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 or 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).
- A FAQ vignette continuously updated in each new version.
Goodness-of-fit plots for non censored data
An example with non censored data

```r
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

<table>
<thead>
<tr>
<th></th>
<th>Weibull</th>
<th>gamma</th>
<th>lnorm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov statistic</td>
<td>0.0598</td>
<td>0.104</td>
<td>0.121</td>
</tr>
<tr>
<td>Cramer-von Mises statistic</td>
<td>0.0356</td>
<td>0.114</td>
<td>0.192</td>
</tr>
<tr>
<td>Anderson-Darling statistic</td>
<td>0.2288</td>
<td>0.654</td>
<td>1.136</td>
</tr>
</tbody>
</table>

## Goodness-of-fit criteria

<table>
<thead>
<tr>
<th></th>
<th>Weibull</th>
<th>gamma</th>
<th>lnorm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akaike's Information Criterion</td>
<td>43.7</td>
<td>46.5</td>
<td>51.8</td>
</tr>
<tr>
<td>Bayesian Information Criterion</td>
<td>48.9</td>
<td>51.7</td>
<td>57.0</td>
</tr>
</tbody>
</table>
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

\texttt{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

<table>
<thead>
<tr>
<th></th>
<th>left</th>
<th>right</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>6.0</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>8.0</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>NA</td>
</tr>
</tbody>
</table>

![CDF Plot](image-url)
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

![Graph showing equivalence classes on the first toy example](image-url)
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

CDF
L L L L R R R

CDF
Wang
Turnbull

-1 0 1 2
0.0 0.2 0.4 0.6 0.8 1.0
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 ppcomp() for one distribution
Q-Q plots and P-P plot for the 3 distributions

![Q-Q plot](image1)

![P-P plot](image2)
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

Empirical and theoretical CDFs

Q–Q plot

P–P plot
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

Thank you for your attention!
We are waiting for your feedback on these new tools.