**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 (Wilk *et al.*, 1968). Q-Q plots are based on a dataset’s quantiles, which are values that divide a random variable’s inverse cumulative distribution function (CDF) at regular intervals (“Quantile”, n.d.). For instance, a 2-quantile divides the inverse CDF at Cumulative Probability = 0.5, and a 3-quantile divides the inverse CDF at Cumulative Probability = 0.33 and Cumulative Probability = 0.67.

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 (1,1) (Wilk *et al.*, 1968).

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 (“Interpreting Qqplot”, n.d.). Simulations can be performed using the following R functions:

**Normal Distribution** (“R: The Normal Distribution”, n.d.)

rnorm(n, mean = 0, sd = 1)

*n* is the number of observations

*mean* is the desired mean of your normal distribution

*sd* is the standard deviation of your normal distribution

**Lognormal Distribution** (“Lognormal {Stats}”, n.d.)

rlnorm(n, meanlog = 0, sdlog = 1)

*n* is the number of observations

*mean* is the desired mean of your lognormal distribution

*sd* is the standard deviation of your lognormal distribution

**Poisson Distribution** (“R: The Poisson Distribution”, n.d.)

rpois(n, lambda)

*n* is the number of observations

*lambda* is the variance and mean of the distribution

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 (“Interpreting Qqplot”, n.d.). Moreover, if the sample-theoretical Q-Q plot has a distinct curvature (S-shaped, C-shaped, etc.), it may have heavier tails, it may be skewed, or it may come from a different distribution entirely (“Examples of”, n.d.).

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 (Ghasemi *et al.,* 2012). The Shapiro-Wilk test operates by correlating the sample and theoretical data, and has the most power of all normality tests (Shapiro and Wilk, 1965; Razali and Wah, 2011). The Anderson-Darling test is based on the empirical cumulative distribution function (EDF), and looks at the squared differences between the sample EDF and theoretical EDF (Anderson and Darling, 1952). The Kolmogorov-Smirnov (K-S) test compares the sample CDF to the theoretical CDF (Kolmogorov, 1933; Smirnov, 1948). Note that the K-S test provides conservative estimates of Type I error (Massey, 1951).

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 (1,1) (Wilk *et al.*, 1968).

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 (“Examples of”, n.d.).

To statistically determine the similarity between distributions of both samples, the Kolmogorov-Smirnov two sample test can be employed (“R: Kolmogorov-Smirnov Tests”, n.d.). 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**

Anderson, T. W.; Darling, D. A. "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes". Annals of Mathematical Statistics (1952), 23: 193–212.

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.

"Examples of Interpreting Q-Q Plots." BYU-Idaho. Web. 3 Mar. 2015. <http://emp.byui.edu/BrownD/Stats-intro/dscrptv/graphs/qq-plot\_egs.htm>.

Ghasemi, Asghar, and Saleh Zahediasl. 'Normality Tests For Statistical Analysis: A Guide For Non-Statisticians'. International Journal of Endocrinology and Metabolism 10.2 (2012): 486-489. Web.

"How to Interpret a QQ Plot." Stack Exchange. Web. 3 Mar. 2015. <http://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot>.

"Interpreting QQplot - Is There Any Rule of Thumb to Decide for Non-normality?" StackExchange. Web. 3 Mar. 2015. <http://stats.stackexchange.com/questions/111010/interpreting-qqplot-is-there-any-rule-of-thumb-to-decide-for-non-normality>.

Kolmogorov A (1933). "Sulla determinazione empirica di una legge di distribuzione". G. Ist. Ital. Attuari 4: 83–91.

"Lognormal {stats}." R Documentation. Web. 3 Mar. 2015. <http://www.inside-r.org/r-doc/stats/dlnorm>.

Massey, Frank J. (1951). “The Kolmogorov-Smirnov test for goodness of fit. J. Amer. Stat. Assoc., 46:68-78.

Oztuna, D., Elhanm A., Tuccar, Ersoz (2006). Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions. Turk J Med Sci, 36 (3):171-176.

"Quantile." Wikipedia. Wikimedia Foundation. Web. 3 Mar. 2015. <http://en.wikipedia.org/wiki/Quantile>.

Razali, N. M., Wah, Y. B. (2011). “Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests”. Journal of Statistical Modeling and Analytics 2 (1): 21-33.

"R: Kolmogorov-Smirnov Tests." R Documentation. Web. 3 Mar. 2015. <https://stat.ethz.ch/R-manual/R-patched/library/stats/html/ks.test.html>.

"R: The Normal Distribution." R Documentation. Web. 3 Mar. 2015. <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Normal.html>.

"R: The Poisson Distribution." R Documentation. Web. 3 Mar. 2015. <https://stat.ethz.ch/R-manual/R-patched/library/stats/html/Poisson.html>.

Shapiro, S. S.; Wilk, M. B. "An analysis of variance test for normality (complete samples)". Biometrika (1965), 52 (3–4): 591–611.

Smirnov N (1948). "Table for estimating the goodness of fit of empirical distributions". Annals of Mathematical Statistics 19: 279–281.

Wilk, M. B., and R. Gnanadesikan (1968). 'Probability Plotting Methods For The Analysis Of Data'. *Biometrika* 55(1): 1.