

Quantile-Quantile Plot

Moo K. Chung
mchung@stat.wisc.edu

Update history October 5, 2003; Feb 5, 2007

1 Quantile Function

Checking the normality of imaging data is important when the underlying statistical model assumes the normality of data. But how do we know the data will follow normality? This is easily checked using the quantile-quantile (QQ) plot. The QQ-plot was first introduced by Wilk and Gnanadesikan (1968).

The quantile point q for random variable X is a point that satisfies

$$P(X \leq q) = F_X(q) = p,$$

where F_X is the cumulative distribution function (CDF) of X . Assuming we can find the inverse cdf,

$$q = F_X^{-1}(p).$$

This function is mainly referred to as a quantile function. Then we can define the quantile-quantile plot of two random variables X and Y to be a parametric curve $\mathcal{C}(p)$ parameterized by $p \in [0, 1]$:

$$\mathcal{C}(p) = (F_X^{-1}(p), F_Y^{-1}(p)).$$

2 Empirical Distribution

The CDF $F_X(q)$ measures the proportion of random variable X less than given value q . So by counting the number of measurements less than q , we can empirically estimate the CDF. Let $X_1, \dots, X_n$ be a random sample of size n . Then order them in increasing order

$$X_{(1)} = \min(X_1, \dots, X_n) \leq X_{(2)} \leq \dots \leq X_{(n)} = \max(X_1, \dots, X_n).$$

Suppose $X_{(j)} \leq q < X_{(j+1)}$. This implies that there are j samples that are smaller than q . So we approximate the CDF as

$$\widehat{F}_X(q) = \frac{j}{n}.$$

The j/n -th *sample quantile* is $X_{(j)}$. Some authors define this as the $(j - 0.5)/n$ -th sample quantile. The factor 0.5 is introduced to account for the discretization.

In computer implementation, it is easier to implement the empirical distribution using the *step function* $\mathcal{I}_q(x)$ which is defined as $\mathcal{I}_q(x) = 1$ if $x \leq q$ and $\mathcal{I}_q(x) = 0$ if $x > q$. Then the CDF is estimated as

$$\widehat{F}_X(q) = \frac{1}{n} \sum_{i=1}^n \mathcal{I}_q(X_i),$$

where $\mathcal{I}_q(X_i)$ counts if X_i is less than q . A different possibly more sophisticated estimation can be found in Frigge *et al.* (1989) The American Statistician. **43**:50-54.

3 Normal Probability Plot

As an illustration, let us plot the QQ-plot of two normal distributions. Suppose $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. Let $Z \sim N(0, 1)$ and let $\Phi(z) = P(Z \leq z)$, the CDF of the standard normal distribution. If we denote q_1 and q_2 to be the p -th quantiles for X and Y respectively, we have

$$p = P(X \leq q_1) = P\left(\frac{X - \mu_1}{\sigma_1} \leq \frac{q_1 - \mu_1}{\sigma_1}\right) = \Phi\left(\frac{q_1 - \mu_1}{\sigma_1}\right).$$

Hence the parameterized QQ-plot is given by

$$q_1(p) = \mu_1 + \sigma_1 \Phi^{-1}(p),$$

$$q_2(p) = \mu_2 + \sigma_2 \Phi^{-1}(p).$$

The QQ-plot without the parameter p is then given by

$$\frac{q_1 - \mu_1}{\sigma_1} = \frac{q_2 - \mu_2}{\sigma_2}.$$

This shows the QQ-plot of two normal distributions is the straight line.

This idea can be used to determine the normality of a given sample. We can check how closely the sample quantiles corresponds to the normal

distribution by plotting the QQ-plot of the sample quantiles vs. the corresponding quantiles of a normal distribution. This particular QQ-plot is referred to as the *normal probability plot*. In normal probability plot, we plot the QQ-plot of the sample vs. $N(0, 1)$.