Fitting parametric distributions using **R**: the `fitdistrplus` package

M. L. Delignette-Muller - CNRS UMR 5558
R. Pouillot
J.-B. Denis - INRA MIAJ

useR! 2009,10/07/2009

Unité MIAJouy-en-Josas Mathématiques et Informatique Appliquées
INRA
Ecole Nationale Vétérinaire de Lyon
Université de Lyon
cnrs

Background

Specifying the probability distribution that best fits a sample data among a predefined family of distributions

- a frequent need especially in Quantitative Risk Assessment
- general-purpose maximum-likelihood fitting routine for the parameter estimation step : `fitdistr(MASS)` (Venables and Ripley, 2002)
- possibility to implement other steps using **R** (Ricci, 2005)
- but no specific package dedicated to the whole process
- difficulty to work with censored data

Objective

Build a package that provides functions to help the whole process of specification of a distribution from data

- choose among a family of distributions the best candidates to fit a sample
- estimate the distribution parameters and their uncertainty
- assess and compare the goodness-of-fit of several distributions

that specifically handles different kinds of data

- discrete
- continuous with possible censored values (right-, left- and interval-censored with several upper and lower bounds)

Technical choices

- Skewness-kurtosis graph for the choice of distributions

  (Cullen and Frey, 1999)

- Two fitting methods
  - matching moments

    for a limited number of distributions and non-censored data
  - **maximum likelihood** (mle) using `optim(stats)`

    for any distribution, predefined or defined by the user

    *for **non-censored** or **censored** data*
- Uncertainty on parameter estimations
  - standard errors from the Hessian matrix (only for mle)
  - parametric or non-parametric **bootstrap**
- Assessment of goodness-of-fit
  - chi-squared, Kolmogorov-Smirnov, Anderson-Darling statistics
  - density, cdf, P-P and Q-Q plots

Technical choices

- Skewness-kurtosis graph for the choice of distributions
  (Cullen and Frey, 1999)
- Two fitting methods
  - matching moments
    for a limited number of distributions and non-censored data
  - **maximum likelihood** (mle) using `optim(stats)`
    for any distribution, predefined or defined by the user
    *for **non-censored** or **censored** data*
- Uncertainty on parameter estimations
  - standard errors from the Hessian matrix (only for mle)
  - parametric or non-parametric **bootstrap**
- Assessment of goodness-of-fit
  - chi-squared, Kolmogorov-Smirnov, Anderson-Darling statistics
  - density, cdf, P-P and Q-Q plots

Technical choices

- Skewness-kurtosis graph for the choice of distributions
  (Cullen and Frey, 1999)
- Two fitting methods
  - matching moments
    for a limited number of distributions and non-censored data
  - **maximum likelihood** (mle) using `optim(stats)`
    for any distribution, predefined or defined by the user
    *for **non-censored** or **censored** data*
- Uncertainty on parameter estimations
  - standard errors from the Hessian matrix (only for mle)
  - parametric or non-parametric **bootstrap**
- Assessment of goodness-of-fit
  - chi-squared, Kolmogorov-Smirnov, Anderson-Darling statistics
  - density, cdf, P-P and Q-Q plots

Technical choices

- Skewness-kurtosis graph for the choice of distributions
  (Cullen and Frey, 1999)
- Two fitting methods
  - matching moments
    for a limited number of distributions and non-censored data
  - **maximum likelihood** (mle) using `optim(stats)`
    for any distribution, predefined or defined by the user
    *for **non-censored** or **censored** data*
- Uncertainty on parameter estimations
  - standard errors from the Hessian matrix (only for mle)
  - parametric or non-parametric **bootstrap**
- Assessment of goodness-of-fit
  - chi-squared, Kolmogorov-Smirnov, Anderson-Darling statistics
  - density, cdf, P-P and Q-Q plots

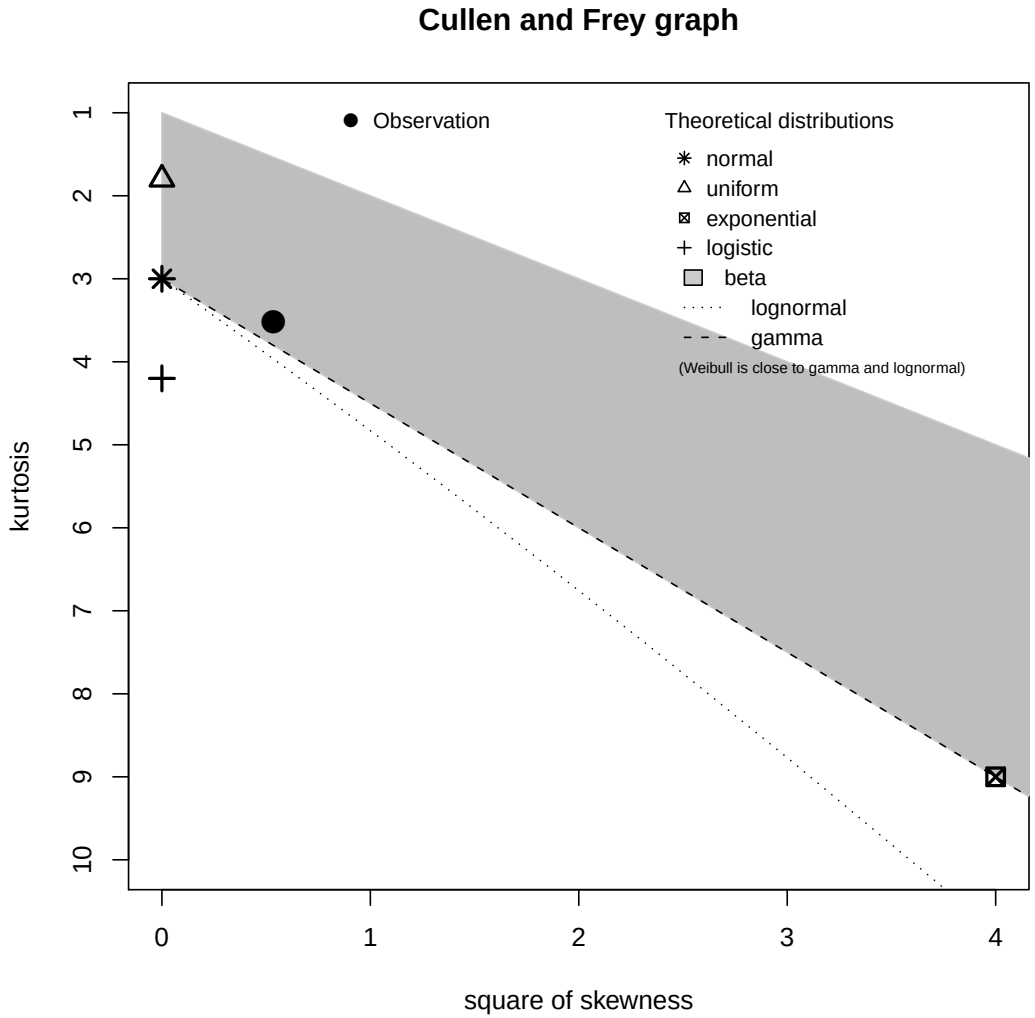
Main functions of `fitdistrplus`

- `descdist`: provides a skewness-kurtosis graph to help to choose the best candidate(s) to fit a given dataset
- `fitdist` and `plot.fitdist`: for a given distribution, estimate parameters and provide goodness-of-fit graphs and statistics
- `bootdist`: for a fitted distribution, simulates the uncertainty in the estimated parameters by bootstrap resampling
- `fitdistcens`, `plot.fitdistcens` and `bootdistcens`: same functions dedicated to continuous data with censored values

Skewness-kurtosis plot for continuous data

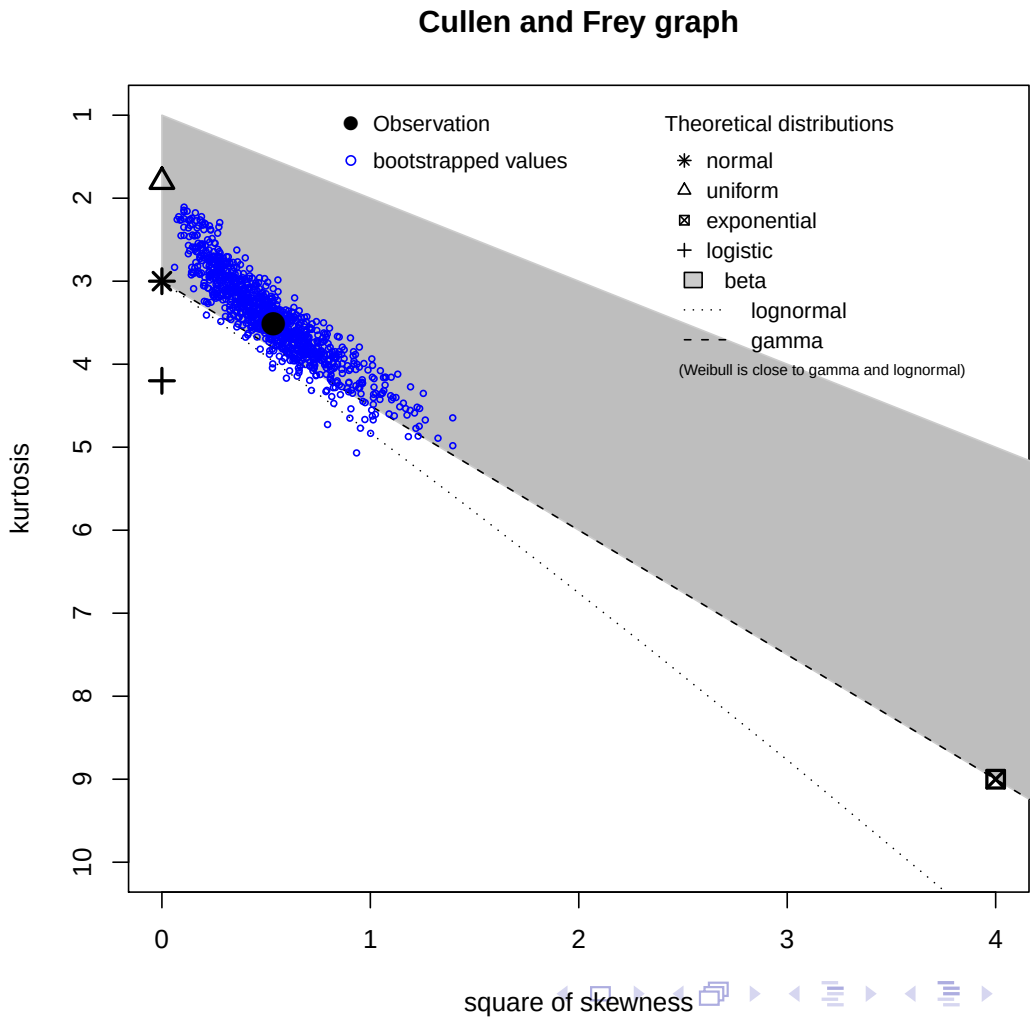
Ex. on consumption data: food serving sizes (g)

```
> descdist(serving.size)
```



Skewness-kurtosis plot for continuous data
with bootstrap option

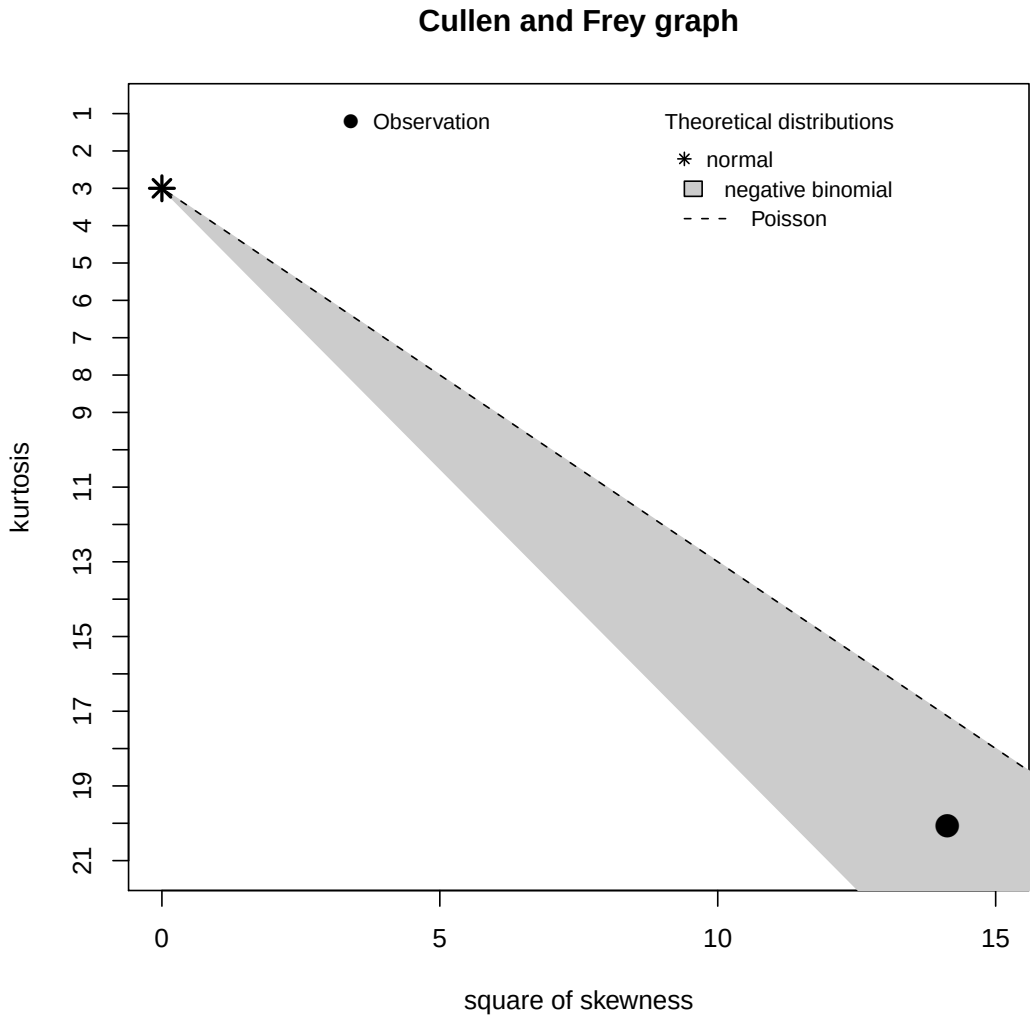
```
> descdist(serving.size,boot=1001)
```



Skewness-kurtosis plot for discrete data

Ex. on microbial data: counts of colonies on small food samples

```
> descdist(colonies.count,discrete=TRUE)
```



Fit of a given distribution by maximum likelihood or matching moments

Ex. on consumption data: food serving sizes (g)

- Maximum likelihood estimation

```
> fg.mle<-fitdist(serving.size,"gamma",method="mle")
> summary(fg.mle)
      estimate Std. Error
shape   4.0083    0.34134
rate    0.0544    0.00494
Loglikelihood:  -1254
```

- Matching moments estimation

```
> fg.mom<-fitdist(serving.size,"gamma",method="mom")
> summary(fg.mom)
      estimate
shape   4.2285
rate    0.0574
```

Fit of a given distribution by maximum likelihood or matching moments

Ex. on consumption data: food serving sizes (g)

- Maximum likelihood estimation

```
> fg.mle<-fitdist(serving.size,"gamma",method="mle")
> summary(fg.mle)
      estimate Std. Error
shape   4.0083    0.34134
rate    0.0544    0.00494
Loglikelihood:  -1254
```

- Matching moments estimation

```
> fg.mom<-fitdist(serving.size,"gamma",method="mom")
> summary(fg.mom)
      estimate
shape   4.2285
rate    0.0574
```

Comparison of goodness-of-fit statistics

Ex. on consumption data: food serving sizes (g)

Comparison of the fits of three distributions using the Anderson-Darling statistics

- Gamma

```
> fitdist(serving.size,"gamma")$ad
[1] 3.566019
```

- lognormal

```
> fitdist(serving.size,"lnorm")$ad
[1] 4.543654
```

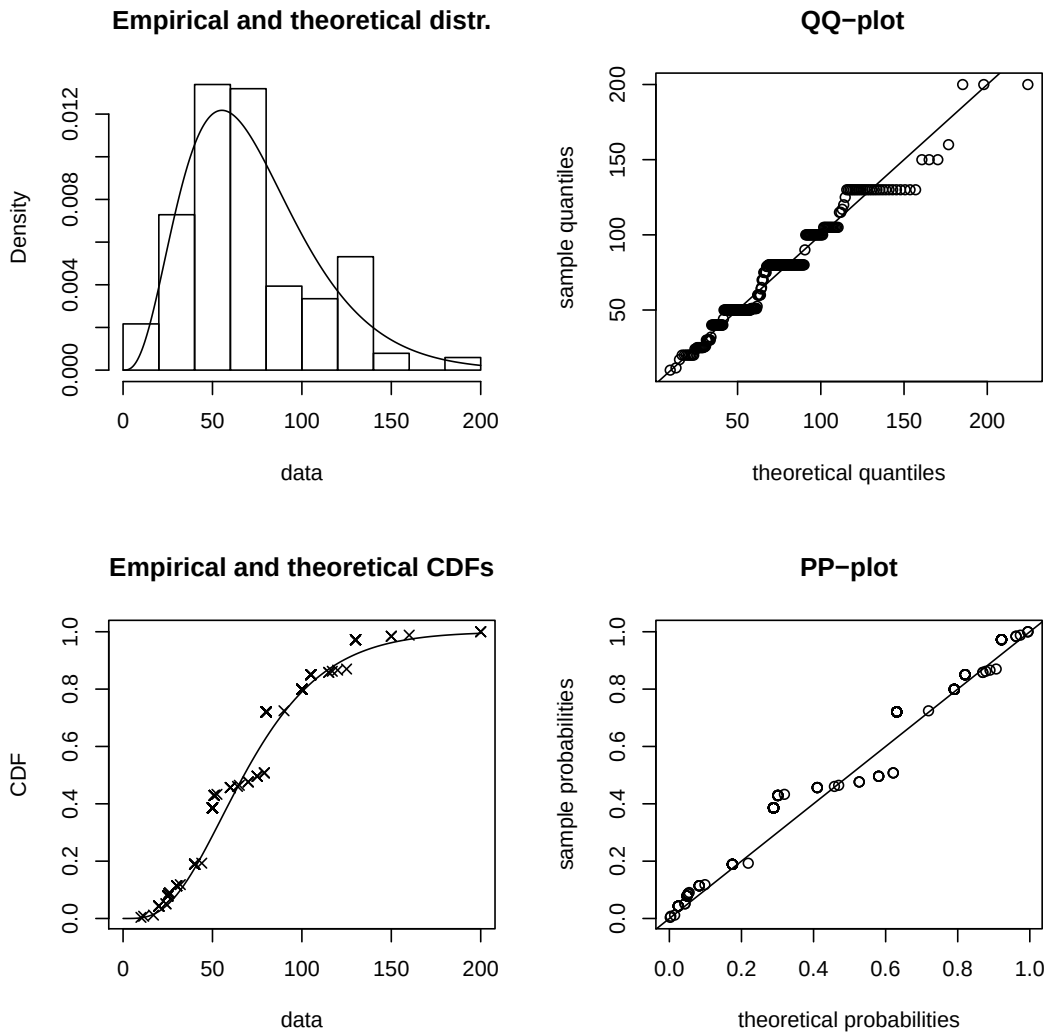
- Weibull

```
> fitdist(serving.size,"weibull")$ad
[1] 3.573646
```

Goodness-of-fit graphs for continuous data

Ex. on consumption data: food serving sizes (g)

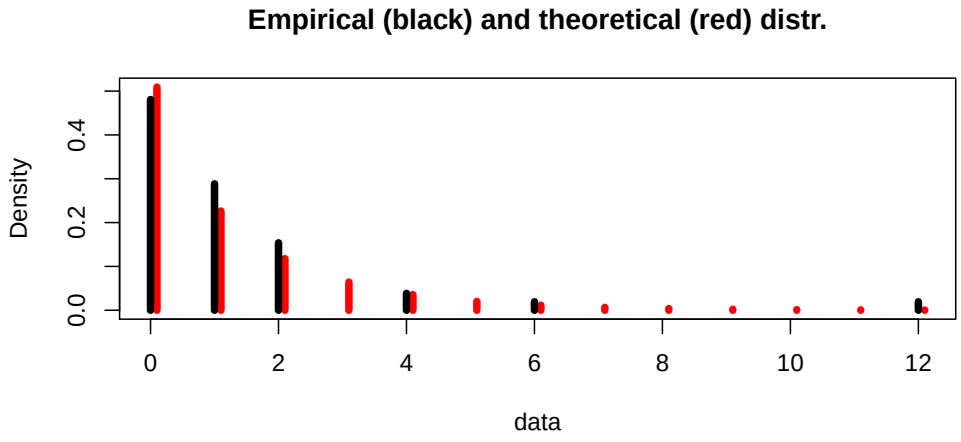
```
> plot(fg.mle)
```

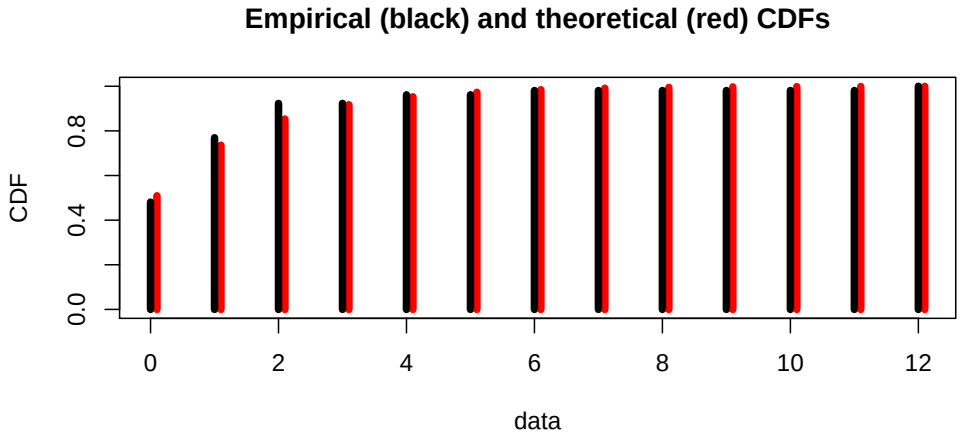


Goodness-of-fit graphs for discrete data

Ex. on microbial data: counts of colonies on small food samples

```
> fnbinom<-fitdist(colonies.count,"nbinom")
> plot(fnbinom)
```





Fit of a given distribution by maximum likelihood to censored data

Ex. on microbial censored data: concentrations in food

- with left censored values (not detected)
- and interval censored values (detected but not counted)

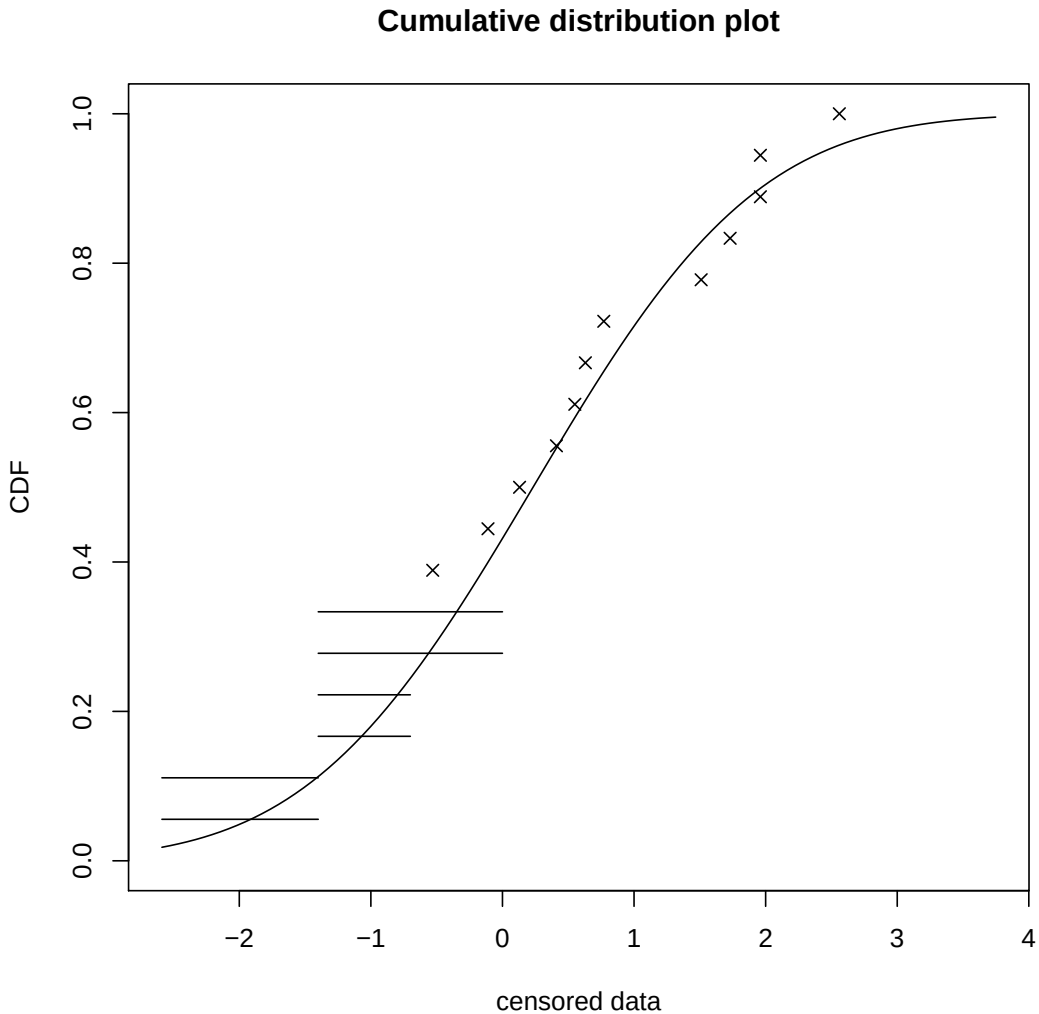
```
> log10.conc
    left right
1   1.73  1.73
2   1.51  1.51
3   0.77  0.77
4   1.96  1.96
5   1.96  1.96
6  -1.40  0.00
7  -1.40 -0.70
8     NA -1.40
9  -0.11 -0.11
...
```

```
> fnorm<-fitdistcens(log10.conc, "norm")
> summary(fnorm)

     estimate Std. Error
mean    0.118      0.332
sd      1.426      0.261

Loglikelihood:  -32.1
```

Goodness-of-fit graphs for censored data

Ex. on microbial censored data: concentrations in food

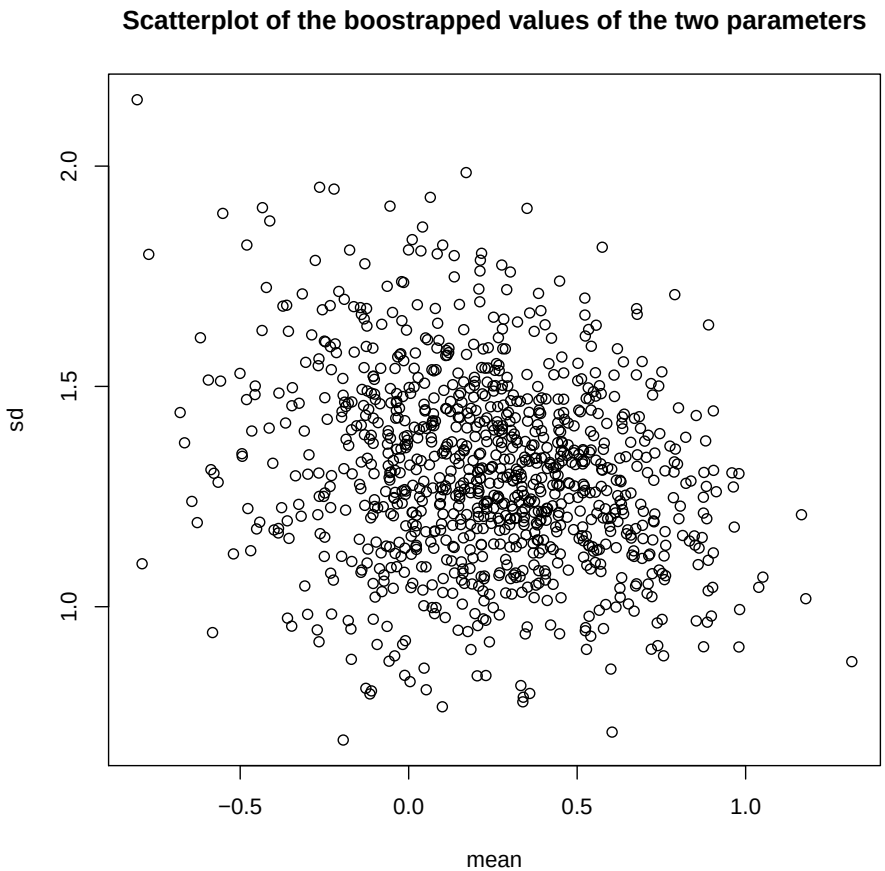
```
> plot(fnorm)
```



Bootstrap resampling

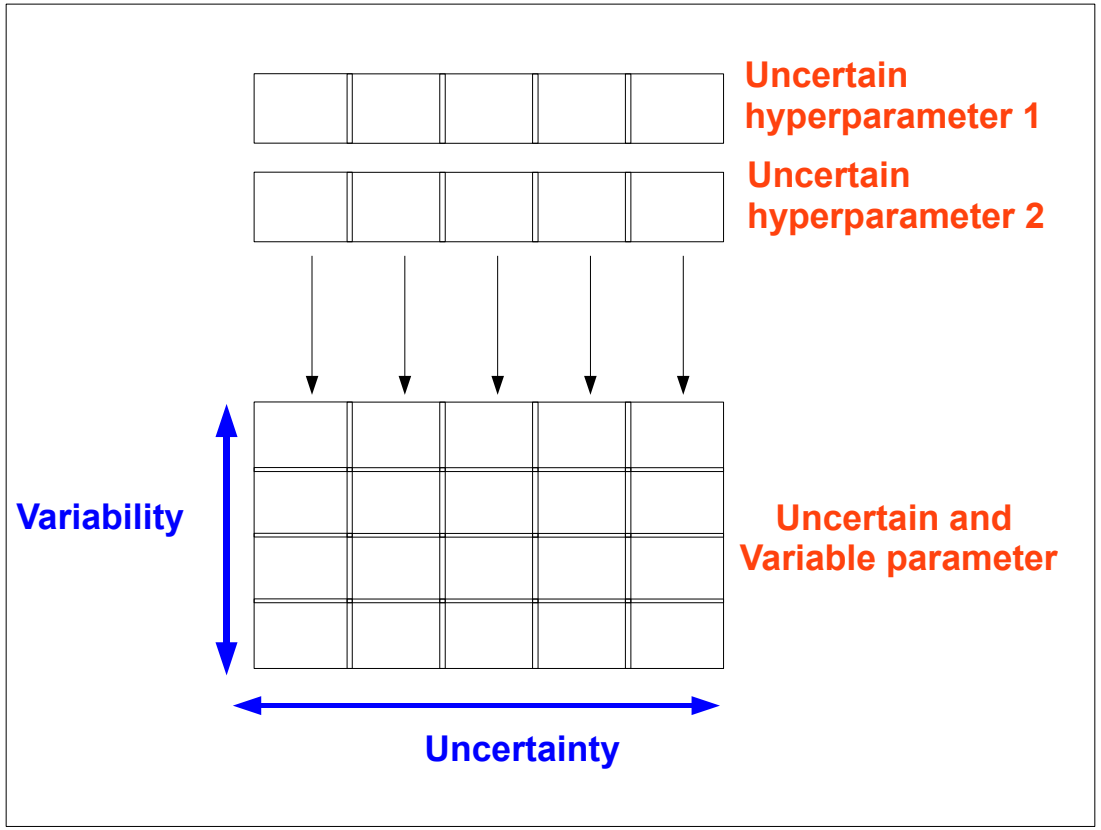
Ex. on microbial censored data

```
> bnorm<-bootdistcens(fnorm)
> summary(bnorm)
Nonparametric bootstrap medians and 95% CI
     Median   2.5% 97.5%
mean  0.233 -0.455 0.875
sd    1.294  0.908 1.776


> plot(bnorm)
```



Use of the bootstrap in risk assessment

The bootstrap sample may be used to take into account uncertainty in risk assessment,
in two-dimensional Monte Carlo simulations,
as proposed in the package `mc2d`.



Conclusion

- `fitdistrplus` could help risk assessment.
  It is a part of a collaborative project with 2 other packages under development, `mc2d` and `ReBaStaBa`:

  The R-Forge project "Risk Assessment with R"

  `http://riskassessment.r-forge.r-project.org/`

- `fitdistrplus` could also be used more largely to help the fit of univariate distributions to data

Conclusion

- `fitdistrplus` could help risk assessment.
  It is a part of a collaborative project with 2 other packages under development, `mc2d` and `ReBaStaBa`:

  The R-Forge project "Risk Assessment with R"
  `http://riskassessment.r-forge.r-project.org/`

- `fitdistrplus` could also be used more largely to help the fit of univariate distributions to data

Still many things to do

`fitdistrplus` is still under development.
Many improvements are planned

- other goodness-of-fit statistics
- other graphs for goodness-of-fit for censored data (Turnbull,...)
- optimized choice of the algorithm used in `optim` for the likelihood maximization
- graphs of likelihood contours (detection of identifiability problems)
- ...

do not hesitate to provide us other improvement ideas !