

Package ‘dbEmpLikeGOF’

February 19, 2015

Version 1.2.4

Date Aug 23, 2013

Title Goodness-of-fit and two sample comparison tests using sample entropy

Author Lori A. Shepherd <las65@buffalo.edu>Albert Vexler <avexler@buffalo.edu>Jeffrey C. Miecznikowski <jcm38@buffalo.edu>

Maintainer Lori A. Shepherd <las65@buffalo.edu>

Description Goodness-of-fit and two sample comparison tests using sample entropy

LazyData yes

License GPL (>= 2)

URL <http://sphhp.buffalo.edu/biostat/research/software/EmpLike.GOF/index.php>

Depends R (>= 2.10)

NeedsCompilation no

Repository CRAN

Date/Publication 2013-08-26 22:03:50

R topics documented:

dbEmpLikeGOF-package	2
Baby	2
dbEmpLikeGOF	3
returnCutoff	5
Snow	6
Tbars	7
testmat	7
Index	8

dbEmpLikeGOF-package *Empirical Likelihood Goodness Of Fit Tests*

Description

Package that has functions to perform empirical likelihood goodness of fit tests for normality, uniformity, and distribution equality.

Author(s)

Jeffrey C. Mieznikowski, Lori A. Shepherd

References

Jeffrey C. Mieznikowski, Albert Vexler, Lori A. Shepherd (2013). dbEmpLikeGOF: An R Package for Nonparametric Likelihood Ratio Tests for Goodness-of-Fit and Two-Sample Comparisons Based on Sample Entropy. *Journal of Statistical Software*, 54(3), 1-19. <http://www.jstatsoft.org/v54/i03/>

Vexler A, Gurevich G, Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. *Computational Statistics and Data Analysis* 54(2010) 531-545.

Gurevich G, Vexler A, A two-sample empirical likelihood ratio test based on samples entropy. *Statistics and Computing*, 2011.

See Also

[dbEmpLikeGOF](#), [returnCutoff](#)

Baby

Baby Dataset

Description

Baby dataset

Format

baby is a vector of data. babyTable is a data.frame

Details

This is a baby boom dataset summarizing the time of birth, sex, and birth weight for 44 babies born in one 24-hour period at a hospital in Brisbane, Australia. The full dataset can be found at *A simple data set for demonstrating common distributions*. Dunn, Peter K. *Journal of Statistics Education*. 7(3) 1999

 dbEmpLikeGOF

Density Based Empirical Likelihood Goodness of Fit

Description

Performs density based empirical likelihood goodness of fit tests for normality, uniformity, and distribution equality.

Usage

```
dbEmpLikeGOF(x,
             y=NA,
             testcall=c("uniform", "normal"),
             delta=0.50,
             delta.equality=0.10,
             num.mc=1000,
             pv1.Table=TRUE,
             vrb=TRUE,
             random.seed.flag=TRUE)
```

Arguments

x	vector of data values
y	an optional vector of data values when testing for distribution equality
testcall	Type of distribution: either uniform or normal
delta	an option for changing the minimizing range for the EL ratio test statistic for the normal and uniform distribution.
delta.equality	an option for changing the minimizing range for the EL ratio test statistic for the two sample distribution equality.
num.mc	number of simulations to use when calculating p-value
pv1.Table	logical indicating if p-value should be calculated based on estimates from stored data tables or by using Monte Carlo techniques
vrb	logical indicating if status messages should be printed
random.seed.flag	logical if set seed should be set

Details

The method employs a density-based empirical likelihood approach to obtain test statistic and p-values for a goodness-of-fit tests for uniformity, normality, and two sample distribution equality.

If both 'x' and 'y' are specified then a two sample distribution is performed to evaluate the null hypothesis of equal distributions.

If only 'x' is specified, then the 'testcall' option must be specified as either 'uniform' (uniform) or 'normal' (normal) denoting whether the distribution of the 'x' vector of observations should be tested against the normal or uniform distribution.

The 'delta' value should remain at the default value of 0.50. The 'delta' value corresponds to the delta in equation 2.10 (normal) or equation 2.3.2 (uniform) in Vexler and Gurevich, 2010. Essentially this setting controls the range over which a minimum is taken to produce the EL ratio test statistic. The range is from 1 to $n^{(1-\text{'delta'})}$ where 'n' represents the number of observations in 'x'.

The 'delta.equality' option specifies the range over which a minimum is taken to produce the EL ratio test statistic for the two sample distribution equality test. The lower endpoint in the range is $n^{(0.5+\text{'delta'})}$ and upper endpoint is $\min(n^{(1-\text{'delta'})}, n/2)$ where 'n' corresponds to the number of observations. Acceptable delta values are in the interval (0,0.25). From our experiences, the two sample distribution test is rather robust to the choice of 'delta.equality'.

The 'pvl.Table' is a binary option where when TRUE, the p-value for the test statistic is determined by imputation from a stored table of test statistics and significance levels for common sample sizes. If 'pvl.Table' is FALSE, then the p-value is determined from Monte-Carlo simulations where the number of resamplings is set by 'num.mc'.

Value

Returns a vector of length 2 with test statistic and p-value.

teststat	the value of the test statistic
pvalue	the p-value for the test

Author(s)

Jeffrey C. Miecznikowski, Lori A. Shepherd

References

Jeffrey C. Miecznikowski, Albert Vexler, Lori A. Shepherd (2013). dbEmpLikeGOF: An R Package for Nonparametric Likelihood Ratio Tests for Goodness-of-Fit and Two-Sample Comparisons Based on Sample Entropy. *Journal of Statistical Software*, 54(3), 1-19. <http://www.jstatsoft.org/v54/i03/>

Vexler A, Gurevich G, Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. *Computational Statistics and Data Analysis* 54(2010) 531-545.

Gurevich G, Vexler A, A two-sample empirical likelihood ratio test based on samples entropy. *Statistics and Computing*, 2011.

Examples

```
x <- rnorm(100)
testNorm <- dbEmpLikeGOF(x, testcall="normal")
testUni <- dbEmpLikeGOF(x, testcall="uniform")
testNorm
testUni
y=rnorm(40)
```

```
testDist <- dbEmpLikeGOF(x,y)
testDist
```

returnCutoff	<i>Estimates the statistic cutoff for a target alpha</i>
--------------	--

Description

estimates the test statistic cutoff for significance

Usage

```
returnCutoff(sample.size,
             testcall=c("uniform", "normal", "distribution.equality"),
             targetalpha=.05,
             num.mc=200,
             delta=0.5,
             delta.equality=0.10,
             pvl.Table=FALSE,
             random.seed.flag=TRUE)
```

Arguments

sample.size	number of observations
testcall	Type of distribution: either uniform, normal, or distribution.equality
targetalpha	The significance level for the test.
num.mc	number of simulations to estimate distribution of statistic
delta	an option for changing the minimizing range for the EL ratio test statistic for the normal and uniform distribution.
delta.equality	an option for changing the minimizing range for the EL ratio test statistic for the two sample distribution equality
pvl.Table	logical indicating if value should be calculated based on estimates from data table or by using monte carlo techniques
random.seed.flag	logical if set seed should be set

Details

This function is designed to return the cut-off for significance for the statistics obtained from the density-based EL tests described in Vexler and Gurevich, 2010 and Gurevich and Vexler, 2011. The significance level for the associated cutoffs are specified by the user in 'targetalpha'.

Note 'sample.size' should be a scalar for the normal and uniform tests, but a vector of length two for 'testcall=distribution.equality' denoting the number of observations for each sample.

The 'delta' value should remain at the default value of 0.50. The 'delta' value corresponds to the 'delta' in equation 2.10 (normal) or equation 2.3.2 (uniform) in Vexler and Gurevich, 2010. Essentially this setting controls the range over which a minimum is taken to produce the EL ratio test statistic. The range is from 1 to $n^{(1-\delta)}$ where 'n' represents the number of observations in 'x'.

The 'delta.equality' option specifies the range over which a minimum is taken to produce the EL ratio test statistic for the two sample distribution equality test. The lower endpoint in the range is $n^{(0.5+\delta)}$ and upper endpoint is $\min(n^{(1-\delta)}, n/2)$ where 'n' corresponds to the number of observations. Acceptable delta values are in the interval (0,0.25). From our experiences, the two sample distribution test is rather robust to the choice of 'delta.equality'.

The 'pvl.Table' is a binary option where when TRUE, the test statistic is determined by imputation from a stored table of test statistics and significance levels for common sample sizes. If 'pvl.Table' is FALSE, then the test statistic is determined from a Monte-Carlo simulation where the number of resamplings is controlled by 'num.mc'.

Value

Returns a statistical cutoff value to assess significance at level 'targetalpha'

Author(s)

Jeffrey C. Miecznikowski, Lori A. Shepherd

References

Jeffrey C. Miecznikowski, Albert Vexler, Lori A. Shepherd (2013). dbEmpLikeGOF: An R Package for Nonparametric Likelihood Ratio Tests for Goodness-of-Fit and Two-Sample Comparisons Based on Sample Entropy. *Journal of Statistical Software*, 54(3), 1-19. <http://www.jstatsoft.org/v54/i03/>

Examples

```
returnCutoff(sample.size=50, testcall="normal")
returnCutoff(sample.size=50, testcall="uniform")
returnCutoff(sample.size=10, testcall="distribution.equality")
returnCutoff(sample.size=c(10,15), testcall="distribution.equality")
```

Snow

Snow Dataset

Description

Snowfall dataset

Format

vector of values

Details

This file contains observations of the annual snowfall amounts in Buffalo, New York. 63 as observed from 1910/11 to 1972/73 as listed in *The autoregressive method: a method of approximating and estimating positive functions*. Carmichael, Jean-Pierre. DTIC Document. 1976

Tbars

TBARS Dataset

Description

TBARS dataset

Format

data is a data.frame. x1 is a vector. y1 is a vector

Details

data from a study evaluating biomarkers related to atherosclerotic coronary heart disease

testmat

Pvalue Data Tables

Description

Stores cutoff information for different target alpha values and various sets of data of varying sample size.

Format

data.frame with columns equal to sample size information and rows equal to different target alpha values.

Details

This file contains cutoff information for different target alpha values and various sets of data of varying sample size. normCut for normal distribution, unifCut for uniform distribution, and distEqCut for two vectors compared to each other. These tables are generated for sample sizes 10, 25, 50, 75, 100, 150, 200, 250, 500, 1000, 2000, 5000, and 10000. The target alphas range from .001 to .999 in increments of .001. The delta for normal and uniform is 0.5 and for distribution equality 0.1.

Note

This dataset is used within the getPval function. There is no need for the user to ever call this dataset.

Index

*Topic **datasets**

Baby, [2](#)

Snow, [6](#)

Tbars, [7](#)

testmat, [7](#)

*Topic **methods**

dbEmpLikeGOF, [3](#)

returnCutoff, [5](#)

*Topic **package**

dbEmpLikeGOF-package, [2](#)

Baby, [2](#)

baby (Baby), [2](#)

babyTable (Baby), [2](#)

data (Tbars), [7](#)

dbEmpLikeGOF, [2](#), [3](#)

dbEmpLikeGOF-package, [2](#)

distEqCut (testmat), [7](#)

EmpLikeGOF (dbEmpLikeGOF-package), [2](#)

normCut (testmat), [7](#)

returnCutoff, [2](#), [5](#)

Snow, [6](#)

snow (Snow), [6](#)

TBARS (Tbars), [7](#)

Tbars, [7](#)

testmat, [7](#)

unifCut (testmat), [7](#)

x1 (Tbars), [7](#)

y1 (Tbars), [7](#)