GTest: a software tool for graphical assessment of empirical distributions' Gaussianity

E. Barca · E. Bruno · D. E. Bruno · G. Passarella

Received: 24 June 2015 / Accepted: 26 January 2016 / Published online: 3 February 2016

Abstract In the present paper, the novel software GTest is introduced, designed for testing the normality of a user-specified empirical distribution. It has been implemented with two unusual characteristics; the first is the user option of selecting four different versions of the normality test, each of them suited to be applied to a specific dataset or goal, and the second is the inferential paradigm that informs the output of such tests: it is basically graphical and intrinsically self-explanatory. The concept of *inference-by-eye* is an emerging inferential approach which will find a successful application in the near future due to the growing need of widening the audience of users of statistical methods to people with informal statistical skills. For instance, the latest European regulation concerning environmental issues introduced strict protocols for data handling (data quality assurance, outliers detection, etc.) and information exchange (areal statistics, trend detection, etc.) between regional and central environmental agencies. Therefore, more and more frequently, laboratory and field technicians will be requested to utilize complex software applications for subjecting data coming from monitoring, surveying or laboratory activities to specific statistical analyses. Unfortunately, inferential statistics, which actually influence the decisional processes for the correct managing of environmental resources, are often implemented in a way which expresses its outcomes in a numerical form with brief comments in a strict statistical jargon (degrees of freedom, level of significance, accepted/rejected H_0, etc.). Therefore, often, the interpretation of such outcomes is really difficult for people with poor statistical knowledge. In such framework, the paradigm of the *visual inference* can contribute to fill in such gap, providing outcomes in self-explanatory graphical forms with a brief comment in the common language. Actually, the difficulties experienced by colleagues and their request for an effective tool for addressing such difficulties motivated us in adopting the *inference-by-eye* paradigm and implementing an easy-to-use, quick and reliable statistical tool. GTest visualizes its outcomes as a modified version of the Q-Q plot. The application has been developed in Visual Basic for Applications (VBA) within MS Excel 2010, which demonstrated to have all the characteristics of robustness and reliability needed. GTest provides true graphical normality tests which are as reliable as any statistical quantitative approach but much easier to understand. The Q-Q plots have been integrated with the outlining of an acceptance region around the representation of the theoretical distribution, defined in accordance with the alpha level of significance and the data sample size. The test decision rule is the following: if the empirical scatterplot falls completely within the acceptance region, then it can be concluded that the empirical distribution fits the theoretical one at the given alpha level. A comprehensive case study has been carried out with simulated and real-world data in order to check the robustness and reliability of the software.

Keywords Statistical inference-by-eye · Normality tests · Q-Q plot · Confidence bands · VBA MS Excel©

E. Barca (✉) · E. Bruno · D. E. Bruno · G. Passarella
Water Research Institute, National Research Council, Viale De Blasio, 5-70125 Bari, Italy
e-mail: emanuele.barca@ba.irsa.cnr.it

Introduction

Statistical procedures and methods became of common use in an ever-wider range of sectors. The increasing use of such methods entailed that an ever-growing number of practitioners has been involved in their usage. Unfortunately, as pointed out by Glantz (2005), practitioners often apply such procedures in a biased fashion, drawing from them sometimes incorrect conclusions (Sutherland et al. 2013). In particular, one of the most common mistakes is the lack of checking about the normality assumption of the working dataset, e.g. for regression, ANOVA or Kriging (Thode 2002; Diggle and Ribeiro 2007). Among all application fields, the environmental sciences have a long tradition of statistical inference for assessing the actual status of natural resources (Masciale et al. 2011) and geostatistics for any pointwise regionalized analysis (Barca and Passarella 2008). Nevertheless, during recent decades, both statistical and geostatistical procedures have been applied even more intensively, driven by an increased concern for environmental issues (Ott 1995; Wheater and Cook 2000; Barca and Passarella 2008; Barca et al. 2008). Nowadays, due to strict environmental regulations, local authorities need to adapt their policies to high-level standards and protocols for data analysis. This need has led people with informal statistical skills to try to apply and understand a wide range of procedures quite suddenly. Nevertheless, the risk of uninformed application of any, even basic, statistical procedure is that biased or completely wrong conclusions may be reached. In this framework, the *inference-by-eye* or visual hypothesis testing can play a key role for filling in the user's knowledge gap. Actually, *inference-by-eye* allows users to read the results of the statistical test by looking at a simple graphical representation without decoding complex numerical tables and indices usually commented by difficult statistical jargon (degree of freedom, accepting/rejecting H_0, level of significance, etc.). What was said above motivated us to adopt the inference-by-eye paradigm and to implement GTest. It is a quick and reliable software tool capable of performing four different graphical normality tests for a given empirical distribution function (EDF) in MS Excel 2010 using VBA as code language. In fact, as stated in the recent scientific literature, Excel has proved to be an ideal environment to develop such kind of tool (Keeling and Pavur 2011; Nash 2006).

The graph selected as explanatory output is a modified version of the quantile-quantile (Q-Q) plot; it is a widespread graphical tool used for testing the level of agreement of an EDF with a known theoretical distribution (Gnanadesikan and Wilk 1968). The Q-Q plot approach fundamentally consists in a graphical comparison between the first quadrant bisector, representing the theoretical distribution, and the scatterplot of the EDF quantiles. The main objective of this work consists in improving the Q-Q plot by also computing and plotting the confidence limit band (CLB) of the EDF at a given α *level* of significance. This graphical arrangement will transform the plot in a rigorous by-eye goodness-of-fit test at the given α level and will remove any subjective inference regarding the evaluation. Practically, the CLB works as an *acceptance region*: only those EDFs whose scatterplot is completely within the acceptance region will positively pass the test at the given α level of significance. The application only requires the dataset to be loaded in a single sheet and the choice of the α value and the test method, which can be easily input by means of an ad hoc dialogue mask. According to the user-selected method, the application creates a new plot sheet with the graphical results and a couple of complementary sheets where all the computations are performed. The four available methods are the simultaneous, pointwise, stabilized and correlation coefficient probability plots based on the Kolmogorov-Smirnov (Hogg and Tanis 1977), binomial distribution (Conover 1980), Michael's transformation (Michael 1983; Royston 1993) and Filliben tests (Filliben 1975), respectively.

Materials and methods

Q-Q plot

The Q-Q plot ("Q" stands for quantile) is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Q-Q plots are commonly used to compare an EDF to a reference theoretical model by means of a *visual goodness of fit* and to mark suspicious values in the case of hypothesis rejection. In this work, a Q-Q plot approach is used to compare a given EDF with the standard normal cumulative distribution function (CDF).

Generally, the X and Y axes of a Q-Q plot represent the theoretical and empirical quantiles of the CDF and the EDF, respectively. Consequently, drawing a Q-Q plot requires a method of coupling ordered values of both distributions. A useful simplification in creating a Q-Q plot consists in comparing the standardized distributions. This consists in moving in the reference origin the empirical distribution central value and in the rescaling of the dataset. The practical use of standardization is that it allows comparing the empirical distribution with a unique theoretical distribution, namely the Gaussian standard. Such choice has two direct consequences, the code's simplification and the moving of the working domain in an interval hardly much larger than [−3, +3] (which covers more than the 99 % of the normal distribution area), approximately. In addition, it sets the geometrical representation of the Q-Q plot exactly in the first quadrant and, by moving the central value to the origin, sets the reference line to the bisector of the first quadrant (standardizing the output graphical fashion).

After the standardization, the EDF quantiles ($e'_{(i)}$) are sorted in ascending order. The corresponding abscissas (i.e. CDF quantiles) are evaluated by means of the following two-step procedure (Hazen 1930):

1. Compute the "plotting position" of the EDF *i*th quantile:

$$p_i = (i–0.5)/n \tag{1}$$

where

p_i = cumulative probability of the EDF *i*th quantile
i = position of the EDF quantile in the ordered list
n = dataset size

In the second step, the inverted standard Gaussian CDF is determined as a function of the previously computed probability:

2. Compute the CDF *i*th quantile

$$c'_{(i)} = \Phi^{-1}(p_i) \tag{2}$$

where

$c'_{(i)}$ = CDF *i*th quantile
i = position of the CDF quantile in the ordered list
Φ^{-1} = inverse standard Gaussian CDF

The EDF can be now represented on the Q-Q plot as a scatterplot of pairs ($c'_{(i)}, e'_{(i)}$), while the theoretical distribution is represented by the first quadrant bisector.

Accepting or rejecting the normal hypothesis by a visual decision rule commonly consists in evaluating *how much* the empirical scatterplot (EDF) departs from the first quadrant bisector (CDF). Nevertheless, this is not an easy task and only the analyst's experience and knowledge can lead to a reliable decision. In fact, the Q-Q plot by itself does not provide either quantitative or objective indices for testing the hypothesis.

Confidence limit bands

A decisive support in comparing EDFs and normal CDFs, graphically, comes from the confidence limit bands (CLBs), which bound an acceptance region about the first quadrant bisector at a given α level of significance. Practically, those EDFs whose scatterplots lie completely within the acceptance region can be considered statistically equivalent to the theoretical reference distribution. Therefore, drawing the acceptance region actually removes any subjectivity in the interpretation of the Q-Q plot. The CLBs are univocally defined as curves representing the confidence interval (CI) limits of each quantile of the normal standard CDF. In practical terms, a CI provides the set of plausible values for a fixed quantile (Beaulieu-Prevost 2006).

The statistical literature reports several methods for calculating the CLBs (Conover 1980; Calzada and Scariano 2002; Michael 1983), which differ from each other in terms of the expected degree of accuracy. In fact, different approaches provide slightly different approximations of the acceptance region boundary problem. Each method is tied to a specific normality test as will be shown in the following section.

Proposed Gaussian test methods

In this paper, four methods based on different theoretical and operative approaches have been applied and compared. Actually, the proposed software allows users to choose one of the methods, case by case. In the following sections, a short discussion on the methods' specific characteristics has been outlined. In general, the technical literature provides a good number of normality tests characterized by different levels of ease and reliability. The methods implemented in GTest share some desirable properties, (i) the possibility of representing test results in the form of a CLB-bounded Q-Q plot, (ii) their diffusion and acceptance within the scientific community and, finally, (iii) their reliability and power in the statistical sense.

The first two approaches perform a *goodness-of-fit test inversion* with or without suitable data transformation, while the third is based on binomial distribution properties. The fourth method, based on Filliben's theory, does not provide a graphical representation of the acceptance region directly. Nevertheless, it can be approximately drawn by applying suitable methods such as resampling or bootstrapping (Stirling 1982; Calzada and Scariano 2002; Michael 1983; Filliben 1975). Finally, it must be remarked that the normality tests often have as a limitation that data are required to be unaffected by ties (repeated values in the data series), which cause biased results. In the present implementation, a suitable procedure is applied in order to also achieve correct results when the distribution being checked is not entirely continuous. In the following, a very short theoretical description of each method is provided, given that these methods are well documented in the technical and scientific literature.

Simultaneous methods

The first two methods, namely the Kolmogorov-Smirnov and stabilized plot methods, are usually referred to as simultaneous since the corresponding CLBs' bounds are univocally described by explicit analytical functions.

Kolmogorov-Smirnov

The Kolmogorov-Smirnov test (*K-S test*) is a non-parametric (distribution free) test which quantifies the distance between the EDF and the CDF (Conover 1980). The *null* distribution associated with this statistic is calculated under the *null hypothesis* that the EDF coincides with the CDF. The scientific literature provides a huge amount of papers related to this test, its technical evolution (Steinskog et al. 2007; Stirling 1982) and its potential graphical development (Calzada and Scariano 2002). Unfortunately, the test was originally conceived for the common normality test, where the characteristic parameters (location and scale) are completely specified a priori and not for its generalized form where the parameters have to be derived from data. This is a well-known weak point of K-S test, because the use of estimated parameters modifies the null hypothesis, considerably reducing the power of the test and providing CLBs at $(1-\alpha)\%$ level of significance but only in average, that is to say that the nominal level of significance is not respected everywhere in the plot (Stirling 1982; D'Agostino and Stephens 1986; Steinskog et al. 2007); nevertheless, it still remains one of the most popular statistical tests and, consequently, it was included in this work.

Equation (3) provides the functions of the CLBs of the normal Q-Q plot obtained by applying the *modified K-S test inversion* approach (Calzada and Scariano 2002):

$$\Phi^{-1}(F_n(v)-D_\alpha) < v < \Phi^{-1}(F_n(v)+D_\alpha); \qquad \forall v \tag{3}$$

with a confidence of $100\cdot(1-\alpha)\%$ and where:

v = empirical quantile
n = dataset size
$F_n(v)$ = empirical CDF
D_α = test rejection threshold
Φ^{-1} = inverted standard Gaussian CDF

Stabilized variance plot test inversion

Like the K-S test, the stabilized plot test is a simultaneous test. It differs substantially from the K-S test for the analytical form of the CLBs' functions (Michael 1983). Equation (4) provides the functions of the CLBs of the normal Q-Q plot obtained by applying the stabilized variance plot approach:

$$\Phi^{-1}\left(\sin^2\left(\arcsin\left(\Phi^{0.5}(v)\right)-\pi D'_\alpha\right)\right) < v < \Phi^{-1}\left(\sin^2\left(\arcsin\left(\Phi^{0.5}(v)\right)+\pi D'_\alpha\right)\right) \tag{4}$$

with a confidence of $100\cdot(1-\alpha)\%$ and where:

v = empirical quantile
D'_α = test rejection threshold
Φ^{-1} = inverted standard Gaussian CDF

Critical values are reported in the tables provided by Michael (1983) and improved by Royston (1993). As will be shown in the case study, this method is well suited for analysing percentage data.

Pointwise method

As described previously, after standardizing and sorting the empirical values, a *plotting position p* (substantially, a cumulative probability estimation) is attributed to each empirical quantile (Eq. (1)). If the discrete nature of the problem is exploited, it can be proved that, once estimated the cumulative probability p associated to any quantiles, the cumulative distribution associated to the generic kth quantile can be defined as follows:

$$P(k) = \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j} = \text{binomial}(k, n, p) \tag{5}$$

As can be viewed, the distribution regulating the probability P of the quantile statistics is inherently *binomial* where the parameters represent:

k = number of successes or quantile index
n = number of successes or quantile index
p = success probability or cumulative empirical quantile probability estimate

Given such theoretical framework, a confidence interval (CI) for any quantile v_{p} can be estimated. In practice, once an α level of significance has been given (e.g. $\alpha = 5\,\%$), the symmetrical two-tailed test can be carried out first computing the following critical probabilities $\alpha_U = 1-(1-\alpha)/2$, and next $\alpha_L = (1-\alpha)/2$ (e.g. $\alpha_U = 0.975$ and $\alpha_L = 0.025$). Successively, equating α_L and α_U to Eq. (5), the indices k_L, k_U can be evaluated. In practice, $k_L = \min\limits_{1\le k\le n} (\text{binomial}(k,\ n,\ p) \ge \alpha_L)$ and $k_U = \max\limits_{1\le k\le n} (\text{binomial}(k,\ n,\ p) \le \alpha_U)$. Finally, the values k_L and k_U will be used to select the corresponding theoretical quantiles $v_{(k_L)}$ and $v_{(k_U)}$, which become the bounds of the CI. A shortcut to find k_L and k_U consists in inverting the binomial distribution, achieving directly the unknown values. p being the cumulative probability associated to the quantile to be estimated, the lower and upper limits of the CI can be computed iteratively for any quantile just changing the corresponding value of k and p. Unfortunately, the method has a drawback; in fact, in some cases, $\boldsymbol{k}$ could not exist as an exact integer value such that binomial$(\boldsymbol{k}, n, p) = \alpha$; consequently, the user could be forced to use an approximate value. Applying the binomial approach in order to estimate the quantile confidence limits is generally known as the pointwise method (Hollander and Wolfe 1999).

Filliben's method

There is abundant scientific literature about plotting positions' best mathematical expression. Equation (1) shows the most used actually in practice, but there are a large number of different equations. The following one allows estimating the median empirical and theoretical quantiles:

$$p_i = \begin{cases} 1-0.5^{(1/n)} & i=1 \\ (i-0.3175)/(n+0.365) & i=2,\ 3,\ \ldots,\ n{-}1 \\ 0.5^{(1/n)} & i=n \end{cases} \tag{6}$$

and

$$c'_{(i)} = \Phi^{-1}(p_i) \tag{7}$$

Filliben provided a quantitative statistics for testing the normal hypothesis. He simply used the linear correlation coefficient between the CDF and EDF quantiles. This parameter, usually called Filliben's correlation coefficient or probability plot correlation coefficient (PPCC), just arises from the concept of the *Q-Q* plot and overcomes the visual inspection providing a single numerical value able to quantify the goodness of fit. Filliben suggested that the advantage of the PPCC rests in its conceptual simplicity; the joint use of the Q-Q plot and correlation coefficient, which measures the linear agreement of an EDF with a CDF, is the best suited to accomplish the task of comparing quantitatively the theoretical CDF with the EDF. Regarding the nature of Filliben's test, since perfect normality implies perfect correlation (i.e. a correlation value of 1), we are only interested in rejecting normality for correlation values that are too low. That is, this is a lower one-tailed test. The computed PPCC is finally compared with the critical value uniquely individuated by n and α of the table provided by Filliben (1975) for testing its significance. The table of critical values was improved successively by Looney and Gulledge (1985) and Devaney (1997); the table provided by Devaney is that actually used in the present work.

What is said above entails that Filliben's method does not allow one to define directly the CLB lines. Nevertheless, they can be drawn approximately by suitable methods such as resampling or bootstrapping (Rochowicz 2010). Among several available bootstrapping techniques, the non-parametrical technique has been applied in this work because of its straightforward implementation in MS Excel.

Final remarks

A problem can rise in representing graphically a normality test; in fact, when the dataset size n overcomes a certain threshold, the cumulative probability of some quantiles located at the extreme of the two tails of the EDF becomes too low or too high. Consequently, the computation of the lower or the upper CI limit of an extreme quantile can result *undetermined*. As an example, for the K-S case, the basic test relationship is the following:

$$1-\alpha \approx P[F_n(v)-D_\alpha(n) \le \Phi(v) \le F_n(v) + D_\alpha(n);\ \forall v] \quad (8)$$

When the $F_n(v)$, corresponding to the cumulative probability of an extreme left tail empirical quantile, is too close to zero, the value $F_n(v)-D_\alpha(n)$ can be negative. In such case, the CI lower limit should be equated to $-\infty$ (or, better, to $-|A|$ with A an arbitrarily large value), since $\lim_{x\to 0} \Phi^{-1}(x) \to -\infty$. Analogously, at the opposite tail, when $F_n(v)+D_\alpha(n)$ is greater than one, the CI upper limit should be equated to $+\infty$ (or, better, to $|A|$ with A an arbitrarily large value), since $\lim_{x\to 1} \Phi^{-1}(x) \to +\infty$. In practice, since the Q-Q plot needs finite values to be drawn, consequently, the points with CIs partially unbounded will not be represented. Notwithstanding that, such points will be accounted for in the output computational sheets. Finally, GTest is designed for managing the presence of "N/A" (or −999.00) occurrences in the dataset.

GTest validation

The four methods provided by the proposed software have been validated at the 5 % significance level using 400 simulated Gaussian and non-Gaussian datasets in order to verify whether and how they provide the expected results. Table 1 summarizes the validation results expressed as a percentage ratio of test successes.

TEST1 runs the software with 200 Gaussian series (null hypothesis true), so we would expect the software to almost always accept the null hypothesis in accordance with the given level of significance. On the other hand, TEST2 runs it with 200 non-Gaussian series (null hypothesis false), so the software should be able to reject the null hypothesis a considerable number of times.

Table 1 Simulated data results

	Positive trials	
	Test 1	Test 2
Method	$1-\alpha$ (%)	$1-\beta$ (%)
K-S	100	30.5
Binom	97	74.5
Stabplot	96.5	78
Filliben	94	99

Finally, a TEST3 has been performed on real-world data. The selected test dataset is of particular interest since the data are expressed as percentage values, a kind of data often misused by practitioners.

Running of GTest

GTest is an MS Excel workbook, in .xlsm format, containing the VBA macro, and three service worksheets containing tables for computing critical values for each allowed test. These sheets are for internal software use only and they must not be deleted or modified.

Executing a GTest run is actually easy and consists in the following few, trivial steps:

1. Open GTest workbook.
2. Create a new worksheet.
3. Paste or type the dataset values in the active worksheet.
4. Run the macro.

At step 3, some rules need to be respected:

1. Even though GTest performs only univariate tests, multivariate datasets can be provided at this step.
2. The dataset needs to be organized as a case (rows) by variable (columns) table.
3. The maximum dataset size is constrained only by the spreadsheet limits.
4. The minimum allowed number of variable cases is six.
5. Variables (columns) do not need to have the same number of cases.
6. The first row of the dataset must contain variable headers.
7. Only quantitative variables are allowed.
8. Cases are not allowed for empty data.

Once the macro starts (step 4), a dialogue mask appears (Fig. 1), where the variable to be tested, the test method and the α level of significance can be input and the procedure started.

The test procedure consists in the following two computational modules. The first one standardizes the EDF values ($e'_{(i)}$) and computes the CDF quantiles ($c'_{(i)}$), while the second calculates the CLBs' ordinates at each CDF quantile. The results of these computational modules are saved in two separate spreadsheets, respectively.

Finally, a third module of GTest is dedicated to plotting the computed results (Fig. 2). As Fig. 2 clearly shows, GTest creates a new plot sheet showing the theoretical distribution (standardized Gaussian) as the bisector of the first quadrant, the standardized empirical quantiles as a scatterplot, and the related lower and upper CLBs.

The easy visual decision rule for accepting or rejecting the Gaussian hypothesis consists in evaluating whether the EDF scatterplot is completely within the acceptance region or whether even a single point falls outside the CLBs. In the first case, the EDF is considered statistically equivalent to the Gaussian distribution at the given level of significance.

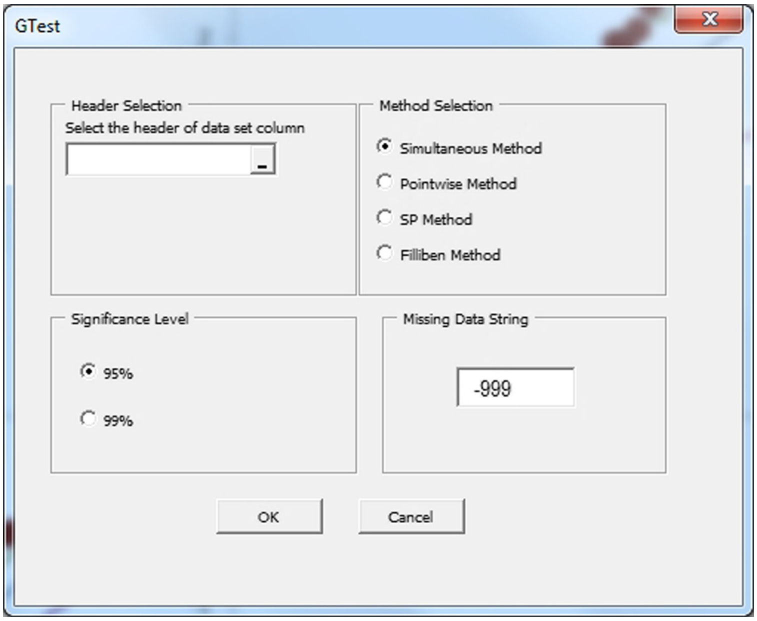
Case study

GTest has been validated using 400 simulated Gaussian and exponentially distributed datasets. GTest validation consisted in assessing the rate of success in recognizing the actual distribution. Afterward, a real case application was performed. The dataset used for testing the software consists of soil data: seven data series of soil parameters were used: content of clay (%), silt (%), sand (%) and organic matter (OM) (%), field capacity (FC) (%), permanent wilting point (PWP) (%) and pH (−). Soil data (2144 values in total) were obtained from the soil-parameter database of Apulia Region (southeastern Italy) and other datasets produced in various public research projects (Castrignanò et al. 2010). In particular, the dataset refers to a plain area of about 1979 km^2 called Capitanata (Fig. 3), which is the main agricultural area in the region. The processed data belong to the B soil horizon. The initial depth of the considered horizon ranges from 15 to 70 cm and the final depth from 38 to 110 cm. The objective of the real-world test is to compare the behaviour of the implemented methods against the specificity of soil data.

Results

TEST1 results

Table 1 shows the results of TEST1 in the second column. As already mentioned above, all the 200 data series of this test are Gaussian by construction, so it is now possible to estimate the probability of a type I error



GTest

Header Selection

Select the header of data set column

Method Selection

● Simultaneous Method

○ Pointwise Method

○ SP Method

○ Filliben Method

Significance Level

● 95%

○ 99%

Missing Data String

-999

Fig. 1 GTest input mask

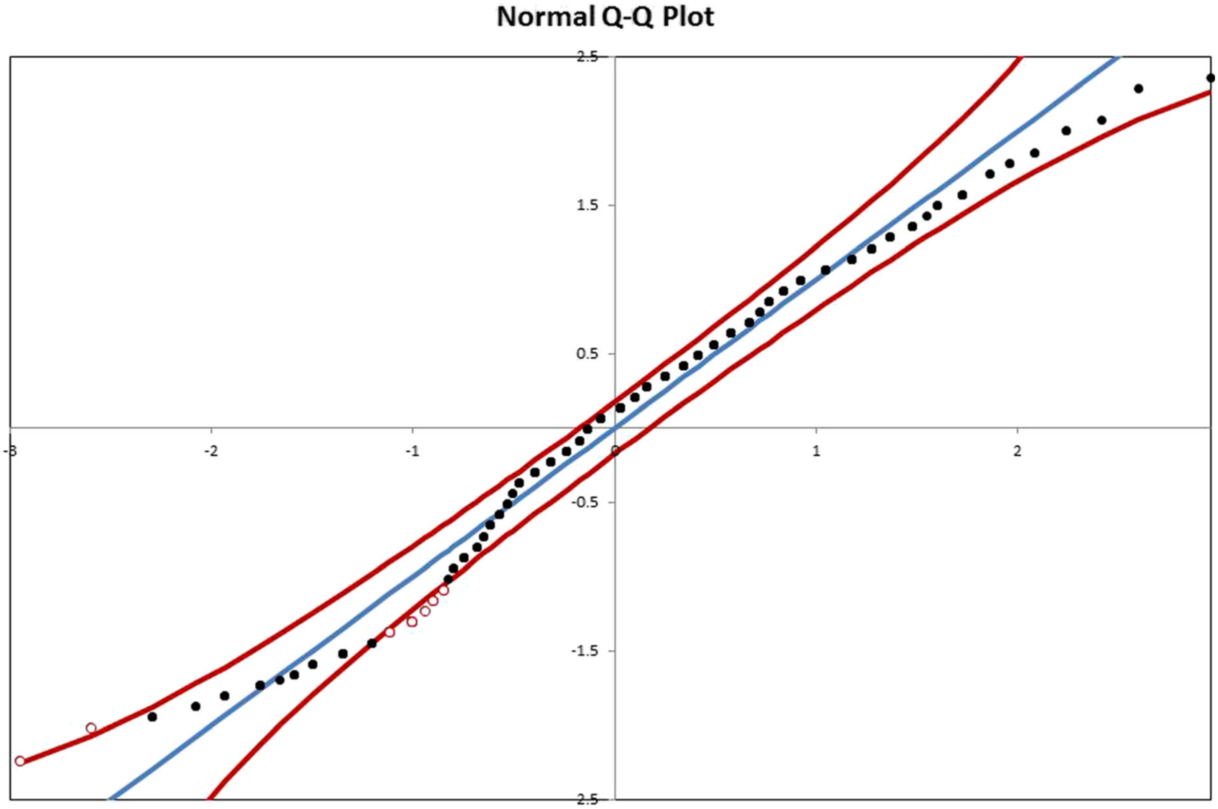


Fig. 2 Normal *Q*-*Q* plot

α (i.e. the probability of rejecting the null hypothesis when it is true) or rather its complementary probability (of accepting the null hypothesis when it is true) and to compare it with the assigned level of significance.

As Table 1 clearly shows, the K-S method accepts the null hypothesis for all the tested series, while for the binomial, stabilized Plot and Filliben methods, it is accepted in 97, 96.5 and 99 % of cases, respectively. Therefore, in general, the software is effectively capable of assessing the Gaussian behaviour of a given dataset, given that the percentage of positive trials is always high, independently of the method.

These last three methods seem to perform better than the K-S method, as the related positive trials are almost equal in percentage terms to the test level of significance (95 %). In contrast, the K-S method behaves as a conservative method. However, this is a well-known characteristic of this specific method, as reported in the technical literature (Steinskog et al. 2007).

In conclusion, we can say that the proposed software passed TEST1 while also correctly highlighting the specific characteristics of any of the methods applied.

TEST2 results

Column III of Table 1 shows the results of TEST2. This time all the data series are non-Gaussian (exponential with parameter $\lambda = 1$) by construction, so it is now possible to assess the probability of the type II error β (the probability of accepting the null hypothesis when it is false), or rather its complementary probability (of rejecting the null hypothesis when it is false), which is useful for evaluating the power of any test (Greene 2000). Obviously, we now expect the number of positive trials, in percentage terms ($1-\beta$), to be high.

In this case, the Filliben method seems to perform better than the others, given the high value of positive trials (99 %). The binomial and stabilized plot methods also perform well, producing, respectively, positive trials in 74.5 and 78 % of cases. Finally, the K-S method exhibits less power than the other methods, producing positive trials in only 30.5 % of cases, as expected (Razali and Wah 2011). Consequently, in this case too, we consider TEST2 passed.

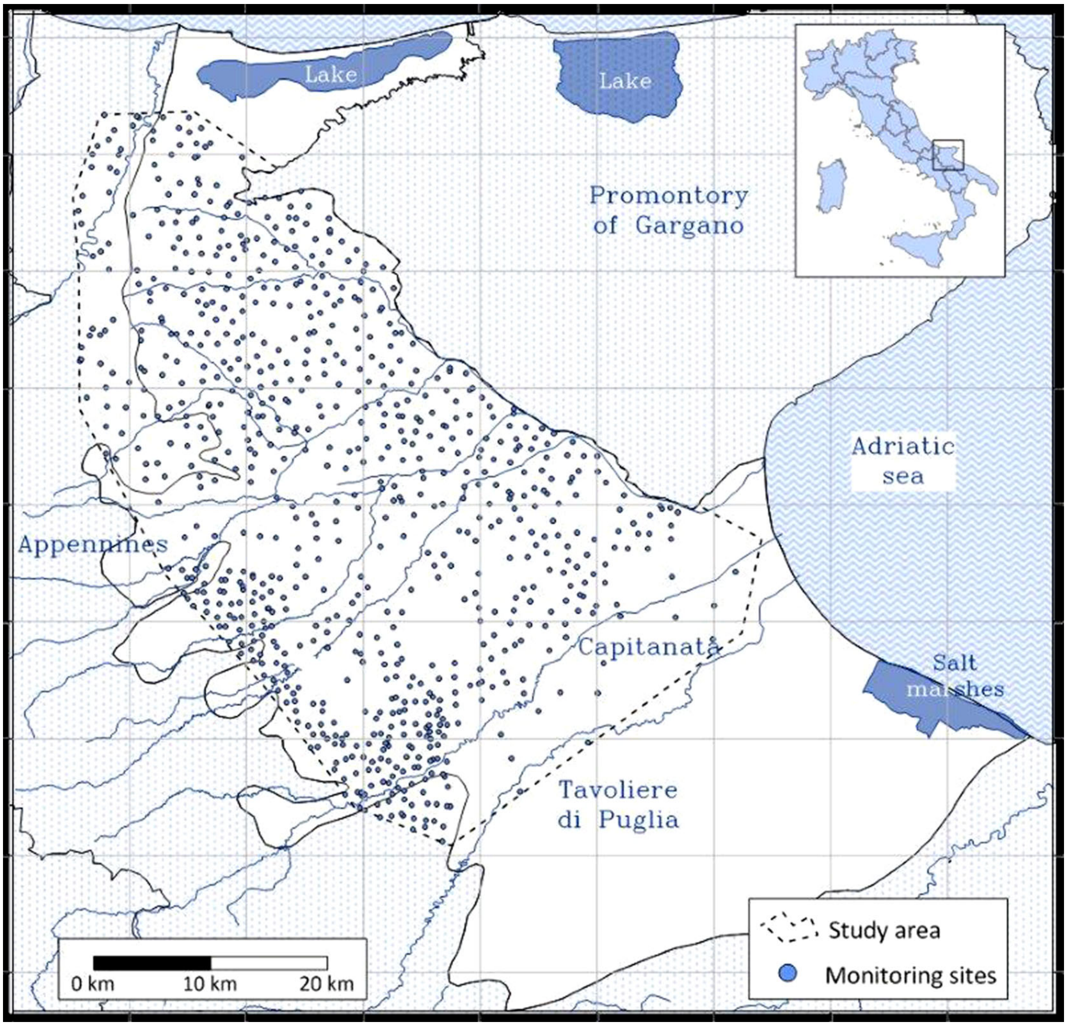


Fig. 3 Study area—Capitanata

TEST3 results

We did not have a priori knowledge of the related EDFs. Nevertheless, we expected an almost normal behaviour of most of the distributions. In fact, according to Table 2, all variables except pH show values of skewness and kurtosis near zero and very similar values of the mean and median (in this case including Ph), suggesting an approximately symmetrical behaviour of the EDFs. From Table 3, it is possible to summarize some characteristics of the four methods implemented. Let us premise that the number of tests is too low to make statistically meaningful inferences, but from a purely descriptive standpoint, it is evident that the stabilized plot method behaves differently from all of the other methods. In fact, the stabilized plot method shows the normality of the analysed datasets in more than 85 % of the cases. On the contrary, the remaining methods show the normality in only 43 % of the cases, even though the departure from the reference distribution is, in most of the cases, very slight. The different behaviour of the stabilized plot method depends jointly on the kind of processed data (percentage data) and the functional data transformation used by the method itself. As shown by Reinard (2006), percentage data cannot vary freely because they range from 0 to 100 %, so an arcsine-square root transformation will convert percentages into scores that are less skewed than the original data, equalizing the local variance of data. The example used showed that there are methods for testing normality that are more suited to the task at hand than others depending on the data to be processed: the sample size, degree of uncertainty contained in the measurements, and kind of data itself. Let us now summarize

Table 2 Report of the main descriptive statistics regarding the seven data series of soil parameters

Statistics	Sand (%)	Silt (%)	Clay (%)	OM (%)	pH (−)	FC (%)	PWP (%)
Mean	25.80	32.15	42.04	1.82	8.03	40.86	24.22
Std error	0.84	0.78	0.82	0.04	0.03	0.34	0.34
Median	24.00	34.00	42.00	1.72	8.08	41.09	24.49
Std deviation	14.95	13.94	14.70	0.80	0.39	6.08	5.93
Kurtosis	0.64	−0.75	−0.35	0.63	5.21	−0.38	−0.40
Skewness	0.76	−0.18	0.06	0.40	−1.10	−0.24	−0.38
Range	85.00	64.00	76.00	5.21	3.02	30.91	27.03
Minimum	0.00	1.00	7.00	0.20	6.09	22.92	7.10
Maximum	85.00	65.00	83.00	5.41	9.11	53.83	34.13
Count	320	320	320	373	180	312	312

the results of all the tests. First of all, it is evident that the tests carried out show the different behaviours of the four methods. However, it is not surprising that different methods based on different theories produce different results, that is, different acceptance regions. Concerning the K-S method, a strong tendency of this method to accept the null hypothesis is evident, even when it is false (high value of positive trials for TEST1 and low value for TEST2). However, the theory explains that this method strongly requires accurate knowledge of the parameters of the reference distribution (mean and standard deviation in the case of a normal distribution), which is seldom available in practical applications, when parameters are obviously estimated from samples. In this case, the K-S method tends to be conservative (i.e. the actual level of significance is greater than that given, and thus, the null hypothesis is rejected less often than would theoretically be correct); this may be a reasonable explanation for the greater propensity of the K-S method to accept the null hypothesis of normality compared with the other methods for both simulated and true data. The binomial method and, in particular, the stabilized plot method definitely behave better than the first one, although they are probably sensitive to the length of the data series (number of cases). The Filliben method deserves a separate mention, because it appears to be more restrictive than the other methods when acting on true data. With regard to the power of respective tests, it is known that a test with a power greater than 0.8 (or $\beta \leq 0.2$) is considered statistically robust (Mazen et al. 1985). Therefore, on comparing the results with information known from the literature, particularly Michael (1983) and Filliben (1975), it is possible to see that the methods implemented actually perform well, with the exception of the K-S method (D'Agostino and Stephens 1986), which was, however, chosen for its widespread use, as mentioned earlier. In Fig. 4 one can appreciate the magnified detail of the overlap of all four acceptance regions provided by the implemented methods for the clay variable. A wider or tighter acceptance region characterizes more or less conservative methods in accordance with their theoretical features. Table 4 summarizes the more appropriate use of the four methods according to the user's objective or the kind of dataset to be tested.

Table 3 Results of all methods applied to the seven data series of soil properties are reported

	Sand	Silt	Clay	OM	Ph	FC	PWP
K-S	NG	NG	G	G	NG	G	NG
Binom	NG	NG	G	G	NG	G	NG
Stabplot	G	NG	G	G	G	G	G
Filliben	NG	NG	G	G	NG	G	NG

G Gaussian, *NG* not Gaussian

Finally, the results from the tests show that the software presented:

- Has a strong capacity to provide robust information about the normal behaviour of a given distribution
- Produces a close correspondence of the results of methods implemented with their theoretical description
- Provides a clear graphical representation in addition to an extremely easy-to-use interface

This leads us to conclude that the software presented in this paper can be considered a solid, reliable tool for assessing the normality of a given distribution.

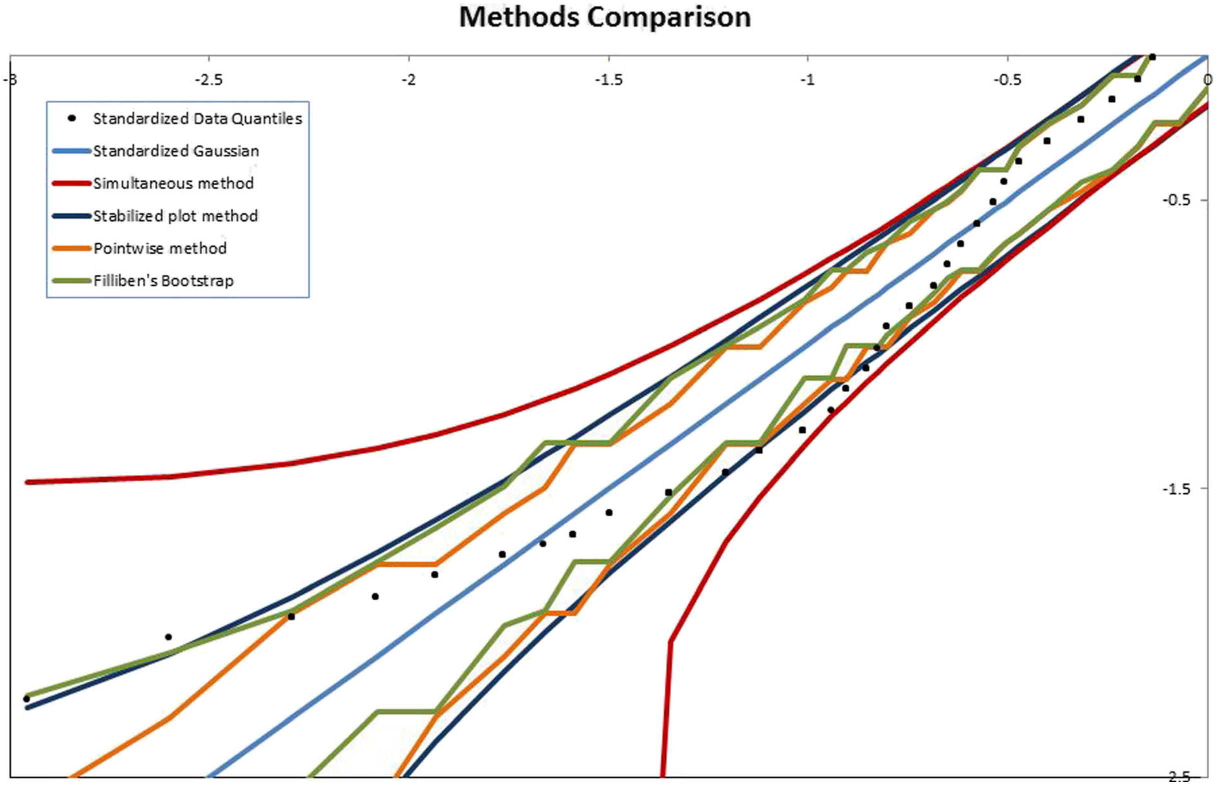


Fig. 4 Comparison of methods' acceptance regions

Conclusions

This paper presents software for testing the assumption of the normality of a given empirical distribution function (EDF) in graphical form. The software has been developed in VBA for MS Excel© and provides results in the form of normal Q-Q plots with the related confidence limit bands (CLBs), which are assessed by means of four different algorithms. The acceptance region bounded by the CLBs makes the understanding of the test straightforward and objective. The four algorithms come from different approaches and offer different levels of accuracy. The first is based on the *binomial distribution* and the second on the *Kolmogorov-Smirnov test*, while the last one represents the so-called *stabilized plot* based on a suitable data transformation. A fourth method is also provided, based on Filliben's *probability plot correlation coefficient* where the Q-Q plot is associated with a bootstrapped acceptance region. The software GTest runs with a simple graphical interface which allows the user to select one of the four available methods, define the α level of significance and perform calculations using the powerful statistical tools available in MS Excel©. The possibility of comparing results from different tests should allow the user to be confident of accepting the right hypothesis. Furthermore, given that real data can be affected by different degrees of error and uncertainty, the user can decide to apply conservative or restrictive methods on a case-by-case basis with regard to the supposed level of uncertainty.

Table 4 Suggestions for the application of the proper test to a given dataset or objective

Method	Objective
K-S	Find an approximate normal dataset
Binom	Multipurpose
Stabplot	Test a percentage dataset
Filliben	Find a strict normal dataset

This paper also presents comprehensive software testing based on 400 runs of simulated data series whose distributions were respectively normal and exponential by construction. The tests show the software to be particularly affordable and reliable while providing results as expected. The four graphical methods behaved as described in the scientific literature. Specifically, the binomial and stabilized plot methods perform very well, while the K-S method behaves conservatively, as expected. Finally, the implementation of Filliben's method seems to behave substantially well. A practical application of the proposed software on a dataset of seven soil properties shows good results according to the expected behaviour of the soil variables. In particular, the stabilized plot was shown to be very suitable for application to percentage soil data, showing the usefulness of having different tools to tackle the same problem. In the light of the reported results, the authors consider the software to be extremely reliable as well as easy to use. Nevertheless, further studies are now underway in order to improve the software's computational power, user interface and graphical appearance.

References

Barca, E., & Passarella, G. (2008). Spatial evaluation of the risk of groundwater quality degradation: a comparison between disjunctive kriging and geostatistical simulation. *Environmental Monitoring and Assessment, 137*(1–3), 261–273.

Barca, E., Passarella, G., & Uricchio, V. F. (2008). Optimal extension of the rain gauge monitoring network of the Apulian regional consortium for agricultural defense. *Environmental Monitoring and Assessment, 145*(1–3), 375–386.

Beaulieu-Prevost, D. (2006). Confidence intervals: from tests of statistical significance to confidence intervals, range hypotheses and substantial effects. *Tutorial in Quantitative Methods for Psychology, 2*(1), 11–19.

Calzada M. E., Scariano S. M. (2002).Visual EDF software to check the normality assumption. *Electronic Proceedings of the Fifteenth Annual International Conference on Technology in Collegiate Mathematics*. Orlando, Florida, 31 October – 3 November 2002, Paper C022.

Castrignanò, A., De Benedetto, D., Girone, G., Guastaferro, F., & Sollitto, D. (2010). Characterization, delineation and visualization of agro-ecozones using multivariate geographical clustering. *Italian Journal of Agronomy, 5*, 121–132.

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). New York: John Wiley and Sons.

D'Agostino R., Stephens M. (1986). *Goodness-of-fit techniques*. Marcel Decker

Devaney J. (1997), Equation discovery through global self-referenced geometric intervals and machine learning. Ph.D thesis, George Mason University, Fairfax, VA.

Diggle P. J., Ribeiro P. J. Jr (2007). *Model-based geostatistics*. Springer Series in Statistics

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics (American Society for Quality), 17*(1), 111–117.

Glantz S. (2005) Primer of biostatistics. McGraw-Hill (6 ed).

Gnanadesikan, R., & Wilk, M. B. (1968). Probability plotting methods for the analysis of data. *Biometrika, 55*(1), 1–17.

Greene, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River: Prentice Hall.

Hazen, A. (1930). *Flood flows. A study of frequencies and magnitudes*. New York: Wiley.

Hogg, R. V., & Tanis, E. A. (1977). *Probability and statistical inference*. New York: MacMillan Publishing.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd ed.). New York: Wiley.

Keeling, K. B., & Pavur, R. J. (2011). Statistical accuracy of spreadsheet software. *The American Statistician, 65*(4), 265–273.

Looney, S. W., & Gulledge, T. R., Jr. (1985). Use of the correlation coefficient with normal probability plots. *The American Statistician, 39*(1), 75–79.

Masciale, R., Barca, E., & Passarella, G. (2011). A methodology for rapid assessment of the environmental status of the shallow aquifer of "Tavoliere di Puglia" (Southern Italy). *Environmental Monitoring and Assessment, 177*(1–4), 245–261.

Mazen A., Magid M., Hemmasi M., Lewis M. F. (1985). In search of power: a statistical power analysis of contemporary research in strategic management. *Academy of Management Proceedings*, 30–34.

Michael, J. R. (1983). The stabilized probability plot. *Biometrika, 70*(1), 11–17.

Nash, J. C. (2006). Spreadsheets in statistical practice—another look. *The American Statistician, 60*(3), 207–289.

Ott W. R. (1995). *Environmental statistics and data analysis*. Lewis Publishers

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21–33.

Reinard J. C. (2006). *Communication research statistics*. Sage Publications

Rochowicz, J. A., Jr. (2010). Bootstrapping analysis, inferential statistics and EXCEL. *Spreadsheets in Education (eJSiE), 4*(3), 1–23.

Royston, P. (1993). Graphical detection of non-normality by using Michael's statistic. *Journal of the Royal Statistical Society: Series C: Applied Statistics, 42*(1), 153–158.

Steinskog, D. J., Tjøstheim, D. B., & Kvamstø, N. G. (2007). A cautionary note on the use of the Kolmogorov-Smirnov test for normality. *American Meteorological Society, 135*(3), 1151–1157. doi:10.1175/MWR3326.1.

Stirling W. D. (1982). Enhancements to aid interpretation of probability plots. *The Statistician*, 31(3)

Sutherland, W. J., Spiegelhalter, D., & Burgman, M. (2013). Policy: twenty tips for interpreting scientific claims. *Nature, 503*, 335–337. doi:10.1038/503335a.

Thode, H. C., Jr. (2002). *Testing for normality*. New York: Marcel Dekker. ISBN 0-8247-9613-6.

Wheater, C. P., & Cook, P. A. (2000). *Using statistics to understand the environment. Introductions to Environment Series* (1st ed.). London: Routledge. 246 p. ISBN 0-415-19887-9.