Fitting parametric univariate distributions to non-censored or censored data using the R package **fitdistrplus**

Marie Laure Delignette-Muller and Christophe Dutang

March 18, 2013

The package **fitdistrplus** provides functions for fitting univariate distributions on different types of data (continuous censored or non-censored data and discrete data) and allowing different estimation methods (maximum likelihood, moment matching, quantile matching and maximum goodness-of-fit estimation). Outputs of `fitdist` and `fitdistcens` functions are S3 objects, for which kind generic methods are provided, including `summary`, `plot` and `quantile`. This package also provides various functions to compare the fit of several distributions to a same data set and can handle bootstrap of parameter estimates. Detailed examples are given in food risk assessment, ecotoxicology and insurance contexts.
Keywords: probability distribution fitting, bootstrap, censored data, maximum likelihood, moment matching, quantile matching, maximum goodness-of-fit.

1 Introduction

Fitting distributions to data is a very common task in statistics and consists in choosing a probability distribution that gives a good representation of a statistical variable, as well as finding parameter estimates of that distribution. It requires judgment and expertise and generally needs an iterative process of distribution choice, parameter estimation, and quality of fit assessment. The `fitdistr` function in the R package **MASS** ([62]) is a well known general-purpose maximum-likelihood fitting routine for the parameter estimation step in R [51]. Other steps of the process may be developed using R, e.g. [53]. In this paper, we present our R package **fitdistrplus** ([15]) implementing several methods of fitting univariate parametric distribution. Our first objective by developping this package was to provide R users a set of functions dedicated to help this overall process.

The `fitdistr` function estimates distribution parameters by maximizing the log-likelihood using the `optim` function. In some cases, other estimation methods could be prefered, such as maximum goodness-of-fit estimation (also called minimum distance estimation), and proposed in the R package **actuar** with three different goodness-of-fit distances, see [16]. While developping the **fitdistrplus** package, our second objective was thus to extend function `fitdistr` by providing various estimation methods in addition to maximum likelihood estimation (MLE). Functions were developped to enable moment matching estimation (MME), quantile matching estimation (QME), and maximum goodness-of-fit estimation (MGE) using eight different distances. Moreover, the **fitdistrplus** package offers the possibility to specify a user-supplied function for optimization, useful in cases where classical optimization techniques not included in `optim` are more adequate.

In applied statistics, it is frequent to have to fit distributions to censored data [32, 24, 9, 36, 10]. The **MASS** `fitdistr` function does not enable maximum likelihood estimation with this type data. Some packages deal with censored data, especially survival data [60] (see [25, 30] for examples), but those packages generally focused on specific models, enabling the fit of only one distribution or a restricted family of distributions. Our third objective was thus to provide R users a function to estimate univariate distribution parameters from censored data, whatever the type of censoring.

Few packages on CRAN provide estimation procedures for a general distribution and a general type of data. The **distrMod** package of [35] provides an object-oriented (S4) implementation of probability models and includes distribution fitting procedures for a given minimization criterion. In **fitdistrplus**, we use the standard S3 class system, we believe simpler than the full object-oriented S4 model for most R users. Furthermore, the **distrMod** package does not allow to fit censored data. The `mle` function of **stats4** package provides a procedure for maximum likelihood estimation whose output has class `"mle"`. Many generic methods are implemented for this type of object, e.g. `confint`, `logLik`,... When designing the **fitdistrplus** package, we also take this into account. Finally, various other packages provide functions to estimate the mode, the moments or the L-moments of a distribution, see the reference manuals of **modeest**, **lmomco** and **Lmoments** packages.

This manuscript reviews the various features of version 1.0-1 of **fitdistrplus**. The package is available from the Comprehensive R Archive Network at `http://cran.r-project.org/package=fitdistrplus`. The development version of the package is located at R-forge as one the packages of the project "Risk Assessment with R" (`http://r-forge.r-project.org/projects/riskassessment/`). The following command will load the package.

```
> library(fitdistrplus)
```

The paper is organized as follows: Section 2 present tools for fitting continuous distributions to classic non-censored data. Section 3 deals with other estimation methods and other types of data, before Section 4 concludes.

2 Fitting distributions to continuous non-censored data

2.1 Choice of candidate distributions

For illustrating the use of various functions of the **fitdistrplus** package with continuous non-censored data, we first use a data set named `groundbeef` which is included in our package. This data set contains pointwise values of serving sizes in grams, collected in a French survey, for ground beef patties consumed by children under 5 years old. It was used in a quantitative risk assessment published in the international journal of food microbiology journal [14].

```
> data(groundbeef)
> str(groundbeef)

'data.frame':        254 obs. of  1 variable:
 $ serving: num  30 10 20 24 20 24 40 20 50 30 ...
```

Before fitting one or more distributions to a data set, it is generally necessary to choose good candidates among a predefined family of distributions. This choice may be guided by the knowledge of stochastic processes governing the modelled variable, but also by the observation of its empirical distribution. To help the user in this choice, we developed functions to plot and characterise the empirical distribution.

First of all, the empirical distribution function and the histogram may be plotted using the classical R functions `ecdf` and `hist`, or by using the `plotdist` function of the **fitdistrplus** package. This function provides two plots (see Figure 1): the left-hand graph is the histogram (on a density level) and the right-hand graph plots the empirical cumulative distribution function (CDF).

```
> plotdist(groundbeef$serving)
```

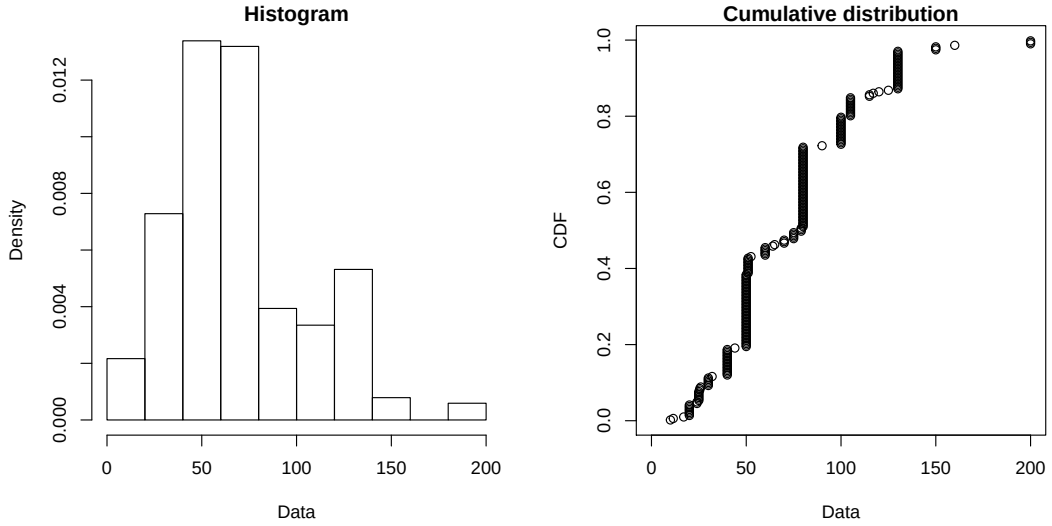

Figure 1: Histogram and CDF plots of an empirical distribution for a continuous variable (serving size from the `groundbeef` data set)

In addition to empirical plots, descriptive statistics may help to choose good candidates to describe a distribution among a family of parametric distributions. Especially the skewness and kurtosis, linked to the third and fourth moments, are useful for this purpose. A non-zero skewness reveals a lack of symmetry of the empirical distribution, while the kurtosis value quantifies the weight of tails in comparison to the normal distribution for which the kurtosis equals 3.

The skewness and kurtosis and their corresponding unbiased estimator from a sample $(X_i)_i \overset{\text{i.i.d.}}{\sim} X$ with observations $(x_i)_i$ are given by

$$sk(X) = \frac{E[(X - E(X))^3]}{Var(X)^{\frac{3}{2}}} \ , \ \widehat{sk} = \frac{\sqrt{n(n-1)}}{n-2} \times \frac{m_3}{m_2^{\frac{3}{2}}}, \tag{1}$$

$$kr(X) = \frac{E[(X - E(X))^4]}{Var(X)^2} \ , \ \widehat{kr} = \frac{n-1}{(n-2)(n-3)}((n+1) \times \frac{m_4}{m_2^2} - 3(n-1)) + 3, \tag{2}$$

where m_2, m_3, m_4 denote empirical moments defined by $m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^r$, with x_i the n observations of variable x and $\overline{x}$ their mean value.

The `descdist` function provides calculations of classical descriptive statistics (minimum, maximum, median, mean, standard deviation), skewness and kurtosis. By default, unbiased estimations of the three last statistics are provided. But, the argument `method` may be used to obtain them without correction for bias. A skewness-kurtosis plot such as the one proposed by [12] is provided by the function `descdist` for the empirical distribution (see Figure 2 for the `groundbeef` data set). On this plot, values for common distributions are displayed in order to help the choice of distributions to fit to data. For some distributions (normal, uniform, logistic, exponential), there is only one possible value for the skewness and the kurtosis. Thus, the distribution is represented by a single point on the plot. For other distributions, areas of possible values are represented, consisting in lines (as for gamma and lognormal distributions), or larger areas (as for beta distribution).

Skewness and kurtosis are known not to be robust. In order to take into account the uncertainty of the estimated values of kurtosis and skewness from data, a bootstrap procedure can be performed by fixing the argument `boot` to an integer above 10. Bootstrap samples of the same size of the original data set are then constructed by random sampling with replacement from that original data set. Values of skewness and kurtosis are computed on that bootstrap samples and reported on the skewness-kurtosis plot. Below is a call to the `descdist` function to describe the distribution of the serving size from the `groundbeef` data set and to draw the corresponding skewness-kurtosis plot (see Figure 2). Looking at the results on this example with a positive skewness and a kurtosis not far from 3, the fit of three common right-skewed distributions could be considered, Weibull, gamma and lognormal distributions.

```
> descdist(groundbeef$serving, boot=1000)


summary statistics
------
min:  10   max:  200
median:  79
mean:  73.65
estimated sd:  35.88
estimated skewness:  0.7353
estimated kurtosis:  3.551
```

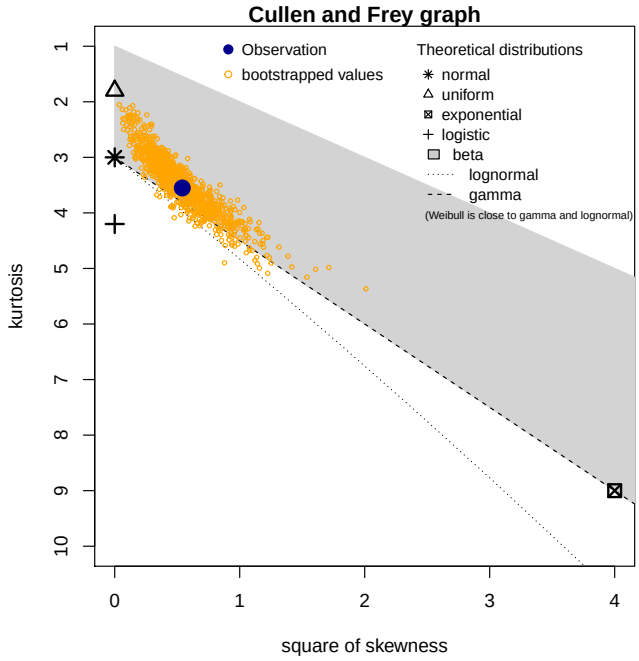

Figure 2: Skewness-kurtosis plot for a continuous variable (serving size from the `groundbeef` data set)

Together with the use of the `plotdist` and `descdist` functions to characterize the empirical distribution, the properties of the modeled variable should be considered, especially its range. Having chosen good candidates for the distribution, we turn to the fit of distributions and comparison of goodness-of-fits.

2.2 Fit of distributions by maximum likelihood estimation

Once selected, one or more parametric distributions $f(.|\theta)$ (with parameter $\theta \in \mathbb{R}^d$) may be fitted to the data set, one at a time, using the `fitdist` function. Under the i.i.d. sample assumption, distribution parameters θ are by default

estimated by maximizing the likelihood defined as:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{3}$$

with x_i the n observations of variable X and $f(.|\theta)$ the density function of the parametric distribution. The other proposed estimation methods are described in Section 3.1.

The `fitdist` function returns the results of the fit of any parametric distribution to a data set as an S3 class object that may be easily printed, summarized or plotted. In order to be used in `fitdist`, a distribution `dist` must be defined by the `d`, `p`, `q` functions, standing respectively for the density, the cumulative distribution and the quantile functions, e.g. `dnorm`, `pnorm` and `qnorm` for the normal distribution. The name of the fitted distribution is specified in the first argument by its classical abbreviation `dist` used in the `d`, `p`, `q` functions, e.g. `"norm"` for the normal distribution. Numerical results returned by the `fitdist` function are (1) the parameter estimates, (2) the estimated standard errors (computed from the estimate of the Hessian matrix at the maximum likelihood solution), (3) the loglikelihood, (4) Akaike and Schwarz information criteria (the so-called AIC and BIC), and (5) the correlation matrix between parameter estimates. Below is a call to the `fitdist` function to fit a Weibull distribution to the serving size from the `groundbeef` data set.

```
> fw <- fitdist(groundbeef$serving, "weibull")
> summary(fw)

Fitting of the distribution ' weibull ' by maximum likelihood
Parameters :
      estimate Std. Error
shape    2.186     0.1046
scale   83.348     2.5269
Loglikelihood:  -1255   AIC:  2514   BIC:  2522
Correlation matrix:
       shape  scale
shape 1.0000 0.3218
scale 0.3218 1.0000
```

The plot of an object of class `"fitdist"` provides four classical goodness-of-fit plots [12]:

- a density plot representing the density function of the fitted distribution along with the histogram of the empirical distribution,
- a CDF plot of both the empirical distribution and the fitted distribution,
- a Q-Q plot representing the empirical quantiles (y-axis) against the theoretical quantiles (x-axis)
- a P-P plot representing the empirical distribution function evaluated at each data point (y-axis) against the fitted distribution function (x-axis).

Unlike `plot.fitdist`, the `denscomp`, `cdfcomp`, `qqcomp` and `ppcomp` functions enable to separately plot each of these four plots, in order to compare the empirical and multiple theoretical distributions fitted on a same data set. These functions must be called with a first argument corresponding to a list of objects of class `fitdist`, and optionaly further arguments to customize the plot (see the reference manual [15] for lists of arguments that may be changed for each plot). In the following example, we compare the fit of a Weibull, a lognormal and a gamma distributions to the `groundbeef` data set (Figure 3).

```
> fg <- fitdist(groundbeef$serving,"gamma")
> fln <- fitdist(groundbeef$serving,"lnorm")
> par(mfrow=c(2, 2))
> denscomp(list(fw,fln,fg), legendtext=c("Weibull", "lognormal", "gamma"))
> qqcomp(list(fw,fln,fg), legendtext=c("Weibull", "lognormal", "gamma"))
> cdfcomp(list(fw,fln,fg), legendtext=c("Weibull", "lognormal", "gamma"))
> ppcomp(list(fw,fln,fg), legendtext=c("Weibull", "lognormal", "gamma"))
```

For CDF, Q-Q and P-P plots, the probability plotting position is defined by default using the Hazen's rule, with probability points of the empirical distribution defined as `(1:n - 0.5)/n`, as recommended by Blom [6]. This plotting position can be easily changed using the arguments `use.ppoints` and `a.ppoints`. When `use.ppoints = TRUE`, the argument `a.ppoints` is passed to the `ppoints` function from the **stats** package to define the probability points of the empirical distribution as `(1:n - a.ppoints)/(n - 2a.ppoints + 1)`. When `use.ppoints = FALSE`, the probability points are simply defined as `1:n / n`.

The density plot and the CDF plot are the most classical goodness-of-fits plots. The two other plots are complementary and can be very informative in some cases. The Q-Q plot emphasizes the lack-of-fit at the distribution

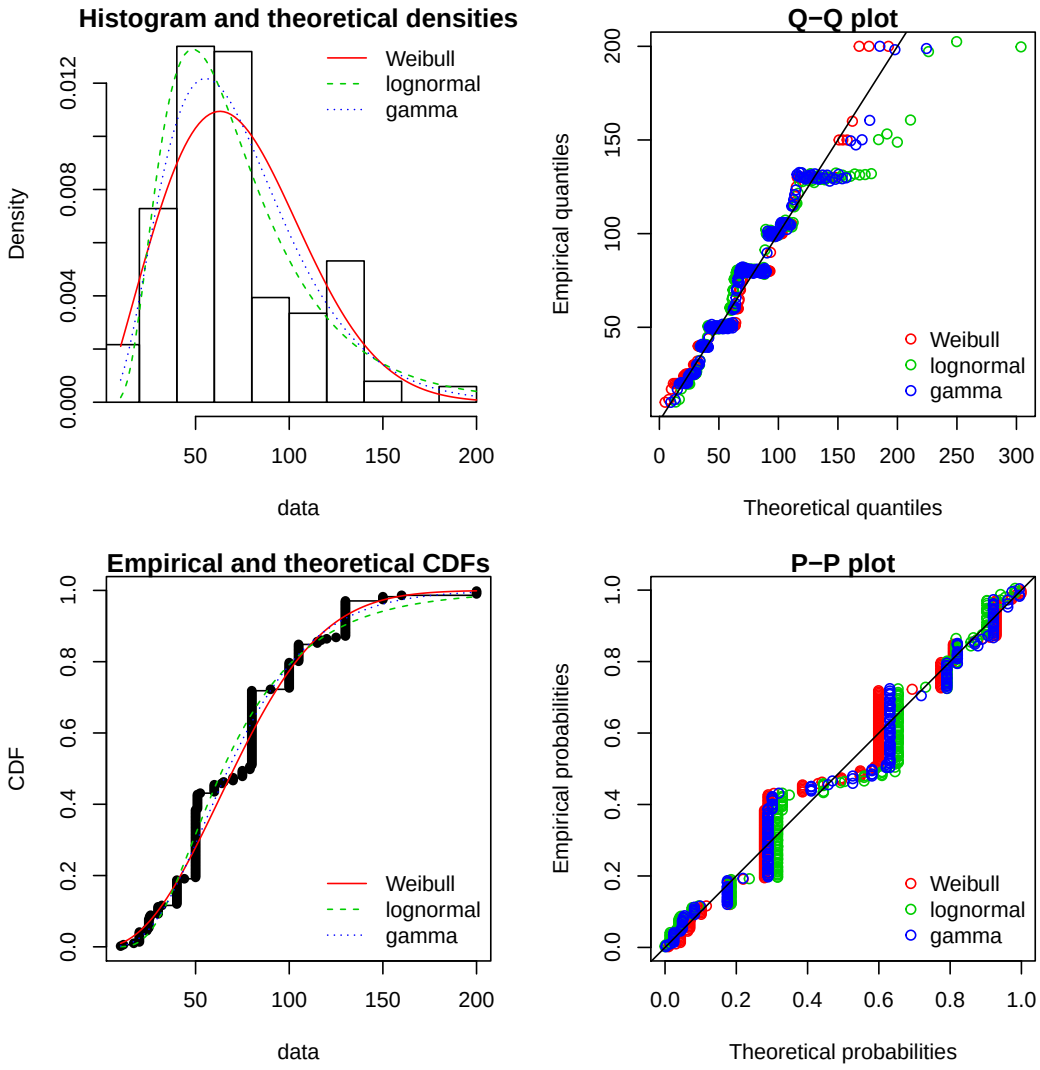

Figure 3: Four Goodness-of-fit plots for various distributions fitted on continuous data (Weibull, gamma and lognormal distributions fitted to serving sizes from the `groundbeef` data set)

tails while the P-P plot emphasizes the lack-of-fit at the distribution center. As an example in Figure 3, none of the three fitted distributions describes the center of the distribution rather better than the two others, but the Weibull and gamma distributions should be prefered for their better description of the right tail of the empirical distribution, especially if the weight of this tail is important in the use of the fitted distribution, as it is in the context of food risk assessment.

To illustrate other features of the **fitdistrplus** package, we will now use another data set named `endosulfan`, which is included in our package. This data set contains acute toxicity values for the organochlorine pesticide endosulfan (geometric mean of LC50 ou EC50 values in $\mu g.L^{-1}$), tested on Australian and non-Australian laboratory-species (arthropods, fish or nonarthropod invertebrates, see [27]).

```
> data(endosulfan)
> str(endosulfan)

'data.frame':        104 obs. of  3 variables:
 $ ATV       : num  0.6 2.8 182.2 0.8 478 ...
 $ Australian: Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 1 ...
 $ group     : Factor w/ 3 levels "Arthropods","Fish",..: 1 1 1 1 1 1 1 1 1 1 ...
```

In ecotoxicology, a lognormal or a loglogistic distribution is often fitted to such a data set in order to characterize the species sensitivity distribution (SSD) for a pollutant. A low percentile of the fitted distribution, generally the 5% percentile, is calculated and named the hazardous concentration 5% (HC5). It is interpreted as the value of the pollutant concentration protecting 95% of the species. But the fit of a lognormal or a loglogistic distribution to the whole `endosulfan` data set is rather bad, especially due to a minority of very high values. We can try to fit this data set by the Pareto distribution or the three-parameter Burr distribution which is an extension of both the loglogistic and the Pareto distribution. Pareto and Burr distributions are provided in package **actuar**. Until here, we did not have to implicity define starting values (in the optimization process) as reasonable starting values are defined within

function `fitdist` for most of the distributions defined in R packages (see the help of `fitdist` for details). For other distributions like the Pareto and the Burr distribution, we have to supply initial values for the distribution parameters in the argument `start` when using the maximum likelihood method. `start` must be a named list with initial values for each parameter (as they appear in the `d`, `p`, `q` functions). Having defined reasonable starting values[1], we can fit various distributions and graphically compare their fits. On this example, the use of functions `cdfcomp` and `qqcomp` is especially interesting to evaluate the goodness-of-fit on the tail of interest while defining an $HC5$ value.

```
> ATV <-endosulfan$ATV
> fendo.ln <- fitdist(ATV, "lnorm")
> library(actuar)
> fendo.ll <- fitdist(ATV, "llogis", start=list(shape=1, scale=500))
> fendo.P <- fitdist(ATV, "pareto", start=list(shape=1, scale=500))
> fendo.B <- fitdist(ATV, "burr", start=list(shape1=0.3, shape2=1, rate=1))
> par(mfrow=c(1, 2))
> cdfcomp(list(fendo.ln, fendo.ll, fendo.P, fendo.B), xlogscale=TRUE,
+           legendtext = c("lognormal","loglogistic","Pareto","Burr"))
> qqcomp(list(fendo.ln, fendo.ll, fendo.P, fendo.B), xlogscale=TRUE, ylogscale=TRUE,
+        legendtext = c("lognormal","loglogistic","Pareto","Burr"))
```

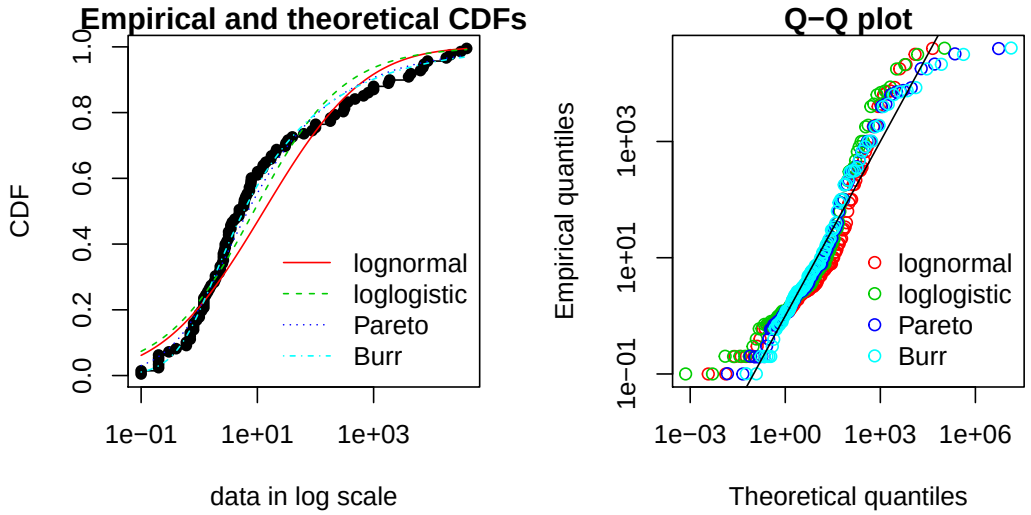

Figure 4: CDF and Q-Q plots to compare the fit of four distributions to acute toxicity values of various organisms for the organochlorine pesticide endosulfan (`endosulfan` data set)

We can see in Figure 4 that none of the fitted distribution correctly describes the right tail observed in the data set. But the left tail seems to be better described by the Burr distribution. Its use could then be considered to estimate the $HC5$ value as the 5% quantile of the distribution. This can be easily done using the `quantile` generic function defined for an object of class `"fitdist"`. Below is this calculation together with the calculation of the empirical quantile for comparison.

```
> quantile(fendo.B, probs = 0.05)

Estimated quantiles for each specified probability (non-censored data)
         p=0.05
estimate 0.2939

> quantile(ATV, probs = 0.05)

 5%
0.2
```

To go further in the comparison of various distributions, we propose the calculation of different goodness-of-fit statistics in our package. The purpose of goodness-of-fit statistics aims to measure the distance between the fitted parametric distribution and the empirical distribution: e.g. the distance between the fitted cumulative distribution function F and the empirical distribution function F_n. When fitting continuous distributions, three goodness-of-fit statistics are classicaly considered: Cramer-von Mises, Kolmogorov-Smirnov and Anderson-Darling statistics. Naming x_i the n observations of a continuous variable X arranged in an ascending order, Table 1 gives the definition and the empirical estimate of the three considered goodness-of-fit statistics.

[1]The `plotdist` function can plot any parametric distribution with specified parameter values in argument `para`. It can thus help to find correct initial values for the distribution parameters in non trivial cases, by iterative calls if necessary (see the reference manual [15] for examples).

Table 1: Goodness-of-fit statistics as defined by Stephens [13].

Statistic	General formula	Computational formula
Kolmogorov-Smirnov (KS)	$\sup \lvert F_n(x) - F(x) \rvert$	$\max(D^+, D^-)$ with $D^+ = \max_{i=1,\ldots,n} \left(\frac{i}{n} - F(x_i)\right); D^- = \max_{i=1,\ldots,n} \left(F(x_i) - \frac{i-1}{n}\right)$
Cramer-von Mises (CvM)	$n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dx$	$\frac{1}{12n} + \sum_{i=1}^{n} \left(F(x_i) - \frac{2i-1}{2n}\right)^2$
Anderson-Darling (AD)	$n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1-F(x))} dx$	$-n - \frac{1}{n} \sum_{i=1}^{n} ((2i-1)(\log(F(x_i)) + \log(1 - F(x_{n+1-i}))))$

They can be computed using the function `gofstat` as defined by Stephens [13].

```
> gofstat(list(fendo.ln, fendo.ll, fendo.P, fendo.B))

Goodness-of-fit statistics
                             1-mle-lnorm 2-mle-llogis 3-mle-pareto 4-mle-burr
Kolmogorov-Smirnov statistic      0.1672       0.1196      0.08488    0.06155
Cramer-von Mises statistic        0.6374       0.3827      0.13926    0.06803
Anderson-Darling statistic        3.4721       2.8316      0.89206    0.52393

Goodness-of-fit criteria
                               1-mle-lnorm 2-mle-llogis 3-mle-pareto 4-mle-burr
Aikake's Information Criterion        1069         1069         1048       1046
Bayesian Information Criterion        1074         1075         1053       1054
```

As giving more weight to distribution tails, Anderson-Darling statistics is of special interest when it matters to equally emphasize the tails as well as the main body of a distribution. This is often the case in risk assessment, e.g. [12, 63]. For this reason, this statistics is often used to select the best distribution among those fitted. Nevertheless, this statistics should be used cautiously when comparing fits of various distributions. We must keep in mind that the weighting of each CDF quadratic difference depends on the theoretical distribution in its definition, see Table 1. Anderson-Darling statistics computed for several distributions fitted on a same data set are thus theoretically difficult to compare. Moreover, such a statistic, as Cramer-von Mises and Kolmogorov-Smirnov ones, does not take into account the complexity of the model. It is not a problem when compared distributions are characterized by the same number of parameters, but it could systematically promote the selection of the more complex distributions in the other case. Looking at classical penalized criteria based on the loglikehood seems thus also interesting, especially to discourage overfitting.

In the previous example, all the goodness-of-fit statistics based on the cdf distance encourage the choice of the Burr distribution, the only one characterized by three parameters, while Akaike and Schwarz information criteria (so called AIC and BIC) respectively gives the preference to the Burr distribution or the Pareto distribution. The choice between these two distributions seems thus less obvious and could be discussed. Even if specifically recommended for discrete distributions, the Chi-squared statistic may also be used for continuous distributions (see Section 3.3 and the reference manual [15] for examples).

2.3 Uncertainty in parameter estimates

The uncertainty in the parameters of the fitted distribution may be simulated by parametric or nonparametric bootstraps using the `boodist` function for non-censored data. This function returns the bootstrapped values of parameters in a S3 class object which may be plotted to visualize the bootstrap region. The medians and the 95 percent confidence intervals of parameters (2.5 and 97.5 percentiles) are printed in the summary. If inferior to the whole number of iterations, the number of iterations for which the function converges is also printed in the summary.

The plot of an object of class `"bootdist"` consists in a scatterplot or a matrix of scatterplots of the bootstrapped values of parameters providing a representation of the joint uncertainty distribution of the fitted parameters (see Figure 5). Below is an example of the use of the `bootdist` function with the previous fit of the Burr distribution to the `endosulfan` data set.

```
> bendo.B <- bootdist(fendo.B, niter=1001)

> summary(bendo.B)

Parametric bootstrap medians and 95% percentile CI
       Median    2.5%  97.5%
shape1 0.1998 0.09871 0.3652
```

```
shape2 1.5847 1.03291 2.9930
rate   1.4955 0.67666 2.8278

The estimation method converged only for  1000  among  1001  iterations
```

```
> plot(bendo.B)
```

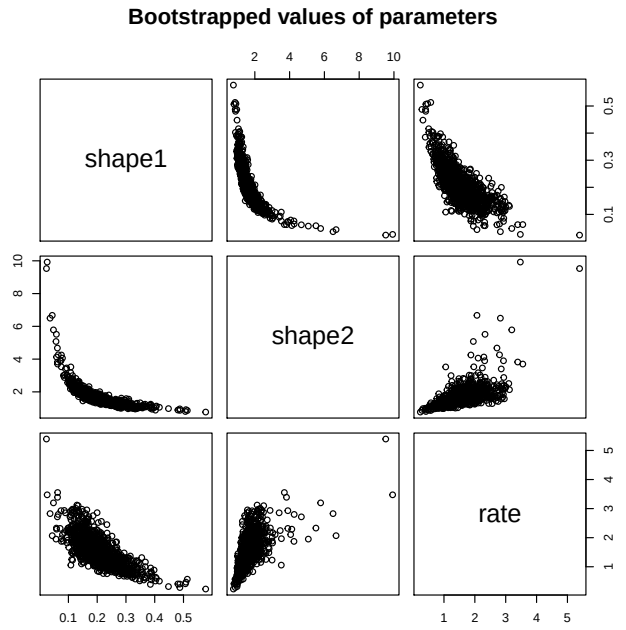

Figure 5: Bootstrappped values of parameters for a fit of the Burr distribution characterized by three parameters (example on the `endosulfan` data set)

Bootstrap samples of parameter estimates are useful especially to calculate confidence intervals on each parameter of the fitted distribution from the marginal distribution of the bootstraped values. It is also interesting to look at the joint distribution of the bootstraped values in a scatterplot (or a matrix of scatterplots if the number of parameters exceeds two) in order to understand the potential structural correlation between parameters (see Figure 5).

The use of the whole bootstrap sample is also of interest in the risk assessment field. Its use enables the characterization of uncertainty in distribution parameters. It can be directly used within a second-order Monte Carlo simulation framework, especially within the package **mc2d** ([49]). One could refer to Pouillot *et al.* ([46]) for an introduction to the use of **mc2d** and **fitdistrplus** packages in the context of quantitative risk assessment.

Bootstrap can also be used to calculate confidence intervals on quantiles of the fitted distribution. For this purpose, a generic `quantile` function is provided for class `bootdist`. By default 95% bootstrap confidence intervals of quantiles are provided. Going back to the previous example from ecotoxicolgy, this function can be used to estimate the uncertainty associated to the HC5 estimation, for example from the previously fitted Burr distribution to the `endosulfan` data set.

```
> quantile(bendo.B, probs = 0.05)

(original) estimated quantiles for each specified probability (non-censored data)
         p=0.05
estimate 0.2939
Median of bootstrap estimates
         p=0.05
estimate 0.3008

two-sided 95 % CI of each quantile
       p=0.05
2.5 %  0.1741
97.5 % 0.5035

The estimation method converged only for  1000  among  1001  bootstrap iterations.
```

3 Advanced topics

3.1 Alternative methods for parameter estimation

Despite maximum likelihood estimation is the default estimation proposed by `fitdist`, other classical estimation methods can be handled to estimate parameters for non-censored data. Thus, this subsection focuses on alternative estimation methods.

One of the alternative for continuous distributions is the maximum goodness-of-fit estimation method also called minimum distance estimation method. In this package this method is proposed with eight different distances: the three classical distances defined in Table 1, or one of the variants of the Anderson-Darling distance proposed by [38] and defined in Table 2. The right-tail AD gives more weight only to the right tail, the left-tail AD gives more weight only to the left tail. Either of the tails, or both of them, can receive even larger weights by using second order Anderson-Darling Statistics.

Table 2: Modified Anderson-Darling statistics as defined by Luceno [38].

Statistic	General formula	Computational formula
Right-tail AD (ADR)	$\int_{-\infty}^{\infty} \frac{(F_n(x)-F(x))^2}{1-F(x)} dx$	$\frac{n}{2} - 2\sum_i F(x_i) - \frac{1}{n}\sum_i((2i-1)ln(1-F(x_{n+1-i})))$
Left-tail AD (ADL)	$\int_{-\infty}^{\infty} \frac{(F_n(x)-F(x))^2}{(F(x))} dx$	$-\frac{3n}{2} + 2\sum_i F(x_i) - \frac{1}{n}\sum_i((2i-1)ln(F(x_i)))$
Right-tail AD 2nd order (AD2R)	$ad2r = \int_{-\infty}^{\infty} \frac{(F_n(x)-F(x))^2}{(1-F(x))^2} dx$	$ad2r = 2\sum_i ln(1-F(x_i)) + \frac{1}{n}\sum_i \frac{2i-1}{1-F(x_{n+1-i})}$
Left-tail AD 2nd order (AD2L)	$ad2l = \int_{-\infty}^{\infty} \frac{(F_n(x)-F(x))^2}{(F(x))^2} dx$	$ad2l = 2\sum_i ln(F(x_i)) + \frac{1}{n}\sum_i \frac{2i-1}{F(x_i)}$
AD 2nd order (AD2)	$ad2r + ad2l$	$ad2r + ad2l$

To fit a distribution by maximum goodness-of-fit estimation, one needs to fix the argument `method` to `"mge"` in the call to `fitdist` and to specify the argument `gof` coding for the chosen goodness-of-fit distance. This function is intended to be used only with continuous non-censored data.

Maximum goodness-of-fit estimation may be useful to give more weight to data at one tail of the distribution. Let us go back to the previous example from ecotoxicology. Instead of trying to find a less classical distribution to correctly fit the empirical distribution especially on its left tail, one could consider the fit of the classical lognormal distribution, but minimizing a goodness-of-fit distance giving more weight to the left tail of the empirical distribution, so as to correctly estimate the 5% percentile. In the following example of `endosulfan` data set, we use left tail Anderson-Darling distances of first or second order (see Figure 6).

```
> fendo.ln.ADL <- fitdist(ATV,"lnorm",method="mge",gof="ADL")
> fendo.ln.AD2L <- fitdist(ATV,"lnorm",method="mge",gof="AD2L")
> cdfcomp(list(fendo.ln, fendo.ln.ADL, fendo.ln.AD2L),
+ xlogscale = TRUE, main = "",
+ legendtext = c("MLE",
+ "Left-tail AD", "Left-tail AD 2nd order"),cex=0.7,
+ xlegend = "bottomright")
```

Comparing the 5% percentiles (HC5) calculated using these three fits to the one calculated from the MLE fit of the Burr distribution, we can observe, on this example, that fitting the lognormal distribution by maximizing left tail Anderson-Darling distances of first or second order enables to approach the value obtained by fitting the Burr distribution by MLE.

```
> ( HC5.estimates <- cbind(empirical = as.numeric(quantile(ATV, probs = 0.05)),
+       Burr = as.numeric(quantile(fendo.B,probs = 0.05)$quantiles),
+       lognormal_MLE = as.numeric(quantile(fendo.ln,probs = 0.05)$quantiles),
+       lognormal_AD2 = as.numeric(quantile(fendo.ln.ADL,probs = 0.05)$quantiles),
+       lognormal_AD2L = as.numeric(quantile(fendo.ln.AD2L,probs = 0.05)$quantiles)) )

     empirical   Burr lognormal_MLE lognormal_AD2 lognormal_AD2L
[1,]       0.2 0.2939       0.07259        0.1959         0.2588
```

Another method commonly used to fit parametric distribution is the moment matching estimation (MME). It consists in finding the value of the parameter θ that equalizes the first theoretical raw moments of the parametric

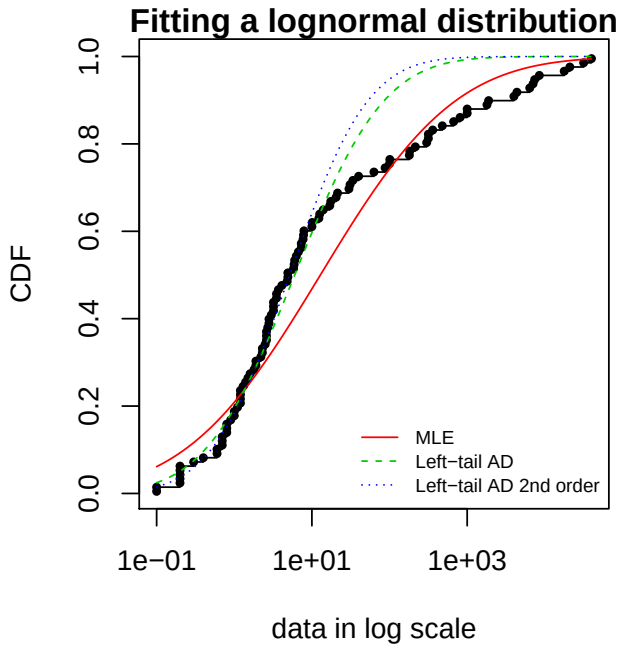

Figure 6: Comparison of a lognormal distribution fitted by MLE and by MGE using two different goodness-of-fit distances : left-tail Anderson-Darling and left-tail Anderson Darling of seond order (example on the `endosulfan` data set)

distribution to the empirical moments (Equation 4)

$$E(X^k|\theta) = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \tag{4}$$

for $k = 1, \ldots, d$, with d the number of parameters to estimate and x_i the n observations of variable X. For moments of order greater or equal than 2, it may also be relevant to match centered moments as given by Equation (5).

$$E\left((X - E(X))^k|\theta\right) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^k \tag{5}$$

This method can be performed by setting the argument `method` to `"mme"` in the call to fitdist. The estimate is computed by a closed-form formula for the following distributions: normal, lognormal, exponential, Poisson, gamma, logistic, negative binomial, geometric, beta and uniform distributions (i.e. base R distributions). In this case, for distributions characterized by one parameter (geometric, Poisson and exponential), this parameter is simply estimated by matching theoretical and observed means, and for distributions characterized by two parameters, these parameters are estimated by matching theoretical and observed means and variances (see e.g. [63]). Otherwise, for not so-common distributions, the equation of moments is solved numerically using the `optim` function by minimizing the sum of squared differences between observed and theoretical moments (see the **fitdistrplus** reference manual [15] for technical details).

To illustrate this method, we use a classical data set from the Danish insurance industry published in [41]. In **fitdistrplus**, the data set is stored in `danishuni` for the univariate version and contains the losse amounts collected at Copenhagen Reinsurance between 1980 and 1990. In actuarial science, it is standard to consider positive heavy-tailed distribution and have a special focus on the right-tail of the distribution. In this numerical experiment, we consider first the lognormal distribution and then the Pareto type II distribution, see e.g. [33].

```
> data(danishuni)
> str(danishuni)

'data.frame':        2167 obs. of  2 variables:
 $ Date: Date, format: "1980-01-03" "1980-01-04" ...
 $ Loss: num  1.68 2.09 1.73 1.78 4.61 ...
```

We can first fit a lognormal distribution on `danishuni` data set by matching moments using a closed-form formula. On the left-hand graph of Figure 7, we compare the fitted distribution function between MME and MLE method. We observe that the moment matching estimation provides a more cautious of the insurance risk as the MME-fitted distribution function (resp. MLE-fitted) underestimates (overestimates) the empirical distribution function for large values of claim amounts.

```
> fdanish.ln.MLE <- fitdist(danishuni$Loss, "lnorm")
> fdanish.ln.MME <- fitdist(danishuni$Loss, "lnorm", method="mme", order=1:2)
> cdfcomp(list(fdanish.ln.MLE, fdanish.ln.MME),
+         legend=c("lognormal MLE", "lognormal MME"), main="Fitting a lognormal distribution",
+         xlogscale=TRUE, datapch=".")
```

In a second time, we choose to fit a Pareto type-II distribution, which gives more weight to the right-tail of the distribution. As the lognormal distribution, the Pareto type-II has two parameters, which allows a fair comparison. The Burr distribution (with its three parameters) would lead to a better fit.

We use the implementation of the **actuar** package providing raw and centered moments for that distribution (in addition to `d`, `p`, `q` and `r` functions, see [23]). Fitting a heavy-tailed distribution for which the first and the second moments do not exist for certain values of the shape parameter requires some cautiousness. This is carried out by providing for the optimization process a lower and an upper bound for each parameter. Our call below immediately calls the L-BFGS-B optimization method in `optim`, since this quasi-Newton allows box constraints[2]. Note that we have to pass a function (called `memp` in our example) for computing the empirical raw moment to `fitdist`, since the user may choose either Equation (4) or (5).

```
> library(actuar)
> fdanish.P.MLE <- fitdist(danishuni$Loss, "pareto",start=c(shape=10, scale=10),
+       lower=2+1e-6, upper=Inf)
> memp <- function(x, order) ifelse(order == 1, mean(x), sum(x^order)/length(x))
> fdanish.P.MME <- fitdist(danishuni$Loss, "pareto", method="mme", order=1:2,
+       memp="memp", start=c(shape=10, scale=10), lower=c(2+1e-6,2+1e-6), upper=c(Inf,Inf))
> cdfcomp(list(fdanish.P.MLE, fdanish.P.MME),
+         legend=c("Pareto MLE", "Pareto MME"), main="Fitting a Pareto distribution",
+         xlogscale=TRUE, datapch=".")
> gofstat(list(fdanish.ln.MLE, fdanish.P.MLE, fdanish.ln.MME, fdanish.P.MME))

Goodness-of-fit statistics
                          1-mle-lnorm 2-mle-pareto 3-mme-lnorm 4-mme-pareto
Kolmogorov-Smirnov statistic      0.1375       0.3124      0.4368       0.3638
Cramer-von Mises statistic       14.7911      37.7166     88.9503      53.0783
Anderson-Darling statistic       87.1933     208.3139    416.2567     272.4729

Goodness-of-fit criteria
                               1-mle-lnorm 2-mle-pareto 3-mme-lnorm
Aikake's Information Criterion        8120         9250        9792
Bayesian Information Criterion        8131         9261        9803
                               4-mme-pareto
Aikake's Information Criterion         9395
Bayesian Information Criterion         9407
```

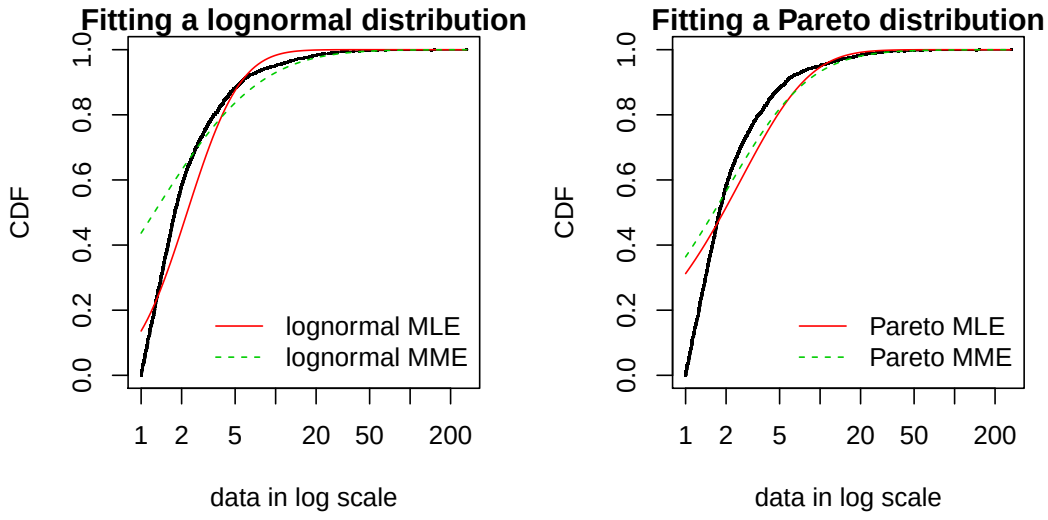

Figure 7: Comparison between MME and MLE when fitting a lognormal or a Pareto distribution to loss data from the `danishuni` data set

[2]That is what the B stands for.

We can see on Figure 7 that matching moments and maximum likelihood fits are far less distant (when looking at the right-tail) for the Pareto distribution than for the lognormal distribution on this data set. Furthermore for these two distributions, the MME method better fits the right-tail of the distribution, which seems logical since empirical moments are influenced by large observed values. In the previous traces, we give the values of goodness-of-fit statistics. Whatever the statistic considered, the MLE-fitted lognormal always provides the best fit to the observed data.

Maximum likelihood and moment matching (the default on many softwares) estimations are certainly the most commonly used method for fitting distributions. But one should keep in mind that these two methods may give very different results. When choosing the matching moment method, the user should be aware of its great sensitivity to outliers. This may be seen as an advantage in our example if the objective is to better describe the right tail of the distribution, but it may be seen as a drawback if the objective is different.

Fitting of a parametric distribution may also be done by matching theoretical quantiles of the parametric distributions (for specified probabilities) against the empirical quantiles. Equation (6) below is thus very similar to Equations (4) and (5)

$$F^{-1}(p_k|\theta) = Q_{n,p_k} \tag{6}$$

for $k = 1, \ldots, d$, with d the number of parameters to estimate (dimension of θ if there is no fixed parameters) and Q_{n,p_k} the empirical quantiles calculated from data for specified probabilities p_k.

Quantile matching estimation is performed by setting the argument `method` to `"qme"` in the call to `fitdist` and adding an argument `probs` defining the probabilities for which the quantile matching is performed. The length of this vector must be equal to the number of parameters to estimate (as the vector of moment orders for MME). Empirical quantiles are computed using the `quantile` function of the **stats** package using the `type` argument equal to 7 by default, but the type of quantile can be easily changed by using the `qty` argument in the call to the `qme` function. The quantile matching is carried out numerically, by minimizing the sum of squared differences between observed and theoretical quantiles.

```
> fdanish.ln.QME1 <- fitdist(danishuni$Loss, "lnorm", method="qme", probs=c(1/3, 2/3))
> fdanish.ln.QME2 <- fitdist(danishuni$Loss, "lnorm", method="qme", probs=c(8/10, 9/10))
> cdfcomp(list(fdanish.ln.MLE, fdanish.ln.QME1, fdanish.ln.QME2),
+         legend=c("MLE", "QME(1/3, 2/3)", "QME(8/10, 9/10)"),
+         main="Fitting a lognormal distribution", xlogscale=TRUE, datapch=".")
```

Above is an example of fitting of a lognormal distribution to `danishuni` data set by matching probabilities ($p_1 = 1/3, p_2 = 2/3$) and ($p_1 = 8/10, p_2 = 9/10$). As expected, the second QME fit gives more weight to the right-tail of the distribution, despite we do not choose the Pareto type-II distribution. Compared to the maximum likelihood estimation, the second QME fit is also more conservative, whereas the first QME fit is less conservative. The quantile matching estimation is of particular interest when we need a focus around particular quantiles, e.g. $p = 99.5\%$ in the Solvency II insurance context or $p = 5\%$ for the HC5 estimation in the ecotoxicology context.

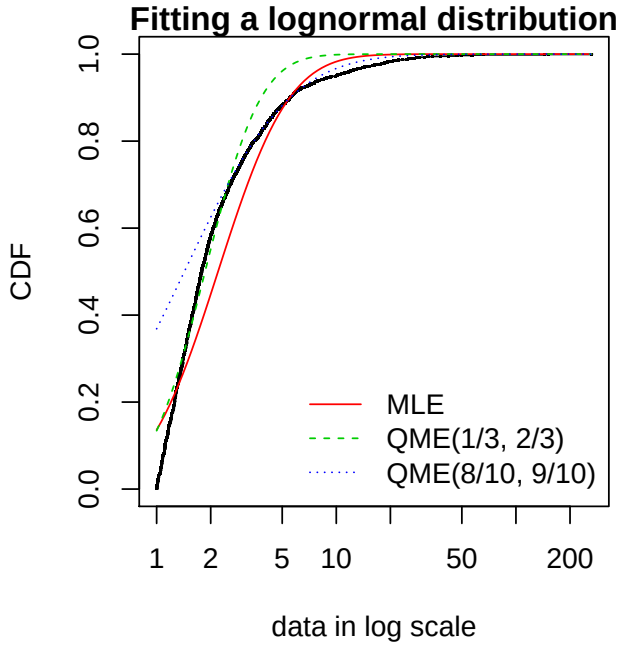

Figure 8: Comparison between QME and MLE when fitting a lognormal distribution to loss data from the `danishuni` data set

3.2 Customization of the optimization algorithm

Each time a numerical minimization (or maximization) is carried out using `fitdist`, the `optim` function of the **stats** package is used by default with the `"Nelder-Mead"` method for distributions characterized by more than one parameter and the `"BFGS"` method for distributions characterized by only one parameter. Sometimes the default algorithm fails to converge. It is then interesting to change some options of the `optim` function or to use another optimization function than `optim` to maximize the likelihood or to minimize a squared difference. The argument `optim.method` can be used in the call to `fitdist` or `fitdistcens`. It will internally be passed to `mledist` and to `optim` (see the help page of `optim` from the package **stats** for details about the different algorithms available proposed by `optim`).

Below are examples of fits of a gamma distribution $\mathcal{G}(\alpha, \lambda)$ to the `groundbeef` data set with various algorithms. Note that the conjugate gradient algorithm (`"CG"`) needs far more iterations to converge (around 2500 iterations) compared to other algorithms (converging in less than 100 iterations).

```
> data(groundbeef)
> fNM <- fitdist(groundbeef$serving, "gamma", optim.method="Nelder-Mead")
> fBFGS <- fitdist(groundbeef$serving, "gamma", optim.method="BFGS")
> fSANN <- fitdist(groundbeef$serving, "gamma", optim.method="SANN")
> fCG <- try(fitdist(groundbeef$serving, "gamma", optim.method="CG", control=list(maxit=10000)))
> if(class(fCG) == "try-error")
+   fCG <- list(estimate=NA)
```

It is also possible to use another function than `optim` to maximize the likelihood. This optimization function must be specified by the argument `custom.optim` in the call to `fitdist`. But before that, it may be necessary to customize this optimization function to meet the following requirements. (1) `custom.optim` function must have (at least) the following arguments: `fn` for the function to be optimized and `par` for the initialized parameters. (2) `custom.optim` should carry out a MINIMIZATION and must return (at least) the following components: `par` for the estimate, `convergence` for the convergence code, `value` for `fn(par)` and `hessian`. Below is an example of code written to wrap the `genoud` function from the **rgenoud** package in order to respect our optimization "template". The **rgenoud** package implements the genetic (stochastic) algorithm.

```
> mygenoud <- function(fn, par, ...)
+ {
+    require(rgenoud)
+    res <- genoud(fn, starting.values=par, ...)
+    standardres <- c(res, convergence=0)
+    return(standardres)
+ }
```

The customized optimization function can then be passed as the argument `custom.optim` in the call to `fitdist` or `fitdistcens`. The following code can for example be used to fit a gamma distribution to the `groundbeef` data set. Note that in this example various arguments are also passed from `fitdist` to `genoud` : `nvars`, `Domains`, `boundary.enforcement`, `print.level` and `hessian`. The code below compares all the parameter estimates ($\hat{\alpha}$, $\hat{\lambda}$) by the different algorithms: shape α and rate λ parameters are relatively similar on this example, roughly 4.00 and 0.05, respectively.

```
> fgenoud <- mledist(groundbeef$serving, "gamma", custom.optim= mygenoud, nvars=2,
+     max.generations=10, Domains=cbind(c(0,0), c(10,10)), boundary.enforcement=1,
+     hessian=TRUE, print.level=0, P9=10)
> cbind(NM = fNM$estimate,
+ BFGS = fBFGS$estimate,
+ SANN = fSANN$estimate,
+ CG = fCG$estimate,
+ fgenoud = fgenoud$estimate)

           NM    BFGS    SANN      CG fgenoud
shape 4.00825 4.22848 4.03305 4.12891 4.00834
rate  0.05442 0.05742 0.05457 0.05607 0.05443
```

3.3 Fitting distributions to other types of data

Analytical methods often lead to semi-quantitative results which are referred to as censored data. Observations only known to be under a limit of detection are called left censored data. Observations only known to be above a limit of quantification are called right censored data. Results known to lie between two bounds are called interval censored data. These two bounds may correspond to a limit of detection and a limit of quantification, or more generally to uncertainty bounds around the observation. Right censored data are also commonly encountered among survival data. A data set may thus contain right, left, or interval censored data, or may be a mixture of these categories, possibly with different upper and lower bounds. Censored data are sometimes excluded from the data analysis or replaced by

a fixed value, which in both cases may lead to biased results. A more recommended approach to correctly model such data is based upon maximum likelihood [32, 24]. We chose this approach in our package.

Censored data can thus contain left censored, right censored and interval censored values, with several lower and upper bounds. Before using the **fitdistrplus** package, such data must be coded into a dataframe with two columns, respectively named `left` and `right`, describing each observed value as an interval. The `left` column contains either `NA` for left censored observations, the left bound of the interval for interval censored observations, or the observed value for non-censored observations. The `right` column contains either `NA` for right censored observations, the right bound of the interval for interval censored observations, or the observed value for non-censored observations. To illustrate the use of package **fitdistrplus** to fit distributions to censored continous data, we use another data set from ecotoxicology, included in our package and named `salinity`. This data set contains acute salinity tolerance (LC50 values in electrical conductivity, $mS.cm^{-1}$) of riverine macro-invertebrates taxa from the southern Murray-Darling Basin in Central Victoria, Australia (see [31]).

```
> data(salinity)
> str(salinity)
```

```
'data.frame':        108 obs. of  2 variables:
 $ left : num  20 20 20 20 20 21.5 15 20 23.7 25 ...
 $ right: num  NA NA NA NA NA 21.5 30 25 23.7 NA ...
```

Using censored data such as those coded in the `salinity` data set, the empirical distribution can be plotted using the `plotdistcens` function. By default, this function uses the Expectation-Maximization approach of Turnbull [61] to compute the overall empirical cdf curve with optional confidence intervals, by calls to `survfit` and `plot.survfit` functions from the **survival** package (Figure 10 shows the Turnbull plot of data together with two fitted distributions). A less rigorous but sometimes more illustrative plot to see the real nature of censored data can be obtained by fixing the argument `Turnbull` to `FALSE` in the call to `plotdistcens` (see Figure 9 for an example and the help page of Function `plotdistcens` for details).

```
> plotdistcens(salinity,Turnbull = FALSE)
```

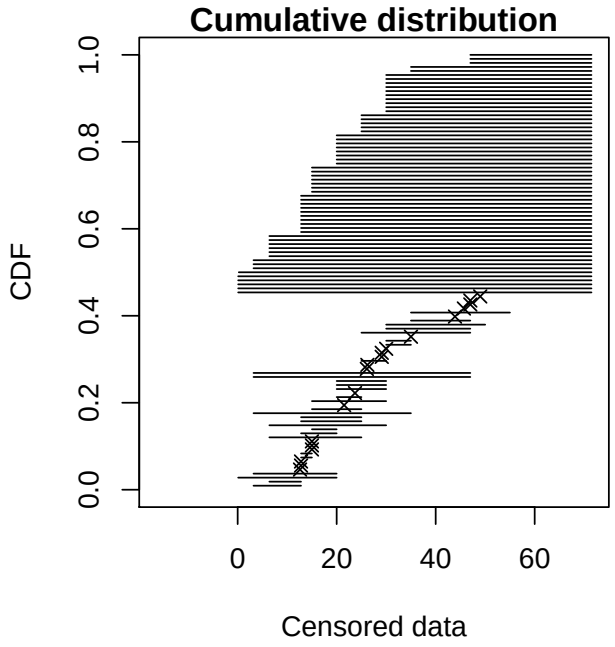

Figure 9: Simple plot of censored data (72-hour acute salinity tolerance of riverine macro-invertebrates from the `salinity` data set) as ordered points and intervals

As for non censored data, one or more parametric distributions can be fitted to the censored data set, one at a time, but using in this case the `fitdistcens` function. This function estimates the vector of distribution parameters θ by maximizing the likelihood for censored data defined as:

$$L(\theta) = \prod_{i=1}^{N_{nonC}} f(x_i|\theta) \times \prod_{j=1}^{N_{leftC}} F(x_j^{upper}|\theta) \times \prod_{k=1}^{N_{rightC}} (1 - F(x_k^{lower}|\theta)) \times \prod_{m=1}^{N_{intC}} (F(x_m^{upper}|\theta) - F(x_j^{lower}|\theta)) \qquad (7)$$

with x_i the N_{nonC} non-censored observations, x_j^{upper} upper values defining the N_{leftC} left-censored observations, x_k^{lower} lower values defining the N_{rightC} right-censored observations, $[x_m^{lower}; x_m^{upper}]$ the intervals defining the N_{intC} interval-censored observations, and F the cumulative distribution function of the parametric distribution.

As `fitdist`, `fitdistcens` returns the results of the fit of any parametric distribution to a data set as an S3 class object that can be easily printed, summarized or plotted. For the `salinity` data set, a lognormal distribution or a loglogistic can be fitted as commonly done in ecotoxicology for such data. As with `fitdist`, for some distributions (see [15] for details), it is necessary to specify initial values for the distribution parameters in the argument `start`. The `plotdistcens` function can help to find correct initial values for the distribution parameters in non trivial cases, by an manual iterative use if necessary.

```
> fsal.ln <- fitdistcens(salinity, "lnorm")
> fsal.ll <- fitdistcens(salinity, "llogis", start=list(shape=5, scale=40))
> summary(fsal.ln)

FITTING OF THE DISTRIBUTION ' lnorm ' BY MAXIMUM LIKELIHOOD ON CENSORED DATA
PARAMETERS
        estimate Std. Error
meanlog   3.3854    0.06487
sdlog     0.4961    0.05455
Loglikelihood:  -139.1   AIC:  282.1   BIC:  287.5
Correlation matrix:
        meanlog  sdlog
meanlog  1.0000 0.2938
sdlog    0.2938 1.0000

> summary(fsal.ll)

FITTING OF THE DISTRIBUTION ' llogis ' BY MAXIMUM LIKELIHOOD ON CENSORED DATA
PARAMETERS
      estimate Std. Error
shape    3.421     0.4158
scale   29.930     1.9447
Loglikelihood:  -140.1   AIC:  284.1   BIC:  289.5
Correlation matrix:
        shape   scale
shape  1.0000 -0.2022
scale -0.2022  1.0000
```

Computations of goodness-of-fit statistics have not yet been developed for fits using censored data but the quality of fit can be judged using Akaike and Schwarz information criteria (AIC and BIC) and the goodness-of-fit CDF plot, respectively provided when summarizing or plotting an object of class `"fitdistcens"`. Function `cdfcompcens` can also be used to compare the fit of various distributions to the same censored data set. Its call is similar to the one of `cdfcomp`. Below is an example of comparison of the two fitted distributions to the `salinity` data set (see Figure 10).

```
> cdfcompcens(list(fsal.ln, fsal.ll),
+     legendtext=c("lognormal", "loglogistic "))
```

Function `bootdistcens` is the equivalent of `bootdist` for censored data, except that it only proposes nonparametric bootstrap, as it is not obvious to simulate censoring within a parametric bootstrap resampling procedure. The generic function `quantile` can also be applied, as for continuous non-censored data, to an object of class `"fitdistcens"` or `"bootdistcens"`.

In addition to the fit of distributions to censored or non censored continuous data, our package can also accomodate discrete variables, such as count numbers, using the functions developped for continuous non-censored data. These functions will provide somewhat different graphs and statistics, taking into account the discrete type of the modeled variable. The discrete nature of the variable is automatically recognized when a classical distribution is fitted to data (binomial, negative binomial, geometric, hypergeometric and Poisson distributions) but must be indicated by fixing argument `discrete` to `TRUE` in the call to functions in other cases. The `toxocara` data set included in the package corresponds to the observation of such a discrete variable. It reports numbers of *Toxocara cati* parasites present in digestive tract, from a random sampling of feral cats living on Kerguelen island ([19]). We use it to illustrate the case of discrete data.

```
> data(toxocara)
> str(toxocara)

'data.frame':        53 obs. of  1 variable:
 $ number: int  0 0 0 0 0 0 0 0 0 0 ...
```

The fit of a discrete distribution to discrete data by maximum likelihood estimation requires the same procedure as for continuous non-censored data. As an example, using the `toxocara` data set, Poisson and negative distributions can be easily fitted.

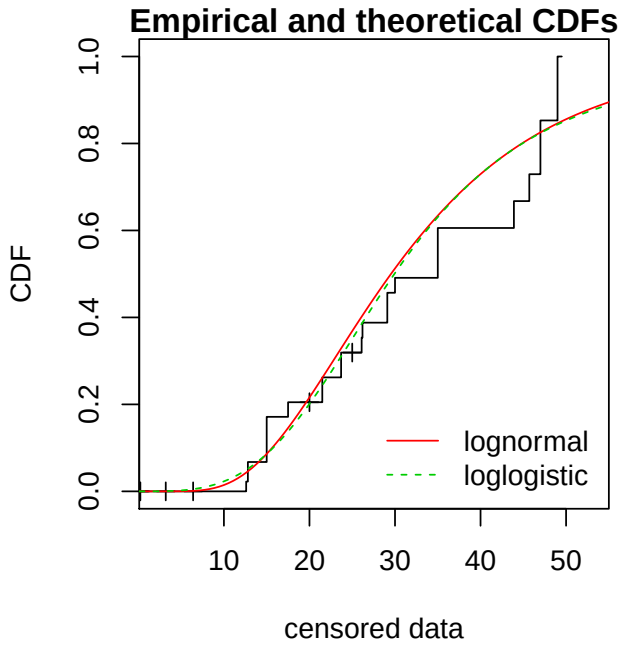

Figure 10: Goodness-of-fit CDF plot for fits of a lognormal and a loglogistic distribution to censored data: LC50 values from the `salinity` data set

```
> (ftoxo.P <- fitdist(toxocara$number, "pois"))

Fitting of the distribution ' pois ' by maximum likelihood
Parameters:
       estimate Std. Error
lambda    8.679     0.4047

> (ftoxo.nb <- fitdist(toxocara$number, "nbinom"))

Fitting of the distribution ' nbinom ' by maximum likelihood
Parameters:
     estimate Std. Error
size   0.3971    0.08289
mu     8.6803    1.93501
```

For discrete distributions, the plot of an object of class `"fitdist"` simply provides two goodness-of-fit plots comparing empirical and theoretical distributions in pdf and in cd f(Figure 11). Function `cdfcomp` can also be used to compare several plots to the same data set, as follows for the previous fits (Figure 12).

```
> plot(ftoxo.P)
```

```
> cdfcomp(list(ftoxo.P,ftoxo.nb),legendtext=c("Poisson", "negative binomial"))
```

When fitting discrete distributions, the Chi-squared statistic is computed by the `gofstat` function using cells defined by the argument `chisqbreaks` or cells automatically defined from the data in order to reach roughly the same number of observations per cell, roughly equal to the argument `meancount`, or sligthly more if there are some ties. The choice to define cells from the empirical distribution (data), and not from the theoretical distribution, was done to enable the comparison of Chi-squared values obtained with different distributions fitted on a same data set. If arguments `chisqbreaks` and `meancount` are both omitted, `meancount` is fixed in order to obtain roughly $(4n)^{2/5}$ cells, with n the length of the data set [63]. Using this default option with the fits of the Poisson and the negative binomial distribution to `toxocara` data set are compared as follows, giving the preference to the negative binomial distribution, from both Chi-squared statistics and information criteria:

```
> gofstat(list(ftoxo.P,ftoxo.nb))

Chi-squared statistic:  31257 7.486
Degree of freedom of the Chi-squared distribution:  5 4
Chi-squared p-value:  0 0.1123
   the p-value may be wrong with some theoretical counts < 5
Chi-squared table:
```

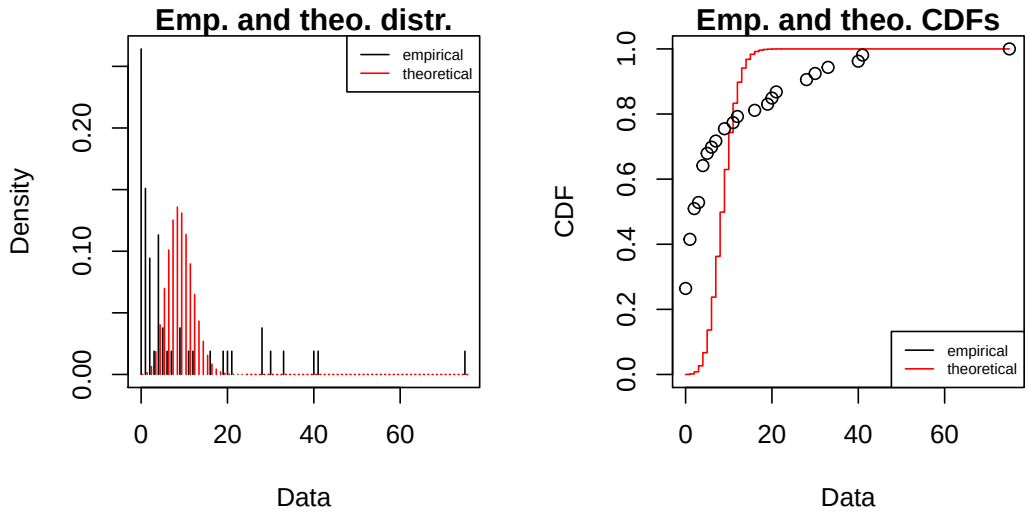

Figure 11: Plot of the fit of a discrete distribution (Poisson distribution fitted to numbers of *Toxocara cati* parasites from the `toxocara` data set)

```
      obscounts theo 1-mle-pois theo 2-mle-nbinom
<= 0         14        0.009014            15.295
<= 1          8        0.078237             5.809
<= 3          6        1.321767             6.845
<= 4          6        2.131298             2.408
<= 9          6       29.827829             7.835
<= 21         6       19.626224             8.271
> 21          7        0.005631             6.537

Goodness-of-fit criteria
                               1-mle-pois 2-mle-nbinom
Aikake's Information Criterion       1017        322.7
Bayesian Information Criterion       1019        326.6
```

4 Conclusion

In this paper, we present the R statistical package **fitdistrplus** for distribution fitting and demonstrate its use and usefulness. Our main objective while developing this package was to provide tools for helping R users to fit distributions to data. We have been encouraged to pursue our work by feedbacks from users of our package in various areas as food or environmental risk assessment, epidemiology, ecology, molecular biology, genomics, bioinformatics, hydraulics, mechanics, financial and actuarial mathematics or operations research. Indeed, this package is already used a lot by practionners and academics for simple MLE fits in [35, 28, 55, 64, 34, 40, 43, 56, 57, 22, 39], for MLE fits and goodness-of-fit statistics in [21, 58, 3, 5, 7, 45, 47], for MLE fits and bootstrap in [11, 42, 44, 59, 26], for MLE fits, bootstrap and goodness-of-fit statistics in [52], for MME fit in [37], for censored MLE and bootstrap in [36, 50, 29, 10, 48], for graphic analysing in [2, 54], for grouped-data fitting methods in [20] and more generally in [18, 9, 8, 1, 17, 4]. Many extensions of this package are planned in the future: we target to extend to censored data some methods for the moment only available for non-censored data, especially concerning goodness-of-fit evaluation and fitting methods. We will also enlarge the choice of fitting methods for non-censored data, by proposing new goodness-of-fit distances (e.g. distances based on quantiles) for maximum goodness-of-fit estimation and new types of moments (e.g. limited expected values) for moment matching estimation. At last, we will consider the case of multivariate distribution fitting.

5 Acknowledgments

The package would not have been at this stage without the stimulating contribution of Régis Pouillot and Jean-Baptiste Denis, especially for its conceptualization.

References

[1] Ö. Aktaş and M. Sjöstrand. Cornish-fisher expansion and value-at-risk method in application to risk management of large portfolios. Master's thesis, School of Information Science, Computer and Electrical Engineering, Halmstad University, 2011.

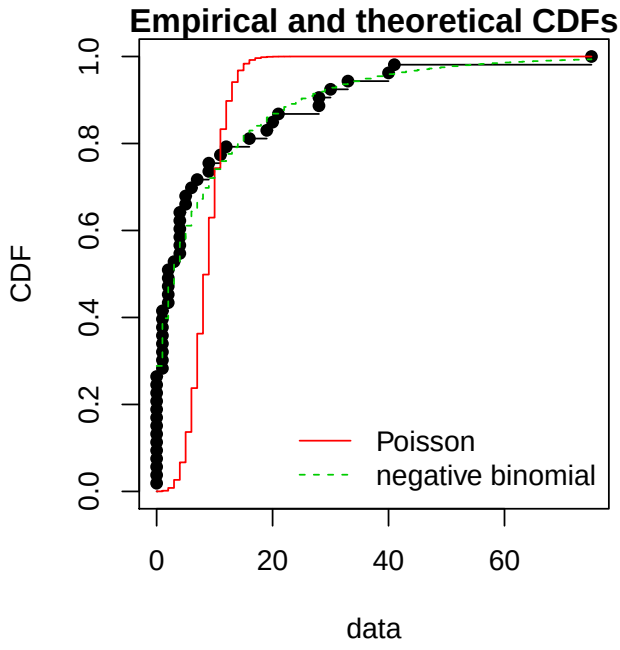

Figure 12: Comparison of the fits of a negative binomial and a Poisson distribution to numbers of *Toxocara cati* parasites from the `toxocara` data set)

[2] P. Anand, K. Yeturu, and N. Chandra. Pocketannotate: towards site-based function annotation. *Nucleic Acids Research*, 40(W1 W400-W408):1–9, 2012.

[3] A. Bagaria, V. Jaravine, Y.J. Huang, G.T. Montelione, and P. Güntert. Protein structure validation by generalized linear model root-mean-square deviation prediction. *Protein Science*, 21(2):229–238, 2012.

[4] R.O. Bakos. *Poszméh együttesek összehasonlító vizsgálata a cserépfalusi fás legeló klönbözó növényborítású területein.* PhD thesis, Hungarian Veterinary Archive, 2011.

[5] R. Benavides-Piccione, I. Fernaud-Espinosa, V. Robles, R. Yuste, and J. DeFelipe. Age-based comparison of human dendritic spine structure using complete three-dimensional reconstructions. *Cerebral Cortex*, 2012.

[6] G. Blom. *Statistical Estimates and Transformed Beta Variables.* Wiley, New York, 1959.

[7] N. Breitbach, K. Böhning-Gaese, I. Laube, and M. Schleuning. Short seed-dispersal distances and low seedling recruitment in farmland populations of bird-dispersed cherry trees. *Journal of Ecology*, 100(6):1349–1358, 2012.

[8] J.P. Brooks, D.J. Edwards, T.P. Sorrell, S. Srinivasan, and R.L. Diehl. Simulating calls for service for an urban police department. In *the 2011 Winter Simulation Conference*, pages 1770–1777, 2011.

[9] P. Busschaert, A.H. Geeraerd, M. Uyttendaele, and J.F. Van Impe. Estimating distributions out of qualitative and (semi)quantitative microbiological contamination data for use in risk assessment. *International Journal of Food Microbiology*, 138:260–269, 2010.

[10] N. Commeau, E. Parent, M.-L. Delignette-Muller, and M. Cornu. Fitting a lognormal distribution to enumeration and absence/presence data. *International Journal of Food Microbiology*, 155:146–152, 2012.

[11] N. J. Croucher, S. R. Harris, L. Barquist, J. Parkhill, and S. D. Bentley. A high-resolution view of genome-wide pneumococcal transformation. *PLoS Pathog*, 8(6):e1002745, 2012.

[12] A.C. Cullen and H.C. Frey. *Probabilistic Techniques in Exposure Assessment.* Plenum Publishing Co., New York, first edition, 1999.

[13] R.B. D'Agostino and M.A. Stephens. *Goodness-of-Fit Techniques.* Dekker, New York, first edition, 1986.

[14] M. L. Delignette-Muller, M. Cornu, and AFSSA Stec Study Grp. Quantitative Risk Assessment for *Escherichia coli* O157:H7 in Frozen Ground Beef Patties Consumed by Young Children in French Households. *International Journal of Food Microbiology*, 128(1, SI):158–164, NOV 30 2008. 5th International Conference on Predictive Modelling in Foods, Natl Tech Univ Athens, Athens, GREECE, SEP 16-19, 2007.

[15] M.L. Delignette-Muller, R. Pouillot, J.B. Denis, and C. Dutang. *fitdistrplus: Help to Fit of a Parametric Distribution to Non-Censored or Censored Data*, 2013. R package version 1.0-1.

[16] C. Dutang, V. Goulet, and M. Pigeon. actuar: an R package for Actuarial Science. *Journal of Statistical Software*, 25(7), 2008.

[17] M. Eik and H. Herrmann. Raytraced images for testing the reconstruction of fibre orientation distributions. In *the Estonian Academy of Sciences*, volume 61, pages 128–136, 2012.

[18] M. Eling. Fitting insurance claims to skewed distributions: Are the skew-normal and the skew-student good models? *Insurance: Mathematics and Economics*, 51(2012):239–248, 2012.

[19] E Fromont, L Morvilliers, M Artois, and D Pontier. Parasite Richness and Abundance in Insular and Mainland Feral Cats: Insularity or Density? *Parasitology*, 123(Part 2):143–151, AUG 2001.

[20] C.H.Y. Fu, H. Steiner, and S.G. Costafreda. Predictive neural biomarkers of clinical response in depression: A meta-analysis of functional and structural neuroimaging studies of pharmacological and psychological therapies. *Neurobiology of Disease*, 2012.

[21] P. Garcia. Analyse statistique des pannes du réseau HTA. Master's thesis, Université de Strasbourg, 2012.

[22] J.P. González-Varo, J.V. López-Bao, and J. Guitián. Functional diversity among seed dispersal kernels generated by carnivorous mammals. *Journal of Animal Ecology*, 81, 2012.

[23] V. Goulet. *actuar: An R Package for Actuarial Science, version 1.1-5*. École d'actuariat, Université Laval, 2012.

[24] D.R. Helsel. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. Wiley, 1st edition, 2005.

[25] S.S. Hirano, M.K. Clayton, and C.D. Upper. Estimation of and temporal changes in means and variances of populations of *Pseudomonas syringae* on snap bean leaflets. *Phytopathology*, 84(9):934–940, 1994.

[26] K. Hoelzer, R. Pouillot, D. Gallagher, M.B. Silverman, J. Kause, and S. Dennis. Estimation of *Listeria monocytogenes* transfer coefficients and efficacy of bacterial removal through cleaning and sanitation. *International Journal of Food Microbiology*, 157(2):267–277, 2012.

[27] G.C. Hose and P.J. Van den Brink. Confirming the Species-Sensitivity Distribution Concept for Endosulfan Using Laboratory, Mesocosm, and Field Data. *Archives of environmental contamination and toxicology*, 47(4):511–520, OCT 2004.

[28] S. Jaloustre, M. Cornu, E. Morelli, V. Noel, and M.L. Delignette-Muller. Bayesian modeling of *Clostridium perfringens* growth in beef-in-sauce products. *Food microbiology*, 28(2):311–320, 2011.

[29] I. Jongenburger, M.W. Reij, E.P.J. Boer, M.H. Zwietering, and L.G.M. Gorris. Modelling homogeneous and heterogeneous microbial contaminations in a powdered food product. *International Journal of Food Microbiology*, 157(1):35–44, 2012.

[30] D. Jordan. Simulating the sensitivity of pooled-sample herd tests for fecal salmonella in cattle. preventive veterinary medicine. *Preventive Veterinary Medicine*, 70(1-2):59–73, 2005.

[31] B.J. Kefford, E.J. Fields, C. Clay, and D. Nugegoda. Salinity tolerance of riverine macroinvertebrates from the Southern Murray-Darling Basin. *Mar. Freshwater Res*, 58(1019-1031), 2007.

[32] J.P. Klein and M.L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2nd edition, 2003.

[33] S.A. Klugman, H.H. Panjer, and G.E. Willmot. *Loss Models: from Data to Decisions*. Wiley, New York, third edition, 2009.

[34] F.H. Koch, D. Yemshanov, R.D. Magarey, and W.D. Smith. Dispersal of invasive forest insects via recreational firewood: A quantitative analysis. *Journal of Economic Entomology*, 105(2):438–450, 2012.

[35] M. Kohl and P. Ruckdeschel. R package distrMod: S4 Classes and Methods for Probability Models. *Journal of Statistical Software*, 35(10):1–27, 2010.

[36] A. Leha, T. Beissbarth, and K. Jung. Sequential interim analyses of survival data in DNA microarray experiments. *BMC Bioinformatics*, 12(127):1–14, 2011.

[37] K.L. Luangkesorn, B.A. Norman, Y. Zhuang, M. Falbo, and J. Sysko. Practice summaries: designing disease prevention and screening centers in Abu Dhabi. *Interfaces*, 42(4):406–409, 2012.

[38] A. Luceno. Fitting the Generalized Pareto Distribution to Data using Maximum Goodness-of-fit Estimators. *Computational Statistics and Data Analysis*, 51(2):904–917, NOV 15 2006.

[39] J.N. Mandl, J.P. Monteiro, N. Vrisekoop, and R.N. Germain. T cell-positive selection uses self-ligand binding strength to optimize repertoire recognition of foreign antigens. *Immunity*, 2013.

[40] N. Marquetoux, M. Paul, S. Wongnarkpet, C. Poolkhet, W. Thanapongtham, F. Roger, C. Ducrot, and K. Chalvet-Monfray. Estimating spatial and temporal variations of the reproduction number for highly pathogenic avian influenza H5N1 epidemic in Thailand. *Preventive Veterinary Medicine*, 106(2):143–151, 2012.

[41] A.J. McNeil. Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bull.*, 1997.

[42] D. Méheust, P. Le Cann, T. Reponen, J. Wakefield, and S. Vesper. Possible application of the environmental relative moldiness index in France: a pilot study in Brittany. *International Journal of Hygiene and Environmental Health*, 2012.

[43] R.P.V. Nordan. An investigation of potential methods for topology preservation in interactive vector tile map applications. Master's thesis, Norwegian University of Science and Technology, 2012.

[44] P.W. Orellano, J.I. Reynoso, A. Grassi, A. Palmieri, O. Uez, and O. Carlino. Estimation of the serial interval for pandemic influenza (pH1N1) in the most southern province of Argentina. *Iranian Journal of Public Health*, 41(12):26–29, 2012.

[45] A.C. Callau Poduje. Bivariate analysis and synthesis of flood events for the design of hydraulic structures – a case study for Argentina. Master's thesis, Leibniz Universitat Hannover, 2012.

[46] R. Pouillot and M.L. Delignette-Muller. Evaluating Variability and Uncertainty Separately in Microbial Quantitative Risk Assessment using two R Packages. *International Journal of Food Microbiology*, 142(3):330–340, SEP 1 2010.

[47] R. Pouillot and M.L. Delignette-Muller. Evaluating variability and uncertainty separately in microbial quantitative risk assessment using two R packages. *International Journal of Food Microbiology*, 142(2010):330–340, 2010.

[48] R. Pouillot, M.L. Delignette-Muller, and M. Cornu. Case study: *L. monocytogenes* in cold-smoked salmon. 2011.

[49] R. Pouillot, M.L. Delignette-Muller, and J.B. Denis. *mc2d: Tools for Two-Dimensional Monte-Carlo Simulations*, 2011. R package version 0.1-12.

[50] R. Pouillot, K. Hoelzer, Y. Chen, and S. Dennis. Estimating probability distributions of bacterial concentrations in food based on data generated using the most probable number (MPN) method for use in risk assessment. *Food Control*, 29(2):350–357, 2012.

[51] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[52] A.J.d.S. Rebuge. Business process analysis in healthcare environments. Master's thesis, Universidade Tecnica de Lisboa, 2012.

[53] Ricci, V. Fitting distributions with **R**. Contributed Documentation available on CRAN, 2005.

[54] A. S. Rosa. Funções de predição espacial de propriedades do solo. Master's thesis, Universidade Federal de Santa Maria, 2012.

[55] H. Sak and C. Haksoz. A copula-based simulation model for supply portfolio risk. *Journal of Operational Risk*, 2011.

[56] C.F. Scholl, C.C. Nice, J.A. Fordyce, Z. Gompert, and M.L. Forister. Larval performance in the context of ecological diversification and speciation in lycaeides butterflies. *International Journal of Ecology*, 2012(2012):1–13, 2012.

[57] J.P. Suuronen, A. Kallonen, M. Eik, J. Puttonen, Ritva Serimaa, and Heiko Herrmann. Analysis of short fibres orientation in steel fibre-reinforced concrete (SFRC) by X-ray tomography. *Journal of Materials Science*, 2012.

[58] T. Tarnczi, V. Fenyves, and Z. Bcs. The business uncertainty and variability management with real options models combined two dimensional simulation. *International Journal of Management Cases*, 13(3):159–167, 2011.

[59] A. Tello, B. Austin, and T.C. Telfer. Selective pressure of antibiotic pollution on bacteria of importance to public health. *Environmental Health Perspectives*, 120(8):1100–1106, 2012.

[60] T. Therneau. *survival: Survival Analysis, Including Penalized Likelihood*, 2011. R package version 2.36-9.

[61] B.W. Turnbull. Nonparametric Estimation of a Survivorship Function with Doubly Censored Data. *Journal of the American Statistical Association*, 69(345):169–173, 1974.

[62] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, 4 edition, 2010.

[63] D. Vose. *Quantitative Risk Analysis. A Guide to Monte Carlo Simulation Modelling*. Wiley, New York, first edition, 2010.

[64] T. Wilson. What were they thinking: modeling think times for performance testing. *CMG Journal Information*, 2011.