Mathematical Statistics

The mathematical statistics is different from the applied statistics. It builds probability models.

In statistics the aim is to associate the examined sample with a model. This connection is never clear and conclusive because an infinite number of data is not available: the connection between the data and a model is an asymptotic process.

It is easier to create models for centrality and variability indexes than for extreme percentiles, because the second ones are more sensitive to little variations.

The nature of the concept of probability, in an empirical sense, can be stated like follows:
In a sequence of trials, done in the same conditions, the frequency of an event comes nearer to the probability of the same event. The approximations tends to ameliorate when the number of trials increases.

With the same number of possible cases, the probability of an event increases when the number of favourable cases increases. It can also happen when, with the same number of favourable cases, the number of possible cases decreases.

The first (in the historical sense) **definition of classic probability:**
The probability of a random event is the ratio between the number of favourable cases and the number of possible cases, as long as they have the same probability.

The concept of **classic probability**, founded on a **mathematical probability** or **prior**, has been the first one to be defined.

It is the probability to obtain heads or tails with a coin, to have a number from 1 to 6 with a dice or in more throwings, to foresee the arrival orders in a competition with various contestants who respect the previous four conditions.

No experimental datum is required. The results are prior known, and it is not necessary to wait for gatherings or observations. **It is only necessary the logic reasoning in order to accurately calculate the probability**.

If the coin, the dice or the match are not loaded, the experimental tests will leave the expected data only for negligible quantities, determined by random events or observational errors.

The estimation of a **prior probability** has serious restrictions in the experimental reasearch. In order to calculate the probability of an event, it is necessary to previously know the various probabilities of all the possible events.

For this reason, this approach cannot be always used.

As estimation of the **probability of an experimental event** it can be used its **frequency**.

The probability of a random event is the limit to which it tends when the number of observations increases, in a series of experiences recurring in the same conditions.

$$P = \lim_{n \to \infty} f$$

It also exists another kind of probability called **posterior probability**. This is not a topic of this course.

If F is the relative frequence of an event in a population, it is generally possible to observe that, when the number of observations (n) increases, the frequency (f) of the sample tends to become similar to the real one or to that of the population (F).

This statement cannot be proved either with mathematical instruments, because it refers to osbervational data, or in an empirical way, because, in the reality, it is not possible to repeat an experiment infinite times.

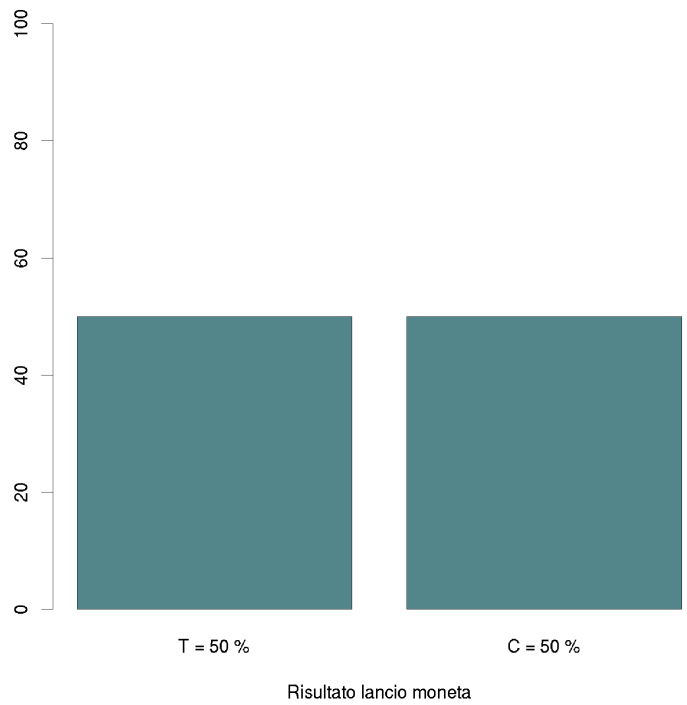
Anyway, it is a **statistical regularity**, called **law of large numbers**, that is the experimental basis both of every statistics theory and of mathematical reasoning.

In these cases, it is a matter of **frequentist probability**, of a **posterior probability**, of **law of large numbers**, or of **statistical probability**.

It is possible to answer to a lot of empirical questions in only one way:

To conceive a series of observations or experimentations, in unrelieved or statistically checked conditions, in order to point out the frequency.

Irrespective of the probability is empirical or mathematical, there are some **probability distributions** able to explain the phenomenon.



The chart represents a discrete variables distribution. It is a mathematical representation that deals with data.

Numerical outcomes for a random phenomenon can be described by mathematical instruments called **random variables** (or **unpredictable**).

Random variables are "mathematical models" which gives probability values to possible numerical values. Their corresponding **probability distributions can be of two kinds: discrete or continuous.**

In the **discrete random variables**, the argument values are the natural numbers: 0, 1, 2, . . . , n. They are necessary to calculate the probability of events that have a discrete number (finite or infinite) of recurrences.

A **random variable** is **continuous** when its distribution is continuous. With this continuous variable, it is estimated the probability to extract not a single value but equal or greater values (or equal or less values). The probabilities are calculated only for intervals values in the case of continuous random varibales. For specific values, they are always zero.

Among the discrete distribution, two are really important in statistics for their quality:

- ⋆ Binomial
- ⋆ Poisson

Both the distributions explain phenomenon related to the behavior of binary phenomenon: error - success, goes - does not goes, defective - not defective.

In particular, they define the number of successes that can be verified in N trials, given a certain probability of success of each trial.

The two functions are different because:

- ⋆ the Binomial deals with medium-high probability events;
- ⋆ the Poisson deals with very low probability events.

The **binomial is a discrete and finite theoretical distribution**, for events classified with a **binary variable** and it is defined by the following density function:

$$p_n(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

where:

$x =$ number of successes

$p =$ probability of success of a single trial

$n =$ number of trials

For example, if the $\boldsymbol{p}$ parameter is the probability of the event "throwing a coin it gives heads", and then $p = 0.5$, it is possible to calculate the probability that on $\boldsymbol{n}$=10 throwings, the number of successes $\boldsymbol{x}$ is equal to 8:

$$P_{10}(8) = \frac{10!}{8! \cdot 2!} \cdot 0.5^8 \cdot 0.5^2 = 0.04 \rightarrow 4\%$$

The probability distribution of the binomial depends on 2 parameters: $\boldsymbol{p}$ and $\boldsymbol{n}$.

If p and $\boldsymbol{q = 1 - p}$ are equal to 0.5 the distribution is always symmetric, independently of n. If p is highly greater or less than q, the distribution is asymmetric. This asymmetry tends to decrease when n increases.

The binomial distribution is usually used for p values that go from 0.1 to 0.9. For p values excluded from this interval, the Poisson distribution is used when n is not high.

When n is so high that even the result of $\boldsymbol{n \cdot p}$ is high, the binomial one is used, for each value of p.

When a sample has big dimensions, the probabilities estimation can be obtained by the normal distribution.

When the number of data n is really high and the probability p is really small, the binomial distribution has some practice disadvantages. They were really important before the introduction of the automatic calculation: to raise really small frequencies to high powers and the calculation of factors for high numbers, make the manual calculation almost impossible.

For n that tends to infinite and p that tends to 0, in a way that $n \cdot p$ is constant, Poisson demonstrated that:

$$P_i = \frac{\mu^i}{i!} e^{-\mu}$$

if: $n \to \infty$ $p \to 0$ $n \cdot p = \mu$

The Poisson is a discrete theoretical distribution and it is totally defined by only one parameter, the mean μ.

Even in the Poisson distribution, the μ expected mean is given by $n \cdot p$.

So that the frequencies or probabilities calculated with the Poisson law are exact, μ must be a constant parameter for the whole distribution.

Also in this distribution $\sigma^2 = n \cdot p \cdot q$.

It is possible to easily verify that the variance is equal to the mean: if we use the just illustrated three conditions, the Poisson distribution variance will be given by:

$$\sigma^2 = \lim_{\substack{n\to\infty \\ p\to 0}} (n \cdot p \cdot q) = \lim_{\substack{n\to\infty \\ p\to 0}} (n \cdot p) \cdot q = \lim_{\substack{n\to\infty \\ p\to 0}} k \cdot (1-p) = k = \mu$$

Or rather **if p tends to 0, q tends to 1; and then $n \cdot p = n \cdot p \cdot q$.**

The Poisson distribution law is also called **law of rare events**, because the probability that the event happens is really low.

It is also called **law of small numbers**, because the absolute frequency of these events is expressed by a small number, even in an high number of trials.

The Poisson distribution has a very asymmetric shape and the most frequent and probable class is zero, when μ is less than 1. It is still asymmetric for values of μ less than 3.

A mean equal to 6-7 establishes a symmetric distribution of probabilities and it is well approximated by the normal (or Gaussian) distribution.

The Poisson distribution can be used instead of the binomial one, for p less than 0.05 and n greater than 100.

The Poisson distribution is used for events that happen both in the space and in the time. For example, the number of not wanted particles in a little space or the number of events that can happen in a little time.

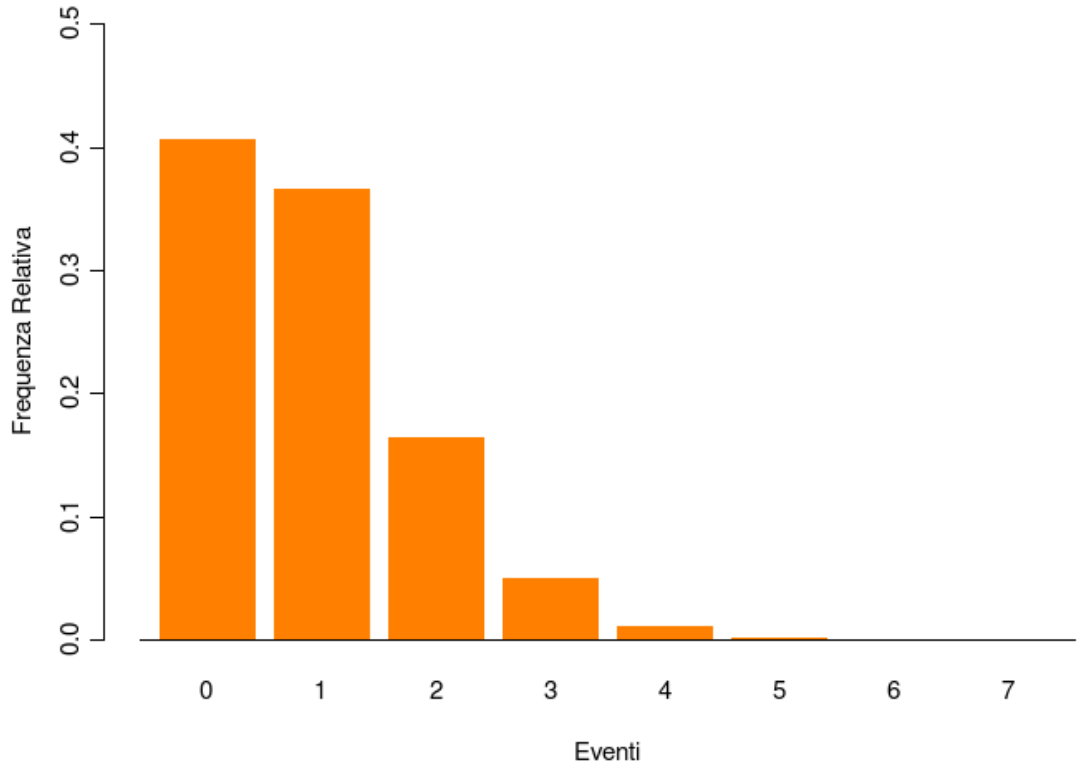
In a great number of books, the mean number of events is not shown with μ but with $\boldsymbol{\lambda}$, particularly when it deals with temporal events.

In order to have the Poisson distribution, a random variable must have **three requirements: stationarity, not-multiplicity, independence**.

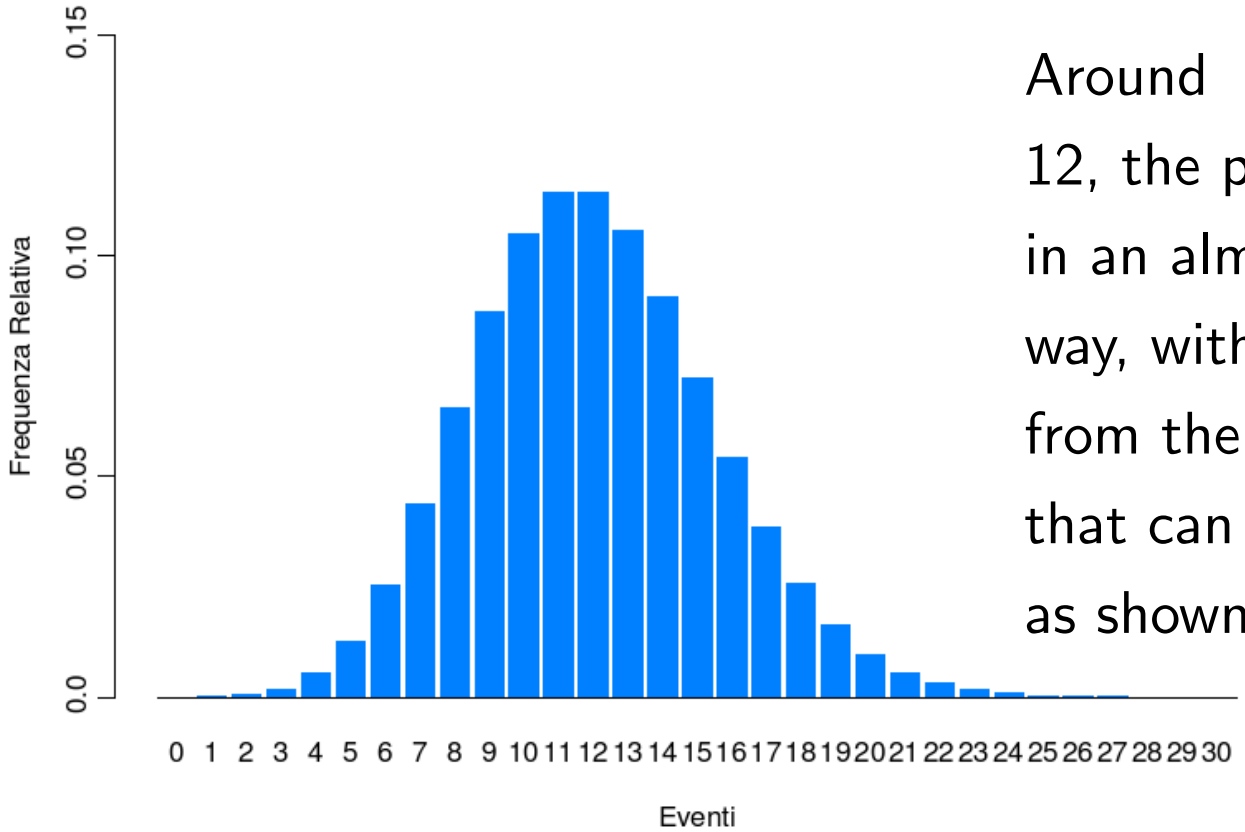
A random variable :

- ⋆ is **stationary**, when the probability of each event in a time interval $(t, t+h)$ or in an infinitesimal space is approximately constant, equal to λh for each t.
- ⋆ It has the **not-multiplicity** requirement, when the probability that two or more events happen in the same infinitesimal time or space interval λh is highly smaller than in the case of only one event. This means that the probability to have two or more events is highly smaller than to have only one event;
- ⋆ is **independent** if, in an interval and in a finite space, the events are independent one from the others and the number of events that happened in an interval is independent from that of the others.

As the chart highlights, the probabilities distribution with mean(μ) 0.9 is highly asymmetric, with right asymmetry.



The probabilities distribution with $\mu = 12$ has an "almost normal" shape: the higher probabilities regard the events around the mean 12.



Around 12, the probabilities decrease in an almost asymmetric way, with differences from the normal distribution that can be ignored, as shown in the graphic.

It is supposed to throw two dices with twelve faces for N times and to sum the result of the two dices.

We repeat the experiments for P times.

Each of the P times we write down how many times the sum is equal to two and to thirteen.

Example: Poisson and Binomial simulations

```
n = 1000
r2 = numeric(n)
r13 = numeric(n)
for ( i in 1:n) {
d1 = sample(1:12, n, replace = T)
d2 = sample(1:12, n, replace = T)
d = d1+d2
r2[i] = length(d[d == 2])
r13[i] = length(d[d == 13])}
par(mfrow = c(1, 2))
hist(r2, prob = T, ylim = c(0, .16), main = "Poisson: sum = 2")
hist(r13, prob = T, ylim = c(0, .16), main = "Binomiale: sum = 13")
```

All the previous models give the theoretical distribution of discrete random variables. When it is necessary to describe **continuous and positive random variables**, as measurements or time, the most useful models are the ones that follow.

Among them, the most frequent distribution and the most useful one for the experimental research, which is the **basis of the parametric statistics**, is the **normal or Gaussian distribution**.

The most important continuous distribution is the normal curve.

The name **normal curve** comes from the convinction (not always correct) that a great number of phenomenon, both the biological ones and the physical one, are usually distributed according the **Gaussian curve**.

Its name **curve of accidental errors**, mainly spread in the phisical subjects, comes from the experimental observations where the distribution of errors, committed when the same size is measured many times, results to be well approximated by this curve.

The normal distribution is defined by two parameters:

- μ, location parameter, that establishes the position of the distribution;
- σ, scale parameter, that establishes the size of the distribution.

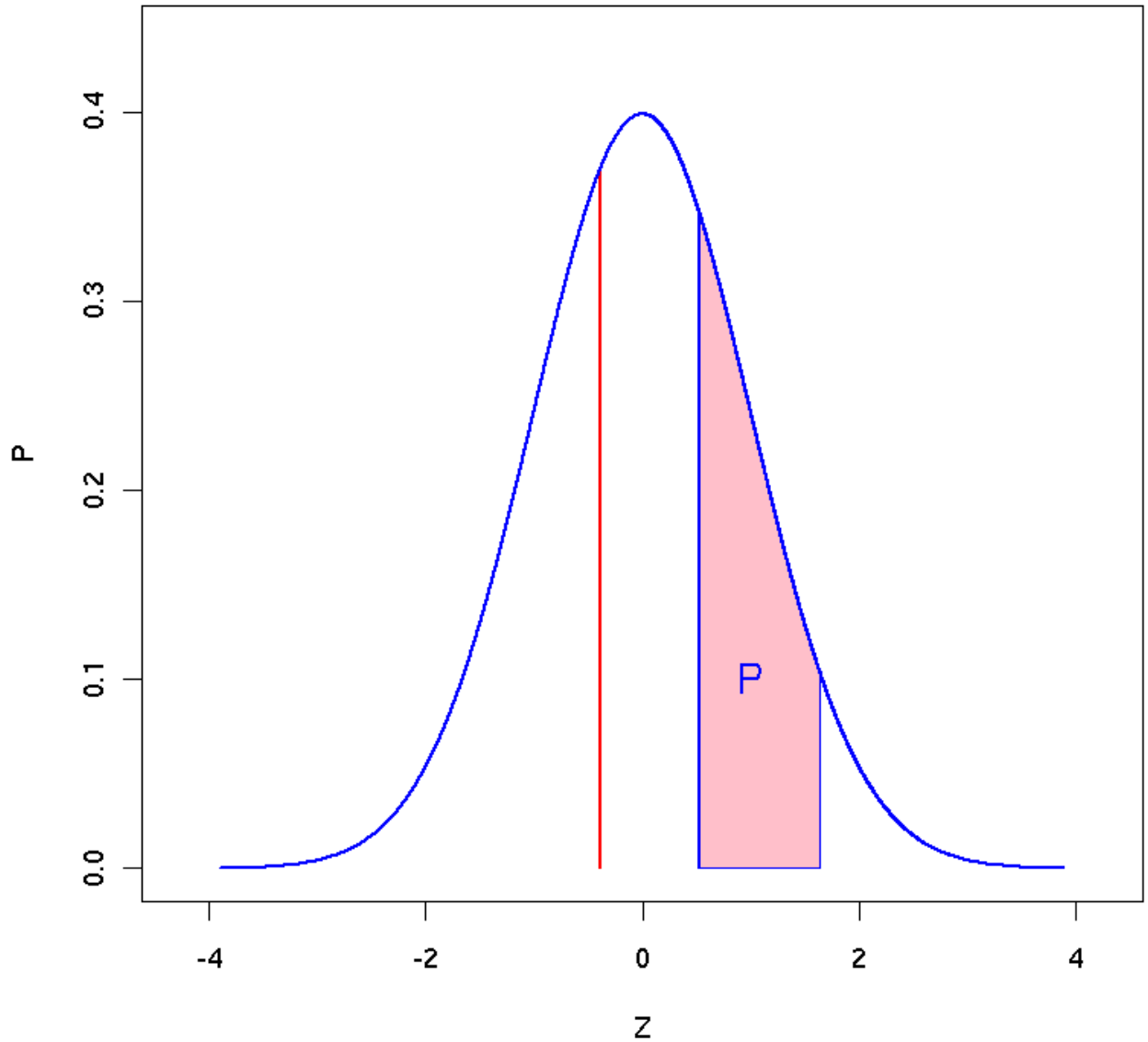
It is defined from $-\infty$ to $+\infty$, and, as it is a probability distribution, its integral from $-\infty$ to $+\infty$ is equal to 1.

At the extremes, it asymptotically tends to zero, and in the tails over the value 6σ, the area underlying is so limited that, with a good approximation, can be considered cut off.

From the theoretical point of view, the normal distribution well represents observational errors, the noise and other phenomenon using a mathematical model.

The **probability** that the variable has a value between two specific values, corresponds to the measure of the area underlying the curve included between the two values (the area represented in the following picture).

The probability that the variable as a precise value is equal to zero, as the area (integral) included between the value and itself is null (the segment represented in the picture).


P
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Z
P

From the mathematical point of view, the Gaussian distribution can be considered as the limit of the binomial distribution:

- ⋆ for n that tends to infinite;
- ⋆ while both p and q don't tend to 0 (this condition distringuishes the Gaussian distribution from the Poisson one).

If n tends to infinite and p stays constant, the mean $(n \cdot p)$ is approximated to infinite and makes the distribution without any practice applications.

On the contrary, the considered variable, that with few data was qualified by discrete unities, can be expressed by always smaller unities. It will become acceptable to define it as a continuous quantity.
The Gaussian distribution can be considered also the limit of the Poisson distribution, when i e μ become really high.

The mathematical formula of the normal distribution is the following one:

$$y = f(x) = N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

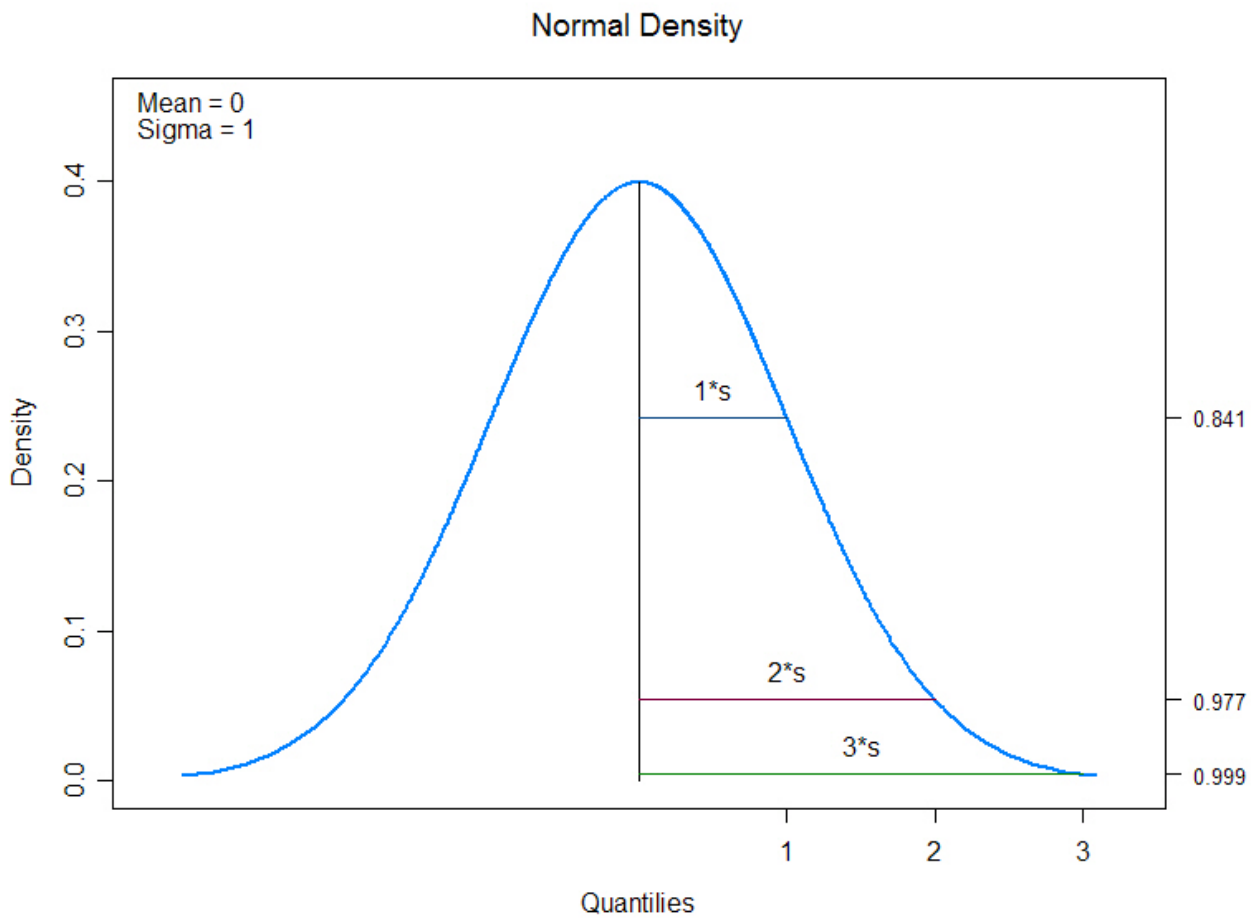
that is the expression of the **probability density function** (or relative frequencies function) of the normal distribution.

In less mathematical words, **it allows to estimate the y value (the value of the y-axis or the height of the curve) for each x value (the x-axis value)**.

The μ and σ values completely define the normal density function.

Infinite normal density curve exist.

This is the graphical representation of a Normal.



The most important characteristics of the normal distribution are: a relatively higher frequencies of the central values and gradually lower frequencies towards the extremes.

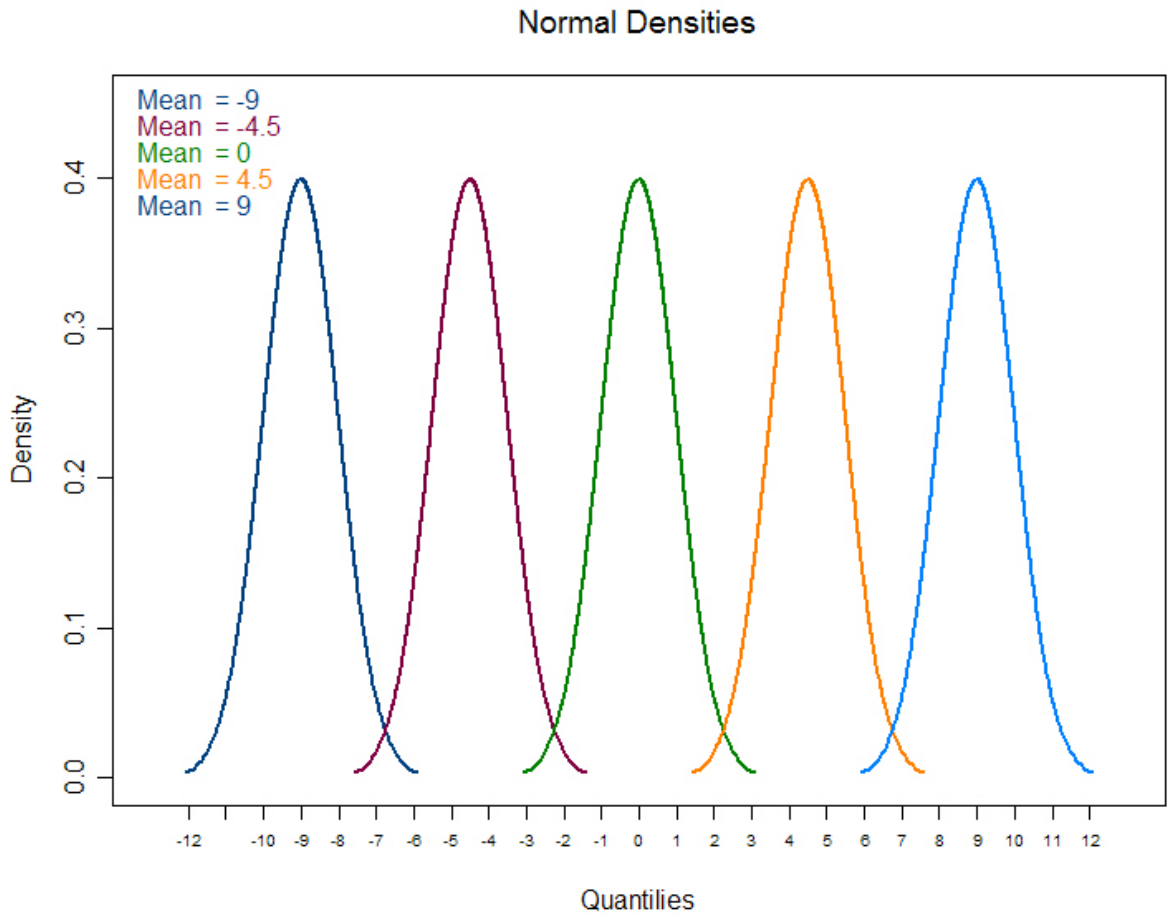
The density function is symmetric compared to/around the mean: it increases from $-\infty$ to the mean and then it decreases up to $+\infty$.

It has two inflection points: the first one, rising, in the point $\mu - \sigma$; the second one, descending, in the point $\mu + \sigma$.

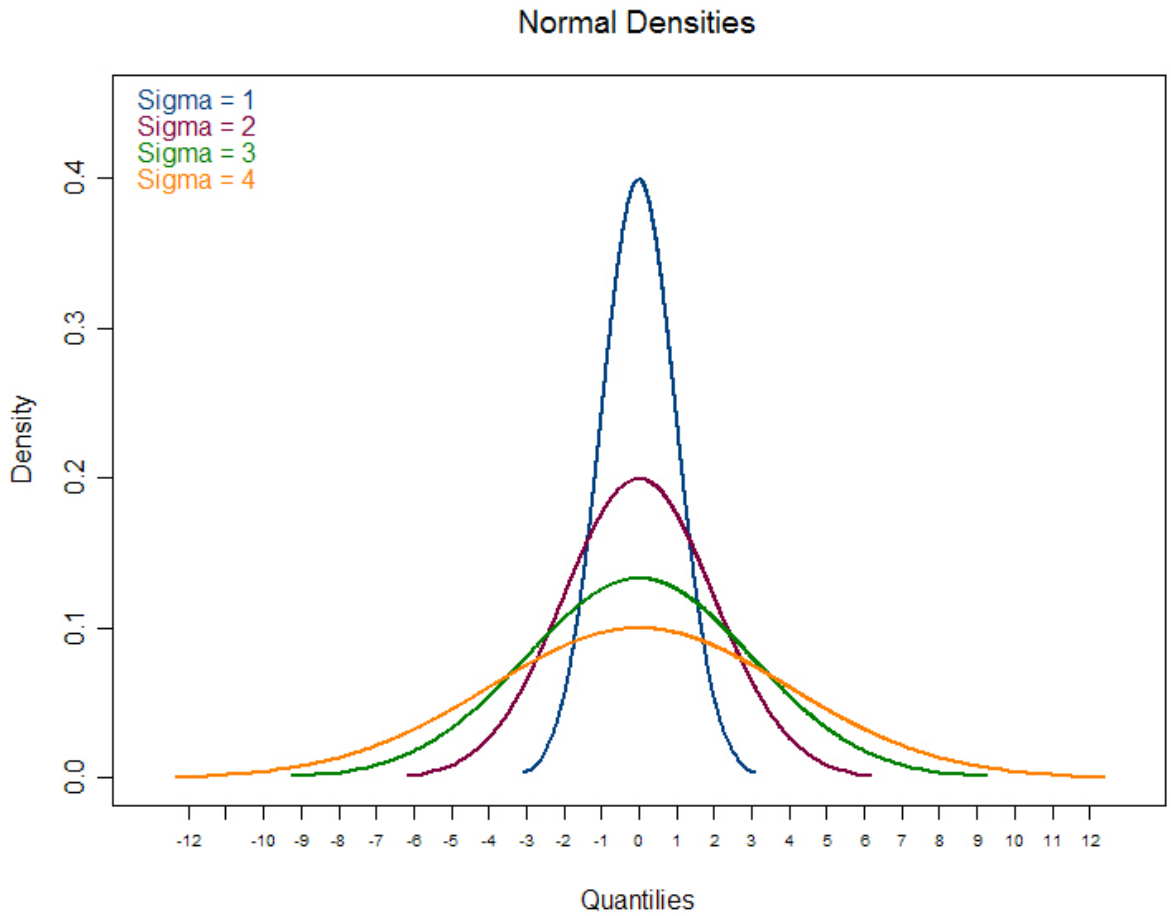
In every normal curve, the mean, the mode and the median coincide.

If the distribution is normal, in order to know its distribution, it is enough to know two parameters of a series of data, the mean μ and the variance σ^2.

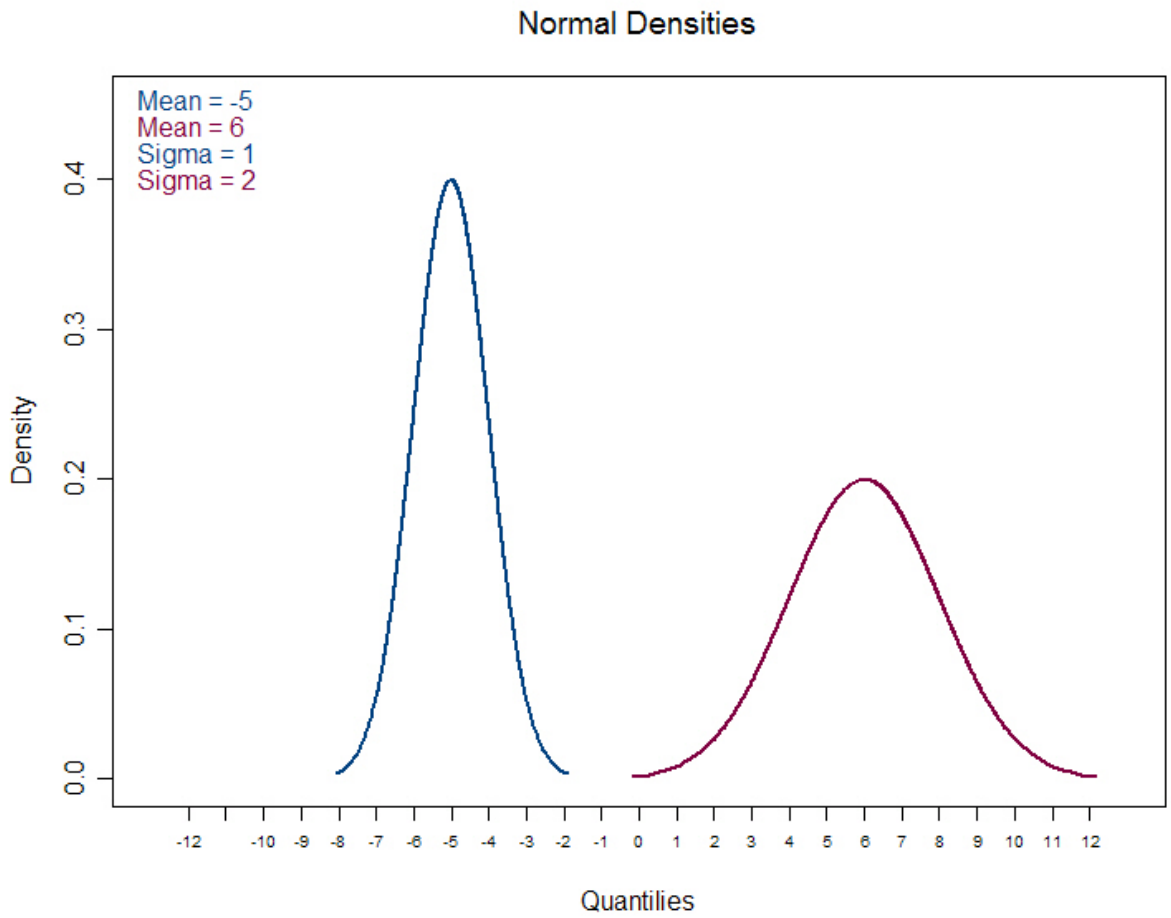
If μ changes and σ stays constant, the result is an infinite number of normal curves with the same shape and dimension, but with the symmetric axis in a different point.



If, on the contrary, μ stays constant and σ changes, all the infinite curves have the same symmetric axis; but they are almost flat, according to the σ value.



These are two normal distributions which are different both for their mean (μ) and for the dispersion of data (σ).



The infinite shapes of the normal distribution, determined by the combination of the differences of the mean and the variance, can be all connected to the same shape.

It is the **standard normal distribution** or **reduced normal**, which is obtained by the transformation of the variable given by:

$$Z = \frac{X - \mu}{\sigma}$$

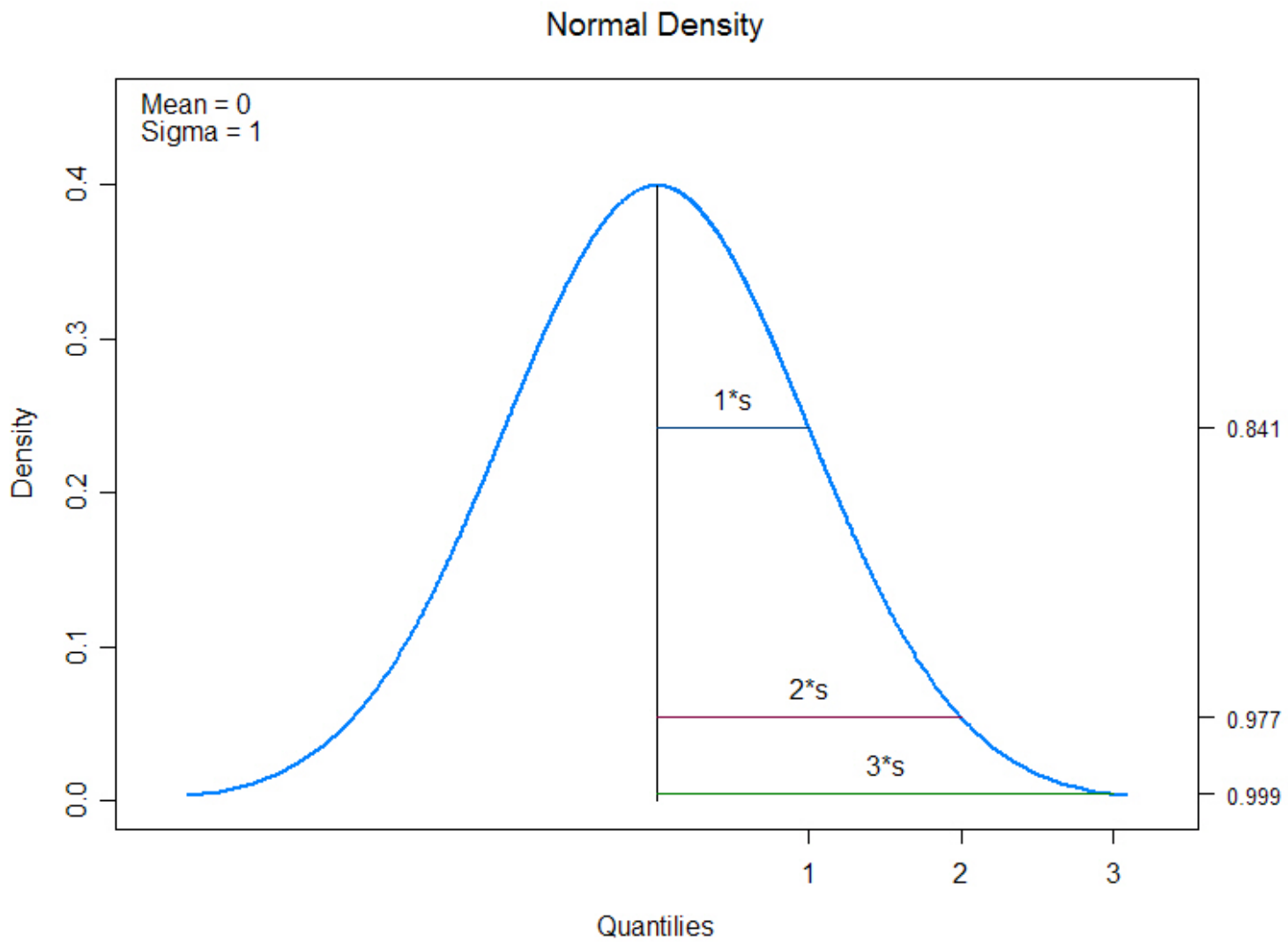
The **standardizing** is a transformation that consists in:

- ⋆ making the mean equal to zero ($\mu = 0$), because the mean is subtracted to every value;
- ⋆ taking the standard deviation (σ) as unit of measurement ($\sigma = 1$) of the new variable.

After the change of the variable, in the reduced normal, the probability density is given by:

$$y = f(z) = N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

This formula highlights how **the distribution's shape does not depend either on the mean or on the variance of the original distribution**.


Normal Density
Mean = 0
Sigma = 1
Density
0.0
0.1
0.2
0.3
0.4
1*s
2*s
3*s
0.841
0.977
0.999
1
2
3
Quantilies

In the statistics practice, **the most useful characteristics of the normal distribution** are not the ratio between x-axis and y-axis, explained before, **but the relations between the distance from the mean and the probability density represented by the curve**.

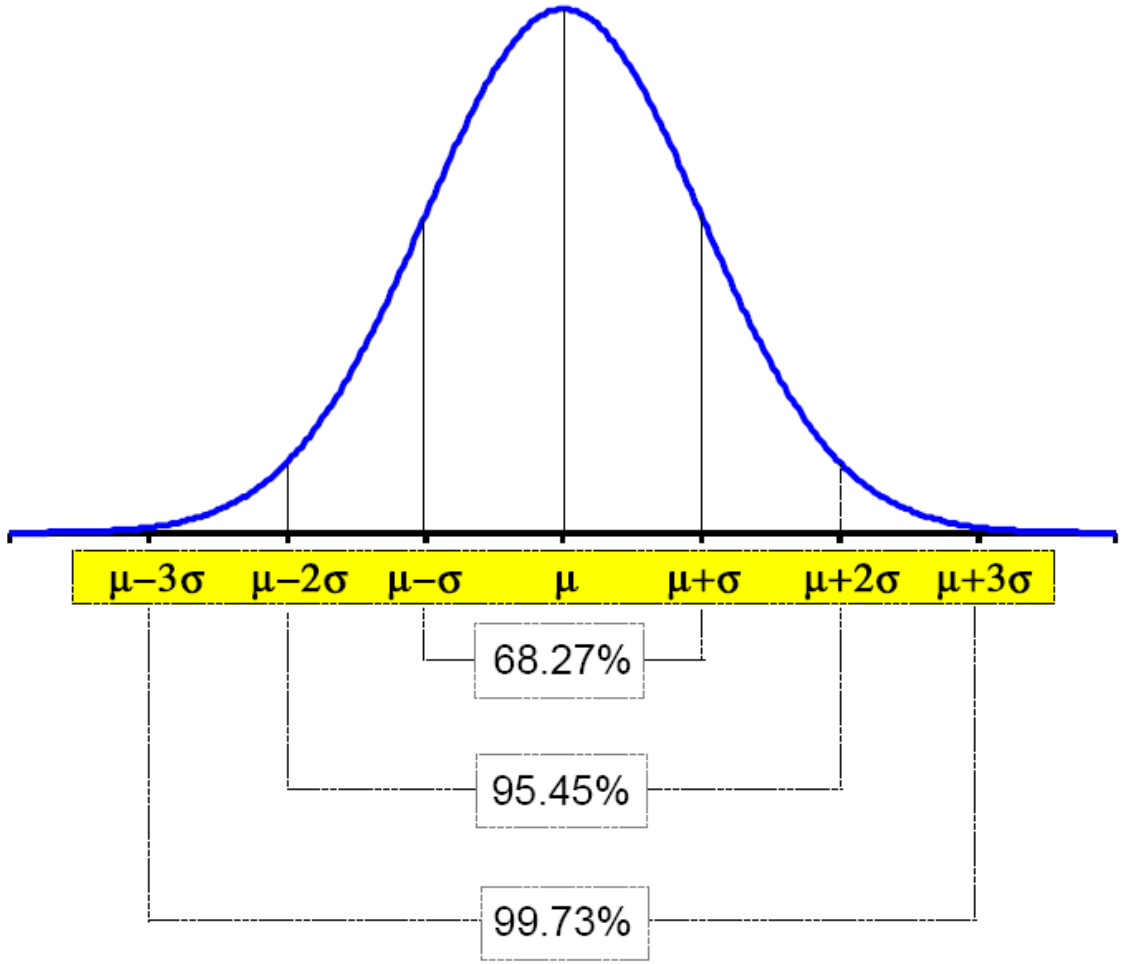
In an easiest way, it is possible to define how many data are included between the mean and a specific value. This is obtained by measuring the distance from the μ mean in σ standard deviations unities.

The fraction of the cases included:

- ⋆ between $\mu + \sigma$ and $\mu - \sigma$ is equal to 68.27%;
- ⋆ that between $\mu + 2\sigma$ and $\mu - 2\sigma$ is equal to 95.45%;
- ⋆ that between $\mu + 3\sigma$ and $\mu - 3\sigma$ is equal to 99.73%.

In pratice, in the normal curve, almost all the data are included around the mean whose wideness is $\boldsymbol{3\sigma}$.

The relation between the percentage of data represented by the curve and the dimensions of the interval between two values, is a relevant characteristic in the applied statistics:



Various distributions, which are not compulsory normal or far from the normality, can become or be considered like that when:

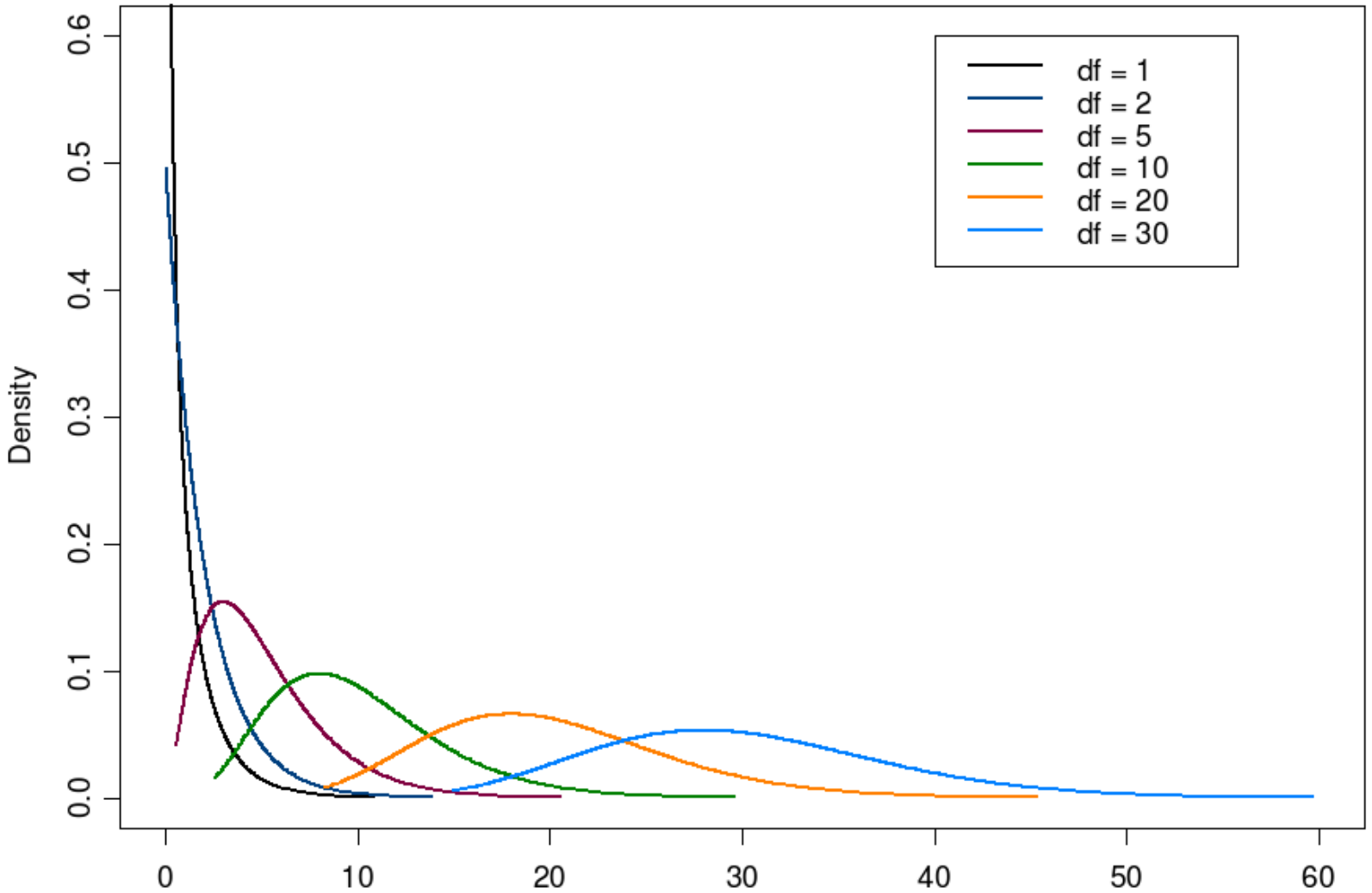
- ⋆ some of their parameters tend to infinite (**asymptotically normal**);
- ⋆ are "almost normal" (**approximations**);
- ⋆ can be appropriately transformed (**tranformations**).

Given n independent random variables $Z_1,\ Z_2,\ \ldots,\ Z_n$, normally distributed with $\mu = 0$ and $\sigma = 1$, the $\boldsymbol{\chi^2_\nu}$, where $\nu = n$, is a random variable obtained by the sum of their squares.

The density function of the χ^2 is obtained **only from the parameter ν**, the number of degrees of freedom (df).

The distribution is defined between 0 and $+\infty$, and when the number of degrees of freedom increases, it tends to have a shape similar to that of the normal.

Graphical representation of the distribution χ^2 for different values of df.



The **Student's t**-distribution takes into consideration the relations among the mean and the variance estimations, **in little size samples**, when the sample variance is used.

The choice between the use of the normal distribution and the Student's t-distribution in the comparison among means comes from the knowledge of the σ^2 variance of the population or from the fact that it is unknown. In this last case, it is necessary to use the s^2 sample variance.

The Student's t random variable is defined as the ratio between a standard normal random variable and the square root of a χ^2 divided by its degrees of freedom:

$$t_\nu = \frac{Z}{\sqrt{\chi^2_\nu / \nu}}$$

If a series of n observations $(X_1, \cdots, X_n)$ is drawn by a normal distribution, it is possible to show that the ratio

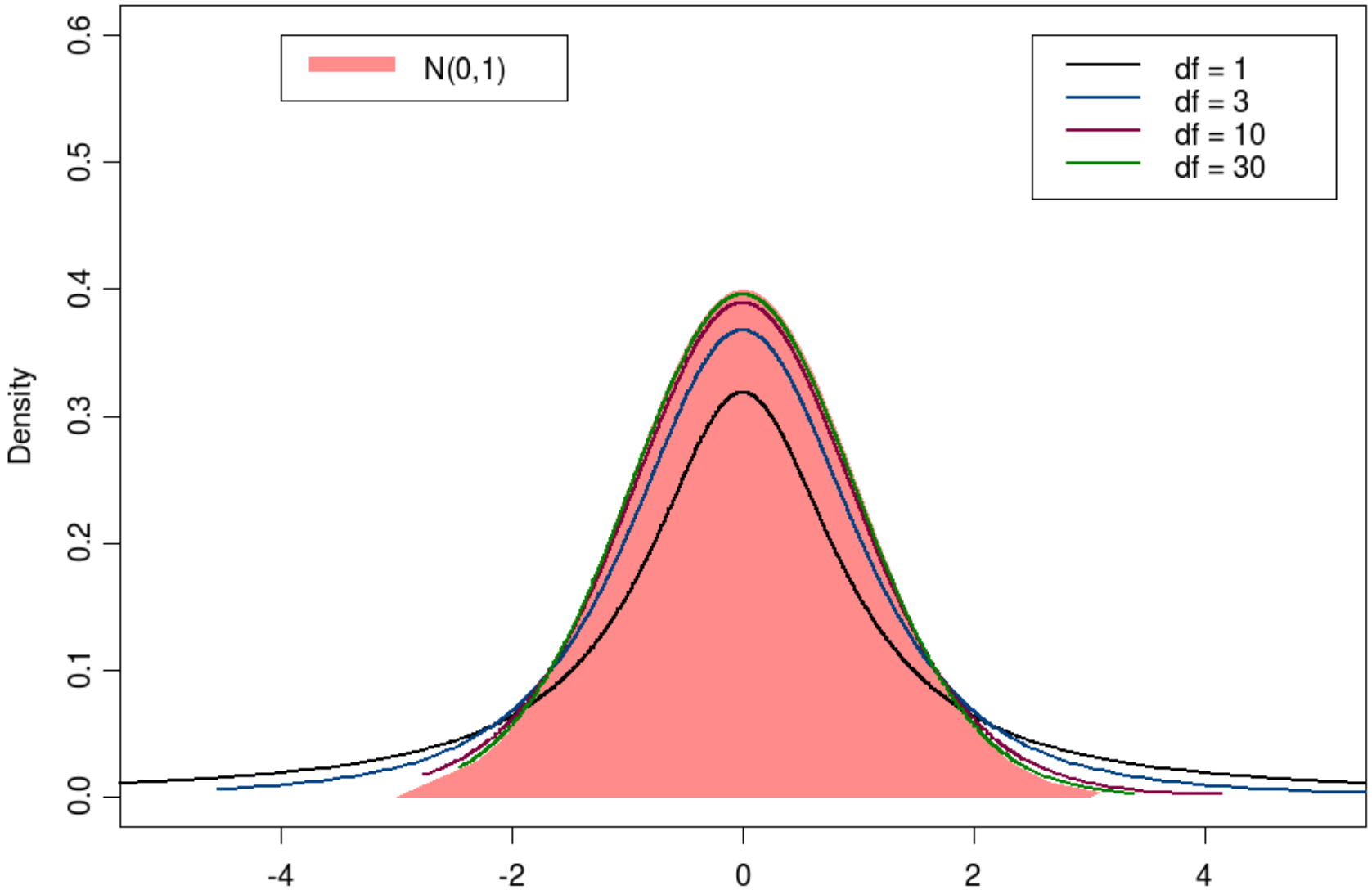
$$t_{n-1} = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

is distributed as a Student's t with $n-1$ degrees of freedom.

The corresponding curve is defined between $-\infty$ and $+\infty$, symmetric, little lowly that the normal and with higher frequencies in the extremes, when the number of df (ν) is really small.

Starting from a number of degrees of freedom around 30, the Student's t-distribution tends to a normal distribution.

This is a graphical representation of a Student's t distribution:



Another interesting distribution, from which it is founded the inference of a big part of the parametric statistics, is the **F-distribution**.

It corresponds to the distribution of the **ratio of 2 chi-squared indipendent random variables** (A e B), **divided by its corresponding degrees of freedom** (m e n).

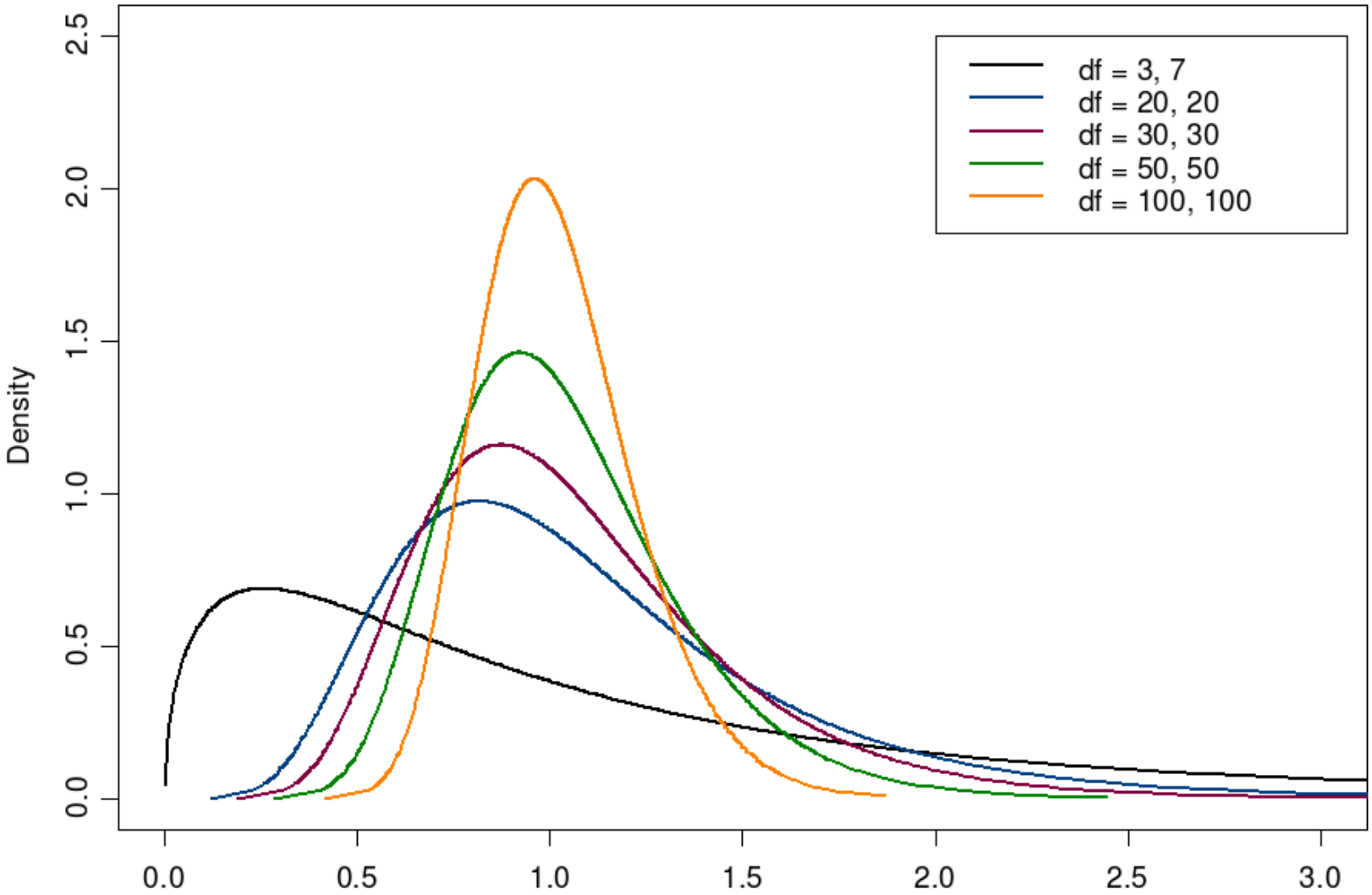
This **ratio F is defined between 0 and** $+\infty$**.**

The F distribution represents a **generalization** of the t distribution.

The t and χ^2 distribution have only one parameter, the F has **two parameters**: the number of degrees of freedom both of the numerator and of the denominator.

The square of a random variable t of Student with n degrees of freedom is equal to a F of Fisher distribution with degrees of freedom 1 **and** n**.**

This is the graphical representation of a F-distribution



A distribution which is highly used in the risk analysis and for extreme events, is the **lognormal distribution**.

$$y = f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \cdot exp\left\{-\frac{(log(x) - \mu)^2}{2 \cdot \sigma^2}\right\}$$

It corresponds to normal data distribution transformed by the exponential function.

That is: lognormal data transformed by a natural logarithm are distributed as a normal. Normal data transformed by the exponential are distributed as a lognormal.

The lognormal distribution **is defined between 0 and** $+\infty$, extremes excluded.

The lognormal distribution mean is:

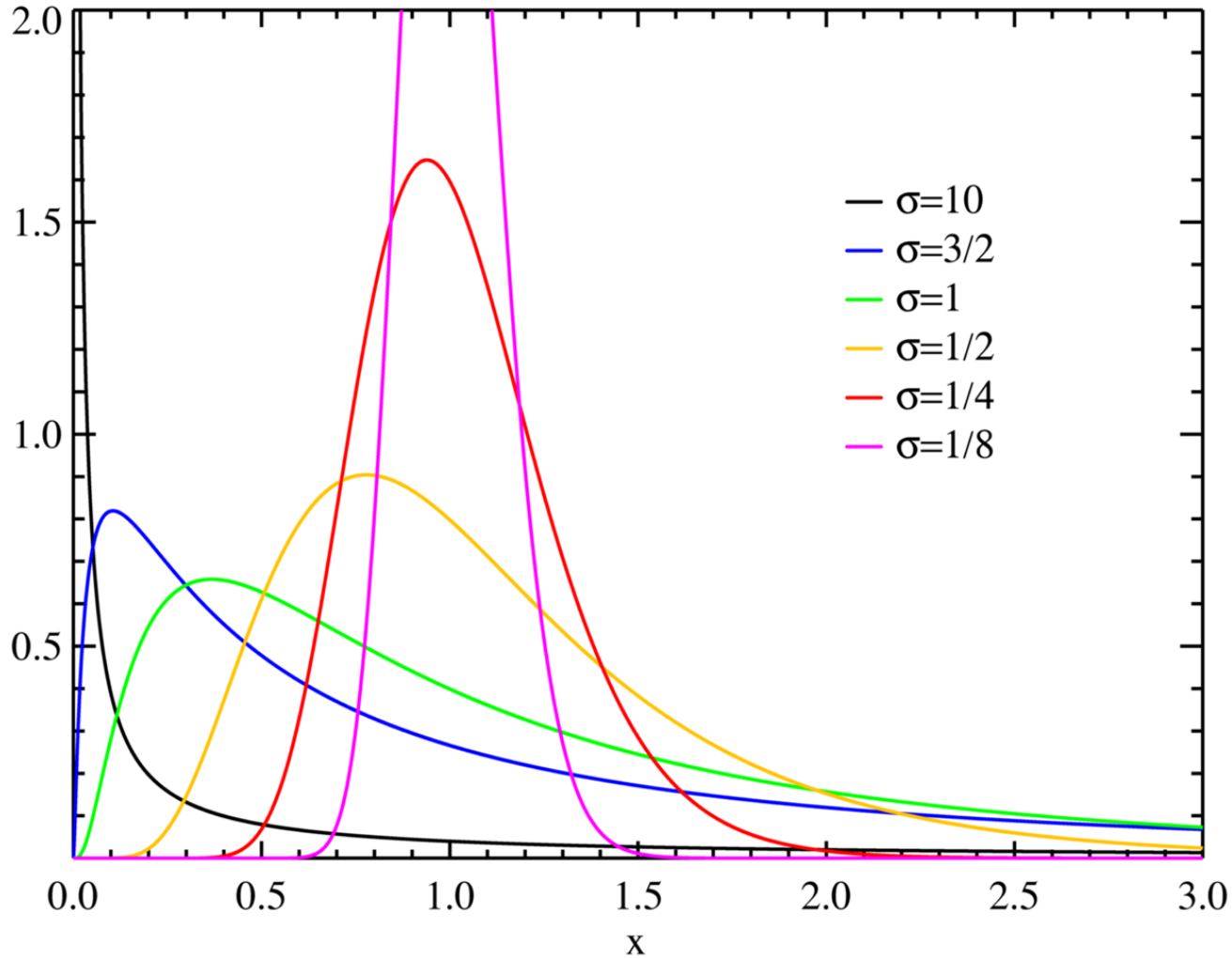
$$E[X] = e^{\mu+\sigma^2/2}$$

The lognormal distribution variance is:

$$Var[X] = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$$

Using the lognormal distribution is quite difficult because the parameters have a big variability for little variations of a right tale. In order to modify the right "tale" 's behavior of the lognormal distribution, it is often necessary a big variation of the parameters' values.
This leads to an "instability" of the parameters' estimations (an "high" value more or less can lead to estimated parameters' values really different among them).

Graphical representation of the lognormal, when σ values change:



The **Weibull distribution** is another distribution which is highly used in the risk analysis and for extreme events.

$$y = f(x; \lambda, k) = \frac{k}{\lambda} \left(\frac{x}{\lambda} \right)^{k-1} \cdot e^{-\left(\frac{x}{\lambda}\right)^{k-1}}$$

Where $k > 0$ is called the **shape** parameter and $\lambda > 0$ is called the **scale** parameter.

The Weibull distribution **is defined between 0 (included) and $+\infty$.**

The Weibull distribution mean is:

$$E[X] = \lambda\Gamma\left(1 + \frac{1}{k}\right)$$

The Weibull distribution variance is:

$$Var[X] = \lambda^2\Gamma\left(1 + \frac{2}{k}\right) - \left[\lambda\Gamma\left(1 + \frac{1}{k}\right)\right]^2$$

The Weibull distribution, as the lognormal one, is difficult to use because of its parameters' variability for little variations of the right tale. In order to modify the behavior of the right "tail" of the Weibull distribution, it is often necessary a big variation of the parameters' values. This leads to an "instability" of the parameters' estimations.

Graphical representation of the Weibull distribution, when k values change: