- Theoretical Quantiles
- QQ Plot

Theoretical Quantiles, again

Now, if q_α is the α-quantile of some Distribution, and X is a r.v. from that Distribution, then

$$\mathbb{P}(X \leq q_\alpha) \geq \alpha \qquad \text{and} \qquad \mathbb{P}(X \geq q_\alpha) \geq 1 - \alpha.$$

Note: Here we are taking inequalities, and not, say, $\mathbb{P}(X \leq q_\alpha) = \alpha$, since, in the Discrete r.v. case, we can have no q_α with exact equality. Say, if $X \sim \textit{Bernoulli}(0.2)$, and $\alpha = 0.4$, then no q_α exists with $\mathbb{P}(X \leq q_\alpha) = \alpha$.

Note: If $\alpha = 0.5$, we call $q_\alpha = q_{0.5}$ to be the **Median of the Distribution**. So if we consider a Continuous r.v. and draw the PDF of that r.v., then the Median is the (leftmost) point dividing the area under the PDF curve into 50%-50% portions.

Theoretical Quantiles, again

Later we will use a lot quantiles. When constructing Confidence Intervals or Hypothesis Testing, we will use Quantiles of the Normal Distribution, t-Distribution, χ^2-Distribution.

Say, later, by z_α we will denote the α-quantile of the Standard Normal Distribution, $\mathcal{N}(0,1)$.

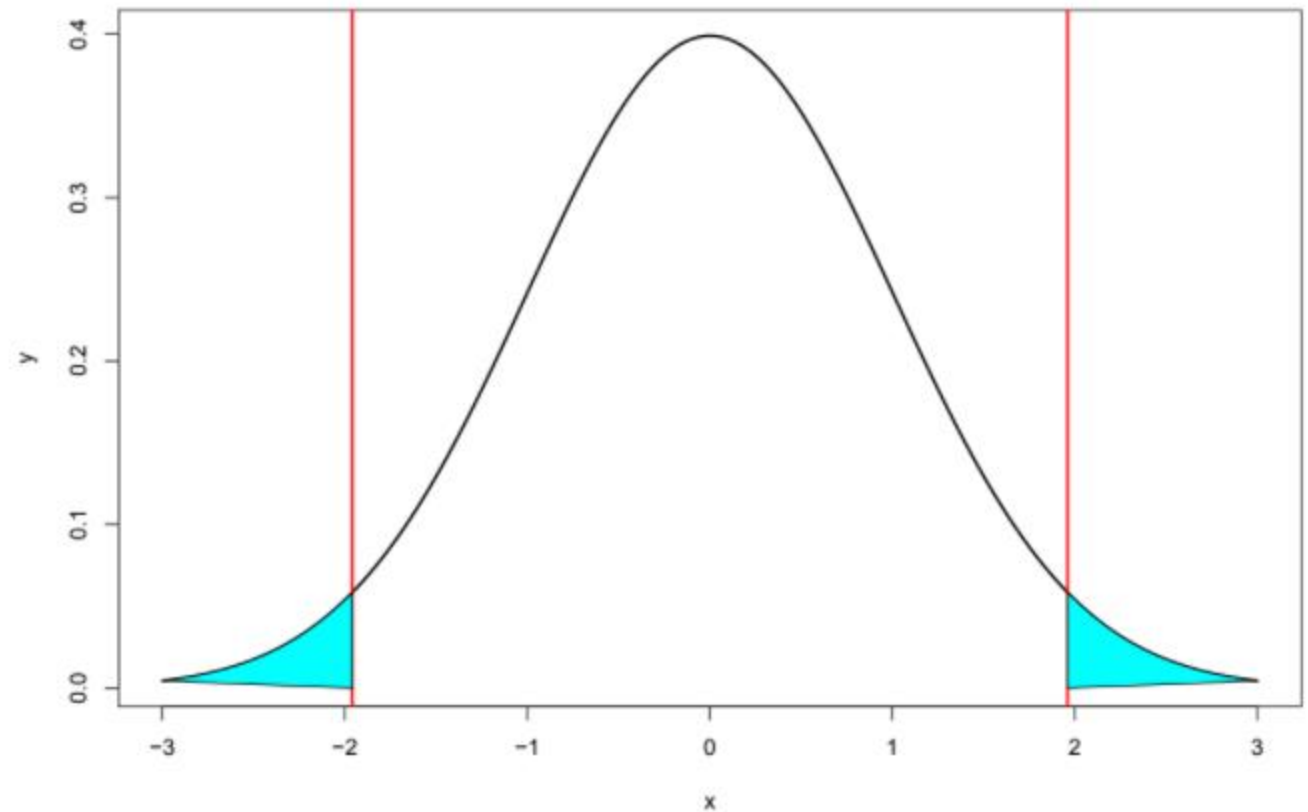
Say, we will take $\alpha \in (0,1)$ and find two points $a, b \in \mathbb{R}$ such that for $X \sim \mathcal{N}(0,1)$

$$\mathbb{P}(X \leq a) = \mathbb{P}(X \geq b) = \frac{\alpha}{2}.$$

The idea is to find a symmetric (in fact, the smallest length) interval $[a,b]$ such that for a Standard Normal r.v. X, the chances of $X \notin [a,b]$ are small, are exactly α.

Graphically

```r
alpha <- 0.05; z.alpha <- qnorm(alpha/2, mean = 0, sd = 1)
x <- seq(-3,3, by = 0.01)
y <- dnorm(x, mean = 0, sd = 1)
plot(x,y, type = "l", xlim = c(-3,3), lwd = 2)
abline(v = z.alpha, lwd = 2, col = "red")
abline(v = -z.alpha, lwd = 2, col = "red")
polygon(c(x[x<=z.alpha], z.alpha),c(y[x<=z.alpha],0),col="cyan")
polygon(c(x[x>=-z.alpha], -z.alpha),c(y[x>=-z.alpha],0),col="cyan")
```



Theoretical Quantiles, again

Then, it is easy to see, if $\alpha \in (0, 0.5)$ because of the symmetry, that $b = -a$, and

$$a = z_{\alpha/2}.$$

So

$$b = -z_{\alpha/2} = z_{1-\alpha/2}$$

Note: Please be careful when using Normal Tables. Usually, there is a picture above the table, on which you can find the explanation of the process. Just search "Normal tables" in Google Images.

Q-Q Plots, Data vs Data

Now, assume we have two Datasets, not necessarily of the same size:

$$x: \ x_1, x_2, ..., x_n \qquad \text{and} \qquad y: \ y_1, y_2, ..., y_m$$

Question: Are x and y coming from the same Distribution?

Q-Q Plot helps to answer to this question visually. To draw the Q-Q Plot for Datasets, we take some levels of quantiles, say, for some k,

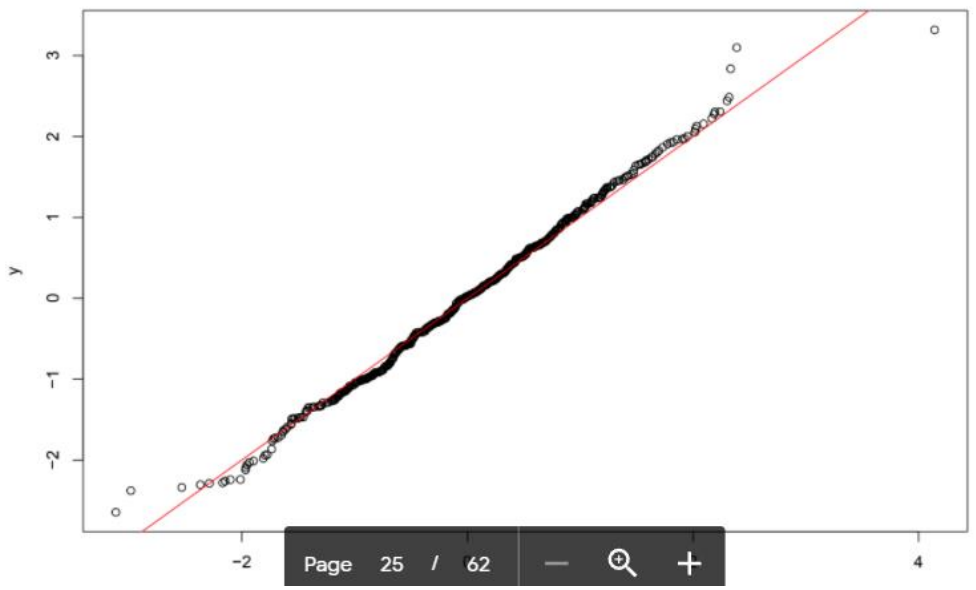
$$\alpha = \frac{1}{k}, \frac{2}{k}, ..., \frac{k-1}{k}$$

and then draw the points (q_α^x, q_α^y).

Idea: If x and y are coming from the same Distribution, then the Quantiles of x and y need to be approximately the same, $q_\alpha^x \approx q_\alpha^y$, so geometrically, the points (q_α^x, q_α^y) need to be close to the bisector line.

Example, Q-Q Plots, Data vs Data

```
x <- rnorm(1000)
y <- rnorm(500)
qqplot(x,y)
abline(0,1, col="red")
```



Example: Assume

$$x: \ -1, 2, 1, 2, 3, 2, 1 \qquad y: \ 0, 3, 4, 1, 1, 1, 1, 2$$

Draw the Q-Q Plot for x and y.

Q-Q Plots, Data vs Theoretical Distribution

Assume now we have a Dataset x and a Theoretical Distribution (say, given by its CDF F or PDF f). The Problem is to estimate visually if the Dataset comes from that Distribution.

Example: Say, is the following Dataset

```
##  [1] -0.378 -0.613 -0.241 -0.276 -0.072 -0.655  0.222 -0
## [11] -0.905 -0.671 -0.943  0.483  0.987 -0.191 -0.282  0
```

from a Normal Distribution?

To answer this question, we again take some levels of quantiles, say, for some k,

$$\alpha = \frac{1}{k}, \frac{2}{k}, ..., \frac{k-1}{k}$$

and then draw the points (q_α^F, q_α^x), where q_α^F is the α-quantile of the Theoretical Distribution, and q_α^x is the α-quantile of x.

Idea: If x is from the Distribution given by F, then we need to have $q_\alpha^F \approx q_\alpha^x$, so, graphically, the point will be close to the bisector.

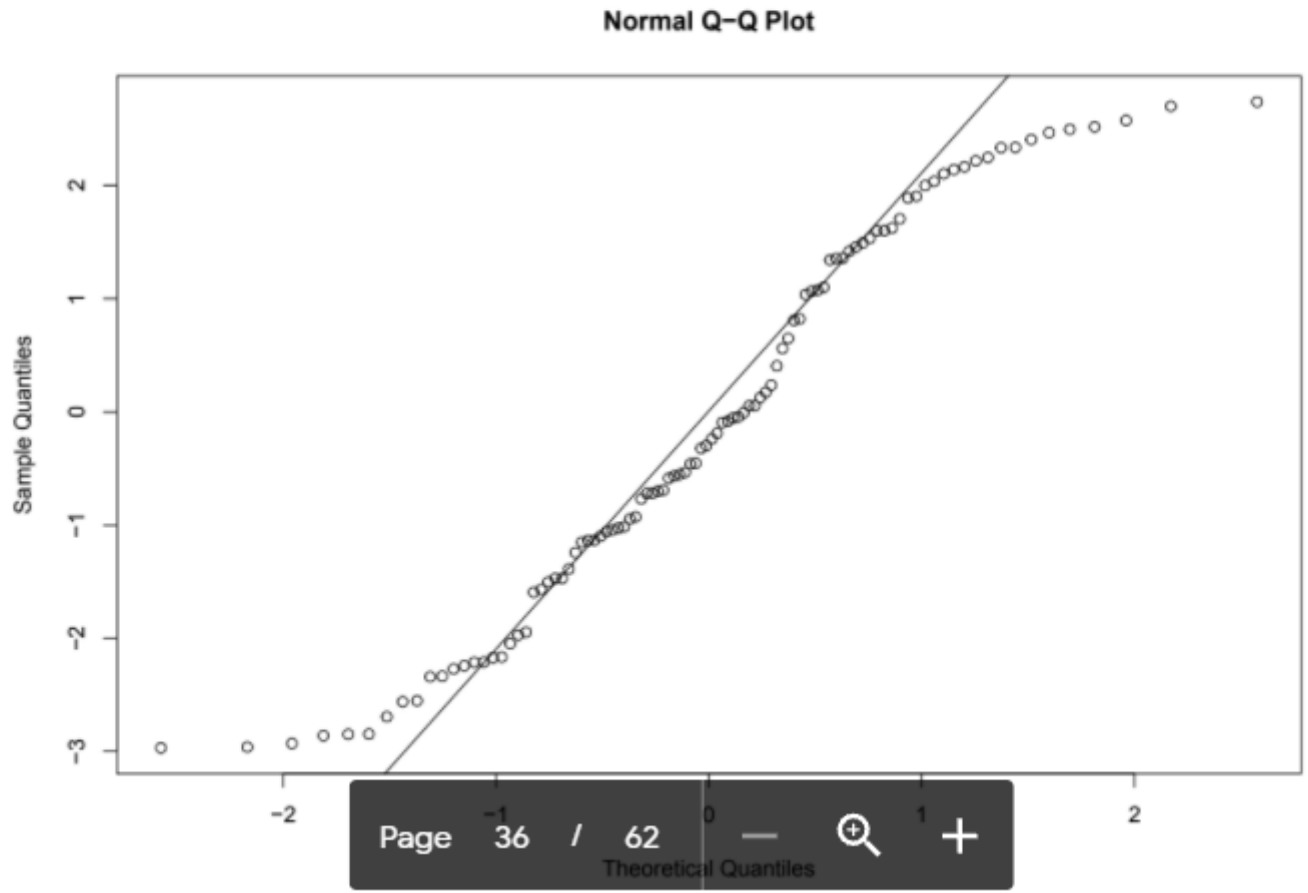
In **R**, we have a function `qqnorm` which plots the Q-Q Plot for the Dataset x vs the Normal Distribution. Unformtunately, we do not have this kind of function for other standard distributions, say, Uniform. But one can use the `qqplot(x,y)` command, by generating y from the given Distribution[1].

Another **R** command is `qqline` which adds a line passing (by default) through the first and third Quartiles,

$$(q^F_{0.25}, q^x_{0.25}) \qquad \textit{and} \qquad (q^F_{0.75}, q^x_{0.75}).$$

Here are some experiments with qqnorm

```
x <- runif(100,-3,3)
qqnorm(x)
qqline(x)
```



Important Note

It is important, that, using `qqnorm`, we can check if our Dataset comes from a Normal Distribution, *with some mean and variance*. I mean, the above idea was, say, to check if given Dataset x comes from given Distribution, say, $\mathcal{N}(2, 3^2)$.

But, for the Normal Distribution, we can use the fact that all Normal Distributions can be obtained from the Standard Normal, by scaling and shifting. This means that the Quantiles of any Normal Distribution can be obtained by a linear transform from the Standard Normal Quantiles[2].

So if, say, x is a sample from $\mathcal{N}(2, 3^2)$, then

- when doing a Q-Q Plot of x vs $\mathcal{N}(2, 3^2)$, the Quantiles will be on the bisector;
- when doing a Q-Q Plot of x vs $\mathcal{N}(0, 1)$, the Quantiles will be on some line (can you find the line equation?);

Q-Q Plots, Theoretical vs Theoretical Distribution

Assume now we have two Theoretical Distributions (say, given by their CDFs F and G). The Problem is to estimate visually which Distribution has fatter tails.

To answer this question, we again take some levels of quantiles, say, for some k,

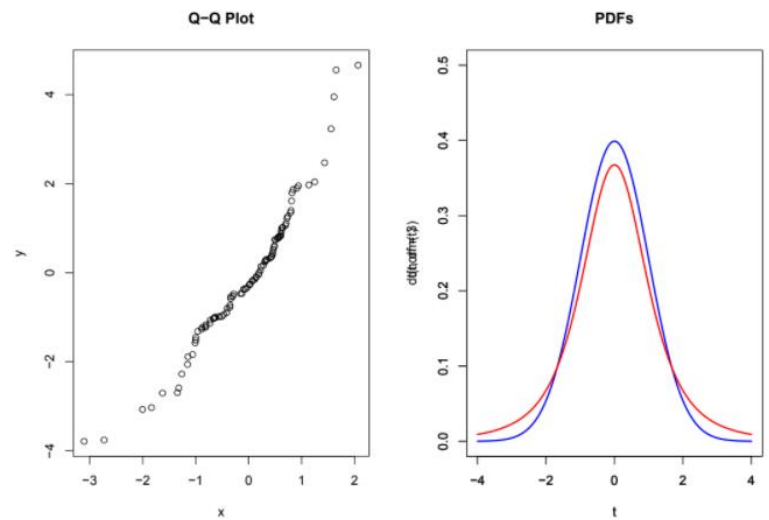
$$\alpha = \frac{1}{k}, \frac{2}{k}, ..., \frac{k-1}{k}$$

and then draw the points (q_α^F, q_α^G), where q_α^F is the α-quantile of the Theoretical Distribution with the CDF F, and q_α^G is the α-quantile of the Theoretical Distribution with the CDF G.

Idea: If G has fatter tails on both sides than F, then we will have graphically some cubic-function graph shape Quantiles.

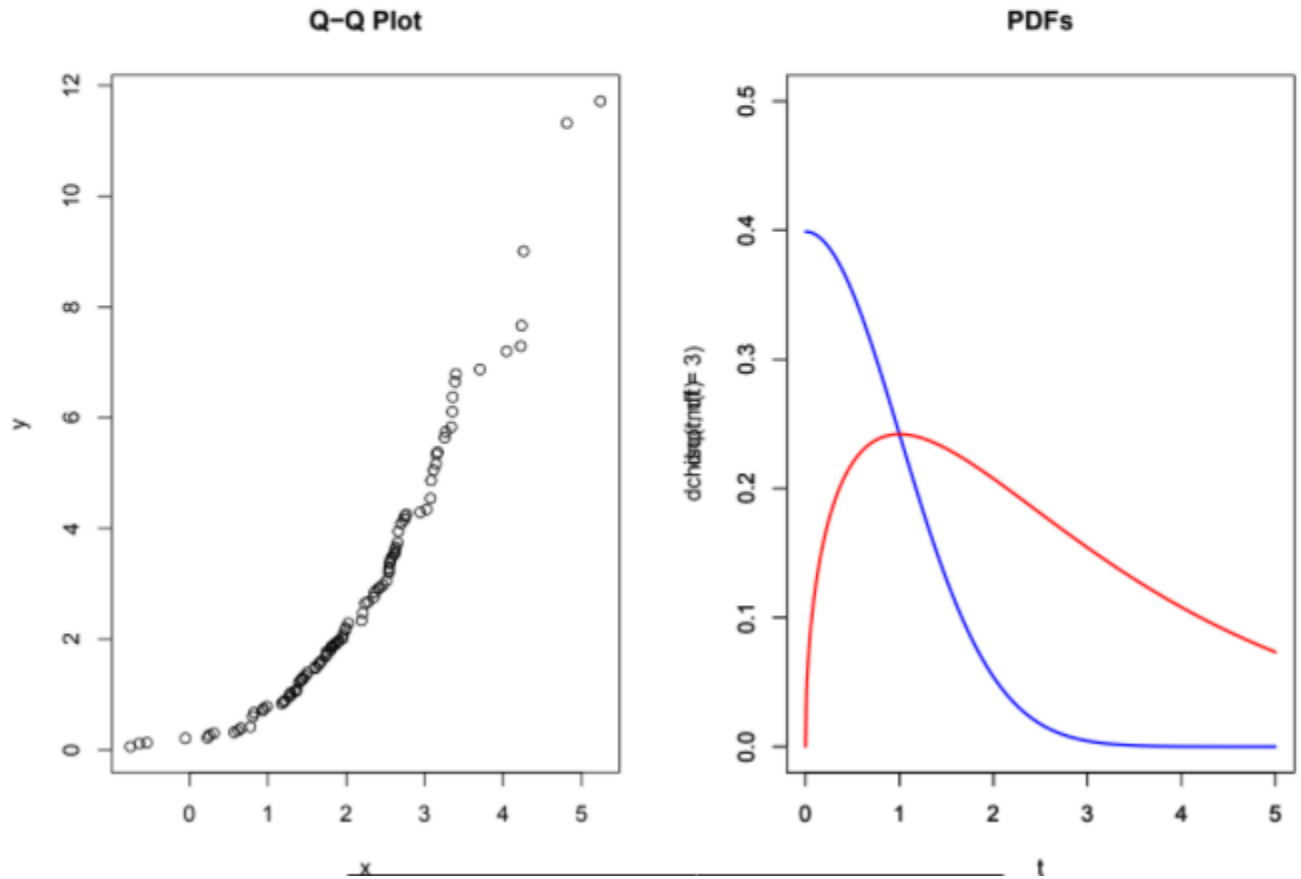
Some Experiments

```
par(mfrow = c(1,2))
x <- rnorm(100, mean=0, sd=1); y <- rt(100, df = 3)
qqplot(x,y, main = "Q-Q Plot")
t <- seq(-4,4,0.01)
plot(t, dnorm(t), type = "l", xlim = c(-4,4), ylim = c(0, 0.5), col ="blue", lwd = 2, main = "PDFs")
par(new = TRUE)
plot(t, dt(t, df = 3), type = "l", xlim = c(-4,4), ylim = c(0, 0.5), col ="red", lwd = 2)
```



Some Experiments

```
par(mfrow = c(1,2))
x <- rnorm(100, mean=2, sd=1); y <- rchisq(200, df = 3)
qqplot(x,y, main = "Q-Q Plot")
t <- seq(0,5,0.01)
plot(t, dnorm(t), type = "l", xlim = c(0,5), ylim = c(0, 0.5), col ="blue", lwd = 2,  main = "PDFs")
par(new = TRUE)
plot(t, dchisq(t, df = 3), type = "l", xlim = c(0,5), ylim = c(0, 0.5), col ="red", lwd = 2)
```



Addition, Q-Q Plot with a Confidence Band

```
require(qqplotr)
x <- data.frame(variable = rnorm(200))
ggplot(data = x, mapping = aes(sample = variable)) + stat_qq_band() +
  stat_qq_line() + stat_qq_point() + labs(x = "Theoretical Quantiles", y = "Sample Quantiles")
```