sciendo

DOI: 10.2478/bile-2020-0015

Biometrical Letters
Vol. 57 (2020), No. 2, 237-251

Comparative evaluation of goodness of fit tests for normal distribution using simulation and empirical data

Achilleas Anastasiou[1], **Alex Karagrigoriou**[1], **Anastasios Katsileros**[2]

[1]Department of Statistics and Actuarial-Financial Mathematics, Laboratory of Statistics and Data Analysis, University of the Aegean, e-mail: sasd20003@sas.aegean.gr, alex.karagrigoriou@aegean.gr
[2]Department of Crop Science, Laboratory of Plant Breeding and Biometry, Agriculture University of Athens, e-mail: katsileros@aua.gr

Summary

The normal distribution is considered to be one of the most important distributions, with numerous applications in various fields, including the field of agricultural sciences. The purpose of this study is to evaluate the most popular normality tests, comparing the performance in terms of the size (type I error) and the power against a large spectrum of distributions with simulations for various sample sizes and significance levels, as well as through empirical data from agricultural experiments. The simulation results show that the power of all normality tests is low for small sample size, but as the sample size increases, the power increases as well. Also, the results show that the Shapiro–Wilk test is powerful over a wide range of alternative distributions and sample sizes and especially in asymmetric distributions. Moreover the D'Agostino–Pearson Omnibus test is powerful for small sample sizes against symmetric alternative distributions, while the same is true for the Kurtosis test for moderate and large sample sizes.

Key words: type I error, power, normality tests, normal and alternative distributions, simulation

1. Introduction

Plant breeding is defined as the science of changing plant traits in order to produce new genotypes with desired characteristics. The traits can be divided into two distinct types based on their effects on the plant phenotype: qualitative and quantitative traits. In qualitative traits, the phenotypic expression is determined by single genes and the phenotypic distribution is

discrete, while in quantitative traits, the phenotypic expression is a combination of the effect of multiple genes and the effect of the environment, and this results in the phenotypic distribution being continuous. For the identification of the superior genotypes from the phenotype and the evaluation of their quantitative traits, the researcher uses carefully designed experiments and analytical methods.

Nearly all inferential statistics (t-tests, ANOVA, simple or multiple regression, quality control charts, etc.) rely upon the assumption of normality, often together with homoscedasticity and independence. Take for instance the context of plant breeding mentioned above, where it is assumed that the continuous distribution of quantitative traits (crop yield, plant height, etc.) will be a normal distribution. However, many studies report that the distributions of quantitative traits in crops deviate from normality (Hennessy, 2009; Limpert and Stahel, 2011) and exhibit both skewness, either positive (Chen and Miranda, 2008) or negative (Atwood et al., 2003; Ramirez et al., 2003), and nonnormal kurtosis (Day, 1965; Buccola, 1986; Moss and Shonkwiller, 1993). Although one-way ANOVA is considered a robust test against normality when the sample sizes are small and the normality assumption is violated, the results may be incorrect or misleading (Glass et al., 1972). In the case that the assumption of normality is not satisfied, the researcher applies an appropriate data transformation, which often results in contradictory conclusions (Stroup, 2014), or uses non-parametric or distribution-free tests, which have less statistical power than parametric tests (Dixon, 1954).

Thus, for researchers and practitioners in various often diverse fields, a need arises to validate the assumption of normality in order to ensure the appropriateness of the selected statistical technique as well as the accuracy of the results obtained. This validation can be explored with diagnostic plots such as Q-Q plots, box-plots and histograms. These diagnostic plots are useful, but although ways of choosing among them are available in the literature (see e.g. Atkinson and Riani, 2012; Fox, 1991), some expertise that comes with practice, knowledge and experience is required for their interpretation. Most of the time, statistical tests are used to confirm conclusions based on visual inspection of graphical methods. It should be noted, though, that limitations and/or disadvantages are not uncommon in their implementation. Indeed, for instance, although statistical tests have the advantage of making an objective judgment of normality most of the time, they have the disadvantage of being frequently insensitive in situations with

small sample sizes or oversensitive in cases with large sample sizes. It turns out that in at least some of such cases one can identify graphical tools having the advantage of allowing relatively better judgment to assess normality than statistical tests, which in turn implies that decisions should be made without neglecting the adequacy/efficiency of graphs (Kozak and Piepho, 2018). At the same time statistical tests should not be overrated and statistical reasoning should be exercised in all instances, not only for choosing the most appropriate test, but also for choosing between a statistical or a graphical tool for assessing normality.

There are plenty of normality tests, with different assumptions and applications, available in the literature, and their goodness of fit properties have been examined by many researchers (Seier, 2002; Yazici and Yolacan, 2007; Krauczi, 2009; Romao et al., 2010; Yap and Sim, 2010; Adefisoye et al., 2016; Islam, 2019). Hence, the researcher faces a fundamental problem, namely how to choose the most suitable test for his/her dataset.

This paper has as its purpose the comparison of various statistical tests using various types of simulation data sets and identification of how these types of data influence the selection of the optimal normality test. We also wish to explore which are the conditions and the assumptions that convert every test separately into the most powerful normality test for agriculture datasets. For illustrative purposes two real examples on experimental data analysis, set in the ANOVA context, are presented to demonstrate, in practice, similarities and dissimilarities among the statistical tests used in this comparative study for assessing normality.

2. Materials and methods

The normality tests evaluated in this study can be classified into three main categories. From the first category, the empirical distribution function (EDF) tests, the Lilliefors (LL), Cramer von Mises (CvM) and Anderson–Darling (AD) tests were evaluated. The second category includes regression and correlation tests, from which the Shapiro–Wilk (SW) and Shapiro–Francia (SW) tests were evaluated. The third category includes moment tests, from which the Skewness (SK), Kurtosis (KU), D'Agostino–Pearson Omnibus (DA), Jarque Bera (JB) and Adjusted Jarque Bera (JBadj) tests were evaluated.

Since a theoretical comparison is not feasible, a simulation procedure was used to evaluate these normality tests in testing if a random sample

of n independent observations comes from a normal population N (μ, σ^2), where μ is the mean and σ^2 the variance.

The null and alternative hypotheses are:

H_0: The distribution is normal
H_1: The distribution is not normal

Initially 10,000 simulations were carried out, in which samples of different size (n = 10, 20, 30, 40, 50, 100, 200, 500, 1000 and 2000) were generated from the standard normal distribution. Then the empirical probability of Type I error (size), which is defined as the number of times the null hypothesis of normality is rejected divided by the total number of simulations (10,000), was evaluated.

Moreover a simulation procedure was carried out in which samples of different size (n= 10, 20, 30, 40, 50, 100, 200 and 500) were generated from five alternative symmetric distributions (Beta, Logistic, t(3), t(5) and t(10)) (Figure 1a), five asymmetric distributions (Gamma, Chi-square, Exponential, Weibull and Lognormal) (Figure 1b), and three bimodal (Figure 1c) and contaminated (Figure 1d) normal distributions. The empirical power of the test, which is calculated as the ratio of the number of times the null hypothesis is rejected over 10,000 (the number of simulations) when the alternative hypothesis of non-normality is true, was evaluated. Tukey's contaminated normal model (Tukey, 1960) was used to create the contaminated and bimodal distributions.

$$F(x) = (1-\epsilon)N(\mu_1, \sigma_1^2) + N(\mu_2, \sigma_2^2) \quad \text{and} \quad 0 < \epsilon < 1. \tag{1}$$

In addition, empirical data from two wheat evaluation experiments with negative and positive skewness in distribution were used to confirm the results of the study. The first experiment comprised eight bread wheat varieties and was laid down according to a randomized complete block design with three replicates, and the second experiment comprised five durum wheat varieties and was laid down according to a randomized complete block design with four replicates. The experiments took place at the experimental station of the Agricultural University of Athens, in the area of Copais, in the growing seasons 2011–2012 and 2014–2015.

The simulations were performed with the statistical software R 3.4, and the nortest, normtest, distr and extradistr packages were used.

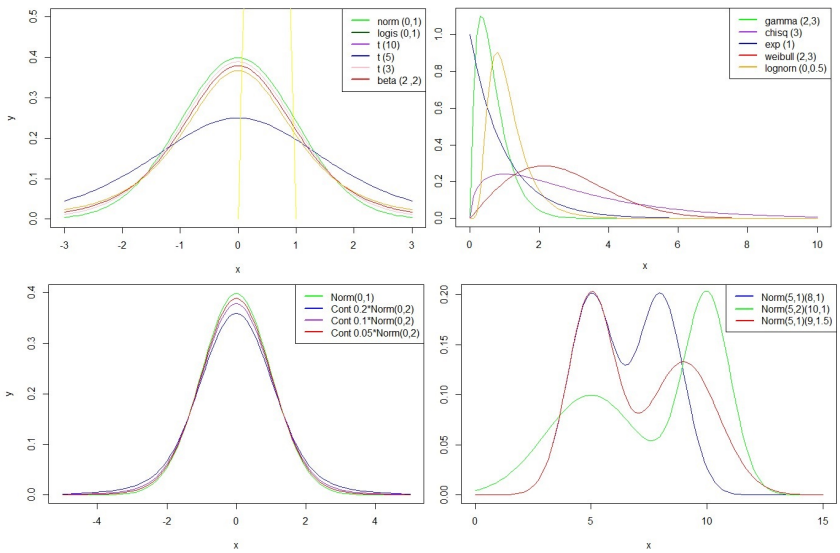

Figure 1. Probability density functions of normal and alternative distributions

3. Results

Size of the test

The empirical probabilities of type I error, for a=0.01 and 0.05, appear in Table 1. All normality tests performed well against normal distribution for all sample sizes used, except the D'Agostino–Pearson Omnibus test, which showed an increased probability of Type I error in small sample sizes.

Power of the test – Symmetric alternatives

Table 2 summarizes the empirical powers with significance level a = 0.05 for the symmetric alternate distributions. The general and expected pattern was observed that as the sample size increases the power of the test also increases. Under the alternative Beta (2, 2) distribution, which is a short-tailed symmetric distribution, the Kurtosis test had the highest power for small sample sizes and the D'Agostino–Pearson Omnibus test had the highest power for moderate and large sample sizes, followed by the Kurtosis and Shapiro–Wilk tests.

In the case of the t(10) (Student-t test with 10 degrees of freedom) distribution, which is a symmetric distribution with a slightly higher kurtosis than the normal, all moment and correlation-regression tests had the great-

Table 1. The empirical probability of type I error of tests for normality

N(0, 1) α=0.01										
N	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	1.5	1.2	1	0.8	0.7	1.1	1.4	2.4	1.1	0.7
20	1.1	0.9	1.1	0.7	1.5	0.7	1	2.2	0.9	1.2
30	1.1	0.8	1.3	1.6	0.9	0.5	1.2	1.6	0.3	1.5
40	1.8	1.5	1	1.3	1.1	1.1	0.7	2.5	0.6	1.3
50	1.2	1.2	1.6	0.6	1	0.7	1.3	1.8	1	0.8
100	1.2	1.1	1.2	0.8	1.3	1.2	0.6	1.4	0.9	1.5
200	1.4	1	1.2	1.2	1.1	1	1.1	1.3	1	0.9
500	0.8	0.8	1.3	1.2	1	0.8	1.5	1.4	1	1.2
1000	0.8	1.1	0.7	1	1.4	0.6	0.8	1.5	0.7	0.7
2000	0.9	1.5	0.9	1.1	1.1	1.5	1	0.8	0.7	0.7
N(0, 1) α=0.05										
N	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	5.4	4.7	4.1	4.8	4.5	4.5	5.6	6.2	5.2	5.4
20	4.2	5.9	6	4.3	4.5	5.1	5	5.2	4.7	6.1
30	5.1	5.7	5.3	5.1	5.2	6.4	5.3	6.3	4.6	3.9
40	5.1	5.7	5.8	6	6	6.1	6.5	5.9	4.2	5.8
50	5.2	4.8	4.2	5.1	5	3.9	5	4.2	5.3	5.2
100	4.5	4.5	5.4	4	4.3	5.3	5.3	5.3	4.6	4.8
200	5.1	5.6	4.9	5.1	5.1	3.6	4.1	5.1	4.6	6.1
500	5.4	4.8	5.2	4.9	5.8	5.5	4.8	4.4	4.9	3.9
1000	5.3	4.3	5	5.5	5.9	5.3	4.9	4.7	5.4	5.2
2000	4.2	3.5	3.7	5.6	4.3	5.4	4.5	4.3	4.6	5.3

est power for small sample sizes, and the D'Agostino–Pearson Omnibus test had the highest power for moderate and large sample sizes, followed by the Jarque Bera test.

Under the logistic (5, 2) distribution, which is also a symmetric distribution with a slightly higher kurtosis than the normal, all moment tests had the highest power for small sample sizes, followed by correlation-regression tests, and the Kurtosis test had the highest power for moderate and large sample sizes, followed by the Adjusted Jarque Bera test.

For a t(5) distribution, which is a symmetric long-tailed distribution, the D'Agostino–Pearson Omnibus test and all moment tests had the highest power for small sample sizes, followed by correlation-regression tests, and the Kurtosis test had the highest power for moderate and large sample sizes, followed by the Adjusted Jarque Bera test.

In the situation where the alternative distribution is a t(3) distribution, which is a symmetric distribution with higher kurtosis than the normal, the D'Agostino–Pearson Omnibus test and all moment tests had the highest power for small sample sizes, followed by correlation-regression tests and the Kurtosis test, while the Shapiro–Francia and Kurtosis tests had the highest

power for moderate and large sample sizes, followed by the Adjusted Jarque Bera test.

Finally, we observe that the Skewness test fails in almost all of the symmetric distributions considered (in some more than in others).

Power of the test – Asymmetric alternatives

The empirical powers for asymmetric distributions with high skewness and kurtosis appear in Table 3. Under Weibull (2, 3) and Gamma (2, 3) distributions, the Shapiro–Wilk test had the highest power for all sample sizes, followed by the Shapiro–Francia and Skewness tests. For a Chi-squared distribution with 3 degrees of freedom, the Cramer von Mises test had the highest power for all the sample sizes, followed by the Shapiro–Wilk and Lilliefors tests. Observe that the Kurtosis test appears to have the worst performance among all tests considered.

The power results for the Lognormal (0, 0.5) distribution show that the Shapiro–Francia and Shapiro–Wilk tests had the highest power for all sample sizes, followed by the Skewness test. Under the Exponential (1) distribution, the Shapiro–Wilk and Shapiro–Francia tests had the highest power for the entire range of sizes considered.

Power of the test – bimodal and contaminated alternatives

The empirical powers for bimodal distributions are presented in Table 4. Under bimodal distributions with negative and positive skewness, the Shapiro–Wilk and Shapiro–Francia tests had the highest power for all sample sizes, followed by the Skewness test. For a bimodal distribution being a mixture of two normal distributions with the same variance but different means, the Shapiro–Wilk and Shapiro–Francia tests had the highest power for all sample sizes, followed by the Anderson–Darling and Skewness tests. Observe that the Kurtosis test is less effective than all others for all of the bimodal distributions examined.

The empirical powers for contaminated normal distributions appear in Table 5. Under a contaminated distribution with high kurtosis, the D'Agostino–Pearson Omnibus and Kurtosis tests had the highest power for moderate and large sample sizes, followed by the Skewness test.

Under a contaminated distribution with medium kurtosis, the D'Agostino–Pearson Omnibus and Kurtosis tests had the highest power for small sample sizes, while the Kurtosis, Jarque Bera and Adjusted Jarque Bera tests had the highest power for moderate and large sample sizes. When the contamination is low the D'Agostino–Pearson Omnibus test had the highest power

Table 2. The empirical powers (a = 0.05) for symmetric alternative distributions

Beta (2, 2) Sk=0, Ku=0.82										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	3.1	4.9	4.4	5	2.6	1.8	6.4	2.7	2.5	1.7
20	3.6	5.1	5.8	4.3	2.5	0.9	7.2	3	0.8	0.2
30	4.7	7.4	7.2	7.9	2.4	0.3	8.3	8.6	0	0.2
40	6.9	10.4	7.4	11.3	4.2	0.2	11.3	15.4	0	0
50	8	14.1	10.3	15.9	5	0.3	17.8	25.3	0.1	0
100	14.5	29.8	25.3	45.2	22.4	0.3	55.6	64.9	4	1.1
200	35.7	69.7	56.8	92.2	75.8	0.2	95.9	96.3	67.2	56.3
500	83.9	99.9	97.8	100	100	0	100	100	100	100
t(10) Sk=0, Ku=0.82										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	5.4	6.4	6.5	6.1	8.5	8.4	5.6	8.6	6.4	8.2
20	6.8	8.2	7.4	11.5	11.6	10.9	11.6	11.8	12.9	12.5
30	8.1	10.6	9.8	13.7	15	13.3	15	15.9	14.2	14.9
40	8.4	11	9.1	12.3	17	13.4	19.3	16.5	19	17.4
50	9.4	12.8	13	14	17.4	17.4	19.7	20	20.1	21.4
100	11.4	16.6	17.1	23.9	28.7	18.6	31.3	30.4	25.5	29.2
200	14.5	23.3	20.3	37.1	40.3	19.3	47.2	43.3	44.9	46.1
500	27.8	50.2	40.4	65.2	71.6	23.6	75.7	68	73.8	72.5
Logistic (5, 2) Sk=0, Ku=-1.1										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	7.6	8	6.9	7.1	9.2	8.1	5.5	10.3	9.8	9.3
20	7.2	10.9	11	11.9	13	15.0	14.4	15.2	15.4	15.1
30	10.2	11.8	9.8	14.5	16.6	15.7	16	18.6	19	17.9
40	9.2	13.8	13.4	16.8	20.2	14.9	23.7	20.8	20.9	21.2
50	10.6	15.3	15.1	17	23	18.9	26.4	23.1	24.8	24.4
100	15.1	22	21.1	29.4	37.8	20.8	39.5	33.8	38.2	39.3
200	27.1	39.4	35.5	48.9	54.9	25.4	61.5	50	58.6	61.3
500	52.8	76.8	70.3	83.7	87.3	28	91.8	85.6	88	90.4
t(5) Sk=0, Ku=-2.5										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	8.9	10.5	9.7	10.4	11.9	13.1	9.3	14.7	14.6	13.8
20	12.7	16	15.5	18.6	21.6	20.7	21	23.5	20.9	21.4
30	15.3	22.2	19.4	28.2	28.1	24.9	28.7	31.9	31.4	33.5
40	19.5	27.4	23.4	28.5	36.1	27.9	37.2	37.1	37.2	38.9
50	21.4	29.3	26.4	34.2	40.9	30.3	43.1	41.2	41.3	41.7
100	33	48.1	41.5	57.4	63.9	43.5	65.7	60.3	62.8	67.2
200	54.9	73.6	67.1	82.1	85.4	45.3	87.5	82.2	86.5	86.2
500	88.3	97.7	96	99	99.4	55.7	99.8	98.7	99.3	99.9
t(3) Sk=0, Ku=-7.5										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	16.2	22.3	17	19.1	20.6	19.4	15.5	24	23.3	23.3
20	24.1	31.4	27.9	32.2	36.9	33.1	37.6	37.5	37.9	38.3
30	36.6	44.8	39.3	47	49.6	41.2	49.5	50.2	52.6	54.2
40	40.8	52.9	50.6	53.6	57.4	45.9	64.4	59.5	58.7	62.7
50	49.9	59.2	54.9	61.6	70.7	50.3	71	66.7	69.1	70.7
100	73.4	83.5	82.6	88.3	92.6	61.6	91.8	86.6	89.6	90
200	93.4	98.7	97.7	98.8	99.4	73.2	98.8	98.6	99	99.4
500	100	100	100	100	100	80.2	100	100	100	100

Table 3. The empirical powers (a = 0.05) for asymmetric distributions

Weibull (2,3) Sk=-2.7, Ku=-12										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	5.7	7.5	7.7	8.5	8.5	7.7	7.3	9.6	8.4	7
20	8.4	13.3	12	13.6	14.4	16.2	8.1	12.6	11.2	11.2
30	13	17.4	16.3	25.1	17.9	20.8	9.6	15.7	17	14.6
40	17.8	22.5	22	30.1	27.4	30.1	11.4	25.9	21.2	20.4
50	20.2	29.1	27.9	42.4	35.1	35.8	13.1	26.4	24	23.6
100	39	60	50.3	77.8	74.2	68.4	14.6	56.1	53.4	47.1
200	68.3	93.8	84.8	99.5	99	95.5	17.5	94.6	95	93.1
500	98.2	100	100	100	100	100	29.2	100	100	100
Gamma (2,3) Sk=1.3, Ku=-2.6										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	16.9	22.6	19.2	23.8	22.5	21.6	12.4	21.1	19.1	17.4
20	30.9	46.3	45.1	52.3	50.9	46.5	23.5	37.1	41.1	39.1
30	46.4	66.4	60.2	75.5	70.1	69.7	34	56	58.2	49.8
40	59.1	79.1	74.3	87	83.5	80	42.7	69.1	69.3	66.3
50	68.3	88.8	83.2	95.5	92.1	90	48.5	79.2	80.6	78.4
100	95.5	99.8	99.1	100	99.9	99.6	72	99.4	99.5	99
200	100	100	100	100	100	100	99.5	100	100	100
500	100	100	100	100	100	100	100	100	100	100
Chisquare (3) Sk=1.5, Ku=-3.2										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	27.9	27.8	30.9	32.8	26.9	19.9	25.2	25.2	21.3	19.3
20	60.3	52.6	67.9	65	56.4	31.5	46.5	51.2	44.7	39.5
30	78.5	73.6	84	81.7	73.9	44.8	65.5	70.6	61.7	58.9
40	90.7	85.4	95.1	93.2	89.3	50.7	78	84.1	74.8	73.6
50	96.2	93.7	99.1	97.7	94.8	59.2	89.7	90.7	86.3	84.4
100	100	99.9	100	100	99.7	80.8	100	99.8	100	98.7
200	100	100	100	100	100	96.8	100	100	100	100
500	100	100	100	100	100	100	100	100	100	100
Lognormal (0,0.5) Sk=1.5, Ku=-3.7										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	17	23.2	22.7	23.5	24.9	26.4	18.1	24.8	23.2	18.3
20	31.6	47.5	41.8	51	51.2	48	28.4	44.7	43.9	39.3
30	45.9	65.1	60.9	71	70.7	65.5	40.5	59.8	61.8	58.3
40	61.2	78.2	74.1	84.9	82.6	81.2	49.4	72.5	73.6	71.4
50	67.6	88.4	81.4	91.3	92.7	87	58	83	82.3	79.7
100	94.7	99.4	98.7	99.7	99.7	99.9	80	98.9	98.6	99.1
200	100	100	100	100	100	100	95.9	100	100	100
500	100	100	100	100	100	100	100	100	100	100
Exponential (1) Sk=1.8, Ku=-4.6										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	16.2	22.3	17	19.1	20.6	19.4	15.5	24	23.3	23.3
20	24.1	31.4	27.9	32.2	36.9	33.1	37.6	37.5	37.9	38.3
30	36.6	44.8	39.3	47	49.6	41.2	49.5	50.2	52.6	54.2
40	40.8	52.9	50.6	53.6	57.4	45.9	64.4	59.5	58.7	62.7
50	49.9	59.2	54.9	61.6	70.7	50.3	71	66.7	69.1	70.7
100	73.4	83.5	82.6	88.3	92.6	61.6	91.8	86.6	89.6	90
200	93.4	98.7	97.7	98.8	99.4	73.2	98.8	98.6	99	99.4
500	100	100	100	100	100	80.2	100	100	100	100

Table 4. The empirical powers (a = 0.05) for Bimodal distributions

Bimodal 0.5xN(5,2)+0.5xN(10,1) Sk=–.43, Ku=–.96										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	5.7	7.5	7.7	8.5	8.5	7.7	7.3	9.6	8.4	7
20	8.4	13.3	12	13.6	14.4	16.2	8.1	12.6	11.2	11.2
30	13	17.4	16.3	25.1	17.9	20.8	9.6	15.7	17	14.6
40	17.8	22.5	22	30.1	27.4	30.1	11.4	25.9	21.2	20.4
50	20.2	29.1	27.9	42.4	35.1	35.8	13.1	26.4	24	23.6
100	39	60	50.3	77.8	74.2	68.4	14.6	56.1	53.4	47.1
200	68.3	93.8	84.8	99.5	99	95.5	17.5	94.6	95	93.1
500	98.2	100	100	100	100	100	29.2	100	100	100
Bimodal 0.5xN(5,1)+0.5xN(9,1.5) Sk=.27, Ku=–.96										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	16.9	22.6	19.2	23.8	22.5	21.6	12.4	21.1	19.1	17.4
20	30.9	46.3	45.1	52.3	50.9	46.5	23.5	37.1	41.1	39.1
30	46.4	66.4	60.2	75.5	70.1	69.7	34	56	58.2	49.8
40	59.1	79.1	74.3	87	83.5	80	42.7	69.1	69.3	66.3
50	68.3	88.8	83.2	95.5	92.1	90	48.5	79.2	80.6	78.4
100	95.5	99.8	99.1	100	99.9	99.6	72	99.4	99.5	99
200	100	100	100	100	100	100	99.5	100	100	100
500	100	100	100	100	100	100	100	100	100	100
Bimodal 0.5xN(5,1)+0.5xN(8,1) Sk=0, Ku=–.94										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	19.3	27.9	27.8	30.9	32.8	26.9	19.9	25.2	25.2	21.3
20	39.5	60.3	52.6	67.9	65	56.4	31.5	46.5	51.2	44.7
30	58.9	78.5	73.6	84	81.7	73.9	44.8	65.5	70.6	61.7
40	73.6	90.7	85.4	95.1	93.2	89.3	50.7	78	84.1	74.8
50	84.4	96.2	93.7	99.1	97.7	94.8	59.2	89.7	90.7	86.3
100	98.7	100	99.9	100	100	99.7	80.8	100	99.8	100
200	100	100	100	100	100	100	96.8	100	100	100
500	100	100	100	100	100	100	100	100	100	100

for small and moderate sample sizes, followed by Adjusted Jarque Bera, which together with the Kurtosis test is powerful for moderate and large sample sizes.

Experimental data analysis

In this section we present, for illustrative purposes, the analysis of two experiments that took place at the Copais experimental station of the Agricultural University of Athens, Greece. The data refer to the growing seasons 2011–2012 and 2014–2015.

Regarding the empirical data of the experiments, the RCBD model was considered, the associated residuals were calculated, and their distribution, accounting for both treatments and blocks, was tested with diagnostic tools as well as normality tests. Diagnostic tools (histogram and QQ plot) indicate

Table 5. The empirical powers (a = 0.05) for Contaminated distributions

Contaminated Sk=0, Ku=-1.36										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	7.2	10	7.6	10.1	12	10.8	8.2	12.8	11.4	11.6
20	9.6	13	13.1	13	15.8	16	17	19.8	18.5	19.6
30	10.6	14.9	13.7	19.5	22.9	21.8	26.2	22	26.4	25.4
40	12.8	19.8	17.4	25.7	30.2	24.2	30.6	28.9	28.2	27.8
50	12.4	23.1	18	26.7	35.2	28.9	35.3	31.5	35.4	36.6
100	19.7	33.8	27.9	47.1	52.4	29.9	56	49.1	53.7	57.2
200	30.6	54	43.9	67.2	77.4	32.8	79.5	73	79.7	77
500	64.4	88.7	80.7	96.1	97.1	37.3	98.3	96.9	97.3	97.8
Contaminated Sk=0, Ku=1.16										
n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
10	5.7	7.7	8.7	8	10.9	8.1	5.5	11.3	8.2	8.9
20	6.8	11.2	9.7	10.7	10.2	13.4	13.9	15	14.3	13.4
30	9.2	10.4	11.3	14.8	18.9	21	19.1	19.3	17	20.3
40	9.5	13	11	18.5	21.7	18.7	24.8	23.3	25.7	20.5
50	11.1	16.3	11.8	23.7	24.1	21.2	26	25.2	29.6	28.8
100	12.5	20.5	16	33.7	38.1	25	42.6	38.3	42.6	44
200	15.8	28.8	22.2	48.4	61.8	30.6	61.2	58.3	61.1	59.2
500	31.2	56	43.4	83.3	87.9	35.6	89.1	86.3	87.8	88.6
Contaminated Sk=0, Ku=0.7										
n	LL	AD	CvM	SW	SF	SK	KU	DA	JB	JBadj
10	6.2	6.7	5.7	6.1	7.3	6.5	6.6	8.5	7.5	7.4
20	5.6	8.8	7.3	10.3	7.7	10.2	11.2	11.1	10.6	10.8
30	6	7.8	8.4	10	10.8	11.7	12.9	14.6	14.5	13.9
40	7.3	9.1	8.9	12	15.6	12.9	15	17.4	15.3	14.9
50	7.1	11.9	10.6	14	16.5	17.2	19.2	19.1	18.1	20.2
100	8	12.3	10.7	20.1	26.2	20.4	29.2	25.1	28.3	30.8
200	10.8	14.6	11.5	35	38.3	26.7	41.5	39.2	45	39.9
500	16.4	24	19	58.5	65.4	29.7	68.3	63.2	64.9	66

that the distributions of the residuals were not symmetric (Figure 2 and 3). The residual distribution for the first set of data, with sample size 20, was positively skewed (0.77) and had kurtosis coefficient 1.76, while for the second set of data, with sample size 24, the residual distribution was negatively skewed (-0.63) and the kurtosis coefficient was -0.50. The results of the normality tests are presented in Tables 6 and 7. In both cases, the moment and EDF tests show conflicting results. The Shapiro–Wilk and Shapiro–Francia tests show similar results, in not rejecting the normality of the error distribution in the first data set and marginally rejecting it in the second data set. Also, notice that the outcome of SW and SF for the second data set is aligned with the D'Agostino–Pearson Omnibus test, which is preferable in cases with a higher coefficient of kurtosis.

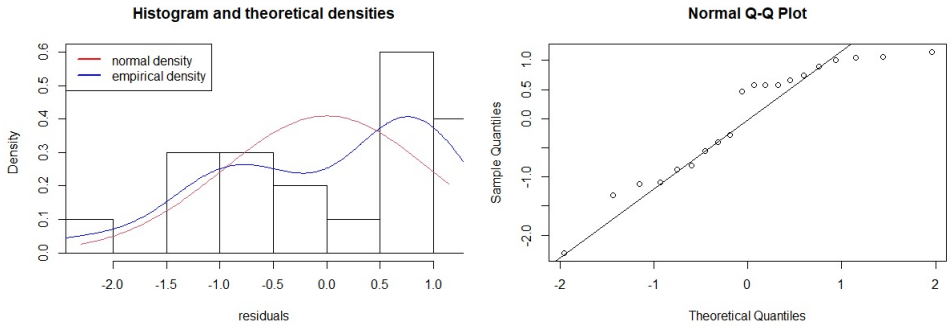

Figure 2. Histogram and QQ plot of the first experiment (n=20).

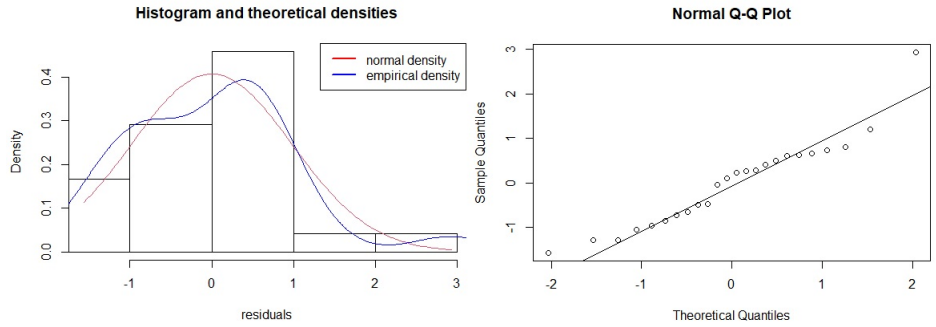

Figure 3. Histogram and QQ plot of the second experiment (n=24).

Table 6. Empirical Data (Skewness = -0.63, Kurtosis = -0.50)

n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
20	0.007	0.026	0.023	0.034	0.048	0.176	0.422	0.361	0.174	0.221
	*	*	*	*	*	ns	ns	ns	ns	ns

Table 7. Empirical Data (Skewness = 0.77, Kurtosis = 1.76)

n	**LL**	**AD**	**CvM**	**SW**	**SF**	**SK**	**KU**	**DA**	**JB**	**JBadj**
24	0.348	0.232	0.358	0.101	0.063	0.096	0.077	0.054	0.053	0.047
	ns	ns	ns	ns	ns	ns	ns	ns	ns	*

4. Conclusion

This comprehensive study evaluates the performance of ten normality tests against normal and alternative distributions for different sample sizes and significance levels. When the distribution is normal, the normality tests performed well with respect to Type I error, while in the case of alternative distributions, the power of the test depends on the type of distribution, the sample size and the significance level. By increasing both the sample size and the significance level, the power of these tests also increases. All the moment tests, expect the Skewness test, are the most powerful against symmetric distributions, followed by the Shapiro–Wilk test. The test that stands out among the moment tests is the D'Agostino–Pearson Omnibus test, which is the most powerful for small samples for both symmetric and contaminated distributions. The Skewness test, unlike other moment tests, is powerful only for asymmetric distributions. The empirical distribution function tests are suitable for asymmetric distributions, but the Shapiro–Wilk test is the one that produces most accurate results in this class of distributions, followed by the Skewness test. The most powerful of the EDF tests is the Anderson–Darling test, followed by Cramer von Mises' test, while the Lilliefors test had the lowest power.

All in all, researchers must combine descriptive estimators and diagnostic plots and exercise statistical reasoning in order to identify the distribution of the data and to select the appropriate normality test. This is required because otherwise structural problems of overuse of statistical tests arise (Kozak and Piepho, 2018). According to the results of the extensive simulation study in the previous section, if the researcher suspects that the distribution is symmetric with moderate or high kurtosis coefficient, and the sample size is small, then the recommended test is D'Agostino–Pearson Omnibus, while for larger sample sizes the Kurtosis test is recommended. In all other cases the recommended test is either Shapiro–Wilk or Shapiro—Francia, both of which are available in most statistical software. The findings of this paper are in good agreement with Yap and Sim (2011).

The conclusions of this study may prove very helpful in the analysis of data from quantitative trait experiments in the future, particularly for plant breeding purposes.

References

Adefisoye J.O., Golam Kibria B.M., George F. (2016): Performances of Several Univariate Tests of Normality: An Empirical Study. Journal of Biometrics & Biostatistics 7: 322.

Atwood J, S.haik S., Watts M. (2003): Are Crop Yields Normally Distributed? A Reexamination. American Journal of Agricultural Economics 85: 888–901.

Atkinson A.C., Riani M. (2012): Robust diagnostic regression analysis. New York: Springer Science & Business Media.

Buccola S.T. (1986): Testing for Nonnormality in Farm Net Returns. American Agricultural Economics Association: 334–343.

Chen S., Miranda M. (2008): Modeling Texas Dryland Cotton Yields, With Application to Crop Insurance Actuarial Rating. Journal of Agricultural and Applied Economics 40(1): 239–252.

Day R.H. (1965): Probability Distributions of Field Crop Yields. Journal of Farm Economics 47(3): 713–741.

Dixon W. (1954): Power under normality of several nonparametric tests. Annals of Mathematical Statistics 25: 610–614.

Fox J. (1991): Regression diagnostics: An introduction, Vol. 79. Newbury Park, CA: Sage.

Glass G.V., Peckham P.D., Sanders J.R. (1972): Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. Review of Educational Research 42(3): 237–288.

Hennessy D. (2009): Crop Yield Skewness and the Normal Distribution. Journal of Agricultural and Resource Economics 34(1): 34–52.

Islam T.U. (2017): Stringency-based ranking of normality tests. Communications in Statistics - Simulation and Computation 46: 655–668.

Jarque C.M., Bera A.K. (1987): A Test for Normality of Observations and Regression Residuals. International Statistical Review 55: 163–172.

Kozak M., Piepho H-P. (2018): What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. Journal of Agronomy and Crop Science 204: 86–98.

Krauczi E. (2009): A study of the quantile correlation test for normality. TEST 18: 156–165.

Limpert E., Stahel W.A. (2011): Problems with Using the Normal Distribution – and Ways to Improve Quality and Efficiency of Data Analysis. PLoS ONE 6(7): e21403.

Moss C.B., Shonkwiler J.S. (1993): Estimating Yield Distributions with a Stochastic Trend and Nonnormal Errors. American Journal of Agricultural Economics 75(4): 1056—1062.

Stroup W.W. (2015): Rethinking the Analysis of Non-Normal Data in Plant and Soil Science. Agronomy Journal 107: 811–827.

R Core Team R (2020): A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ramirez O.A., Misra S. (2003): Field J. Crop-Yield Distributions Revisited. American Journal of Agricultural Economics 85: 108—120.

Romao X., Delgado R., Costa A. (2010): An empirical power comparison of univariate goodness-of-fit tests for normality. Journal of Statistical Computation and Simulation 80: 1–47.

Seier E. (2002): Comparison of Tests for Univariate Normality. InterStat Statistical Journal 1: 1–17.

Tukey J.W. (1960): A survey of sampling from contaminated distributions. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, Stanford Univ. Press: 448–485.

Yap B.W., Sim C.H. (2011): Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation 81: 1–15.

Yazici B., Yolacan S. (2007): A comparison of various tests of normality. Journal of Statistical Computation and Simulation 77: 175–183.