Package 'GofCens'

August 20, 2021

Type Package

Title Goodness-of-Fit Methods for Right-Censored Data

Version 0.91

Date 2021-08-05

Description Graphical tools and goodness-of-fit tests for right-censored data:
1. Kolmogorov-Smirnov, Crámer-von Mises, and Anderson-Darling tests based on the empirical distribution function for complete data and their extensions for right-censored data.
2. Generalized chi-squared-type tests based on the squared difference between observed and expected counts using random cells with right-censored data.
3. A series of graphical tools such as probability or cumulative hazard plots to guide the decision about the parametric model that best fits the data.

License GPL (>= 2)

Encoding UTF-8

Depends R (>= 3.5.0), eha, actuar, fitdistrplus

Imports survival, survsim, grid, ggplot2, gridExtra

NeedsCompilation no

Author Klaus Langohr [aut, cre],
Mireia Besalú [aut],
Guadalupe Gómez [ctb]

Maintainer Klaus Langohr <klaus.langohr@upc.edu>

Repository CRAN

Date/Publication 2021-08-20 12:50:02 UTC

R topics documented:

`GofCens-package` *Goodness-of-Fit Methods for Right-Censored Data.*

Description

This package implements both graphical tools and goodness-of-fit tests for right-censored data. It has implemented

1. Kolmogorov-Smirnov, Crámer-von Mises, and Anderson-Darling tests based on the empirical distribution function for complete data and their extensions for right-censored data.
2. Generalized chi-squared-type tests based on the squared difference between observed and expected counts using random cells with right-censored data.
3. A series of graphical tools such as probability or cumulative hazard plots to guide the decision about the parametric model that best fits the data.

Details

Package:	GofCens
Type:	Package
Version:	0.91
Date:	2021-08-05
License:	GPL (>= 2)

Author(s)

Klaus Langohr, Mireia Besalú, Guadalupe Gómez

Maintainer: Klaus Langohr <klaus.langohr@upc.edu>

`chisqcens1` *General chi-squared statistics for right-censored data.*

Description

`chisqcens1` computes the general chi-squared statistic for right-censored data introduced by Kim (1993).

Usage

```
chisqcens1(times, cens = rep(1, length(times)), M,
           distr = c("exponential", "gumbel", "weibull", "normal",
                     "lognormal", "logistic", "loglogistic", "beta",
                     "uniform"),
           betaLimits = c(0, 1), igumb = c(10, 10), degs = 4,
           params = list(shape = NULL, shape2 = NULL, location = NULL,
                         scale = NULL))
```

Arguments

`times`	Numeric vector of times until the event of interest.
`cens`	Status indicator (1, exact time; 0, right-censored time). If not provided, all times are assumed to be exact.
`M`	Number indicating the number of cells that will be considered.
`distr`	A string specifying the name of the distribution to be studied. The possible distributions are the exponential (`"exponential"`), the Weibull (`"weibull"`), the Gumbel (`"gumbel"`), the normal (`"normal"`), the lognormal (`"lognormal"`), the logistic (`"logistic"`), the loglogistic (`"loglogistic"`), the beta (`"beta"`), and the uniform (`"uniform"`) distribution.
`betaLimits`	Two-components vector with the lower and upper bounds of the Beta distribution. This argument is only required, if the beta distribution is considered.
`igumb`	Two-components vector with the initial values for the estimation of the Gumbel distribution parameters.
`degs`	Integer indicating the number of decimal places of the numeric results of the output.
`params`	List specifying the parameters of the theoretical distribution. By default, parameters are set to `NULL` and estimated with the maximum likelihood method. This argument is only considered, if all parameters of the studied distribution are specified.

Details

The function implements the test introduced by Kim (1993) and returns the value of the test statistic.

The cell boundaries of the test are obtained via the quantiles, which are based on the Kaplan-Meier estimate of the distribution function. In the presence of right-censored data, it is possible that not all quantiles are estimated, and in this case, the value of `M` provided by the user is reduced.

The parameter estimation is acomplished with the `fitdistcens` function of the **fitdistrplus** package.

Value

A list containing the following components

`Statistic`	Value of the test statistic.
`Distribution`	Null distribution.

Parameters	The values of the parameters of the null distribution. If the user has set the parameters manually, these will be the returned parameters, otherwise the maximum likelihood estimates are returned.
CellNumber	Vector with two values: the original cell number introduced by the user and the final cell number used.

Author(s)

K. Langohr, M. Besalú, G. Gómez.

References

J. H. Kim. *Chi-Square Goodness-of-Fit Tests for Randomly Censored Data.* In: The Annals of Statistics, 21 (3) (1993), 1621-1639.

See Also

chisqcens2 for the computation of the p-value.

Examples

```
# Complete data
set.seed(123)
chisqcens1(time = rgumbel(1000, 12, scale = 4), M = 8, distr = "gumbel")

# Censored data
library(survival)
chisqcens1(aml$time, aml$status, M = 6, distr = "weibull")

data(nba)
chisqcens1(nba$survtime, nba$cens, 10, "logis", degs = 2)
chisqcens1(nba$survtime, nba$cens, 10, "beta", betaLimits = c(0, 70))
```

---

chisqcens2 *General chi-squared test for right-censored data.*

Description

chisqcens2 computes the general chi-squared statistic for right-censored data introduced by Kim (1993). The p-value is computed using bootstrapping.

Usage

```
chisqcens2(times, cens, M,
           distrData = c("weibull", "lognormal", "loglogistic"),
           distrCens = c("weibull", "lognormal", "loglogistic", "uniform"),
           BS = 1000, degs = 4,
           params = list(shape = NULL, location = NULL, scale = NULL))
```

Arguments

times	Numeric vector of times until the event of interest.
cens	Status indicator (1, exact time; 0, right-censored time).
M	Number indicating the number of cells that will be considered.
distrData	A string specifying the name of the distribution to be studied. The possible distributions are the Weibull ("weibull"), the lognormal ("lognormal"), and the loglogistic ("loglogistic") distribution.
distrCens	A string specifying the name of the distribution of the censoring times. The possible distributions are the Weibull ("weibull"), the lognormal ("lognormal"), the loglogistic ("loglogistic"), and the uniform ("uniform") distribution.
BS	Number of randomly censored samples under the null distribution. Default value: BS = 1000.
degs	Integer indicating the number of decimal places of the numeric results of the output.
params	List specifying the parameters of the theoretical distribution. By default, parameters are set to NULL and estimated with the maximum likelihood method. This argument is only considered, if all parameters of the studied distribution are specified.

Details

The function implements the test introduced by Kim (1993).

Different from the function chisqcens1, this function does not only provide the value of the test statistic, but also a p-value. For this purpose, random censored samples are generated under the null distribution and a given distribution of the censoring times, and the value of the test statistic is obtained for each sample. The empirical distribution of the test statistics obtained is used to determine the p-value.

The random censored samples are generated with the simple.surv.sim function of the **survsim** package.

Value

A list containing the following components

Statistic	Value of the test statistic.
pvalue	p-value computed using bootstrapping.
distrData	Null distribution.
distrCens	Distribution of the censoring times.
Parameters	The values of the parameters of the null distribution. If the user has set the parameters manually, these will be the returned parameters, otherwise the maximum likelihood estimates are returned.
CellNumber	Vector with two values: the original cell number introduced by the user and the final cell number used.

Warning

If the amount of data is large, the execution time of the function can be elevated. The parameter BS can limit the number of random censored samples generated and reduce the execution time. Also, notice that this function only works with right-censored data.

Author(s)

K. Langohr, M. Besalú, G. Gómez.

References

J. H. Kim. *Chi-Square Goodness-of-Fit Tests for Randomly Censored Data.* In: The Annals of Statistics, 21 (3) (1993), 1621-1639.

See Also

chisqcens1 for the computation of the test statistic of different distributions.

Examples

```
## Not run:
set.seed(123)
library(survsim)
n <- 50
datos <- simple.surv.sim(n, Inf,  dist.ev = "lnorm", 3, 2,
                         dist.cens = "lnorm",3, 2)
chisqcens2(datos$stop, datos$status, M = 8, "lognormal", "lognormal")
chisqcens2(datos$stop, datos$status, M = 8, "lognormal", "unif")

chisqcens2(nba$survtime, nba$cens, 10, "loglog", BS = 100)

## End(Not run)
```

---

cumhazPlot — *Cumulative hazard plots to check the goodness of fit of parametric models*

Description

cumhazPlot uses the cumulative hazard plot to check if a certain distribution is an appropiate choice for the data.

Usage

```
cumhazPlot(times, cens = rep(1, length(times)), distr = "all6", colour = 1,
           betaLimits = c(0, 1), igumb = c(10, 10), ggplo = FALSE, m = NULL,
           prnt = TRUE, decdig = 7, ...)
```

Arguments

`times`	Numeric vector of times until the event of interest.
`cens`	Status indicator (1, exact time; 0, right-censored time). If not provided, all times are assumed to be exact.
`distr`	A string specifying the names of the distributions to be studied. The possible distributions are the exponential (`"exponential"`), the Weibull (`"weibull"`), the Gumbel (`"gumbel"`), the normal (`"normal"`), the lognormal (`"lognormal"`), the logistic (`"logistic"`), the loglogistic (`"loglogistic"`), and the beta (`"beta"`) distribution. By default, `distr` is set to `"all6"`, which means that the cumulative hazard plots are drawn for the Weibull, loglogistic, lognormal, Gumbel, logistic, and normal distributions.
`colour`	Colour of the points. Default colour: black.
`betaLimits`	Two-components vector with the lower and upper bounds of the Beta distribution. This argument is only required, if the beta distribution is considered.
`igumb`	Two-components vector with the initial values for the estimation of the Gumbel distribution parameters.
`ggplo`	Logical to use or not the **ggplot2** package to draw the plots. Default is `FALSE`.
`m`	Optional layout for the plots to be displayed.
`prnt`	Logical to indicate if the maximum likelihood estimates of the parameters of all distributions considered should be printed. Default is `TRUE`.
`decdig`	Number of significant (see `signif`) digits to print when printing the parameter estimates. It is a suggestion only.
`...`	Optional arguments for function `par`, if `ggplo = FALSE`.

Details

The cumulative hazard plot is based on transforming the cumulative hazard function Λ in such a way that it becomes linear in t or $\log(t)$. This transformation is specific for each distribution. The function uses the data to compute the Nelson-Aalen estimator of the cumulative hazard function, $\widehat{\Lambda}$, and the maximum likelihood estimators of the parameters of the theoretical distribution under study. If the distribution fits the data, the plot is expected to be a straight line.

The parameter estimation is acomplished with the `fitdistcens` function of the **fitdistrplus** package.

Value

`params`	A list with the maximum likelihood estimates of the parameters of all distributions considered.

Author(s)

K. Langohr, M. Besalú, G. Gómez.

Examples

```
# Complete data and default distributions
set.seed(123)
x <- rlogis(1000, 50, 5)
cumhazPlot(x, lwd = 2)

# Censored data comparing three distributions
data(nba)
cumhazPlot(nba$survtime, nba$cens, distr = c("expo", "normal", "gumbel"))
```

`gofcens` *Kolmogorov-Smirnov, Crámer-von Mises, and Anderson-Darling statistics for complete and right-censored data*

Description

`gofcens` computes the Kolmogorov-Smirnov, Crámer-von Mises, and Anderson-Darling statistics for complete and right-censored data against eight possible distributions.

Usage

```
gofcens(times, cens = rep(1, length(times)),
        distr = c("exponential", "gumbel", "weibull", "normal",
                  "lognormal", "logistic", "loglogistic", "beta"),
        betaLimits = c(0, 1), igumb = c(10, 10), degs = 4,
        params = list(shape = NULL, shape2 = NULL, location = NULL,
                      scale = NULL))
```

Arguments

`times`	Numeric vector of times until the event of interest.
`cens`	Status indicator (1, exact time; 0, right-censored time). If not provided, all times are assumed to be exact.
`distr`	A string specifying the name of the distribution to be studied. The possible distributions are the exponential (`"exponential"`), the Weibull (`"weibull"`), the Gumbel (`"gumbel"`), the normal (`"normal"`), the lognormal (`"lognormal"`), the logistic (`"logistic"`), the loglogistic (`"loglogistic"`), and the beta (`"beta"`) distribution.
`betaLimits`	Two-components vector with the lower and upper bounds of the Beta distribution. This argument is only required, if the beta distribution is considered.
`igumb`	Two-components vector with the initial values for the estimation of the Gumbel distribution parameters.
`degs`	Integer indicating the number of decimal places of the numeric results of the output.
`params`	List specifying the parameters of the theoretical distribution. By default, parameters are set to `NULL` and estimated with the maximum likelihood method. This argument is only considered, if all parameters of the studied distribution are specified.

Details

Fleming et al. (1980) proposed a modified Kolmogorov-Smirnov test to be used with right-censored data. Koziol and Green (1976) proposed a Crámer-von Mises statistic for randomly censored data. This function reproduces this test for a given survival data and a theorical distribution. In presence of ties, different authors provide slightly different definitions of the product-limit estimator, what might provide different values of the test statistic.

When dealing with complete data, we recommend the use of functions `ks.test` of the **stats** package and `cvm.test` and `ad.test` of the **goftest** package.

Value

A list containing the following components

`Tests statistics`
: Values of the Kolmogovor-Smirnov, Crámer-von Mises, and Anderson-Darling test statistics

`Distribution`
: Null distribution

Author(s)

K. Langohr, M. Besalú, G. Gómez.

References

T. R. Fleming et al. *Modified Kolmogorov-Smirnov test procedure with application to arbitrarily right-censored data*. In: Biometrics 36 (1980), 607-625.

J. A. Koziol and S. B. Green. *A Crámer-von Mises statistic for randomly censored data*. In: Biometrika, 63 (3) (1976), 465-474.

A. N. Pettitt and M. A. Stephens. *Modified Crámer-von Mises statistics for censored data.* In: Biometrika, 63 (2) (1976), 291-298.

See Also

ks.test (Package `stats`), cvm.test (Package `goftest`), and ad.test (Package `goftest`) for complete data, and KScens for the Kolmogorov-Smirnov test for right-censored data, which returns the p-value.

Examples

```
# Complete data
set.seed(123)
gofcens(times = rweibull(1000, 12, scale = 4), distr = "weibull")

# Censored data
library(survival)
gofcens(aml$time, aml$status, distr = "norm")

data(nba)
gofcens(nba$survtime, nba$cens, "lognorm")
gofcens(nba$survtime, nba$cens, "norm")
```

KScens *Kolmogorov-Smirnov test for complete and right-censored data*

Description

KScens computes the Kolmogorov-Smirnov statistic and p-value for complete and right-censored data against eight possible distributions.

Usage

```
KScens(times, cens = rep(1, length(times)),
       distr = c("exponential", "gumbel", "weibull", "normal",
                 "lognormal", "logistic", "loglogistic", "beta"),
       betaLimits = c(0, 1), igumb = c(10, 10), degs = 4,
       params = list(shape = NULL, shape2 = NULL, location = NULL,
                     scale = NULL))
```

Arguments

times	Numeric vector of times until the event of interest.
cens	Status indicator (1, exact time; 0, right-censored time). If not provided, all times are assumed to be exact.
distr	A string specifying the name of the distribution to be studied. The possible distributions are the exponential ("exponential"), the Weibull ("weibull"), the Gumbel ("gumbel"), the normal ("normal"), the lognormal ("lognormal"), the logistic ("logistic"), the loglogistic ("loglogistic"), and the beta ("beta") distribution.
betaLimits	Two-components vector with the lower and upper bounds of the Beta distribution. This argument is only required, if the beta distribution is considered.
igumb	Two-components vector with the initial values for the estimation of the Gumbel distribution parameters.
degs	Integer indicating the number of decimal places of the numeric results of the output.
params	List specifying the parameters of the theoretical distribution. By default, parameters are set to NULL and estimated with the maximum likelihood method. This argument is only considered, if all parameters of the studied distribution are specified.

Details

Fleming et al. (1980) proposed a modified Kolmogorov-Smirnov test to use with right-censored data. This function reproduces this test for a given survival data and a theorical distribution. The p-value is computed following the results of Koziol and Byar (1975) and the output of the function follows the notation of Fleming et al. (1980).

In presence of ties, different authors provide slightly different definitions of $\widehat{F}_n(t)$, with which other values of the test statistic might be obtained.

When dealing with complete data, we recommend to use the function `ks.test` of the **stats** package.

The parameter estimation is acomplished with the `fitdistcens` function of the **fitdistrplus** package.

Value

A list containing the following components:

`p-value`	Estimated p-value.
`A`	Value of the modified Kolmogorov-Smirnov statistic.
`F(y_m)`	Estimation of the image of the last recorded time.
`y_m`	Last recorded time.
`distr`	Null distribution.
`param`	List with the maximum likelihood estimates of the parameters of the distribution under study.

Author(s)

K. Langohr, M. Besalú, G. Gómez.

References

T. R. Fleming et al. *Modified Kolmogorov-Smirnov test procedure with application to arbitrarily right-censored data*. In: Biometrics 36 (1980), 607-625.

J.A. Koziol and P. Byar. *Percentage Points of the Asymptotic Distributions of One and Two Sample K-S statistics for Truncated or Censored Data*. In: Technometrics 17 (4) (1975), 507-510.

See Also

ks.test (Package `stats`) for complete data and gofcens for Crámer von-Mises and Anderson-Darling statistics for right-censored data.

Examples

```
# Complete data
set.seed(123)
KScens(times = rweibull(1000, 12, scale = 4), distr = "weibull")

# Censored data
library(survival)
KScens(aml$time, aml$status, distr = "norm")

data(nba)
KScens(nba$survtime, nba$cens, "logis", degs = 2)
KScens(nba$survtime, nba$cens, "beta", betaLimits = c(0, 70))
```

nba *Survival times of former NBA players.*

Description

Survival times of former NBA players after their NBA career.

Usage

```
data("nba")
```

Format

A data frame with 3501 observations on the following 3 variables.

`id` Player ID

`survtime` Time (in years) from end of NBA career until either death or April 15, 2014.

`cens` Death indicator (1, exact survival time; 0, right-censored survival time).

Details

The survival times of former NBA players were analyzed by Martínez et al. (2019).

Source

J. A. Martínez, K. Langohr, J. Felipo, and M. Casals. *Mortality of NBA players: Risk factors and comparison with the general US population.* In: Applied Sciences, 9 (3) (2019).

Examples

```
data(nba)
cumhazPlot(nba$survtime, nba$cens)
```

probPlot *Probability plots to check the goodness of fit of parametric models*

Description

`probPlot` provides four types of probability plots: P-P plot, Q-Q plot, Stabilised probability plot, and Empirically Rescaled plot to check if a certain distribution is an appropiate choice for the data.

Usage

```
probPlot(times, cens = rep(1, length(times)),
         distr = c("exponential", "gumbel", "weibull", "normal",
                   "lognormal", "logistic", "loglogistic", "beta"),
         plots = c("PP", "QQ", "SP", "ER"),
         colour = c("green4", "deepskyblue4", "yellow3", "mediumvioletred"),
         betaLimits = c(0, 1), igumb = c(10, 10), mtitle = TRUE, ggplo = FALSE,
         m = NULL, prnt = TRUE, decdig = 7,
         params = list(shape = NULL, shape2 = NULL, location = NULL,
                       scale = NULL), ...)
```

Arguments

`times`	Numeric vector of times until the event of interest.
`cens`	Status indicator (1, exact time; 0, right-censored time). If not provided, all times are assumed to be exact.
`distr`	A string specifying the name of the distribution to be studied. The possible distributions are the exponential (`"exponential"`), the Weibull (`"weibull"`), the Gumbel (`"gumbel"`), the normal (`"normal"`), the lognormal (`"lognormal"`), the logistic (`"logistic"`), the loglogistic (`"loglogistic"`), and the beta (`"beta"`) distribution.
`plots`	Vector stating the plots to be displayed. Possible choices are the P-P plot (`"PP"`), the Q-Q plot (`"QQ"`), the Stabilised Probability plot (`"SP"`), and the Empirically Rescaled plot (`"ER"`). By default, all four plots are displayed.
`colour`	Vector indicating the colours of the displayed plots. The vector will be recycled if its length is smaller than the number of plots to be displayed.
`betaLimits`	Two-components vector with the lower and upper bounds of the Beta distribution. This argument is only required, if the beta distribution is considered.
`igumb`	Two-components vector with the initial values for the estimation of the Gumbel distribution parameters.
`mtitle`	Logical to add or not the title "Probability plots for a `distr` distribution" to the plot. Default is `TRUE`.
`ggplo`	Logical to use or not the **ggplot2** package to draw the plots. Default is `FALSE`.
`m`	Optional layout for the plots to be displayed.
`prnt`	Logical to indicate if the maximum likelihood estimates of the parameters should be printed. Default is `TRUE`.
`decdig`	Number of significant (see `signif`) digits to print when printing the parameter estimates. It is a suggestion only.
`params`	List specifying the parameters of the theoretical distribution. By default, parameters are set to `NULL` and estimated with the maximum likelihood method. This argument is only considered, if all parameters of the studied distribution are specified.
`...`	Optional arguments for function `par`, if `ggplo = FALSE`.

Details

By default, function `probPlot` draws four plots: P-P plot, SP plot, Q-Q plot, and EP plot. Following, a description is given for each plot.

The **Probability-Probability plot** (P-P plot) depicts the empirical distribution, $\widehat{F}(t)$, which is obtained with the Kaplan-Meier estimator if data are right-censored, versus the theoretical cumulative distribution function (cdf), $\widehat{F_0}(t)$. If the data come from the chosen distribution, the points of the resulting graph are expected to lie on the identity line.

The **Stabilised Probability plot** (SP plot), proposed by Michael (1983), is a transformation of the P-P plot. It stabilises the variance of the plotted points. If $F_0 = F$ and the parameters of F_0 are known, $\widehat{F_0}(t)$ corresponds to the cdf of a uniform order statistic, and the arcsin transformation stabilises its variance. If the data come from distribution F_0, the SP plot will resemble the identity line.

The **Quartile-Quartile plot** (Q-Q plot) is similar to the P-P plot, but it represents the sample quantiles versus the theoretical ones, that is, it plots t versus $\widehat{F}_0^{-1}(\widehat{F}(t))$. Hence, if F_0 fits the data well, the resulting plot will resemble the identity line.

A drawback of the Q-Q plot is that the plotted points are not evenly spread. Waller and Turnbull (1992) proposed the **Empirically Rescaled plot** (EP plot), which plots $\widehat{F}_u(t)$ against $\widehat{F}_u(\widehat{F}_0^{-1}(\widehat{F}(t)))$, where $\widehat{F}_u(t)$ is the empirical cdf of the points corresponding to the uncensored observations. Again, if $\widehat{F}_0$ fits the data well, the ER plot will resemble the identity line.

By default, all four probability plots are drawn and the maximum likelihood estimates of the parameters of the chosen parametric model are returned. The parameter estimation is acomplished with the `fitdistcens` function of the **fitdistrplus** package.

Value

`outp`	List with the maximum likelihood estimates of the parameters of the distribution under study.

Author(s)

K. Langohr, M. Besalú, G. Gómez.

References

J. R. Michael. *The Stabilized Probability Plot*. In: Biometrika 70 (1) (1983), 11-17.

L.A. Waller and B.W. Turnbull. *Probability Plotting with Censored Data*. In: American Statistician 46 (1) (1992), 5-12.

Examples

```
# P-P, Q-Q, SP, and EP plots for complete data
set.seed(123)
x <- rlnorm(1000, 3, 2)
probPlot(x)
probPlot(x, distr = "lognormal")

# P-P, Q-Q, SP, and EP plots for censored data using ggplot2
```

```
library(survival)
probPlot(aml$time, aml$status, ggplo = TRUE)

# P-P, Q-Q and SP plots for censored data and lognormal distribution
data(nba)
probPlot(nba$survtime, nba$cens, "lognorm", plots = c("PP", "QQ", "SP"),
         ggplo = TRUE, m = matrix(1:3, nr = 1))
```

Index