Fitting parametric distributions using R: the `fitdistrplus` package

Marie Laure Delignette-Muller[1,2,3,*] **, Regis Pouillot**[4] **, Jean-Baptiste Denis**[5]

1. University of Lyon, Lyon, France
2. CNRS UMR5558, Villeurbanne, France
3. National veterinary school of Lyon, Marcy l'Etoile, France
4. 4515 Willard Ave., Chevy Chase, MD 20815, U.S.
5. MIA-INRA, Jouy-en-Josas, France.

* Contact author: ml.delignette@vet-lyon.fr

Keywords: fitting distributions, maximum-likelihood, goodness-of-fit, bootstrap, censored data

Fitting distributions consists of selecting the best fitting probability distribution from a predefined family of distributions. This practice is specially needed in the domain of Quantitative Risk Assessment. It requires judgment and expertise and generally needs an iterative process of distribution choice, parameters estimation, and quality of fit evaluation. The function `fitdistr` in `MASS` (Venables and Ripley, 2002) is a general-purpose maximum-likelihood fitting routine for the parameter estimation step. Other steps of the process may be developed using `R` (Ricci, 2005) but, to our knowledge, no specific package has been implemented for that purpose.

The package `fitdistrplus` provides several functions to help the fit of a univariate parametric distribution to data. Data may be continuous or discrete, and a specific approach is proposed for each of these two types of data. Continuous data may contain censored values (right-, left- and interval-censored with several upper and lower bounds) as frequently obtained as microbial or chemical analysis outputs used in risk assessment. More precisely, `fitdistrplus` is a set of integrated functions specifically written to:

- Help choose the best parametric distribution that fits a given dataset, using a skewness-kurtosis plot;
- For a given distribution, estimate the parameters using the maximum likelihood method or the method of matching moments and provide goodness-of-fit graphs (empirical and theoretical distributions plot in density and in cdf, P-P plot and Q-Q plot) and statistics (Chi-squared, Kolmogorov-Smirnov and Anderson-Darling statistics) to assess the fit;
- For a fitted distribution, simulate the uncertainty in the estimated parameters by parametric or non-parametric bootstrap resampling. This method may be used in risk assessment for describing an input by a distribution reflecting variability, conditionally to hyperparameters that are considered uncertain.

This package was first built to help the specification of distributions in quantitative risk assessment, but could be used more largely as an help to fit distributions to data, as it provides larger possibilities than the function `fitdistr`. While `fitdistrplus` is already available on the `CRAN`, new graphs for goodness-of-fit for censored data, new goodness-of-fit statistics and tests for non-censored and censored data are currently under development within the R-Forge project "Risk assessment with R"(Pouillot *et al.*, 2008).

References

Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth edition. Springer, New-York, U.S.

Ricci, V. (2005). Fitting distributions with R
`http://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf`

Pouillot, R., Delignette-Muller M.L., Denis, J.-B. (2008). Risk Assessment with R,
`http://riskassessment.r-forge.r-project.org/`.