GofCens package: Goodness-of-fit methods for complete and right-censored data

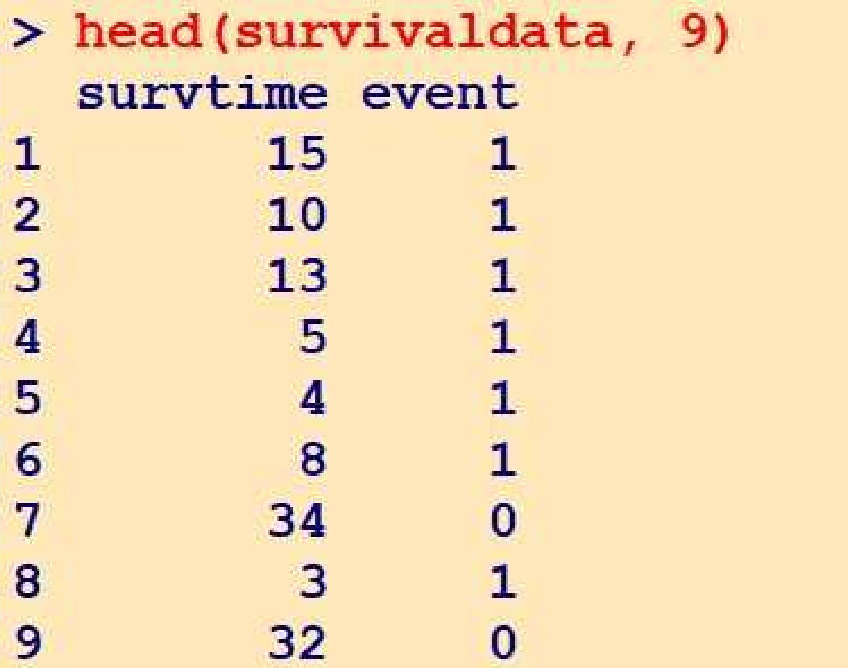
Klaus Langohr Mireia Besalú Matilde Francisco
Arnau García Guadalupe Gómez Melis

GRBIO
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
IRIS

Introduction

Motivation of GofCens package

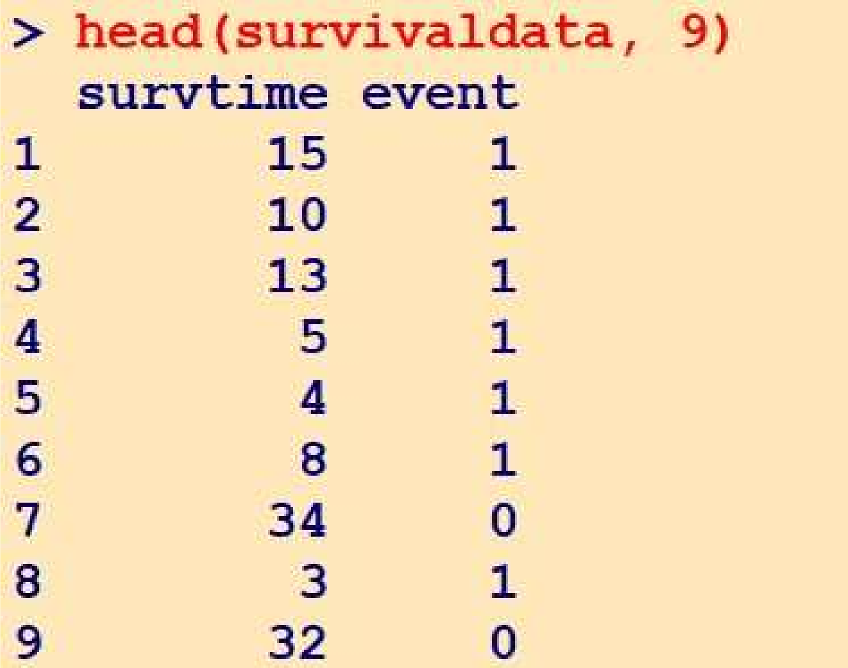
Given a sample of right-censored data ...



```
> head(survivaldata, 9)
  survtime event
1       15     1
2       10     1
3       13     1
4        5     1
5        4     1
6        8     1
7       34     0
8        3     1
9       32     0
```

Motivation of GofCens package

Given a sample of right-censored data ...



```
> head(survivaldata, 9)
  survtime event
1       15     1
2       10     1
3       13     1
4        5     1
5        4     1
6        8     1
7       34     0
8        3     1
9       32     0
```

... which R package offers graphical tools or statistical tests to identify an appropriate parametric model for the data?

A little bit of history

Master's theses

- **Anna Febrer Galvany** (2015):
  Analytical and Graphical Goodness of Fit Methods for Parametric Survival Models with Right-censored Data.
  Supervisors: G. Gómez Melis & K. Langohr

- **Mireia Besalú Mayol** (2016)
  Testing Goodness-of-Fit of Parametric Survival Models for Right-censored Data
  Supervisors: G. Gómez Melis & K. Langohr

- **Daniel García Carrasco** (2017)
  Goodness-of-fit R package for Right-censored data
  Supervisors: G. Gómez Melis, K. Langohr & M. Besalú

The GofCens package

Five functions that implement **statistical tests**

`KScens()`	Kolmogorov-Smirnov test
`CvMcens()`	Cramér-von Mises test
`ADcens()`	Anderson-Darling test
`gofcens()`	KS, CvM, and AD test
`chisqcens()`	χ^2-type test

The GofCens package

Five functions that implement **statistical tests**

`KScens()`	Kolmogorov-Smirnov test
`CvMcens()`	Cramér-von Mises test
`ADcens()`	Anderson-Darling test
`gofcens()`	KS, CvM, and AD test
`chisqcens()`	χ^2-type test

Three functions that implement **graphical tools**

`probPlot()`	Probability and Q-Q plots
`cumhazPlot()`	Cumulative hazard plots
`kmPlot()`	KME with overlaid parametric $S(t)$

The GofCens package

Five functions that implement **statistical tests**

`KScens()`	Kolmogorov-Smirnov test
`CvMcens()`	Cramér-von Mises test
`ADcens()`	Anderson-Darling test
`gofcens()`	KS, CvM, and AD test
`chisqcens()`	χ^2-type test

Three functions that implement **graphical tools**

`probPlot()`	Probability and Q-Q plots
`cumhazPlot()`	Cumulative hazard plots
`kmPlot()`	KME with overlaid parametric $S(t)$

Data set **nba** on mortality data of NBA players (Martínez *et al.* 2022).

Statistical Tests

Notation and hypothesis

Notation:

- $T_1, \ldots, T_n$: independent random variables with a common distribution $F(\cdot)$
- $C_1, \ldots, C_n$: independent censoring random variables with a common distribution $H(\cdot)$
- $Y_i = \min(T_i, C_i)$, $i = 1, \ldots, n$: observed variables
- $\delta_i = \mathbf{1}\{T_i \leq C_i\}$, $i = 1, \ldots, n$: event indicator

Notation and hypothesis

Notation:

- $T_1, \ldots, T_n$: independent random variables with a common distribution $F(\cdot)$
- $C_1, \ldots, C_n$: independent censoring random variables with a common distribution $H(\cdot)$
- $Y_i = \min(T_i, C_i)$, $i = 1, \ldots, n$: observed variables
- $\delta_i = \mathbf{1}\{T_i \leq C_i\}$, $i = 1, \ldots, n$: event indicator

Null hypothesis of goodness-of-fit test:

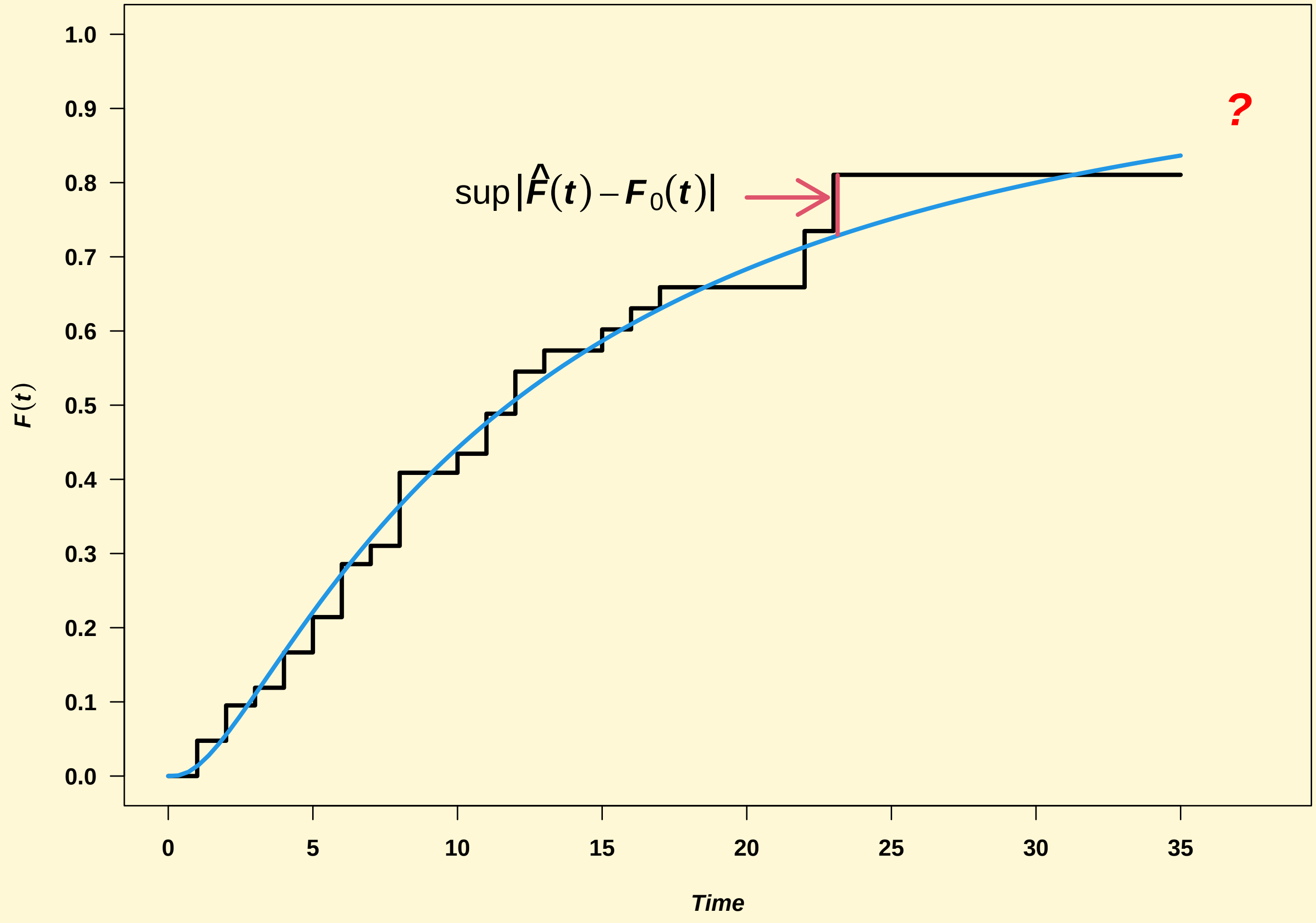
$$H_0 \colon F(t) = F_0(t; \theta),$$

where θ represents unknown parameter vector.

Theoretical foundation

- **T. R. Fleming, J. R. O'Fallon, P. C. O'Brien, and D. P. Harrington**. Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics*, 36(4):607–625, 1980.
- **J. A. Koziol and D. P. Byar**. Percentage points of the asymptotic distributions of one and two sample K-S statistics for truncated or censored data. *Technometrics*, 17(4):507–510, 1975.
- **J. A. Koziol and S. B. Green**. A Cramér-von Mises statistic for randomly censored data. *Biometrika*, 63 (3):465–474, 1976.
- **A. N. Pettitt and M. A. Stephens**. Modified Cramér-von Mises statistics for censored data. *Biometrika*, 63(2):291–298, 1976.

Kolmogorov-Smirnov test



Kolmogorov-Smirnov test

Kolmogorov-Smirnov statistic with right-censored data:

$$\widehat{D}_n = \sup_{0 \leq t \leq t_m} |\hat{F}_n(t) - F_0(t)|$$

$$= \sup_{0 \leq t \leq t_m} \left| \int_0^t \frac{\hat{S}_n(t) S_0(s)}{\hat{S}_n(s)} d\left[\hat{\Lambda}_n(s) - \Lambda_0(s)\right] \right|$$

Asymptotic distribution:

$$\lim_{n \to \infty} P(\sqrt{n}\widehat{D}_n \leq k) =$$

$$\sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2k^2} P\left(\left| Z - 2jk\sqrt{\frac{1-T}{T}} \right| < k\sqrt{\frac{1}{T - T^2}} \right),$$

Kolmogorov-Smirnov test

Kolmogorov-Smirnov statistic with right-censored data:

$$\widehat{D}_n = \sup_{0 \le t \le t_m} |\hat{F}_n(t) - F_0(t)|$$

$$= \sup_{0 \le t \le t_m} \left| \int_0^t \frac{\hat{S}_n(t) S_0(s)}{\hat{S}_n(s)} d\left[\hat{\Lambda}_n(s) - \Lambda_0(s)\right] \right|$$

Asymptotic distribution:

$$\lim_{n \to \infty} P(\sqrt{n}\widehat{D}_n \le k) =$$

$$\sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2k^2} P\left(\left| Z - 2jk\sqrt{\frac{1-T}{T}} \right| < k\sqrt{\frac{1}{T-T^2}} \right),$$

Function `KScens()` computes test statistic and p-value.

Bootstrapping can be used, too.

Cramér-von Mises-type test

Cramér-von Mises statistic:

$$M_n = n \int_{-\infty}^{+\infty} \left(\hat{F}_n(t) - F_0(t)\right)^2 dF_0(t)$$

Modified Cramér-von Mises statistic:

$$\widehat{M}_n = n_r \sum_{j=1}^{n_r+1} \hat{F}_n(u_{(j-1)})(u_{(j)} - u_{(j-1)}) \left(\hat{F}_n(u_{(j-1)}) - (u_{(j)} + u_{(j-1)})\right) + \frac{n_r}{3}$$

Cramér-von Mises-type test

Cramér-von Mises statistic:

$$M_n = n \int_{-\infty}^{+\infty} \left(\hat{F}_n(t) - F_0(t)\right)^2 dF_0(t)$$

Modified Cramér-von Mises statistic:

$$\widehat{M}_n = n_r \sum_{j=1}^{n_r+1} \hat{F}_n(u_{(j-1)})(u_{(j)} - u_{(j-1)}) \\ \left(\hat{F}_n(u_{(j-1)}) - (u_{(j)} + u_{(j-1)})\right) + \frac{n_r}{3}$$

Problem: Distribution of $\widehat{M}_n$ is not straightforwardly implemented $\Longrightarrow$ p-value cannot be computed directly.

Solution: Function `CvMcens()` employs bootstrapping.

Anderson-Darling test

Anderson-Darling statistic:

$$A_n = n \int_{-\infty}^{+\infty} (\hat{F}_n(t) - F_0(t))^2 \frac{dF_0(t)}{F_0(t)(1 - F_0(t))}.$$

Modified Anderson-Darling statistic:

$$\begin{aligned}\widehat{A}_n = -n_r + n_r \sum_{j=1}^{n_r} (\hat{F}_n(u_{(j-1)}) - 1)^2 \left[\log|1 - u_{(j-1)}| - \log|1 - u_{(j)}|\right] \\ + n_r \sum_{j=1}^{n_r - 1} \hat{F}_n^2(u_{(j)}) \left[\log|u_{(j+1)}| - \log|u_{(j)}|\right] - n_r \log|u_{(n)}|.\end{aligned}$$

Anderson-Darling test

Anderson-Darling statistic:

$$A_n = n \int_{-\infty}^{+\infty} (\hat{F}_n(t) - F_0(t))^2 \frac{dF_0(t)}{F_0(t)(1 - F_0(t))}.$$

Modified Anderson-Darling statistic:

$$\begin{aligned}\widehat{A}_n = -n_r + n_r \sum_{j=1}^{n_r} (\hat{F}_n(u_{(j-1)}) - 1)^2 \left[\log|1 - u_{(j-1)}| - \log|1 - u_{(j)}|\right] \\ + n_r \sum_{j=1}^{n_r - 1} \hat{F}_n^2(u_{(j)}) \left[\log|u_{(j+1)}| - \log|u_{(j)}|\right] - n_r \log|u_{(n)}|.\end{aligned}$$

Problem: Distribution of $\widehat{A}_n$ is not straightforwardly implemented $\Longrightarrow$ p-value cannot be computed directly.
Solution: Function `ADcens()` employs bootstrapping.

Bootstrapping for right-censored data

Computation of p-value for $H_0\colon F(t) = F_0(t;\theta)$ via bootstrapping with $b = 1, \ldots, B$ bootstrap samples:

1. Generation of $T_1^b, \ldots, T_n^b$ from $F_0(t;\theta)$.

Bootstrapping for right-censored data

Computation of p-value for $H_0\colon F(t) = F_0(t;\theta)$ via bootstrapping with $b = 1, \ldots, B$ bootstrap samples:

1. Generation of $T_1^b, \ldots, T_n^b$ from $F_0(t;\theta)$.
2. Generation of $C_1^b, \ldots, C_n^b$ from $\hat{H}_n(t)$.

Bootstrapping for right-censored data

Computation of p-value for $H_0\colon F(t) = F_0(t;\theta)$ via bootstrapping with $b = 1, \ldots, B$ bootstrap samples:

1. Generation of $T_1^b, \ldots, T_n^b$ from $F_0(t;\theta)$.
2. Generation of $C_1^b, \ldots, C_n^b$ from $\hat{H}_n(t)$.
3. Generation of $Y_1^b, \ldots, Y_n^b$ and $\delta_1^b, \ldots, \delta_n^b$.

Bootstrapping for right-censored data

Computation of p-value for $H_0\colon F(t) = F_0(t;\theta)$ via bootstrapping with $b = 1, \ldots, B$ bootstrap samples:

1. Generation of $T_1^b, \ldots, T_n^b$ from $F_0(t;\theta)$.
2. Generation of $C_1^b, \ldots, C_n^b$ from $\hat{H}_n(t)$.
3. Generation of $Y_1^b, \ldots, Y_n^b$ and $\delta_1^b, \ldots, \delta_n^b$.
4. MLE of $\hat{\theta}_n^b$ given $(Y_i^b, \delta_i^b), i = 1, \ldots n$.

Bootstrapping for right-censored data

Computation of p-value for $H_0\colon F(t) = F_0(t;\theta)$ via bootstrapping with $b = 1, \ldots, B$ bootstrap samples:

1. Generation of $T_1^b, \ldots, T_n^b$ from $F_0(t;\theta)$.
2. Generation of $C_1^b, \ldots, C_n^b$ from $\hat{H}_n(t)$.
3. Generation of $Y_1^b, \ldots, Y_n^b$ and $\delta_1^b, \ldots, \delta_n^b$.
4. MLE of $\hat{\theta}_n^b$ given $(Y_i^b, \delta_i^b), i = 1, \ldots n$.
5. Computation of the test statistic $(\hat{G}_n^{\hat{\theta}_n})_b$.

Bootstrapping for right-censored data

Computation of p-value for $H_0\colon F(t) = F_0(t;\theta)$ via bootstrapping with $b = 1, \ldots, B$ bootstrap samples:

1. Generation of $T_1^b, \ldots, T_n^b$ from $F_0(t;\theta)$.
2. Generation of $C_1^b, \ldots, C_n^b$ from $\hat{H}_n(t)$.
3. Generation of $Y_1^b, \ldots, Y_n^b$ and $\delta_1^b, \ldots, \delta_n^b$.
4. MLE of $\hat{\theta}_n^b$ given $(Y_i^b, \delta_i^b), i = 1, \ldots n$.
5. Computation of the test statistic $(\hat{G}_n^{\hat{\theta}_n})_b$.

P-value:

$$p = \frac{1}{B+1}\left(\sum_{i=1}^{B} \mathbf{1}\{\hat{G}_n \leq (\hat{G}_n^{\hat{\theta}_n})_b\} + 1\right)$$

Bootstrapping for right-censored data

Computation of p-value for $H_0\colon F(t) = F_0(t)$ via bootstrapping with $b = 1, \dots, B$ bootstrap samples:

1. Generation of $T_1^b, \dots, T_n^b$ from $F_0(t; \hat{\theta}_n)$.
2. Generation of $C_1^b, \dots, C_n^b$ from $\hat{H}_n(t)$.
3. Generation of $Y_1^b, \dots, Y_n^b$ and $\delta_1^b, \dots, \delta_n^b$.
4. MLE of $\hat{\theta}_n^b$ given $(Y_i^b, \delta_i^b), i = 1, \dots n$.
5. Computation of the test statistic $(\hat{G}_n^{\hat{\theta}_n})_b$.

P-value:

$$p = \frac{1}{B+1}\left(\sum_{i=1}^{B} \mathbf{1}\{\hat{G}_n \leq (\hat{G}_n^{\hat{\theta}_n})_b\} + 1\right)$$

Graphical Tools

Probability plots

Function `probPlot()` implements four probability and Q-Q plots for right-censored data.

Plot	**Abscissa**	**Ordinate**
P-P plot	$\widehat{F}_n(t)$	$\widehat{F}_0(t)$
Q-Q plot	t	$\widehat{F}_0^{-1}(\widehat{F}_n(t))$
SP plot	$\frac{\pi}{2}\arcsin\left(\widehat{F}_n(t)^{\frac{1}{2}}\right)$	$\frac{\pi}{2}\arcsin\left(\widehat{F}_0(t)^{\frac{1}{2}}\right)$
ER plot	$\widehat{F}_u(t)$	$\widehat{F}_u(\widehat{F}_0^{-1}(\widehat{F}_n(t)))$

Cumulative hazard plots

Function `cumhazPlot()` implements cumulative hazard plots for 8 distributions.

Distribution	$\Lambda(t)$	**Plot**
Exponential	βt	$\widehat{\Lambda}(t)$ vs t
Weibull	$(\beta t)^{\alpha}$	$\ln \widehat{\Lambda}(t)$ vs $\ln t$
Gumbel	$e^{\frac{t-\mu}{\beta}}$	$\ln \widehat{\Lambda}(t)$ vs t
Normal	$-\ln\left(1-\Phi\left(\frac{t-\mu}{\beta}\right)\right)$	$\Phi^{-1}\left(1-e^{\widehat{\Lambda}(t)}\right)$ vs t
Lognormal	$-\ln\left(1-\Phi\left(\frac{\ln t-\mu}{\beta}\right)\right)$	$\Phi^{-1}\left(1-e^{\widehat{\Lambda}(t)}\right)$ vs $\ln t$
Logistic	$\ln\left(1+e^{-\frac{t-\mu}{\beta}}\right)$	$\ln\left(e^{\widehat{\Lambda}(t)}-1\right)$ vs t
Loglogistic	$\ln\left(1+\left(\frac{t}{\beta}\right)^{\alpha}\right)$	$\ln\left(e^{\widehat{\Lambda}(t)}-1\right)$ vs $\ln t$
4-Param. Beta	$-\ln\left(1-\frac{B_{\frac{t-a}{b-a}}(\alpha,\gamma)}{B(\alpha,\gamma)}\right)$	$F^{-1}_{B(\alpha,\beta)}\left(1-e^{\widehat{\Lambda}(t)}\right)$ vs t

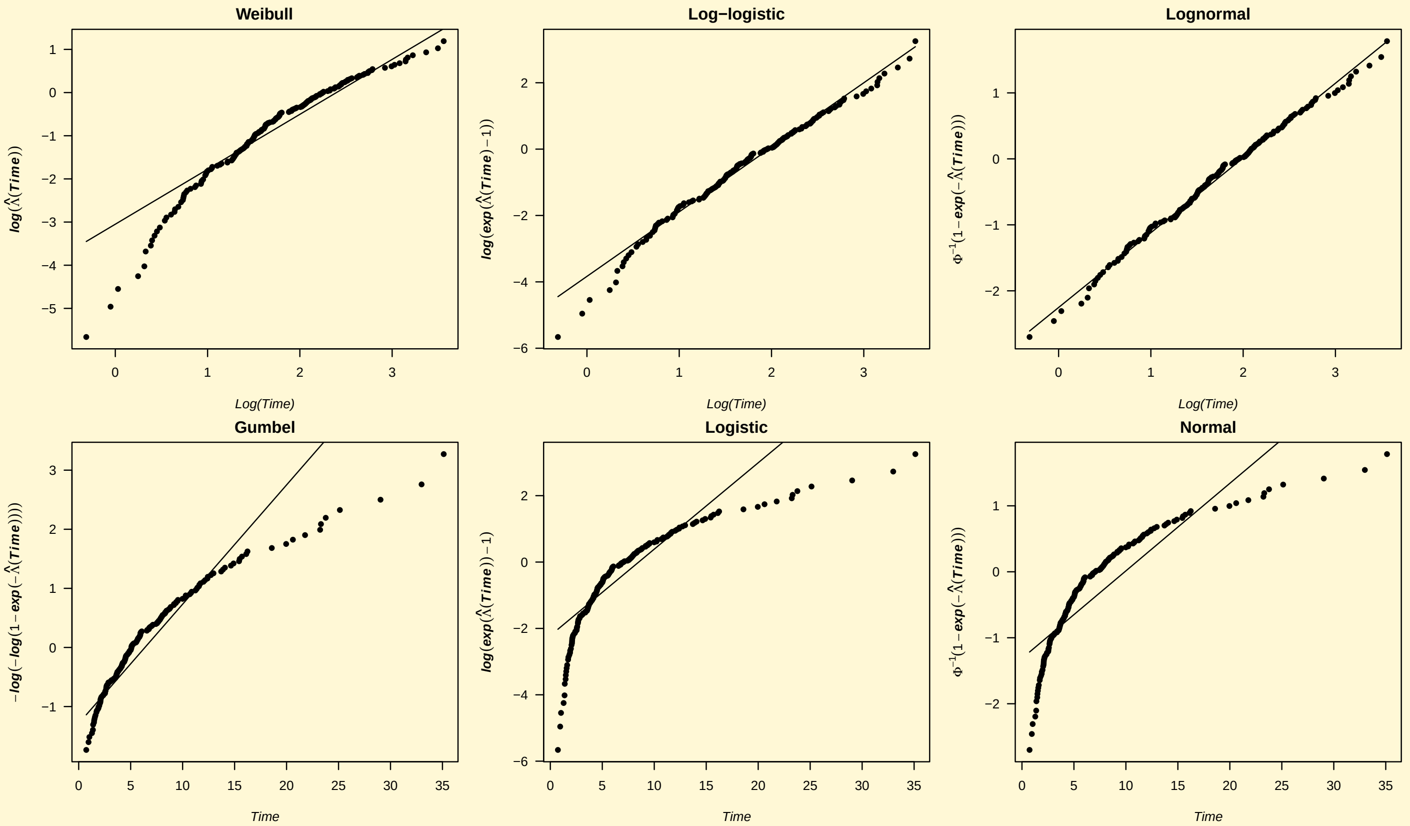
Illustration

Data from log-normal distribution

```
> set.seed(123)
# Survival times from log-normal distribution
> survt <- round(rlnorm(300, 2, 1), 2)
# Censoring times from exponential distribution
> censt <- round(rexp(300, 1 / 20), 2)

# Observed survival and event indicator
> times <- pmin(survt, censt)
> delta <- as.numeric(survt <= censt)
> table(delta)
> delta
  0   1
106 194
```

Cumulative hazard plots

```
> cumhazPlot(times, delta, font.lab = 4, cex.lab = 1.3,
+   cex.lab = 1.3, degs = 2, prnt = TRUE)
```

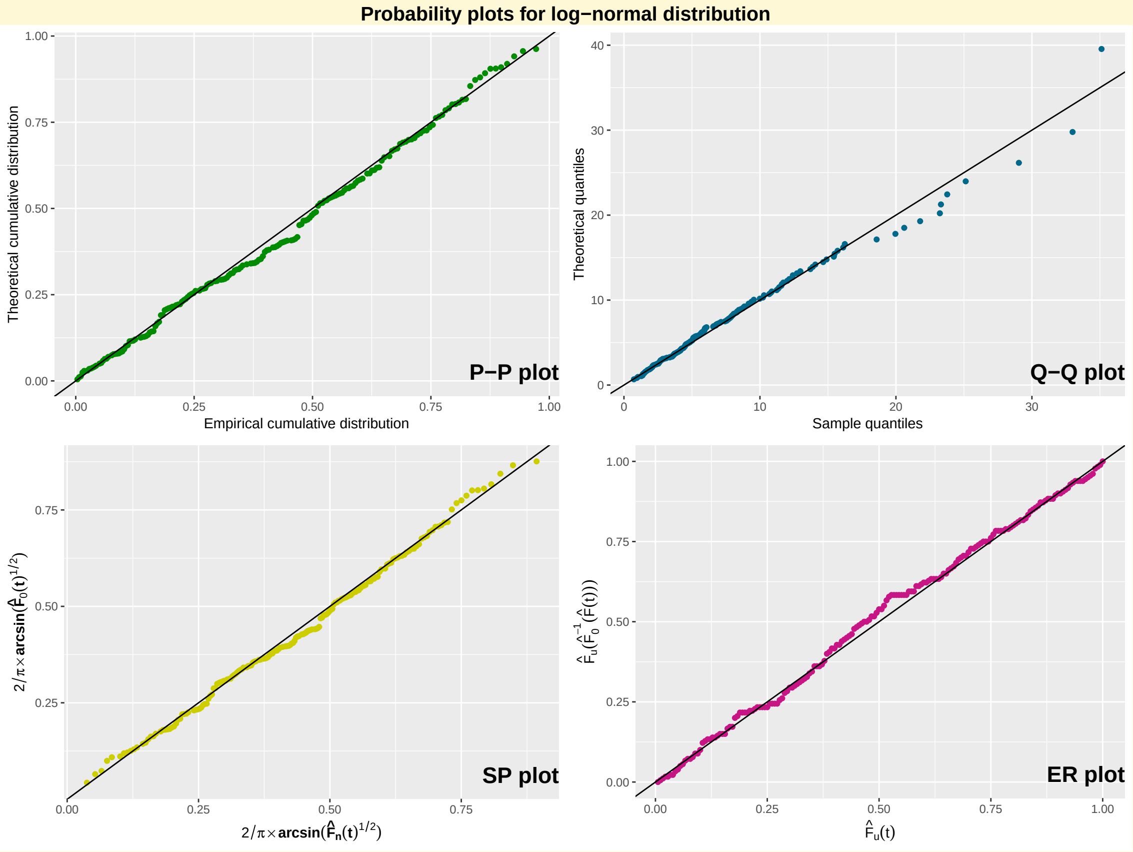


Output of `cumhazPlot()`

```
Parameter estimates

weibull
   Shape (se): 1.27 (0.07)
   Scale (se): 10.98 (0.62)
   AIC: 1304.05
   BIC: 1311.45

loglogistic
   Shape (se): 1.94 (0.11)
   Scale (se): 7.18 (0.42)
   AIC: 1272.48
   BIC: 1279.89

lognormal
Location (se): 1.99 (0.06)
   Scale (se): 0.88 (0.05)
   AIC: 1266.9
   BIC: 1274.31
:
```

Probability plots

```
> probPlot(times, delta, distr = "lognormal",
+    ggp = TRUE)
```



Kolmogorov-Smirnov test

```
> summary(KScens(times, delta, distr = "weibull",
+   boot = FALSE))

Distribution: weibull

KS Test results:
      A p-value
  1.391   0.042

Parameter estimates (se):
shape                scale
1.273 (0.067)        10.98 (0.621)

AIC: 1304.046
BIC: 1311.453
```

Function `gofcens()`

```
> set.seed(123)
> summary(gofcens(Surv(times, delta) ~ 1, distr = "lognormal",
+   params0 = list(location = 1.8, scale = 1)), print.AIC = FALSE,
+   print.BIC = FALSE, print.infoBoot = TRUE)

Distribution: lognormal

Null hypothesis:
location    scale
     1.8      1.0

Test statistics
   KS   CvM    AD
1.830 0.632 3.818

p-values
   KS   CvM    AD
0.004 0.006 0.006

Parameter estimates (se):
location          scale
1.988 (0.058)     0.883 (0.045)

Number of bootstrap samples: 999
```

Final comments

Final comments

- GofCens provides useful tools for assessing goodness of fit with right-censored data for many distributions.
- Version 1.5 is available on both CRAN and GitHub.

- Functions support complete and right-censored data.
- Test functions can be used for any distribution if `pname`, `dname`, and `rname` functions exist.
- Performance comparisons of the tests are currently under study.

Final comments

- GofCens provides useful tools for assessing goodness of fit with right-censored data for many distributions.
- Version 1.5 is available on both CRAN and GitHub.

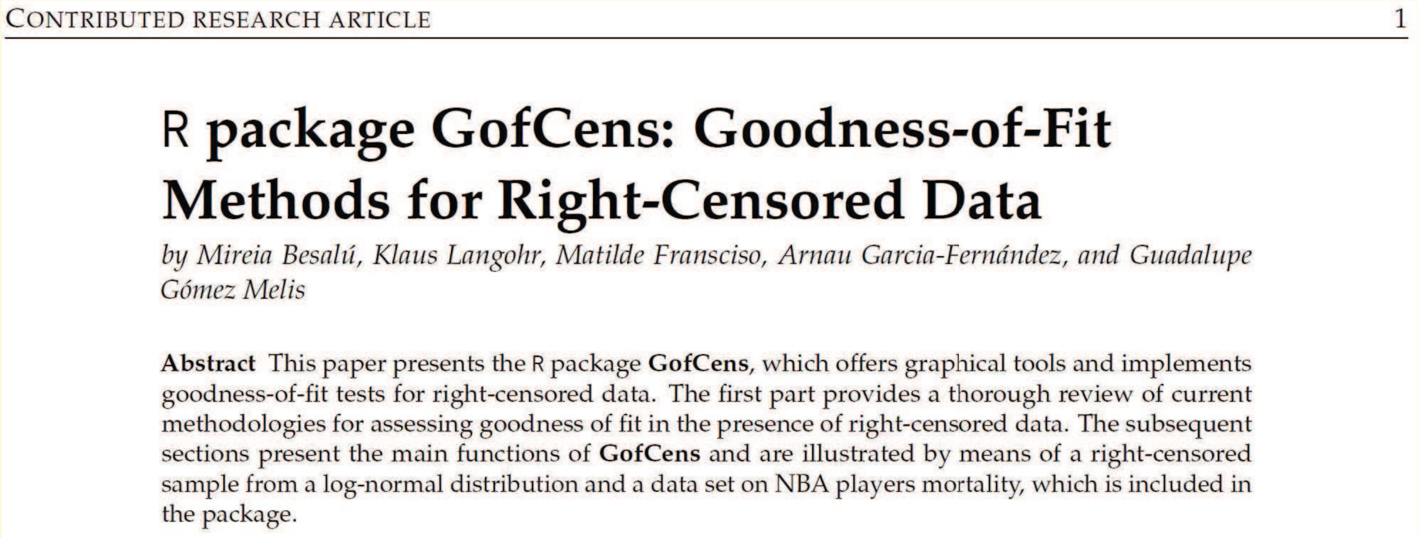
- Functions support complete and right-censored data.
- Test functions can be used for any distribution if `pname`, `dname`, and `rname` functions exist.
- Performance comparisons of the tests are currently under study.

Possible package extensions:

- Adaption of the functions to handle left-truncated data.
- Adaption to handle interval-censored data.
- Residual analysis of the accelerated failure time model.

Good news

Accepted for publication in the R Journal:



CONTRIBUTED RESEARCH ARTICLE 1

R package GofCens: Goodness-of-Fit Methods for Right-Censored Data

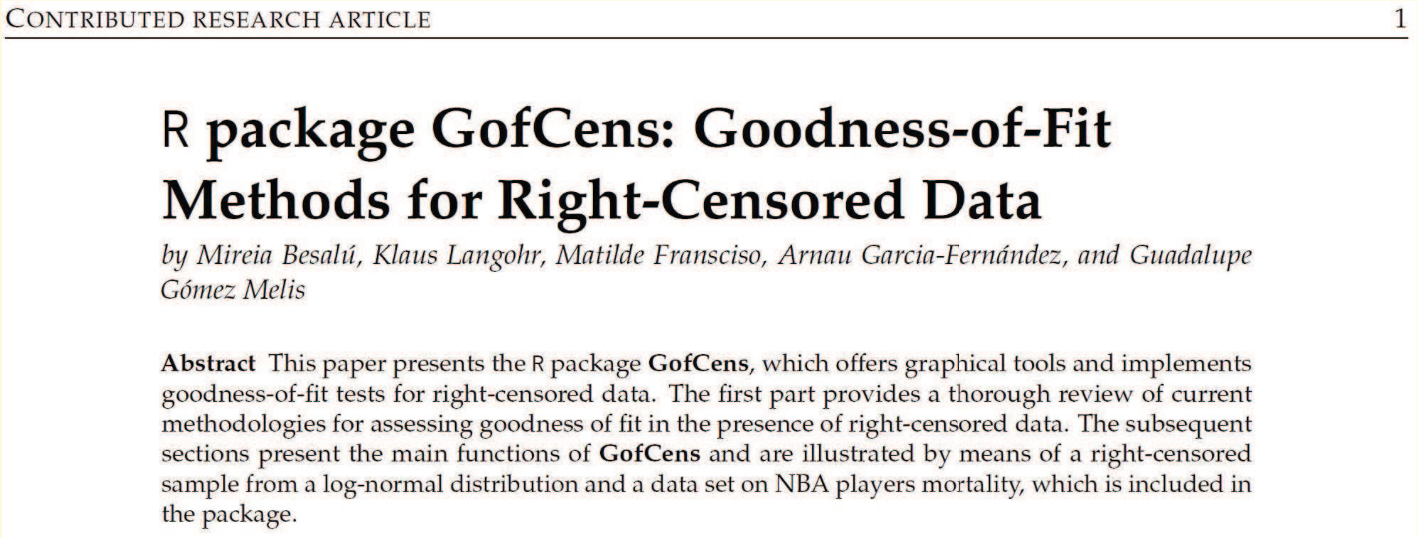
by Mireia Besalú, Klaus Langohr, Matilde Fransciso, Arnau Garcia-Fernández, and Guadalupe Gómez Melis

Abstract This paper presents the R package **GofCens**, which offers graphical tools and implements goodness-of-fit tests for right-censored data. The first part provides a thorough review of current methodologies for assessing goodness of fit in the presence of right-censored data. The subsequent sections present the main functions of **GofCens** and are illustrated by means of a right-censored sample from a log-normal distribution and a data set on NBA players mortality, which is included in the package.

Good news

Introduction
Statistical Tests
Graphical Tools
Illustration
Final comments

Accepted for publication in the R Journal:



CONTRIBUTED RESEARCH ARTICLE 1

R package GofCens: Goodness-of-Fit Methods for Right-Censored Data

by Mireia Besalú, Klaus Langohr, Matilde Fransciso, Arnau Garcia-Fernández, and Guadalupe Gómez Melis

Abstract This paper presents the R package **GofCens**, which offers graphical tools and implements goodness-of-fit tests for right-censored data. The first part provides a thorough review of current methodologies for assessing goodness of fit in the presence of right-censored data. The subsequent sections present the main functions of **GofCens** and are illustrated by means of a right-censored sample from a log-normal distribution and a data set on NBA players mortality, which is included in the package.

Publication will be announced in the GRBIO Newsletter.

Acknowledgements

Grants:

- RED2024-153680-T: Biostatnet 2025: Fortaleciendo la Investigación de Excelencia en Bioestadística a Nivel Nacional e Internacional Ministerio de Ciencia, Innovación y Universidades.
- PID2023-148033OB-C21: Statistics for Health Sciences: Advances in Survival Analysis and Clinical Trials. Financiado por MICIU/AEI /10.13039/501100011033 y por FEDER, UE.
- 2021 SGR 01421: GRBIO: Grup de Recerca en Bioestadística i Bioinformàtica. Agència de Gestió d'Ajuts Universitaris i de Recerca.

Moltes gràcies!

Thank you very much!

¡Muchas gracias!