Testing for Normality from the Parametric Seven-Number Summary

José Moral De La Rubia

School of Psychology, Universidad Autónoma de Nuevo León, Nuevo León, Mexico
Email: jose.morald@uanl.edu.mx, jose_moral@hotmail.com

Abstract

The objective of this study is to propose the Parametric Seven-Number Summary (PSNS) as a significance test for normality and to verify its accuracy and power in comparison with two well-known tests, such as Royston's W test and D'Agostino-Belanger-D'Agostino K-squared test. An experiment with 384 conditions was simulated. The conditions were generated by crossing 24 sample sizes and 16 types of continuous distributions: one normal and 15 non-normal. The percentage of success in maintaining the null hypothesis of normality against normal samples and in rejecting the null hypothesis against non-normal samples (accuracy) was calculated. In addition, the type II error against normal samples and the statistical power against normal samples were computed. Comparisons of percentage and means were performed using Cochran's Q-test, Friedman's test, and repeated measures analysis of variance. With sample sizes of 150 or greater, high accuracy and mean power or type II error (≥0.70 and ≥0.80, respectively) were achieved. All three normality tests were similarly accurate; however, the PSNS-based test showed lower mean power than K-squared and W tests, especially against non-normal samples of symmetrical-platykurtic distributions, such as the uniform, semi-circle, and arcsine distributions. It is concluded that the PSNS-based omnibus test is accurate and powerful for testing normality with samples of at least 150 observations.

Keywords

Normality Tests, Parametric Omnibus Test, Quantiles, Accuracy, Potency

1. Introduction

1.1. The Summary of Five and Seven Numbers and Its Graphic Application

In the first edition of the *Elements of Statistics*, Bowley [1] introduced

stem-and-leaf and cumulative frequency graphs, advocated the use of simple random sampling, highlighted the descriptive importance of position measures, as reflected in the five-number summary (minimum value, the three quartiles and maximum value), and developed the quartile coefficient of skewness [2].

In the first edition of the *Elementary Manual of Statistics*, Bowley [3] expanded the summary from five to seven numbers, by including the first and last deciles, allowing two extremes of deviation to be defined. Very low scores are below the 10th percentile and very high scores are above the 90th percentile. The central tendency or shoulder zone extends into the interquartile range, low scores are between the 10th and 25th percentiles, and high scores are between the 75th and 90th percentiles. Thus, the seven-number summary is not only useful for calculating statistics on central tendency (median), variation (absolute and relative ranges), skewness (quartile coefficient) and kurtosis (percentile coefficient), but also to interpret the scores [4].

The visual representation of the five-number summary led Spear to develop the range bar graph among her charting techniques [5] [6], and Tukey to create the box-and-whisker plot among his graphical tools for exploratory data analysis [7] [8] [9]. Thanks to computational statistics and Tukey's exploratory data analysis, the use of the box-and-whisker plot to evaluate symmetry, kurtosis, the presence of outliers, and fit or proximity to the normal distribution was widely extended [10]. With Tukey [9], outliers are defined by fences. The lower fence is placed at a distance of one and a half times the interquartile range, starting from the first quartile. On the other hand, the upper fence is placed at a distance of one and a half times the interquartile range from the third quartile. If scores are found below the lower fence, then there are outliers in the left tail. Similarly, if there are scores above the upper fence, then there are outliers in the right tail [11]. McGill, Tukey, and Larsen [12] highlighted the possibility of representing two or more variables on a box-and-whisker plot and introduced three variants of the basic design. The first variant incorporates a visual measurement of the sample size across the width of the boxes. The second variant includes an indication of the rough significance of differences between medians across notches, and the third variant combines the characteristics of the two previous graphs.

Other modifications to Tukey's box-and-whisker plot have been proposed, such as delimiting the whiskers with the extreme deciles [13] [14] or the 5th and 95th or 2nd and 98th percentiles [15], give a smooth profile to the boxes to represent their density in the so-called vase plot [16], expand the surface of smooth edges for all the data in the so-called violin plot [17], and include a horizontal line for the mean and give more detail to the representation of the density of the data in the bean plot [18]. In turn, there are two-dimensional forms of the box-and-whisker plot [19] [20] [21] [22] and a version for bivariate data named bag-and-bolster plot [23].

1.2. Assessment of Normality

For the evaluation of normality, there are three basic strategies: one uses de-

scriptive statistics, another applies tests of significance, and the third analyzes graphs. From the descriptive statistics, it is observed if the 95% confidence intervals of the arithmetic mean, median, and mode overlap; and the 95% confidence intervals of the coefficients of skewness and excess kurtosis (based on standardized central moments) include zero. In the case of a variable of positive values, it can be verified if the lower limit of the 95% confidence interval of Pearson's coefficient of variation is less than a half, and that of the quartile coefficient of dispersion is less than a third. In addition, it can be checked if approximately 68.3% of the data is concentrated at a standard deviation above and below the arithmetic mean, 95.4% at two standard deviations, and 99.7% at three standard deviations [24].

The hypothesis-testing strategy has five variants. The first variant is based on the goodness of fit or difference between the observed and expected absolute frequencies for k class intervals, such as the Pearson's chi-squared test [25] and Woolf's G test [26]. The second variant is based on the largest difference between the empirical and theoretical cumulative distributions of each sample data, such as the Kolmogorov [27] and Smirnov [28] D test and Kuiper's V test [29]. The third variant s very similar to the second and minimizes the squared vertical distance between empirical and theoretical cumulative distributions, such as the Cramer [30] and von Mises [31] W-squared test, Anderson and Darling's A-squared test [32], and Watson's U-squared test [33]. The fourth variant is based on the correction between the observed and expected quantiles, such as the W tests of Shapiro and Wilk [34], Royston [35], and Chen and Shapiro [36], the W' tests of Shapiro and Francia [37] and Royston [38], and the D test D'Agostino [39] [40]. Finally, the fifth variant is based on standardized central moments, such as the tests of D'Agostino and Pearson [41], Jarque and Bera [42], Urzua [43], and D'Agostino, Belanger, and D'Agostino [44].

The graphical strategy observes if the histogram has a bell shape [45], the normal quantile-quantile plot shows an alignment of the cloud of points on a 45-degree fit line [46], and the box-and-whisker plot has symmetry in its elements and the absence of atypical cases [9] [44].

In the graphical evaluation of normality, Cleveland [13] proposed to include the extreme deciles (10th and 90th percentiles) in the box-and-whisker plot. Note that when rounded to two decimal places, the 2nd, 8th, 25th, 50th, 75th, 92nd, and 98th percentiles of a standard normal distribution are almost uniformly spaced. The second percentile corresponds to −2.05, the eighth to −1.34, the twenty-fifth to −0.67, the fiftieth to 0, the seventy-fifth to 0.67, the ninety-second to 1.34, and the ninety-eighth to 2.05. Within the sequence, a distance of 0.67 separates them, except at the two extremes that present a discrepancy of fewer than 5 tenths. To be exact, the quantiles of orders 0.022 and 0.978 should be used. These percentiles are close to those of the seven-number summary: minimum or 0 percentile, 10th percentile, 50th percentile, 75th percentile, 90th percentile, and maximum or 100th percentile. Since it is a specific property of

the normal distribution, which is a key distribution for parametric statistics, this group of percentiles is called a parametric seven-number summary, unlike the one defined by Bowley [3], which is more general and is named a non-parametric seven-number summary [47] [48]. Based on this property of the normal distribution, the box-and-whisker plot, with notches at the 2nd and 98th percentiles, allows us to visualize whether the distance between the notches and the hinges (quartiles) are constant and there are no outliers, which would show that the sample data present a good fit to normality [4].

1.3. Proposal of a New Test from the Parametric Seven-Number Summary

Currently, the parametric seven-number summary is used for the visual assessment of normality, but the present study proposes its application through a test of statistical significance. A constant distance or amplitude of two thirds between the numbers is stipulated, so that the seven points on the ordinate axis of the Probit function or quantile function of the standard normal distribution would be: -2, $-1.\hat{3}$, $-0.\hat{6}$, 0, $0.\hat{6}$, $1.\hat{3}$ and 2, and the orders or values on the abscissa axis would be: 0.023, 0.091, 0.252, 0.5, 0.748, 0.909 and 0.977.

The sampling distribution of the quantiles of a distribution that has its finite moments corresponds to a normal distribution [49]. Given a sequence of n samples of a random variable X with density function $f_X(x)$ and finite moments, the sampling distribution of the quantile of order p, $q_X(p)$, is a normal distribution, whose mean is the population quantile $\varrho_X(p)$ and its variance is the quotient between the product of the order of the quantile and its complement (numerator) and the product of the sample size and the square of the density function evaluated in the population quantile (denominator). The standard deviation or error (*se*) is the root of this quotient:

$$se_{q_X(p)} = \sqrt{\frac{p(1-p)}{nf_X^2\left[\varrho_X(p)\right]}}$$

Using this formula, the standard errors of the seven quantiles that constitute the parametric seven-number summary can be obtained, which are standard normal distribution quantiles. See **Table 1**.

The standard errors in **Table 1** allow us to estimate by intervals and to check whether the sample quantiles are equivalent to those expected for a standard normal distribution.

$$Z \sim N(0,1)$$

$$P\left(\varrho_Z(p) \in [LL, UL]\right) = 1 - \alpha$$

$$P\left(\varrho_Z(p) \in \left[q_Z(p) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{nf_Z^2(\varrho_Z(p))}}, q_Z(p) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{nf_Z^2(\varrho_Z(p))}}\right]\right) = 1 - \alpha$$

LL = lower limit of the interval for the population quantile of order *p*.

Table 1. Value in the Probit function, cumulative distribution function, density function and standard error of the parametric seven-number summary.

$\varrho_Z(p) = \Phi^{-1}(p)$	p-order	$f_Z(z_p) = \varphi(z_p)$	$se_{q_{Z(p)}}$
-2	0.023	0.054	$2.762/\sqrt{n}$
$-1.\hat{3}$	0.091	0.164	$1.755/\sqrt{n}$
$-0.\hat{6}$	0.252	0.319	$1.360/\sqrt{n}$
0	0.500	0.399	$1.253/\sqrt{n}$
$0.\hat{6}$	0.748	0.319	$1.360/\sqrt{n}$
$1.\hat{3}$	0.909	0.164	$1.755/\sqrt{n}$
2	0.977	0.054	$2.762/\sqrt{n}$

Note. $\varrho_Z(p) = \Phi^{-1}(p)$ = Probit function, quantile function or inverse function of the cumulative distribution function of a variable Z with standard normal distribution $N(0, 1)$ that gives the population quantiles z_p, that is, the values of Z; p-order = order of the quantile z_p of a standard normal distribution; $f_Z(z_p) = \varphi(z_p)$ = density function of a standard normal distribution; and $se_{q_{Z(p)}}$ = standard error of the sample quantile of order p calculated from a random sample of size n drawn from a standard normal distribution.

UL = upper limit of the interval for the population quantile of order p.

$\varrho_Z(p)$ = population quantile of order p, where ϱ (lowercase koppa) is the archaic Greek letter corresponding to the lowercase Latin letter q.

$q_Z(p)$ = sample quantile or point estimator of the population quantile of order p.

n = sample size.

$f_Z^2(\varrho_Z(p)) = f_Z^2(z_p)$ = square of the height of point $\varrho_Z(p) = z_p$ in a standard normal distribution.

$z_{1-\alpha/2}$ = quantile of order $1 - (\alpha/2)$ of a standard normal distribution, where α is the significance level, usually 0.05 ($z_{0.975} = 1.96$).

Statistical hypotheses. null hypothesis: $H_0 : q_Z(p) = \varrho_Z(p)$ and alternative hypothesis: $H_1 : q_Z(p) \neq \varrho_Z(p)$.

$q_Z(p)$ = sample quantile or point estimator of the population quantile of order p.

$\varrho_Z(p)$ = hypothetical or population value of the quantile of order p.

Assumptions. A large random sample drawn from a normal distribution with unknow parameters $N(\mu, \sigma^2)$.

Test statistics and sampling distribution:

$$Z_{q_Z(p)} = \frac{q_Z(p) - \varrho_Z(p)}{\sqrt{\dfrac{p(1-p)}{nf_Z^2[\varrho_Z(p)]}}} = z_{q_Z(p)}$$

$$Z_{q_Z(p)} \sim N(0,1)$$

Decision. Let $P\left(Z \le \left|z_{q_{Z}(p)}\right|\right)$ be the probability of obtaining a value less than or equal to the absolute value of the test statistic in a standard normal distribution. If $2 \times \left(1 - P\left(Z \le \left|z_{q_{Z}(p)}\right|\right)\right) \le \alpha$, the null hypothesis is accepted in a two-tailed test at a significance level of α. Conversely, if $2 \times \left(1 - P\left(Z \le \left|z_{q_{Z}(p)}\right|\right)\right) > \alpha$, is rejected.

Knowing how to standardize the sample quantiles of a normal population, the planting of the PSNS as an omnibus test for normality is continued. Let X be a random variable supposedly with a normal distribution of unknown parameters $N\left(\mu_X, \sigma_X^2\right)$. A random sample of size n is obtained (at least 20 cases). The quantiles of X are estimated from this sample. By constituting the point estimate of the population quantiles, the calculation can be performed using rule 9 of linear interpolation [50], which provides approximately unbiased quantiles in the case of normality [51]. Let's see the procedure of this rule:

The n sample data of X are sorted in ascending order: $x_{(i)}; i = 1, 2, \cdots, n$

$$h = 0.375 + p \times (n + 0.25)$$

$$\hat{\varrho}_X(p) = q_X(p) = x_{(i = \lfloor h \rfloor)} + (h - \lfloor h \rfloor)\left(x_{(i = \lfloor h \rfloor + 1)} - x_{(i = \lfloor h \rfloor)}\right)$$

In case h is less than 1, the sample value in the first order $x_{(1)}$ or minimum value of the sample is taken as quantile. In case h is greater than n, the sample value in the nth order $x_{(n)}$ or maximum value of the sample is taken.

Next, the mean and variance of the population that is assumed to follow a normal distribution are estimated, and the quantiles are standardized so that they follow a standard normal distribution $N(0,1)$ in case the sample data have been drawn from a normal distribution $N\left(\mu_X, \sigma_X^2\right)$ or at least with finite moments and large sample size.

$$\hat{\mu}_X = \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\hat{\sigma}_X^2 = s_X^2 = \frac{\sum_{i=1}^{n} \left(x_i - \hat{\mu}_X^2\right)^2}{n-1}$$

$$\hat{\sigma}_X = s_X = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \hat{\mu}_X^2\right)^2}{n-1}}$$

$$z_{q_X(p)} = \frac{\hat{\varrho}_X(p) - \hat{\mu}_X}{\hat{\sigma}_X} = \frac{q_X(p) - \overline{x}}{s_X}$$

The same result is obtained if, first, the n sample data of X (z_x) are standardized, and then the sample quantiles are calculated on these standardized data.

$$z_{q_X(p)} = q_Z(p)$$

It is tested whether the standardized quantiles are equivalent to those expected for a normal distribution. For this purpose, the standardized differences between standardized sample quantiles the expected values under the null hypothesis of the normal distribution are calculated.

$$Z_{q_Z(p)} = \frac{q_Z(p) - \varrho_Z(p)}{se_{q_Z(p)}} = z_{q_Z(p)}$$

$$Z_{q_Z(p)} \sim N(0,1)$$

If the $z_{q_Z(p)}$ values are squared and added, the resulting test statistic follows a chi-squared distribution with seven degrees of freedom, since it is the sum of squares of seven variables random with standard normal distribution. Thus, an omnibus test of normality is obtained, and the test statistic is denoted by *SS*.

$$SS = \sum_{i=1}^{7} z_{q_Z(p)}^2 \sim \chi_7^2$$

$$SS = \frac{q_Z(p=0.023)+2}{2.762/\sqrt{n}} + \frac{q_Z(p=0.091)+4/3}{1.755/\sqrt{n}} + \frac{q_Z(p=0.252)+2/3}{1.360/\sqrt{n}}$$
$$+ \frac{q_Z(p=0.5)}{1.253/\sqrt{n}} + \frac{q_Z(p=0.748)-2/3}{1.360/\sqrt{n}} + \frac{q_Z(p=0.909)-4/3}{1.755/\sqrt{n}}$$
$$+ \frac{q_Z(p=0.977)-2}{2.762/\sqrt{n}}$$

Let $P(\chi_7^2 \geq SS)$ be the probability of obtaining a value greater than or equal to the absolute value of the test statistic in a chi-squared distribution with seven degrees of freedom. If $P(\chi_7^2 \geq SS) \geq \alpha$, the null hypothesis of normal distribution, $H_0 : X \sim N(\mu_X, \sigma_X^2)$, is accepted in a two-tailed test at a significance level of α. If $P(\chi_7^2 \geq SS) < \alpha$, it is rejected.

Here is an example calculation using the PSNS-based omnibus test. Let us be a random sample with 20 data points: −0.23, −1.39, 0.38, 0.52, −0.49, 0.28, −0.04, 0.11, 1.03, −0.33, −0.33, 0.06, 0.16, 0.29, −0.16, −1.06, 0.54, 0.88, −1.64 and −0.31. Check that it has been collected from a standard normal distribution at a significance level (α) of 5%. These data were generated by inverse transform sampling from a standard uniform distribution U (0, 1) and the Probit or inverse function of the cumulative distribution function of a standard normal distribution: $x_i = \Phi^{-1}(u_i)$. To simplify the calculations, the sample data were rounded to two decimal places.

The data is sorted in ascending order: $x_{(1)} = -1.64$, $x_{(2)} = -1.39$, $x_{(3)} = -1.06$, $x_{(4)} = -0.49$, $x_{(5)} = -0.330$, $x_{(6)} = -0.33$, $x_{(7)} = -0.31$, $x_{(8)} = -0.23$, $x_{(9)} = -0.16$, $x_{(10)} = -0.04$, $x_{(11)} = 0.06$, $x_{(12)} = 0.11$, $x_{(13)} = 0.16$, $x_{(14)} = 0.28$, $x_{(15)} = 0.29$, $x_{(16)} = 0.38$, $x_{(17)} = 0.52$, $x_{(18)} = 0.54$, $x_{(19)} = 0.88$ and $x_{(20)} = 1.03$.

The sample quantiles of the orders corresponding to the PSNS are calculated using Blom's rule or the Hyndman-Fan linear interpolation rule 9, which provides approximately unbiased estimates if the sample is drawn from an (assumed) normal distribution.

$$h = 0.375 + p \times (n + 0.25) = 0.375 + 0.0228 \times 20.25 = 0.8367 < 1$$

$$q(p = 0.0228) = x_{(i = \lceil h \rceil)} = x_{(1)} = -1.64$$

$$h = 0.375 + p \times (n + 0.25) = 0.375 + 0.0912 \times 20.25 = 2.2220$$

$$q_X\left(p=0.0912\right)=x_{\left(i=\lfloor h\rfloor\right)}+\left(h-\lfloor h\rfloor\right)\left(x_{\left(i=\lfloor h\rfloor+1\right)}-x_{\left(i=\lfloor h\rfloor\right)}\right)$$

$$=x_{(2)}+0.222\times\left(x_{(3)}-x_{(2)}\right)$$

$$=-1.39+0.222\times\left(-1.06-\left(-1.39\right)\right)$$

$$=-1.3167$$

$$h=0.375+p\times\left(n+0.25\right)=0.375+0.2525\times20.25=5.488$$

$$q_X\left(p=0.2525\right)=x_{(5)}+0.488\times\left(x_{(6)}-x_{(5)}\right)$$

$$=-0.33+0.488\times\left(-0.33-\left(-0.33\right)\right)$$

$$=-0.33$$

$$h=0.375+p\times\left(n+0.25\right)=0.375+0.5\times20.25=10.5$$

$$q_X\left(p=0.5\right)=x_{(10)}+0.5\times\left(x_{(11)}-x_{(10)}\right)$$

$$=-0.04+0.5\times\left(0.06-\left(-0.04\right)\right)$$

$$=0.01$$

$$h=0.375+p\times\left(n+0.25\right)=0.375+0.7475\times20.25=15.512$$

$$q\left(p=0.7475\right)=x_{(15)}+0.5\times\left(x_{(16)}-x_{(15)}\right)$$

$$=0.29+0.512\times\left(0.38-0.29\right)$$

$$=0.3361$$

$$h=0.375+p\times\left(n+0.25\right)=0.375+0.9088\times20.25=18.778$$

$$q_X\left(p=0.9088\right)=x_{(18)}+0.5\times\left(x_{(19)}-x_{(18)}\right)$$

$$=0.54+0.778\times\left(0.88-0.54\right)$$

$$=0.8045$$

$$h=0.375+p\times\left(n+0.25\right)=0.375+0.9772\times20.25=20.1643>20$$

$$q_X\left(p=0.9772\right)=x_{\left(i=\lfloor h\rfloor\right)}=x_{(20)}=1.03$$

The sample mean and standard deviation ($\bar{x}$ and s, respectively) are calculated.

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}=-\frac{1.73}{20}=-0.0865$$

$$s=\sqrt{\frac{\sum_{i=1}^{n}\left(x_i-\hat{\mu}_X^2\right)^2}{n-1}}=\sqrt{\frac{8.975255}{19}}=0.6873$$

The sample quantiles are standardized. For the quantile of order 0.0228, its standardized value is −2.2603. For the rest of the quantiles see **Table 2**.

$$z_{q_X}\left(p=0.0228\right)=\frac{q_X\left(p=0.0228\right)-\bar{x}}{s}=\frac{-1.64-\left(-0.0865\right)}{0.6873}=-2.2603$$

The standard error of each sample quantile is calculated under the assumption that the sample data were drawn from a normal distribution. The standard error corresponding to the sample quantile of order 0.0228 is 0.6175. For the rest of the quantiles see **Table 2**.

Table 2. Testing for normality using the PSNS-based test.

	Theoretical calculation			Empirical calculation			
z_p	$p = \Phi(z_p)$	$\varphi(z_p)$	$q_X(p)$	$z_{q_X}(p) = q_Z(p)$	$se_{q_Z(p)}$	$z_{q_Z(p)}$	$z^2_{q_Z(p)}$
-2	0.0228	0.0540	-1.64	-2.2603	0.6175	-0.4215	0.1777
$-1.\hat{3}$	0.0912	0.1640	-1.3167	-1.7899	0.3925	-1.1633	1.3532
$-0.\hat{6}$	0.2525	0.3194	-0.33	-0.3543	0.3041	1.0272	1.0552
0	0.5000	0.3989	0.01	0.1404	0.2802	0.5010	0.2510
$0.\hat{6}$	0.7475	0.3194	0.3361	0.6148	0.3041	-0.1704	0.0290
$1.\hat{3}$	0.9088	0.1640	0.8045	1.2964	0.3925	-0.0941	0.0089
2	0.9772	0.0540	1.03	1.6245	0.6175	-0.6081	0.3698
Σ							3.2448

Note. z_p = quantiles equispaced at a distance of two thirds in a standard normal distribution $N(0, 1)$, $p = \Phi(z_p)$ = order of the quantile in a standard normal distribution, $\varphi(z_p)$ = point probability or height of the quantile in a standard normal distribution, $q_X(p)$ = sample quantile of order p calculated by the linear interpolation rule 9 (Hyndman & Fan, 1996) or Blom's rule (1958), $z_{q_X}(p) = z_{q_Z}(p)$ = standardized value of the quantile sample that is equivalent to calculating the quantile from the standardized data, $se_{q_Z(p)}$ = standard error of the sample quantile under a normality assumption, $z_{q_Z(p)}$ = standardized difference with respect to the expected value under the hypothesis null of normal distribution, $z^2_{q_Z(p)}$ = square of the standardized difference with respect to the expected value under the null hypothesis of normal distribution, Σ = sum per column.

$$se_{q_X(p=0.0228)} = \sqrt{\frac{p(1-p)}{nf_Z^2(z_p)}} = \sqrt{\frac{0.0228 \times (1-0.0228)}{20 \times 0.0540^2}} = 0.6175$$

The standardized sample quantile (minuend) and the expected value under the normal distribution (subtrahend) are subtracted, and the difference (numerator) is divided by the standard error of the sample quantile (denominator). In the case of the sample quantile of order 0.0228, this standardized distance is -0.4220. For the rest of the quantiles see **Table 2**.

$$z_{q_Z(p=0.0228)} = \frac{q_Z(p) - \varrho_Z(p)}{se_{q_Z(p)}} = \frac{-2.2603 - 2}{0.6175} = -0.4220$$

These standardized distances are squared and summed to obtain the test statistic or sum of squares (*SS*) of the standardized distance. As can be seen in **Table 2**.

$$SS = \sum_{i=1}^{7} z^2_{q_Z(p)} = 3.2448$$

The p value is obtained as the probability of having a value greater than or equal to *SS* in a chi-squared distribution with seven degrees of freedom. Since

the probability is greater than the significance level of 0.05, the null hypothesis of normality is accepted. Precisely, the test statistic is less than the critical value of a quantile of order 0.95 in a chi-squared distribution with seven degrees of freedom:

$$P\left(\chi_7^2 \geq 3.2448\right) = 0.8615 > \alpha = 0.05$$

$$SC = 3.245 < {}_{0.95}\chi_7^2 = 14.0671$$

The type II error (β) or probability of maintaining the null hypothesis conditional on the alternative hypothesis being true ($SS >$ critical value $= {}_{0.95}\chi_7^2 = 14.07$) is greater than 0.50. In addition, the power (ϕ) or the probability of rejecting the null hypothesis conditional on the alternative hypothesis being true ($SS >$ critical value) is close to 0.20. Consequently, the correct decision is to hold the null hypothesis. To calculate these probabilities, the cumulative distribution function of a non-central chi-squared distribution with seven degrees of freedom (df) and a Non-Centrality Parameter (NCP) located at the value of the SS test statistic is used. This distribution function is evaluated at the critical value: ${}_{0.95}\chi_7^2 = 14.07$ (true alternative hypothesis).

$$\beta = \chi_{df=7,NCP=SS}^2\left({}_{1-\alpha}\chi_{df}^2\right) = \chi_{7,3.245}^2\left({}_{0.95}\chi_7^2 = 14.0671\right) = 0.7932$$

$$\phi = 1 - \beta = 1 - \chi_{df=7,NCP=SS}^2\left({}_{0.95}\chi_7^2\right) = 1 - 0.7932 = 0.2068$$

1.4. Objectives

Once the new omnibus test of normality has been proposed, based on the seven-number parametric summary (PSNS), the objective of the study is to compare it with the W test [35] and the K-squared test [44], which are the most powerful tests for normality in simulation studies [52]. The comparison is made both in the precision to maintain the null hypothesis of normality against normal samples and to reject it against non-normal samples and in the type II error level in the first case and the statistical power in the second case. It should be noted that in a recent study performed by Khatun [53], the Shapiro and Wilk test was the most powerful for testing normality with nine different sample sizes (10, 20, 25, 30, 40, 50, 100, 200, and 300), followed by the Shapiro-Francia and Anderson-Darling tests. The least powerful test was that of Jarque and Bera. However, the present study does not use the version of Jarque and Bera [42], but that of D'Agostino *et al.* [44], which is a more powerful variety of moment-based normality tests [54] [55].

In applied statistics in the field of social sciences and other sciences, it is common to find small samples of less than 30 observations and to want to use parametric tests, for which some distributional assumptions, usually of normality, are required to be met. In case of non-compliance, non-parametric tests should be chosen [56]. Thus, having an accurate and powerful normality test for small samples is important [53]. However, the proposed PSNS-based normality test, built on the asymptotic normality of the sample quantile distribution, is not

expected to be useful for this purpose, showing low precision and power with small samples.

2. Materials and Methods

From a standard continuous uniform distribution U (0, 1), 24 samples of different sizes were drawn (n = 20, 30, 40, 50, 60, 70, 80, 90 100, 125, 150, 175, 200, 250, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, and 2000) to generate by the inverse transform sampling 16 continuous distributions: one standard normal and 15 non-normal [57].

- Standard normal distribution with location parameter μ of 0 and squared scale parameter σ^2 of 1: $N(\mu = 0, \sigma^2 = 1)$. This distribution is unimodal ($Mo = 0$), symmetric (Pearson's coefficient of asymmetry based on the standardized third central moment: $\beta_1 = \mu_3/\mu_2^{3/2} = 0$), mesokurtic (Pearson's excess kurtosis [B_2] that is the standardized fourth central moment [$\beta_2 = \mu_4/\mu_2^2$] minus the value of the standardized fourth central moment in a standard normal distribution: $B_2 = \beta_2 - 3 = 0$), has finite moments, and its domain corresponds to the interval $(-\infty, +\infty)$.

$$x_i \in X \sim N(0,1) \text{ and } u_i \in U \sim U(0,1)$$

$$x_i = \Phi^{-1}(u_i)$$

Φ^{-1} = quantile function or inverse cumulative distribution function of a standard normal distribution, also called a Probit function.

Non-normal distributions:

- Standard logistic distribution with location parameter μ of 0 and scale parameter s of 1: Logistic ($\mu = 0$, $s = 1$). This distribution is unimodal ($Mo = 0$), symmetric ($\beta_1 = 0$) and leptokurtic ($B_2 = 1.2$), has finite moments, and its domain corresponds to the interval $(-\infty, + \infty)$.

$$x_i \in X \sim \text{Logistic}(0,1) \text{ and } u_i \in U \sim U(0,1)$$

$$x_i = \ln\left(\frac{u_i}{1-u_i}\right)$$

- Standard Laplace distribution with location parameter μ of 0 and scale b of 1: Laplace ($\mu = 0$, $b = 1$). This distribution is unimodal ($Mo = 0$), symmetric ($\beta_1 = 0$) and leptokurtic ($B_2 = 3$), has finite moments, and its domain corresponds to the interval $(-\infty, +\infty)$.

$$x_i \in X \sim \text{Laplace}(0,1) \text{ and } u_i \in U \sim U(0,1)$$

$$x_i = -1 \times \text{sign}\left(u_i - \frac{1}{2}\right) \times \ln\left(1 - 2 \times \left|p - \frac{1}{2}\right|\right)$$

- Standard Student's t distribution with five degrees of freedom: $t(\nu = 5)$. This distribution is unimodal ($Mo = 0$), symmetric ($\beta_1 = 0$) and leptokurtic ($B_2 = 6$), at least its first four moments are finite, and its domain corresponds to the interval $(-\infty, +\infty)$.

$$x_i \in X \sim t(5) \quad \text{and} \quad u_i \in U \sim U(0,1)$$

$$x_i = F_X^{-1}(u_i)$$

F_X^{-1} = quantile function or inverse cumulative distribution function of X.

- Standard Cauchy distribution with location parameter x_0 of 0 and scale parameter γ of 1: Cauchy ($x_0 = 0$, $\gamma = 1$). This distribution is a unimodal ($Mo = 0$), symmetric (Bowley's Quartile Coefficient of Skewness: $QCS = (P_{25} + P_{75} - 2 \times P_{50})/(P_{75} - P_{25}) = 0$) and leptokurtic (Kelley's Percentile Coefficient of Kurtosis: $PCK = (P_{75} - P_{25})/[2 \times (P_{90} - P_{10})] = 0.162$; corrected, or centered at 0 based on the expected value for a normal distribution PCK: $CPCK$) = $CPK - 0.263 = -0.101$), has no finite moments, and its domain corresponds to the interval $(-\infty, +\infty)$.

$$x_i \in X \sim \text{Cauchy}(0,1) \quad \text{and} \quad u_i \in U \sim U(0,1)$$

$$x_i = \tan\left[\pi \times \left(u_i - \frac{1}{2}\right)\right]$$

- Generalized beta distribution with shape parameters α and β of 1.5 and threshold parameters -2 as minimum value a and 2 as maximum value b or Wigner's semicircle distribution with radius r of 2: Beta ($\alpha = 1.5$, $\beta = 1.5$, $a = -2$, $b = 2$) $\equiv$ Semicircle ($r = 2$). This distribution is unimodal ($Mo = 0$), symmetric ($\beta_1 = 0$) and platykurtic ($B_2 = -1$), has finite moments, and its domain corresponds to the interval $[-2, 2]$.

$$x_i \in X \sim \text{Beta}(1.5, 1.5, -2, 2) \equiv \text{Semicircle}(r = 2) \quad \text{and} \quad u_i \in U \sim U(0,1)$$

$$x_i = F_X^{-1}(u_i)$$

F_X^{-1} = quantile function or inverse cumulative distribution function of X.

- Uniform distribution with threshold parameters -2 as minimum value a and 2 as maximum value b: $U(a = -2, b = 2)$. This distribution has no mode, but does have finite moments, is symmetric ($\beta_1 = 0$), platykurtic ($B_2 = -1.2$), and its domain corresponds to the interval $[-2, 2]$.

$$x_i \in X \sim U(-2, 2) \quad \text{and} \quad u_i \in U \sim U(0,1)$$

$$x_i = -2 + 4 \times u_i$$

- Generalized arcsine distribution with threshold parameters -2 as minimum value a and 2 as maximum value b: Arcsine ($a = -2$, $b = 2$). This distribution is bimodal ($Mo_1 = -2$ and $Mo_2 = 2$), symmetric ($\beta_1 = 0$) and platykurtic ($B_2 = -1.5$), has finite moments, and its domain corresponds to the interval $[-2, 2]$.

$$x_i \in X \sim \text{Arcsine}(-2, 2) \quad \text{and} \quad u_i \in U \sim U(0,1)$$

$$x_i = -2 + 4 \times \sin^2\left(\frac{\pi}{2}u_i\right)$$

The angles of u_i are measured in units of radians.

- Triangular distribution with threshold parameters -2 as minimum value a and 2 as maximum value c and with modal value or location parameter b of

2: Triangular ($a = -2$, $b = 2$, $c = 2$). This distribution is unimodal ($Mo = 2$), negative asymmetric ($\beta_1 = -1$) and platykurtic ($B_2 = -0.6$), has finite moments, and its domain corresponds to the interval $[-2, 2]$.

$$x_i \in X \sim \text{Triangular}(-2, 2, 2) \quad \text{and} \quad u_i \in U \sim U(0,1)$$

$$x_i = -2 + 4 \times \sqrt{u_i}$$

- Standard Fisher's Z distribution with shape parameters ν_1 of 3 and ν_2 of 9: Z ($\nu_1 = 3$, $\nu_2 = 9$). This distribution is unimodal ($Mo = 0$), negative asymmetric ($\beta_1 = -0.60$) and leptokurtic ($B_2 = 1.03$), has finite moments, and its domain corresponds to the interval $(-\infty, +\infty)$.

$$x_i \in X \sim Z(3,9) \quad \text{and} \quad u_i \in U \sim U(0,1)$$

$$x_i = 0.5 \times \ln\left(F_Y^{-1}(u_i)\right); Y \sim F(\nu_1 = 3, \nu_2 = 9)$$

F_Y^{-1} = quantile function or inverse cumulative distribution function of Y.

- Weilbull distribution with shape parameter α of 2 (increasing failure rate) and scale parameter β of 2: Weibull ($\alpha = 2$, $\beta = 2$). This distribution is unimodal ($Mo = \sqrt{2}$), negative asymmetric ($\beta_1 = -2.09$) and leptokurtic ($B_2 = 1.53$), has finite moments, and its domain corresponds to the interval $[0, +\infty)$.

$$x_i \in X \sim \text{Weibull}(2,2) \quad \text{and} \quad u_i \in U \sim U(0,1)$$

$$x_i = 2 \times \sqrt{-\ln(1 - u_i)}$$

- Program Evaluation and Review Technique (PERT) distribution with threshold position parameter −2 as minimum value a and 4 as maximum value c, modal value or location parameter b of 0 and scale parameter λ of 4: PERT ($a = -2$, $b = 0$, $c = 4$, $\lambda = 4$), which is equivalent to a beta distribution with four parameters: Beta ($\alpha = 7/3$, $\beta = 11/3$, $a = -2$, $b = 4$). This distribution is unimodal ($Mo = 0$), positive asymmetric ($\beta_1 = 0.302$) and platykurtic ($B_2 = -6/11$), has finite moments, and its domain corresponds to the interval $[-2, 4]$.

$$x_i \in X \sim \text{PERT}(a = -2, b = 0, c = 4, \lambda = 4)$$

$$\equiv \text{Beta}(\alpha = 7/3, \beta = 11/3, a = -2, b = 4)$$

$$y \quad u_i \in U \sim U(0,1)$$

$$x_i = F_X^{-1}(u_i)$$

F_X^{-1} = quantile function or inverse cumulative distribution function of X.

- Standard Rayleigh distribution with scale parameter σ of 1: Rayleigh ($\sigma = 1$). This distribution is unimodal ($Mo = 1$), positive asymmetric ($\beta_1 = 0.63$) and leptokurtic ($B_2 = 0.25$), has finite moments, and its domain corresponds to the interval $[0, +\infty)$.

$$x_i \in X \sim \text{Rayleigh}(1) \quad \text{and} \quad u_i \in U \sim U(0,1)$$

$$x_i = \sqrt{-2 \times \ln(1 - u_i)}$$

- Chi-squared distribution with four degrees of freedom (shape parameter): χ^2 ($v = 4$). This distribution is unimodal ($Mo = 2$), positive asymmetric ($\beta_1 = \sqrt{2}$) and leptokurtic ($B_2 = 3$), has finite moments, and its domain corresponds to the interval $[0, +\infty)$.

$$x_i \in X \sim \chi^2(4) \quad \text{and} \quad u_i \in U \sim U(0,1)$$

$$x_i = F_X^{-1}(u_i)$$

F_X^{-1} = quantile function or inverse cumulative distribution function of X.

- Log-normal distribution, also known as Galton's distribution, with location parameter μ of 0 and squared scale parameter σ^2 of a quarter: Lognormal ($\mu = 0$, $\sigma^2 = 0.25$). This distribution is unimodal ($Mo = 0.78$), positive asymmetric ($\beta_1 = 1.75$) and leptokurtic ($B_2 = 5.90$), has finite moments, and its domain corresponds to the interval $(0, +\infty)$.

$$x_i \in X \sim \text{Lognormal}(0, 0.25) \quad \text{and} \quad u_i \in U \sim U(0,1)$$

$$x_i = e^{\frac{\Phi^{-1}(u_i)}{2}}$$

Φ^{-1} = quantile function or inverse cumulative distribution function of a standard normal distribution, also called Probit function.

- F distribution with shape parameters v_1 and v_2 of 9: $F(v_1 = 9$ and $v_2 = 9)$. This distribution is unimodal ($Mo = 7/11$), positive asymmetric ($\beta_1 = 4.39$) and leptokurtic ($B_2 = 98.81$), has finite moments, and its domain corresponds to the interval $[0, +\infty)$.

$$x_i \in X \sim F(v_1 = 9, v_2 = 9) \quad \text{and} \quad u_i \in U \sim U(0,1)$$

$$x_i = F_X^{-1}(u_i)$$

F_X^{-1} = quantile function or inverse cumulative distribution function of X.

For each of the 368 samples (23 sample sizes by 16 distributions), the null hypothesis of normality was tested using the PSNS-based omnibus test, as well as the W-test [35] and the K-squared test [44]. Once the probability value (p) has been calculated, it is decided whether or not the null hypothesis is maintained at a significance level of 5% (if $p \geq 0.05$, the null hypothesis of normality is accepted; if $p < 0.05$, is rejected).

To verify the accuracy of the PSNS-based test when contrasting the null hypothesis of normality, a dichotomous qualitative variable named *correct decision* (A) was created based on the expectation (maintain the null hypothesis of normality with the normal distribution and reject it with the 15 other distributions): A = {0 = no, 1 = yes}. The accuracy or probability of being correct when maintaining the null hypothesis in case of normality and rejecting it in case of non-normality was calculated for each of the three statistical tests for normality (PSNS-based, W, and K-squared) through the proportion of the number of correct decisions and number of trials. The equivalence of this proportion of correct decisions among the three tests ($k = 3$ tests) was tested using the Q test of Cochran [58]. The comparison of proportions was carried out both for each of the 24

sample sizes (n = 16 distributions) and for each of the 16 distributions (n = 24 sample sizes). The effect size was estimated by the eta-squared: $\hat{\eta}_Q^2 = Q/(n(k-1))$, where $Q \sim \chi_{k-1}^2$ [59]. This calculation through a quotient is analogous to that used to calculate Kendall's concordance coefficient W, with the test statistic following a chi-squared distribution in the numerator and the product between the sample size and the degrees of freedom in the denominator: $W = \chi_r^2/(n(k-1))$, where $\chi_r^2 \sim \chi_{k-1}^2$ [60] [61]. It is worth noting that there is a correspondence between Kendall's W and the average of Spearman's rank correlation coefficient (r_S): $W = (\overline{r}_S(k-1)+1)/k$ [62]. This opens the possibility of interpreting the effect size in Cochran's Q test from the cut-off points established for the association strength measured by Spearman's coefficient [63]. When k = 3, $\overline{r}_S < 0.10$ $\rightarrow$ $W \equiv \hat{\eta}_Q^2 < 0.4$ show very small agreement, $\overline{r}_S \in [0.10, 0.30)$ $\rightarrow$ $W \equiv \hat{\eta}_Q^2 \in [0.40, 0.53)$ small, $\overline{r}_S \in [0.30, 0.50)$ $\rightarrow$ $W \equiv \hat{\eta}_Q^2 \in [0.53, 0.67)$ medium, $\overline{r}_S \in [0.50, 0.70)$ $\rightarrow$ $W \equiv \hat{\eta}_Q^2 \in [0.67, 0.80)$ large, and $\overline{r}_S \geq 0.70$ $\rightarrow$ $W \equiv \hat{\eta}_Q^2 \geq 0.80$ very large. Pairwise comparisons were made using McNemar's exact (binomial) test [64]. Benjamini and Yekutieli's correction was applied to control the rate of false-positive discoveries with correlated or paired data [65]. With this correction, once the three probability values (p_i-value) are ordered in ascending order, p_1-value < p_2-value < p_3-value, the paired differences are significant with values less than 0.009, 0.018 and 0.027, respectively.

On the other hand, the statistical power of the three statistical tests for normality (k = 3) was calculated and their power averages were compared to the 24 sample sizes (n = 24) and the 15 non-normal distributions (n = 15). The type II error or conservative error (β) is the probability of maintaining the null hypothesis conditioned to the fact that the alternative hypothesis (corresponding to the value of test statistic) is true, $\beta = P$ (H$_0$ is accepted|H$_1$ is true). Its complement provides the statistical power (ϕ) or probability of rejecting the null hypothesis conditional on the alternative hypothesis being true $\phi = P$ (H$_0$ is rejected|H$_1$ is true). In the PSNS-based test, the type II error was calculated through the cumulative distribution function of a non-central chi-squared distribution with seven degrees of freedom and the SS statistic as a non-centrality parameter (NCP), evaluated at the critical value or quantile of order 0.95 for a chi-squared distribution with seven degrees of freedom, so the power is obtained using the following formula:

$$\phi = 1 - \chi_{df=7,NCP=SS}^2\left(_{0.95}\chi_7^2\right)$$

In the K-squared test [44], type II error was calculated through cumulative distribution function of a non-central chi-squared distribution with two degrees of freedom and the K-squared statistic as a non-centrality parameter, evaluated in the critical value or quantile of order 0.95 for a chi-squared distribution with two degrees of freedom, so the power is obtained using the following formula:

$$\phi = 1 - \chi_{df=2,NCP=K^2}^2\left(_{0.95}\chi_2^2\right)$$

In the W test [35], type II error was calculated as in the one-sample Z test for

a mean (alternative hypothesis to the right tail):

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu = \mu_1$$

$$\mu_1 > \mu_0$$

$$\beta = P\left(Z \leq \frac{\overline{x}_{Crit} - \mu_1}{\sigma_\mu} \right) = P\left[Z \leq \frac{\left(\mu_0 + Z_{1-\frac{\alpha}{2}} \times \sigma_\mu \right) - \mu_1}{\sigma_\mu} \right]$$

$$Z_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2} \right) = \text{quantile of order } 1 - (\alpha/2) \text{ in a standard normal distri-}$$

bution.

Applied to the Royston's test, which applies a logarithmic transformation to the W statistic so that it follows a standard normal distribution, the power would be obtained by the following formula:

$$\phi = 1 - P\left[Z \leq \frac{\left(\mu_Y + Z_{1-\frac{\alpha}{2}} \times \sigma_Y \right) - \ln(1-W)}{\sigma_Y} \right]$$

Given a random sample of size n.

$$\mu_Y = 0.0038915 \times \ln^3(n) - 0.083751 \times \ln^2(n) - 0.31082 \times \ln(n) - 1.5861$$

$$\sigma_Y = e^{0.0030302 \times \ln^2(n) - 0.082676 \times \ln(n) - 0.4803}$$

$$Z_Y = \frac{Y - \mu_Y}{\sigma_Y} = \frac{\ln(1-W) - \mu_Y}{\sigma_Y} \sim N(0,1)$$

$W = r^2_{x_{(i)}a_i}$ = the square of the correlation between the empirical quantiles or values ordered in ascending order (x_i) and the means or expected values of the corresponding order statistics for a normal distribution, these expected values being standardized and normalized (a_i).

In Royston's W test, the decision is unilateral to the right tail, so when $P(Z \geq Z_Y) \geq \alpha$, the null hypothesis of normality is maintained [35].

If the normality assumption was fulfilled, the power averages in the 15 non-normal distributions ($n = 15$) among the three statistical tests for normality ($k = 3$) were compared for each of the 24 samples using the repeated measures analysis of variance (ANOVA). This type of ANOVA is an omnibus parametric technique for comparing repeated means based on the decomposition of the total variance and the assumptions of normal distribution in the variables compared and homogeneity of variance in the differences between variables. The first distributional assumption was tested using the original W test [34]. The assumption of sphericity or equivalence of the variances of the differences among the three statistical tests for normality was tested using the test of John [66], Na-

gao [67], and Sugiura [68]. In case of non-compliance with the sphericity, a correction was applied to the degrees of freedom. The degrees of freedom were multiplied by an epsilon correction factor. When the Greenhouse-Geisser epsilon factor [69] was less than 0.75, it was used as a correction, otherwise, the Huynh-Feldt epsilon factor [70] was used. The effect size was estimated by the partial eta squared and was interpreted from the cut-off points suggested by Cohen [71] for ANOVA: <0.01 trivial, [0.01, 0.06) small, [0.06, 0.14) medium and ≥ 0.14 large. Pairwise comparisons were made using Student's paired-samples t-test as the sphericity assumption was not met in any case [72]. The Benjamini-Yekutieli correction was applied to control for the rate of false-positive discoveries with correlated, or paired data [65].

Faced with the breach of the normality assumption, the differences among the means were tested using Friedman's test [60] [73] [74] with the modification of Agresti and Pendergast [75]. The effect size was estimated by Kendall's W from Friedman's original test statistic: $W = \chi_r^2 / (n(k-1))$ [60] [61], and it was interpreted from its correspondence with the average of the Spearman's rank correlation coefficient [62] and the cut-off points to interpret the strength of association of Hopkins [63]. When $k = 3$, values of $W < 0.4$ show a very small concordance, in the interval [0.40, 0.53) small, in the interval [0.53, 0.67) medium, in the interval [0.67, 0.80) large, and ≥0.80 very large. Pairwise comparisons were made using Wilcoxon's signed-rank test [76] with the calculation of exact probability and the correction of Benjamini and Yekutieli [65] was applied to control the rate of false-positive discoveries.

The significance level was set at 5% in the omnibus and normality tests. Calculations were performed using Excel 2019 [77] and Real Statistics Resource Pack for Excel [78].

3. Results

3.1. Comparison of Accuracy by Sample Size and Type of Distribution

Table 3 shows the omnibus comparison of the proportions of correct decision (accuracy) when testing the null hypothesis of normality among the PSNS-based, W, and K-squared tests for each of the 24 sample sizes (the 16 distributions grouped) using Cochran's Q test, as well as pairwise comparisons using McNemar's exact test. Table 4 shows this same information for each of the 16 probability distributions (the 24 sample sizes grouped).

The proportion of correct decisions was equivalent in 18 of the 24 sample sizes (75%), which corresponds to the samples of 20, 60, 70, 90, 100, and 150, as well as to the samples of 200 or more data. With 20 sample data, the three tests had the same accuracy, which was very low (0.19). All three detected normality and were able to reject normality with the Cauchy and Laplace distributions. However, they failed to reject the null hypothesis of normality with all other distributions, even with the F (9, 9) and arcsine (−2, 2) distributions that are highly

Table 3. Comparison of the proportions of correct decision when contrasting normality by sample size.

n	Accuracy			Cochran's test		Effect size	McNemar's exact test	
	PSNS	W	K^2	Q	p-value	$\hat{\eta}_Q^2$	Pair	p-value
20	0.188	0.188	0.188	-	-	-		
30	0.313	0.563	0.500	6.5	0.039*	0.203	PSNS – W	0.063
							PSNS – K^2	0.125
							W – K^2	1
40	0.313	0.375	0.500	4.667	0.097	0.146		
50	0.250	0.500	0.563	8.400	0.015*	0.263	PSNS – W	0.063
							PSNS – K^2	0.125
							W – K^2	1
60	0.563	0.813	0.813	5.333	0.069	0.167		
70	0.438	0.563	0.625	2	0.368	0.063		
80	0.438	0.625	0.688	6.5	0.039*	0.203	PSNS – K^2	0.125
							PSNS – W	0.25
							W – K^2	1
90	0.438	0.563	0.563	4	0.135	0.125		
100	0.500	0.688	0.563	4.667	0.097	0.146		
125	0.688	0.875	0.875	6	<0.050*	0.188	PSNS – W	0.25
							PSNS – K^2	0.25
							W – K^2	1
150	0.813	0.938	0.875	3	0.223	0.094		
175	0.750	1	0.875	6	<0.050*	0.188	PSNS – W	0.125
							PSNS – K^2	0.5
							W – K^2	0.5
200	0.813	1	0.938	4.667	0.097	0.146		
250	0.813	0.875	0.938	3	0.223	0.094		
300	0.875	1	1	4	0.135	0.125		
400	0.813	1	0.938	4.667	0.097	0.146		
500	0.938	1	1	2	0.368	0.063		
600	0.938	1	1	2	0.368	0.063		
700	0.938	1	1	2	0.368	0.063		
800	0.938	1	1	2	0.368	0.063		
900	1	1	1	-	-	-		

Continued

1000	1	1	1	-	-	-		
1500	1	1	1	-	-	-		
2000	1	1	1	-	-	-		
Total	0.698	0.815	0.810	68.035	<0.001*	0.096	PSNS − K²	0.008*
							PSNS − W	0.016*
							W − K²	1

Note. n = size of the random sample for 16 different distributions, Total = the 24 samples with different sample sizes are grouped, Accuracy = proportion of correct decision when contrasting normality by the PSNS-based, Royston's W, and D'Agostino *et al.*'s K-squared tests, Q = statistic of Cochran's Q test, p value = right-tailed asymptotic probability with an asterisk in case of significant difference at a significance level of 0.05, $\hat{\eta}_Q^2$ = eta squared coefficient of Serlin *et al.* [59] as a measure of effect size. Pairwise comparisons using McNemar's exact test: Pair = compared pairs ordered in ascending order by their p values, p-value = two-tailed exact probability by the binomial distribution with an asterisk in case of significant difference at a significance level of 0.05 corrected by the Benjamini-Yekutieli procedure for false-positive discovery rate in correlated samples (0.009 for the first or lowest p-value, 0.018 for the second, and 0.027 for the third), and indicated in italics if the difference is significant at a significance level of 0.05 without correction.

Table 4. Comparison of the proportions of correct decision when contrasting normality by type of distribution.

Distribution	Accuracy			Cochran's test		Effect size	McNemar's exact test	
	PSNS	*W*	*K²*	*Q*	*p*-value	$\hat{\eta}_Q^2$	Pair	*p*-value
W	0.500	0.667	0.625	5.2	0.074	0.108		
T	0.750	0.875	0.833	4.667	0.097	0.097		
ZF	0.667	0.750	0.708	3	0.223	0.063		
Sin⁻¹	0.875	0.958	0.958	4	0.135	0.083		
U	0.708	0.833	0.833	6	<0.050*	0.125	PSNS − W	0.25
							PSNS − K²	0.25
							W − K²	1
SC	0.500	0.792	0.833	14.25	0.001*	0.297	PSNS − K²	0.008*
							PSNS − W	0.016*
							W − K²	1
N (0, 1)	1	1	1					
Logistic	0.250	0.583	0.542	14.25	0.001*	0.297	PSNS − W	0.008*
							PSNS − K²	0.016*
							W − K²	1
Laplace	0.750	0.792	0.833	3	0.223	0.063		
Student	0.583	0.708	0.792	7.6	0.022*	0.158	PSNS − K²	0.063

Continued

							W – PSNS	0.25
							W – K²	0.5
Cauchy	1	1	1					
PERT	0.458	0.542	0.625	4.8	0.091	0.100		
Rayleigh	0.542	0.833	0.708	10.571	0.005*	0.220	PSNS – W	0.016
							PSNS – K²	0.125
							W – K²	0.25
χ^2	0.833	0.875	0.875	0.667	0.717	0.014		
LogN	0.833	0.917	0.875	3	0.223	0.063		
F	0.917	0.917	0.917					

Note. Distribution = type of distribution in the 24 random samples with different sizes: W = Weibull (α = 2, β = 2), T = Triangular (a = −2, b = 2, c = 2), ZF = Fisher's Z (ν_1 = 3, ν_2 = 9), Sin^{-1} = Arcsine (a = −2, b = 2), U = Uniform (a = −2, b = 2), SC = Semicircle (r = 2), N (μ = 0, σ^2 = 1), Logistic (μ = 0, s = 1), Laplace (μ = 0, b = 1), Student's t (ν = 5), Cauchy (x_0 = 0, γ = 1), Rayleigh (σ = 1), χ^2 = Chi-squared (ν = 4), LogN = Lognormal (μ = 0, σ^2 = 0.25), F (ν_1 = 9, ν_2 = 9). Accuracy = proportion of correct decision when contrasting normality by the PSNS-based, Royston's W, and D'Agostino *et al.*'s K-squared test, Q = statistic of Cochran's Q test, p-value = right-tailed asymptotic probability with an asterisk in case of significant difference at a significance level of 0.05, $\hat{\eta}_Q^2$ = eta squared coefficient of Serlin *et al.* [59] as a measure of effect size. Pairwise comparisons using McNemar's exact test: Pair = compared pairs ordered in ascending order by their probability values, p-value = two-tailed exact probability by the binomial distribution with an asterisk in case of significant difference at a significance level of 0.05 corrected by the Benjamini-Yekutieli procedure for false-positive discovery rate in correlated samples (0.009 for the first or lowest p-value, 0.018 for the second, and 0.027 for the third), and indicated in italics if the difference is significant at a significance level of 0.05 without correction.

deviated from normality. With a sample of 125, the W and K-squared tests achieved high accuracy (≥0.80), and with 150 or more, all three tests accomplished this high level. Accuracy was perfect in all three tests from a sample size of 900. Differences by Cochran's omnibus Q test appeared with sample sizes of 30, 40, 50, 80, 125, and 175. In these six comparisons, the effect size ($\hat{\eta}_Q^2$) was less than 0.27, which corresponds to a very small effect size when interpreted from the analogy of this coefficient with Kendall's coefficient of agreement W and the relation of Kendall's W with the Spearman's rank correlation coefficient. Thus, when making pairwise comparisons, no difference was significant with or without Benjamini-Yekutieli correction for the false-positive discovery rate. When grouping the 24 sample sizes and 16 distributions (total), the proportion of correct decisions for the PSNS-based test was 0.70, for the W test 0.82, and the K-squared test 0.81. The difference among these proportions was significant by the omnibus test. The accuracy of PSNS-based test was significantly lower than that of K-squared and W tests when these last two tests had a statistically equivalent accuracy. See Table 3.

By distributions, the three tests discriminated the Cauchy distribution and detected the normal distribution with the 24 sample sizes, that is, its accuracy was 1. With no statistically significant difference between the three tests, the accuracy was greater than 0.90 with the F distribution; greater than 0.80 with the arcsine distribution, chi-squared distribution with four degrees of freedom, and lognormal distribution; and greater than 0.70 with the triangular distribution. The lowest efficacy appeared with the logistic distribution that closely resembled the normal distribution. With this latter distribution, the PSNS-based test has lower efficacy than the other two tests. See Table 4.

The accuracy of the three tests was equivalent in 11 of the 16 (68.75%) distributions. There was a significant difference by omnibus Cochran's Q test in the Wigner's semicircle, logistic, Student's t, Rayleigh, and uniform distributions. However, the effect size was very small ($\hat{\eta}_Q^2 < 0.30$). When making pairwise comparisons with the Benjamini-Yekutieli correction, the PSNS-based test was less accurate than the W and K-squared tests when rejecting non-normality in the Wigner's semicircle and logistic distributions. Without the correction, its efficiency would also be significantly lower than that of the W test in the Rayleigh distribution. The pairwise differences, even without correction, were not significant in the uniform and Student's t distributions. See Table 4.

3.2. Comparison of Potency by Sample Size and Type of Distribution

As a preliminary step to comparing the central tendency of power when contrasting the null hypothesis of normality using the PSNS-based, W, and K-squared tests, it was verified whether the power distributions follow the normal probability model. If the distributional assumption was accepted, a parametric test, such as the repeated measures ANOVA, was used to compare the central tendency; if the assumption was rejected, a non-parametric test was used, such as Friedman's test.

As can be seen in Table 5, the hypothesis of normality in the average power distribution of the three statistical tests for normality was maintained with sample sizes of 30, 50, 60, 70, 90, and 100 (when grouping the 15 non-normal distributions). With none of the 14 non-normal distributions (when grouping the 24 sample sizes), this hypothesis was accepted. This assumption was also not fulfilled with the normal distribution (when grouping the 24 sample sizes), although in this case the average type II error distribution of the PSNS-based, W, and K-squared tests was tested.

Table 6 shows the comparisons among the average power (in 15 non-normal distributions) of the PSNS-based, W, and K-squared tests for 18 different sample sizes without assuming normality, since this distributional assumption was rejected. Table 7 shows these comparisons for the six sample sizes in which the normality assumption was held. Finally, in Table 8, there are comparisons of the average power (in the 24 sample sizes) of the PSNS-based, W, and K-squared tests for the 15 types of non-normal distribution without assuming normality.

Table 5. Testing for normality through the Shapiro and Wilk's test in the distribution of average power by sample size and type of distribution.

n	PSNS		W		K²		Dist.	PSNS		W		K²	
	SW	p	SW	p	W	p		SW	p	SW	p	SW	p
20	0.73	0.003	0.80	0.018	0.86	0.086	W	0.84	0.001	0.76	<0.001	0.80	<0.001
30	**0.86**	**0.104**	**0.94**	**0.551**	**0.84**	**0.056**	T	0.70	<0.001	0.58	<0.001	0.80	<0.001
40	0.80	0.022	0.82	0.033	0.88	0.140	ZF	0.81	<0.001	0.70	<0.001	0.71	<0.001
50	**0.89**	**0.187**	**0.86**	**0.094**	**0.87**	**0.124**	Sin⁻¹	0.60	<0.001	0.36	<0.001	0.32	<0.001
60	**0.85**	**0.078**	**0.90**	**0.232**	**0.85**	**0.082**	U	0.76	<0.001	0.61	<0.001	0.50	<0.001
70	**0.90**	**0.248**	**0.86**	**0.090**	**0.87**	**0.136**	SC	0.83	0.001	0.76	<0.001	0.64	<0.001
80	0.85	0.082	0.83	0.041	0.92	0.373	N	0.84	0.001	0.81	<0.001	0.81	0.001
90	**0.86**	**0.087**	**0.84**	**0.055**	**0.85**	**0.078**	Log	0.91	0.043	0.82	0.001	0.87	0.004
100	**0.86**	**0.097**	**0.85**	**0.066**	**0.86**	**0.086**	Laplace	0.74	<0.001	0.68	<0.001	0.67	<0.001
125	0.81	0.027	0.36	<0.001	0.75	0.005	Student	0.84	0.001	0.66	<0.001	0.68	<0.001
150	0.67	0.001	0.67	0.001	0.68	0.001	Cauchy	0.21	<0.001	0.21	<0.001	0.22	<0.001
175	0.82	0.035	0.59	<0.001	0.61	<0.001	PERT	0.86	0.003	0.83	0.001	0.85	0.002
200	0.77	0.010	0.58	<0.001	0.55	<0.001	Ray	0.81	<0.001	0.72	<0.001	0.79	<0.001
250	0.70	0.002	0.55	<0.001	0.59	<0.001	χ²	0.63	<0.001	0.54	<0.001	0.62	<0.001
300	0.84	0.059	0.59	<0.001	0.55	<0.001	LogN	0.61	<0.001	0.48	<0.001	0.51	<0.001
400	0.67	0.001	0.35	<0.001	0.44	<0.001	F	0.42	<0.001	0.44	<0.001	0.48	<0.001
500	0.58	<0.001	0.31	<0.001	0.31	<0.001							
600	0.55	<0.001	0.29	<0.001	0.32	<0.001							
700	0.46	<0.001	0.29	<0.001	0.30	<0.001							
800	0.44	<0.001	0.29	<0.001	0.32	<0.001							
900	0.40	<0.001	0.29	<0.001	0.29	<0.001							
1000	0.49	<0.001	0.29	<0.001	0.28	<0.001							
1500	0.37	<0.001	0.36	<0.001	0.30	<0.001							
2000	0.62	<0.001	0.46	<0.001	0.30	<0.001							
Total	0.76	<0.001	0.65	<0.001	0.67	<0.001							

Note. n = size of the random sample in each of the 15 non-normal distributions that provide the power data when testing normality, either by the PSNS-based, Royton's W, or D'Agistino *et al.*'s K-squared tests, SW = statistic of the original test of Shapiro and Wilk [35] and p value = one-tailed exact probability of the Shapiro-Wilk test, Distr. = type of random distribution in each of the 24 sample sizes that provide the power data, except for error II in the standard normal distribution $N(0, 1)$, when normality is tested, either by PSNS-based, W, or K-squared tests. The distributions appear in ascending order by degree of skewness first and excess kurtosis second: W = Weibull ($a = 2$, $\beta = 2$), T = Triangular ($a = -2$, $b = 2$, $c = 2$), ZF = Fisher's Z ($v_1 = 3$, $v_2 = 9$), Sin⁻¹ = Arcsine ($a = -2$, $b = 2$), U = Uniform ($a = -2$, $b = 2$), SC = Semicircle ($r = 2$), N ($\mu = 0$, $\sigma^2 = 1$), Log = Logistic ($\mu = 0$, $s = 1$), Laplace ($\mu = 0$, $b = 1$), Student's t ($v = 5$), Cauchy ($x_0 = 0$, $\gamma = 1$), PERT ($a = -2$, $b = 0$, $c = 4$, $\lambda = 4$), Ray = Rayleigh ($\sigma = 1$), χ^2 = Chi-squared ($v = 4$), LogN = Lognormal ($\mu = 0$, $\sigma^2 = 0.25$), F ($v_1 = 9$, $v_2 = 9$), Total = the 15 samples with different types of non-normal distribution are grouped. It is marked in bold when the null hypothesis of normality is maintained in the three distributions of average potency.

Table 6. Comparison of the average power by sample size without assuming normality.

n	Average power			Friedman's test		$W \to \overline{r}_s$	Wilcoxon's exact test		
	PSNS	W	K^2	F	p-value		Pair	$|MD|$	p-value
						0.431	$W - K^2$	0.100	<0.001*
20	0.283	0.243	0.344	100.609	0.001*	↓	$PSNS - K^2$	0.061	0.107
						0.147	$PSNS - W$	0.045	0.188
						0.658	$W - K^2$	0.193	<0.001*
40	0.385	0.392	0.585	260.909	<0.001*	↓	$PSNS - K^2$	0.200	<0.001*
						0.487	$PSNS - W$	0.007	0.847
						0.298	$PSNS - K^2$	0.166	0.007*
80	0.495	0.605	0.661	50.937	0.007*	↓	$PSNS - W$	0.110	0.015*
						0.053	$W - K^2$	0.056	0.359
						0.622	$PSNS - W$	0.216	<0.001*
125	0.731	0.947	0.849	230.019	<0.001*	↓	$PSNS - K^2$	0.118	0.007*
						0.433	$W - K^2$	0.098	0.011*
						0.477	$PSNS - W$	0.057	0.004*
150	0.822	0.879	0.876	130.584	<0.001*	↓	$PSNS - K^2$	0.054	0.009*
						0.216	$W - K^2$	0.003	0.855
						0.390	$PSNS - W$	0.085	<0.001*
175	0.850	0.934	0.876	80.941	0.001*	↓	$PSNS - K^2$	0.026	0.244
						0.085	$W - K^2$	0.058	0.426
						0.441	$PSNS - W$	0.077	0.001*
200	0.885	0.963	0.935	110.023	<0.001*	↓	$PSNS - K^2$	0.049	0.003*
						0.162	$W - K^2$	0.028	0.855
						0.357	$PSNS - W$	0.053	*0.010*
250	0.811	0.864	0.908	70.778	0.002*	↓	$PSNS - K^2$	0.097	0.013*
						0.036	$W - K^2$	0.044	0.715
						0.559	$PSNS - W$	0.083	<0.001*
300	0.889	0.974	0.975	170.744	<0.001*	↓	$PSNS - K^2$	0.082	0.002*
						0.339	$W - K^2$	0.001	0.426
						0.479	$PSNS - W$	0.068	0.003*
400	0.888	0.961	0.958	120.875	<0.001*	↓	$PSNS - K^2$	0.070	0.006*
						0.219	$W - K^2$	0.002	1
						0.435	$PSNS - W$	0.059	<0.001*
500	0.932	0.991	0.991	100.769	< 0.001*	↓	$PSNS - K^2$	0.059	0.012*
						0.153	$W - K^2$	0.0001	0.542

Continued

						0.472	PSNS – K^2	0.038	0.003*
600	0.957	0.988	0.995	120.498	<0.001*	↓	PSNS – W	0.031	0.003*
						0.208	W – K^2	0.008	0.094
						0.521	PSNS – W	0.060	0.004*
700	0.940	1	0.998	150.211	<0.001*	↓	PSNS – K^2	0.058	0.004*
						0.282	W – K^2	0.002	0.010*
						0.416	PSNS – K^2	0.043	0.004*
800	0.957	0.998	0.999	90.964	0.001*	↓	PSNS – W	0.041	0.010*
						0.124	W – K^2	0.001	0.068
						0.368	PSNS – W	0.009	0.004*
900	0.990	0.999	0.998	80.139	0.002*	↓	PSNS – K^2	0.008	0.012*
						0.052	W – K^2	0.001	0.695
						0.493	PSNS – W	0.008	0.004*
1000	0.988	0.996	0.998	130.632	<0.001*	↓	PSNS – K^2	0.010	0.004*
						0.240	W – K^2	0.002	0.105
						0.558	W – K^2	<0.001	0.002*
1500	0.990	1	1	170.698	<0.001*	↓	PSNS – W	0.010	0.004*
						0.337	PSNS – K^2	0.010	0.004*
2000	1	1	1	20.771	0.080	<0.001			
						0.297	W – K^2	0.080	<0.001*
Total	0.744	0.805	0.824	1510.798	<0.001*	↓	PSNS – W	0.061	<0.001*
						0.055	PSNS – K^2	0.020	<0.001*

Note. n = size of the 15 samples with non-normal distributions, Total = the 15 samples with non-normal distributions are grouped, average power when contrasting the null hypothesis of normality by the tests of the PSNS-based, Royston's W, and D'Agostino *et al.*'s K-squared tests, F = value of the Friedman's test statistic with the Agresti-Pendergast modification, p-value = right-tailed probability value with an asterisk in case of significant difference at a significance level of 0.05, $W \to \bar{r}_s$ = Kendall's coefficient of agreement as a measure of effect size [62] and its correspondence with the average of the Spearman's rank correlation coefficient between the PSNS-based, W, and K-squared tests. Pairwise comparisons through the Wilcoxon's exact test: Pair = compared pairs ordered in ascending order by their p values, $|MD|$ = mean difference in absolute value, p value = two-tailed exact probability with an asterisk in case of significant difference at a significance level of 0.05 corrected using the Benjamini-Yekutieli procedure for false-positive discovery rate in correlated samples (0.009 for the first or lowest p-value, 0.018 for the second, and 0.027 for the third), and it is indicated in italics if the difference is significant at a significance level of 0.05 without correction.

Table 7. Comparison of the mean power by sample size assuming normality.

n	Average potency			JNS test		ANOVA				Comparisons	
	PSNS	W	K²	χ²	p-value	Stat.	F	p-value	$\hat{\eta}_p^2$	Pair	p-value
30	0.407	0.482	0.537	380.36	<0.001	HF	50.38	0.011	0.278	PSNS – K²	0.009*
						ε = 1				PSNS – W	0.103
										W – K²	0.108
50	0.378	0.484	0.551	330.96	<0.001	HF	70.36	0.004	0.345	PSNS – K²	0.006*
						ε = 0.91				PSNS – W	*0.035*
										W – K²	0.075
60	0.589	0.757	0.793	28	<0.001	GH	80.01	0.006	0.364	PSNS – W	<0.001*
						ε = 0.72				PSNS – K²	0.009*
										W – K²	0.528
70	0.567	0.651	0.713	200.20	<0.001	GH ε = 0.59	10.91	0.186	0.120		
90	0.513	0.583	0.637	380.42	<0.001	HF	30.93	0.031	0.219	PSNS – K²	*0.030*
						ε = 1				PSNS – W	0.092
										W – K²	0.219
100	0.613	0.641	0.631	390.39	<0.001	HF	00.30	0.744	0.021		
						ε = 1					

Note. n = size of the 15 samples with non-normal distributions, mean of the power (in 15 non-normal distributions) when testing the null hypothesis of normality using the PSNS-based, Royston's W, and D'Agostino *et al.*'s K-squared tests, JNS test = test for sphericity [66] [67] [68]: χ^2 = test statistic, p–value = probability value to the right tail in a chi-squared distribution with two degrees of freedom. ANOVA = repeated measures analysis of variance: Test: HF = analysis of intragroup variance with the Huynh-Feldt correction for sphericity when the epsilon value is greater than or equal to 0.75 ($\varepsilon \geq 0.75$), and GH = with Greenhouse-Geisser correction for sphericity when the value of epsilon is less than 0.75 ($\varepsilon < 0.75$), p value = probability value to the right tail in a distribution $F(2 \times \varepsilon, 11 \times \varepsilon)$, $\hat{\eta}_p^2$ = partial eta squared coefficient. Comparisons with Student's t-test for paired samples, p-value = two-tailed exact probability with an asterisk in case of significant difference at a significance level of 0.05 corrected using the Benjamini-Yekutieli procedure for false-positive discovery rate in correlated samples (0.009 for the first or lowest p-value, 0.018 for the second, and 0.027 for the third), and it is indicated in italics if the difference is significant at a significance level of 0.05 without correction.

This last table includes the comparisons of the average type II error of the PSNS-based, W, and K-squared tests when applied to the 24 samples of different sizes generated to follow a standard normal distribution.

The PSNS-based, W, and K-squared tests had an average power of less than 0.70 with sample sizes of 100 or less. With a sample of 125, the PSNS-based test had an average power of 0.73, and the other two tests were greater than 0.80.

Table 8. Comparison of the average power or type II error by type of distribution without assuming normality.

Distr.	Average power				Friedman's test		$W \rightarrow \bar{r}_s$	Wilcoxon's exact test		
	PD	PSNS	W	K^2	F	p-value		Pair	$\|MD\|$	p-value
W	ϕ	0.622	0.698	0.712	170.12	<0.001*	0.427	PSNS – K^2	0.090	<0.001*
							↓	PSNS – W	0.076	0.004*
							0.141	W – K^2	0.014	0.684
T	ϕ	0.818	0.884	0.805	220.17	<0.001*	0.491	PSNS – W	0.066	<0.001*
							↓	W – K^2	0.079	0.001*
							0.237	PSNS – K^2	0.012	0.165
ZF	ϕ	0.715	0.760	0.780	130.30	<0.001*	0.366	PSNS – K^2	0.065	0.003*
							↓	W – K^2	0.020	_0.027_
							0.049	PSNS – W	0.045	0.056
Sin^{-1}	ϕ	0.874	0.936	0.950	310.25	<0.001*	0.576	PSNS – K^2	0.076	<0.001*
							↓	W – K^2	0.014	0.001*
							0.364	PSNS – W	0.062	0.001*
U	ϕ	0.751	0.838	0.897	810.37	<0.001*	0.780	W – K^2	0.059	<0.001*
							↓	PSNS – K^2	0.145	<0.001*
							0.670	PSNS – W	0.087	<0.001*
SC	ϕ	0.609	0.750	0.847	720.17	<0.001*	0.758	PSNS – K^2	0.239	<0.001*
							↓	W – K^2	0.098	<0.001*
							0.637	PSNS – W	0.141	<0.001*
N (0, 1)	β	0.807	0.902	0.828	140.11	<0.001*	0.380	PSNS – W	0.095	<0.001*
							↓	W – K^2	0.074	0.002*
							0.070	PSNS – K^2	0.021	0.375
Logistic	ϕ	0.473	0.576	0.636	90.06	<0.001*	0.283	PSNS – K^2	0.163	<0.001*
							↓	W – K^2	0.060	0.005*
							0.076	PSNS – W	0.103	0.061
Laplace	ϕ	0.788	0.795	0.844	60.69	0.003*	0.225	PSNS – K^2	0.055	_0.016_
							↓	W – K^2	0.049	0.190
							0.163	PSNS – W	0.006	0.198
Student	ϕ	0.639	0.716	0.784	190.53	<0.001*	0.459	PSNS – K^2	0.146	<0.001*
							↓	W – K^2	0.069	0.003*
							0.189	PSNS – W	0.077	0.089

Continued

Distr.	PD	PSNS	W	K²	F	p-value	$W \to \bar{r}_s$	Pair	|MD|	p-value
Cauchy	ϕ	0.992	0.998	0.999	30.05	0.057	0.117			
							0.429	PSNS – K²	0.127	0.002*
PERT	ϕ	0.550	0.621	0.677	170.31	<0.001*	↓	PSNS – W	0.071	0.004*
							0.144	W – K²	0.056	0.056
							0.465	PSNS – W	0.103	<0.001*
Ray	ϕ	0.712	0.815	0.782	200.01	<0.001*	↓	PSNS – K²	0.070	0.003*
							0.198	W – K²	0.033	0.060
							0.245	PSNS – W	0.029	*0.012*
χ^2	ϕ	0.851	0.880	0.864	70.46	0.002*	↓	PSNS – K²	0.013	0.132
							0.133	W – K²	0.016	0.395
							0.268	PSNS – W	0.043	0.002*
LogN	ϕ	0.850	0.892	0.885	80.43	0.001*	↓	PSNS – K²	0.035	0.008*
							0.098	W – K²	0.008	0.623
							0.171	PSNS – W	0.005	0.144
F	ϕ	0.916	0.910	0.902	40.75	0.013*	↓	PSNS – K²	0.013	0.233
							0.244	W – K²	0.008	0.782

Note. Distr. = type of distribution followed by the 24 samples with different sizes: W = Weibull ($\alpha = 2$, $\beta = 2$), T = Triangular ($a = -2$, $b = 2$, $c = 2$), ZF = Fisher's Z ($\nu_1 = 3$, $\nu_2 = 9$), Sin^{-1} = Arcsine ($a = -2$, $b = 2$), U = Uniform ($a = -2$, $b = 2$), SC = Semicircle ($r = 2$), N ($\mu = 0$, $\sigma^2 = 1$), Logistic ($\mu = 0$, $s = 1$), Laplace ($\mu = 0$, $b = 1$), Student's t ($\nu = 5$), Cauchy ($x_0 = 0$, $\gamma = 1$), PERT ($a = -2$, $b = 0$, $c = 4$, $\lambda = 4$), Ray = Rayleigh ($\sigma = 1$), χ^2 = Chi-squared ($\nu = 4$), LogN = Lognormal ($\mu = 0$, $\sigma^2 = 0.25$) and F ($\nu_1 = 9$, $\nu_2 = 9$). *PD* = probabilities for the decision in the statistical test: ϕ = power or probability that the test statistic remains in the rejection region conditioned to the alternative hypothesis (parameter at the critical value), that is, to reject the null hypothesis when is false, and β = type II error or probability that the test statistic remains in the region of acceptance conditioned to the alternative hypothesis (parameter at the critical value), that is, to maintain the null hypothesis when it is false. Average power = arithmetic means of power (ϕ) or type II error (β) when contrasting the null hypothesis of normality using the PSNS-based, Royston's W, and D'Agostino *et al.*'s K-squared tests in 24 random samples of different sizes with the same distribution. *F* = value of the Friedman's test statistic with the Agresti-Pendergast modification, *p*-value = probability to the right tail in a Fisher-Snedecor F distribution with degrees of freedom 2 and 28, it is marked with an asterisk in case of significant difference at a significance level of 0.05. $W \to \bar{r}_s$ = Kendall's W concordance coefficient as a measure of effect size [62] and its correspondence with the average of Spearman's rank correlation coefficients between the PSNS-based, W, and K-squared tests. Pairwise comparisons through the Wilcoxon's exact test: Pair = compared pairs ordered in ascending order by their *p* values, \|*MD*\| = mean difference in absolute value, *p*-value = two-tailed exact probability with an asterisk in case of significant difference at the significance level of 0.05 corrected by the Benjamini-Yekutieli procedure for false-positive discovery rate in correlated samples (0.009 for the first or lowest *p*-value, 0.018 for the second, and 0.027 for the third), and it is indicated in italics if the difference is significant at a significance level of 0.05 without correction.

With samples of 150 and more, all three tests had an average power greater than 0.80. Except for the sample sizes of 70, 100, and 2000, there was a significant difference in the average power among PSNS-based, W, and K-squared tests through the omnibus test. With sizes of 20 and 40, the K-squared test had higher average power than the W test, the average power of the PSNS-based test being equivalent to that of the other two tests. In the other nineteen significant pairwise comparisons, PSNS-based test either had significantly lower power than W and K-squared test (with size sample of 60, 80, 125, 150, 200, 300, 400, 500, 600, 700, 800, 900, 1000, and 1500) or lower than one of them (than the W test with a sample size of 175 and the K-squared test with sample sizes of 30, 50, 90 and 250). On the other hand, the power was equivalent between W and K-squared test, except with the size of 125 in which the W test had more power than the K-squared test and with the size of 1500 in which the K-squared test had more power than the W test. With the sample size of 2000, the average power was unit with no difference among the three tests. When grouping the 15 non-normal distributions and the 24 sample sizes, the PSNS-based test had an average power of 0.74, W test of 0.81, and K-squared test of 0.82. By the omnibus test, there was a significant difference with a very small effect size (Kendall's W = 0.30 → $\overline{r_S}$ = 0.05). In the pairwise comparisons, the K-squared test had higher power than the other two tests, and the PSNS-based test had lower power than the other two tests. See Table 6 and Table 7.

Regarding the distributions, there was the only equivalence of average power with the Cauchy distribution by the omnibus test. In the F distribution, the omnibus test showed a significant difference, but the effect size was very small, and the pairwise comparisons did not detect any significant difference, even without the Benjamini-Yekutieli correction for false-positive discovery rate. With the Laplace and chi-squared distributions, the omnibus test also shows a significant difference, but not the pairwise comparisons with the correction. Without the correlation, the K-squared test had higher power than the PSNS-based test in the Laplace distribution (p = 0.016) and the W test had higher power than the PSNS-based test in the chi-squared distribution (p = 0.012). In the other distributions, there were significant differences by pairs and the PSNS-based test that had the lowest average power. The effect size was large ($W \geq 0.67$) in the uniform and semicircle distributions, medium (≥ 0.53) in the arcsine distribution, and there were significant differences between the three tests in the three distributions. The K-squared test had the highest average power, the W test intermediate, and the PSNS-based test the lowest. The average power was equivalent between K-squared and W tests and greater than that of the PSNS-based test in the Weibull, Rayleigh, and LogNormal distributions. In the triangular and normal distributions, the average power of the W test was greater than that of the other two that had an equivalent average power. In Fisher's Z, logistic, and Student's t distribution, the mean power of the K-squared test was higher than those of the other two tests that had an equivalent mean power. See Table 8.

4. Discussion

The present study takes up the proposal to test whether a sequence of sample data has been drawn from a normal distribution using seven quantiles that are equally spaced, at a distance of two-thirds, from the central point in a standard normal distribution [4]. The proposal is an adaptation of the seven-number summary for an assessment of normality [47]. As the summary refers to the normal distribution, it is qualified as parametric [48]. Until now, its application in the form of an omnibus test of significance had not been considered. In this work, a very simple proposal is posed, considering that the standard error and distribution of the quantiles from their asymptotic approximation to the normal distribution is known [49]. To calculate the test statistic, first, the sample quantiles corresponding to the seven equidistant population quantiles in a normal distribution are calculated, second, they are standardized with the sample mean and standard deviation, and third, the standardized difference of each sample quantile with the expected quantile is calculated on the basis that the sampling distribution of quantiles is asymptotically normal. The sum of squares of these differences constitutes the test statistic that follows a chi-squared distribution with seven degrees of freedom under the null hypothesis of normality and large sample size [57]. In turn, the non-central chi-squared distribution allows calculating the type II error and statistical power [79].

Once the proposal is made, it is necessary to answer whether the new test for normality based on PSNS has accuracy, that is, whether its proportion of correct decisions when maintaining the null hypothesis against a sample drawn from a normal distribution and rejecting the null hypothesis against samples drawn from non-normal distributions is at least 0.70 and preferably greater than 0.80 [80]. It is also necessary to find out whether its probability of rejecting the null hypothesis of normality when it is false (power) in samples from non-normal populations is high, that is, greater than or equal to 0.80 [81]. On the contrary, does the PSNS-based test show a potency is very low (<0.20) when the sample is drawn from a normal distribution? There are many tests of significance for normality with proven accuracy and power, such as the W test of Royston [35] and the K-squared test of D'Agostino *et al.* [44], so it is required to know whether the PSNS-based test is equivalent in accuracy and power to these tests for normality.

The proposal has a first large sample requirement. The minimum criterion for a large sample is variable, so large samples can be considered those greater than or equal to 20, 30, 40, 100, 200, 500, or 1000 [82]. It was decided to start with a small size of 20 and go up to 2000, contemplating a total of 24 sample sizes. Within each sample size, 16 distributions were considered, one normal and 15 non-normal. The latter include a negative asymmetric and platykurtic distribution, namely Triangular ($a = -2$, $b = 2$, $c = 2$); two negative asymmetric and leptokurtic distributions, namely Weibull ($\alpha = 2$, $\beta = 2$) and Fisher's Z ($\nu_1 = 3$, $\nu_2 = 9$); three symmetric and platykurtic distributions, namely Arcsine ($a = -2$, $b =$

2), $U(a = -2, b = 2)$ and Wigner's semicircle ($r = 2$); four symmetric and lepto-kurtic distributions, namely logistic ($\mu = 0$, $s = 1$), Laplace ($\mu = 0$, $b = 1$), Student's t ($v = 5$), and Cauchy ($x_0 = 0$, $\gamma = 1$), a positive asymmetric and platykurtic distribution, namely PERT ($a = -2$, $b = 0$, $c = 4$, $\lambda = 4$); and four positive asymmetric and leptokurtic distributions, namely Rayleigh ($\sigma = 1$), χ^2 ($v = 4$), Log-normal ($\mu = 0$, $\sigma^2 = 0.25$), and F ($v_1 = 9$, $v_2 = 9$).

It has been pointed out that simulation results have little generalizability because real situations are much more complex than the simplified reality assumed by the simulation [83]. Nevertheless, this remark refers to explanatory models of real complex phenomena and does not apply to the type of research presented [84]. To achieve the objectives of the proposed study, it is necessary to know exactly the population from which each sample is drawn. The simplest and most direct method to achieve this is simulation. Simulation using random extractions allows its generalization to real situations with variables that follow with good fit some of the considered distribution models [84]. On the contrary, carrying out the study using empirical samples of variables with clearly established distributions would have provided less certainty, as well as being more expensive and less generalizable [85]. Therefore, simulation was chosen as the best option.

It should be noted that faced with distribucions distribution without a closed-form density function, but infinitely divisible, there exists a simulation option, namely the minimum Hellinger distance (SMHD) estimation. Luong and Bilodeau [86] have shown that the SMHD estimators are consistent and high efficiency in general with less regularity conditions needed than the maximum likelihood estimators. In adittion, they have asymptotic efficiency and normality, as well as the potential to be robust.

As expected, the accuracy of the PSNS-based omnibus test is very low with sample sizes from 20 to 50. Nevertheless, it is also low for the W and K-squared tests. Even the accuracy is equivalent between the three tests, either by the omnibus test or the pairwise comparisons. Although the W and K-squared tests are supposed to be powerful with these sample sizes [24] [44], the present results agree with recent studies with small samples [52] [53] [54]. On the other hand, the average power is low, although there is a difference between the PSNS-based, W, and K-squared tests with these sample sizes (20 to 50). The K-squared test shows the highest average power. With samples of 20 and 40, the average power of the PSNS-based test is equivalent to that of the W test; however, with sizes 30 and 50, the PSNS-based test has the lowest accuracy. This pattern continues up to 100. Namely, a pattern of low accuracy in the three tests with equivalent proportions by the omnibus test or the pairwise comparisons when the omnibus test is significant, but with a low average power difference, showing the PSNS-based test the lowest average power and the K-squared test the highest.

With a size of 125, the accuracy and average power of the W and K-squared tests are high, while the PSNS-based test is not. A sample size of 150 is required for the PSNS-based omnibus test to have high accuracy and power. Conse-

quently, the large sample size requirement implies a sample of at least 150 participants. From 200 participants, the accuracy of the PSNS-based test is equivalent to the other two tests using the omnibus test, but its power is lower than that of the K-squared and W tests when these last two tests have equivalent power, except with a sample of 1,500, in which the K-squared test outperforms the W test.

As expected, the logistic distribution, which is the closest to the normal distribution [87], is difficult to discriminate for the three tests with small sample sizes, so sizes of 500 or larger are required for W and K-squared test to be high average power and 600 or larger are required for the PSNS-based test to be high average power. The low average power situation also appears with the PERT ($a = 2$, $b = 0$, $c = 4$, $\lambda = 4$) distribution, which is a very slightly positive asymmetry ($\beta_1 = 0.30$) and platykurtic ($B_2 = -0.55$) distribution. A sample size of 250 is required for the PSNS-based and K-squared tests to achieve high average power and 175 for the W test to achieve high power.

All three tests have high and equivalent power in discriminating the Cauchy and F ($v_1 = 9$, $v_2 = 9$) distributions, which are quite far from normal distribution due to leptokurtosis, as well as in discriminating the Laplace and chi-squared (with four degrees of freedom) distributions, which are also leptokurtic, although the statical equivalence in average power with the latter is through pairwise comparisons corrected for false-positive discovery rate.

With significant difference, the PSNS-based test has lower power than the K-squared and W tests with symmetrical-platykurtic distributions, such as the uniform, semicircle, and arcsine distributions; the effect size is large with the first two and medium with the last. With the triangular, Rayleigh, Student's t, PERT, and Weibull distributions, the effect size is small, with the PSNS-based test showing less power than the W and K-squared tests, which do not differ in power. With the logistic and lognormal distributions, the effect size is very small, and the same differential pattern appears. With Fisher's Z distribution, the effect size is very small, but the K-squared test has a differentially higher power than the PSNS-based test when the PSNS-based and W tests do not differ in power.

As for the normal distribution, the three tests not only have an accuracy of 100%, but also a mean power of less than 0.20 and a type II error of more than 0.80, making them sensitive to the presence of normality. The W test is the one that generates the highest error by maintaining the null hypothesis of normality conditional on the alternative hypothesis being true (type II error) with a significant difference compared to the K-squared and PSNS-based tests that present an equivalent level of error. Therefore, the W test was the most sensitive to normality, as previously reported in other studies [24] [44] [52] [53] [54].

5. Conclusion

The proposed seven-number parametric summary as a significance test for normality is accurate (correct decisions greater than 0.70) and powerful (mean

power greater than 0.80) with samples of 150 or more. It is less powerful than the W and K-squared tests up to a sample size of 2000, before which the power of all three tests is unity. The greatest difference in mean power appears with the symmetrical-platykurtic distributions. It should be noted that the W and K-squared tests also show low power with sample sizes below 100, especially when faced with logistic, PERT, and Weibull distributions.

6. Limitations

A limitation of the study is that it consists of a single repetition of the experiment, hence it constitutes a pilot test of the proposal. In turn, repeating the experiment a large number of times, such as 500 or 1000 times, makes it possible to calculate the sensitivity (probability of detecting cases), specificity (probability of rejecting non-cases), positive predictive value (probability that at giving a positive test is a case), and negative predictive value (probability that a negative test is a non-case) in each of the 384 conditions (16 types of distribution and 24 sample sizes). For these estimations, the case is normality with normal samples, but the case definition should be reversed with samples drawn from non-normal distributions.

7. Suggestions

If the PSNS is contemplated, it is recommended to apply the omnibus test proposed in this study with large samples ($n \geq 150$). Furthermore, it is suggested to replicate this study simulating a large number of samples per condition, even extending the conditions by including the hyperbolic secant, hyperexponential, Pareto, exponentially modified Gaussian, truncated normal, and raised cosine distributions.

With longer than 150 observations, the based-PSNS test can be used with due caution, which involves examining the histogram, boxplot, and normal quantile-quantile plot. It should be noted that Royston's W and D'Agostino *et al.*'s K-squared tests also show deficiencies in precision and power with samples less than 100 against non-normal distributions. However, the three normality tests are accurate and powerful with small samples taken from a normal distribution.

Acknowledgements

The author would like to thank the referee for their helpful comments.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Bowley, A.L. (1901) Elements of Statistics. P. S. King, London.

[2] Friendly, M. (2007) A Brief History of Data Visualization. In: Chen, C., Härdle, W.

and Unwin, A., Eds., *Handbook of Computational Statistics: Data Visualization*, Springer-Verlag, Heidelberg, 1-34.

[3] Bowley, A.L. (1910) Elementary Manual of Statistics. P. S. King, London.

[4] Surhone, L.M., Timpledon, M.T. and Marseken, S.F. (2010) Seven-Number Summary: Descriptive Statistics, Summary Statistics, Five-Number Summary, Box Plot, Parametric Statistics. Betascript Publishing, Beau-Bassin.

[5] Spear, M.E. (1952) Charting Statistics. McGraw-Hill, New York.

[6] Spear, M.E. (1969) Practical Charting Techniques. McGraw-Hill, New York.

[7] Tukey, J.W. (1970) Exploratory Data Analysis. Limited Preliminary Edition, Addison-Wesley, Reading.

[8] Tukey, J.W. (1972) Some Graphic and Semi-Graphic Displays. In: Bancroft, T.A. and Brown, S.A., Eds, *Statistical Papers in Honor of George W. Snedecor*, Iowa State University Press, Aimes, 293-316.

[9] Tukey, J.A. (1977) Exploratory Data Analysis. Addison and Wesley, Reading.

[10] Wickham, H. and Stryjewski, L. (2011) 40 Years of Boxplots.
http://vita.had.co.nz/papers/boxplots.pdf

[11] Cox, N.J. (2009) Speaking Stata: Creating and Varying Box Plots. *Stata Journal*, **9**, 478-496. https://doi.org/10.1177/1536867X0900900309

[12] McGill, R., Tukey, J.W. and Larsen, W.A. (1978) Variations of Box Plots. *The American Statistician*, **32**, 12-16. https://doi.org/10.1080/00031305.1978.10479236

[13] Cleveland, W.S. (1985) Elements of Graphing Data. Wadsworth, Monterey.

[14] Gotelli, N.J. and Ellison, A.M. (2004) A Primer of Ecological Statistics. Sinauer Associates, Inc., Sunderland.

[15] Reimann, C., Filzmoser, P., Garrett, R.G. and Dutter, R. (2008) Statistical Data Analysis Explained: Applied Environmental Statistics with R. John Wiley and Sons, New York. https://doi.org/10.1002/9780470987605

[16] Benjamini, Y. (1988) Opening the Box of a Boxplot. *The American Statistician*, **42**, 257-262. https://doi.org/10.1080/00031305.1988.10475580

[17] Hintze, J.L. and Nelson, R.D. (1998) Violin Plots: A Box-Plot Density Trace Synergism. *The American Statistician*, **52**, 181-184.
https://doi.org/10.1080/00031305.1998.10480559

[18] Kampstra, P. (2008) Beanplot: A Boxplot for Visual Comparison of Distributions. *Journal of Statistical Software Code Snippets*, **28**, 1-9.
https://doi.org/10.18637/jss.v028.c01

[19] Hyndman, R. (1996) Computing and Graphing Highest Density Regions. *The American Statistician*, **50**, 120-126.
https://doi.org/10.1080/00031305.1996.10474359

[20] Muth, S.Q., Potterat, J.J. and Rothenberg, R.B. (2000) Birds of Feather: Using a Rotational Box Plot to Assess Ascertainment Bias. *International Journal of Epidemiology*, **29**, 899-904. https://doi.org/10.1093/ije/29.5.899

[21] Phuyal, S., Rashid, M. and Sarkar, J., (2021) GIplot: An R Package for Visualizing the Summary Statistics of a Quantitative Variable.
https://cran.r-project.org/web/packages/GIplot/index.html

[22] Sarkar, J. and Rashid, M. (2021) IVY Plots and Gaussian Interval Plots. *Teaching Statistics*, **43**, 85-90. https://doi.org/10.1111/test.12257

[23] Rousseeuw, P.J., Ruts, I. and Tukey, J.W. (1999) The Bagplot: A Bivariate Boxplot.

The American Statistician, **53**, 382-387.
https://doi.org/10.1080/00031305.1999.10474494

[24] Mishra, P., Pandey, C.M., Singh, U., Gupta, A., Sahu, C. and Keshri, A. (2019) Descriptive Statistics and Normality Tests for Statistical Data. *Annals of Cardiac Anaesthesia*, **22**, 67-72. https://doi.org/10.4103/aca.ACA_157_18

[25] Pearson, K. (1900) On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Series 5, **50**, 157-175. https://doi.org/10.1080/14786440009463897

[26] Woolf, B. (1957) The Log-Likelihood Ratio Test (The G-Test); Methods and Tables for Tests of Heterogeneity in Contingency Tables. *Annals of Human Genetics*, **21**, 397-409. https://doi.org/10.1111/j.1469-1809.1972.tb00293.x

[27] Kolmogorov, A.N. (1933) Sulla Determinizione Empirica di una Legge di Distribuzione. *Giornale dell Istituto Italiano degli Attuari*, **4**, 83-91.

[28] Smirnov, N.V. (1939) On the Estimation of the Discrepancy between Empirical Curves of Distributions for Two Independent Samples. *Moscow University Mathematics Bulletin*, **2**, 3-26.

[29] Kuiper, N.H. (1960) Tests Concerning Random Points on a Circle. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, Series A, **63**, 38-47. https://doi.org/10.1016/S1385-7258(60)50006-0

[30] Cramer, H. (1928) On the Composition of Elementary Errors. *Scandinavian Actuarial Journal*, **1**, 13-74. https://doi.org/10.1080/03461238.1928.10416862

[31] Von Mises, R.E. (1928) Wahrscheinlichkeit, Statistik und Wahrheit [Probability, Statistics and Truth]. Verlag von Julius Springer, Wien. https://doi.org/10.1007/978-3-662-36230-3

[32] Anderson, T.W. and Darling, D.A. (1952) Asymptotic Theory of Certain "Goodness-of-Fit" Criteria Based on Stochastic Processes. *Annals of Mathematical Statistics*, **23**, 193-212. https://doi.org/10.1214/aoms/1177729437

[33] Watson, G.S. (1961) Goodness-of-Fit Tests on a Circle. *Biometrika*, **48**, 109-114. https://doi.org/10.1093/biomet/48.1-2.109

[34] Shapiro, S.S. and Wilk, M.B. (1965) An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, **52**, 591-611. https://doi.org/10.1093/biomet/52.3-4.591

[35] Royston, J.P. (1992) Approximating the Shapiro-Wilk W-Test for Non-Normality. *Statistics and Computing*, **2**, 117-119. https://doi.org/10.1007/BF01891203

[36] Chen, L. and Shapiro, S.S. (1995) An Alternative Test for Normality Based on Normalized Spacings. *Journal of Statistical Computation and Simulation*, **53**, 269-288. https://doi.org/10.1080/00949659508811711

[37] Shapiro, S.S. and Francia, R.S. (1972) An Approximate Analysis of Variance Test for Normality. *Journal of the American Statistical Association*, **67**, 215-216. https://doi.org/10.1080/01621459.1972.10481232

[38] Royston, J.P. (1993) A Toolkit of Testing for Non-Normality in Complete and Censored Samples. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **42**, 37-43. https://doi.org/10.2307/2348109

[39] D'Agostino, R.B. (1971) An Omnibus Test of Normality for Moderate and Large Size Samples. *Biometrika*, **58**, 341-348. https://doi.org/10.1093/biomet/58.2.341

[40] D'Agostino, R.B. (1972) Small Sample Probability Points for the D Test of Normal-

ity. *Biometrika*, **59**, 219-221. https://doi.org/10.2307/2334638

[41] D'Agostino, R.B. and Pearson, E.S. (1973) Tests for Departure from Normality. Empirical Results for the Distributions of b_2 and $\sqrt{b_1}$. *Biometrika*, **60**, 613-622. https://doi.org/10.1093/biomet/60.3.613

[42] Jarque, C.M. and Bera, A. (1987) A Test for Normality of Observations and Regression Residuals. *International Statistical Review*, **55**, 163-172. https://doi.org/10.2307/1403192

[43] Urzua, C. (1996) On the Correct Use of Omnibus Tests for Normality. *Economics Letters*, **53**, 247-251. https://doi.org/10.1016/S0165-1765(96)00923-8

[44] D'Agostino, R.B., Belanger, A. and D'Agostino Jr., R.B. (1990) A Suggestion for Using Powerful and Informative Tests of Normality. *The American Statistician*, **44**, 316-321. https://doi.org/10.1080/00031305.1990.10475751

[45] Moore, D. (1986) Tests of Chi-Squared Type. In: D'Agostino, R.B. and Stephens, M.A., Eds., *Goodness-of-Fit Techniques*, Marcel Dekker, New York, 63-95. https://doi.org/10.1201/9780203753064-3

[46] Wilk, M.B. and Gnanadesikan, R. (1968) Probability Plotting Methods for the Analysis of Data. *Biometrika*, **55**, 1-17. https://doi.org/10.2307/2334448

[47] Estaki, M., Jiang, L., Bokulich, N.A., McDonald, D., González, A., Kosciolek, T., Martino, C., Zhu, Q., Birmingham, A., Vázquez-Baeza, Y., Dillon, M.R., Bolyen, E., Caporaso, J.G. and Knight, R. (2020) QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data. *Current Protocols in Bioinformatics*, **70**, e100. https://doi.org/10.1002/cpbi.100

[48] Milligan, C., Montufar, J., Regehr, J. and Ghanney, B. (2016) Road Safety Performance Measures and AADT Uncertainty from Short-Term Counts. *Accident Analysis & Prevention*, **97**, 186-196. https://doi.org/10.1016/j.aap.2016.09.013

[49] Stuart, A. and Ord, J.K. (1994) Kendall's Advanced Theory of Statistics. Volume 1. Distribution Theory. Sixth Edition, Edward Arnold, London.

[50] Hyndman, R.J. and Fan, Y. (1996) Sample Quantiles in Statistical Packages. *The American Statistician*, **50**, 361-365. https://doi.org/10.1080/00031305.1996.10473566

[51] Blom, G. (1958) Statistical Estimates and Transformed Beta Variables. John Wiley and Sons, New York.

[52] Yap, B.W. and Sim, C.H. (2011) Comparisons of Various Types of Normality Tests. *Journal of Statistical Computation and Simulation*, **81**, 2141-2155. https://doi.org/10.1080/00949655.2010.520163

[53] Khatun, N. (2021) Applications of Normality Test in Statistical Analysis. *Open Journal of Statistics*, **11**, 113-122. https://doi.org/10.4236/ojs.2021.111006

[54] Alonso, J.C. and Montenegro, S. (2015) A Monte Carlo Study to Compare 8 Normality Tests for Least-Squares Residuals Following a First Order Autoregressive Process. *Estudios Gerenciales*, **31**, 253-265. https://doi.org/10.1016/j.estger.2014.12.003

[55] Sanchez-Espigares, J.A., Grima, P. and Marco-Almagro, L. (2019) Graphical Comparison of Normality Tests for Unimodal Distribution Data. *Journal of Statistical Computation and Simulation*, **89**, 145-154. https://doi.org/10.1080/00949655.2018.1539085

[56] Mukasa, E.S., Christospher, W., Ivan, B. and Kizito, M. (2021) The Effects of Parametric, Non-Parametric Tests and Processes in Inferential Statistics for Business

Decision Making. *Open Journal of Business and Management*, **9**, 1510-1526. https://doi.org/10.4236/ojbm.2021.93081

[57] Gupta, S.C. and Kapoor, V.K. (2020) Fundamentals of Mathematical Statistics. Twelfth Edition, Sultan Chand & Sons, New Delhi.

[58] Cochran, W.G. (1950) The Comparison of Proportions in Matched Samples. *Biometrika*, **37**, 256-266. https://doi.org/10.1093/biomet/37.3-4.256

[59] Serlin, R.C., Carr, J. and Marascuilo, L.A. (1982) A Measure of Association for Selected Nonparametric Procedures. *Psychological Bulletin*, **92**, 786-790. https://doi.org/10.1037/0033-2909.92.3.786

[60] Friedman, M. (1940) A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *The Annals of Mathematical Statistics*, **11**, 86-92. https://doi.org/10.1214/aoms/1177731944

[61] Kendall, M.G. and Babington-Smith, B. (1939) The Problem of m Rankings. *The Annals of Mathematical Statistics*, **10**, 275-287. https://doi.org/10.1214/aoms/1177732186

[62] Kendall, M.G. and Gibbons, J.D. (1990) Rank Correlation Methods. Oxford University Press, New York.

[63] Hopkins, K.D. (2006) A New View of Statistics. A Scale of Magnitudes for Effect Statistics. http://www.sportsci.org/resource/stats/effectmag.html

[64] McNemar, Q. (1947) Note on the Sampling Error of the Difference between Correlated Proportions and Percentages. *Psychometrika*, **12**, 153-157. https://doi.org/10.1007/BF02295996

[65] Benjamini, Y. and Yekutieli, D. (2001) The Control of the False Discovery Rate in Multiple Testing under Dependency. *Annals of Statistics*, **29**, 1165-1188. https://doi.org/10.1214/aos/1013699998

[66] John, S. (1972) The Distribution of a Statistic Used for Testing Sphericity of Normal Distributions. *Biometrika*, **59**, 169-173. https://doi.org/10.1093/biomet/59.1.169

[67] Nagao, H. (1973) On Some Test Criteria for Covariance Matrix. *Annals of Statistics*, **1**, 700-709. https://doi.org/10.1214/aos/1176342464

[68] Sugiura, N. (1972) Locally Best Invariant Test for Sphericity and the Limiting Distributions. *Annals of Mathematical Statistics*, **43**, 1312-1316. https://doi.org/10.1214/aoms/1177692481

[69] Greenhouse, S.W. and Geisser, S. (1959) On Methods in the Analysis of Profile Data. *Psychometrika*, **24**, 95-112. https://doi.org/10.1007/BF02289823

[70] Huynh, H. and Feldt, L.S. (1976) Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs. *Journal of Educational Statistics*, **1**, 69-82. https://doi.org/10.3102/10769986001001069

[71] Cohen, J. (1988) Statistical Power Analysis for Behavioral Sciences. Second Edition, Lawrence Erlbaum Associates, Hillsdale.

[72] Howell, D.C. (2010) Statistical Methods for Psychology. Seventh Edition, Wadsworth Cengage Learning, Belmont.

[73] Friedman, M. (1937) The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, **32**, 675-701. https://doi.org/10.1080/01621459.1937.10503522

[74] Friedman, M. (1939) A Correction. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, **34**, 109. https://doi.org/10.2307/2279169

[75] Agresti, A. and Pendergast, J. (1986) Comparing Mean Ranks for Repeated Measures Data. *Communications in Statistics—Theory and Methods*, **15**, 1417-1433. https://doi.org/10.1080/03610928608829193

[76] Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, **1**, 80-83. https://doi.org/10.2307/3001968

[77] Microsoft Corporation (2019) Microsoft Excel 2019 for Windows. https://office.microsoft.com/excel

[78] Zaiontz, C. (2021) Real Statistics Resource Pack Software. Release 7.6. http://www.real-statistics.com

[79] Ramachandran, K.M. and Tsokos, C.P. (2020) Mathematical Statistics with Applications in R. Third Edition, Academic Press, Hoboken.

[80] Fay, M.P., Sachs, M.C. and Miura, K. (2018) Measuring Precision in Bioassays: Rethinking Assay Validation. *Statistics in Medicine*, **37**, 519-529. https://doi.org/10.1002/sim.7528

[81] Aberson, C.L. (2019) Applied Power Analysis for the Behavioral Science. Second Edition, Routledge, New York. https://doi.org/10.4324/9781315171500

[82] Brysbaert, M. (2019) How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition*, **2**, 16. https://doi.org/10.5334/joc.72

[83] Meredith, J. (1998) Building Operations Management Theory through Case and Field Research. *Journal of Operations Management*, **16**, 441-454. https://doi.org/10.1016/S0272-6963(98)00023-0

[84] Morris, T.P., White, I.R. and Crowther, M.J. (2019) Using Simulation Studies to Evaluate Statistical Methods. *Statistics in Medicine*, **38**, 2074-2102. https://doi.org/10.1002/sim.8086

[85] Overton, C.E., Stage, H.B., Ahmad, S., Curran-Sebastian, J., Dark, P., Das, R., Fearon, E., Felton, T., Fyles, M., Gent, N., Hall, I., House, T., Lewkowicz, H., Pang, X., Pellis, L., Sawko, R., Ustianowski, A., Vekaria, B. and Webb, L. (2020) Using Statistics and Mathematical Modelling to Understand Infectious Disease Outbreaks: COVID-19 as an Example. *Infectious Disease Modelling*, **5**, 409-441. https://doi.org/10.1016/j.idm.2020.06.008

[86] Luong, A. and Bilodeau, C. (2018) Asymptotic Normality Distribution of Simulated Minimum Hellinger Distance Estimators for Continuous Models. *Open Journal of Statistics*, **8**, 846-860. https://doi.org/10.4236/ojs.2018.85056

[87] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995) Continuous Univariate Distributions. Second Editon, John Wiley and Sons, New York.