
A modified Q-Q plot for large sample sizes¹

Gráfico Q-Q modificado para grandes tamaños de muestra

Jorge Iván Vélez^a
jorge.velez@anu.edu.au

Juan Carlos Correa Morales^b
jccorrea@unal.edu.co

Abstract

The Q-Q plot is a graphical tool for assessing the goodness-of-fit of observed data to a theoretical distribution in which every single observation in the data is represented by a symbol. In many occasions, due to either natural variations of the data or to a large sample size, the Q-Q plot could be interpreted as a sign of failure of the proposed model. One alternative is to consider a special set of characteristics of the data such as the sample quantiles that, jointly with its theoretical counterparts, allow the user to effectively compare both. We propose and illustrate a modified Q-Q plot that helps to visualise the differences between the observed quantiles and their corresponding theoretical values, and overcome some technical problems of the traditional Q-Q plot.

Palabras clave: Q-Q Plot, statistical graphics, Goodness-of-Fit.

Resumen

El gráfico Q-Q es una herramienta para determinar si los datos observados se ajustan a una distribución de probabilidad teórica, en el que cada observación en los datos es representada por un símbolo. En muchas ocasiones, debido a variaciones naturales en los datos o un gran tamaño de muestra, el gráfico Q-Q puede interpretarse como una falla en el modelo probabilístico propuesto. Una alternativa es considerar un conjunto de características de los datos tales como los cuantiles muestrales que, en conjunto con su equivalente teórico, permitan al usuario comparar ambos de manera efectiva. Proponemos e ilustramos un gráfico

¹Vélez, J.I., Correa, J.C. (2015), A modified Q-Q plot for large sample sizes. *Comunicaciones en Estadística*, **8**(2), 163-172.

^aArcos-Burgos Group, Department of Genome Sciences, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia; Neurosciences Research Group, University of Antioquia, Medellín, Colombia; Research Group in Statistics, National University of Colombia at Medellín, Medellín Colombia

^bAssociate Professor, Department of Statistics, National University of Colombia at Medellín, Medellín, Colombia; Research Group in Statistics, National University of Colombia at Medellín, Medellín, Colombia

Q-Q modificado que permite visualizar las diferencias entre los cuantiles observados y teóricos, y remediar algunas dificultades técnicas del gráfico tradicional.

Keywords: Gráfico QQ, gráficos estadísticos, bondad de ajuste.

1 Introduction

Graphical data analysis is an important step towards the understanding of a statistical problem (Tukey 1977, Fienberg 1979, Wainer 1981, Tufte 1983, Tukey 1990, Wainer 1990). Despite its apparent simplicity and the fact that the design of effective statistical graphics has extensively been discussed by several authors (Cleveland 1985, Burn 1993), it is often the case that a graphic conveys the wrong impression and leads to the misinterpretation of the information represented there (Cleveland & McGill 1985, Wainer 1984).

The Q-Q plot is a common tool to analyse the goodness-of-fit of sample data to a theoretical distribution (Wilk & Gnanadesikan 1968, Easton & McCulloch 1990, Marden 1998, Marden 2004, Dhar et al. 2014); it allows the user to compare a theoretical model, represented by a 45° slope line, with an empirical quantile function represented by all sample points¹. However, there are several drawbacks with the Q-Q plot. First, sample variation, sometimes, makes it difficult to interpret the plot, especially the behaviour in the tails (DasGupta 1985). We illustrate this situation in Example 1 where data from a normal distribution is generated, but natural variation in the data leads to wrongly conclude, based on the Q-Q plot, that the normal distribution is not a suitable probabilistic model for the data. Second, the computational difficulties when the sample size is large, and third, the well-known difficulties due to human perception (Cleveland 1985, Cleveland & McGill 1985). For instance, it has been shown that the comparison between two functions with different slopes is a difficult task unless one of them has zero slope, and that human perception is not even close to acceptable when comparing differences between two functions unless one of them is constant (Cleveland 1985, Cleveland & McGill 1985).

In this paper, we propose a modification of the traditional Q-Q plot such that some of the known issues (Nair 1982, Rosenkrantz 2000) and drawbacks previously mentioned are overcome. This modification uses the sample percentiles and compare them with their expected value under the theoretical distribution F_0 . Subsequently, we plot these differences and, to make even simpler the analysis of the resulting plot, compute and plot point-wise confidence intervals for them. Our proposal is illustrated through three simulated data sets, and an implementation of this approach is provided in R (R Core Team 2014).

¹In the case of normally-distributed data, with mean μ and variance σ^2 , the model is represented by the equation $y = \mu + \sigma x$, with x and y the theoretical and sample quantiles, respectively. If a straight line of equation $y = x$ is desired, then the data needs to be standardised. Another example is the QQ-plot for the Weibull distribution where the slope and intercept depend on the parameters of the distribution. However, after estimating these parameters, the slope of the resulting line associated with the QQ-plot has a 45° slope.

2 Modified Q-Q plot

In general terms, we are interested in checking whether the data comes from a pre-specified theoretical distribution F_0 by testing a hypothesis of the form

$$H_0 : X \sim F_0 \quad (1)$$

against a suitable alternative hypothesis, say H_1 . In the expression above, X is a random variable, F_0 is the hypothetical cumulative distribution function of the model that depends on the parameter vector θ , and it is assumed that F_0 has inverse F_0^{-1} . If θ is unspecified, we estimate it by using a consistent procedure. Now, assuming a large sample size, $F_\theta \approx F_{\hat{\theta}}$.

Let $X = (x_1, \dots, x_n)$ be a random sample of size n from an unknown distribution F , $\alpha \in (0, 1)$ be the type I error probability, $\xi_{p_i} = F_0^{-1}(p_i)$ be the i th theoretical percentile, $\hat{\xi}_{p_i}$ its sample estimator, and

$$\Delta_i = \hat{\xi}_{p_i} - \xi_{p_i} \quad (2)$$

the difference between what is observed and what is to be expected under the probability distribution F_0 . Under the normality assumption, $E[\Delta_i] = 0$ as $n \rightarrow \infty$.

Our proposal of a modified Q-Q plot is based on the following result in Serfling (1980, pp. 80). Let $0 < p_1 < \dots < p_k < 1$. Suppose that F has a density f in the neighborhoods of $\xi_{p_1}, \dots, \xi_{p_k}$, and that f is positive and continuous at $\xi_{p_1}, \dots, \xi_{p_k}$. Then $(\hat{\xi}_{p_1}, \dots, \hat{\xi}_{p_k})$ is asymptotically normal with mean vector $(\xi_{p_1}, \dots, \xi_{p_k})$ and covariance $n^{-1}\sigma_{ij}$, where

$$\sigma_{ij} = \frac{p_i(1-p_j)}{f(\xi_{p_i})f(\xi_{p_j})} \quad \text{for } i \leq j$$

is the ij -th element of the variance covariance matrix of X , and $\sigma_{ij} = \sigma_{ji}$.

Now, to check whether F_0 is a plausible model, we proceed as follows:

Define $0 < p_1 < \dots < p_k < 1$, and compute the sample percentiles $\hat{\xi}_{p_i}$, $i = 1, \dots, k$.

Compute the theoretical percentiles $\xi_{p_i} = F_0^{-1}(p_i)$, $i = 1, \dots, k$.

Compute Δ_i , $i = 1, 2, \dots, k$, as in (2).

Calculate the standard deviation of each sample percentile as

$$\sigma_{\hat{\xi}_{p_i}} = \sqrt{\frac{p_i(1-p_i)}{n \{f_0(F_0^{-1}(p_i))\}^2}},$$

and the $100(1 - \alpha)\%$ confidence interval as

$$(-z_{1-\alpha/2} \sigma_{\hat{\xi}_{p_i}}, z_{1-\alpha/2} \sigma_{\hat{\xi}_{p_i}}) \quad (3)$$

with z_γ the γ percentile of the standard normal distribution.

Plot Δ_i against p_i , and draw the corresponding confidence interval at each p_i , $i = 1, 2, \dots, k$. Empirical evaluation suggests that using $2 \leq k \leq 9$ is sufficient to determine whether the distribution of interest fits the data well.²

3 Examples

In this section we illustrate our approach with four sets of simulated data.³

Example 1: Normally distributed data. Let $X = (x_1, \dots, x_n)$ be a random sample of size $n = 100$ from a standard normal distribution, e.g., $X \sim N(0, 1)$. In figure 1 we present both the classic and modified Q-Q plots; the former was constructed using the `qqnorm()` and `qqline()` functions of R, and the latter using our implementation in the same statistical language.

Does the data follow a Normal distribution? Observe that in figure 1(a) the behaviour of the sample points is erratic and misleading, especially in the tails, which may lead us to conclude that the data is not normally distributed. Conversely, the modified Q-Q plot using the deciles is easier to read and interpret, and the conclusion is direct: the data follows a Normal distribution. In addition, the confidence interval around each decile gives a better sense of the sample variation, which is not possible using the classical Q-Q plot. A Shapiro-Wilks normality test confirms what we already knew ($W = 0.9793$, p -value = 0.1167).

Example 2: Exponentially distributed data. In this example, we simulate $n = 50$ observations from an Exponential distribution with parameter $\lambda = 1$, e.g., $X \sim \text{Exp}(1)$. As in Example 1, we constructed both the classic and modified Q-Q plots for this data (see figure 1(c)–(d) for more information), the latter using the 10th, 20th, 30th, 40th and 50th percentiles.

Note that even when the data is generated from an Exponential distribution, figure 1(c) seems not to show this properly. For instance, several points are below the expected value under theoretical distribution being tested, which may lead the data analyst to wrongly conclude that the data is not exponentially distributed. Despite natural variation in the sample, the modified Q-Q plot in figure 1(d), on the contrary, shows that the difference between the observed and theoretical

²Here, a value of $k = 5$ implies the comparison of the 20th, 40th, 60th and 80th percentiles of the sample distribution with those of the theoretical distribution of interest, F_0 .

³R code for generating the results presented herein is at the end of this document.

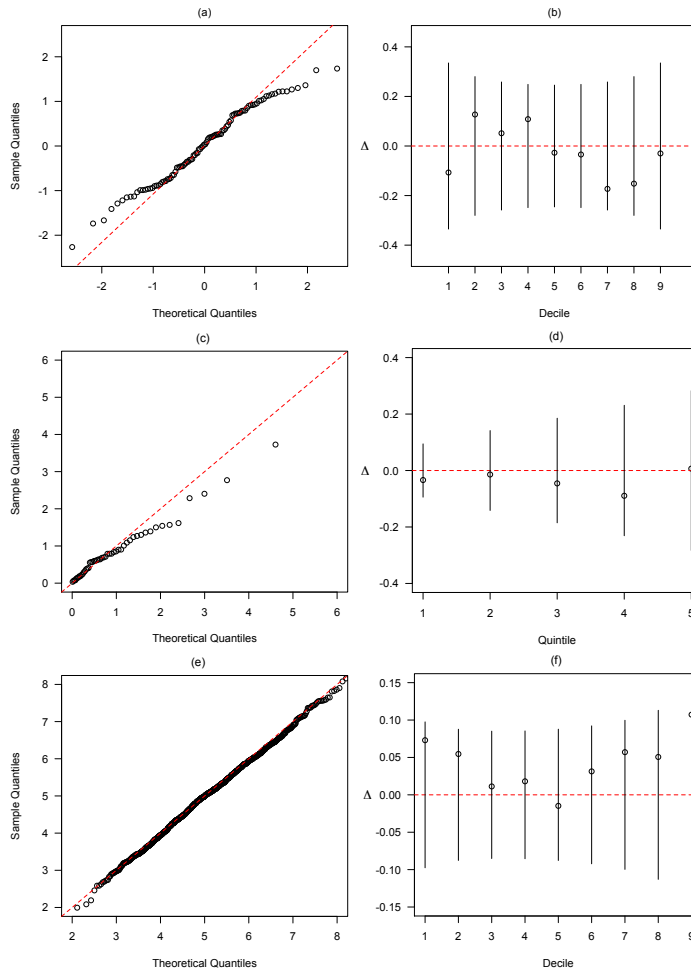


Figure 1: *Classic (left) and modified (right) Q-Q plots for the simulated data in the Examples section⁵. Source: Own elaboration.*

percentiles under the Exponential distribution is within the limits of the 95% confidence intervals. Hence, we can conclude that the simulated data comes, in fact, from an Exponential distribution. It is also worth noting the “triangular” shape of the 95% confidence intervals compared to the “U” shape displayed in Example 1.

Example 3: Gamma distributed data. Consider $n = 1000$ observations from a Gamma distribution with parameters $\alpha = 20$ and $\beta = 4$, e.g., $X \sim \text{Gamma}(20, 4)$. The classic and modified versions of the Q-Q plot are presented in figures 1(e)–(f).

In this example, the sample size can be considered “large” compared to that used in examples 1 and 2. Three important points are worth noting: (1) both the classic and modified Q-Q plots indicate that, in fact, the sample data follows a Gamma distribution; (2) in the classic Q-Q plot, the behaviour in the tails is not as bad as it was in examples 1 and 2 (this is consequence of a considerable increase in the sample size); (3) the modified Q-Q plot, once again, shows that the expected difference between the theoretical and sample percentiles is within the limits of the 95% confidence intervals.

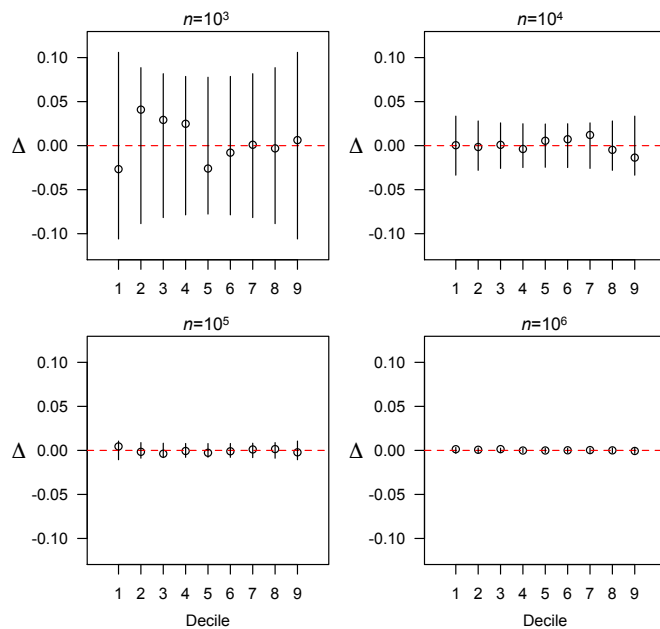


Figure 2: Modified Q-Q plot for different sample sizes when $X \sim N(0, 1)^7$. Source: Own elaboration.

Example 4: Normally distributed data with variable sample size. Here we consider samples of size $n = 10^3, 10^4, 10^5$, and 10^6 from a standard normal distribution. Our results are presented in figure 2. Note that, as n increases, $\Delta \rightarrow 0$. Hence, the length 95% confidence intervals decreases.

4 Discussion

Q-Q plots have become one of the most used graphical tools for verifying whether a particular statistical distribution (e.g., the normal distribution) fits the data. In this paper, we have described and exemplified a modification of the Q-Q plot, using percentiles, that overcomes some of the technical problems of the classic Q-Q plot (and which have been addressed long time ago but many users are unaware of).

Three main aspects of this modification are noteworthy. First, the graphical representation is similar to that in residual plots used in regression analysis and hence can be interpreted likewise. However, it is important to take into account that in our procedure the horizontal line at zero represents the theoretical distribution, and both the observed percentiles and their corresponding confidence intervals fluctuate around it. Second, if the difference between the observed and theoretical percentile falls outside the $100(1 - \alpha)\%$ confidence interval, it indicates that the probability distribution F_0 does not fit the data well in that percentile. Clearly, this type of comparison cannot be done using the traditional Q-Q plot. Third, a similar approach based on the difference between the theoretical and observed values of the cumulative distribution F_0 was proposed by van der Loo (2010) to detect univariate outliers; the method is robust against chosen parameter settings and uses a Q-Q plot to show these differences and label data points as “not belonging to the bulk” (van der Loo 2010). Likewise, Ueda (1996/2009) proposed the detection of discordant outliers using the cumulative distribution of Normal distribution. A recent simulation study concluded that Ueda’s method is sensitive to outliers when the distribution is not symmetric, and that such sensitivity increases with the sample size (Marmolejo-Ramos et al. 2015).

Three examples are shown to illustrate the use of the QQ-plot proposed herein. Of particular interest is Example 3 in which simulated data from a Gamma distribution is analysed to determine whether, in fact, this distribution fits the data. Although in this example the parameters and the distribution generating the data were known, the example reinforces the usefulness of the method. A natural improvement of the proposed plot is the automatic estimation of the parameters for some known distributions.

Future research plans can be aimed at calculating, instead of point-wise, simultaneous $100(1 - \alpha)\%$ confidence intervals (Roy & Bose 1953, Schaffer 1995, Benjamini & Hochberg 1995, Hsu 1996) for the percentiles in the modified Q-Q plot, as well as to systematically study, using statistical simulation, its performance. The derivation of a formal statistical test based on the modified Q-Q plot may also be an area of research worth pursuing.

5 Computational details

R code for all examples is available from the first author by request, or can be downloaded in https://dl.dropboxusercontent.com/u/9601860/qplotcode/suppl_material.pdf.

6 Acknowledgements

We are grateful to Mr. Jairo Ángel from the Instituto de Matemática e Estatística at the Universidade de São Paulo, Brazil for critical reading of this manuscript. We also thank the Editor and two anonymous reviewers for their insightful comments and suggestions, which vastly improved a previous version of this manuscript. JIV was supported by the Eccles Scholarship in Medical Sciences, the Fenner Merit Scholarship, and the Australian National University (ANU) High Degree Research Scholarship. JIV thanks Dr. Mauricio Arcos-Burgos from ANU for his support. ⁸

Recibido: 10 de diciembre del 2014

Aceptado: 31 de marzo del 2015

References

- Benjamini, Y. & Hochberg, Y. (1995), ‘Controlling the false discovery rate: A practical and powerful approach to multiple testing’, *Journal of the Royal Statistical Society, Series B (Methodological)* **57**(1), 389 – 400.
- Burn, D. A. (1993), ‘Designing Effective Statistical Graphs’, *Handbook of Statistics* **9**, 828–833.
- Cleveland, W. & McGill, R. (1985), ‘Graphical Perception and Graphical Methods for Analyzing Scientific Data’, *Science* **229**(4716), 828–833.
- Cleveland, W. S. (1985), *The Elements of Graphing Data*, 1 edn, Wadsworth: Monterey.
- DasGupta, A. (1985), *The Use and Abuse of the Q – Q Plot: Some Asymptotic Theory*, Technical Report #95-30, Department of Statistics, Purdue University.
- Dhar, S. S., Chakraborty, B. & Chaudhuri, P. (2014), ‘Comparison of Multivariate Distributions using Quantile-Quantile Plots and Related Tests’, *Bernoulli* **20**(3), 1484–1506.
*<http://projecteuclid.org/euclid.bj/1402488947>
- Easton, G. S. & McCulloch, R. E. (1990), ‘A Multivariate Generalization of Quantile-Quantile Plots’, *Journal of the American Statistical Association* **85**(410), 376–386.
- Fienberg, S. E. (1979), ‘Graphical Methods in Statistics’, *The American Statistician* **33**(4), 165–178.

⁸The authors have indicated they have no financial relationships relevant to this article to disclose.

- Hsu, J. C. (1996), *Multiple Comparison: Theory and Methods*, Chapman & Hall, Great Britain.
- Marden, J. I. (1998), 'Bivariate QQ plots and Spider Web plots', *Statistica Sinica* **8**, 813–826.
- Marden, J. I. (2004), 'Positions and QQ Plots', *Statistical Science* **19**(4), 606–614.
- Marmolejo-Ramos, F., Vélez, J. I. & Romão, X. (2015), 'Automatic detection of discordant outliers via the Ueda's method', *Journal of Statistical Distributions and Applications* **2**(1), 8.
*<http://www.jsdajournal.com/content/2/1/8>
- Nair, V. N. (1982), 'Q-Q Plots with Confidence Bands for Comparing Several Populations', *Scandinavian Journal of Statistics* **9**(4), 1993–200.
- R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*<http://www.R-project.org/>
- Rosenkrantz, W. A. (2000), 'Confidence Bands for Quantile Functions: A Parametric and Graphic Alternative for Testing Goodness of Fit', *The American Statistician* **54**(3), 1985–190.
- Roy, S. N. & Bose, R. C. (1953), 'Simultaneous confidence interval estimation', *The Annals of Mathematical Statistics* **24**(4), 513–536.
- Schaffer, J. P. (1995), 'Multiple Hypothesis Testing: A Review', *Annu. Rev. Psychol.* **46**, 561–84.
- Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley: New York.
- Tufte, E. (1983), *The Visual Display of Quantitative Information*, Graphics Press: Cheshire.
- Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley Publishing Company: Reading, Massachusetts.
- Tukey, J. W. (1990), 'Data-Based Graphics: Visual Display in the Decades to Come', *Statistical Science* **5**(3), 327–339.
- Ueda, T. (1996/2009), 'A simple method for the detection of outliers', *Electronic Journal of Applied Statistical Analysis* **2**(1), 67–76.
- van der Loo, M. P. (2010), *Distribution-based outlier detection for univariate data*, Technical Report 10003, Statistics Netherlands, The Hague/Heerlen.
- Wainer, H. (1981), 'Graphical Data Analysis', *Annual Review of Psychology* **32**(1), 191–204.

- Wainer, H. (1984), 'How to Display Data Badly', *The American Statistician* **38**(2), 137–147.
- Wainer, H. (1990), 'Graphical Visions from William Playfair to John Tukey', *Statistical Science* **5**(3), 340–346.
- Wilk, M. B. & Gnanadesikan, R. (1968), 'Probability Plotting Methods for the Analysis of Data', *Biometrika* **55**(1), 1–17.