Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 925-930
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Influence of the Fitted Straight Line for Confidence Bands Algorithm in Q-Q Plots

Sonia Castillo-Gutiérrez, María Dolores Estudillo-Martínez, Emilio Damián Lozano-Aguilera

Department of Statistics and Operations Research, University of Jaén, Jaén, Spain
Email: socasti@ujaen.es

How to cite this paper: Castillo-Gutiérrez, S., Estudillo-Martínez, M.D. and Lozano-Aguilera, E.D. (2021) Influence of the Fitted Straight Line for Confidence Bands Algorithm in Q-Q Plots. *Open Journal of Statistics*, **11**, 925-930.
https://doi.org/10.4236/ojs.2021.116054

Received: October 8, 2021
Accepted: November 27, 2021
Published: November 30, 2021

Abstract

Confidence bands in a Normal Q-Q Plot allow us to detect non-normality of a data set rigorously, and in such a way that the conclusion does not depend on the subjectivity of the observer of the graph. In the construction of the graph, it is usual to fit a straight line to the plotted points, which serves both to check the hypothesis of normality (linear configuration of the plotted points) and to produce estimates of the parameters of the distribution. We can opt for different types of lines. In this paper, we study the influence of five types of fitted straight lines in a Normal Q-Q Plot used for construction the confidence bands based on the exact distribution of the order statistics.

Keywords

Normal Q-Q Plot, Confidence Bands, Fitted Straight Line, Goodness of Fit

1. Introduction

Normal probability plots and, in particular, Normal Q-Q Plots, are used to determine if a set of observations derives from a normal distribution. For this, it is necessary that the plotted points on the graph have a rectilinear configuration.

Normal Q-Q Plot compares the empirical quantiles of sample data, *i.e.*, the ordered sample data, $Q_x(p_i) = x_{(i)}$, with the corresponding quantiles of a theoretical distribution, *i.e.*, the normal distribution, $Q_t(p_i) = \Phi^{-1}(p_i)$. Therefore, the plotted points on the graph are the pairs $\left(\Phi^{-1}(p_i), x_{(i)}\right)$ where Φ is the standard normal cumulative distribution function and $p_i, i = 1, \cdots, n$ are the plotting positions. In the literature, several definitions of plotting positions are available [1] [2].

In the development of this paper, we will use the definition proposed by Yu and Huang [3]:

$$p_i = \frac{i - 0.326}{n + 0.348}, \quad i = 1, \cdots, n. \tag{1}$$

On a Normal Q-Q Plot, we can represent a straight line enabling us to take a decision about the straight form of the points on the graph and determine if the hypothesis of normality is verified. There are also different lines that we can represent on the graph [4].

The main problem of this graphical technique is that the observer of the graph may affect the conclusion. That is why this technique is often called "informal technique". To avoid this problem, the confidence bands or acceptance region [5] are used to determine whether or not a data set has a normal distribution, so that the conclusion is the same regardless of the observer of the graph. Some of the confidence bands depend on the straight line represented on the Normal Q-Q Plot to be able to be constructed.

Therefore, the plotting positions, the fitted straight line and the confidence bands are key elements in a Normal Q-Q Plot. Due to the high number of combinations of these three elements that exist, it is necessary to analyze the influence that the use of different combinations can have on the final conclusion. In this study, we will focus on the analysis of five types of straight lines and on the confidence bands based on the exact distribution of the order statistics [5].

Here, we focus on the normal distribution. However, the study can be extended to any distribution of interest.

This paper is organized as follows: in Section 2, we explain the five straight lines that we have used in this study. Section 3 presents the confidence bands based on the exact distribution of the order statistics. In Section 4, two examples illustrate the performance provided. Finally, in the last section, the conclusions of this study are presented.

2. Fitted Straight Lines in a Q-Q Plot

In this section, we carry out a review of some of the straight lines which can be fitted in a Q-Q Plot [4] and that we will use in our study to verify the influence they have on the confidence bands.

1) Straight line that passes through the first and third quartiles. This procedure consists of locating a point on the graph corresponding to the first quartile and another corresponding to the third quartile and joining these two points.

2) The least-squares line. The straight line, in our case, will take the form:

$$x = \mu + \sigma z \tag{2}$$

and the estimation of μ and σ will be obtained by using the unweighted least squares method. The solution in the case of normal distribution is the following:

$$\tilde{\sigma} = \frac{\sum z_i x_{(i)}}{\sum z_i^2}, \quad \tilde{\mu} = \overline{x} \tag{3}$$

and the fitted straight line is: $x = \tilde{\mu} + \tilde{\sigma} z$, where $x_{(i)}$ are the ordered observations and z_i are the N (0, 1) quantiles in the plotting positions p_i.

3) Straight line with slope the quasi-standard deviation s and constant the average of the data set. This method consists of fitting the straight line to the plotted points: $x = \overline{x} + sz$ where $\overline{x}$ is the average of the observations.

4) Theil-Sen's line [6]. The slopes of the lines passing through all possible pairs of points are calculated. Then, the median of all previous slopes is taken as an estimate of the slope. For the calculation of the constant, n constants of the lines through each of the points and the previously estimated slope are calculated. The estimated constant of the straight line will be the median of the n constants obtained.

5) Tukey's line [7]. This method consists of dividing the set of observations into three equal parts and calculating the median for each of them and determining the straight line from the three medians. The steps to obtain Tukey's line of general expression $x = a + by$ are the following:

a) Given the observations: $\left(z_1, x_{(1)}\right), \cdots, \left(z_n, x_{(n)}\right)$, they are divided into three groups with an approximately equal number of elements according to the variable *z*.

b) For each group the median is calculated by obtaining the following points:

$$\left(\tilde{z}_L, \tilde{x}_L\right), \left(\tilde{z}_C, \tilde{x}_C\right), \left(\tilde{z}_R, \tilde{x}_R\right) \tag{4}$$

where $\tilde{z}_L$ is the median of the left group, $\tilde{z}_C$ is the median of the central group and $\tilde{z}_R$ is the median of the right group of the observations of *z*. Similar to the observations of *x*.

c) The slope of Tukey's line is calculated by the following expression:

$$b = \frac{\tilde{x}_R - \tilde{x}_L}{\tilde{z}_R - \tilde{z}_L} \tag{5}$$

d) The constant of Tukey's line is calculated by the following expression:

$$a = \frac{\left(\tilde{x}_R + \tilde{x}_C + \tilde{x}_L\right) - b\left(\tilde{z}_R + \tilde{z}_C + \tilde{z}_L\right)}{3} \tag{6}$$

3. Confidence Bands Based on the Exact Distribution of the Order Statistics

The procedure to obtain the confidence bands based on the exact distribution of the order statistics is [5]:

Step 1 Fix the significance level *α*.

Step 2 Draw a Normal Q-Q Plot and fit a straight line. The fitted straight line provides an estimate of the parameters μ and σ of the normal distribution.

Step 3 Determine, for each *i*, $i = 1, \cdots, n$, the values $\Phi_i\left(p_1^{(i)}(\alpha)\right)$ and $\Phi_i\left(p_2^{(i)}(\alpha)\right)$ as the quantiles of order $\alpha/2$ and $1-\alpha/2$ of a $Beta(i, n-i+1)$ distribution.

Step 4 Determine the values $p_1^{(i)}(\alpha)$ and $p_2^{(i)}(\alpha)$, for each *i*, as the value Φ^{-1} in the quantiles calculated in the previous step. Φ is the distribution function of a normal distribution with parameters μ and σ. The values of μ and σ are the values obtained in Step 2.

Step 5 Plot, for each *i,* vertically, an interval centered on the corresponding point of the fitted straight line with the lower end of the band as the point $p_1^{(i)}(\alpha)$ and the upper end as the point $p_2^{(i)}(\alpha)$.

Step 6 Join the points calculated in the preceding step to obtain a band.

Step 7 Reject the hypothesis of normality if at least $\alpha\%$ of the observations fall outside the confidence bands.

4. Examples

In this section, we show two examples of how to construct Normal Q-Q Plot using confidence bands. First, considering simulated data and, secondly, with real data. The examples have been made using R [8].

4.1. Example 1

Table 1 shows a simulated size 30 sample of a Cauchy distribution.

Figure 1 shows a Normal Q-Q Plot constructed from the above observations. The plotting position considered, p_i, is that of Yu and Huang [3]. However, any other plotting position could be used to construct the Normal Q-Q Plot. The plot also represents the confidence bands based on the exact distribution of the order statistics. To obtain these confidence bands, we have considered a straight

Table 1. Simulated sample of a Cauchy distribution

0.64526129	−0.23126506	−2.89827739	0.00419841	−0.33761901	2.76499664
−0.41264397	−2.00742570	−0.12076996	0.44170899	0.36764003	0.23110927
0.38854176	2.61359225	−0.00184472	2.79458447	−1.65745029	0.07538504
0.90409775	2.40883027	2.06627019	−0.20223268	−1.35559281	−1.99385347
0.34705189	−4.50703372	0.02023511	−0.29384267	0.48452742	1.40517152

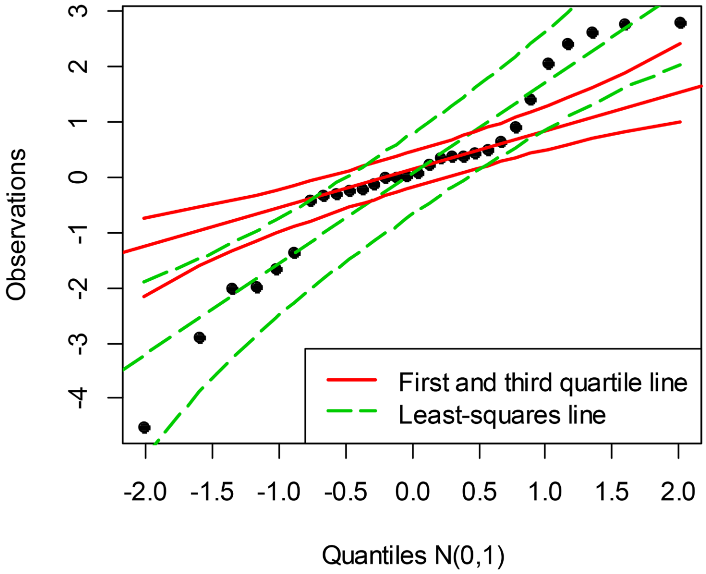

Figure 1. Normal Q-Q Plot with confidence bands using simulated data.

line that passes through the first and third quartiles and the least-squares line. It can be observed that the hypothesis of normality of observations is rejected according to the confidence bands obtained by considering the straight line that passes through the first and third quartiles, but it is not rejected according to that obtained by the least-squares line, although the data comes from a Cauchy distribution.

4.2. Example 2

The data set shown in **Table 2** comes from Bickel and Doksum [9] and lists the elapsed times spent above a certain high level for a series of 66 wave records taken at San Francisco Bay.

Following the same procedure as in the previous example, we have obtained **Figure 2**.

In **Figure 2**, we can observe that the hypothesis of normality is rejected according to the confidence bands obtained by considering the straight line that passes through the first and third quartiles (there are 5 points outside the confidence bands, more than $\alpha = 5\%$ of the data). Instead, it is not rejected according to that obtained by the least-squares line (there are 3 points outside the proposed confidence bands, less than $\alpha = 5\%$ of the data).

Table 2. Data set from Bickel and Doksum.

2.968	2.097	1.611	3.038	7.921	5.476	9.858	1.397	0.155	1.301	9.054
1.958	4.058	3.918	2.019	3.689	3.081	4.229	4.669	2.274	1.971	10.379
3.391	2.093	6.053	4.196	2.788	4.511	7.300	5.856	0.860	2.093	0.703
1.182	4.114	2.075	2.834	3.968	6.480	2.360	5.249	5.100	4.131	0.020
1.071	4.455	3.676	2.666	5.457	1.046	1.908	3.064	5.392	8.393	0.916
9.665	5.564	3.599	2.723	2.870	5.453	4.091	3.716	6.156	2.039	1.582

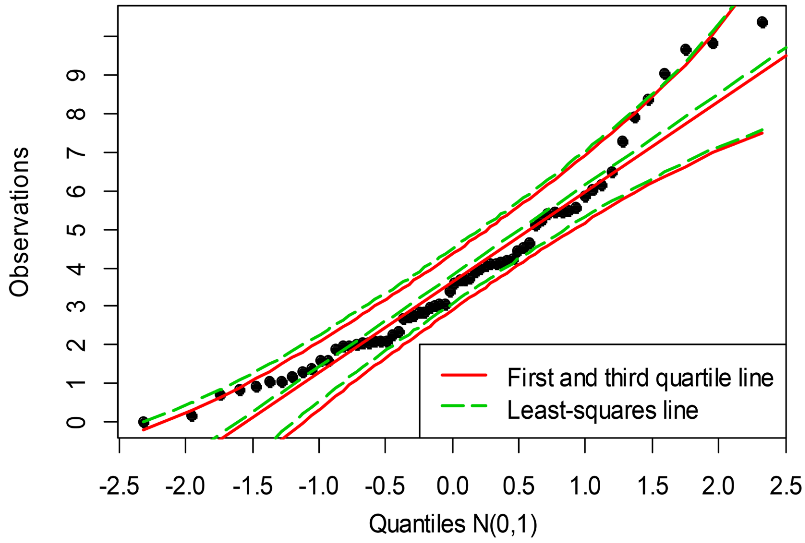

Figure 2. Normal Q-Q Plot with confidence bands using real data.

5. Conclusions

The aim of this work has been to analyze the influence of different types of straight lines that can be represented in a Normal Q-Q Plot at the moment of detecting the non-normality of a set of observations. Confidence bands represented in Q-Q Plot depend on the fitted straight line, so if we change the straight line, the confidence bands also change, and the conclusion may be different.

There are three elements that can vary in a Normal Q-Q Plot: plotting positions, confidence bands and straight lines. We have focused on the plotting positions proposed by Yu and Huang [3]. In [5] out of the three graphic techniques compared, the best method proves to be the confidence bands based on the exact distribution of the order statistics, so in this study, we have used such confidence bands. Therefore, we have fixed these two elements and we have compared the graphics obtained with five types of straight lines. The final conclusion is that the election of straight line for construction of confidence bands in a Normal Q-Q Plot it can change the decision about whether or not the data comes from a Normal distribution. Therefore, special care must be taken about the line to choose when building a Normal Q-Q Plot.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Castillo-Gutiérrez, S., Lozano-Aguilera, E. and Estudillo-Martínez, M.D. (2012) Selection of a Plotting Position for a Normal Q-Q Plot. R Script. *Journal of Communication and Computer*, **9**, 243-250.

[2] Cunnane, C. (1978) Unbiased Plotting Positions. A Review. *Journal of Hydrology,* **37**, 205-222. https://doi.org/10.1016/0022-1694(78)90017-3

[3] Yu, G.-H. and Huang, C.-C. (2001) A Distribution Free Plotting Position. *Stochastic Environmental Research and Risk Assessment,* **15**, 462-476. https://doi.org/10.1007/s004770100083

[4] Castillo-Gutiérrez, S., Lozano-Aguilera, E. and Estudillo-Martínez, M.D. (2012) A New Proposal to Adjust a Straight Line to a Normal Q-Q Plot. *Journal of Mathematics and System Science*, **2**, 327-333.

[5] Estudillo-Martínez, M.D., Castillo-Gutiérrez, S. and Lozano-Aguilera, E. (2013) New Confidence Bands on Q-Q Plots to Detect Non-Normality. *International Journal of Computer Mathematics,* **90**, 2137-2146. https://doi.org/10.1080/00207160.2013.792920

[6] Theil, H. (1950) A Rank Invariant Method for Linear and Polynomial Regression Analysis I, II, III. *Proceedings of Koninklijke Nederlandse Akademie van Wetenschappen Serie A*, **53**, 386-392.

[7] Tukey, J.W. (1977) Exploratory Data Analysis. Addison-Wesley.

[8] R Development Core Team (2008) R: A Language and Environment for Statistical Computing, Vienna, Austria. http://www.r-project.org/

[9] Bickel, P.J. and Doksum, K.A. (1977) Mathematical Statistics: Basic Ideas and Selected Topics. Holden-Day, San Francisco.