Use of the library `fitdistrplus` to specify a distribution from non-censored or censored data

Marie Laure Delignette-Muller, Régis Pouillot, Jean-Baptiste Denis and Christophe Dutang

July 18, 2009

Here you will find some easy examples of use of the functions of the library **fitdistrplus**. The aim is to show you by examples how to use these functions to help you to specify a parametric distribution from data corresponding to a random sample drawn from a theoretical distribution that you want to describe. For details, see the documentation of each function, using the R help command (ex.: `?fitdist`). Do not forget to load the library using the function `library` before testing following examples.

```
> library(fitdistrplus)
```

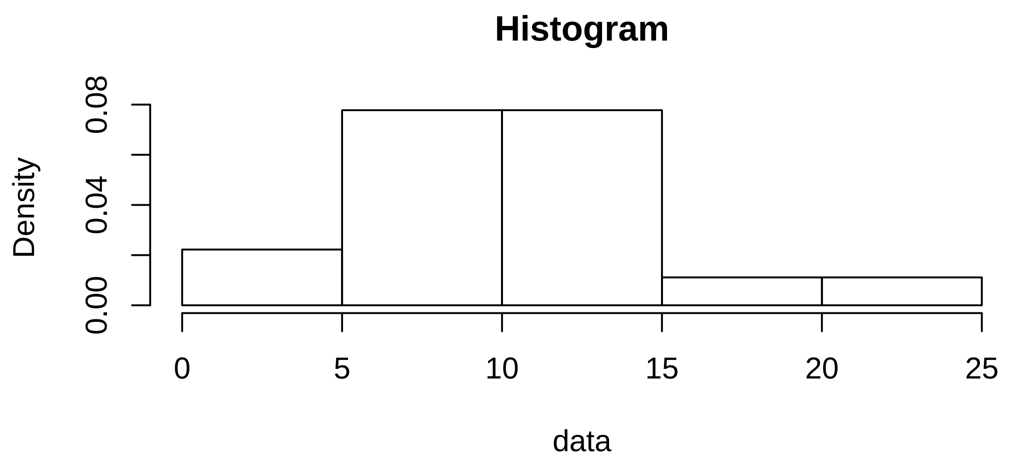
Contents

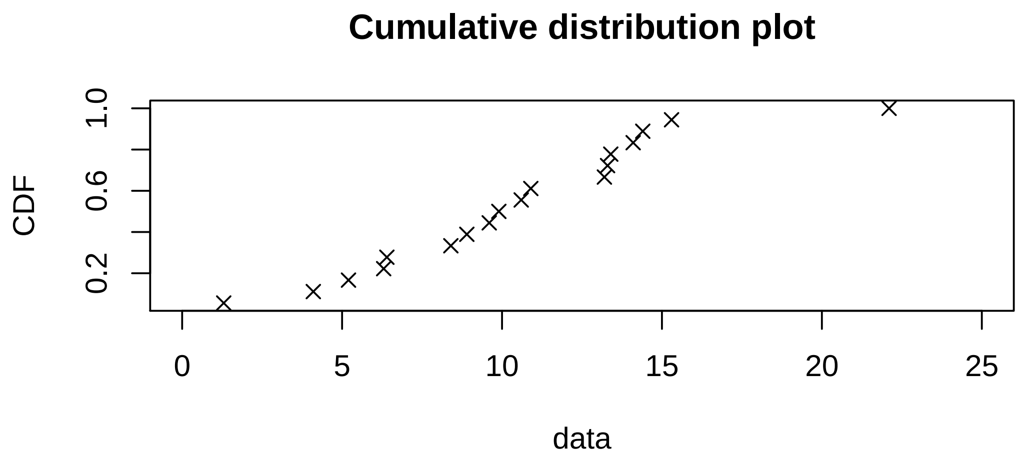
1 Specification of a distribution from non-censored continuous data

1.1 Graphical display of the observed distribution

First of all, the observed distribution may be plotted using the function `plotdist`.

```
> x1 <- c(6.4, 13.3, 4.1, 1.3, 14.1, 10.6, 9.9, 9.6, 15.3, 22.1,
+     13.4, 13.2, 8.4, 6.3, 8.9, 5.2, 10.9, 14.4)
> plotdist(x1)
```

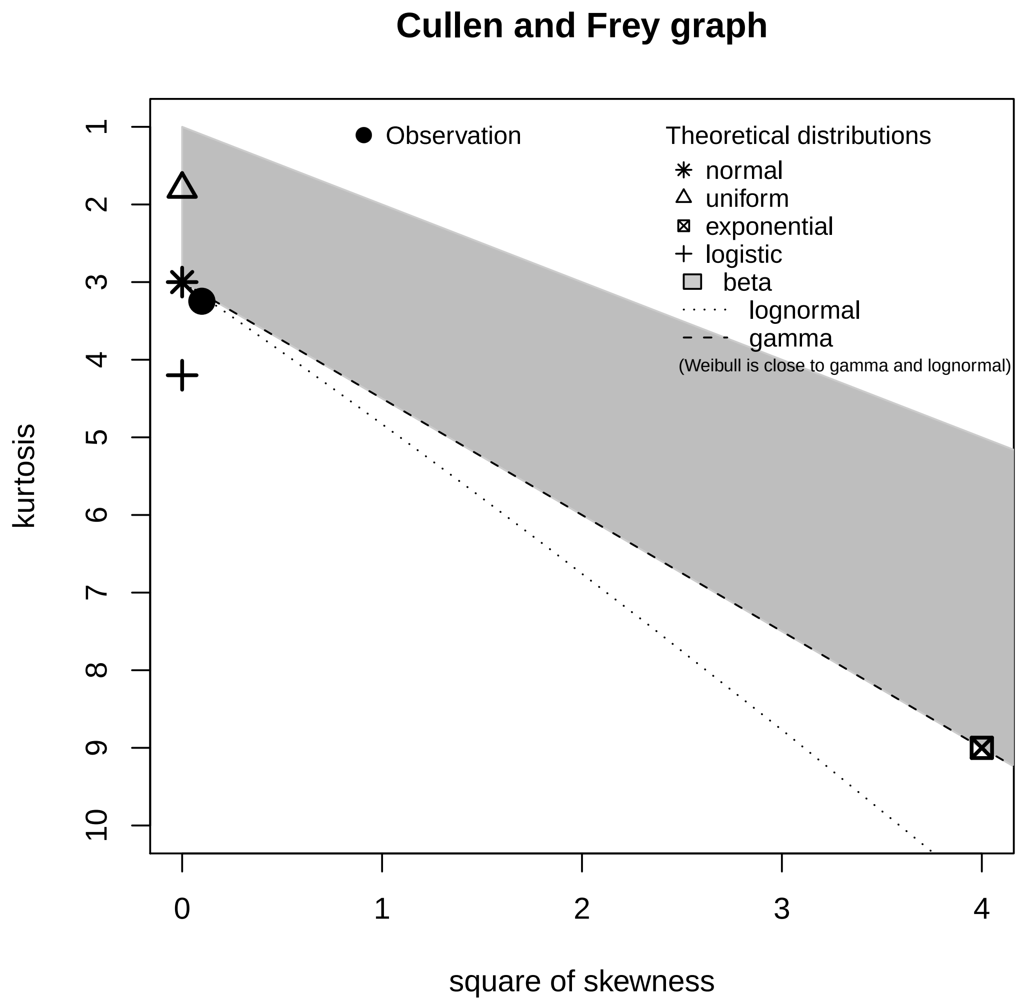




1.2 Characterization of the observed distribution

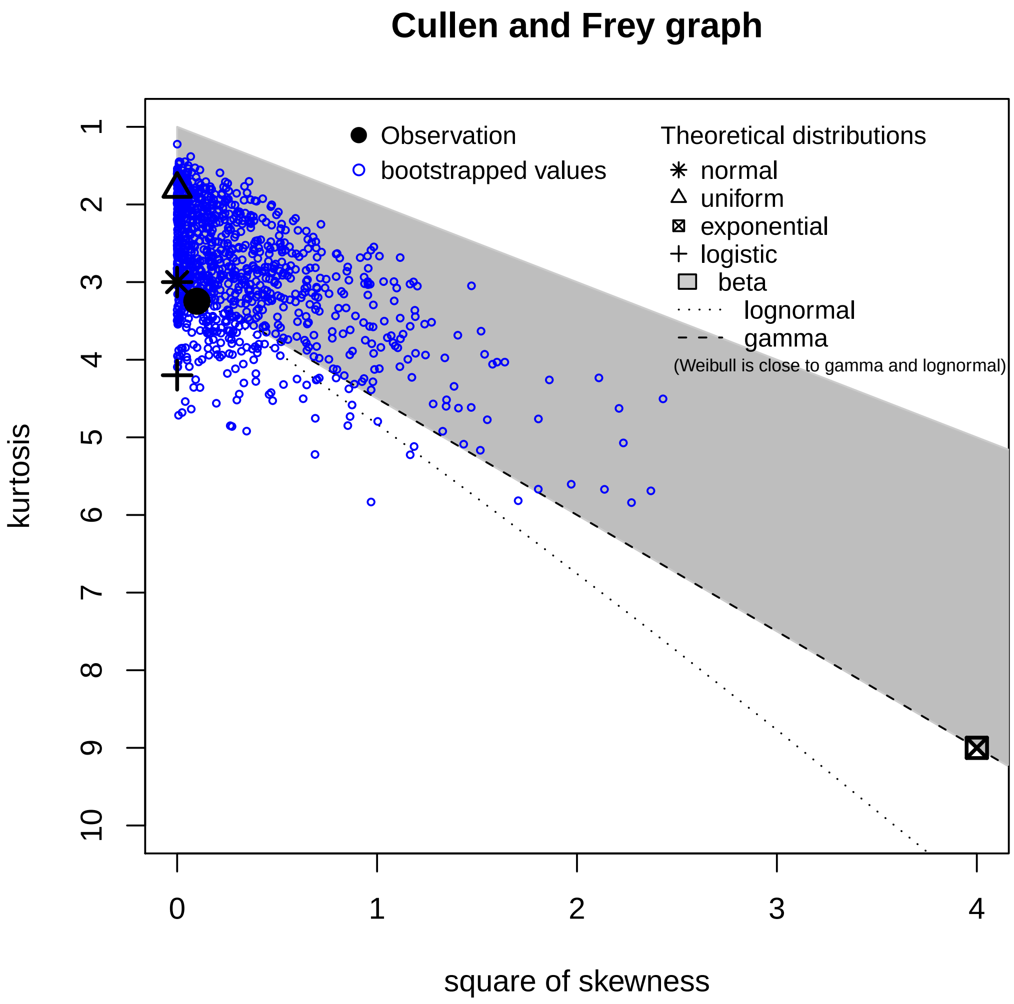
Descriptive parameters of the empirical distribution may be computed using the function `descdist`. This function will also provide by default a skewness-kurtosis plot which may help you to select which distribution(s) to fit among the potential candidates.

```
> descdist(x1)

summary statistics
------
min:  1.3   max:  22.1
median:  10.2
mean:  10.4
sample sd:  4.75
sample skewness:  0.314
sample kurtosis:  3.25
```



In order to take into account the uncertainty of the estimated values of kurtosis and skewness, the data set may be boostrapped by fixing the argument `boot` to an integer above 10 in `descdist`. `boot` values of skewness and kurtosis corresponding to the boot nonparametric bootstrap samples are then computed and reported in blue color on the skewness-kurtosis plot.

```
> descdist(x1, boot = 1000)

summary statistics
------
min:  1.3   max:  22.1
median:  10.2
mean:  10.4
sample sd:  4.75
sample skewness:  0.314
sample kurtosis:  3.25
```



1.3 Fitting of a distribution

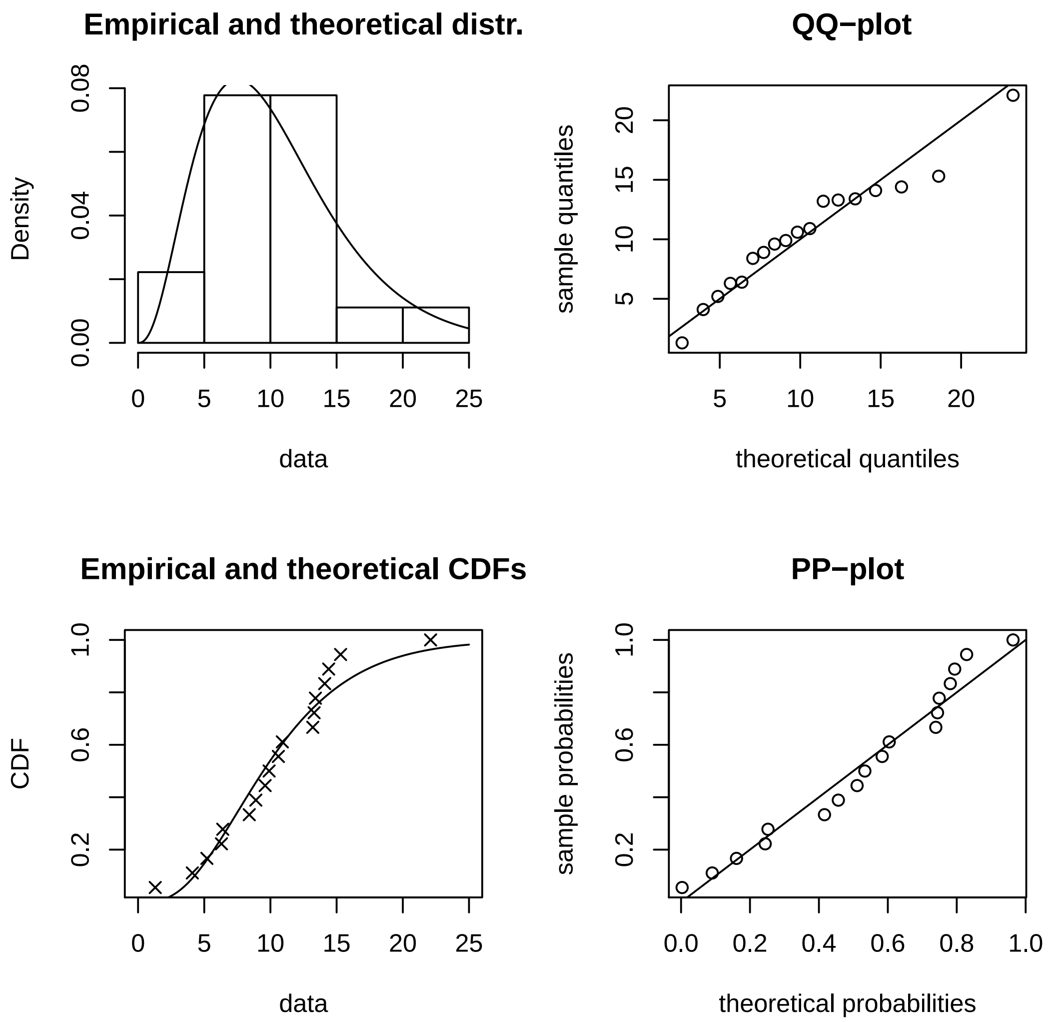
One or more parametric distributions may then be fitted to the data set, one at a time, using the function `fitdist`. This function uses the maximum likelihood method if the argument `method="mle"` (or if it is omitted) or the matching moments method if the argument `method="mom"`. When fitting continuous[1] distributions, Kolmogorov-Smirnov and Anderson-Darling statistics are computed and corresponding tests are performed when possible. Even if less appropriate for continuous distributions, the Chi-squared statistic is also computed when possible. For this calculation, cells are defined by the argument `chisqbreaks` or automatically defined from the data set and from the argument `meancount` (the approximate mean count per cell) which is fixed to $(4n)^{2/5}$ if omitted (with n the length of the data set). For more details, see the help of the function `fitdist`. Four goodness of fit plots are also provided.

Below is the result of a fit of a gamma distribution by maximum likelihood.

```
> f1g <- fitdist(x1, "gamma")
> plot(f1g)
> summary(f1g)

FITTING OF THE DISTRIBUTION ' gamma ' BY MAXIMUM LIKELIHOOD
PARAMETERS
      estimate Std. Error
shape    3.575      1.140
rate     0.343      0.118
Loglikelihood:  -54.4
Correlation matrix:
      shape  rate
shape 1.000 0.931
rate  0.931 1.000

------
```

[1]The reference book on continuous distribution is Johnson et al. (1994).

```
GOODNESS-OF-FIT STATISTICS

_____________ Chi-squared_____________
Chi-squared statistic:  7.93
Degree of freedom of the Chi-squared distribution:  3
Chi-squared p-value:  0.0475
!!! the p-value may be wrong
            with some theoretical counts < 5 !!!

!!! For continuous distributions, Kolmogorov-Smirnov and
      Anderson-Darling statistics should be prefered !!!

_____________ Kolmogorov-Smirnov_____________
Kolmogorov-Smirnov statistic:  0.138
Kolmogorov-Smirnov test: not calculated

_____________ Anderson-Darling_____________
Anderson-Darling statistic:  0.457
Anderson-Darling test:  not rejected
```



Below is the result of another fit of the same distribution by matching moments.

```
> f1gbis <- fitdist(x1, "gamma", method = "mom")
> summary(f1gbis)

FITTING OF THE DISTRIBUTION ' gamma ' BY MATCHING MOMENTS
PARAMETERS
      estimate
shape    4.810
rate     0.462
```

```
------
GOODNESS-OF-FIT STATISTICS

_____________ Chi-squared_____________
Chi-squared statistic:  7.27
Degree of freedom of the Chi-squared distribution:  3
Chi-squared p-value:  0.0637
!!! the p-value may be wrong
            with some theoretical counts < 5 !!!

!!! For continuous distributions, Kolmogorov-Smirnov and
      Anderson-Darling statistics should be prefered !!!

_____________ Kolmogorov-Smirnov_____________
Kolmogorov-Smirnov statistic:  0.144
Kolmogorov-Smirnov test: not calculated

_____________ Anderson-Darling_____________
Anderson-Darling statistic:  0.471
Anderson-Darling test:  not rejected
```

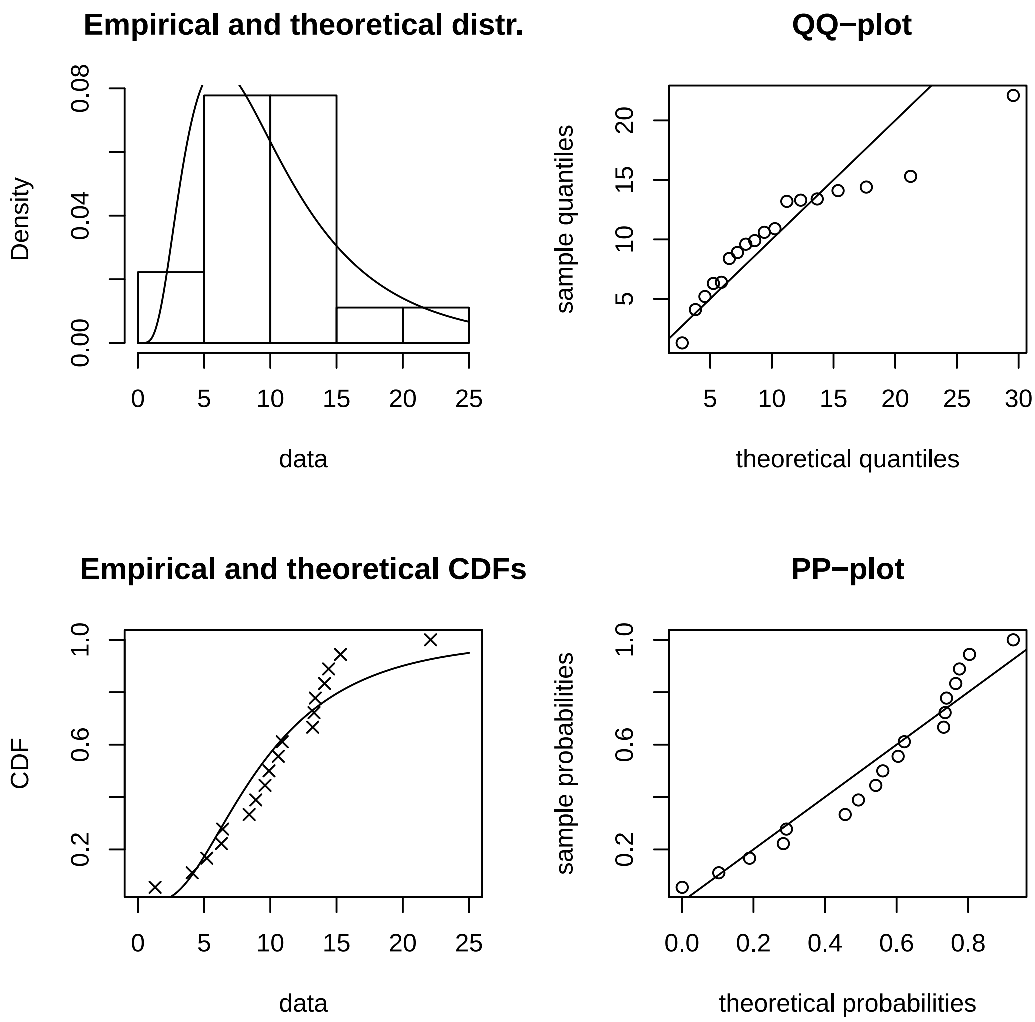
As can be seen in this returned summary, the automatic definition of the cells required to calculate the Chi-squared statistic does not give theoretical counts large enough to validate the use of the test in this example. It is often the case for small data sets. The observed and theoretical counts may be printed as below :

```
> f1g$chisqtable

        obscounts theocounts
<= 5.2  3.0000000  2.8950753
<= 8.4  3.0000000  4.5964712
<= 9.9  3.0000000  2.1080076
<= 13.2 3.0000000  3.7064776
<= 14.1 3.0000000  0.7577383
> 14.1  3.0000000  3.9362300
```

Below is the fit of a lognormal distribution.

```
> f1l <- fitdist(x1, "lnorm")
> plot(f1l)
> summary(f1l)

FITTING OF THE DISTRIBUTION ' lnorm ' BY MAXIMUM LIKELIHOOD
PARAMETERS
        estimate Std. Error
meanlog    2.197      0.147
sdlog      0.622      0.104
Loglikelihood:  -56.5
Correlation matrix:
          meanlog     sdlog
meanlog  1.00e+00 -2.70e-11
sdlog   -2.70e-11  1.00e+00

------
GOODNESS-OF-FIT STATISTICS

_____________ Chi-squared_____________
Chi-squared statistic:  11.1
Degree of freedom of the Chi-squared distribution:  3
Chi-squared p-value:  0.0110
!!! the p-value may be wrong
            with some theoretical counts < 5 !!!

!!! For continuous distributions, Kolmogorov-Smirnov and
      Anderson-Darling statistics should be prefered !!!

_____________ Kolmogorov-Smirnov_____________
```

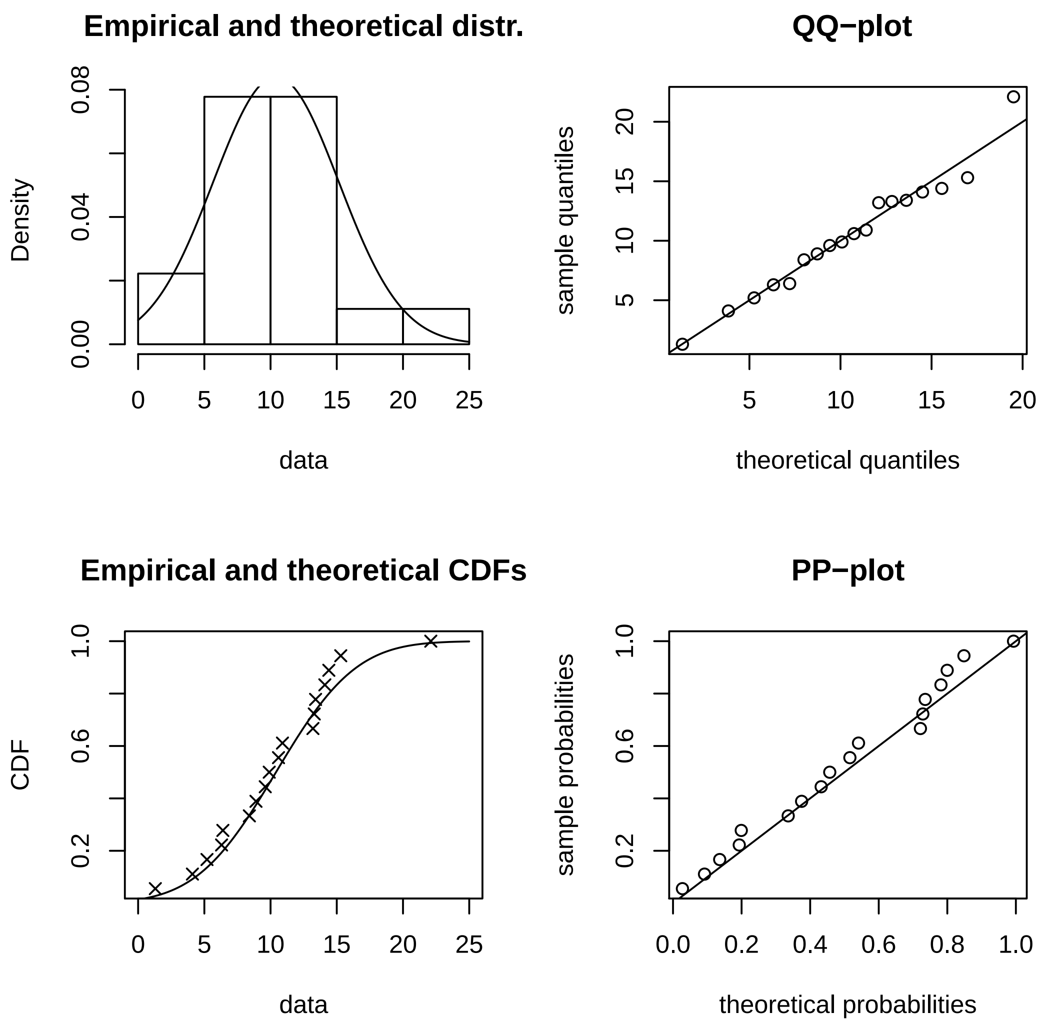
```
Kolmogorov-Smirnov statistic:  0.178
Kolmogorov-Smirnov test: not calculated

_____________ Anderson-Darling_____________
Anderson-Darling statistic:  0.793
Anderson-Darling test:  rejected
```



Below is the fit of a normal distribution.

```
> f1n <- fitdist(x1, "norm")
> plot(f1n)
> summary(f1n)

FITTING OF THE DISTRIBUTION ' norm ' BY MAXIMUM LIKELIHOOD
PARAMETERS
     estimate Std. Error
mean    10.41      1.119
sd       4.75      0.791
Loglikelihood:  -53.6
Correlation matrix:
          mean        sd
mean  1.00e+00 -1.57e-09
sd   -1.57e-09  1.00e+00

------
GOODNESS-OF-FIT STATISTICS

_____________ Chi-squared_____________
Chi-squared statistic:  4.83
Degree of freedom of the Chi-squared distribution:  3
```

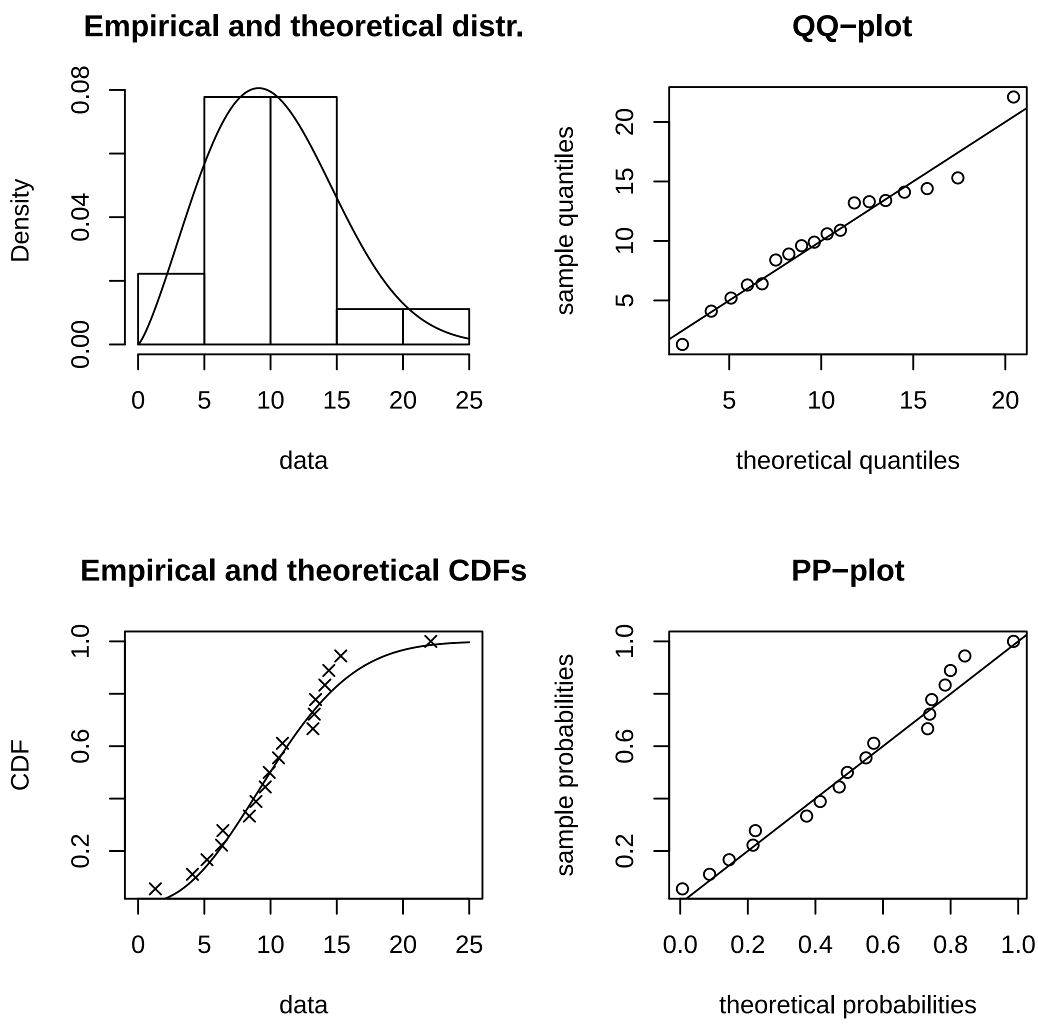
```
Chi-squared p-value:  0.185
!!! the p-value may be wrong
            with some theoretical counts < 5 !!!

!!! For continuous distributions, Kolmogorov-Smirnov and
      Anderson-Darling statistics should be prefered !!!

_____________ Kolmogorov-Smirnov_____________
Kolmogorov-Smirnov statistic:  0.110
Kolmogorov-Smirnov test: not calculated

_____________ Anderson-Darling_____________
Anderson-Darling statistic:  0.226
Anderson-Darling test:  not rejected
```



Below is the fit of a Weibull distribution.

```
> f1w <- fitdist(x1, "weibull")
> plot(f1w)
> summary(f1w)

FITTING OF THE DISTRIBUTION ' weibull ' BY MAXIMUM LIKELIHOOD
PARAMETERS
      estimate Std. Error
shape     2.29      0.426
scale    11.70      1.264
Loglikelihood:  -53.5
Correlation matrix:
      shape scale
shape   1.0   0.3
```

```
scale   0.3   1.0

------
GOODNESS-OF-FIT STATISTICS

_____________ Chi-squared_____________
Chi-squared statistic:  5.87
Degree of freedom of the Chi-squared distribution:  3
Chi-squared p-value:  0.118
!!! the p-value may be wrong
            with some theoretical counts < 5 !!!

!!! For continuous distributions, Kolmogorov-Smirnov and
      Anderson-Darling statistics should be prefered !!!

_____________ Kolmogorov-Smirnov_____________
Kolmogorov-Smirnov statistic:  0.121
Kolmogorov-Smirnov test: not calculated

_____________ Anderson-Darling_____________
Anderson-Darling statistic:  0.282
Anderson-Darling test:  not rejected
```



The values of the Anderson-Darling statistic (or another result of the fit: see the help of `fitdist` for details) for the different fittings may be extracted and compared to help the selection of a distribution :

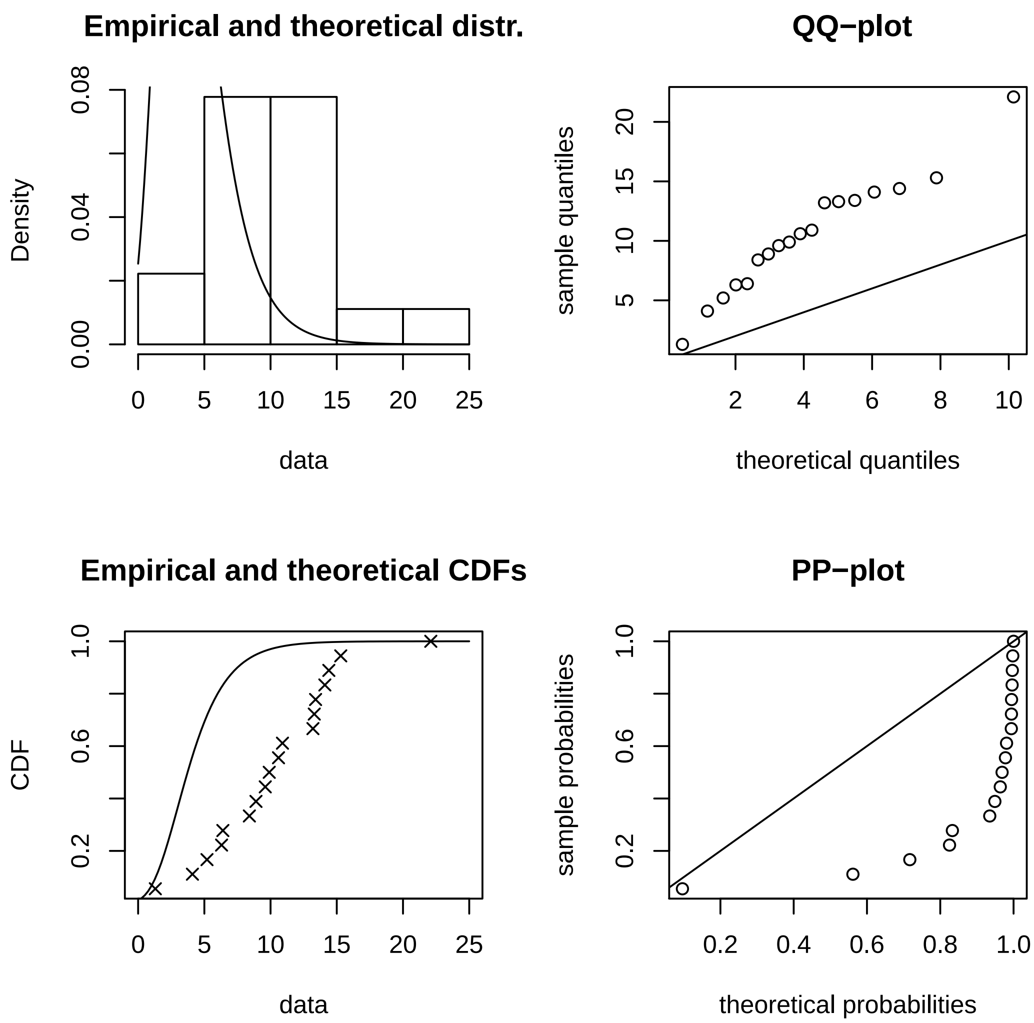
```
> anderson <- list(lnorm = f1l$ad, gamma = f1g$ad, norm = f1n$ad,
+     weibull = f1w$ad)
> anderson

$lnorm
```

```
[1] 0.7925666

$gamma
[1] 0.4567361

$norm
[1] 0.225598

$weibull
[1] 0.2821827
```

For some distributions (see the help of `fitdist` for details), it is necessary to specify initial values for the distribution parameters in the argument `start` when using the maximum likelihood method. `start` must be a named list of parameters initial values. The names of the parameters in `start` must correspond exactly to their definition in R or to their definition in a previous R code. The function `plotdist` may help to find correct initial values for the distribution parameters in non trivial cases, by an manual iterative use if necessary.

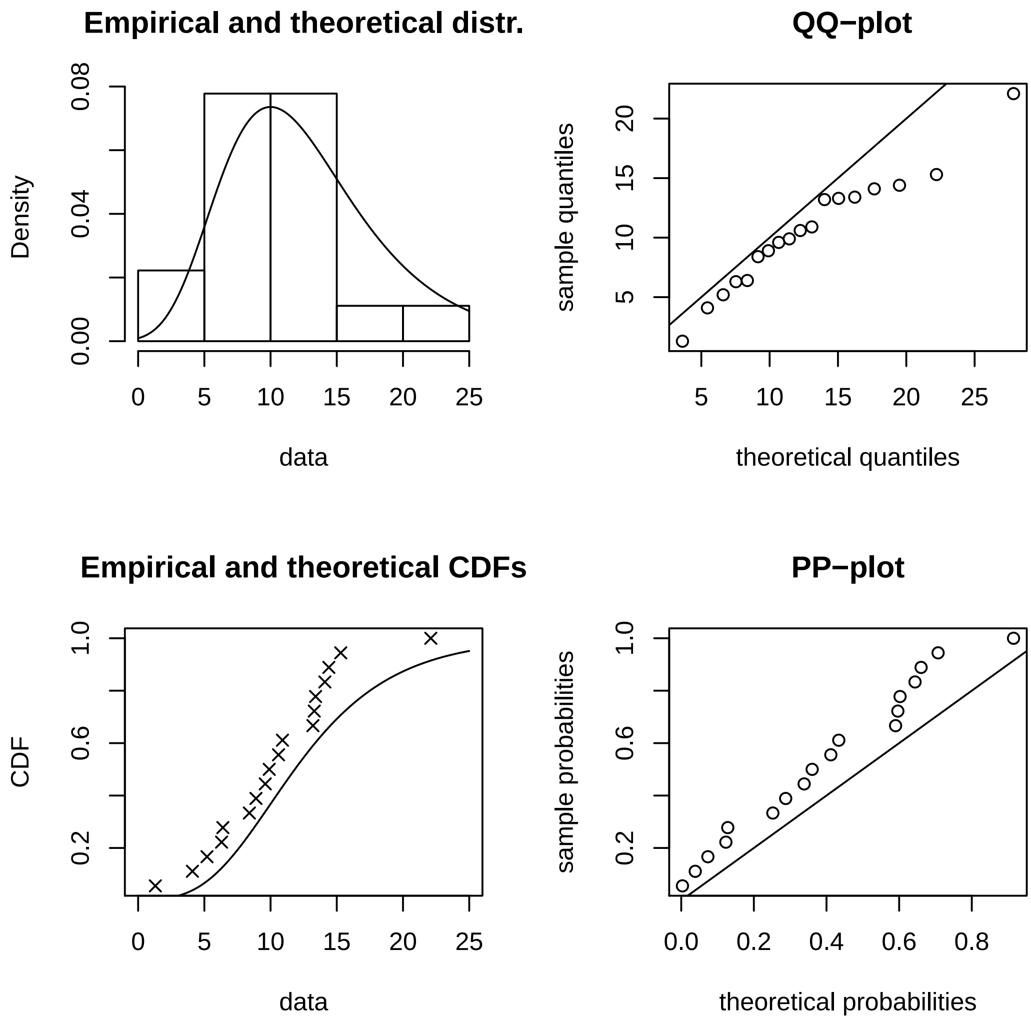
For example, below is the definition of the Gumbel distribution (also named extreme value distribution) and a first plot of the data set with the Gumbel distribution with arbitrary values for parameters.

```
> dgumbel <- function(x, a, b) 1/b * exp((a - x)/b) * exp(-exp((a -
+     x)/b))
> pgumbel <- function(q, a, b) exp(-exp((a - q)/b))
> qgumbel <- function(p, a, b) a - b * log(-log(p))
> plotdist(x1, "gumbel", para = list(a = 3, b = 2))
```



The same data set may be plotted with a Gumbel distribution with modified values for parameters.

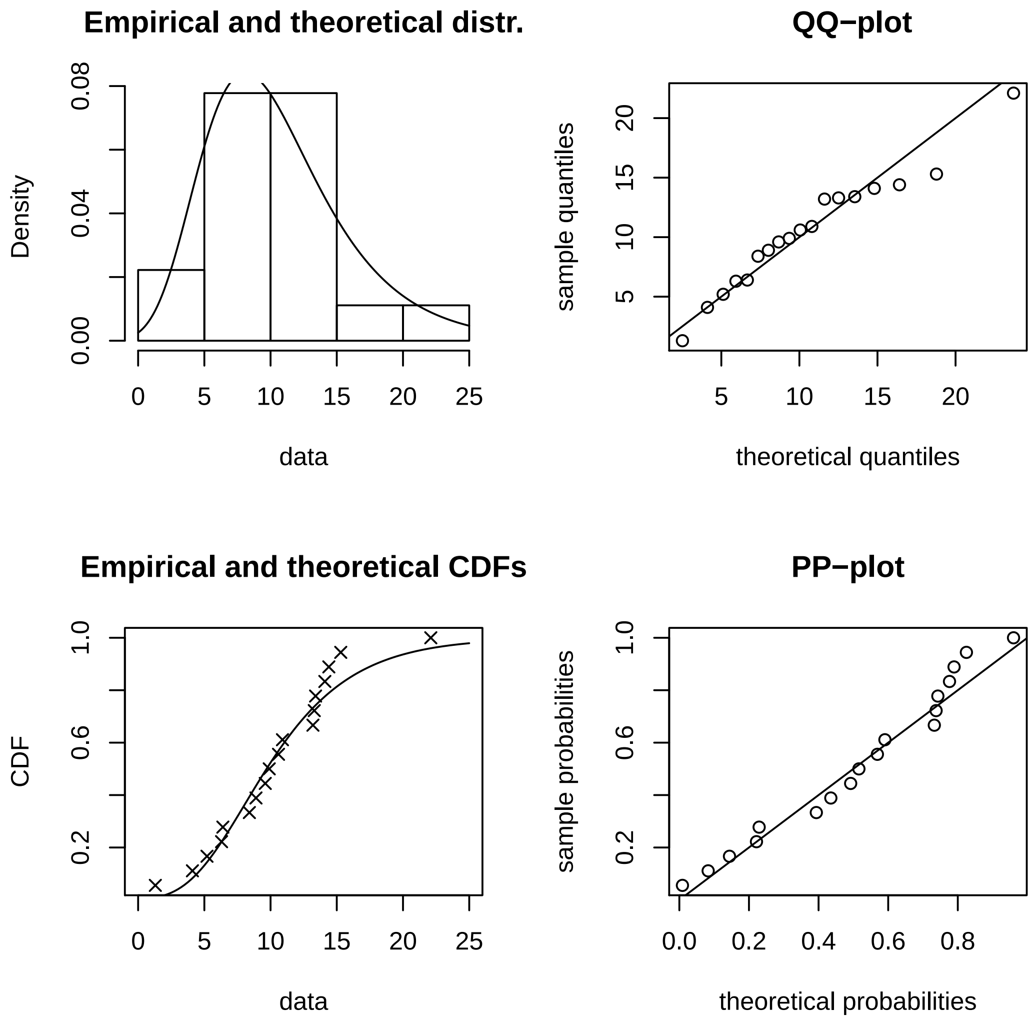
```
> plotdist(x1, "gumbel", para = list(a = 10, b = 5))
```



And a Gumbel distribution may be fitted to data with these values for initial parameter values.

```
> fgu <- fitdist(x1, "gumbel", start = list(a = 10, b = 5))
> plot(fgu)
> summary(fgu)

FITTING OF THE DISTRIBUTION ' gumbel ' BY MAXIMUM LIKELIHOOD
PARAMETERS
  estimate Std. Error
a     8.09      1.092
b     4.38      0.766
Loglikelihood:  -54.1
Correlation matrix:
      a     b
a 1.000 0.330
b 0.330 1.000

------
GOODNESS-OF-FIT STATISTICS

_____________ Chi-squared_____________
Chi-squared statistic:  7.56
Degree of freedom of the Chi-squared distribution:  3
Chi-squared p-value:  0.056
!!! the p-value may be wrong
            with some theoretical counts < 5 !!!

!!! For continuous distributions, Kolmogorov-Smirnov and
      Anderson-Darling statistics should be prefered !!!
```

```
_____________ Kolmogorov-Smirnov_____________
Kolmogorov-Smirnov statistic:  0.121
Kolmogorov-Smirnov test: not calculated

_____________ Anderson-Darling_____________
Anderson-Darling statistic:  0.34
Anderson-Darling test: not calculated
```

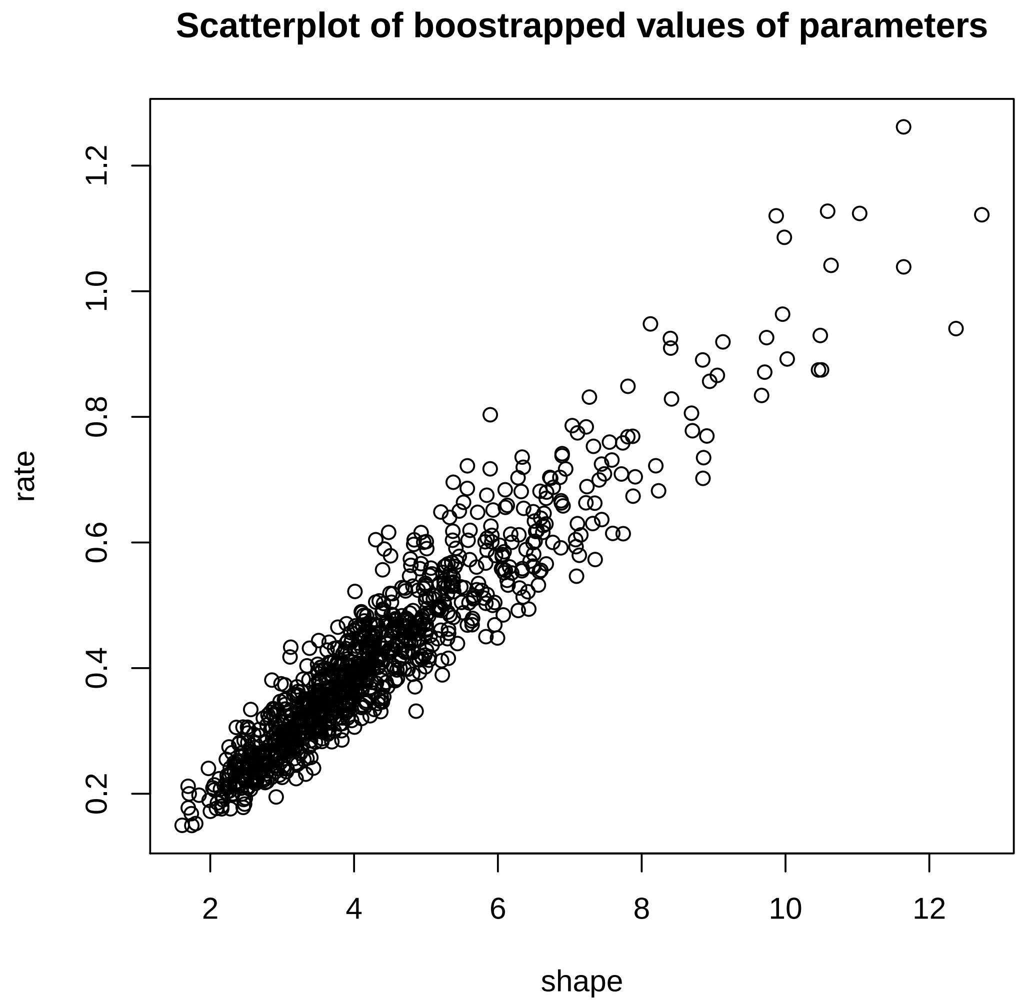


1.4 Simulation of the uncertainty by boostrap

The uncertainty in the parameters of the fitted distribution may be simulated by parametric or nonparametric boostrap using the function `boodist`. This function returns the boostrapped values of parameters which may be plotted to visualize the bootstrap region. It also calculates the 95 percent confidence intervals for each parameter from the 2.5 and 97.5 percentiles of the boostrap values of each parameter (see the help of the function `bootdist` for details).

Below is an example of the use of this function with the previous fit of the gamma distribution.

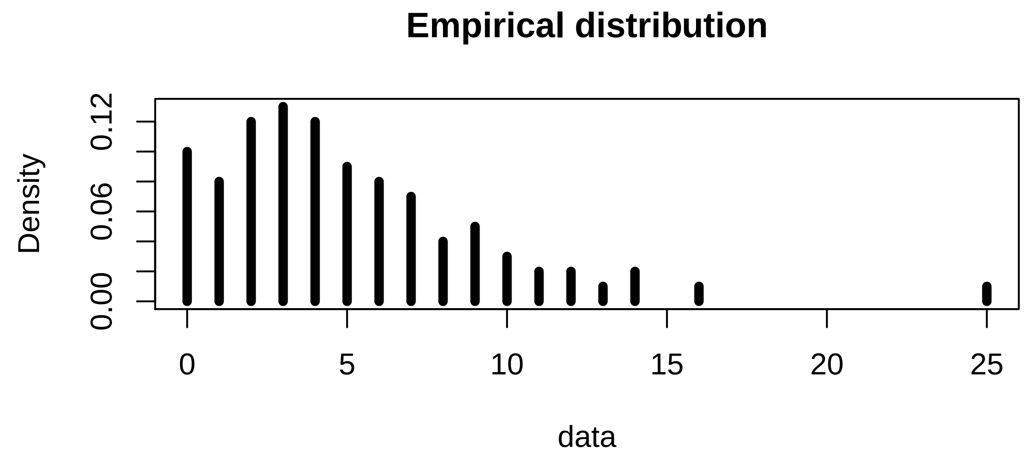
```
> b1g <- bootdist(f1g)
> plot(b1g)
> summary(b1g)

Parametric bootstrap medians and 95% CI
      Median  2.5% 97.5%
shape  3.987 2.117  8.27
rate   0.381 0.197  0.82

Maximum likelihood method converged for  999  among  999  iterations
```

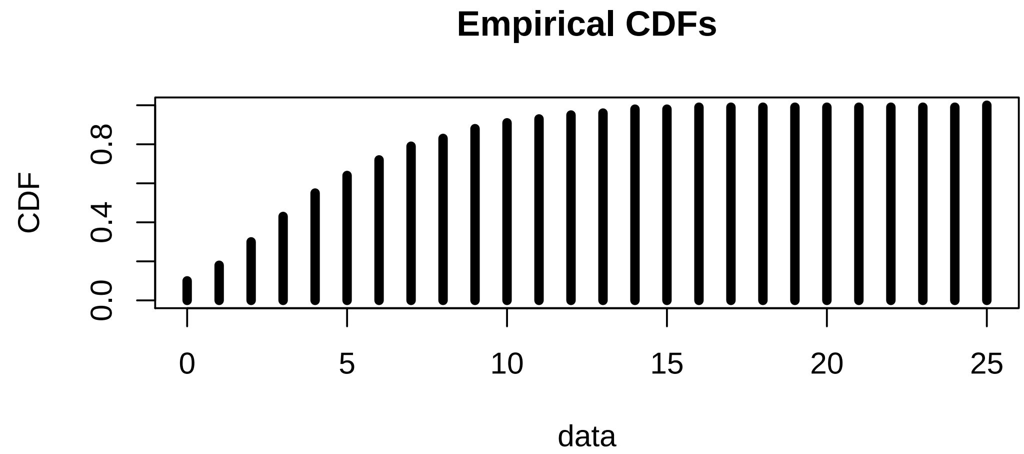


2 Specification of a distribution from non-censored discrete data

A discrete data set may be considered as a continuous one for example for a large data set from a binomial distribution converging to a normal one. A discrete plot of the distribution may also be provided, fixing the argument discrete of the function `plotdist` to `TRUE`.

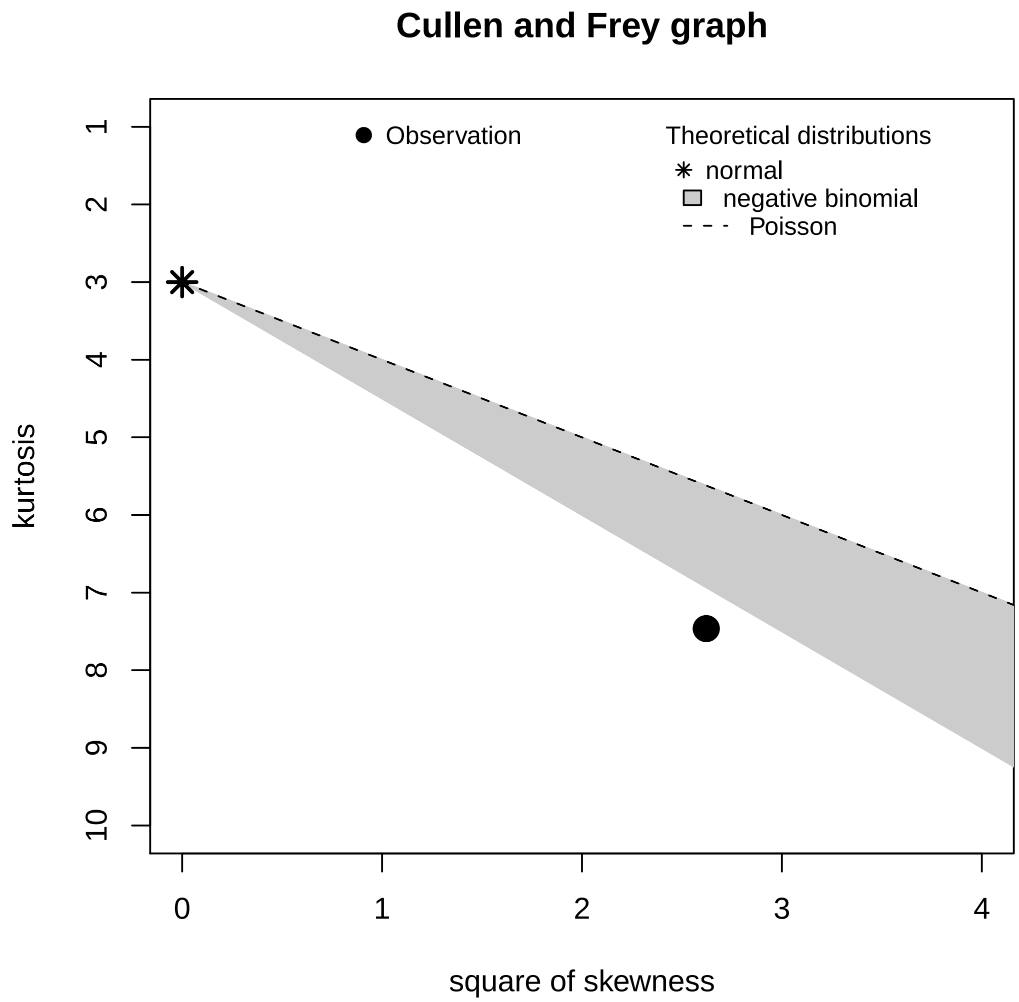
```
> x2 <- rnbinom(n = 100, size = 2, prob = 0.3)
> plotdist(x2, discrete = TRUE)
```





As for continuous distributions, descriptive parameters of the empirical distribution may be computed using the function `descdist` which also provides a skewness-kurtosis plot which may help you to choose which distribution(s) to fit.

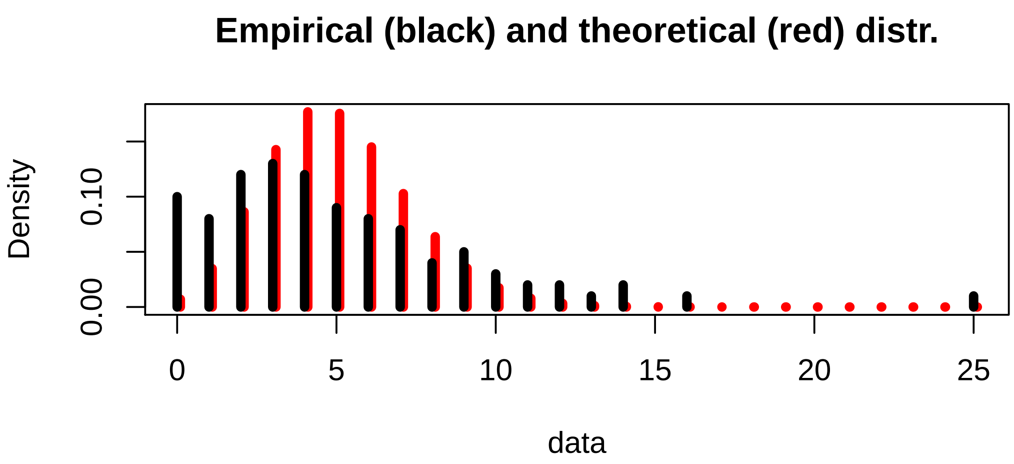
```
> descdist(x2, discrete = T)

summary statistics
------
min:  0   max:  25
median:  4
mean:  4.96
sample sd:  4.12
sample skewness:  1.62
sample kurtosis:  7.46
```

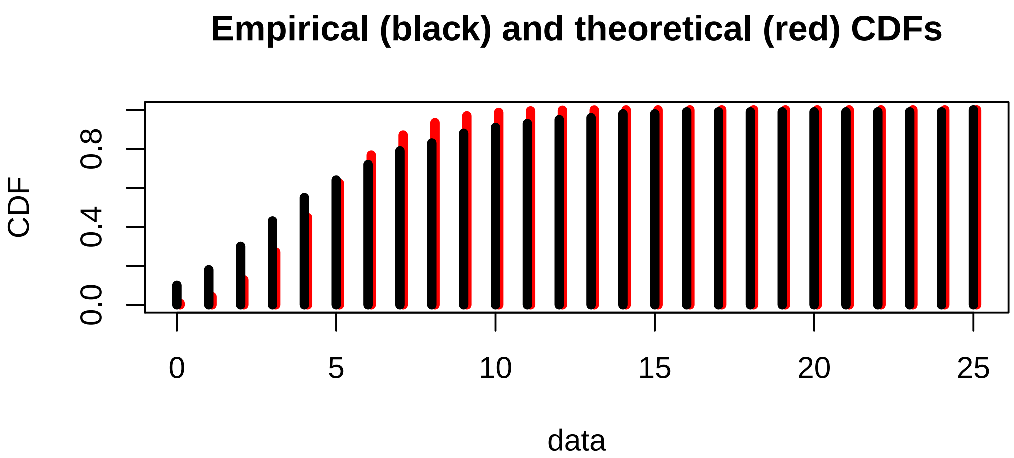


As for continuous distributions, one or more parametric distributions may then be fitted to the data set by maximum likelihood or matching moments.

Below is the result of the fit of a Poisson distribution with the bootstrap simulations.

```
> f2p <- fitdist(x2, "pois")
> plot(f2p)
> summary(f2p)

FITTING OF THE DISTRIBUTION ' pois ' BY MAXIMUM LIKELIHOOD
PARAMETERS
       estimate Std. Error
lambda     4.96      0.223
Loglikelihood:  -315
------
GOODNESS-OF-FIT STATISTICS

_____________ Chi-squared_____________
Chi-squared statistic:  165
Degree of freedom of the Chi-squared distribution:  6
Chi-squared p-value:  5.45e-33
!!! the p-value may be wrong
            with some theoretical counts < 5 !!!

> b2p <- bootdist(f2p)
> summary(b2p)

Parametric bootstrap medians and 95% CI
Median   2.5%  97.5%
  4.96   4.53   5.41

Maximum likelihood method converged for  999  among  999  iterations
```

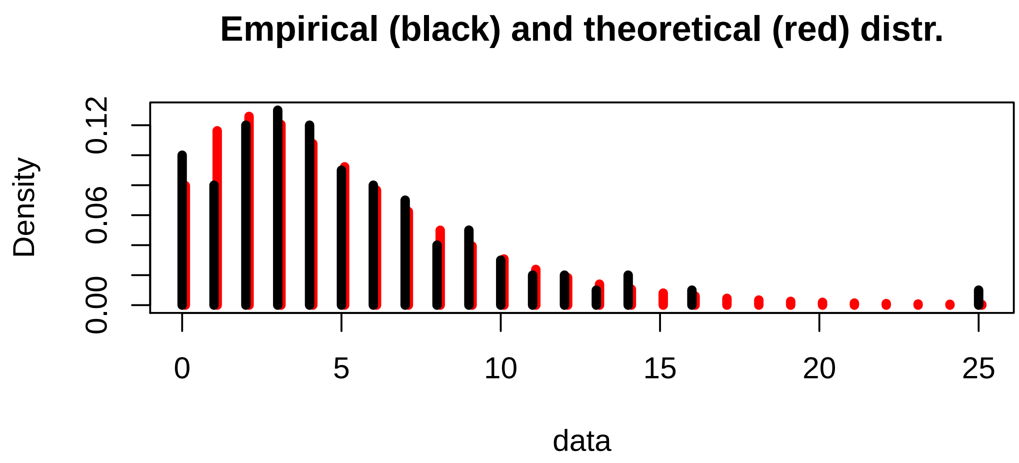


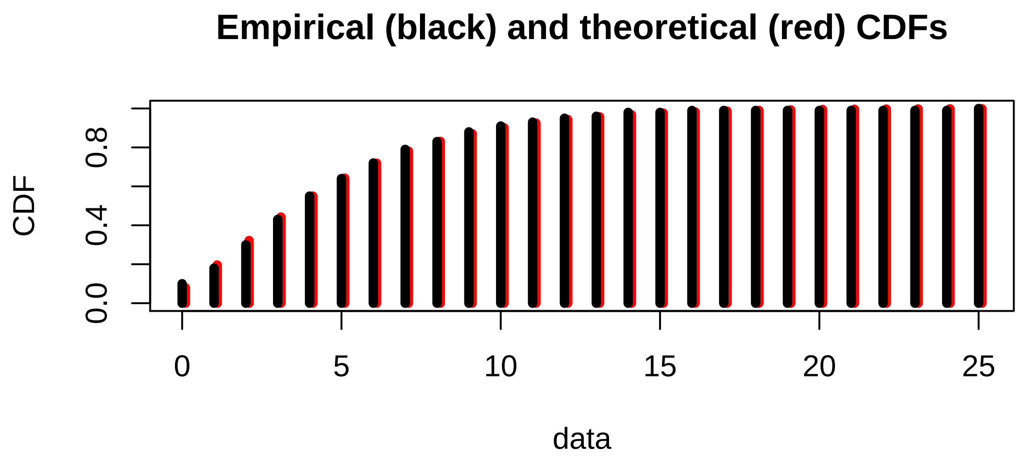


Below is the result of the fit of a negative binomial distribution with the boostrap simulations.

```
> f2n <- fitdist(x2, "nbinom")
> plot(f2n)
> summary(f2n)

FITTING OF THE DISTRIBUTION ' nbinom ' BY MAXIMUM LIKELIHOOD
PARAMETERS
     estimate Std. Error
size     2.07      0.441
mu       4.96      0.411
Loglikelihood:  -263
Correlation matrix:
         size       mu
size 1.00e+00 8.77e-05
mu   8.77e-05 1.00e+00

------
GOODNESS-OF-FIT STATISTICS

_____________ Chi-squared_____________
Chi-squared statistic:  1.63
Degree of freedom of the Chi-squared distribution:  5
Chi-squared p-value:  0.898

> b2n <- bootdist(f2n)
> summary(b2n)

Parametric bootstrap medians and 95% CI
     Median 2.5% 97.5%
size   2.10 1.43  3.37
```

```
mu     4.95 4.20  5.76

Maximum likelihood method converged for  999  among  999  iterations
```





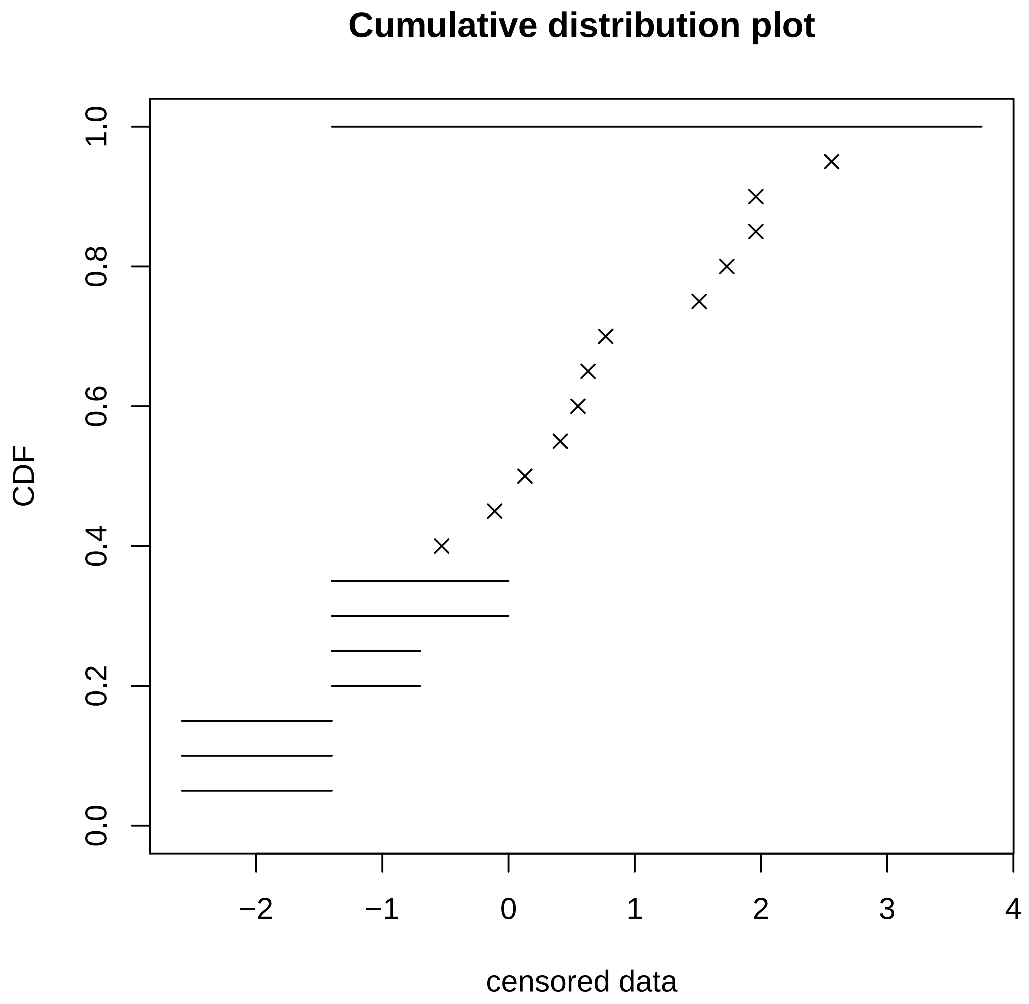
3 Specification of a distribution from censored data

Censored data may contain left censored, right censored and interval censored values, with several lower and upper bounds. 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.

3.1 Graphical display of the observed distribution

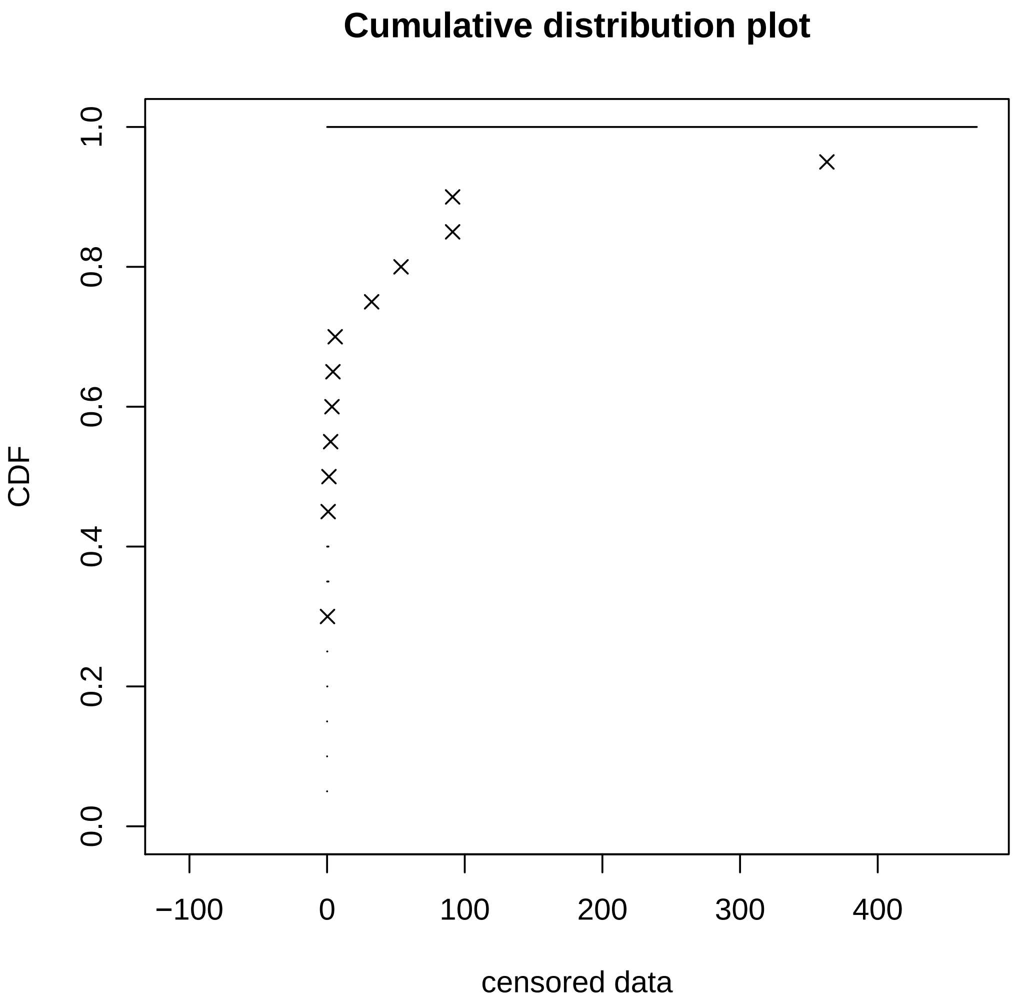
First of all, the observed distribution may be plotted using the function `plotdistcens`. Data are reported directly as segments for interval, left and right censored data, and as points for non-censored data. For more details, see the help of the function `plotdistcens`.

```
> d1 <- data.frame(left = c(1.73, 1.51, 0.77, 1.96, 1.96, -1.4,
+     -1.4, NA, -0.11, 0.55, 0.41, 2.56, NA, -0.53, 0.63, -1.4,
+     -1.4, -1.4, NA, 0.13), right = c(1.73, 1.51, 0.77, 1.96,
+     1.96, 0, -0.7, -1.4, -0.11, 0.55, 0.41, 2.56, -1.4, -0.53,
+     0.63, 0, -0.7, NA, -1.4, 0.13))
> plotdistcens(d1)
```



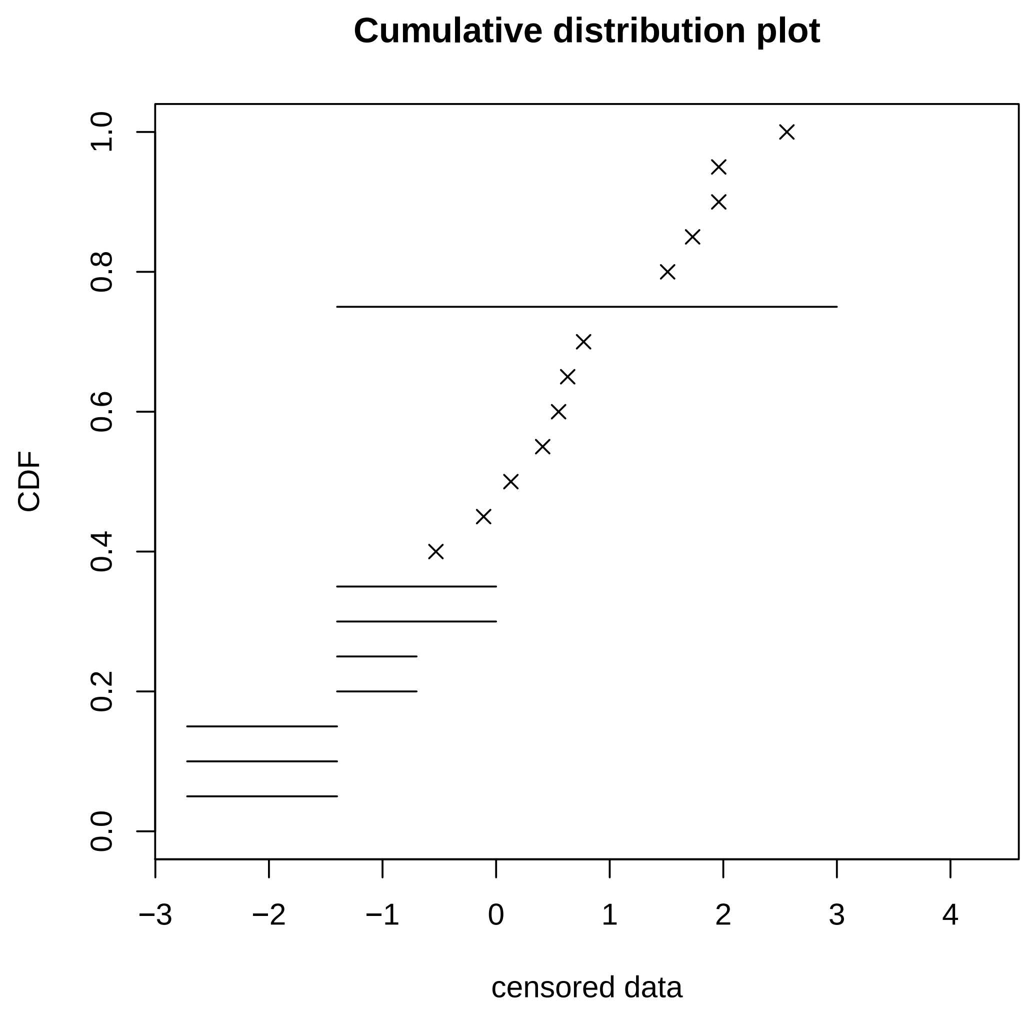
When left or right NA-values correspond to finite value (for example 0 for left NA-values of positive data), the arguments `leftNA` (or `rightNA`) must be affected to this finite value to ensure a correct plot of left (or right) censored observations, as in the example below.

```
> d2 <- data.frame(left = 10^(d1$left), right = 10^(d1$right))
> plotdistcens(d2, leftNA = 0)
```



It is also possible to fix `rightNA` or `leftNA` to a realistic extreme value, even if not exactly known, to obtain a reasonable global ranking of observations, as in the example below for the first dataset.

```
> plotdistcens(d1, rightNA = 3)
```

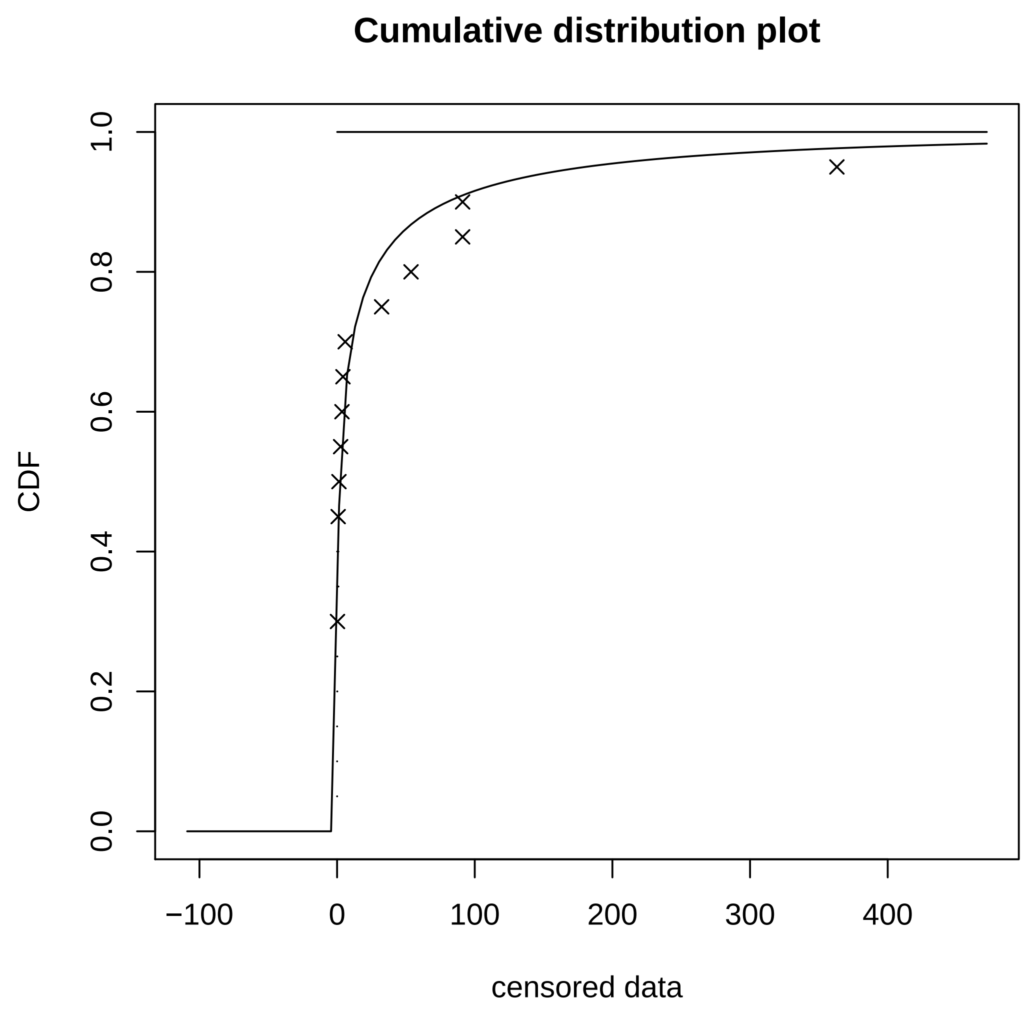


3.2 Fitting of a distribution

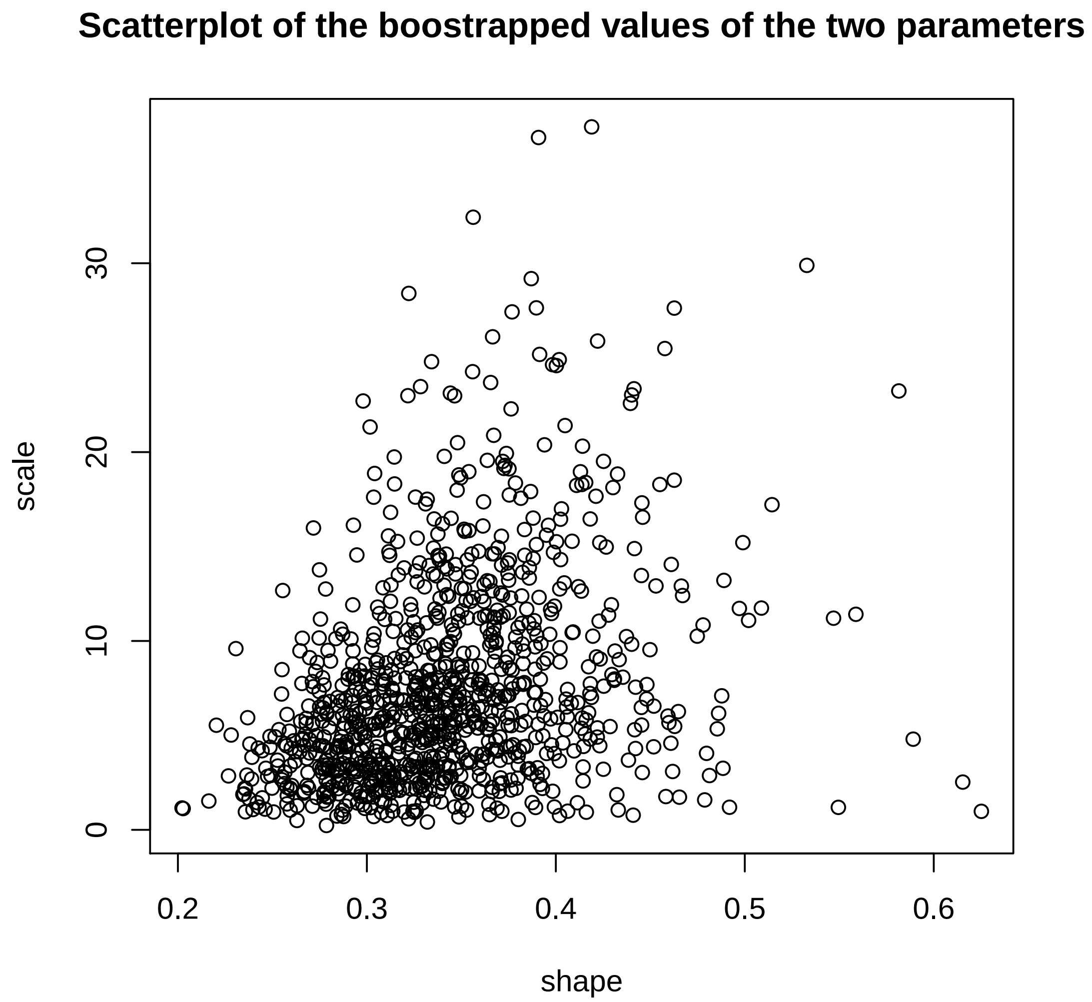
One or more parametric distributions may then be fitted to the censored data set, one at a time, using the function `fitdistcens`. This function always uses the maximum likelihood method. For more details, see the help of the function `fitdistcens`. Only one goodness of fit plot is provided for censored data, in cumulative frequencies. The uncertainty in the parameters of the fitted distribution may be simulated by nonparametric boostrap only, using the function `boodistcens`.

Below is the result of a fit of a Weibull distribution by maximum likelihood and the results of the corresponding boostrap simulations.

```
> f2w <- fitdistcens(d2, "weibull")
> summary(f2w)

FITTING OF THE DISTRIBUTION ' weibull ' BY MAXIMUM LIKELIHOOD ON CENSORED DATA
PARAMETERS
      estimate Std. Error
shape    0.324     0.0613
scale    6.124     4.5872
Loglikelihood:  -68.5
Correlation matrix:
      shape scale
shape 1.000 0.326
scale 0.326 1.000

> plot(f2w, leftNA = 0)
```



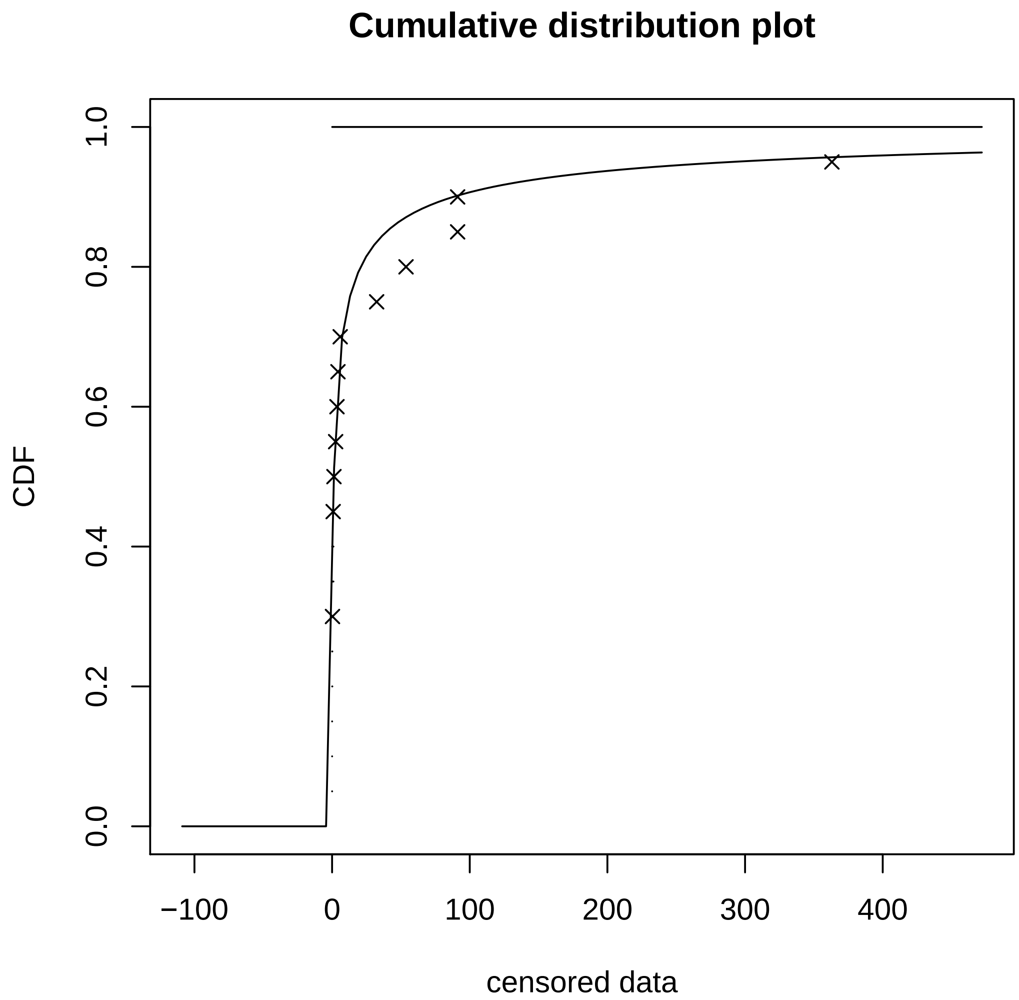

```
> b2w <- bootdistcens(f2w)
> summary(b2w)

Nonparametric bootstrap medians and 95% CI
      Median 2.5%  97.5%
shape  0.338 0.25  0.469
scale  5.970 1.25 23.212

Maximum likelihood method converged for  999  among  999  iterations

> plot(b2w)
```



Goodness of fit statistics are not computed for fit on censored data, so the quality of fit may only be estimated from the loglikelihood and the goodness of fit plot.

Below is the fit of a lognormal distribution to the same censored data set.

```
> f2l <- fitdistcens(d2, "lnorm")
> summary(f2l)

FITTING OF THE DISTRIBUTION ' lnorm ' BY MAXIMUM LIKELIHOOD ON CENSORED DATA
PARAMETERS
        estimate Std. Error
meanlog     0.27      0.764
sdlog       3.28      0.600
Loglikelihood:  -68.7
Correlation matrix:
        meanlog   sdlog
meanlog  1.0000 -0.0739
sdlog   -0.0739  1.0000

> plot(f2l, leftNA = 0)
```

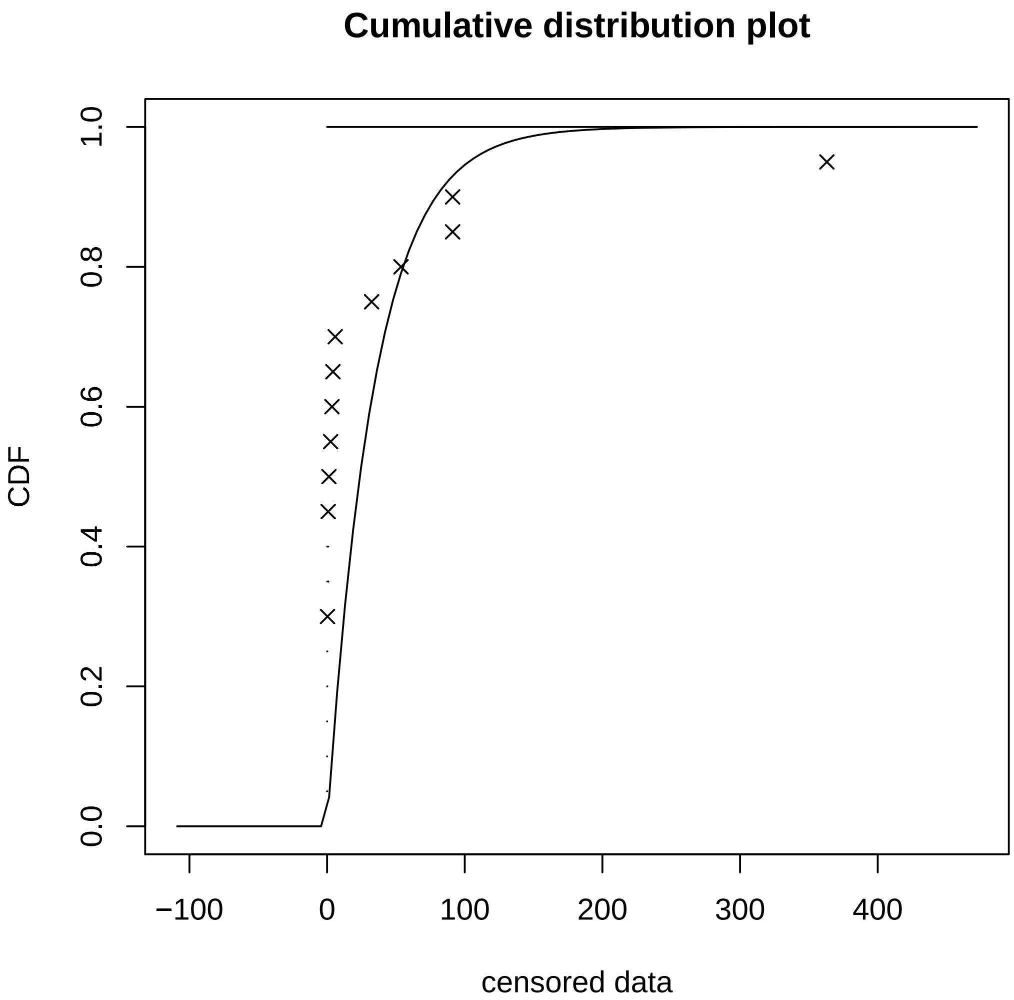


Below is the fit of an exponential distribution.

```
> f2e <- fitdistcens(d2, "exp")
> summary(f2e)
```

```
FITTING OF THE DISTRIBUTION ' exp ' BY MAXIMUM LIKELIHOOD ON CENSORED DATA
PARAMETERS
     estimate Std. Error
rate   0.0292    0.00668
Loglikelihood:  -99.6
```

```
> plot(f2e, leftNA = 0)
```



As with `fitdist`, for some distributions (see the help of `fitdistcens` for details), it is necessary to specify initial values for the distribution parameters in the argument `start`. `start` must be a named list of parameters initial values. The names of the parameters in `start` must correspond exactly to their definition in R or to their definition in a previous R code. The function `plotdistcens` may help to find correct initial values for the distribution parameters in non trivial cases, by an manual iterative use if necessary, as explained previously for non-censored continuous data.

References

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994), *Continuous univariate distributions*, John Wiley. 4