Graphical Comparison of Normality Tests for Unimodal Distribution Data

José A. Sánchez-Espigares
josep.a.sanchez@upc.edu
Department of Statistics and Operations Research,
Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona, 08028, Spain

Pere Grima
pere.grima@upc.edu
Department of Statistics and Operations Research,
Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona, 08028, Spain

Lluís Marco-Almagro
lluis.marco@upc.edu
Department of Statistics and Operations Research,
Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona, 08028, Spain

Graphical Comparison of Normality Tests for Unimodal Distribution Data

José A. Sánchez-Espigares, Pere Grima, Lluís Marco-Almagro
Department of Statistics and Operations Research,
Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona, 08028, Spain

ABSTRACT

A methodology is proposed to compare the power of normality tests with a wide variety of alternative unimodal distributions. It is based on the representation of a distribution mosaic in which kurtosis varies vertically and skewness horizontally. The mosaic includes distributions such as exponential, Laplace or uniform, with normal occupying the center. Simulation is used to determine the probability of a sample from each distribution in the mosaic being accepted as normal. We demonstrate our proposal by applying it to the analysis and comparison of some of the most well-known tests.

KEYWORDS: Goodness-of-Fit, Test Power, Graphical Techniques, Comparing Normality Tests, Anderson-Darling, Shapiro-Wilk

1 Introduction

There are a wide variety of normality tests, from the classic Kolmogorov-Smirnov to the widely used Shapiro-Wilk and Anderson-Darling tests. Books are dedicated exclusively to normality tests, such as Thode [1], who describes dozens of them. The problem is not yet closed, as new tests and modifications of existing ones continue to emerge (see, for example, Desgagné and de Micheaux [2]).

The fact that so many exist surely indicates that not one is better than all the others in all circumstances. The comparison of normality tests has been addressed in articles such as Farell and Rogers-Stewart [3], where 14 test types are compared to 48 possible alternative distributions and the results are presented in a table with the power of the test indicated against each alternative for a significance level of 0.10 and a sample size of n=20. Yacini and Yolocan [4] compare 12 tests against 5 alternative distributions and present the results also in a table where the power of the test is indicated according to the alternative distribution for α=0.05 and sample sizes of n=20, 30, 40 and 50. Romão *et al*. [5] present an exhaustive study in which they describe and analyze the performance of 33 normality tests against data from a wide variety of distributions, presenting the power obtained in tables but also with line charts in which the horizontal axis indicates which test was performed. In a similar manner, Yap and Sim [6] compare 8 types of tests against 9 alternative distributions, presenting the results in tables as well as graphs, where for each alternative distribution the power curves are presented according to the size of the sample.

Our work proposes a methodology for comparing the power of normality tests to a wide variety of alternative unimodal distributions in a highly visual manner and with a single graph. It is based on the representation suggested by Sánchez-Espigares *et al*. [7], who builds a distribution mosaic in which the kurtosis varies vertically and the skewness horizontally. The mosaic includes distributions such as exponential, Laplace and uniform, with normal occupying the center. Simulation is used to determine the probability of a sample from each of the distributions in the mosaic being accepted as normal.

The next section describes the tests to be compared. Next, we describe the characteristics of the distribution mosaic, how the power of the tests is represented in the mosaic and, finally, the studied tests are analyzed and compared.

2 Test Selection

To demonstrate the possibilities of the proposed procedure and also compare some of the most well-known tests, we have chosen three from each of the strategies in which normality tests can be grouped: regression tests, tests based on the empirical distribution function (EDF) and tests based on moments.

Regression tests are based on the fact that the distribution function $F(x)$ of a random variable $X \sim N(\mu, \sigma)$ is a straight line when represented on a normal probability plot (Q-Q plot). Therefore, given an ordered sample of values $x_{(1)} \cdots x_{(n)}$ with $F\left(x_{(i)}\right), \cdots F\left(x_{(n)}\right)$ values of their distribution function, points $\left(x_{(i)}, F(x_{(i)})\right)$ should align approximately according to a straight line in a Q-Q plot, and any departure from that alignment indicates the data's lack of normality. From among the tests based on this idea, we have selected:

- Shapiro-Wilk (SW) is probably the best known and most often used (a detailed description can be found, for instance, in Thode [1]). The original version [8] has some computational limitations, especially for large sample sizes. Royston [9] suggested a transformation of the original statistic that allows it to be applied to sample sizes of up to $n = 2000$ without any demand for great computational resources.
- Shapiro-France (SF) (see Thode [1]) is a variant of SW. When it appeared in 1972 [10], its main advantage was the demand for fewer computational resources than the original SW. This advantage has ceased to be of interest, especially after Royston's contributions to the SW test;

but it continues to be among the most representative of this group of tests that are based on correlation.

- Filliben [11] uses the correlation between the sample order statistics and the estimated median values of the theoretical order statistics. Its main advantage is that the calculations are very easy because there is no need to calculate the expected values of the normal order statistics.

The test statistic in tests based on the EDF is a measure of the discrepancy between the EDF and the theoretical distribution function. The most typical is that of Kolmogorov-Smirnov (KS), which uses the maximum distance – in absolute value – between both distributions. The tests of this type that we analyze here are:

- Lilliefors. This uses the same test statistic as that of KS, but with a different reference distribution, due to the fact that the KS test requires knowledge of the population parameters, while Lilliefors bases its estimation on the sample. Lilliefors deduced the critical values through simulation, but analytical methods for determining them have also been published [12].

- Cramer-von Mises (CvM). The test statistic is determined from the discrepancy between the theoretical distribution function F and the empirical function F_n accumulated throughout all the variation space of x. It is specified in a relatively simple formula (see, for example, [13]). The critical values depend on the size of the sample and the number of known population parameters.

- Anderson-Darling (AD). This is surely the most commonly used of this group. It is similar to the CvM but gives more weight to the discrepancy in the tails of the distribution (see, for example, [13]).

Finally, we have the group that uses a test statistic based on the difference between, on the one hand, the kurtosis and the skewness of the data (third and fourth moment) and, on the other, their theoretical values. For this group, we have selected:

- D'Agostino-Pearson K^2 (DA). The test statistic is a function of the kurtosis and skewness of the sample. It follows a Chi-square distribution with 2 degrees of freedom if the hypothesis of normality is true [14].

- Jarque-Bera (JB). Conceptually similar to the one above. The test statistic is also calculated from the kurtosis and skewness of the sample. Thus, it is also distributed as a Chi-square with 2 degrees of freedom; but when $n < 2000$, the p-value is determined by simulation [15].

- Adjusted Jarque-Bera (AJB). It uses a new test statistic computed from the first four moments about the origin. The p-values are determined by simulation [16].

To apply these tests and analyze their performance, we have used functions that have already been developed and implemented in R statistical software packages [17]. Table 1 indicates which package and function were used for each test.

Table 1: Tests analyzed and the R packages and functions that were used to apply them.

Test	Package	Function
Shapiro-Wilk	`stats`, R Core Team [17]	`shapiro.test(x)`
Shapiro-Francia	`nortest`, Gross and Ligges [18]	`sf.test(x)`
Filliben's	`ppcc`, Pohlert [19]	`ppccTest(x, “qnorm”)`
Lilliefords	`nortest`, Gross and Ligges [18]	`lillie.test(x)`
Cramer-von Mises	`nortest`, Gross and Ligges [18]	`cvm.test(x)`
Anderson-Darling	`nortest`, Gross and Ligges [18]	`ad.test(x)`
D'Agostino-Pearson K_s^2	`fBasics`, Wuertz et al. [20]	`dagoTest(x)`
Jarque-Bera	`normtest`, Gavrilov and Pusev [21]	`jb.norm.test(x)`
Adjusted Jarque-Bera	`normtest`, Gavrilov and Pusev [21]	`ajb.norm.test(x)`

3 Distribution mosaic. Representation of the power of a test

Based on a Skewed Exponential Power Distribution (SEPD) used by Zhu and Zinde-Walsh [22], Sánchez-Espigares *et al.* [7] consider the probability density function that is used to create the mosaic distributions. This function is characterized by the mean and variance of the variable considered and also a third parameter, p, which is related to kurtosis and varies between 1 (double exponential distribution) and 50 (practically a uniform distribution). It also employs a fourth parameter, α, which is related to the asymmetry that varies between 0 (very asymmetric distribution with tail to the right) and 1 (with tail to the left). The values $p = 2$ and $\alpha = 0.5$ correspond to a normal distribution.

We want the number of distributions on each side of the mosaic to be odd so that the normal distribution remains exactly in the center. It is easily deduced that for α to vary between 0 and 1 in equidistant intervals and for $\alpha = 0.5$ to remain in the center, it is sufficient that the i-th position has the value $\alpha = \frac{i-1}{m-1}$, with m being the number of distributions on each side of the mosaic. Regarding the values of p, their determination is not so immediate. Each value must be equal to the previous one raised to a power of $j = \sqrt[\frac{m}{2}-0.5]{\frac{\log 50}{\log 2}}$, except for the second one, which always equals $2^{1/j^{\left(\frac{m}{2}-1.5\right)}}$ [7]. For example, if the mosaic is of size 11x11 ($m = 11$), we have $j = 1.4136$ and the values of α and p will be:

i	1	2	3	4	5	6	7	8	9	10	11
α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
p	1	1.19	1.28	1.41	1.63	2	2.66	3.99	7.08	15.92	50

Figure 1 shows an 11x11 mosaic. The values of μ and σ are the same for all distributions and do not affect their shape but only the scale of the axes, which we do not consider here. Each distribution corresponds to the values of α and p that are indicated. In [7], the R code is included to create mosaics as large as 49x49, although it can easily be changed to obtain larger mosaics.

From each of the distributions that appear in the mosaic, a random sample of size n can be obtained and contrasted against the normal distribution by means of the test whose power we want to analyze. We chose, for example, the Anderson-Darling test and generate 10000 samples of size n=100 from each of the distributions that appear in the mosaic of Figure 1. We consider that the hypothesis of normality is not rejected if the p-value obtained is greater than 0.05; and we annotate on each distribution the proportion of times that the hypothesis of normality would not be rejected with samples from that distribution. The values obtained are indicated in Figure 2. The area is outlined for the distributions with values of this proportion greater than 0.5. In this figure, it can be observed that, if the population from which the sample comes is exponential, the probability of not rejecting the hypothesis of normality is practically null with a sample size of n=100 when applying the AD test. The probability of not rejecting is approximately 18% if the sample comes from a Laplace distribution and around 5-6% if it comes from a uniform distribution.

Naturally, a larger mosaic can be constructed. Figure 3 shows one with 101 distributions on each side, with curves outlining the distributions in which normality is not rejected for the proportion of times indicated. The thickest line corresponds to the proportion p=0.5. The appearance of these curves can also be compared when the sample size is varied.

Figure 4 shows the curves that delimit the areas in which normality is not rejected with a probability of 50%, depending on the sample size, which is indicated on the curve itself. The curve that corresponds to $n = 10$ does not appear in the figure because any of the mosaic distributions for that sample size would be accepted with a probability greater than 50%. If the distribution from which the data come is uniform, a sample of $n = 20$ observations will also result in a greater than 50% probability of not rejecting the hypothesis of normality.

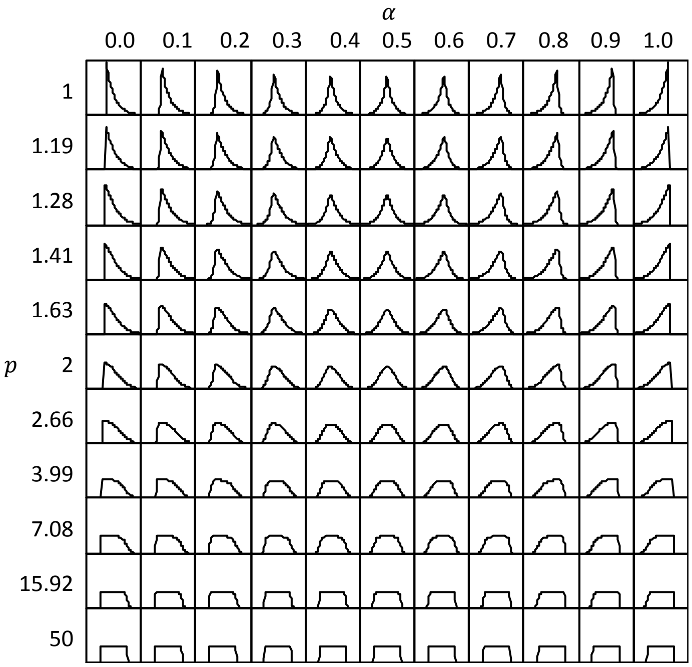
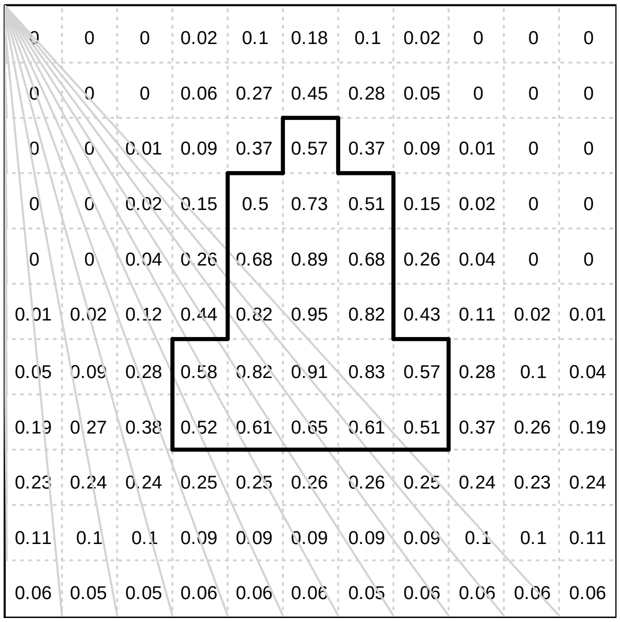


Figure 1: Mosaic of 11x11 distributions with the p and α values that correspond to each one



0	0	0	0.02	0.1	0.18	0.1	0.02	0	0	0
0	0	0	0.06	0.27	0.45	0.28	0.05	0	0	0
0	0	0.01	0.09	0.37	0.57	0.37	0.09	0.01	0	0
0	0	0.02	0.15	0.5	0.73	0.51	0.15	0.02	0	0
0	0	0.04	0.26	0.68	0.89	0.68	0.26	0.04	0	0
0.01	0.02	0.12	0.44	0.82	0.95	0.82	0.43	0.11	0.02	0.01
0.05	0.09	0.28	0.58	0.82	0.91	0.83	0.57	0.28	0.1	0.04
0.19	0.27	0.38	0.52	0.61	0.65	0.61	0.51	0.37	0.26	0.19
0.23	0.24	0.24	0.25	0.25	0.26	0.26	0.25	0.24	0.23	0.24
0.11	0.1	0.1	0.09	0.09	0.09	0.09	0.09	0.1	0.1	0.11
0.06	0.05	0.05	0.06	0.06	0.06	0.05	0.06	0.06	0.06	0.06

Figure 2: The box corresponding to each distribution indicates the proportion of times that the hypothesis of normality is not rejected when applying the Anderson-Darling test to samples of size n=100. The area where this proportion is greater than 0.5 is outlined.

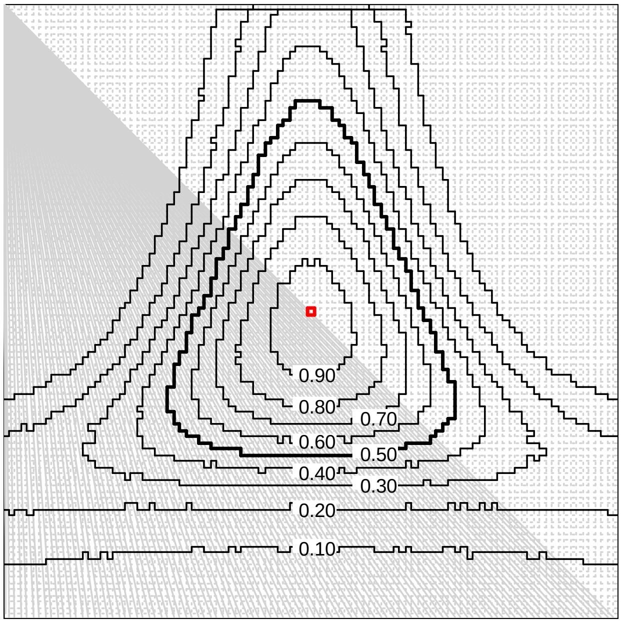


Figure 3: Curves on a 101x101 mosaic that indicate the proportion of times that the hypothesis of normality is not rejected when applying the Anderson-Darling test to samples of size n=100.

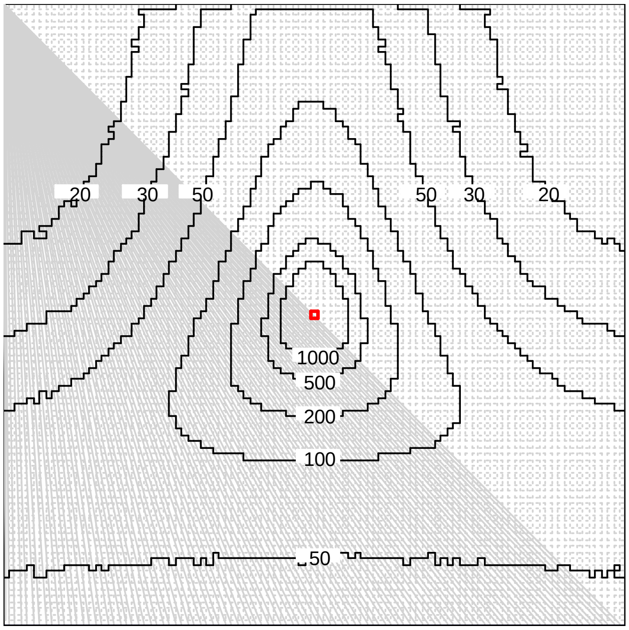


Figure 4: Curves that delimit the distributions for which normality is not rejected more than 50% of the time with the Anderson-Darling test for the sample sizes indicated

4 Comparison of tests

As an example of the possibilities of our method, 10000 samples of $n = 20$, 50 and 100 observations have been generated from each of the distributions that appear in the mosaic. The figures indicate the curves that enclose the distributions for which the hypothesis of normality is not rejected more than 50% of the time. The smaller the surface this curve encloses, that is, the fewer distributions it includes, the better the test is. This is because it is then more likely to reject the hypothesis of normality of a greater number of distributions that are not actually normal (those that are outside the curve).

Looking at the distributions in group 1 (Figure 5), based on correlation and regression measures, we observe that the Shapiro-France and Filliben tests have practically identical performance. With sample sizes of $n = 20$, the curve corresponding to the Shapiro-Wilk test is not very different either. For $n = 50$, and much more clearly for $n = 100$, the Shapiro-Wilk test shows greater power with regard to low kurtosis distributions, since its curve moves away from the lower zone. However, it extends somewhat further into the zone of symmetrical distributions with high kurtosis (towards the Laplace distribution). Overall, and taking the area enclosed by the curves as a measure of the performance of the test, we can say that SW produces the best performance.

Regarding the distributions of group 2 (Figure 6), their ranking is clear. The one that performs best for the 3 sample sizes considered is AD. It is noteworthy that if the Lilliefors test is applied with sample size $n = 20$, the probability of rejecting the null hypothesis for data coming from a uniform distribution is greater than 50%, and only slightly lower if they come from an exponential distribution.

In group 3 (Figure 7), the d'Agostino test clearly performs better when the sample size is n=100, whereas Jarque-Bera test performs slighthly better for n=20. The performance of both is very similar when n=50. Keeping in mind that we generally work with small samples, Jarque-Bera test is best for us, although it is a debatable point since it depends on the sample size. Figure 8 compares those that have been considered the best from each group, and in this case it is very clear that for any of the sample sizes considered the test that performs best is Shapiro-Wilk.

As additional material to this paper, we include a file in html format created with R Markdown with the explanations and the R code to draw the curves that allow comparing the Lilliefords, Cramer-Von Misses and Anderson-Darling tests with n = 100 (Figure 6, right). The values of the probability of rejecting the hypothesis of normality are calculated separately since the computation time is long. We also include an html file, built in the same way as the previous one, with the explanations and the code to calculate those probability values. In order to be able to quickly see the result obtained by the program that draws the curves, three files are also included in txt format with the data for each one of the tests represented.

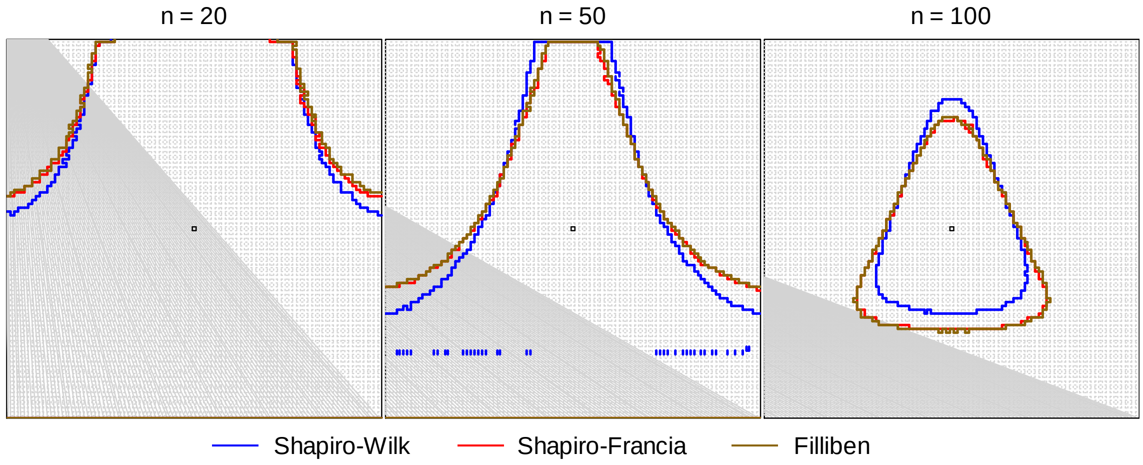


Figure 5: Tests based on correlation measures. If the sample comes from one of the distributions enclosed by the curve, the probability of not rejecting the hypothesis of normality is greater than 50%

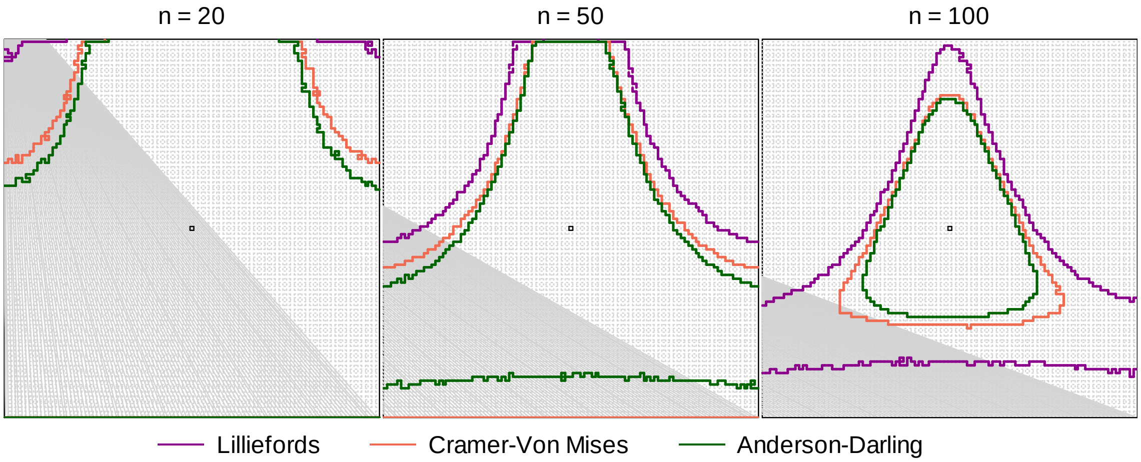


Figure 6: Tests based on EDF. If the sample comes from one of the distributions enclosed by the curve, the probability of not rejecting the hypothesis of normality is greater than 50%.

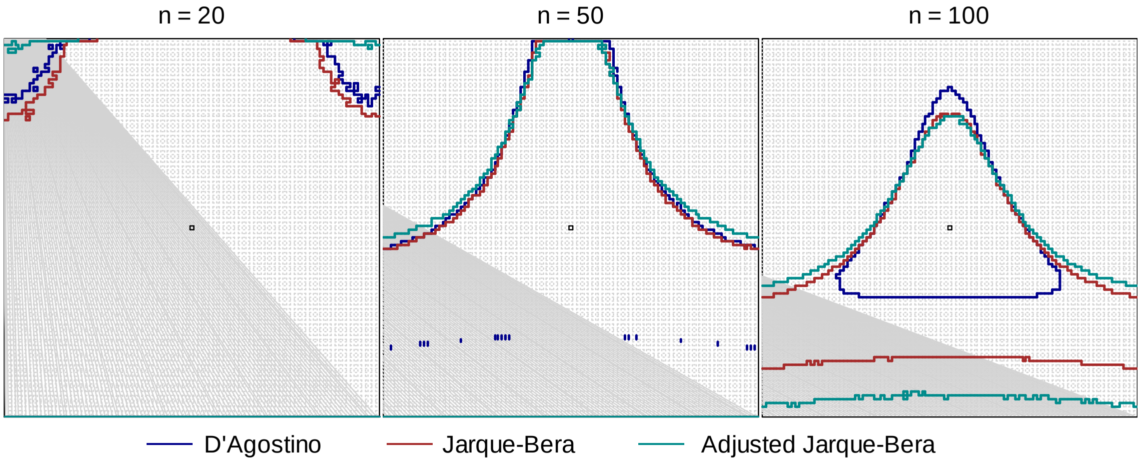


Figure 7: Tests based on moments. If the sample comes from one of the distributions enclosed by the curve, the probability of not rejecting the hypothesis of normality is greater than 50%.

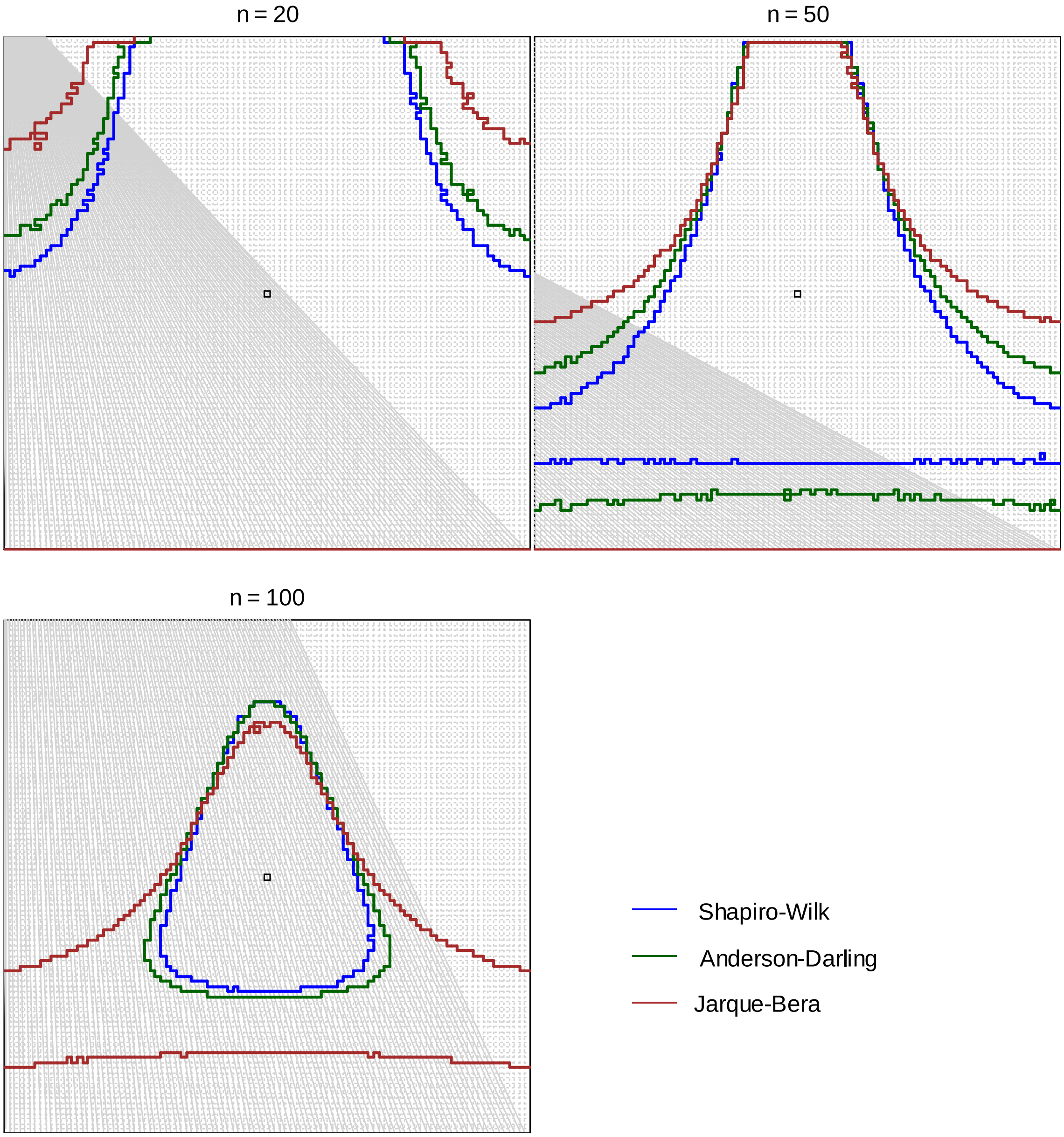


Figure 8: Comparison of the tests considered best for each group.

5 Final remarks

The proposed method allows visualizing the power of a normality test in comparison to a wide range of unimodal distributions. This procedure is especially useful for graphically comparing the power of different tests as well as the influence of the sample size.

This way of visualizing the results allows us to consider the difference between a statistically significant difference and an important difference in the context of normality tests. When a statistical method is valid only under the hypothesis of normality of the data, it is worth asking what deviation from normality is tolerable. For example, if the data belongs to one of the distributions that are next to the normal in a mosaic of 101x101 distributions, would the method be good? The answer is surely yes, but at what distance from normal would that no longer be true? And if we know the answer to this question, then what sample size is necessary for a high probability of rejecting the normality hypothesis if the data come from that distribution? Furthermore, what test performs best for that objective?

The proposed graphics naturally give rise to questions of this type, and they also allow a clearer view of the possibilities and limitations of normality tests.

References

[1] Thode, HC Jr. Testing for Normality. Marcel Dekker, Inc. New York. 2002

[2] Desgagné, A. and de Micheaux, P.L. (2017) A powerful and interpretable alternative to the Jarque–Bera test of normality based on 2nd-power skewness and kurtosis, using the Rao's score test on the APD family, Journal of Applied Statistics, DOI: 10.1080/02664763.2017.1415311.

[3] Farell, PJ and Rogers-Stewart, K (2006). Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test. Journal of Statistical Computation and Simulation. 76:9, 803-816

[4] Yazici, B and Yolacan, S. (2007). A comparison of various tests of normality. Journal of Statistical Computation and Simulation. Vol. 77, No. 2, February 2007, 175–183

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

[6] Yap, BW and Sim, CH (2011). Comparisons of various types of normality tests. Vol. 81, No. 12, December 2011, 2141–2155

[7] Sánchez-Espigares, JA, Grima, P and Marco-Almagro, Ll. (2017): Visualizing Type II Error in Normality Tests, The American Statistician, DOI: 10.1080/00031305.2016.1278035

[8] Shapiro, S.S and Wilk, M.B. (1965): An Analysis of Variance Test for Normality (Complete Samples). Biometrika. Vol. 52, No. 3/4 (Dec., 1965), pp. 591-611

[9] Royston, P. (1991): Approximating the Shapiro-Wilk W-test for non-normality. Statistics and Computing (1992) 2, 117-119.

[10] Shapiro, S.S. and Francia, R.S. "An approximate analysis of variance test for normality", Journal of the American Statistical Association 67 (1972) 215–216.

[11] Filliben, J.J. (1975): The Probability Plot Correlation Coefficient Test for Normality. Technometrics, Vol. 17, No. 1, pp.: 111-116.

[12] Dallal, G.E. and Wilkinson, L. (1986): An analytic approximation to the distribution of Lilliefors' test for normality. The American Statistician, 40, 294–296.

[13] Stephens, M.A. (1986): Tests based on EDF statistics. In: D'Agostino, R.B. and Stephens, M.A., eds.: Goodness-of-Fit Techniques. Marcel Dekker, New York.

[14] D’Agostino, Ralph B.; Albert Belanger; Ralph B. D’Agostino, Jr (1990). "A suggestion for using powerful and informative tests of normality" (PDF). The American Statistician. 44 (4): 316–321. doi:10.2307/2684359.

[15] Jarque, C.M. and Bera, A.K. (1987): A Test for Normality of Observations and Regression Residuals. International Statistical Review, 55, 2, pp. 163-172.

[16] Urzua, C. M. (1996): On the correct use of omnibus tests for normality. — Economics Letters, vol. 53, pp. 247–251.

[17] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[18] Juergen Gross and Uwe Ligges (2015). nortest: Tests for Normality. R package version 1.0-4. https://CRAN.R-project.org/package=nortest

[19] Thorsten Pohlert (2017). ppcc: Probability Plot Correlation Coefficient Test. R package version 1.0. https://CRAN.R-project.org/package=ppcc

[20] Diethelm Wuertz, Tobias Setz and Yohan Chalabi (2017). fBasics: Rmetrics - Markets and Basic Statistics. R package version 3042.89. https://CRAN.R-project.org/package=fBasics

[21] Ilya Gavrilov and Ruslan Pusev (2014). normtest: Tests for Normality. R package version 1.1. https://CRAN.R-project.org/package=normtest

[22] Zhu, D., and Zinde-Walsh, V. (2009), “Properties and Estimation of Asymmetric Exponential Power Distribution,” Journal of Econometrics, 148, 89–99.