*Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.*
*http://www.simfit.org.uk*

The normal distribution has great importance in data analysis because, although experimental measurements never follow normal distributions exactly, many observations are approximately normally distributed, or become so after transformations such as replacing observations by the logarithms. For instance experimental error is often approximately normally distributed

## Definitions

A random variable $X$ is said to normally distributed if the probability density function (pdf) $f(x)$ and cumulative distribution function (cdf) $F(x)$ are

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$
$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

so that the probability of a value occurring in the range $A, B$ with $B > A$ is

$$P(A \le X \le B) = F(B) - F(A).$$

Here $\mu$ is the mean, the standard deviation is $\sigma$, and the variance is $\sigma^2$. Because $\mu$ can have any value at all and $\sigma$ can have any positive value it is useful to consider the standardized variable $Z$ defined as

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

which is normally distributed with mean zero and variance one.

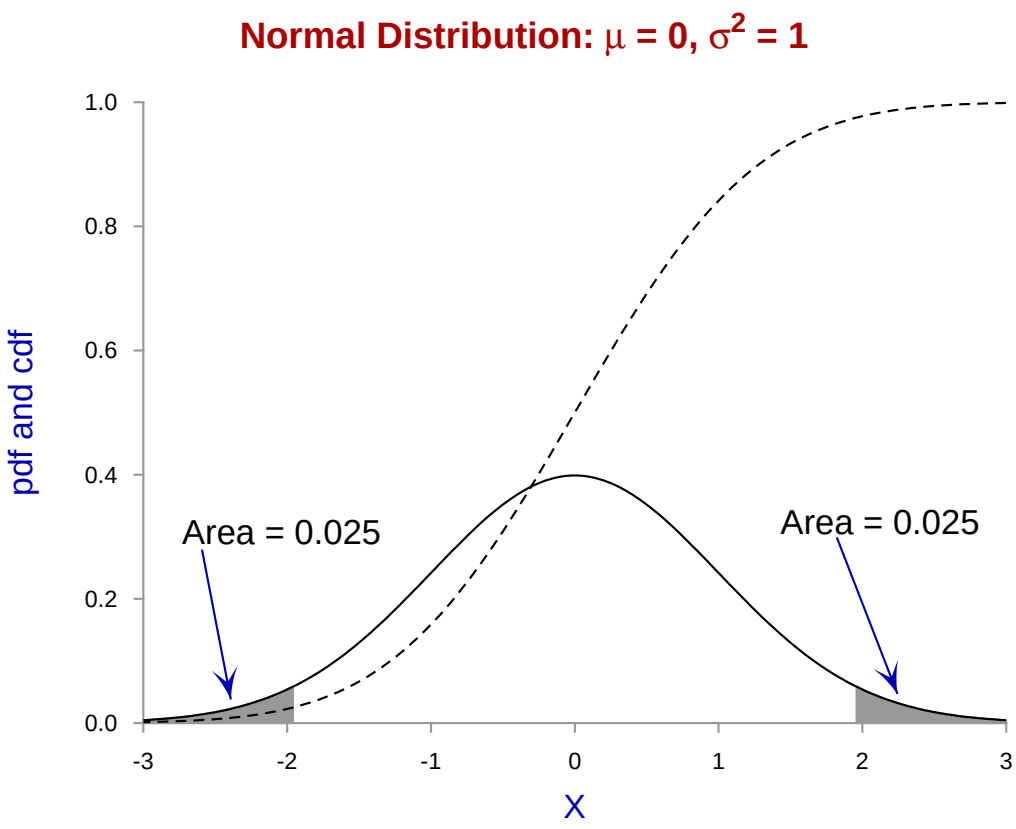
## Simfit program normal

In order to understand this distribution by plotting profiles and calculating deviates we shall consider some of the procedures available using the SimFIT program **normal**. To do this, select menu item `[A/Z]` from the main SimFIT menu, and open program **normal** when the following options will be available

```
Input: mu and sigma
Input: x, calculate pdf(x)
Input: x, calculate cdf(x)
Input: alpha, calculate x
```

as well as options to test if data are normally distributed, to perform power and sample size calculations, or to investigate the multivariate normal distribution.

If the default values for $\mu$ and $\sigma$ are accepted then the following plot can be obtained.




This illustrates that the normal distribution pdf is a symmetrical bell-shaped curve with tails that rapidly decrease after some two standard deviations. The cdf on the other hand is a monotonic sigmoidal curve rising from a minimum value of zero to a maximum value of one. It is in fact the integral of the pdf, that is, the value of $F(x)$ at the value $x$ is simply the area under the pdf curve from $-\infty$ to $x$.

Particular interest attaches to the area in the lower and upper tails of this distribution. In fact, the tails illustrated in this figure are the lower and upper 2.5% points. In other words, the probability of a value occurring in the lower tail is 0.025, the probability of a value occurring in the upper tail is 0.025, so that the probability of a value occurring in either the lower or upper tail is obviously 0.05. Perhaps the best known 2-tail critical points are the 68% and 95% ones, i.e.

$$P(\mu - \sigma \le X \le \mu + \sigma) = 0.68, \quad \text{and} \quad P(\mu - 1.96\sigma \le X \le \mu + 1.96\sigma) = 0.95.$$

Out of interest, it should be pointed out that the tails were shaded in this figure by using the advanced option to transfer the data into the SimFIT program **simplot** followed by assigning the first two lines to be closed polygons which were then colored grey.

Before the widespread availability of computers, values of such critical points were read off from tables, and also the inverses were obtained in this way, that is the values of $X$ calculated from specific values of $F(x)$. We now explain how to do this using SimFIT program **normal**.

## Obtaining critical values

SimFIT program **normal** was used to obtain three values of $-1, 0, 1$ for the pdf, the same three for the cdf, and three critical values for 2.5%, 5.0% and 50.0% as in this table from the results log file, which was archived by SimFIT when program **normal** was closed.

```
pdf values
Current parameters: mu = 0.0E+00, sigma = 1.0E+00, sigma^2 = 1.0E+00
pdf(-1.000E+00) = 2.420E-01
pdf( 0.000E+00) = 3.989E-01
pdf( 1.000E+00) = 2.420E-01

cdf values
Current parameters: mu = 0.0E+00, sigma = 1.0E+00, sigma^2 = 1.0E+00
P(X =< -1.000E+00) = 0.1587 ... P(X >= -1.000E+00) = 0.8413
P(X =<  0.000E+00) = 0.5000 ... P(X >=  0.000E+00) = 0.5000
P(X =<  1.000E+00) = 0.8413 ... P(X >=  1.000E+00) = 0.1587

critical points
Current parameters: mu = 0.0E+00, sigma = 1.0E+00, sigma^2 = 1.0E+00
P(X =<  1.960E+00) = 0.9750 *** P(X >=  1.960E+00) = 0.0250
P(X =<  1.645E+00) = 0.9500 *** P(X >=  1.645E+00) = 0.0500
P(X =<  0.000E+00) = 0.5000 *** P(X >=  0.000E+00) = 0.5000
```

The pdf values illustrate in numbers what is displayed in the graph, that $f(-1) = f(1)$ because of the fact that values equally spaced below and above the mean give the same pdf values due to the symmetry.

The cdf values also illustrate that the areas in the lower and upper tails at values equally spaced below and above the means are equal, and clearly the areas below and above the mean are 0.5.

The critical points illustrated show the same symmetry, but it should be emphasized that using lower and upper critical points for statistical testing would normally require critical points based on the sum of lower and upper tail probabilities, as in a two-tail test. For instance, if a statistical test is conducted to see if an observation is consistent with a certain mean, then the two-tail test would allow for the observation being either extremely low or extremely large. If the analyst was just not prepared to consider such an outcome but would only countenance the possibility of an observation being too large for the null hypothesis, or too small as the case may be, would a one-tail test based on only one of the tails be used.

A trick often resorted to when submitting grant proposals is to do power and sample size calculations using one-tail tests when two-tailed would be more honest. Like claiming a lower variance than is justified experimentally to reduce the sample size required, statistics is always open to such abuse.