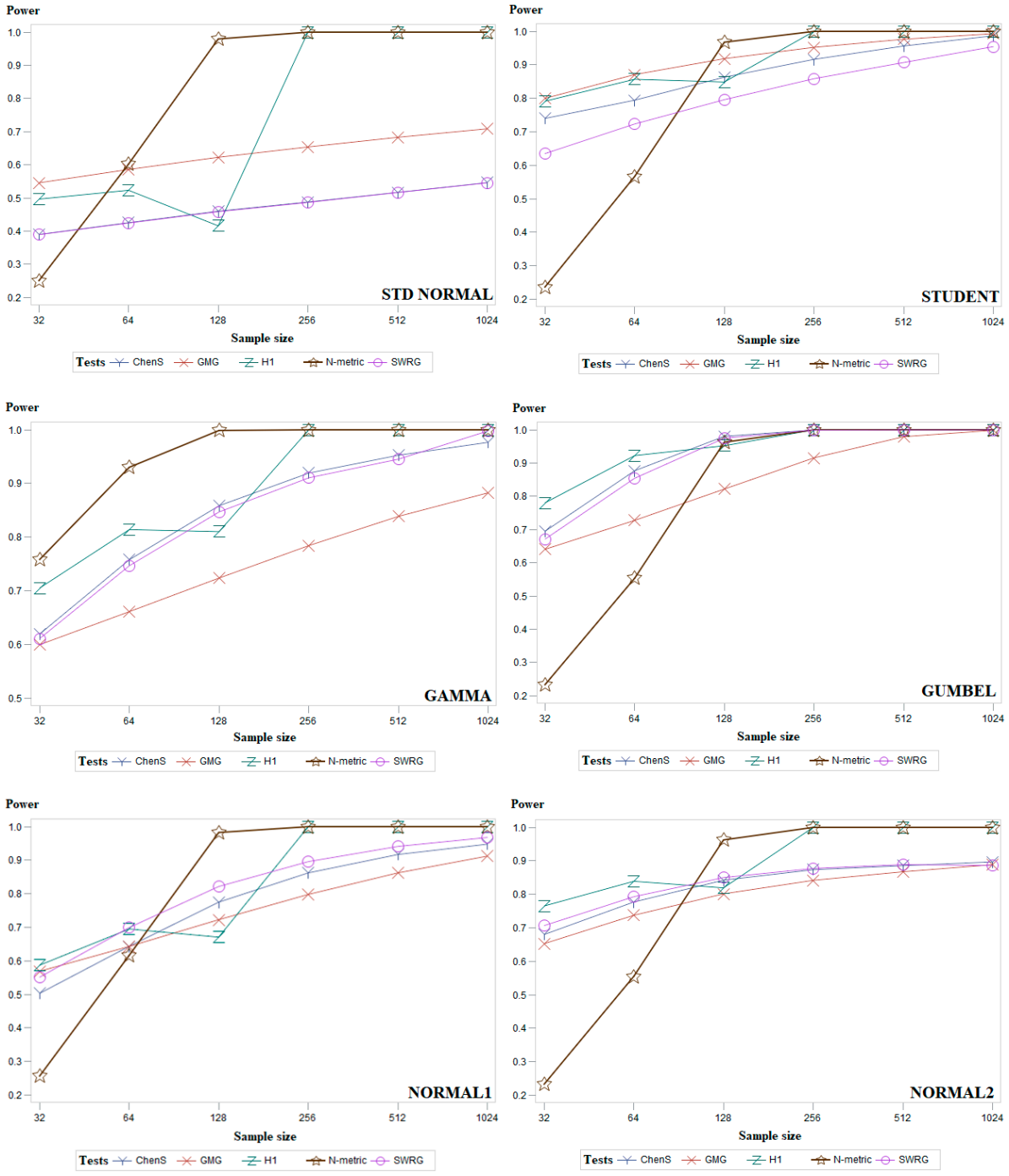

Figure 4. Average empirical power results, for the examined sample sizes, for the groups of symmetric, asymmetric, and modified normal distributions of five powerful goodness-of-fit tests.

6. Conclusions and Future Work

In this study, a comprehensive comparison of the power of popular normality tests was performed. Given the importance of this topic and the extensive development of normality tests, the proposed new normality test, the detailed test descriptions provided, and the power comparisons are relevant. Only univariate data were examined in this study of the power of normality tests (a study with multivariate data is planned for the future).

The study addresses the performance of 40 normality tests, for various sample sizes n for a number of symmetric, asymmetric, and modified normal distributions. A new goodness-of-fit test has been proposed. Its results are compared with other tests.

Based on the obtained modeling results, it was determined that the most powerful tests for the groups of symmetric, asymmetric, and modified normal distributions were *Hosking1* (for smaller sample sizes) and our proposed *N-metric* (for larger sample sizes) test. The power of the *Hosking1* test (for smaller sample sizes) is 1.5 to 7.99 percent higher than the second (by power) test for the groups of symmetric, asymmetric, and modified normal distributions. The power of the *N-metric* test (for larger sample sizes) is 6.2 to 16.26 percent higher than the second (by power) test for the groups of symmetric, asymmetric, and modified normal distributions.

The *N-metric* test is recommended to be used for symmetric data sets of size $n \geq 112$, for asymmetric data sets of size $n \geq 118$, and for bell-shaped distributed data sets of size $n \geq 88$.

Author Contributions: Data curation, J.A. and T.R.; formal analysis, J.A. and T.R.; investigation, J.A. and T.R.; methodology, J.A. and T.R.; software, J.A. and T.R.; supervision, T.R.; writing—original draft, J.A. and M.B.; writing—review and editing, J.A. and M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Generated data sets were used in the study (see in Section 4).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

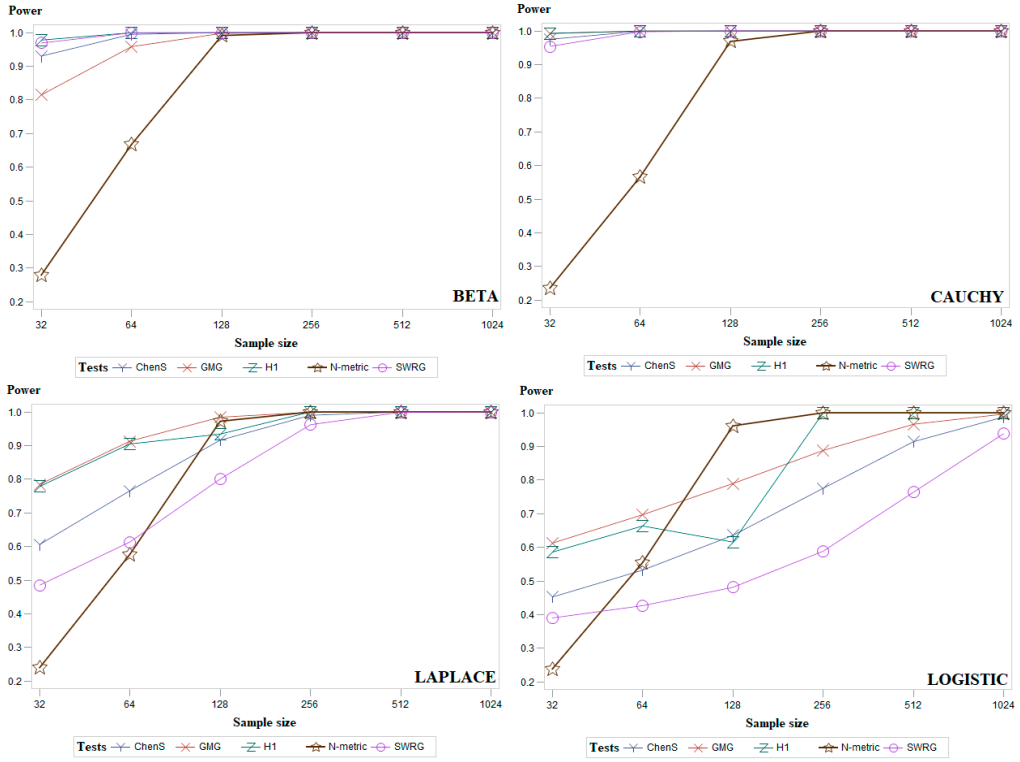

Figure A1. Average empirical power results, for all sample sizes, for the groups of symmetric distributions of five powerful goodness-of-fit tests.

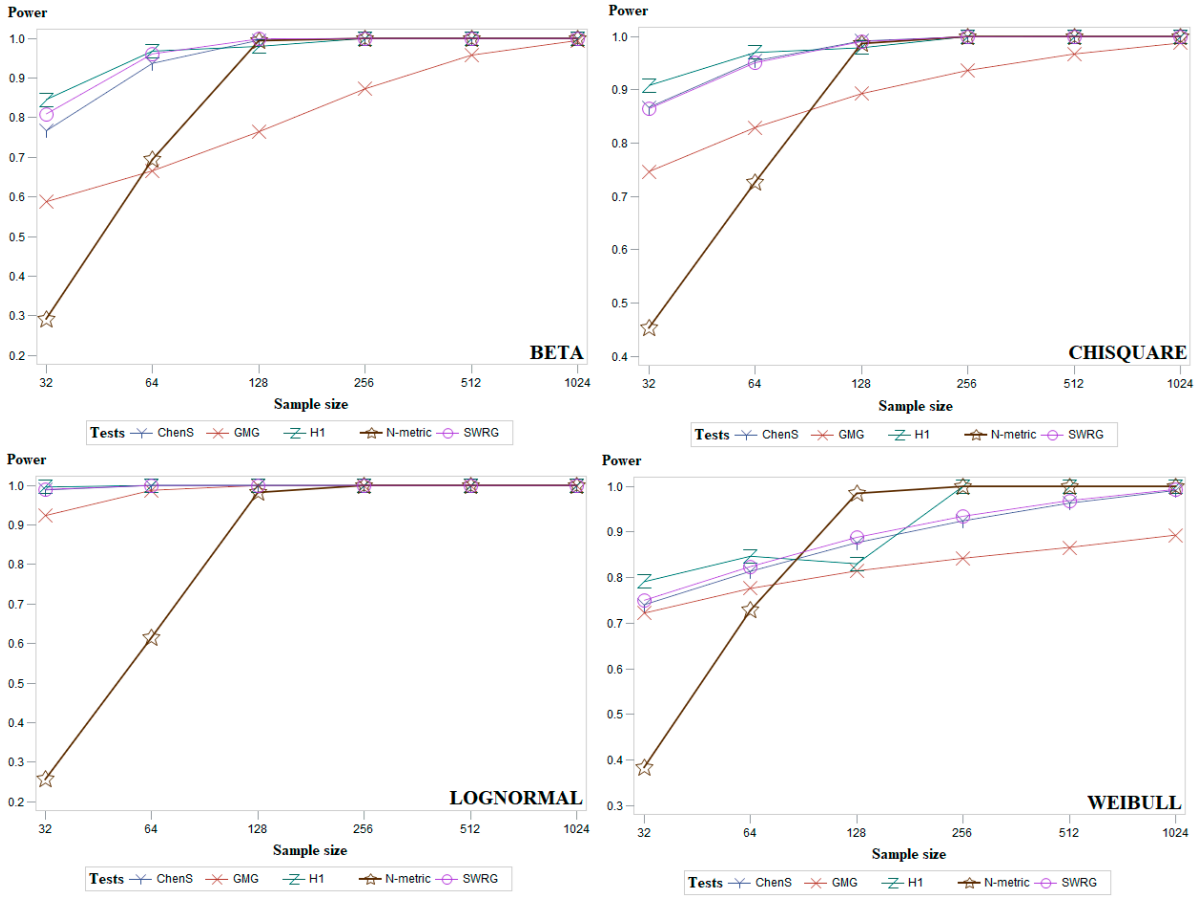

Figure A2. Average empirical power results for the examined sample sizes for the groups of asymmetric distributions of five powerful goodness-of-fit tests.

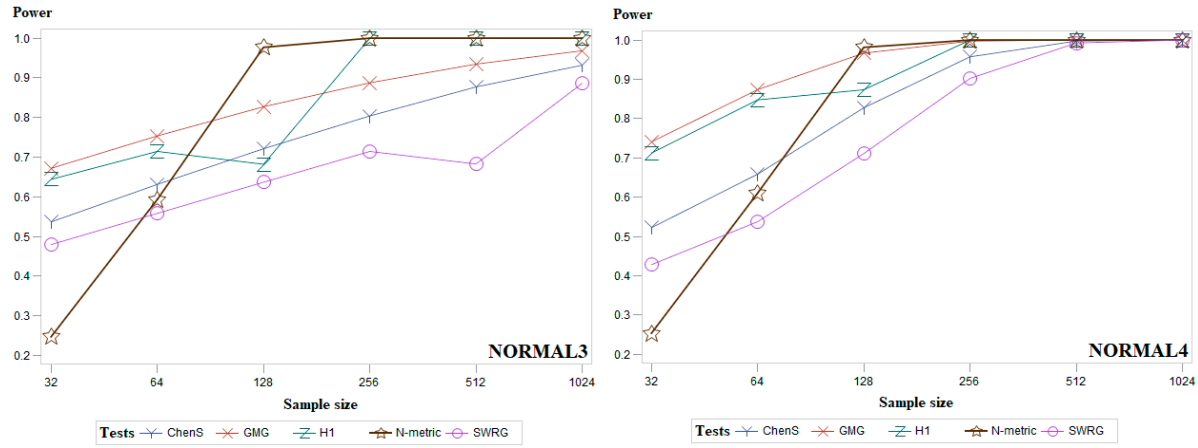

Figure A3. Average empirical power results for the examined sample sizes for the groups of the modified normal distributions of five powerful goodness-of-fit tests.

References

1. Barnard, G.A.; Barnard, G.A. *Introduction to Pearson (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*; Springer Series in Statistics Breakthroughs in Statistics; Springer: Cham, Switzerland, 1992; pp. 1–10.
2. Kolmogorov, A. Sulla determinazione empirica di una legge di distribuzione. *Inst. Ital. Attuari Giorn.* **1933**, *4*, 83–91.
3. Adefisoye, J.; Golam Kibria, B.; George, F. Performances of several univariate tests of normality: An empirical study. *J. Biom. Biostat.* **2016**, *7*, 1–8.

4. Anderson, T.W.; Darling, D.A. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Ann. Math. Stat.* **1952**, *23*, 193–212. [CrossRef]
5. Hosking, J.R.M.; Wallis, J.R. Some statistics useful in regional frequency analysis. *Water Resour. Res.* **1993**, *29*, 271–281. [CrossRef]
6. Cabana, A.; Cabana, E.M. Goodness-of-Fit and Comparison Tests of the Kolmogorov-Smirnov Type for Bivariate Populations. *Ann. Stat.* **1994**, *22*, 1447–1459. [CrossRef]
7. Chen, L.; Shapiro, S.S. An Alternative Test for Normality Based on Normalized Spacings. *J. Stat. Comput. Simul.* **1995**, *53*, 269–288. [CrossRef]
8. Rahman, M.M.; Govindarajulu, Z. A modification of the test of Shapiro and Wilk for normality. *J. Appl. Stat.* **1997**, *24*, 219–236. [CrossRef]
9. Ray, W.D.; Shenton, L.R.; Bowman, K.O. Maximum Likelihood Estimation in Small Samples. *J. R. Stat. Soc. Ser. A* **1978**, *141*, 268. [CrossRef]
10. Zhang, P. Omnibus test of normality using the Q statistic. *J. Appl. Stat.* **1999**, *26*, 519–528. [CrossRef]
11. Barrio, E.; Cuesta-Albertos, J.A.; Matrán, C.; Rodríguez-Rodríguez, J.M. Tests of goodness of fit based on the L2-Wasserstein distance. *Ann. Stat.* **1999**, *27*, 1230–1239.
12. Glen, A.G.; Leemis, L.M.; Barr, D.R. Order statistics in goodness-of-fit testing. *IEEE Trans. Reliab.* **2001**, *50*, 209–213. [CrossRef]
13. Bonett, D.G.; Seier, E. A test of normality with high uniform power. *Comput. Stat. Data Anal.* **2002**, *40*, 435–445. [CrossRef]
14. Psaradakis, Z.; Vávra, M. Normality tests for dependent data: Large-sample and bootstrap approaches. *Commun. Stat.-Simul. Comput.* **2018**, *49*, 283–304. [CrossRef]
15. Zhang, J.; Wu, Y. Likelihood-ratio tests for normality. *Comput. Stat. Data Anal.* **2005**, *49*, 709–721. [CrossRef]
16. Gel, Y.R.; Miao, W.; Gastwirth, J.L. Robust directed tests of normality against heavy-tailed alternatives. *Comput. Stat. Data Anal.* **2007**, *51*, 2734–2746. [CrossRef]
17. Coin, D. A goodness-of-fit test for normality based on polynomial regression. *Comput. Stat. Data Anal.* **2008**, *52*, 2185–2198. [CrossRef]
18. Desgagné, A.; Lafaye de Micheaux, P. A powerful and interpretable alternative to the Jarque–Bera test of normality based on 2nd-power skewness and kurtosis, using the Rao's score test on the APD family. *J. Appl. Stat.* **2017**, *45*, 2307–2327. [CrossRef]
19. Steele, C.M. The Power of Categorical Goodness-Of-Fit Statistics. Ph.D. Thesis, Australian School of Environmental Studies, Warrandyte, Victoria, Australia, 2003.
20. Romão, X.; Delgado, R.; Costa, A. An empirical power comparison of univariate goodness-of-fit tests for normality. *J. Stat. Comput. Simul.* **2010**, *80*, 545–591. [CrossRef]
21. Choulakian, V.; Lockhart, R.; Stephens, M. Cramérvon Mises statistics for discrete distributions. *Can. J. Stat.* **1994**, *22*, 125–137. [CrossRef]
22. Shapiro, S.S.; Wilk, M.B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52*, 591–611. [CrossRef]
23. Lilliefors, H.W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J. Am. Stat. Assoc.* **1967**, *62*, 399–402. [CrossRef]
24. Ahmad, F.; Khan, R.A. A power comparison of various normality tests. *Pak. J. Stat. Oper. Res.* **2015**, *11*, 331. [CrossRef]
25. D'Agostino, R.B.; Pearson, E.S. Testing for departures from normality. I. Fuller empirical results for the distribution of b_2 and $\sqrt{b_1}$. *Biometrika* **1973**, *60*, 613–622.
26. Filliben, J.J. The Probability Plot Correlation Coefficient Test for Normality. *Technometrics* **1975**, *17*, 111–117. [CrossRef]
27. Martinez, J.; Iglewicz, B. A test for departure from normality based on a biweight estimator of scale. *Biometrika* **1981**, *68*, 331–333. [CrossRef]
28. Epps, T.W.; Pulley, L.B. A test for normality based on the empirical characteristic function. *Biometrika* **1983**, *70*, 723–726. [CrossRef]
29. Jarque, C.; Bera, A. Efficient tests for normality, homoscedasticity andserial independence of regression residuals. *Econ. Lett.* **1980**, *6*, 255–259. [CrossRef]
30. Bakshaev, A. Goodness of fit and homogeneity tests on the basis of N-distances. *J. Stat. Plan. Inference* **2009**, *139*, 3750–3758. [CrossRef]
31. Hill, T.; Lewicki, P. *Statistics Methods and Applications*; StatSoft: Tulsa, OK, USA, 2007.
32. Kasiulevičius, V.; Denapienė, G. Statistikos taikymas mokslinių tyrimų analizėje. *Gerontologija* **2008**, *9*, 176–180.
33. Damianou, C.; Kemp, A.W. New goodness of statistics for discrete and continuous data. *Am. J. Math. Manag. Sci.* **1990**, *10*, 275–307.