**Q-Q Plots for Interpretation of IncuCyte Data**

Maia Smith

Feb 28, 2015

Q-Q plots are used to ascertain whether or not two sets of values come from the same distribution. There are two general types of Q-Q plots, those involving comparisons between sample and theoretical data, and those comparing two sample datasets.

With regard to the first type of Q-Q plot, one might hypothesize that one’s data is coming from a particular sort of distribution (ex. normal, log-normal, etc.). By plotting the quantiles of one’s sample data against the quantiles of theoretical data from the hypothesized distribution, the conformation of the sample data to the hypothesized distribution can be visualized. Should the sample and theoretical datasets originate from the same distribution—and provided the sample data is scaled—the plot will show a straight line through (0,0) and (0,1).

Deviants from the straight line can be eyeballed as either outliers or random variants. The best approach for doing so is to simulate many datasets from the specified distribution, and plot each as a Q-Q plot. Comparing these plots to the sample-theoretical plot in question, one may assess the potential significance of outliers. If the sample-theoretical Q-Q plot *does not* stick out from the simulated plots, it likely does not contain outliers. On the other hand, if the sample-theoretical Q-Q plot *does* stick out from the others, it may contain outliers. Moreover, if the sample-theoretical Q-Q plot has a distinct curvature (S-shaped, C-shaped, etc.), it may be coming from a different distribution entirely.

Aside from eyeballing, some statistical tests are available to assess whether the sample data is coming from a normal distribution. These include the Shapiro-Wilk test, the Anderson-Darling test and the Kolmogorov-Smirnov one sample test.

The second type of Q-Q plot is useful when one is questioning whether two sample datasets originate from the same or different distributions. By plotting the quantiles of each dataset against one another, the hypotheses can be confirmed one way or another. As with the sample-theoretical Q-Q plot, if the two datasets are scaled and originating from the same distribution, the plot will show a straight line through (0,0) and (0,1).

To eyeball deviations in a sample-sample Q-Q plot, the procedure may not be so simple as it is for sample-theoretical Q-Q plots. Simulated data is not a possibility, considering neither sample’s underlying distribution is known for certain. Therefore, deviations from a straight line must be somewhat ambiguously estimated as outliers or random variants. If the line is obviously shaped in an S or C fashion, perhaps the two datasets are arising from different distributions.

To statistically determine the similarity in distributions of both samples, the Kolmogorov-Smirnov two sample test can be employed. This test is based upon the empirical distribution functions of the two samples.

Applying these methods to the analysis of our IncuCyte data, we can use the sample-theoretical Q-Q plot to aid in assessing which distribution each curve metric is following, and whether or not outliers are present. We can perform statistical tests for the same effort as reinforcement. Using the sample-sample Q-Q plot can help us determine whether or not two of our curve metrics are following the same distribution. Again, this analysis can be reinforced by statistical tests.

**References**

Wilk, M. B., and R. Gnanadesikan. 'Probability Plotting Methods For The Analysis Of Data'. *Biometrika* 55.1 (1968): 1. Web.

DUDOIT, S., YANG, Y. H., CALLOW, M. J. AND SPEED, T. P. (2001). Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. Statistica Sinica, 12, 111–139.

plot, How. 'How To Interpret A QQ Plot'. Stats.stackexchange.com. N.p., 2015. Web. 2 Mar. 2015.

non-normality?, Interpreting. 'Interpreting Qqplot - Is There Any Rule Of Thumb To Decide For Non-Normality?'. Stats.stackexchange.com. N.p., 2015. Web. 2 Mar. 2015.