STAT 210
Applied Statistics and Data Analysis
Quantile plots

Joaquín Ortega
KAUST

joaquin.ortegasanchez@kaust.edu.sa

Quantiles

Distribution functions and location and scale parameters

Consider a random variable X that has distribution function F_X:

$$F_X(x) = P(X \leq x),$$

and consider a linear transformation of X:

$$Y = aX + b$$

where $a \neq 0$ and $b \in \mathbb{R}$. The distribution function of Y is easy to obtain in terms of F_X. If $a > 0$,

$$F_Y(y) = P(Y \leq y) = P(aX+b \leq y) = P(X \leq \frac{y-b}{a}) = F_X(\frac{y-b}{a})$$

and a similar relation is true for $a < 0$.

We say that b is a **location** parameter while a is a **scale** parameter.

Distribution functions and location and scale parameters

The distribution functions associated to these transformations $aX + b$ are known as a **location and scale family**.

Example

Let $X \sim N(0, 1)$ be a standard normal random variable. Then, it is possible to show that $Y = aX + b$ has also a normal distribution with parameters b and a^2:

$$Y \sim N(b, a^2).$$

We say that the normal distribution is a location and scale family with location parameter the mean and scale parameter the standard deviation.

Quantiles

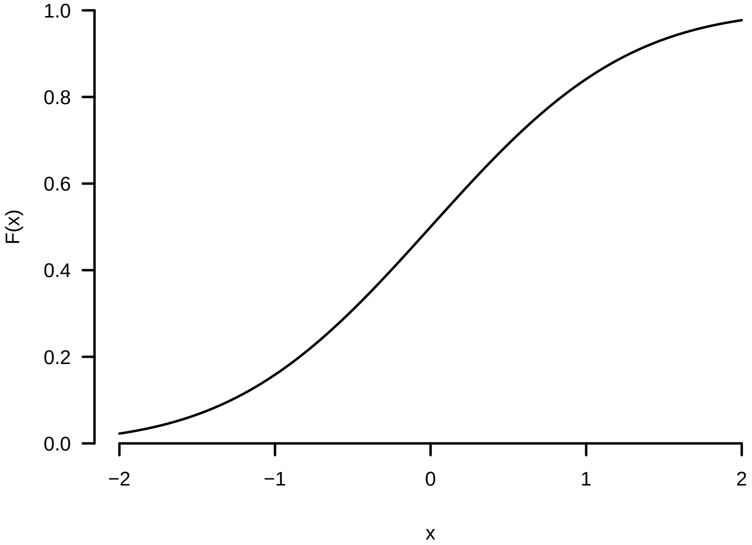
Quantiles divide a probability distribution into sections having equal probabilities.

For example,

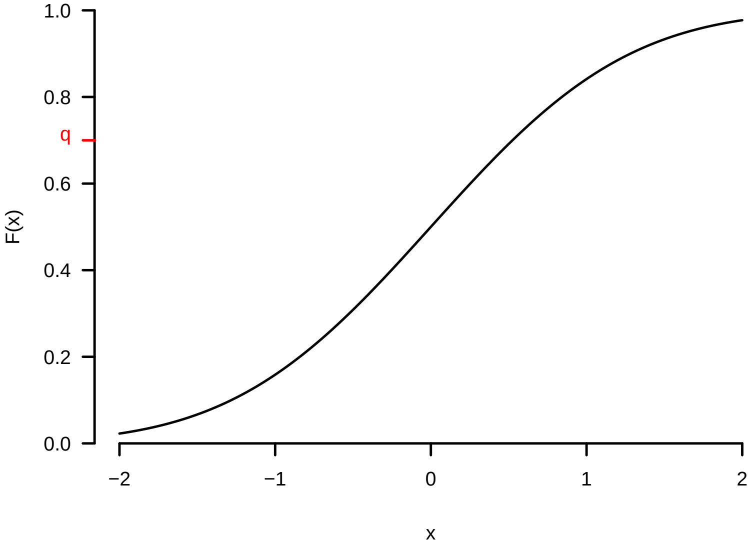
- the median divides the distribution in two
- quartiles divide the distribution in four
- deciles divide the distribution in 10
- percentiles divide the distribution in 100

The generic name for all these quantities is quantile.

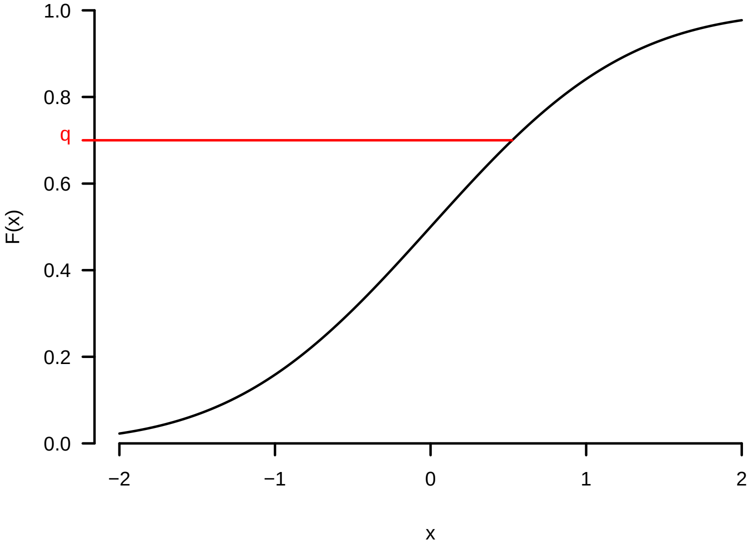
Quantiles



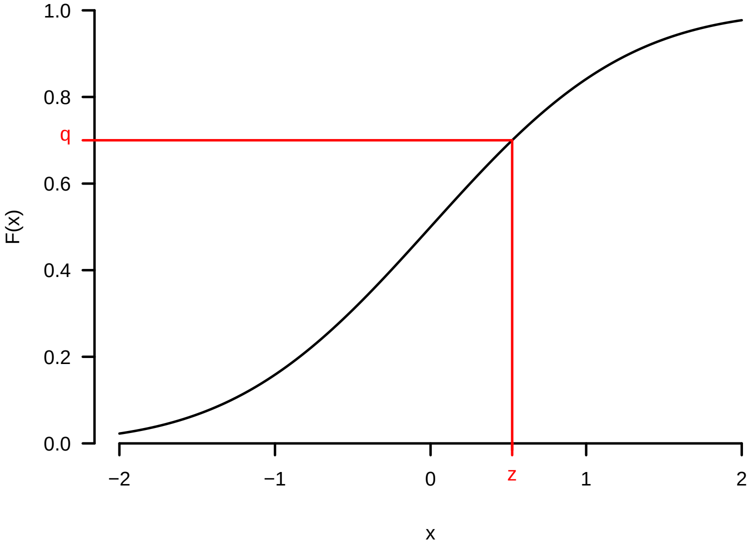
Quantiles



Quantiles



Quantiles



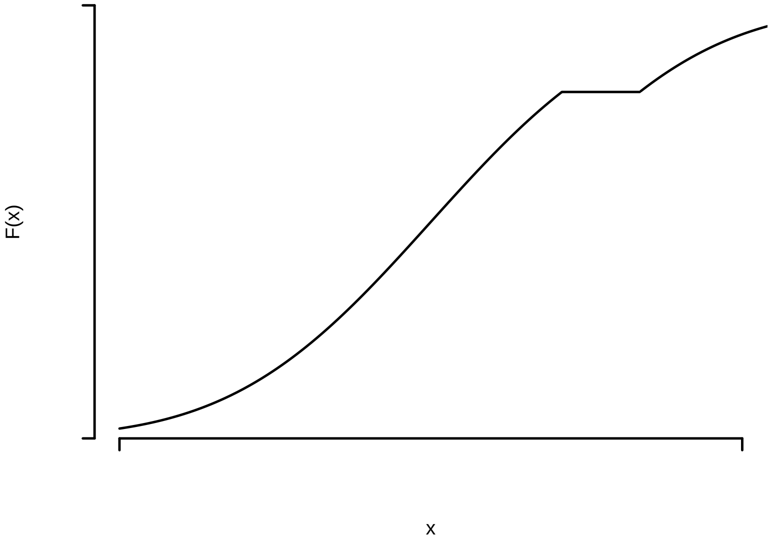
Quantiles

Definition

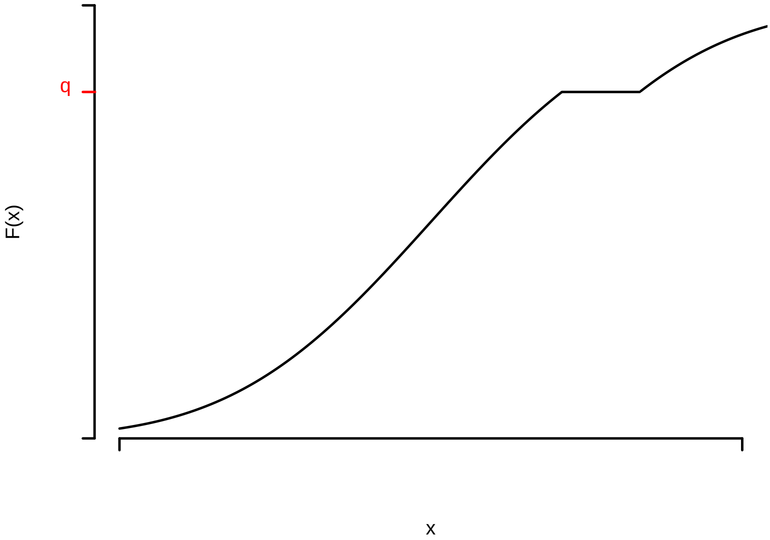
Given a distribution function $F(x)$ that is continuous and strictly increasing, for $0 < q < 1$, the q quantile is the value z such that a fraction q of the distribution is to the left of z:

$$P(X \leq z) = q$$

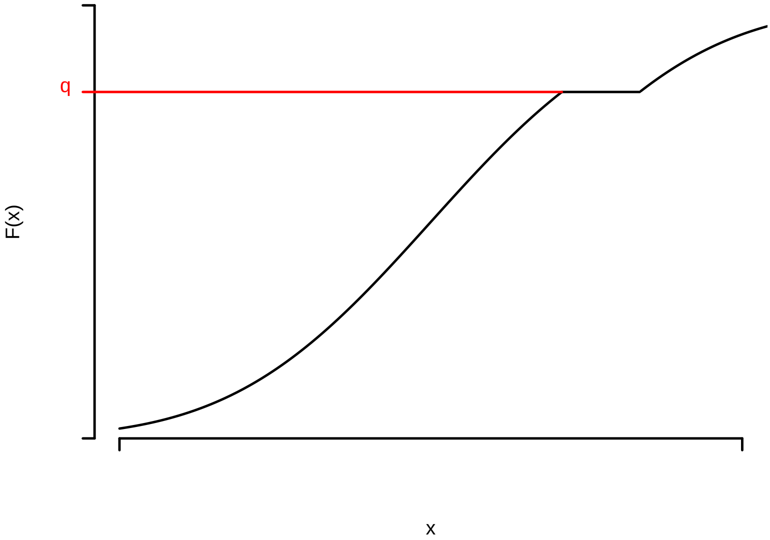
Quantiles



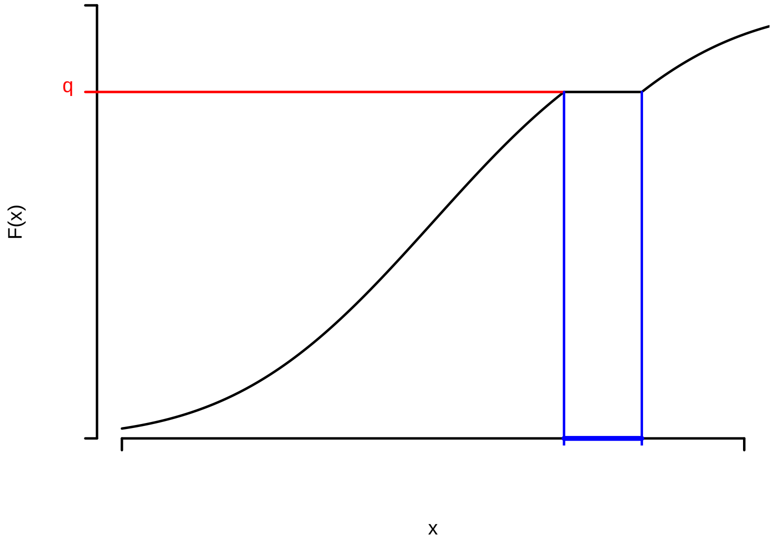
Quantiles



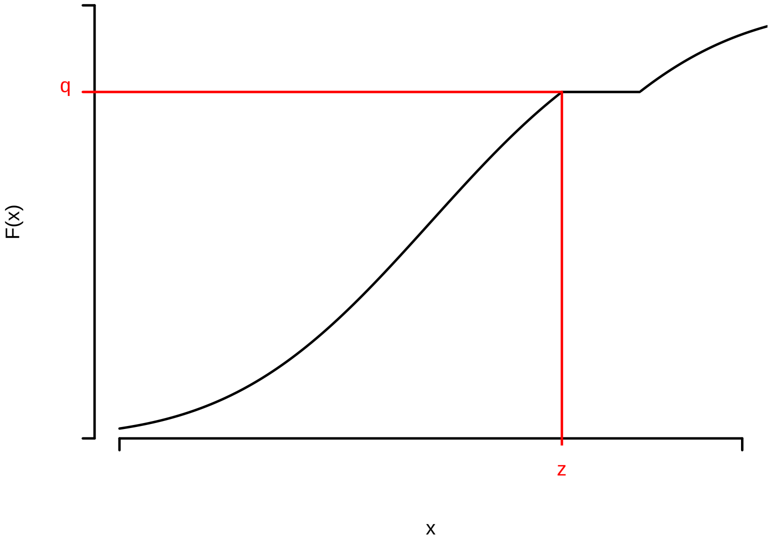
Quantiles



Quantiles



Quantiles



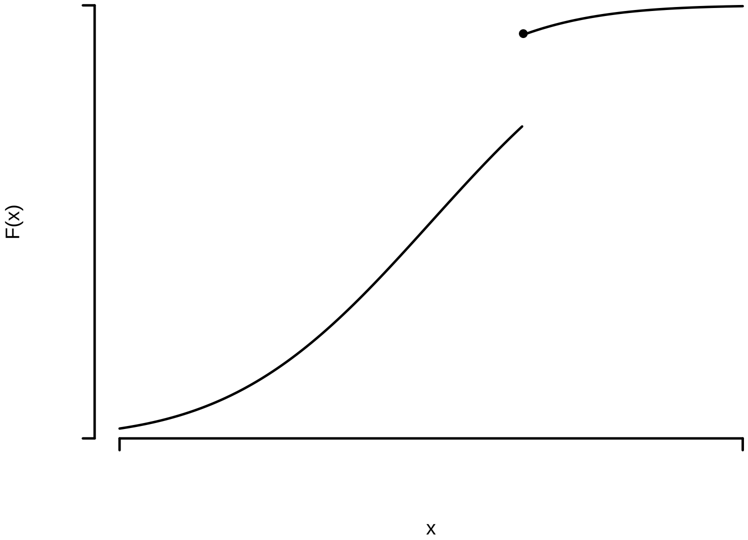
Quantiles

If the quantile is not unique, we take the smallest value for which $F(z) = q$:

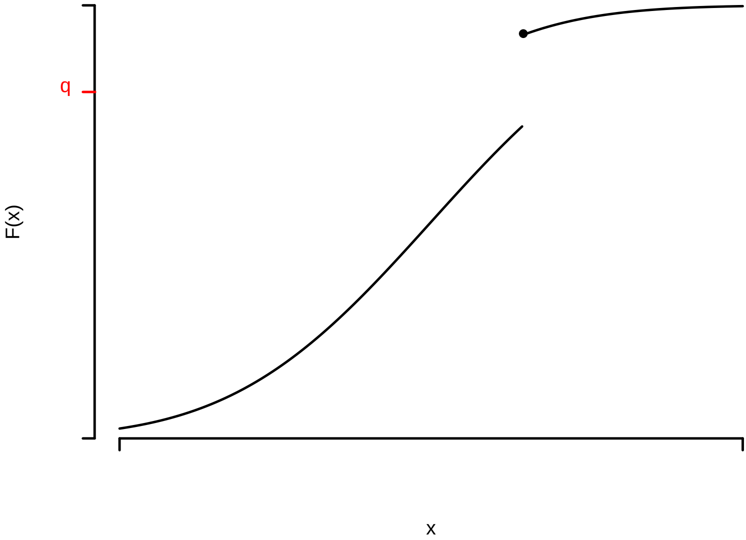
$$z = \inf\{x : F(x) \geq q\}$$

Note that if we are in the earlier case, i.e., the function is continuous and strictly increasing, this definition gives the same value for z as the previous one.

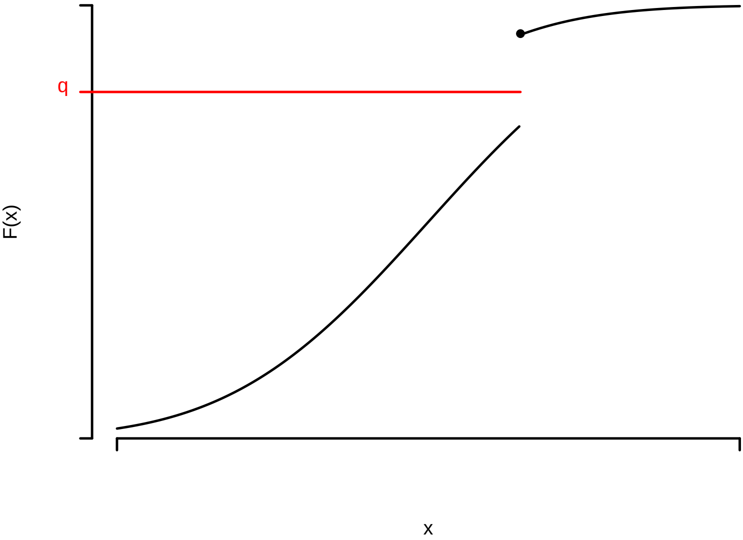
Quantiles



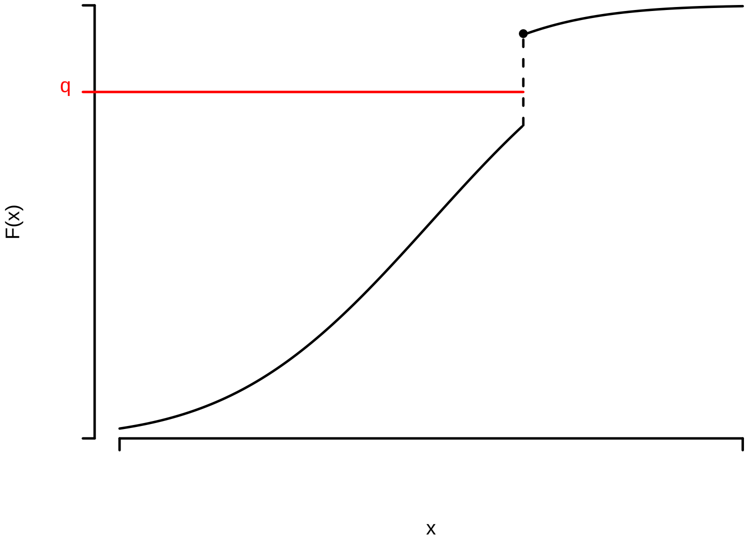
Quantiles



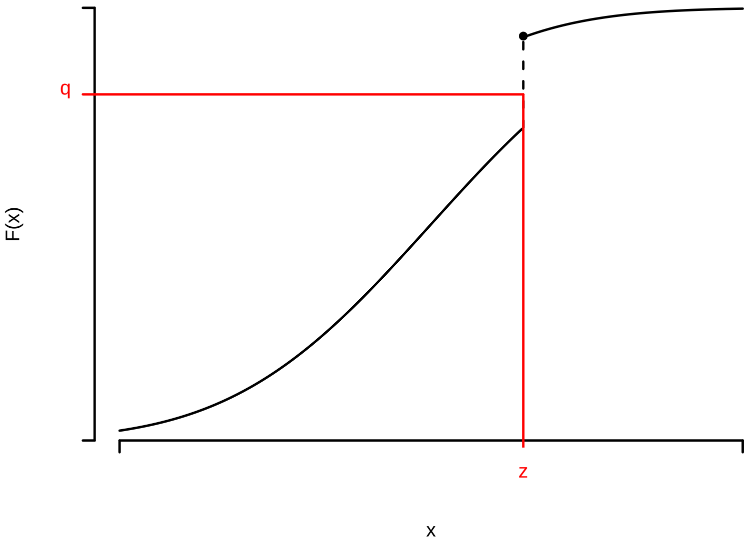
Quantiles



Quantiles



Quantiles



Quantiles

If F is discontinuous then it may happen that there is no value of z that satisfies $F(z) = q$. In this case

$$z = \inf\{x : F(x) \geq q\}$$

Quantile function

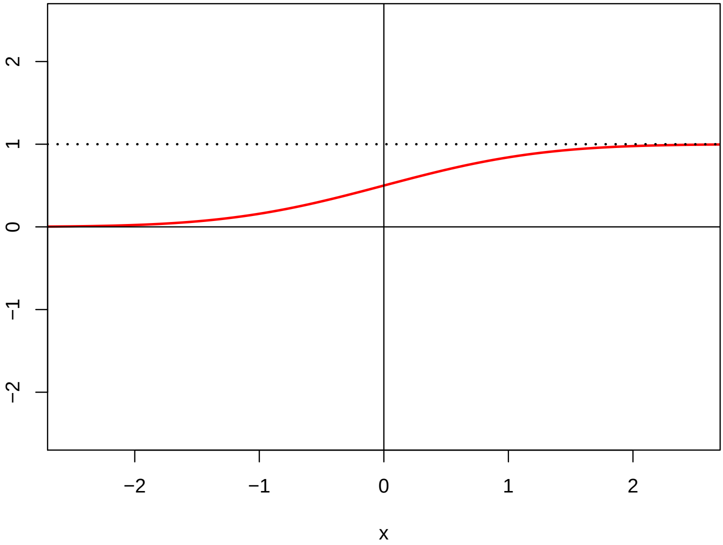
Definition

The **quantile function** Q is the function that, given q, $0 < q < 1$, produces the value $z = \inf\{x : F(x) \geq q\}$.

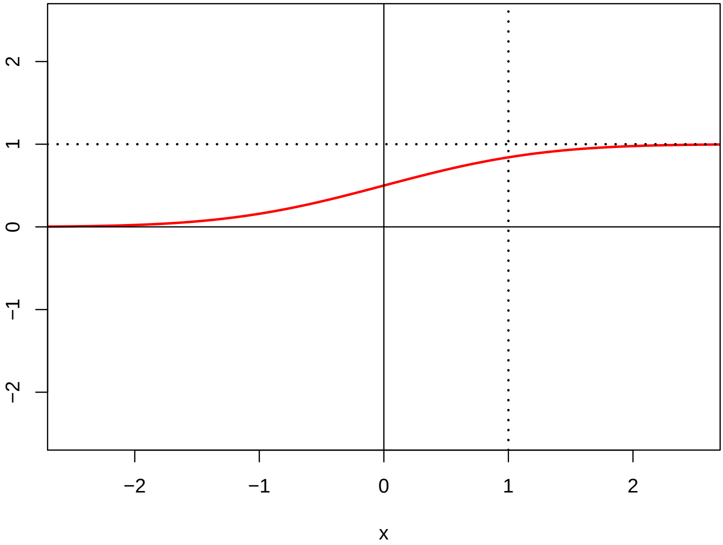
If F is continuous and strictly increasing, then Q is the inverse function of F.

The empirical quantiles are the quantiles of the empirical distribution.

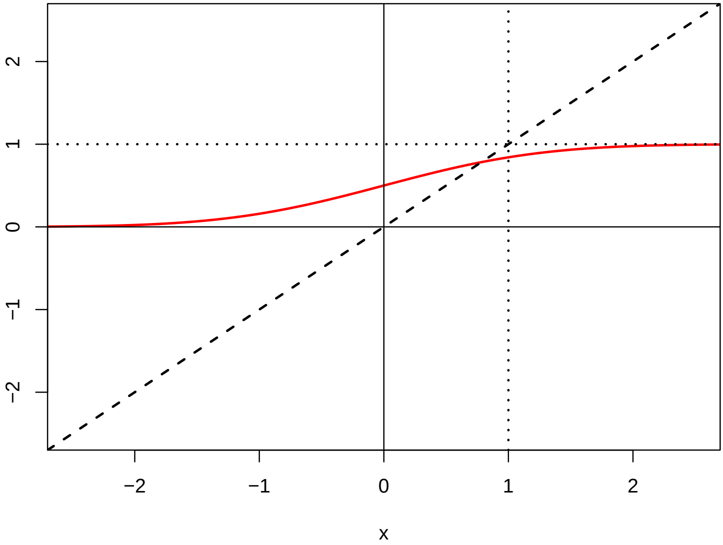
Quantile function



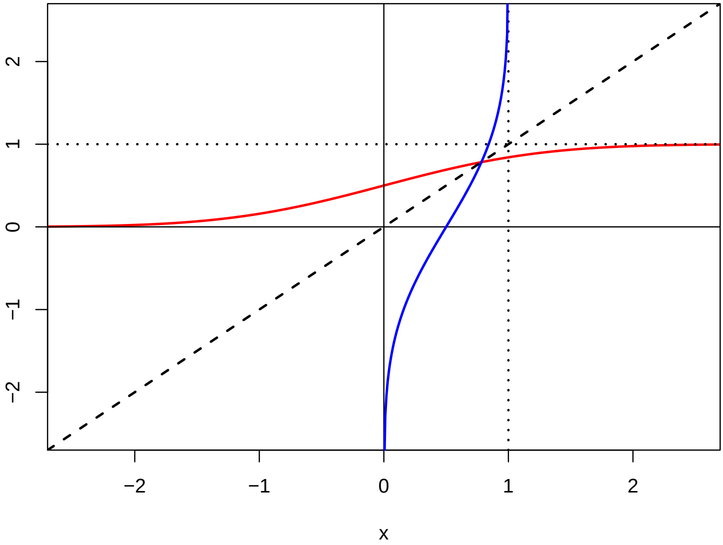
Quantile function



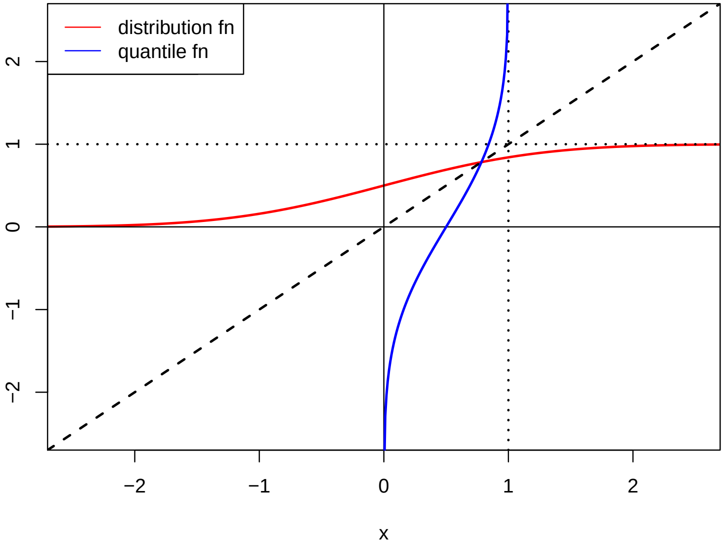
Quantile function



Quantile function



Quantile function



Quantile plots

Quantile plots

The quantile plots, proposed by Wilk and Gnanadesikan in 1968, are a visual tool to compare the distribution of two sets of data or to compare a set of data with a reference distribution.

If the two distributions belong to the same location and scale family, the graph will be approximately a straight line.

Suppose we have two samples of the same size, $x_i, y_i, 1 \leq i \leq n$. The order statistics of the samples are the ordered values: For the x sample, assuming there are no ties, this would be

$$x_{(1)} < x_{(2)} < \cdots < x_{(n-1)} < x_{(n)}$$

Quantile plots

The quantile plot for the two samples is the plot of ordered values of x versus the ordered values of y, if both samples have the same size. If the two samples are not the same size, linear interpolation is used.

In `R` the function for making quantile plots to compare two samples is `qqplot`

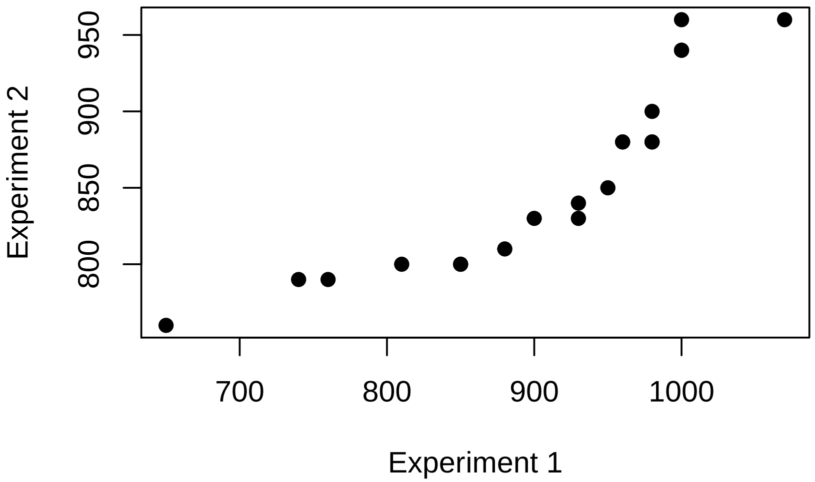
We will use the data for the Michelson-Morley experiment in the `morley` dataset to give some examples

```
data(morley)
str(morley, vec.length = 1)
```

```
## 'data.frame':    100 obs. of  3 variables:
##  $ Expt : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Run  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Speed: int  850 740 900 1070 930 850 950 980 980 880
```

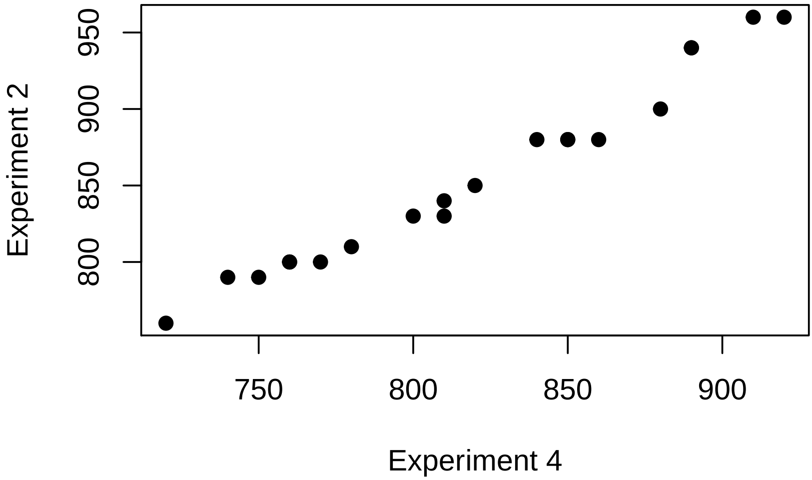
Quantile plots

```
qqplot(morley$Speed[morley$Expt==1],
       morley$Speed[morley$Expt==2],
       xlab='Experiment 1', ylab = 'Experiment 2',pch=19)
```



Quantile plots

```
qqplot(morley$Speed[morley$Expt==4],
       morley$Speed[morley$Expt==2],
       xlab='Experiment 4', ylab = 'Experiment 2',pch=19)
```



Quantile plots

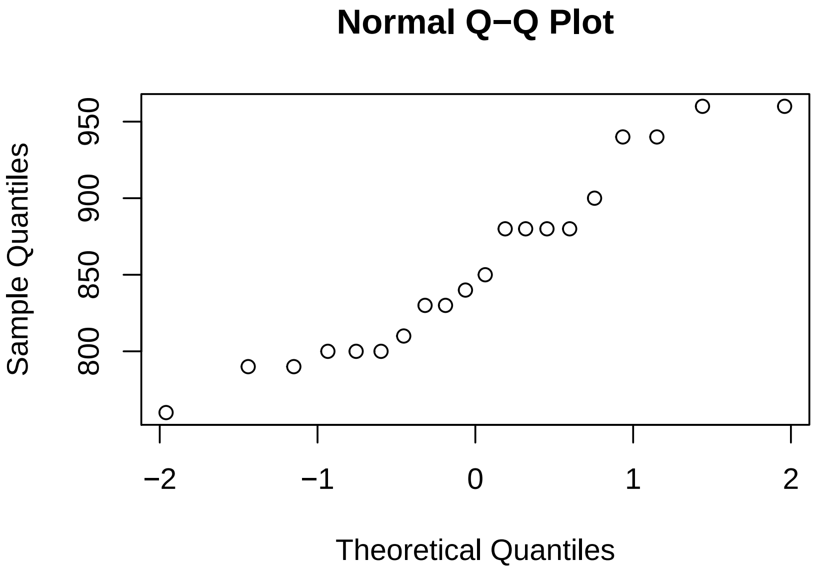
When we want to compare with a reference distribution, the empirical quantiles are plotted against the quantiles calculated from the reference distribution.

In particular, the function `qqnorm` in `R` draws a quantile plot to compare a given data set with the normal distribution.

If the fit is good, the points should appear to be on a straight line.

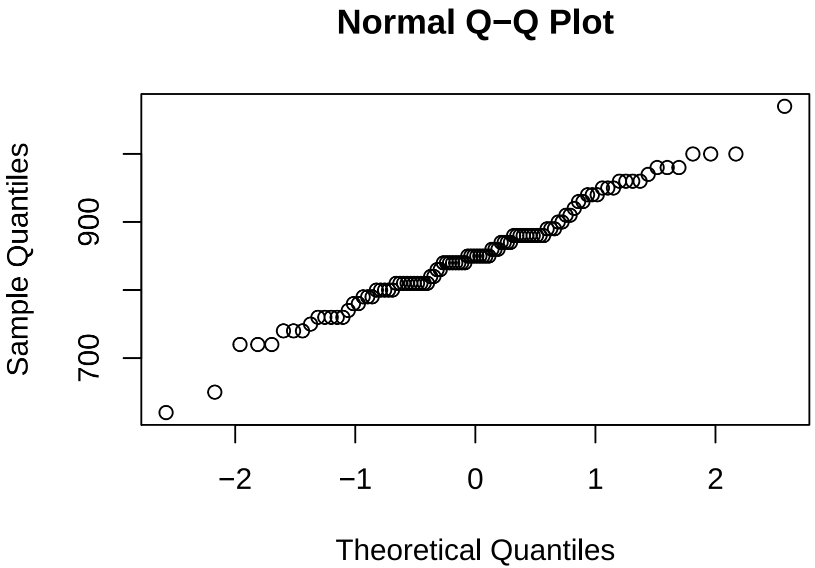
Quantile plots

```
qqnorm(morley$Speed[morley$Expt==2])
```



Quantile plots

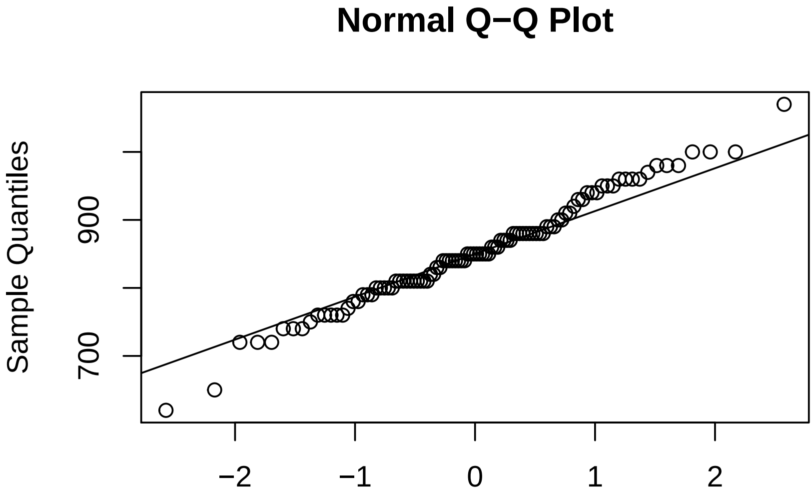
```r
qqnorm(morley$Speed)
```



Quantile plots

As a visual aide, the function qqline() draws a straight line the passes through the two quartiles:

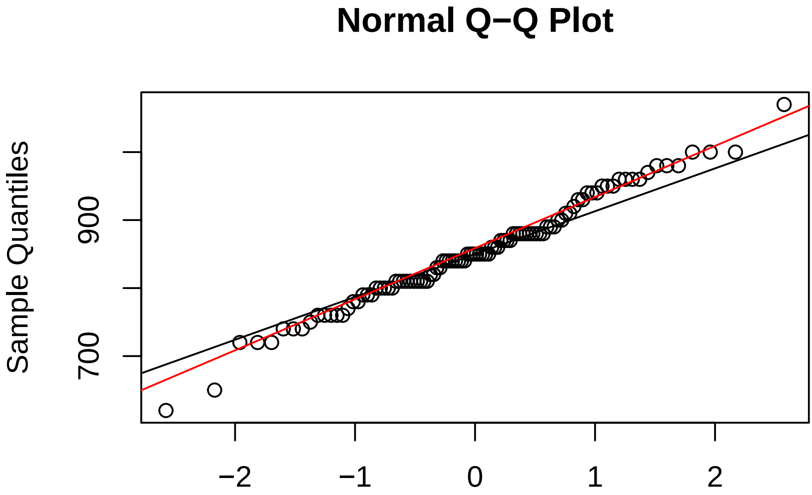
```
qqnorm(morley$Speed)
qqline(morley$Speed)
```



Quantile plots

Beware that this is not the 'best fitting' line:

```
qqnorm(morley$Speed); qqline(morley$Speed)
qqline(morley$Speed,probs = c(0.18,0.8), col='red')
```



Simulations

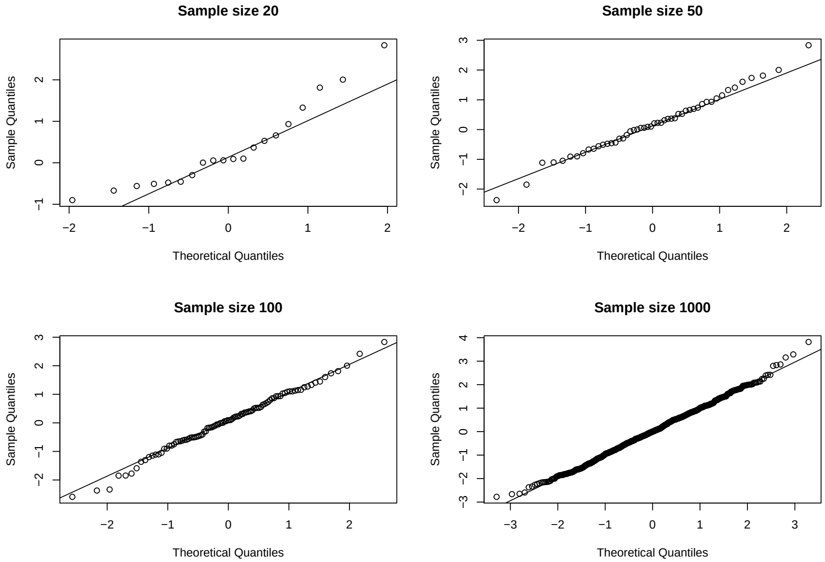
Some Simulations

We simulate 1000 points from the standard normal distribution and extract from it samples of sizes 20, 50 and 100, to observe the effect of size on the quantile plots.

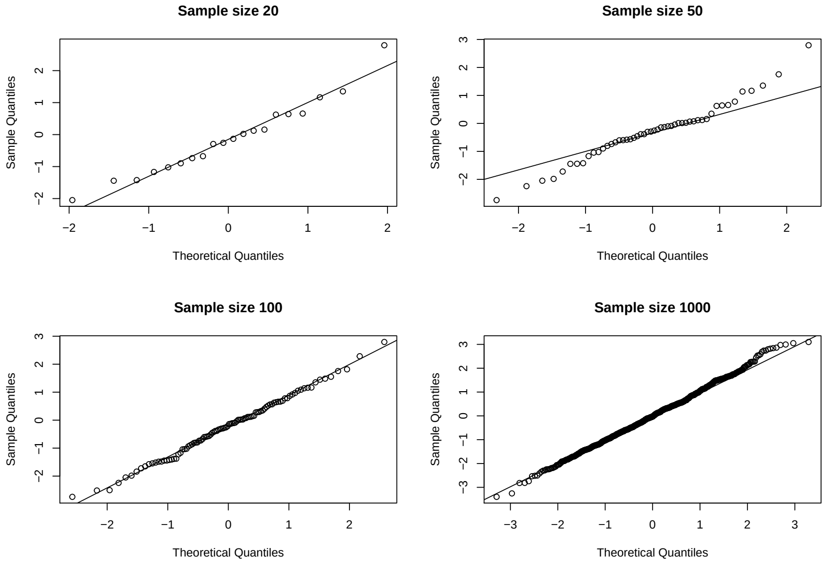
Remember that since we are simulating from a normal distribution, we would expect to observe straight lines.

```
set.seed(1290)
norm1000 <- rnorm(1000); norm100 <- norm1000[1:100]
norm50 <- norm1000[1:50]; norm20 <- norm1000[1:20]
```

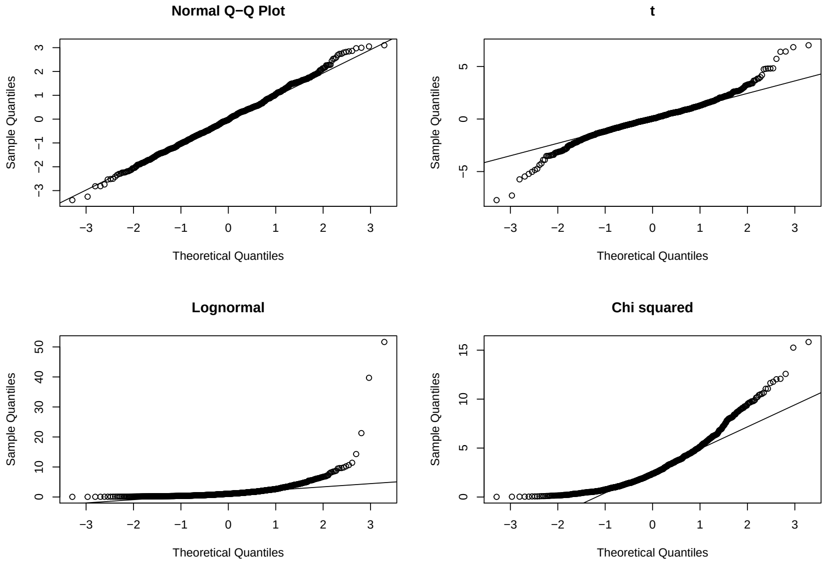
Some Simulations



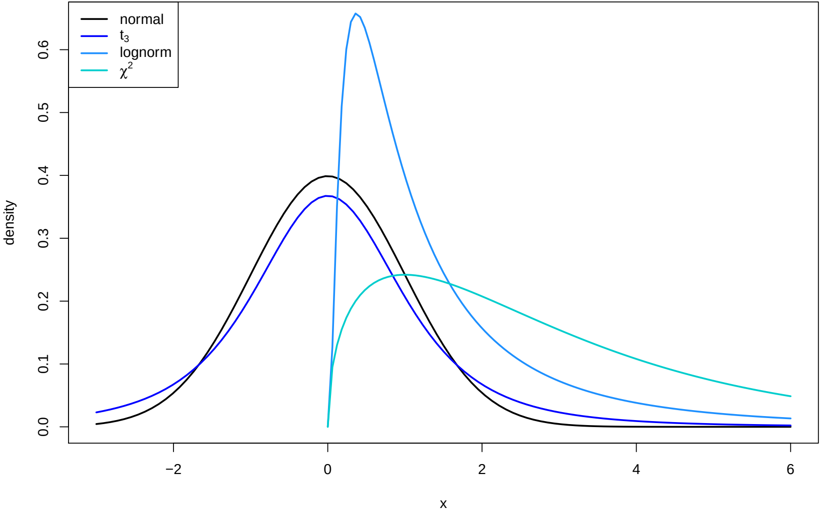
Some Simulations



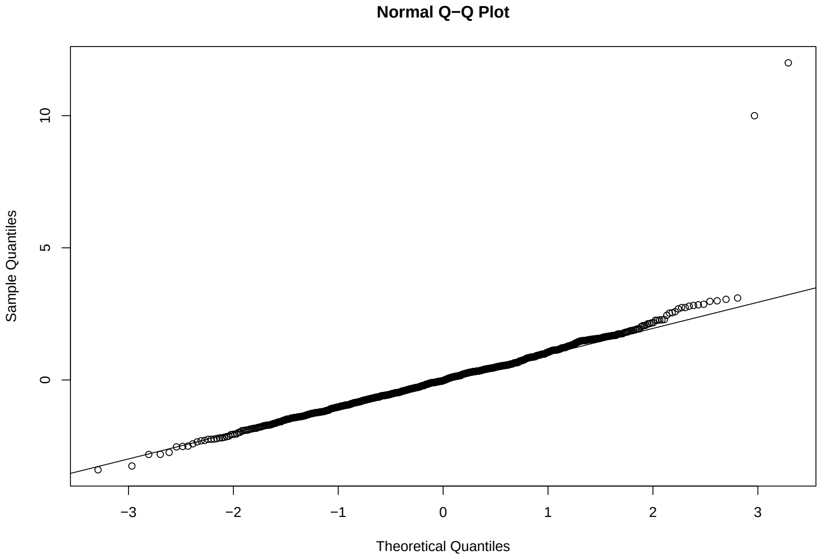
Some Simulations



Some Simulations



Outliers



Normal Q−Q Plot