New goodness-of-fit plots for censored data in the package **fitdistrplus**

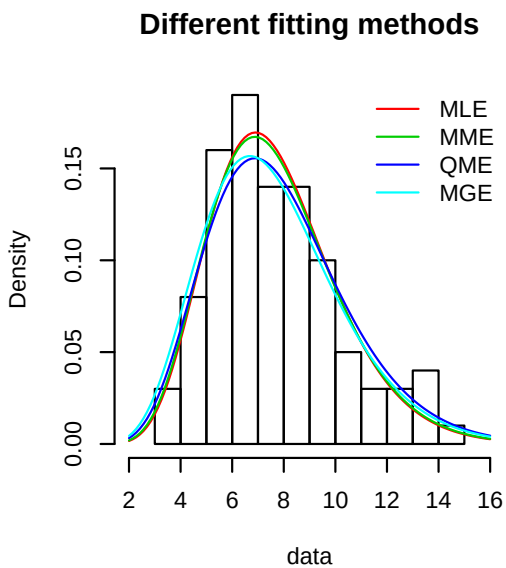
M.L. Delignette-Muller (LBBE, Lyon)
C. Dutang (CEREMADE, Paris) and
A. Siberchicot (LBBE, Lyon)

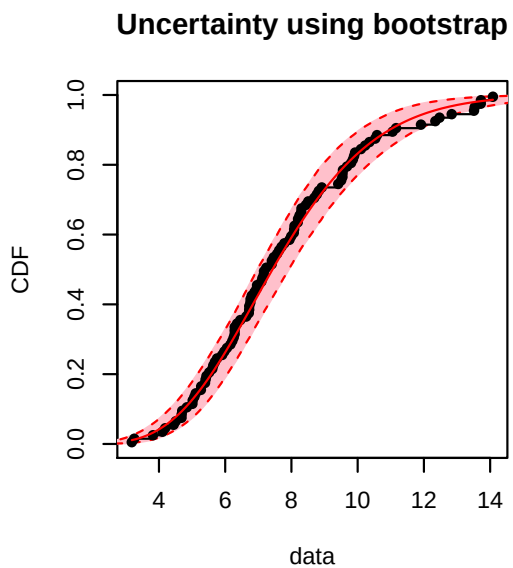
July 6th 2018

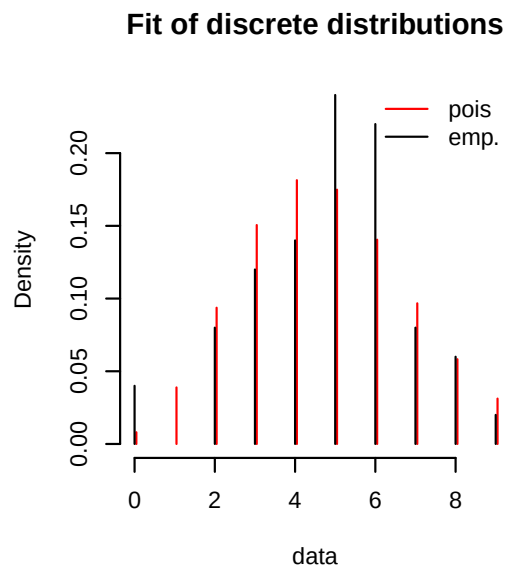
General presentation of the package fitdistrplus

A package to help the **fit of parametric distributions** to univariate discrete our continuous non censored or censored data.

- stable version 1.0-9 on **CRAN** (first release in 2009).
- version 1.0-10 in development on **Rforge** (soon on CRAN).
- Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. **Journal of Statistical Software**, 64(4), 1-34. (311 citations in scholar google)
- A FAQ vignette continuously updated in each new version.







Goodness-of-fit plots for non censored data

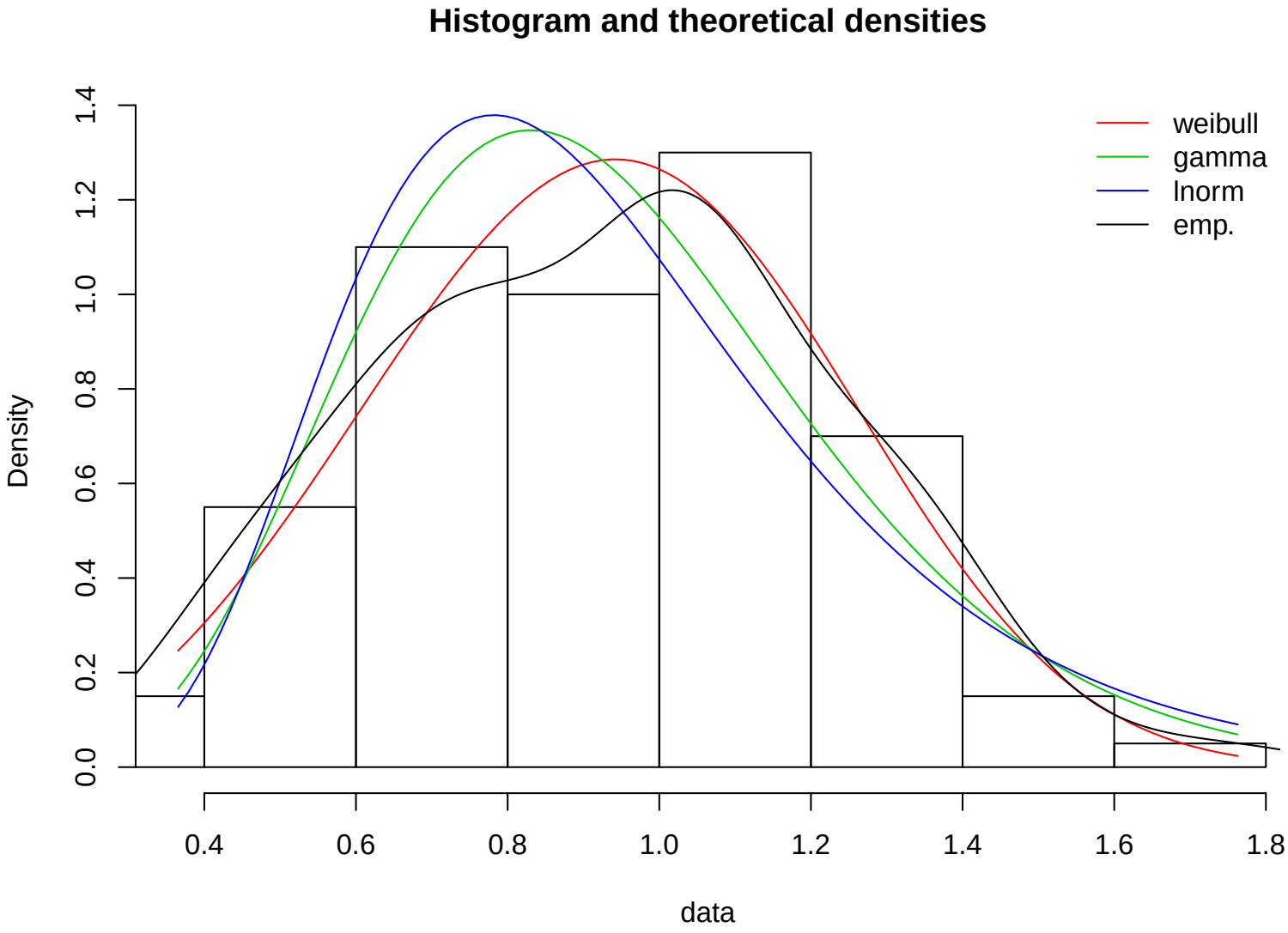
An example with non censored data

```
r <- rweibull(100, shape = 3, scale = 1)
fw <- fitdist(r, "weibull")
fg <- fitdist(r, "gamma")
fl <- fitdist(r, "lnorm")
gofstat(list(fw, fg, fl),
        fitnames = c("Weibull", "gamma", "lnorm"))
```

```
## Goodness-of-fit statistics
##                                 Weibull gamma lnorm
## Kolmogorov-Smirnov statistic     0.0598 0.104 0.121
## Cramer-von Mises statistic       0.0356 0.114 0.192
## Anderson-Darling statistic       0.2288 0.654 1.136
##
## Goodness-of-fit criteria
##                                   Weibull gamma lnorm
## Akaike's Information Criterion       43.7  46.5  51.8
## Bayesian Information Criterion       48.9  51.7  57.0
```

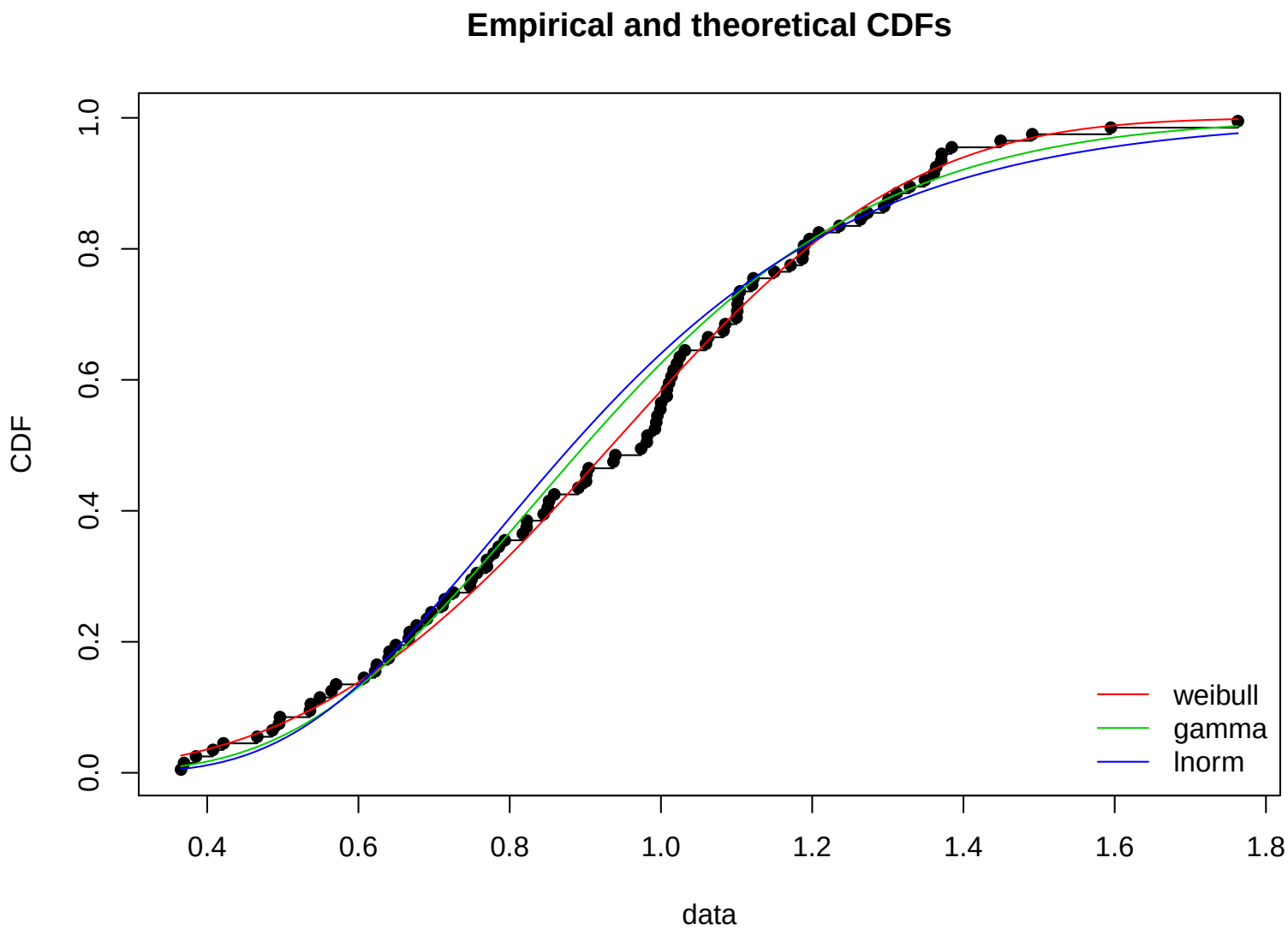
A goodness-of-fit plot in density plot

```
denscomp(list(fw, fg, fl), demp = TRUE, fitlty = 1)
```



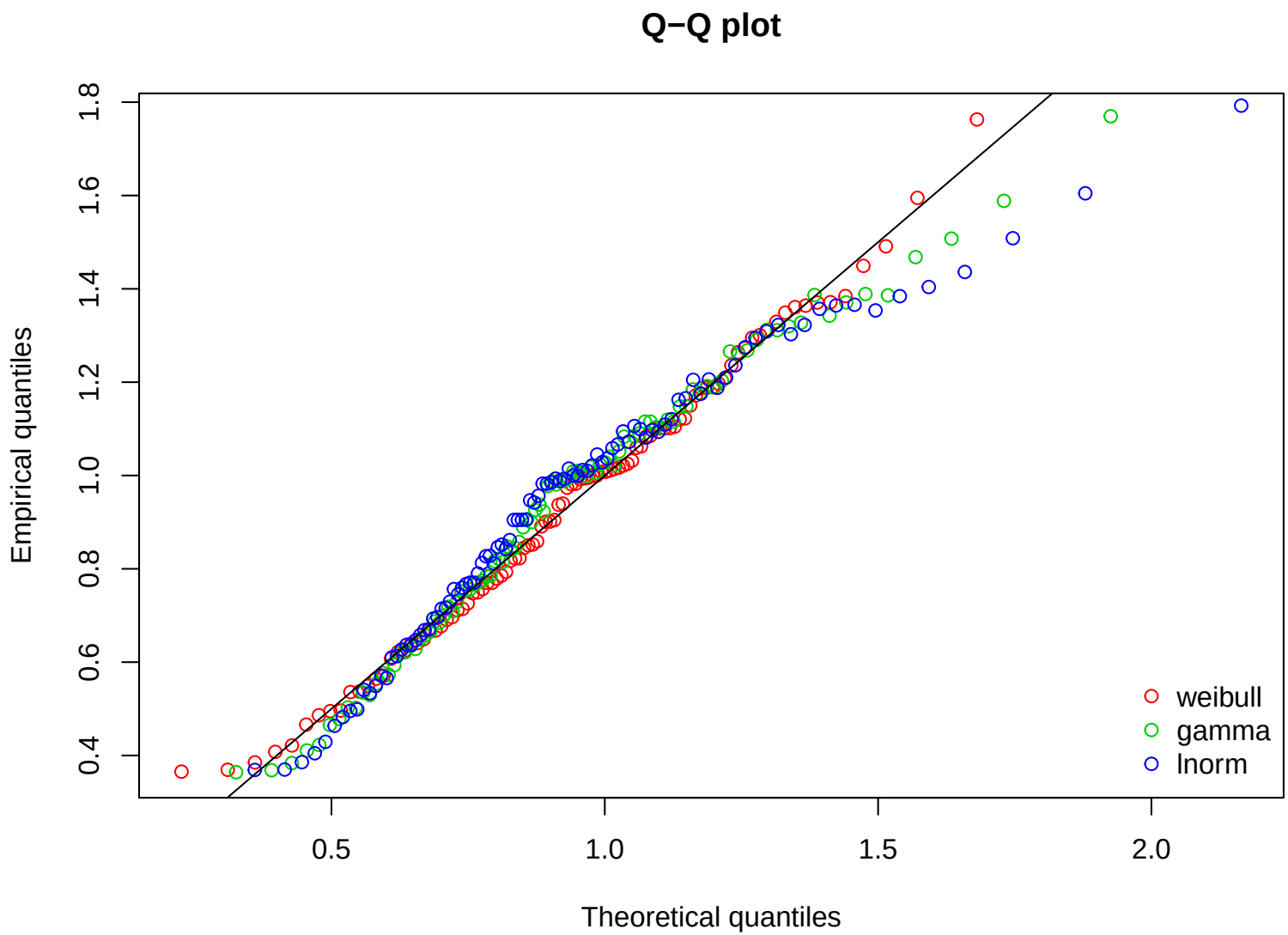
A goodness-of-fit plot in CDF

```
cdfcomp(list(fw, fg, fl), fitlty = 1)
```



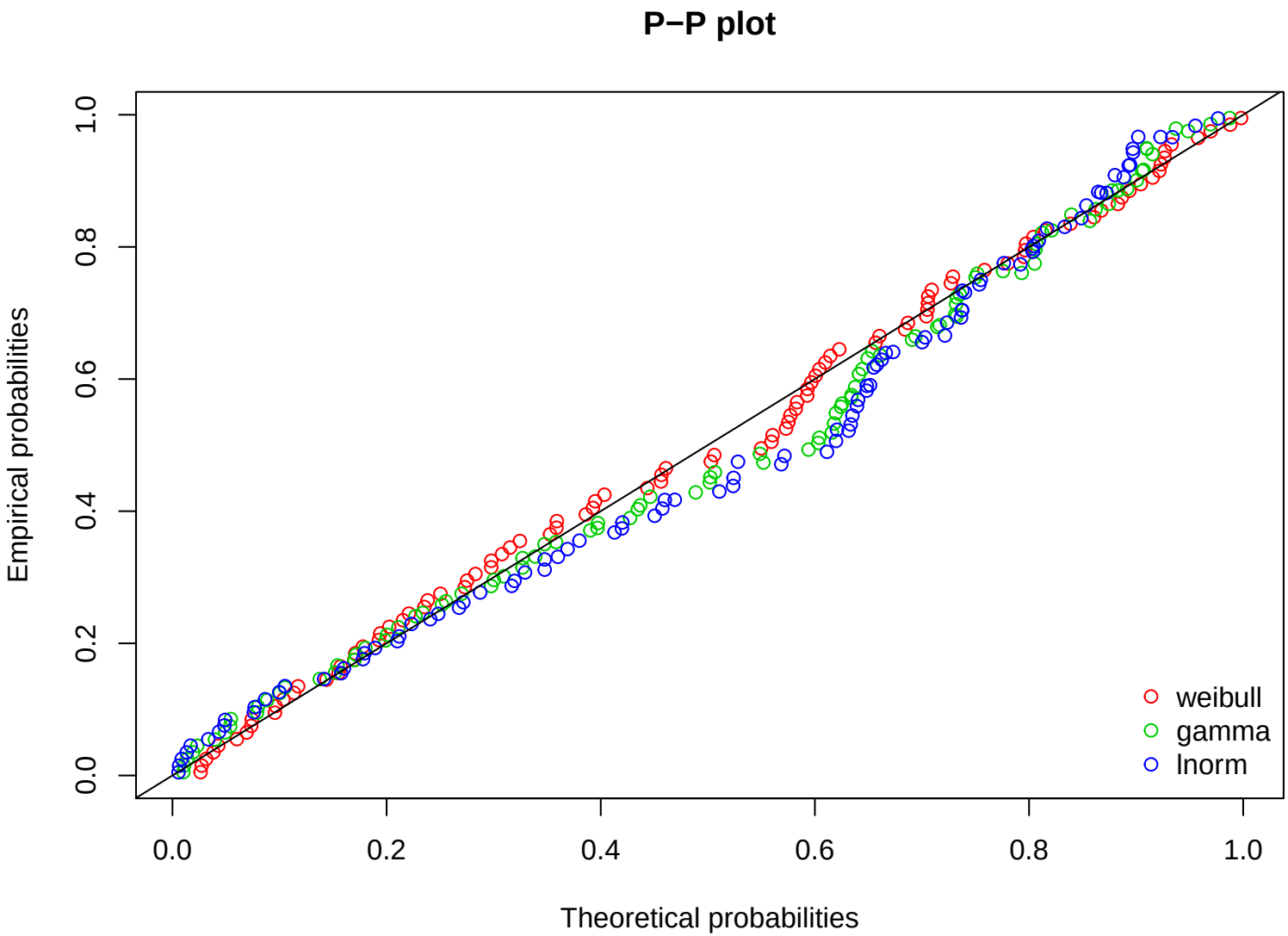
A Q-Q plot which emphasizes differences at tails

```
qqcomp(list(fw, fg, fl))
```



a P-P plot which emphasizes differences in the center

```
ppcomp(list(fw, fg, fl))
```



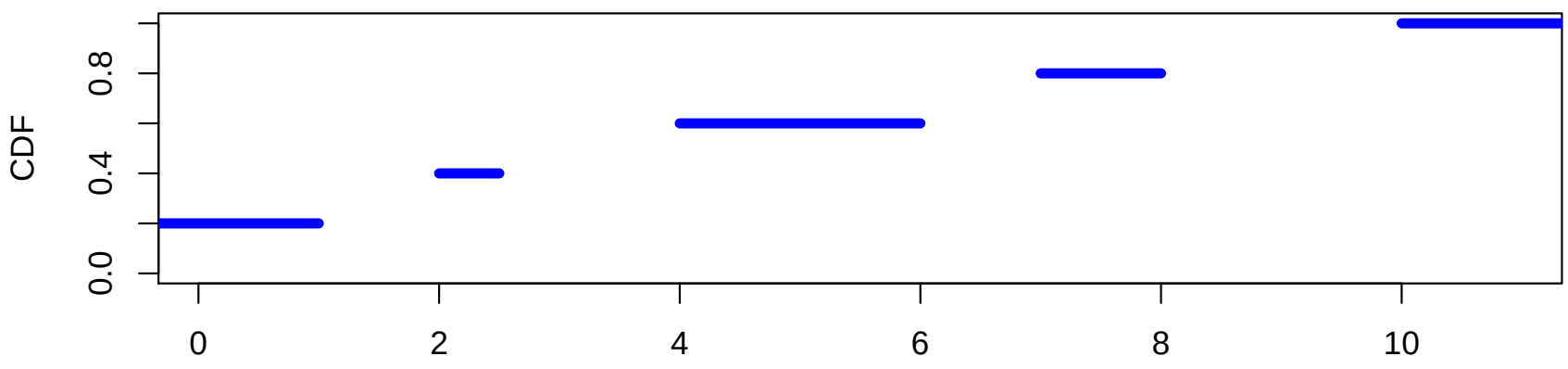
Representation of the ECDF for censored data

How to represent an ECDF from censored data ?

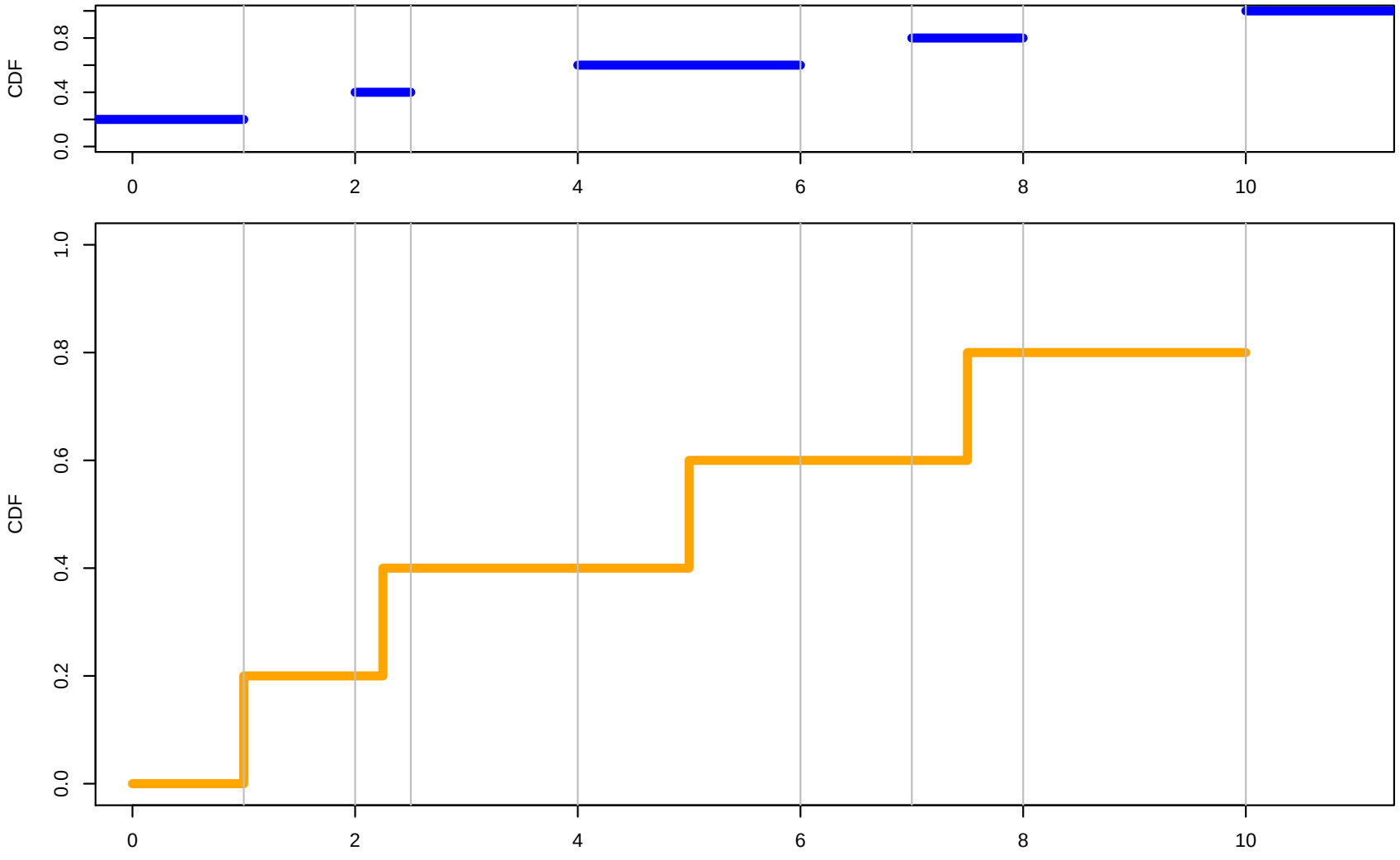
A first toy example with left, right and interval censored data

```
d
```

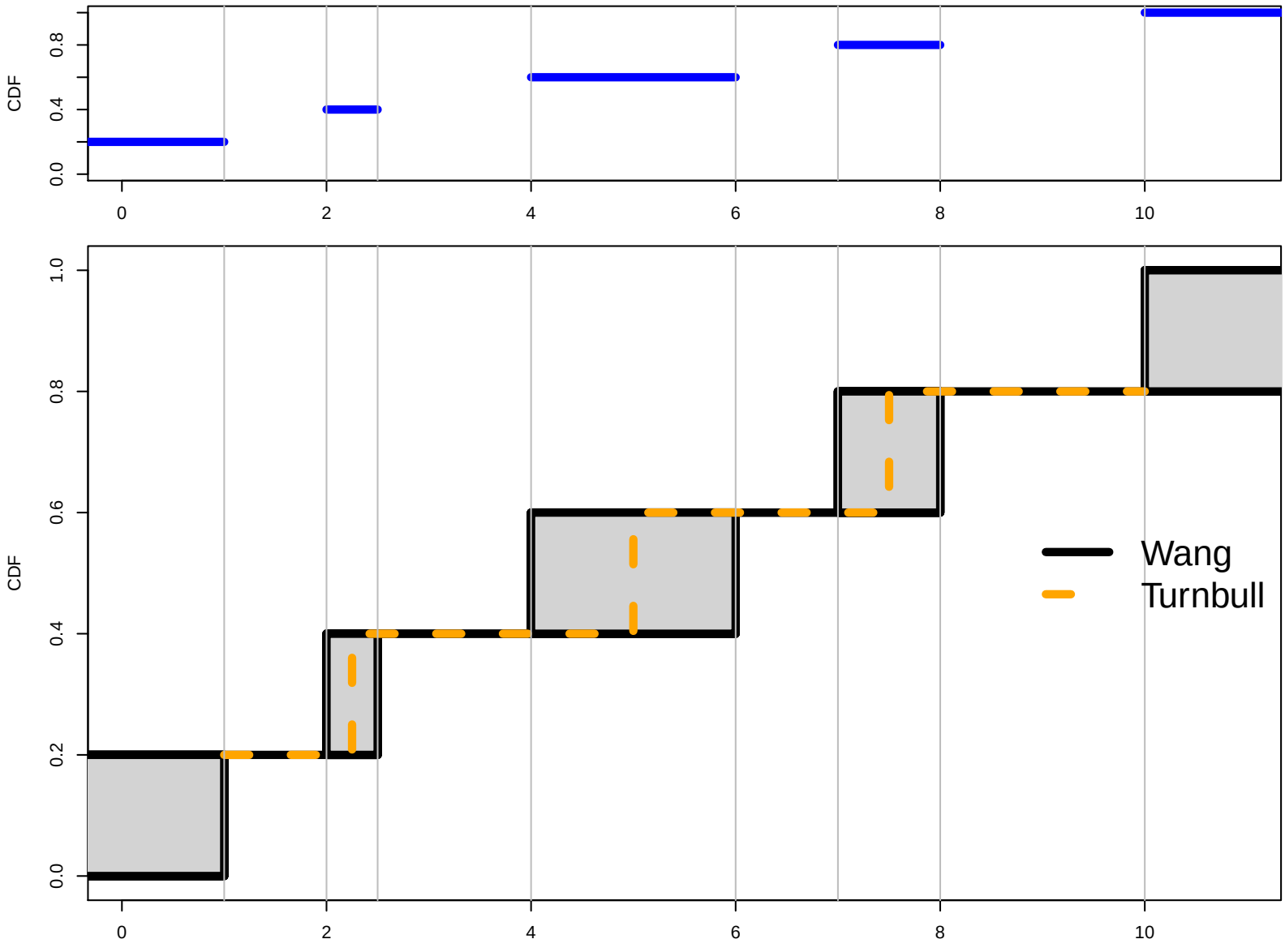
```
##   left right
## 1   NA   1.0
## 2    2   2.5
## 3    4   6.0
## 4    7   8.0
## 5   10    NA
```



Non Parametric Maximum Likelihood Estimation (NPMLE) of the ECDF: the Turnbull plot (package survival) used in former versions of fitdistrplus.



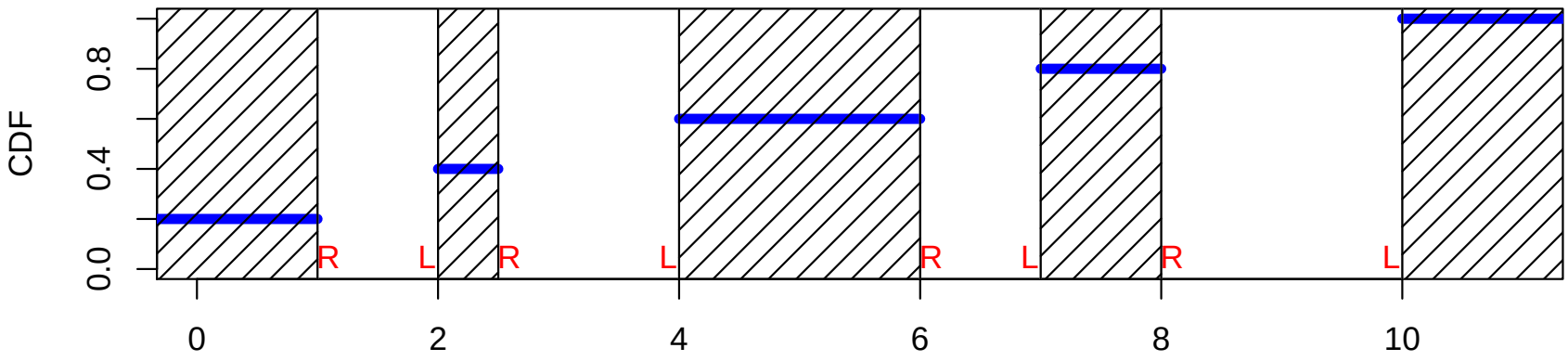
A new algorithm and plot from the package npsurv (Wang)



The two steps of an NPMLE algorithm

1. Identification of **equivalence classes** (also named **Turnbull intervals** or **maximal intersection intervals** or **innermost intervals** or **maximal cliques** of the data) = set of points/intervals under which the ECDF may change (each region between a left bound L immediately followed by a right bound R, even if of null length). The NPMLE is unique only up to these equivalence classes (**non uniqueness** represented by **grey rectangles**).

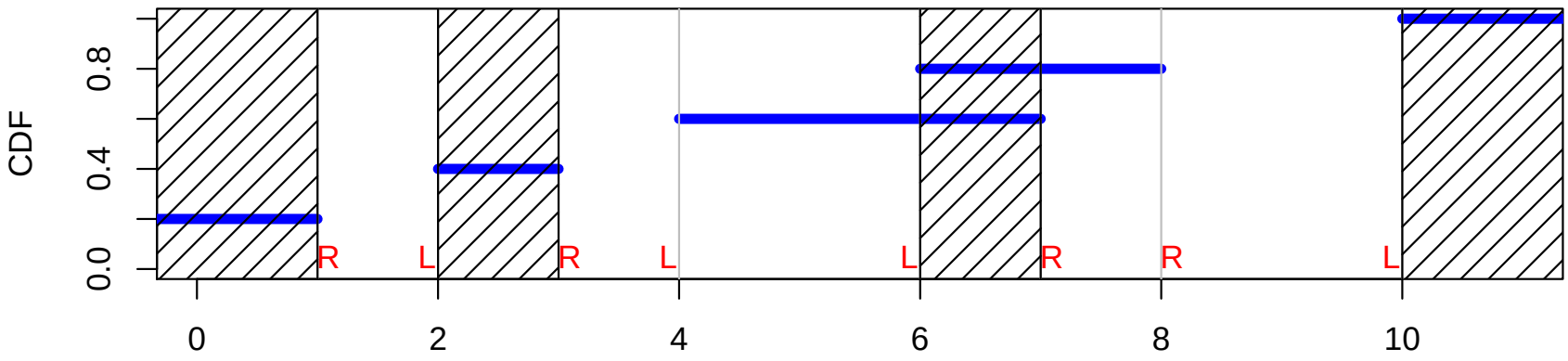
Equivalence classes on the first toy example



The two steps of an NPMLE algorithm

1. Identification of **equivalence classes** (also named **Turnbull intervals** or **maximal intersection intervals** or **innermost intervals** or **maximal cliques** of the data) = set of points/intervals under which the ECDF may change (each region between a left bound L immediately followed by a right bound R, even if of null length). The NPMLE is unique only up to these equivalence classes (**non uniqueness** represented by **grey rectangles**).

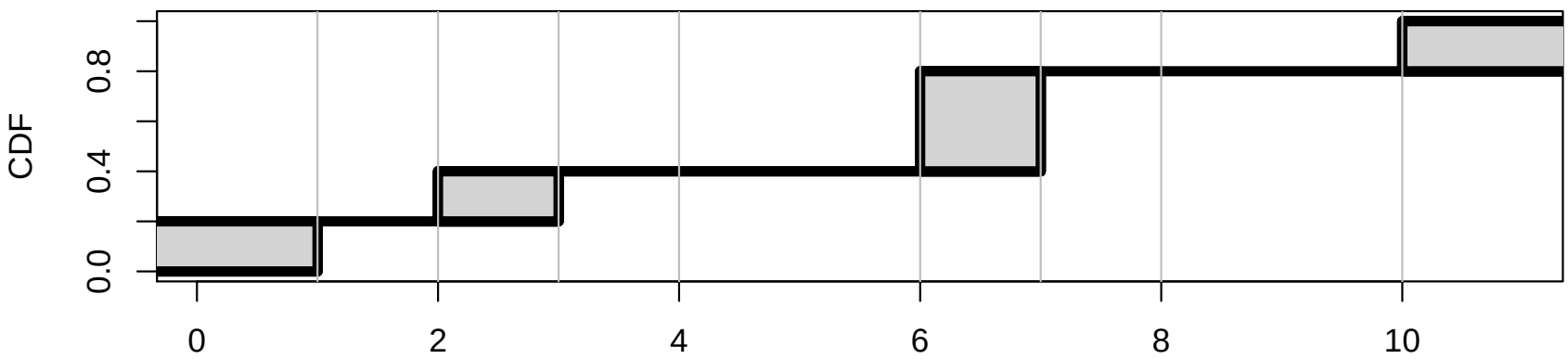
Equivalence classes on a second toy example



The two steps of an NPMLE algorithm

1. Identification of **equivalence classes** (also named **Turnbull intervals** or **maximal intersection intervals** or **innermost intervals** or **maximal cliques** of the data) = set of points/intervals under which the ECDF may change (each region between a left bound L immediately followed by a right bound R, even if of null length). The NPMLE is unique only up to these equivalence classes (**non uniqueness** represented by **grey rectangles**).
2. Assign a **probability mass** to each equivalence class (may be 0).

The Wang plot on the second toy example

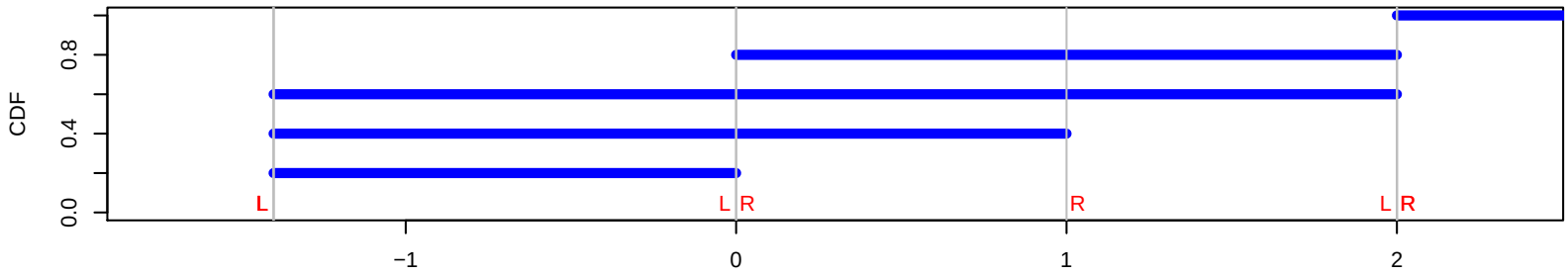


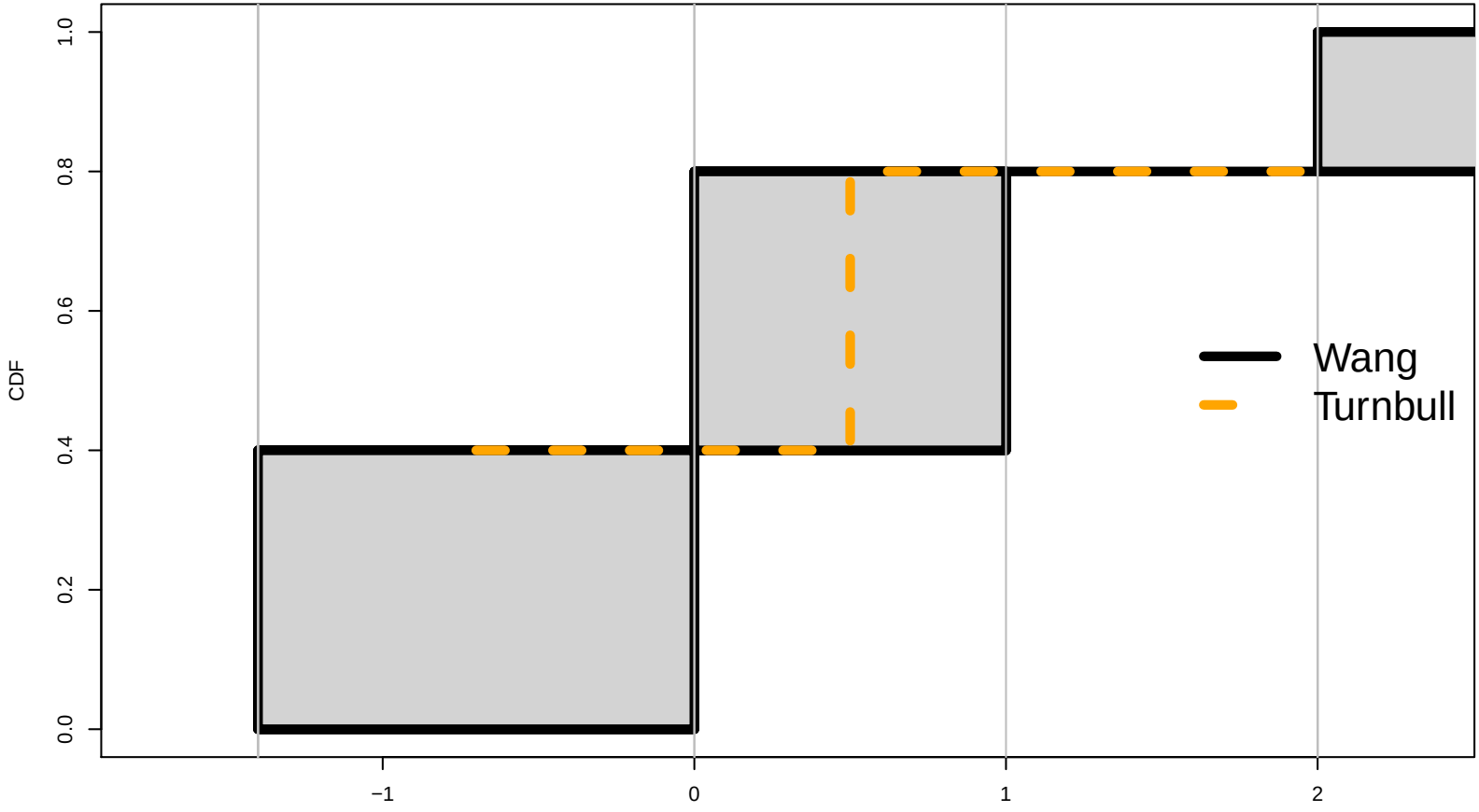
The two steps of an NPMLE algorithm

1. Identification of **equivalence classes** (also named **Turnbull intervals** or **maximal intersection intervals** or **innermost intervals** or **maximal cliques** of the data) = set of points/intervals under which the ECDF may change (each region between a left bound L immediately followed by a right bound R, even if of null length). The NPMLE is unique only up to these equivalence classes (**non uniqueness** represented by **grey rectangles**).
2. Assign a **probability mass** to each equivalence class (may be 0).

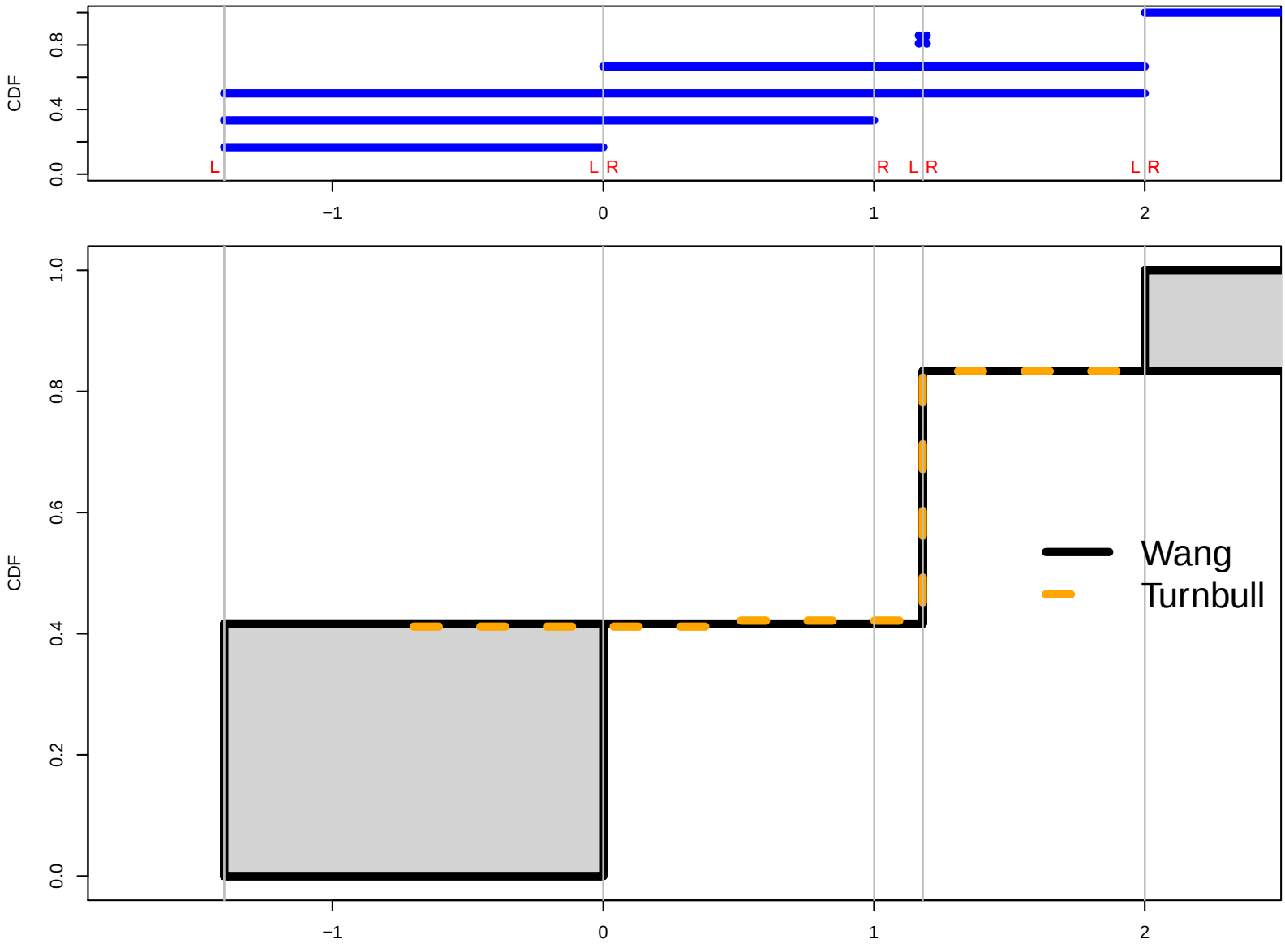
Various algorithms implemented in the packages **Icens**, **interval** and **npsurv** (more or less performant and not all handling left censored data).

A third toy example

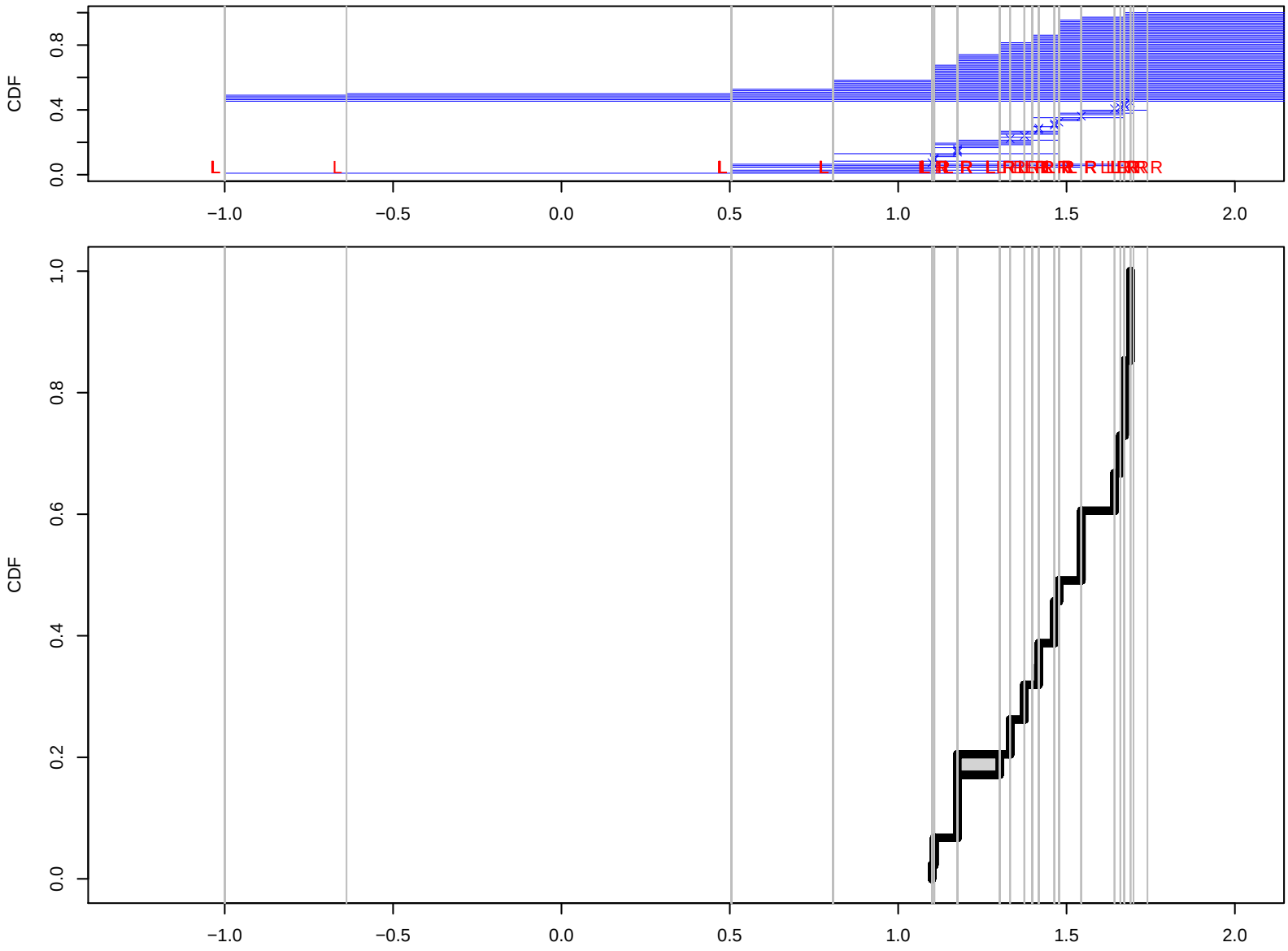




The third toy example with the add of a non censored obs.

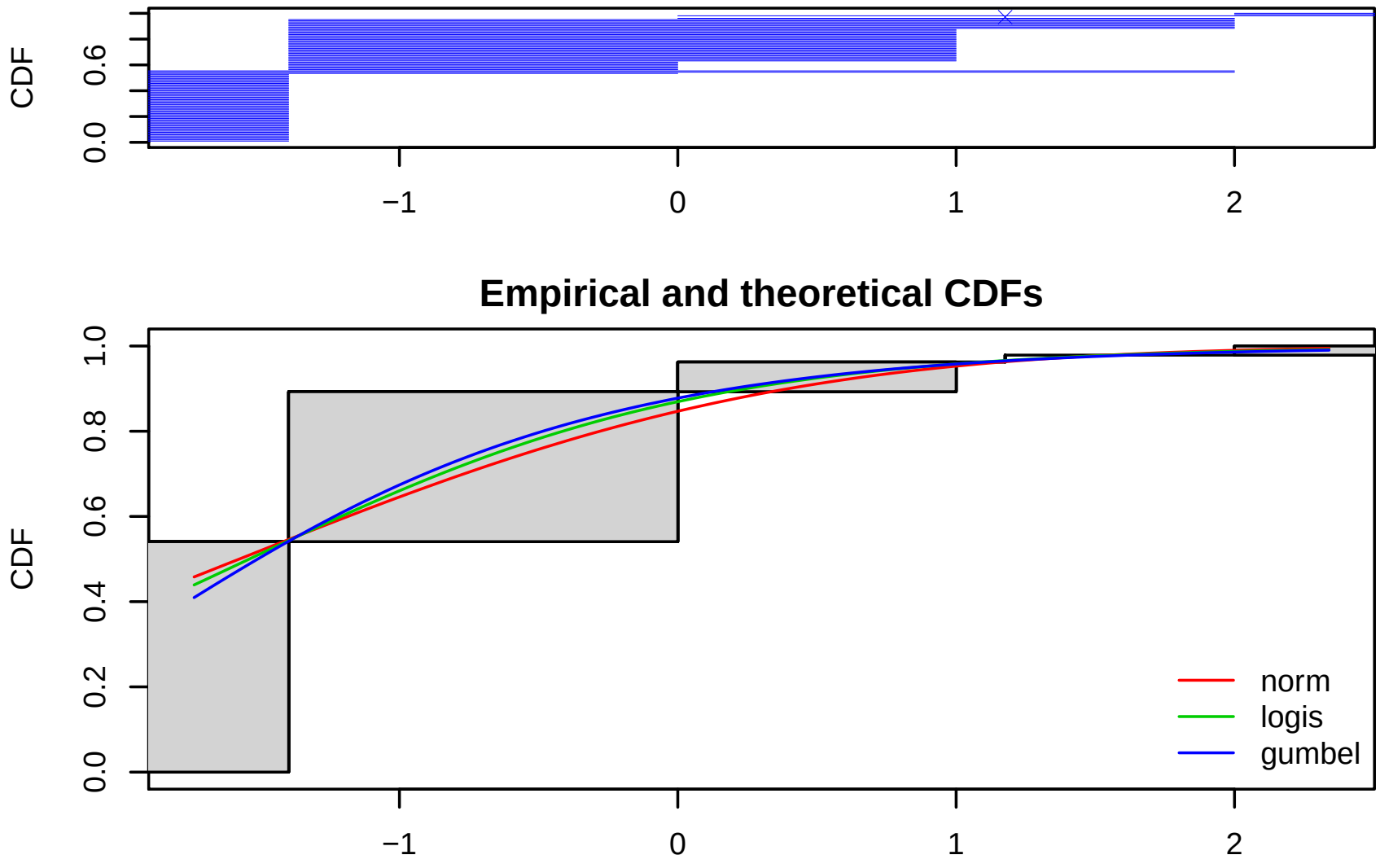


A realistic example: data salinity

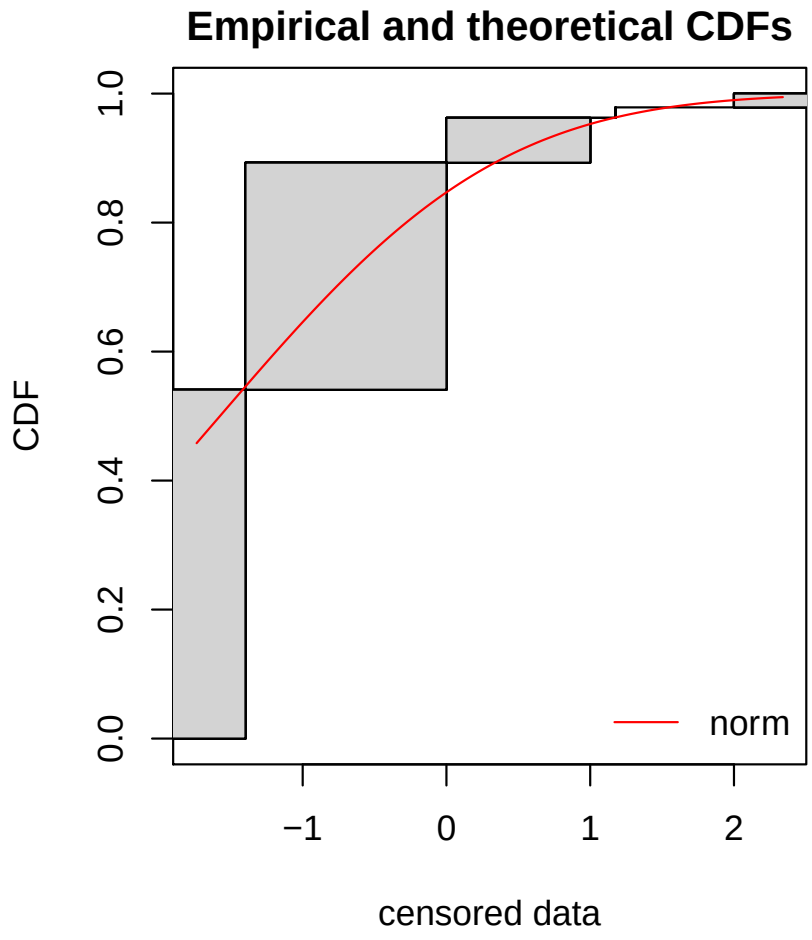


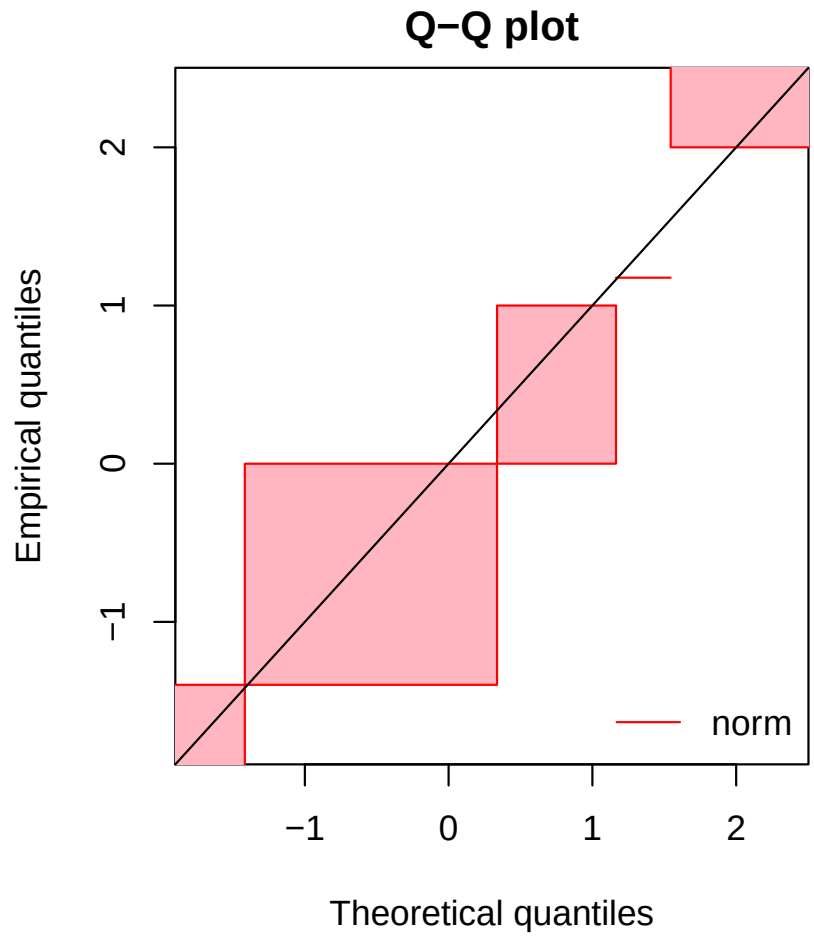
New CDF, Q-Q and P-P plots implemented for censored data

Use of cdfcompcens() to assess the fit of 3 distributions on data smokedfish

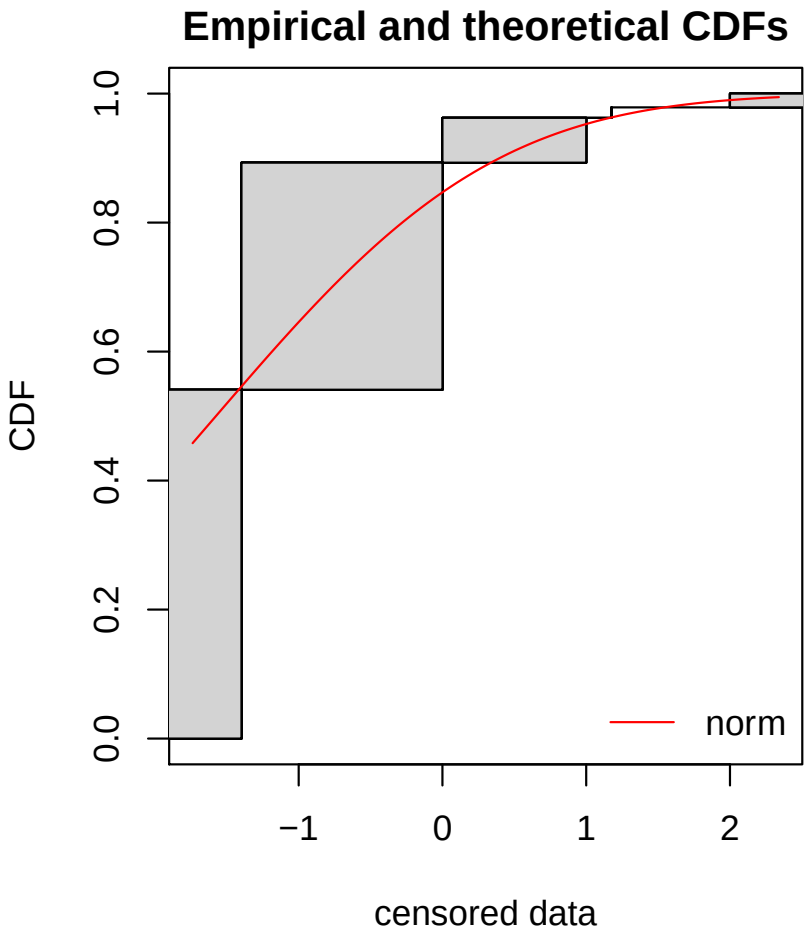


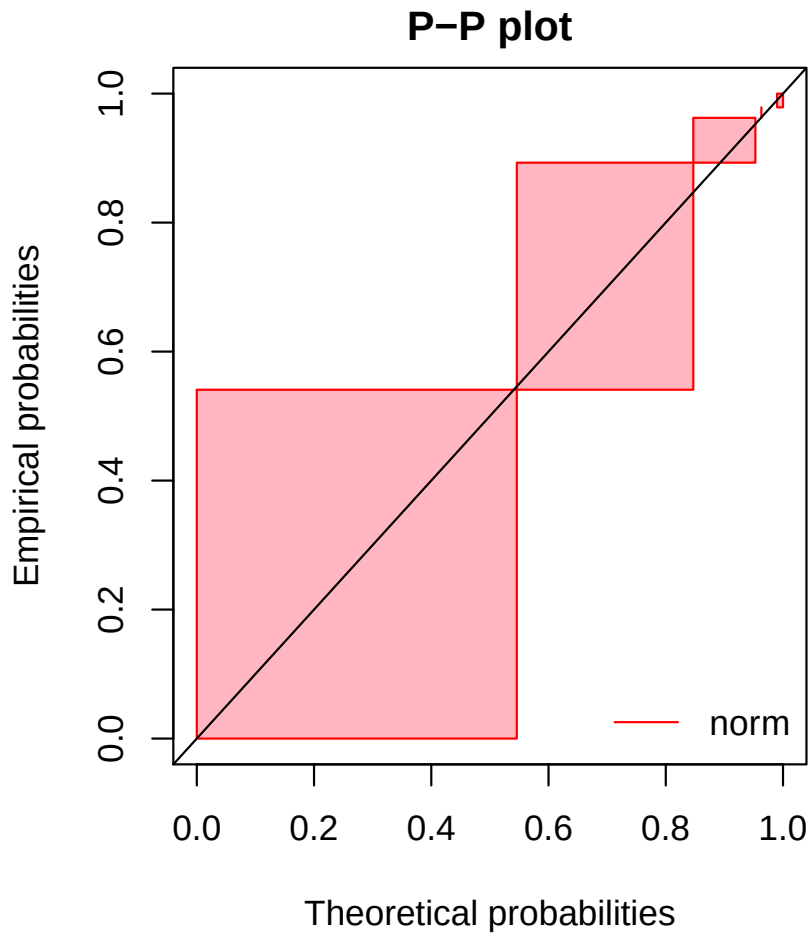
Use of qqcompcens() for one distribution



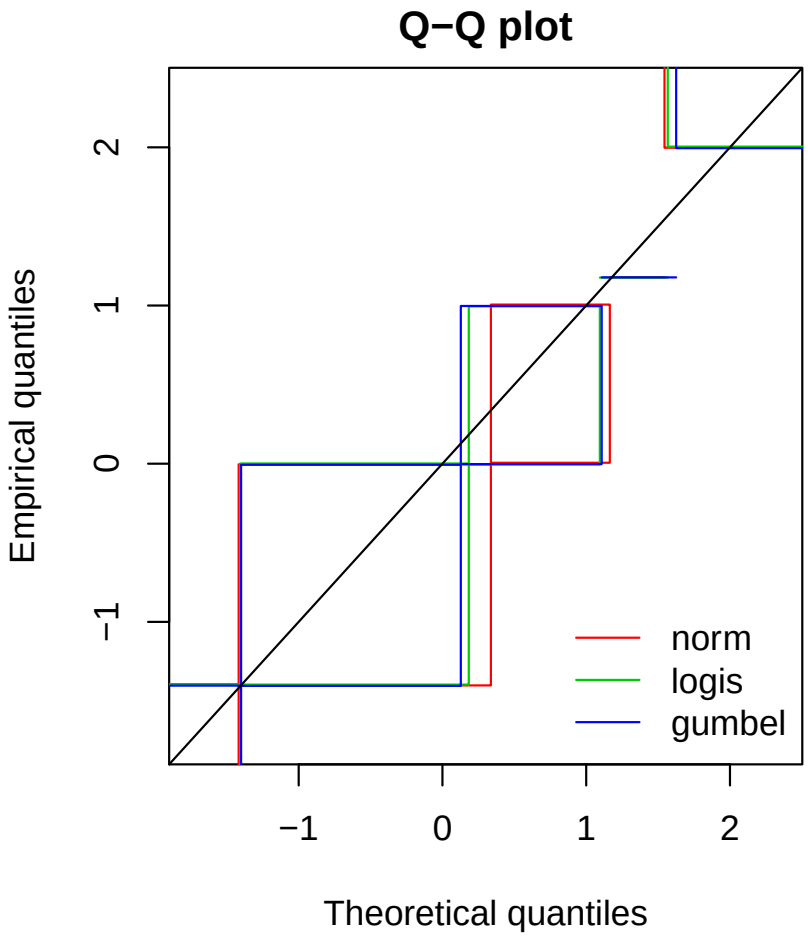


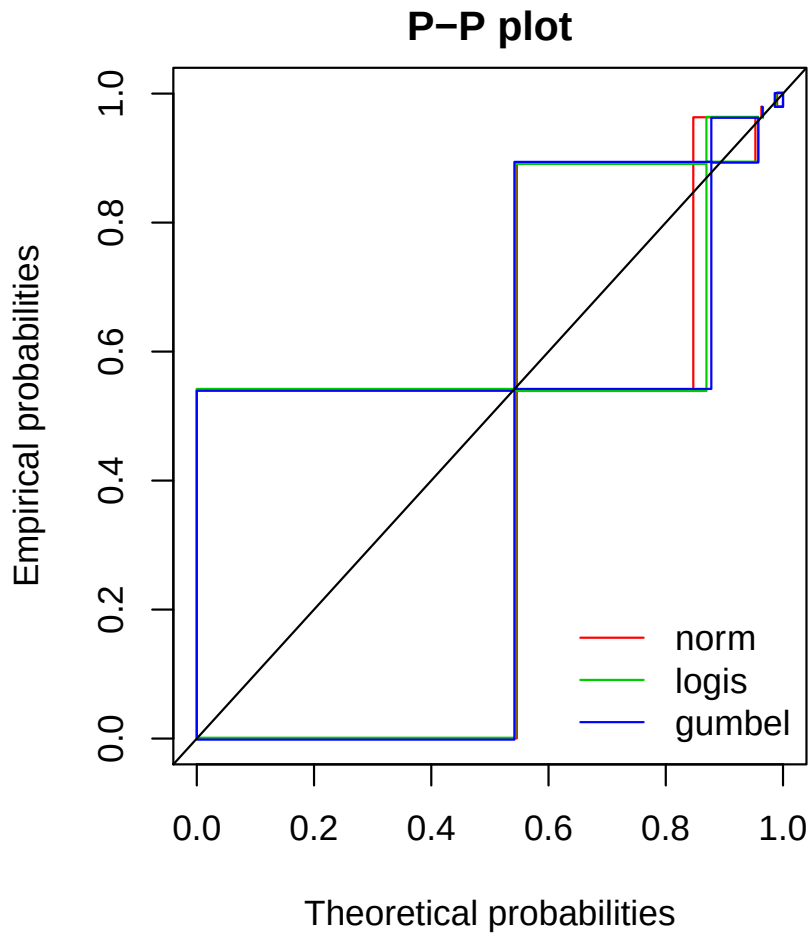
Use of ppcompcens() for one distribution



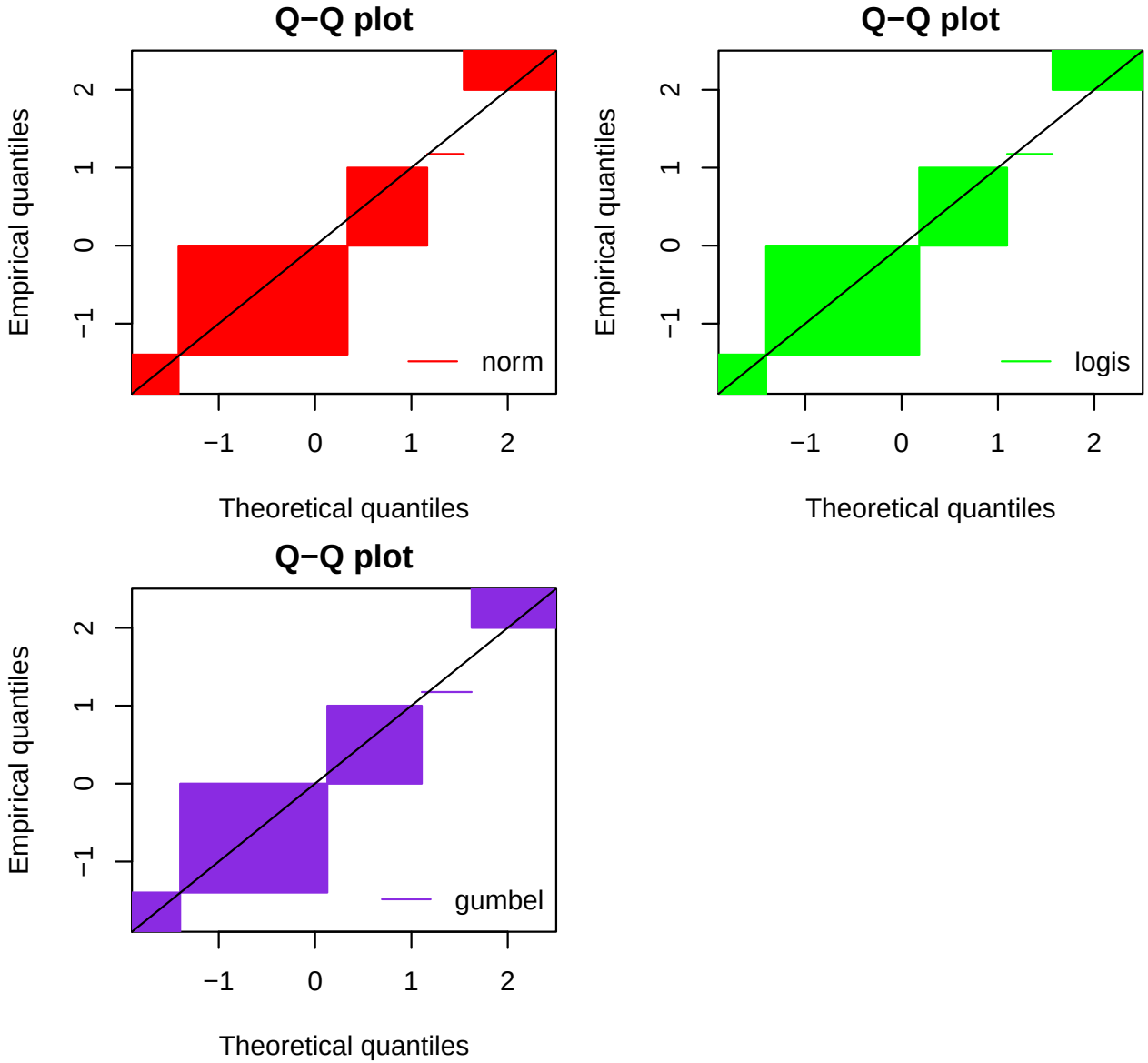


Q-Q plots and P-P plot for the 3 distributions



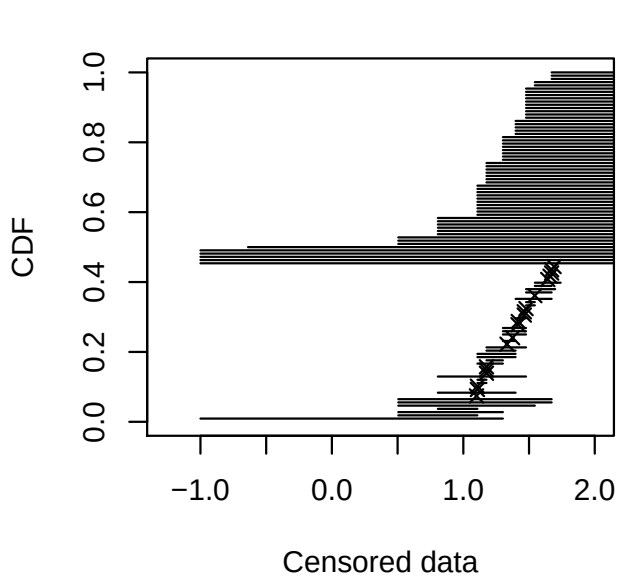


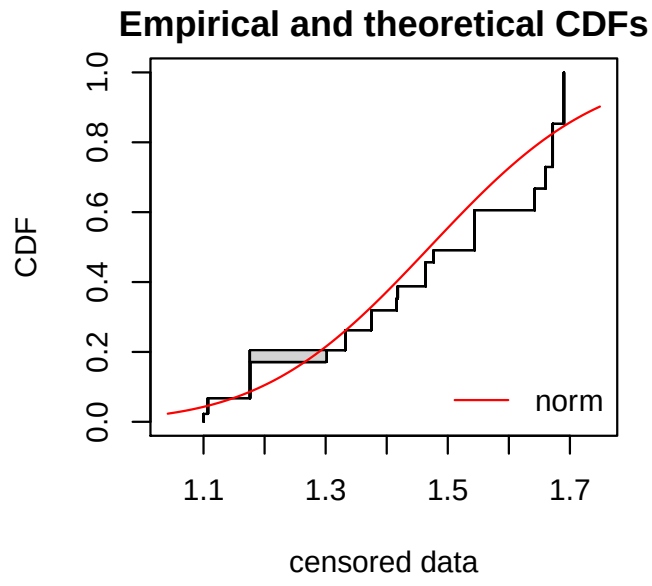
An alternative presentation of the Q-Q plots for the 3 dist.

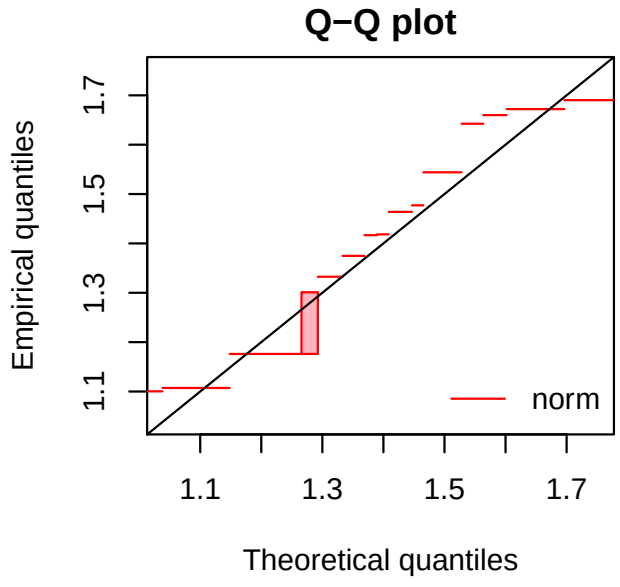


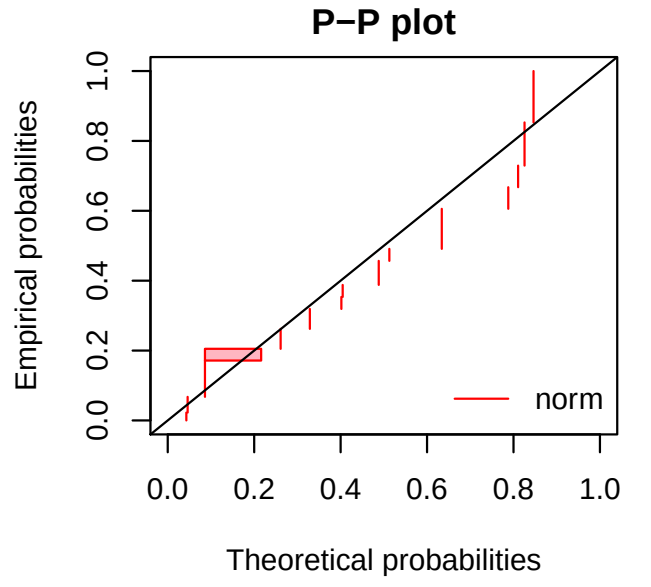
Will be soon implemented in the plotstyle ggplot.

Another example with data salinity









How to use of these new goodness-of-fit plots ?

Example of code:

```
data(smokedfish)
d <-  log10(smokedfish)
# Plot of the NPMLE CDF on censored data
plotdistcens(d)
# Two MLE fits
fitsfn <- fitdistcens(d,"norm")
fitsfl <- fitdistcens(d,"logis")
# Three goodness-of-fit plots for one fit
plot(fitsfn)
# Goodness-of-fit plots for one or more fits
cdfcompcens(list(fitsfn,fitsfl))
qqcompcens(list(fitsfn,fitsfl))
ppcompcens(list(fitsfn,fitsfl))
```

Other recent improvements of fitdistrplus

Version 1.0-8

- add of an optional use of ggplot2 in cdfcomp(), denscomp(), qqcomp() and ppcomp().

Version 1.0-10

- Improvement of goodness-of-fit plots for discrete distributions in denscomp().
- Add of new default starting values for distributions in actuar.

Version 1.0-11

- add of an optional use of ggplot2 in cdfcompcens(), denscompcens() and ppcompcens().

References

- Turnbull BW (1974). Nonparametric estimation of a survivorship function with doubly censored data. Journal of American Statistical Association, 69, 169-173.
- Gentleman, R., & Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. Biometrika, 81(3), 618-623.
- Wang, Y. (2008). Dimension-reduced nonparametric maximum likelihood computation for interval-censored data. Computational Statistics & Data Analysis, 52(5), 2388-2402.
- Wang, Y., & Taylor, S. M. (2013). Efficient computation of nonparametric survival functions via a hierarchical mixture formulation. Statistics and Computing, 23(6), 713-725.
- Wang, Y., & Fani, S. (2018). Nonparametric maximum likelihood computation of a U-shaped hazard function. Statistics and Computing, 28(1), 187-200.

Thank you for your attention !

We are waiting for your feedback on these new tools.