

Probabilistic models

Menu: QCExpert Probabilistic models

The module fits statistical models from various classes by the method of maximum likelihood (MLE, *Maximum Likelihood Estimate*). There are 11 univariate distribution types available for use, 5 of them are symmetric, remaining 6 are asymmetric. For each distribution type, the module fits a model of a given type by looking for the best parameter values via maximization of the likelihood function. Best fitting models found in each distributional class are compared across classes using the following two criteria: likelihood function value and linearity (correlation coefficient) of the P-P plot. Within a distributional class, best fitting parameters are found by the numerical maximization of the likelihood function, or of its logarithm L , respectively,

$$L(A, B, C, \mathbf{x}) = \sum_{i=1}^n \ln f(A, B, C, x_i),$$

where A, B, C are parameters of the given distribution type, \mathbf{x} is the n vector of measured data, and f is probability density of the given distribution, see below. In addition to the parameter estimates for each of the models and some further statistics, the module can produce diagnostic plots and user-required quantiles. The module can be useful in situations where data are not normally distributed. It can be used not only in the situation where the non-normal model to be used is known in advance (e.g. from a physical theory), but even if one is not sure about the distribution type that should be used to fit the data.

Data and parameters

The module is intended for analysis of the data, which come from *symmetrical* or *positively skewed* data. Negatively skewed data need to be multiplied by (-1) before the analysis. The data to be analyzed should be entered in one column, whose name is selected in the *Columns* field. In the *Distribution* field, one can choose which of the available models will be fitted. Symmetric distributions are listed on the left, asymmetric are listed on the right. By default, all 11 types are fitted. *Symmetric* and *Asymmetric* buttons can be used to enter all symmetric, respectively asymmetric distributions available. At least one distribution type has to be selected. Asymmetric distributions' fitting is generally longer. Hence, if we are not specifically interested in fitting asymmetric models, it is better not to choose them. If the data are symmetric, or if they have a positive skewness, skewed distribution fitting can even fail. Such a failure is indicated by the message „Not available“, or „Error“. One can choose in the Data field whether all data, marked only, or unmarked only data should be used for the calculations. If the *Calculate probability* selection is checked, the user has to supply a value in the *X* field. Probability that a random variable with the fitted distribution is smaller than or equal to this value (i.e. the cumulative distribution function value), will be returned by the software. This probability will be listed in Protocol for all models whose fitting is requested. If the *Compute quantiles* selection is checked, the user has to enter a probability, say p (a value between 0 and 1, $0 < p < 1$), for which the quantiles are to be evaluated. The Protocol lists p -th and $(1-p)$ -th quantiles for each of the fitted models.

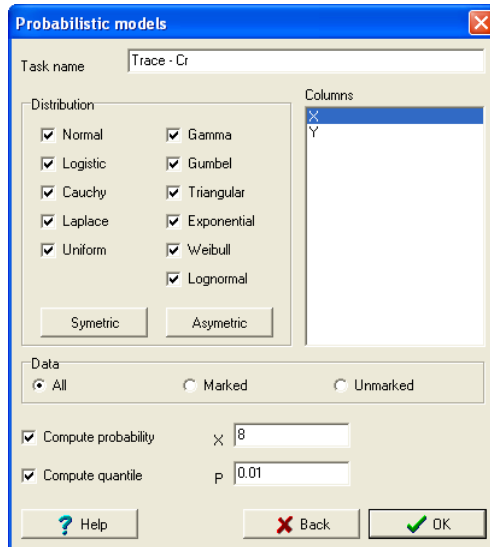


Fig. 1 Dialog panel for the Probabilistic models

Below, we list probability density functions for all available models. For each of them, parameter restrictions are listed (if there are any). Probability density function is generically denoted by $f(x)$, while the cumulative distribution function is denoted by $F(x)$, and the quantile function by $F^{-1}(x)$.

Normal distribution:

$$f(x) = \frac{1}{B\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-A}{B}\right)^2\right), \quad A \in \mathbb{R}, B > 0$$

Uniform distribution:

$$f(x) = \begin{cases} \frac{1}{B-A} & \text{pro } A \leq x \leq B \\ 0 & \text{jinak} \end{cases}, \quad A, B \in \mathbb{R}, A < B$$

Laplace distribution:

$$f(x) = \frac{1}{2B} \exp\left(-\frac{|x-A|}{B}\right), \quad A \in \mathbb{R}, B > 0$$

Logistic distribution:

$$f(x) = \frac{1}{B} \frac{\exp\left(\frac{x-A}{B}\right)}{\left[1 + \exp\left(\frac{x-A}{B}\right)\right]^2}, \quad A \in \mathbb{R}, B > 0$$

Cauchy distribution:

$$f(x) = \frac{1}{\pi B \left[1 + \left(\frac{x-A}{B}\right)^2\right]}, \quad A \in \mathbb{R}, B > 0$$

Asymmetric distributions

Exponential distribution:

$$f(x) = \frac{1}{B} \exp\left(\frac{A-x}{B}\right), \quad x \geq A, B > 0$$

Gamma distribution:

$$f(x) = \frac{1}{B\Gamma(C)} \left(\frac{x-A}{B}\right)^{C-1} \exp\left(\frac{A-x}{B}\right), \quad x > A, B > 0, C > 0$$

Gumbel distribution:

$$f(x) = \frac{1}{B} \exp\left(\frac{A-x}{B}\right) \exp\left(-\exp\left(\frac{A-x}{B}\right)\right), \quad A \in \mathbb{R}, B > 0$$

Lognormal distribution:

$$f(x) = \frac{1}{C(x-A)\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\ln(x-A)-B}{C}\right)^2\right), \quad x > A, B \in \mathbb{R}, C > 0$$

Triangular distribution:

$$f(x) = \begin{cases} \frac{2(x-A)}{(B-A)(C-A)} \text{ pro } x < C \\ \frac{2(B-x)}{(B-A)(B-C)} \text{ pro } x \geq C \end{cases}, \quad x > A, B > C$$

Weibull distribution:

$$f(x) = \frac{C}{B} \left(\frac{x-A}{B}\right)^{C-1} \exp\left(-\left(\frac{x-A}{B}\right)^C\right), \quad x > A, B > 0, C > 0$$

Protocol

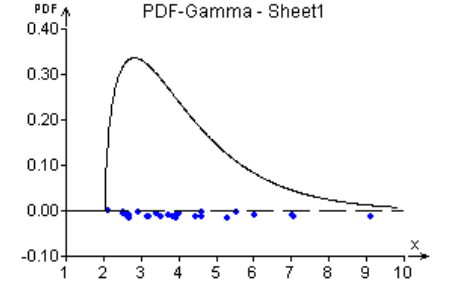
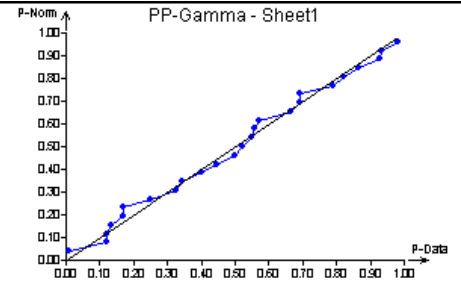
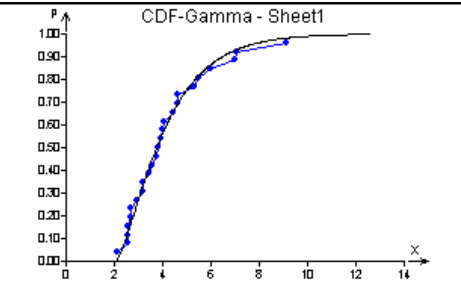
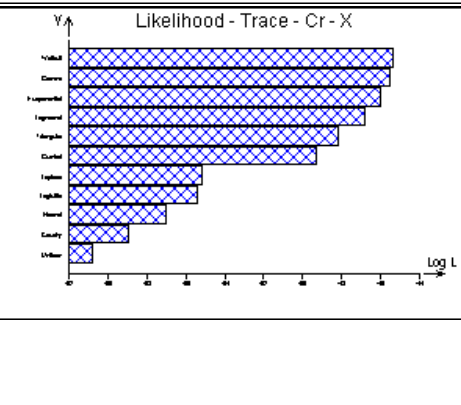
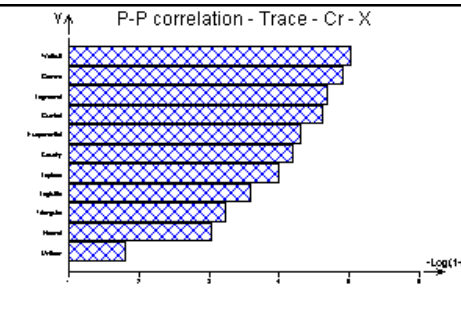
Probabilistic models, Maximum likelihood estimation (MLE)	
Task name:	Name of the spreadsheet containing data.
List of the distributions fitted	This paragraph lists all distributions, which were selected in the Dialog panel and which were fitted to the data subsequently. The list is organized into two parts: Symmetric models and Asymmetric models. For each model, log-likelihood value is listed, together with the correlation coefficient for the P-P plot, and MLE estimates of the parameters.
Loglikelihood	The value of the logarithm of the likelihood function, L (for a given distribution). Log-likelihood is suggested as the main criterion to look at, when comparing how well different distributional models fit the data. Large L values should correspond to better fitting distributions.
P-P correlation	Correlation coefficient, r_p from the P-P plot. This plot is similar to the Q-Q plot. It is true that the higher r_p (closer to one), the better is the empirical probability $i/(n+1)$ approximated by the theoretical model probability $F(x_i)$. The r_p can be looked at when comparing how good is the fit of different distributional types. This is an alternative to the criterion based on the (log)likelihood, and in general, it does not need to give the same results as before.
Parameters	Parameter estimates, obtained by the maximum likelihood method. Various parameters' meaning should be clear from the definitions of

	various distribution types, listed in the paragraph 0.
Sample moments	Various moment estimates.
Average	Arithmetic average of the data. $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Variance	Sample variance. $s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$
Skewness	Sample skewness. $a = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$
Kurtosis	Sample kurtosis. $b = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^4}$
Median	Sample median, $\tilde{x} = (x_{(n/2)} + x_{(n/2+1)})/2$ for n even, $\tilde{x} = x_{((n+1)/2)}$ for n odd.
Moments computed from the fitted models	Moments, median and mode, which are computed analytically as functions of model parameters. If no closed form is available, - appears in the table below.
Expected value	$\mu = \int_{-\infty}^{\infty} x dF(x)$
Variance	$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 dF(x)$
Skewness	$g_1 = \frac{1}{\sigma^3} \int_{-\infty}^{\infty} (x - \mu)^3 dF(x)$
Kurtosis	$g_2 = \frac{1}{\sigma^4} \int_{-\infty}^{\infty} (x - \mu)^4 dF(x)$
Median	$F^{-1}(0.5)$
Mode	Location of the maximum of the probability density function, $\arg \max_x f(x)$.
Quantiles and probabilities	Cumulative distribution function value at x (probability that a random variable with the fitted distribution is smaller than or equal to x). Or, p -th, $(1-p)$ -th quantiles for a given p . These are listed in the table for each of the fitted models when x and/or p is entered in the appropriate field of the Dialog panel (see Fig. 1).
Prob(X)	Cumulative distribution function value at x (probability that a random variable with the fitted distribution is smaller than or equal to x), $F(X)$
Quant(p)	p -th quantile (the value which will not be exceeded with the probability p , that is the value which will be exceeded with the probability $(1 - p)$), $F^{-1}(p)$.
Quant($1 - p$)	$(1-p)$ -th quantile, the value that will be exceeded with the probability p , $F^{-1}(1 - p)$.

Graphs

For each of the distributions selected, probability density curve is plotted, together with the P-P plot, and cumulative distribution function plot. For easier visual judgment of how good the model fit

is, individual data points are plotted below the plotted curve. The last two plots help to compare quality of the fit across different models.

 <p>PDF-Gamma - Sheet1</p>	<p>Probability density function, $f(x)$ plot. It is drawn for each of the models fitted. Individual data points are plotted below the x-axis. Definitions of probability density function for each of the models available are listed above. This plot has no direct diagnostic meaning.</p>
 <p>PP-Gamma - Sheet1</p>	<p>The P-P plot. Values of $i/(n+1)$ are plotted on the horizontal axis, while the fitted model cumulative distribution function evaluated at data points, $F(x_i)$ is plotted on the vertical axis. If the empirical and model distribution were exactly the same, then all the points would lie on the $y=x$ line (it is plotted as well, for comparison). How good is the data distribution approximated by a particular model can be checked informally from visual inspection of how close the data are to the ideal $y=x$ line. More formally, correlation coefficient can be used to measure this closeness. The correlation coefficient is listed in the Protocol, together with the loglikelihood maximum. The correlation is also listed on the <i>Quality of the fit</i> plot, see later.</p>
 <p>CDF-Gamma - Sheet1</p>	<p>Cumulative distribution function plot, $F(x)$ for the selected distributions. The model curves are superimposed by the empirical distribution curve to judge the goodness of model fit visually. Empirical distribution function points have y-coordinates $(x_i; i/(n+1))$.</p>
 <p>Likelihood - Trace - Cr - X</p>	<p>Overall plot which helps to decide, which models fit the data well, if judged by the loglikelihood values, L. maximum L for each of the selected distribution model type is plotted in the form of the bar chart. The bars are ordered from shortest to longest, for better orientation. The larger the maximum L, the better. If two distribution types show similar maximum L values, they should be judged as being equally good. The maximum value of L is (among other things) related to the sample size and to the data variability, so that L maximums should not be directly compared for datasets of different sizes, or with different standard deviations.</p>
 <p>P-P correlation - Trace - Cr - X</p>	<p>Overall plot which helps to judge which models fit the data well, if judged by the transformed correlation coefficient from the P-P plot, namely by $-\ln(1 - r_P)$. The transformation is used because the r_P values do not discriminate among various models very nicely, when judged visually. The higher $-\ln(1 - r_P)$ value, the better. Ordering of fitted distributional types according to the loglikelihood maximums and transformed correlation coefficients are generally not the same.</p>