Mosaic Normality Test

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

ABSTRACT

A procedure is proposed here for jointly visualizing the compatibility of a sample with a family of Skewed Exponential Power Distributions, of which the distributions known as Normal, Exponential, Laplace and Uniform are particular cases. The procedure involves constructing a mosaic that contains these distributions in such a way that the asymmetry varies from left to right and the kurtosis varies from top to bottom. The null hypothesis that the sample belongs to each of the mosaic distributions is tested, with the corresponding box for each distribution being shaded in a gray scale according to the p-value obtained. The location and shape of the shaded area that appears on the mosaic facilitates not only identifying which distributions are compatible with the sample, but also assessing the power of the test performed. All the parameters that define the distribution are considered to be known. The problem remains open for expanding the procedure to cases in which these parameters are not known but must be estimated from the sample. As additional material, a code written in R that carries out the proposed test is included.

KEYWORDS: Normality tests, Goodness-of-Fit Methods, Power test, Graphical Methods

1   Introduction

There are a great variety of tests and procedures to check the normality of the data, so many that it may seem unnecessary to create one more. Thode (2002) lists an exhaustive series of tests and procedures that can be used for assessing normality. These procedures can be divided between those that use graphic techniques and those that use analytical techniques. Graphical methods such as the representation of values in Normal Probability Plot (NPP) allow visualizing the degree of discrepancy compared to the theoretical distribution together with the specific values that stand apart, but they

---

[*] Corresponding author. E-mail: pere.grima@upc.edu

do not offer an objective measure of that discrepancy. Analytical tests do measure this discrepancy, but it is advisable to complement them with some type of graphic representation that allows observing possible anomalous values that could affect the test result.

The analytical methods use different criteria to assess the discrepancy between the available data and the theoretical distribution. Regardless of which criterion is used, they are all based on the same reasoning schema:

1) The null hypothesis states that the data do indeed come from a Normal distribution.

2) A measure of the discrepancy between the data and the theoretical distribution is calculated. This measure can be gathered in various ways, such as: the regression and correlation techniques of Shapiro-Wilk and Shapiro-Francia; the difference between the theoretical and empirical distribution, as in the techniques used by Kolmogorv-Smirnov and Anderson-Darling; and estimating moments, like the methods of D'Agostino and Jarque-Bera.

3) The distribution of this discrepancy measure is known when the data actually come from a normal distribution. This distribution is called the "reference distribution" and may have been obtained either analytically or by simulation.

4) The discrepancy measure is compared with its reference distribution and the p-value p is determined. The hypothesis of normality is rejected if $p < \alpha$, with α being the level of significance chosen for the test.

This is actually the typical schema of hypothesis testing. However, even though general practice places the burden of proof on demonstrating a particular matter of interest (for example, that a certain drug is more effective than a placebo), testing normality assumes that the null hypothesis is true by default. This leads to paradoxes such as: the less data one has, the better it is to pass the test and be able to work with the normality assumption.

On the other hand, when a hypothesis test is carried out, it is usual to test a value that will serve as the parameter of the null hypothesis, such that by varying this value the probability of type II error can be continuously calculated and, thus, the power curve of the test can be drawn. However, this is not possible when testing normality, since the null hypothesis does not refer to the value of a parameter but to a type of distribution. Some works – such as those by Yacini and Yolacan (2007), Romão *et al*. (2010) and Yap and Sim (2011) – compute the probability of type II error for different normality tests using simulated data from different alternative distributions. This information helps in choosing the most convenient test in each case; although, when the test is performed,

it is not easy to visualize what its power curve would be. This difficulty can lead to the probability of type II error going unnoticed and its value being much greater than what the analyst would assume.

1.1 Visualizing type II error in Normality Tests

Sánchez-Espigares *et al.* (2018) propose a procedure for visualizing this type II error probability in a normality test. It is based on representing a mosaic comprised of $m \times m$ boxes, each box having a different distribution. Based on the Skewed Exponential Power Distribution (SEPD) used by Zhu and Zinde-Walsh (2009), they propose the following probability density function to create the distributions of the mosaic:

$$
f(y) = \begin{cases}
\dfrac{B}{\sigma} K(p) \exp\left(-\dfrac{1}{p} \left| \dfrac{A + B\left(\frac{y-\mu}{\sigma}\right)}{2\alpha} \right|^{p} \right) & \text{if } y \leq \mu - \dfrac{A}{B}\sigma\,, \\[2em]
\dfrac{B}{\sigma} K(p) \exp\left(-\dfrac{1}{p} \left| \dfrac{A + B\left(\frac{y-\mu}{\sigma}\right)}{2(1-\alpha)} \right|^{p} \right) & \text{if } y > \mu - \dfrac{A}{B}\sigma\,,
\end{cases}
$$

where:

$$
K(p) = 1 / \left[2p^{1/p} \Gamma(1 + 1/p) \right]\,,
$$

$$
A = \frac{1}{K(p)} \left[(1 - 2\alpha) \frac{p\Gamma(2/p)}{\Gamma^{2}(1/p)} \right]\,,
$$

$$
B = \sqrt{ \frac{1}{[K(p)]^{2}} \left[(3\alpha^{2} - 3\alpha + 1) \frac{p^{2}\Gamma(3/p)}{\Gamma^{3}(1/p)} \right] - \frac{1}{[K(p)]^{2}} \left[(1 - 2\alpha) \frac{p\Gamma(2/p)}{\Gamma^{2}(1/p)} \right]^{2} }\,.
$$

This function is characterized by the mean (μ) and the variance (σ^{2}) of the variable under consideration, as well as two more parameters: p, which is related to kurtosis and can vary between 1 (maximum kurtosis) and infinity (uniform distribution); and α, which is related to asymmetry and varies between 0 (very asymmetric distribution with a tail to the right) and 1 (with a tail to the left). The values $p = 2$ and $\alpha = 0.5$ correspond to the Normal distribution.

For constructing the mosaic, the value of p is varied between 1 and 50 because it is necessary that the maximum value is finite; and when $p = 50$, the distribution is already

practically uniform. The number of distributions on each side of the mosaic should be an odd number so that the Normal distribution remains directly in the center. In order for the values of α to cover their entire range of variation, the i-th value of α must be $\alpha_i = \frac{i-1}{m-1}$, where m is the number of distributions on each side of the mosaic.

Regarding the values of p, we begin with $p = 1$; and then each value is equal to the previous one raised to the power of $j = \sqrt[\frac{m}{2}-0.5]{\frac{\log 50}{\log 2}}$, except for the second one, which is always equal to $2^{1/j^{\left(\frac{m}{2}-1.5\right)}}$. For example, if $m = 9$ (9×9 mosaic), the values of α and p are those indicated in Figure 1, which also shows the shape that each of the distributions will have. The first row includes the exponential distribution ($p = 1$, $\alpha = 0$) and also the Laplace distribution ($p = 1$, $\alpha = 0.5$). The Normal distribution has been boxed in the center. In the last row, the distributions are practically uniform.
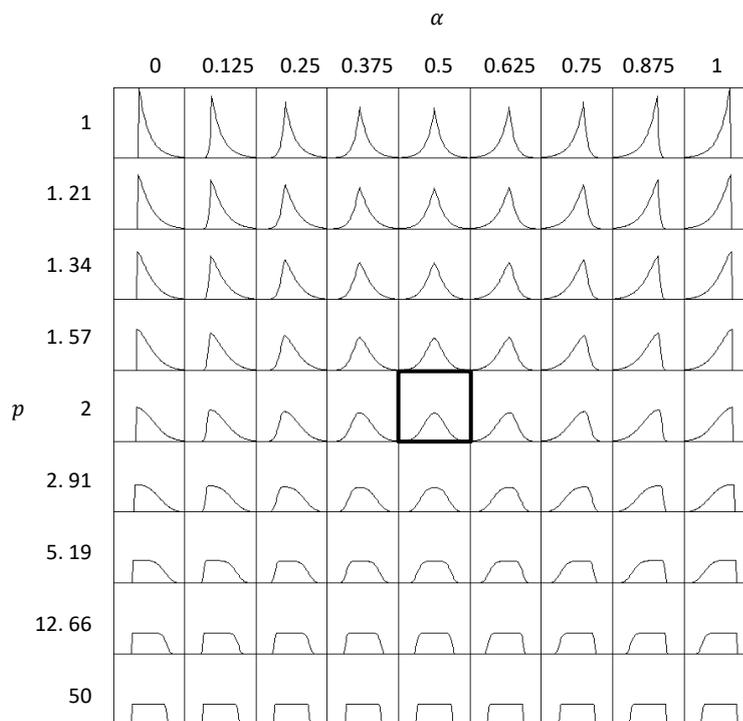

Figure 1: Mosaic of 9x9 distributions, with values of p (vertical) and α (horizontal).

Each distribution in the mosaic is used to generate k random samples of size n, and the normality of each of them is tested. Based on a level of significance that has been previously established (normally equal to 0.05), we calculate the proportion of times that the hypothesis of normality in each of the distributions is not rejected. Finally, each mosaic box (distribution) is shaded in a scale of gray according to the value of that proportion.

For example, upon generating $k = 1000$ samples of size $n = 50$ for each of the distributions that appear in the 9x9 mosaic in Figure 2, the number that appears in each box indicates the proportion of times that the null hypothesis of normality has not been rejected (using the Kolmogorov test). It can be seen that this proportion drops below 5% only for the exponential distributions in the upper corners, showing that the power of the test is low even with a sample that we could consider large (50 observations). A sequential gray scale is used to display the results more clearly. From lighter to darker shades, the ranges are 0-0.05, 0.05-0.25, 0.25-0.50 and 0.50-1.
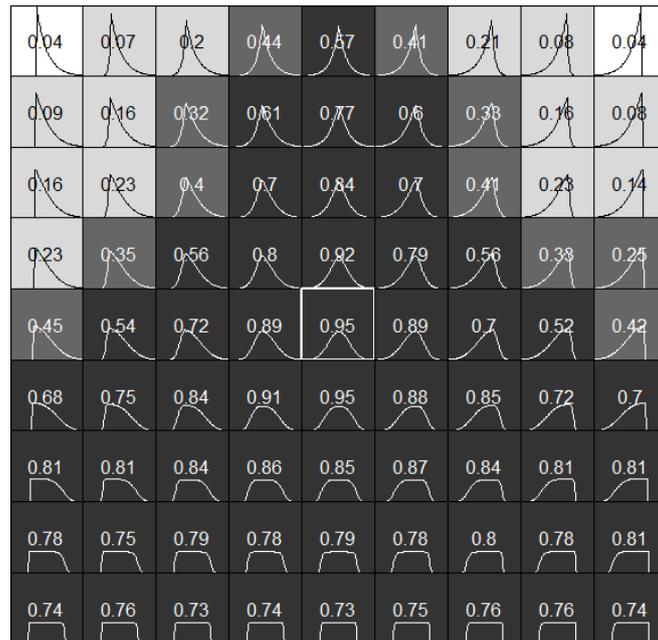

Figure 2: Each box shows the proportion of times that the hypothesis of normality is not rejected (α=0.05) when using the Kolmogorov test on samples of size n=50 coming from the given distributions.

1.2 Our proposal

We propose using the mosaic of distributions not only to visualize the probability of type II error but also to provide a new approach to normality tests. To test the normality of a sample, we propose testing if it fits in each of the distributions that appear in the mosaic and giving a dark color to those boxes (distributions) in which the fitting hypothesis cannot be rejected. In this way, a dark area on the mosaic covers all the distributions that are compatible with the available sample. If the box corresponding to the Normal distribution is within the dark area, the normality hypothesis cannot be rejected. This procedure can be easily incorporated into statistical software packages and has the additional advantage to provide an idea about the power of the test, which is impossible to observe when using conventional tests. Indeed, it is different for a sample to pass the normality test with a small or a large dark area on the mosaic. Clearly, in the first case

In the next sections we discuss some of the goodness of fit measures that can be used in our test. We describe the possibilities of using the graphic representation that we propose as well as its implementation with the statistical software R (R Core Team 2017). After that, we present two case studies in which the advantages of our suggested methodology will be apparent. Finally, we summarize the most relevant aspects of the proposed procedure, highlight its limitations, and suggest some possible improvements.

2   Goodness of fit criteria

We need goodness-of-fit tests that can be applied to any of the distributions that appear in the mosaic. For instance, even though the Shapiro-Wilk test gives excellent results for testing normality, it would not be useful for the remaining distributions. We can use tests based on measuring the discrepancy between the empirical distribution function of the sample, $F_n(x)$, and the theoretical distribution function that is proposed as the null hypothesis, $F(x)$. The values of $F(x)$ can be considered values of a uniform distribution $U(0,1)$ regardless of the original distribution of X, as long as it is continuous. Therefore, whatever the distribution in question, we measure the difference between the values of the empirical distribution of the sample distribution being tested and another sample that is always a random sample of a distribution $U(0,1)$.

In turn, the criterion of Stephens (1986) establishes that adjustment tests based on the Empirical Distribution Function (EDF) can be divided into two classes: the supremum class and the quadratic class. The most typical of each group are the Kolmogorov-Smirnov test (supremum) and the Anderson-Darling test (quadratic). These are the ones we will use. Below, we describe the general characteristics of each.

2.1   Kolmogorov-Smirnov

Given n data points $X_1, X_2, \cdots, X_n$, ordered from the smallest to largest value, the EDF is defined as $F_n(X_i) = k(i)/n$, where $k(i)$ is the number of points less than X_i. This is a step function that increases the value of each ordered data point by $1/n$. The Kolmogorov-Smirnov test statistic is defined as:

$$D_n = \max_{1 \le i \le n} \left(F(X_i) - \frac{i-1}{n}, \ \frac{i}{n} - F(X_i) \right)$$

Note that D_n is the maximum distance between both distributions. Let us take, for example, a sample of 15 random values of a distribution $N(0,1)$ generated with R

(seed=1). Figure 3 shows its EDF (stepped form) and the theoretical cumulative distribution (solid line), and highlights the value of the test statistic D_n.
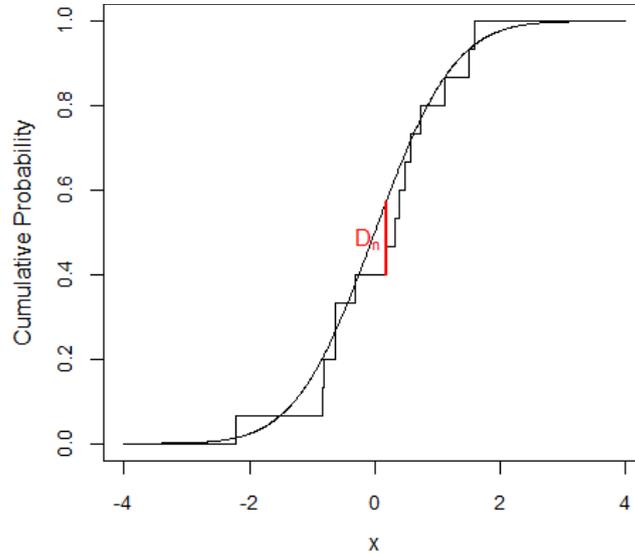

Figure 3: Kolmogorov-Smirnov statistic test (D_n): Maximum difference between theoretical and empirical distribution function

As we have said, the distribution of D_n does not depend on the distribution F, but it does depend on n. Using the same form suggested by Kolmogorov (see, for example, Gibbons and Chakraborti 2003), its asymptotic expression is:

$$\lim_{n \to \infty} P\left(D_n \sqrt{n} \leq x\right) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 x^2}$$

Marsaglia *et al*. (2003) detail a procedure for determining p-values with a precision of at least 13 digits for values of n, from 2 up to at least 16000. This procedure is used in the `ks.test` function, which is included in the `stats` package (R Core Team 2017) of the statistical software R.

2.2   Anderson-Darling test

This is based on the Cramer-von Mises test, which quantifies the difference between F and F_n through the expression:

$$Q = n \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 dF(x)$$

Anderson and Darling (1954) added the weighting factor $\frac{1}{[F(x)(1-F(x)]}$, which gives more weight to the values located at the extremes of the sample. In this way, the test is more

powerful and, in addition, the critical values do not depend on n for $n > 5$. The discrepancy measure, including the weighting factor, can be expressed as (see, for example, Marsaglia and Marsaglia 2004):

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i - 1}{n} \left[\ln\big(F(X_i)\big) + \ln(1 - F(X_{n+1-i})) \right]$$

where $X_1 < \cdots < X_n$ are the values of the ordered sample. As in the Kolmogorov-Smirnov test, the values of $F(X_i)$ can be understood as a random sample of a distribution $U(0,1)$. Marsaglia and Marsaglia (2004) present several methods for deducing the probability $P(A^2 < k)$. The most practical one offers an approximation that has greater than 0.000002 accuracy for probabilities that are less than 0.9, and 0.0000008 accuracy for those that are above. The `ADGofTest` package (Gil Bellosta 2011) of the statistical software R includes the `ad.test` function, which gives p-values for the Anderson-Darling test that are calculated according to one of the methods proposed by Marsaglia and Marsaglia. The manual for this package does not clarify which of the paper's proposed methods is applied, but the accuracy of even the most quick-and-easy one is more than enough.

3   Normality Test using the mosaic of distributions

As said before, given a sample on which we wish to test the normality, our proposal is testing whether it pertains not only to the Normal distribution but to all the distributions that appear in the mosaic. To show the results visually, each box in the mosaic is given a color that is a function of the p-value obtained by testing the distribution contained in it. These colors are:

p-value	Color
< 0.01	White
0.01 – 0.05	Light gray
0.05 – 0.10	Gray
> 0.10	Dark gray

In order to carry this out, we have created a program with the statistical software package R, which is available to interested readers as additional material to this article. The program allows you to choose:

- The size of the mosaic, up to a size of 49×49 (which would be easy to expand).

- Aspects related to the presentation, such as the cut-off points of the p-value in gray scale or whether or not the p-values are shown in each cell.

- The criterion for measuring the discrepancy. As we have mentioned, this may be that of Kolmogorov-Smirnov (KS) or of Anderson-Darling (AD).

- The sample to test. This may be a dataset whose normality we wish to test, or we can also use a randomly generated sample of a Normal, uniform or exponential distribution in order to observe the behavior of the program. In the first case, it is necessary to introduce the population parameters, which are presumably known.

If the criterion for measuring the discrepancy is that of KS, the p-value is calculated with the `ks.test` function of the `stats` package, which is already loaded by default. If the AD criterion is applied, the `ADGofTest` package is used (it is necessary to install the library beforehand). These functions are based on Marsaglia *et al*. (2003) as well as Marsaglia and Marsaglia (2004). Furthermore, as we have mentioned previously, they provide a more than adequate approximation for our objectives.

Below, we will see some examples, all of which are reproducible using the seed indicated in each case. Let us take a sample of $n = 100$ observations of a distribution $N(0,1)$ generated with seed=1. Using the AD criteria, we obtain the mosaic in Figure 4 (right). Naturally, the null hypothesis of normality is not rejected, but neither will be the Laplace distribution (double exponential) located in the first row of the mosaic and in the same column as the normal distribution. However, the exponential distribution (both right- and left-skewed) would be rejected, as well as the uniform distribution. If the KS criterion is used, we have the mosaic in Figure 4 (left). It is clear that this test has less power than that of AD.

Unsurprisingly, the size of the shaded area on the mosaic tends to decrease as the sample size increases. This can be seen in Figure 5, where we represent the mosaics obtained using randomly generated samples (seed=1) of a distribution N (0,1) and with the indicated sample sizes.

This procedure can also be used to test whether the sample comes from any of the distributions that appear in the mosaic. It is enough to observe in which interval can be found the p-value of the box corresponding to the null hypothesis. For example, Figure 6 shows the shaded area obtained from a sample of 100 randomly generated observations (seed=1) from an exponential distribution (top) and a uniform distribution (bottom). Both cases include the histogram representing the distribution of the data.
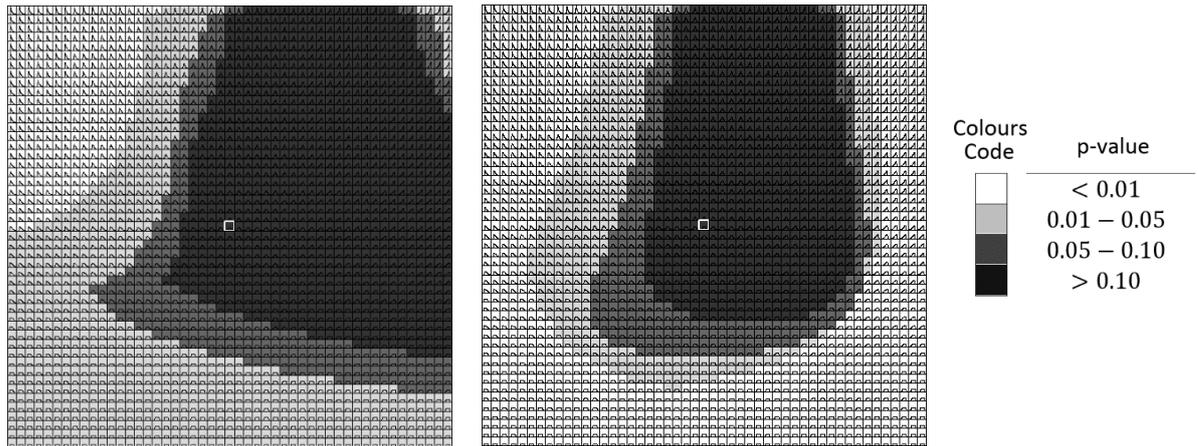
Figure 4: Mosaic test applied to a sample of 100 random numbers of N(0,1) generated with R (seed=1), using the criteria of Kolmogorov-Smirnov (left) and Anderson-Darling (right).
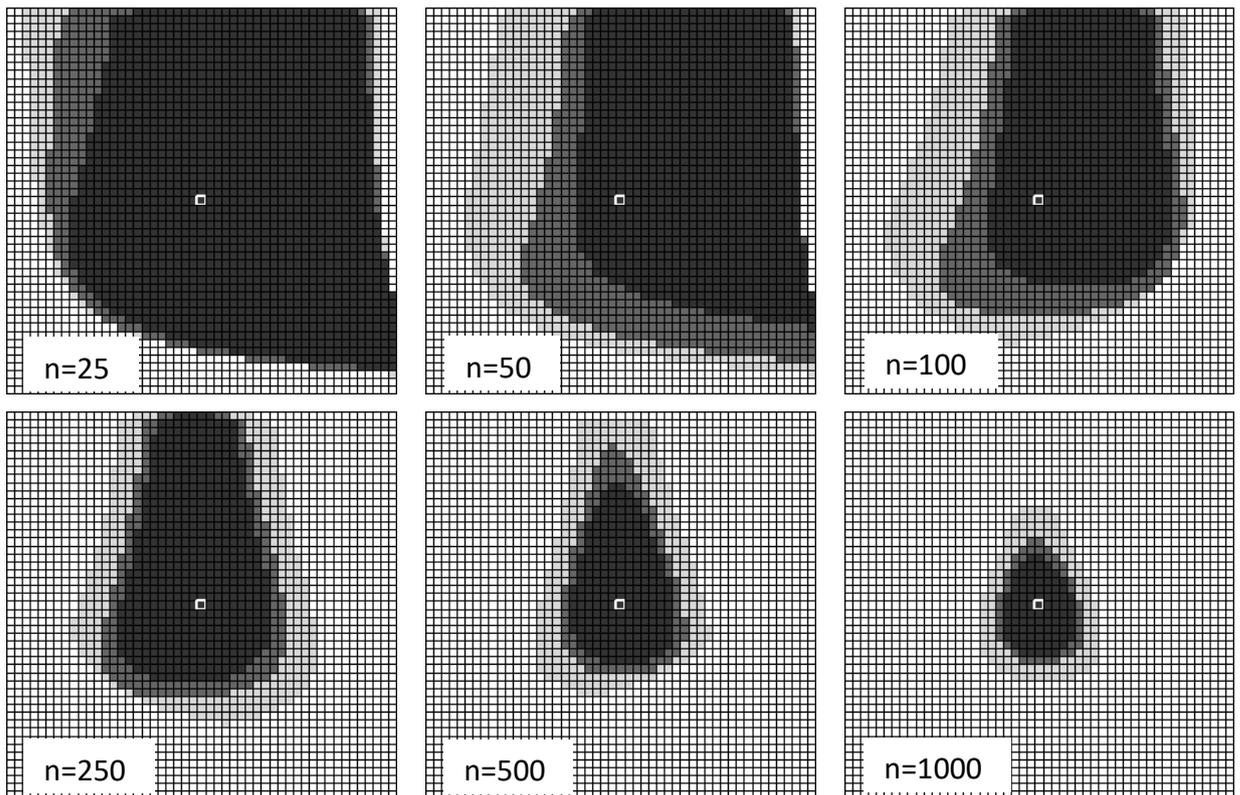

Figure 5: Results obtained using the AD criteria with samples of the indicated size, obtained randomly (with R, seed=1) from a population of N(0,1)
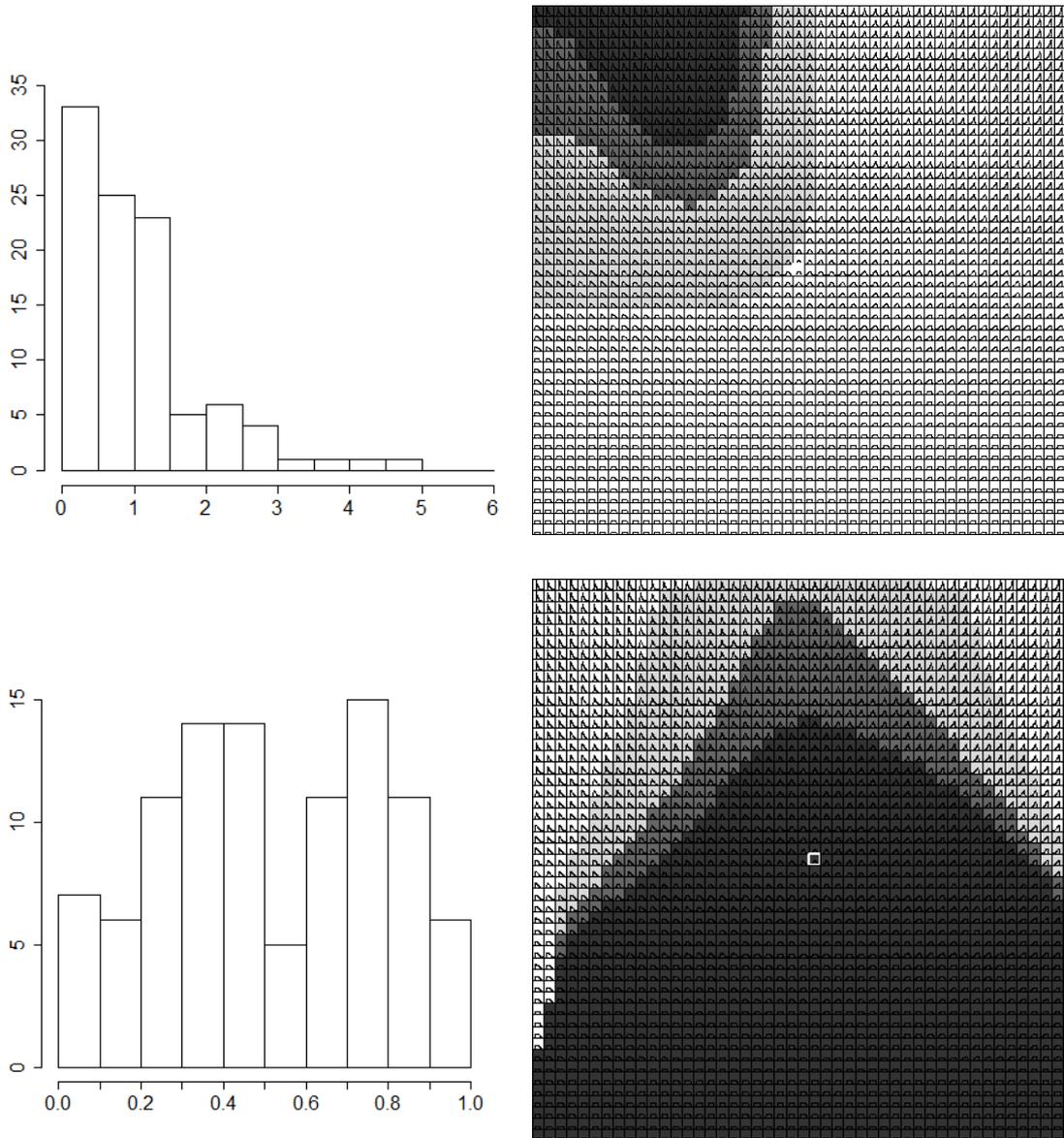
Figure 6: Results obtained using the AD criteria with a sample of n=100 from an exponential distribution (λ=1) (top) and uniform distribution (0,1) (bottom), using R (seed=1). The histogram corresponding to the data is attached.

4 Case Studies

4.1 Leghorn Chick Data

To illustrate normality tests based on EDF statistics, Stephens (1986, p. 98) uses data from Bliss (1946) on the weight in grams of twenty 21-day-old leghorn chicks (Table 1).

Table 1: Leghorn Chick Data (from Bliss, 1946)

156	162	168	182	186	190	190	196	202	210
214	220	226	230	230	236	236	242	246	270

Stephens states that – obviously – the data have been rounded, so in a strict sense the population is discrete. However, with values of this magnitude, the obtained approximation shows a negligible difference from the unrounded values. The values of μ and σ are assumed to be known, and we use those that are proposed by Stephens: $\mu = 200$ and $\sigma = 35$. Applying KS to test normality obtains a test statistic of $D = 0.171$, with a p-value = 0.60. An analogous result is obtained when applying the AD test: in this case, the test statistic is $A^2 = 1.017$ and the p-value is 0.35. Therefore, we would assume that these data come from a normal distribution.

However, when we test whether or not our sample comes from each of the distributions in the mosaic, all boxes take on a darker color when using the KS test, meaning that the obtained p-values are greater than 0.10 for all the distributions. Therefore, using the KS test at this level of significance precludes rejecting the hypothesis of normality, although neither can we reject that the data come from any of the other distributions in the mosaic (which, as we have seen, may be exponential, uniform or Laplace). This situation shows the low value of the KS test when using small sample sizes.

Using the AD test to these same data, we obtain a mosaic with distributions that would be rejected at the usual level of significance $\alpha = 0.05$, although all of them are very far from the normal distribution (Figure 7, left).

4.2 Chest circumferences of Scottish soldiers

Stigler (1986, p. 207) reproduces the famous dataset of chest circumferences of 5758 Scottish soldiers analyzed by Quetelet. Unfortunately, the original data were not preserved, and we have only the tabulated data (Table 2).

We have transformed each of the values in the table (X) into continuous values by replacing them with:

$$X - 0.5 + \frac{1}{f_X+1} \ , \ \ X - 0.5 + \frac{2}{f_X+1} \ , \ \cdots, \ X - 0.5 + \frac{f_X}{f_X+1}$$

where f_X is the observed frequency for the value X. Rounding to the nearest integer, the table's original values are once again obtained.

Table 2: Quetelet's data on the chest circumferences of 5738 Scottish soldiers

Circumference of chest. English inches (X)	Frequency
33	3
34	18
35	81
36	185
37	420
38	749
39	1073
40	1079
41	934
42	658
43	370
44	92
45	50
46	21
47	4
48	1
Total	5738

In this case, the data also pass the normality tests of both KS ($D = 0.009$, p-value= 0.74) and AD ($A^2 = 0.47$; p-value $= 0.48$).

For each distribution in the mosaic, we test the null hypothesis that our dataset pertains to that distribution, and it is rejected for all of them except for the normal distribution and those very close to it. Only a small spot surrounds the normal distribution in the center. Figure 7 (right) shows the spot obtained by applying the AD test. If the KS test is used, the spot is somewhat larger.
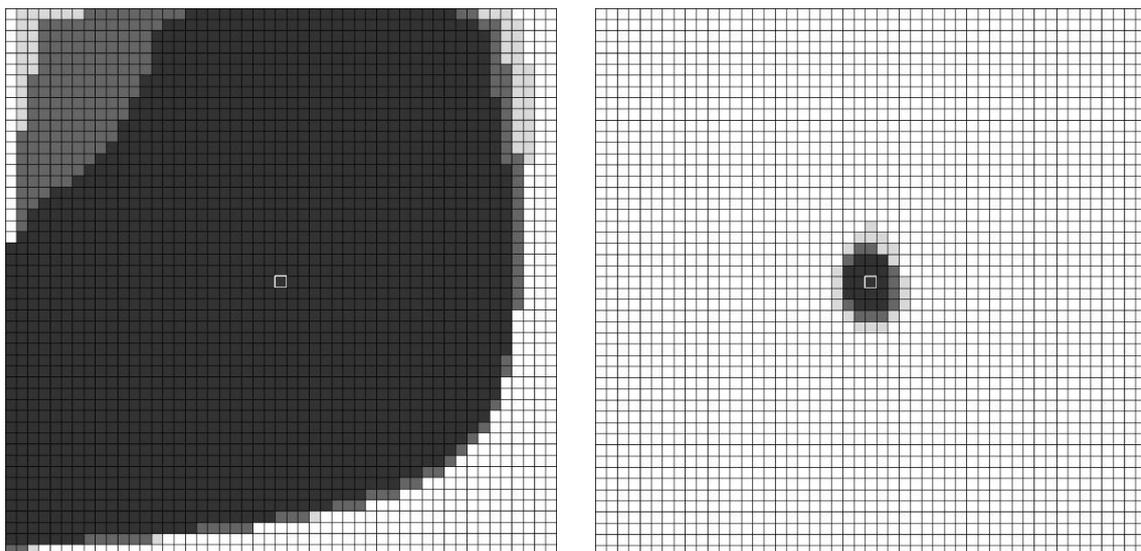

Figure 7: Mosaic obtained by applying the AD test to Bliss data (left) and Quetelet data (right)

Consequently, the result is unlike the previous example and we can state with certainty that these data come from a normal distribution, because we can definitively reject that the data come from distributions that are even very close to the normal. Applying a conventional normality test does not distinguish between the two analyzed cases. However, when using the mosaic of distributions, the two situations are shown to be clearly different.

5   Final remarks

The mosaic of distributions provides a clear idea of whether or not the sample is compatible not only with the distribution being tested as the null hypothesis, but also with all the unimodal distributions around it. A large shaded area on the mosaic (which occurs with small or moderately sized samples) shows that the sample could also come from other distributions that are far away from those being tested, and this is something that is generally not taken into account by analysts.

Note that, although the mosaics used to visualize the type II error (Sánchez-Espigares et al., 2018) and those we propose here are visually similar, their objectives and the information they represent are different. Table 3 summarizes the characteristics of each type of representation.

Table 3: Comparison of mosaics for visualizing type II error and goodness of fit tests

Type II error display	Goodness of fit test
In each mosaic box, we generate k random samples of the distribution that appears in that box.	There is only one sample.
In each box, a normality test is carried out for each of the k generated samples.	The sample is tested to determine if it comes from each of the distributions in the mosaic.
Using a set level of significance for each test, the normality of the sample is or is not rejected. The proportion of times that the hypothesis of normality is not rejected is assigned to each box.	The p-value obtained by testing whether or not the sample comes from the distribution that appears in this box is assigned to that box.
Optionally, the proportion of times that the normality hypothesis has not been rejected is indicated over each box (distribution).	Optionally, the p-value obtained in the test is indicated over each box (distribution).

The boxes are colored in gray scale according to the proportion of times that the normality hypothesis is rejected.

The boxes are colored in gray scale according to the p-value obtained by testing whether the sample data come from the distribution of the box.

The proposed procedure requires knowing the parameters that define the population considered in the null hypothesis. For practical purposes, being able to use the estimates of those parameters obtained from the available sample would be more useful. However, using these estimates changes the shape of the reference distributions, and we do not have their analytical expressions. Simulation techniques are necessary for deducing these expressions and/or calculating the critical values.

Deducing analytical expressions is not an easy task, and simulation requires great computational effort. The critical values depend on the values of α and p that are linked to the distribution, and also on the size of the sample n. Obtaining these critical values at a precision of 4 digits for each of the distributions in the mosaic is, for the moment, beyond our range of possibilities. Nevertheless, we believe this can be addressed in the near future.

References

Anderson, T. W. and D. A. Darling. 1954. A Test of Goodness of Fit. *Journal of the American Statistical Association* 49(268):765-769.

Bliss, C. I. 1946. Collaborative comparison of three ratios for the chick assay of vitamin D. *Journal of the Association of Official Agricultural Chemists* 29:396-408.

Gil Bellosta, C. J. 2011. *ADGofTest: Anderson-Darling GoF test*. R package (version 0.3). https://CRAN.R-project.org/package=ADGofTest

Gibbons, J. D. and S. Chakraborti. 2003. *Nonparametric Statistical Inference*. 4 ed. New York: Marcel Dekker.

Marsaglia, G., W. W. Tsang and J. Wang. 2003. Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, 8(18):1-4. doi: 10.18637/jss.v008.i18

Marsaglia, J. and G. Marsaglia. 2004. Evaluating the Anderson-Darling Distribution. *Journal of Statistical Software* 9(2):1-5. doi: 10.18637/jss.v009.i02.

R Core Team. 2017. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Romão, X., R. Delgado and A. Costa. 2010. An empirical power comparison of univariate goodness-of-fit tests for normality. *Journal of Statistical Computation and Simulation* 80(5): 545-591. doi: 10.1080/00949650902740824

Sánchez-Espigares, J. A., P. Grima, L. Marco-Almagro. 2018. Visualizing Type II Error in Normality Tests. *The American Statistician.* 72(2):158-162. doi: 10.1080/00031305.2016.1278035

Stephens, M. A. 1986. Test Based on EDF Statistics. In *Goodness-of-Fit Techniques*, ed. R. B. D'Agostino and M. A. Stephens, 97-193. New York: Marcel Dekker.

Stigler, S. M. 1986. The History of Statistics. The Measurement of Uncertainty before 1900. Cambridge, MA: Harvard University Press.

Thode, H. C. 2002. *Testing for Normality*. New York: Marcel Dekker.

Yacini, B. and S. Yolacan. 2007. A comparison of various tests of normality. *Journal of Statistical Computation and Simulation.* 77(2): 175-183. doi: 10.1080/10629360600678310.

Yap, B. W. and C. H. Sim. 2011. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12):2141-2155. doi: 10.1080/00949655.2010.520163

Zhu, D. and V. Zinde-Walsh. 2009. Properties and estimation of asymmetric exponential power distribution. *Journal of Econometrics.* 148(1): 89-99. doi: 10.1016/j.jeconom.2008.09.038