Explaining Normal Quantile-Quantile Plots through Animation: The Water-Filling Analogy

Robert A. Stine

Department of Statistics

The Wharton School of the University of Pennsylvania

Philadelphia, PA 19104-6340

September 9, 2016

Abstract

A normal quantile-quantile (QQ) plot is an important diagnostic for checking the assumption of normality. Though useful, these plots confuse students in my introductory statistics classes. A water-filling analogy, however, intuitively conveys the underlying concept. This analogy characterizes a QQ plot as a parametric plot of the water levels in two gradually filling vases. Each vase takes its shape from a probability distribution or sample. If the vases share a common shape, then the water levels match throughout the filling, and the QQ plot traces a diagonal line. An R package `qqvases` provides an interactive animation of this process and is suitable for classroom use.

Key words: Education, diagnostic, simulation.

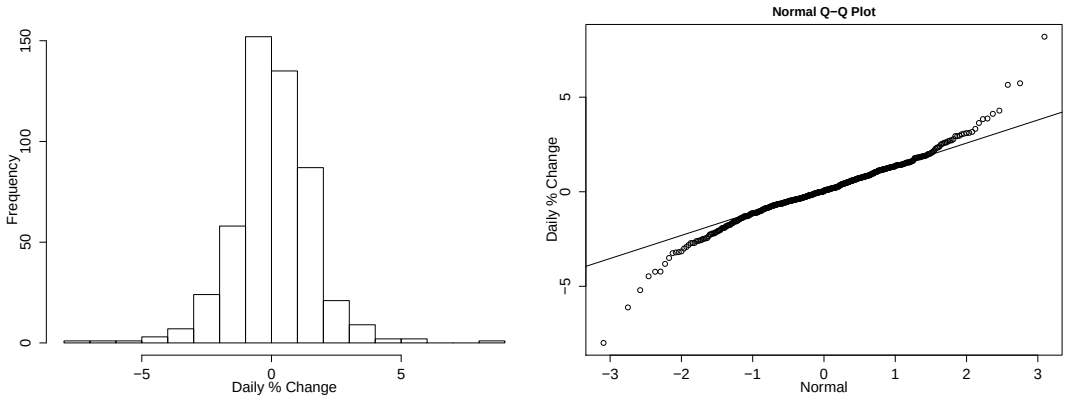


Figure 1: *Though hard to judge from the histogram, the normal QQ plot shows that the distribution of daily percentage changes in the value of Apple stock in 2014-2015 has thicker tails than a normal distribution.*

1 Introduction

Normal QQ plots are an important visual diagnostic, but one that can be hard to explain. Students quickly learn that if the data in a normal QQ plot deviate from a diagonal reference line, then the assumption of normality is questionable. For example, Figure 1 illustrates the difficulty of judging normality from a histogram. The data are daily percentage changes in the value of Apple stock in 2014-2015. The histogram is bell-shaped, but the distribution has thicker tails than anticipated by normality. The deviations from the diagonal line in the normal QQ plot imply that, in the extremes, the data extend farther out than expected under normality. While they recognize its importance, many of my students have treated this plot as a graphical "black box": a useful diagnostic that relies on a magical mechanism. In the spirit of Brown and Kass (2009) and Cobb (2015), this paper offers a heuristic that makes these "fundamental concepts accessible."

2 Water-filling analogy

A normal QQ plot compares the shape of the empirical distribution of a sample to the shape of a normal distribution. To set up the analogy, consider comparing the shape of a continuous distribution to that of the normal. Quantile plots graph percentiles of the distributions and therein lies the difficulty for students. Many of my students

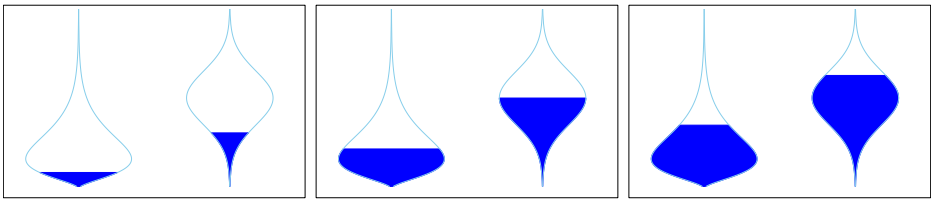


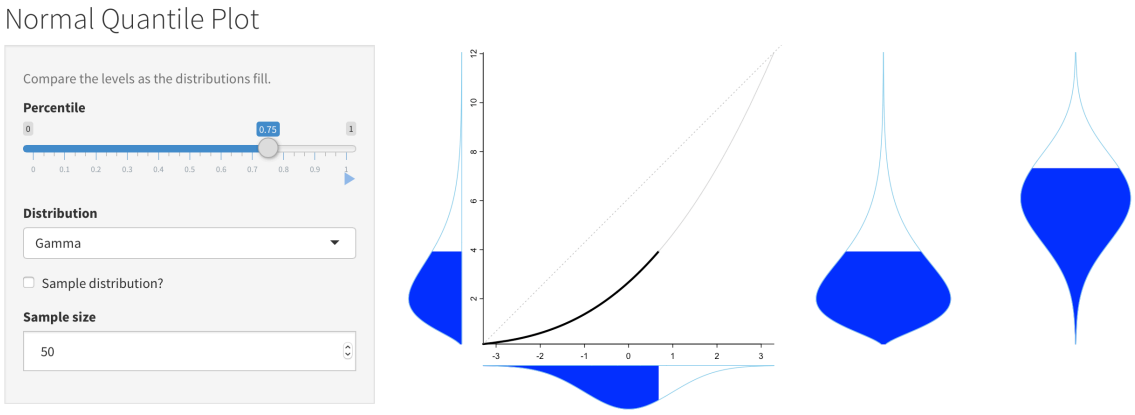
Figure 2: *Different water levels in two vases reflect the different shapes of the underlying distributions. The pair in the left frame are 10% full, then 50% full in the middle frame, and finally 80% full on the right.*

find a cumulative distribution confusing enough without considering its inverse. A simple analogy, however, makes quantile functions more approachable. Namely, quantile functions are analogous to water levels as we fill transparent vases.

Figure 2 illustrates the idea. The "vase" on the left of each frame of Figure 2 is formed by gluing a gamma density function with its mirror image. Similarly, the vase on the right of each frame is a Gaussian vase. Both vases have the same "volume" (really, area). To limit the heights of the vases, both are truncated at the 0.0005 and 0.9995 percentiles. (I call these containers vases because of their resemblance to vase plots (Benjamini, 1988).) Now imagine simultaneously filling the two vases at equal rates with water, as suggested by the sequence of plots in the figure. The left frame shows the two vases, initially 10% full. The middle frame shows them filled to 50%, and the right frame shows them 80% full. If two vases have the same shape, then the water levels match throughout the filling. Otherwise, as in this example, one fills more slowly than the other, and the levels differ. The level in the gamma vase grows slowly at the start of the filling because it has a wide base. As the filling proceeds, the level in the gamma vase eventually catches up with the level in the Gaussian vase because both vases hold the same amount.

In this context, a normal QQ plot is the parametric plot of these water levels. The level of the Gaussian vase determines the coordinate on the x-axis, and the level of the other vase gives the coordinate for the y-axis. Figure 3 shows the normal quantile plot for the two vases in Figure 2, as rendered by the accompanying application. To make the linkage between this graph and the vases explicit, the figure displays halves of the vases (the density functions) along the respective axes. The quantile plot adds

Figure 3: *This view of the open-source application shows the normal QQ plot for a gamma distribution. Adjacent vases reinforce the water-filling analogy as the quantiles increase.*



a diagonal line to make it easier to identify curvature.

Controls in the application shown in Figure 3 allow interactive modifications. The percentile slider controls the water level; for example, moving the slider to the right increases the water level and extends the curve in the plot. Other controls change the distribution that defines the y-axis; choices include a normal distribution, the shown gamma distribution (with shape parameter 3), a beta distribution, t-distributions (with 3 and 6 degrees of freedom), and a mixture of a normal and gamma.

3 Empirical QQ plots

Applying this analogy to the normal QQ plot of data requires more work and imagination for two reasons. First, it would not make sense to fill a "discrete vase" – the water would leak out. Second, normal QQ plots of data should include bands that indicate whether deviations from the diagonal are large enough to imply a significant departure from normality.

To address the first problem, statistics offers a variety of smooth density estimates, but these estimates are unfamiliar to students taking introductory statistics. For example, a kernel density provides a continuous density estimate, and these have been used to enhance boxplots (Hintze and Nelson, 1998). A kernel density estimate, however, diverts attention from the QQ plot to itself. Rather than take that route, then, the accompanying software shows a histogram of data. (The statistics package JMP

adopts a similar presentation.) Because observations in a tall histogram bin are relatively closely packed, the heights of the bins in the histogram inversely convey the average rate of water-filling within an interval.

For the second problem, there are simple approximations that quantify the size of a departure from the reference line. Students recognize that there is a problem when they see deviations like those in Figure 1, but the decision of whether a deviation is "large enough" (statistically significant) is a tough call for a student who is new to QQ plots. Bands remove this subjectivity by indicating how much departure from the reference line can be produced by sampling variation. A student doesn't have to guess whether the data drift far from the reference line; they can see whether points fall outside the bands.

For setting the bounds in the normal quantile plot, a variety of elaborate methods are available. Aldor-Noiman, Brown, Buja, Rolke, and Stine (2013) review several powerful proposals, but the accompanying animation requires limits that can be computed quickly. With that in mind, bounds for the deviation from normality first tabled in Lilliefors (1967) work nicely. These tables adjust for the use of estimated parameters in the normal distribution. The tables give the value c_α such that

$$\lim_n P(\sup_x |F_n(x) - \widehat{\Phi}(x)| \le c_\alpha/\sqrt{n}) \approx 1 - \alpha \ , \tag{1}$$

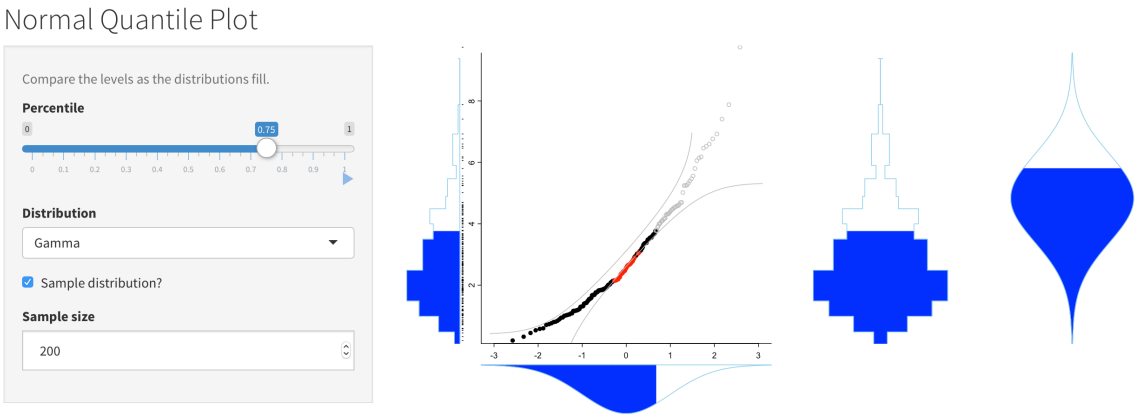
where F_n denotes the empirical distribution of the data and $\widehat{\Phi}$ denotes the cumulative normal distribution with estimated parameters $\overline{X}$ and s^2. The asymptotic critical value for $\sup_x |F_n(x) - \widehat{\Phi}(x)| \approx 0.89/\sqrt{n}$ if $\alpha = 0.05$.

For example, Figure 4 shows an example of an normal QQ plot of a sample of 200 observations from a gamma density, filled to the 75th percentile. Selecting the "Sample distribution?" checkbox in the application dialog produces an empirical QQ plot. A histogram replaces the distribution on the y-axis. Points in this sample drift outside the limits, indicating a statistically significant departure from normality. (These are highlighted in red in the figure.)

4 Discussion

I must admit that you do not need a computer animation to teach quantile plots this way. The water-filling analogy alone seems to take the mystery out of QQ plots. I

Figure 4: *This animation shows the normal QQ plot of a sample of 200 observations from a gamma distribution.*



sketch two vases side-by-side on a blackboard, say that each vase holds a liter, and gradually "fill" the vases by coloring in the levels with chalk or a marker. Students are quick to recognize that the levels – the quantiles – remain equal only if the vases have the same shape. In that case, a graph of the level of one vase versus the level of the other falls along a diagonal line. I then sketch a blackboard version of Figure 2 and ask the students to tell me how the graph of the levels will look. This discussion is also a nice opportunity to convey what is meant by the shape of a distribution.

For those who want to use software, the animation can be run either on-line or installed locally. Those with less interest in R can run the application remotely by pointing their browser to `http://gosset.wharton.upenn.edu:3838/stine/qqvases/`. Readers familiar with R can download the package `qqvases` from the CRAN repository or from links on my web page `www-stat.wharton.upenn.edu/~stine/`. The software exploits `shiny`, a library for R used here to render plots with interactive controls in a browser window (Chang, Cheng, Allaire, Xie, and McPherson, 2015). Running the command `qq_vase_plot` locally allows the user to customize the application, such as adding more distributions and generating QQ plots of data.

In addition to explaining QQ plots, the software can be used to illustrate other fundamental concepts, such as Type I errors (the quantile plot of a sample from a normal distribution has data outside the bands) and power (the bands in an QQ plot become tighter as the sample size grows). It can be surprising to see how hard it is to recognize that small samples from a gamma distribution differ significantly from the

normal. Some students do ask about the origin of the confidence bands. I don't have a simple analogy explaining those, so I use the question as an opportunity to advertise more advanced courses.

It is not unusual to see quantile plots explained with the help of showing distributions along the axes (e.g., the cover and Figure 3.4 of Verzani, 2005), but to my knowledge authors have not exploited the water-filling analogy and resulting animation. This analogy is used to explain normal quantile plots in a textbook citepstinefoster14, but I find that it works better when animated.

References

Aldor-Noiman, S., Brown, L. D., Buja, A., Rolke, W., and Stine, R. A. (2013), "The power to see: A new graphical test of normality," *The American Statistician*, 67, 249–260.

Benjamini, Y. (1988), "Opening the box of a boxplot," *The American Statistician*, 42, 257–262.

Brown, E. N. and Kass, R. E. (2009), "What is statistics?" *The American Statistician*, 63, 105–110.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2015), *shiny: Web Application Framework for R*, r package version 0.11.1.

Cobb, G. (2015), "Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up," *The American Statistician*, 69, 266–281.

Hintze, J. L. and Nelson, R. D. (1998), "Violin plots: A box plot-density trace synergism," *The American Statistician*, 52, 181–184.

Lilliefors, H. W. (1967), "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *Journal of the Amer. Statist. Assoc.*, 62, 399–402.

Verzani, J. (2005), *Using R for Introductory Statistics*, Boca Raton FL: Chapman & Hall/CRC.