13. Introduction to Quantile-Quantile (QQ) Plots

Quantile-Quantile (QQ) Plots I

6:08 / 6:08 1.0x

OK?

And so I'm just plotting those points.

Those guys are given by the H0 I'm testing.

Right?

For H0, I'm testing if F is equal to F0,

so this will be F0.

OK?

And I'm just testing if this is actually a good fit.

If you do another test, something else will come.

End of transcript. Skip to the start.

Video
Download video file

Transcripts
Download SubRip (.srt) file
Download Text (.txt) file

Concept Check: QQ Plots

1/1 point (graded)

A quantile-quantile (QQ) plot is an informal but useful method for goodness of fit testing. Suppose we have an i.i.d. sample $X_1, X_2, \ldots, X_n$ with unknown cdf F^*. We are interested in determining whether or not F^* is the cdf F of some known distribution. For example, we may set $F = \Phi_{0,1}$ to determine whether or not the sample comes from a standard Gaussian $\mathcal{N}(0,1)$.

The **quantile-quantile (QQ) plot** is constructed in the following way from a data set:

1. Reorder the samples to be in increasing order. Denote the reordered sample by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$.

2. Plot the points

$$\left(F^{-1}\left(\frac{1}{n}\right), X_{(1)}\right),\ \left(F^{-1}\left(\frac{2}{n}\right), X_{(2)}\right),\ \ldots,\ \left(F^{-1}\left(\frac{i}{n}\right), X_{(i)}\right),\ \ldots,\ \left(F^{-1}\left(\frac{n-1}{n}\right), X_{(n-1)}\right).$$

Note that above we omit plotting the n'th point because $F^{-1}(n/n) = F^{-1}(1) = \infty$. (In cases where $F^{-1}(1)$ is defined, we do not need to omit that point.)

Which of the following are true about quantile-quantile (QQ) plots? (Choose all that apply.)

- [x] A QQ plot provides a visual method for determining whether or not a data set has a certain distribution. ✔

- [ ] A QQ plot is a rigorous, formal method of goodness of fit testing. For example, it makes sense to talk about the type 1 error of a goodness of fit test.

- [ ] If n is very large and the points on the QQ plot lie very far from the line y = x, then it is reasonable to conclude that F and F^* are close as cdfs.

- [x] Looking at the graphs of the empirical cdf $F_n(t)$ and the cdf F, it can be difficult to tell if the two functions are close. The QQ plot transforms the cdf F and the empirical cdf $F_n(t)$ so that it is easier to compare the two visually. ✔

✔

Solution:

We examine the choices in order.

- The first choice is correct. If the empirical cdf $F_n(t)$ and cdf F are close, then we expect $F^{-1}(i/n) \approx X_{(i)}$. Hence, the data points

$$\left(F^{-1}(1/n), X_{(1)}\right),\ \left(F^{-1}(2/n), X_{(2)}\right),\ \ldots,\ \left(F^{-1}(i/n), X_{(i)}\right),\ \ldots,\ \left(F^{-1}((n-1)/n), X_{(n-1)}\right).$$

  should lie near the line $y = x$. Plotting these points, we can visually compare the distribution of our data to the distribution that has cdf F.

- The second choice is incorrect. In contrast to the methods we have studied before for hypothesis testing, a QQ plot is **not** a formal testing procedure. Rather, it conveniently visualizes the data so that we can get a sense of whether the distribution of our data matches some known distribution (for example, $\mathcal{N}(0,1)$).

- The third choice is incorrect. Rather, if n is very large and the points in the QQ plot are very far from the line $y = x$, this implies that the cdf of $X_1, \ldots, X_n$ and the cdf F are **not** close. More precisely, this would imply that F is **not** well-approximated by $F_n(t)$. On the other hand, if n is very large, then $F_n(t) \approx F^*$ by the Glivenko-Cantelli theorem. Therefore, if the QQ plot lies near the line $y = x$, we see that $F(i/n) \approx F_n(i/n)$ for all $1 \leq i \leq n$. In this situation it would be more reasonable to infer that $X_1, \ldots, X_n$ have cdf F.

- The fourth choice is correct. Indeed, it can be difficult to compare the empirical cdf $F_n(t)$ with another cdf F near the tail values. Taking the QQ plot compares the inverse cdf F^{-1} with a notion of the inverse of the empirical cdf (However, note that, strictly speaking, the empirical cdf is not invertible). This is much easier to visualize, because then we just have to check if the points in the QQ plot lie near the line $y = x$ to see if the distribution of our data is similar to the distribution with cdf F.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Discussion

Show Discussion

Topic: Unit 4 Hypothesis testing:Lecture 16: Goodness of Fit Tests Continued: Kolmogorov-Smirnov test, Kolmogorov-Lilliefors test, Quantile-Quantile Plots / 13. Introduction to Quantile-Quantile (QQ) Plots