

Package ‘DTDA’

January 20, 2025

Type Package

Title Doubly Truncated Data Analysis

Version 3.0.1

Author Carla Moreira, Jacobo de Uña-Álvarez and Rosa Crujeiras

Maintainer Carla Moreira <carlangmm@gmail.com>

Description Implementation of different algorithms for analyzing randomly truncated data, one-sided and two-sided (i.e. doubly) truncated data. It serves to compute empirical cumulative distributions and also kernel density and hazard functions using different bandwidth selectors. Several real data sets are included.

Imports doParallel, foreach

License GPL-2

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2022-01-12 17:12:43 UTC

Contents

ACS	2
ACSred	3
AIDS	4
AIDS.DT	5
ChildCancer	6
densityDT	7
efron.petrosian	9
EqSRounded	11
hazardDT	12

lynden	14
PDearly	16
PDlate	18
Quasars	19
rsm.DT	20
shen	21

Index	24
--------------	-----------

ACS *Acute Coronary Syndrome data*

Description

The data include information of 939 patients with confirmed diagnosis of type 1 (primary spontaneous) acute coronary syndrome (ACS). Patients were consecutively admitted to the Cardiology Department of two tertiary hospitals in Portugal between August 2013 and December 2014. The age at diagnosis is doubly truncated because of the interval sampling.

Usage

`data(ACS)`

Format

A data frame with 939 observations on the following 5 variables.

`X` a numeric vector, age at diagnosis (in years).

`U` a numeric vector, the elapsed time (in years) between birth and the beginning of the study (August 2013).

`V` a numeric vector, the elapsed time (in years) between birth and end of the study (December 2014).

`Sex` a numeric vector, sex of the participants (0 = female, 1 = male).

`diagnosis` a numeric vector, type of diagnosis at discharge 1 - STEMI (ST elevation myocardial infarction) and 2 - NSTEMI (all others diagnoses).

Details

The age at diagnosis X is doubly truncated due to the interval sampling. The length of the sampling interval ($V-U$) is 1.42 years. The NPMLE of the cumulative distribution function of X does not exist or is not unique for this dataset. The necessary and sufficient graphical condition presented by Xiao and Hudgens (2020) to determine the existence and uniqueness of the NPMLE is not satisfied.

References

Araújo C, Laszczyska O, Viana M, Melão F, Henriques A, Borges A, Severo M, Maciel MJ, Moreira I, Azevedo A (2018) Sex differences in presenting symptoms of acute coronary syndrome: the EPIHeart cohort study. *BMJ Open* **8**.

Xiao J and Hudgens MG (2020) On nonparametric maximum likelihood estimation with double truncation. *Biometrika* **106**, 989-996.

See Also[ACSred](#)**Examples**

```
data(ACS)
str(ACS)
```

`ACSred`*Acute Coronary Syndrome reduced data*

Description

The data include information of 917 patients with confirmed diagnosis of type 1 (primary spontaneous) acute coronary syndrome (ACS). Patients were consecutively admitted to the Cardiology Department of two tertiary hospitals in Portugal between August 2013 and December 2014. The age at diagnosis is doubly truncated because of the interval sampling. This dataset is a reduced sample of the original ACS data, guaranteeing the existence and uniqueness of the NPMLE, according to *Xiao and Hudgens (2020)*

Usage

```
data(ACSred)
```

Format

A data frame with 917 observations on the following 5 variables.

`X` a numeric vector, age at diagnosis (in years).

`U` a numeric vector, the elapsed time (in years) between birth and the beginning of the study (August 2013).

`V` a numeric vector, the elapsed time (in years) between birth and end of the study (December 2014).

`Sex` a numeric vector, sex of the participants (0 = female, 1 = male).

`diagnosis` a numeric vector, type of diagnosis at discharge 1 - STEMI (ST elevation myocardial infarction) and 2 - NSTEMI (all others diagnoses).

Details

The age at diagnosis `X` is doubly truncated due to the interval sampling. The length of the sampling interval (`V-U`) is 1.42 years. The NPMLE of the cumulative distribution function of `X` for the complete data does not exist or is not unique for this dataset.

References

Araújo C, Laszczynska O, Viana M, Melão F, Henriques A, Borges A, Severo M, Maciel MJ, Moreira I, Azevedo A (2018) Sex differences in presenting symptoms of acute coronary syndrome: the EPIHeart cohort study. *BMJ Open* **8**.

Moreira C, de Uña-Álvarez J, Santos AC and Barros H (2021) Smoothing Methods to estimate the hazard rate under double truncation. <https://arxiv.org/abs/2103.14153>.

Xiao J and Hudgens MG (2020) On nonparametric maximum likelihood estimation with double truncation. *Biometrika* **106**, 989-996.

See Also

[ACS](#)

Examples

```
data(ACSred)
str(ACSred)
```

AIDS

AIDS Blood Transfusion Data, right truncated

Description

The data include information on the infection and induction times for 258 adults who were infected with HIV virus and developed AIDS by June 30, 1996. The data consist on the time in years, measured from April 1, 1978, when adults were infected by the virus from a contaminated blood transfusion, and the waiting time to development of AIDS, measured from the date of infection. The induction times are right-truncated.

Usage

```
data(AIDS)
```

Format

A data frame with 258 observations on the following 3 variables.

INFTime a numeric vector, the infection time (years).

INDTime a numeric vector, the induction time (years).

V a numeric vector, the time from HIV infection to the end of the study (years).

Source

Klein and Moeschberger (1997) *Survival Analysis Techniques for Censored and truncated data*. Springer.

References

Lagakos SW and Barraj LM and de Gruttola V (1988) Nonparametric Analysis of Truncated Survival Data, with Applications to AIDS. *Biometrika* **75**, 515–523.

Examples

```
data(AIDS)
str(AIDS)
```

AIDS.DT

AIDS Blood Transfusion Data, doubly truncated

Description

The data include information of transfusions cases of transfusion-related AIDS, corresponding to individuals diagnosed prior to July 1, 1986. Only 295 patients with consistent data, for which the infection could be attributed to a single transfusion or short series were included. Since HIV was unknown before 1982, this implies that cases developing AIDS prior to this date were not reported, leading to a doubly truncated data. The incubation time is doubly truncated due to the interval sampling.

Usage

```
data(AIDS.DT)
```

Format

A data frame with 295 observations on the following 4 variables.

X a numeric vector, the induction or incubation time: time elapsed from HIV infection to AIDS (in months).

U a numeric vector, time from 1982 to HIV infection (in months).

V a numeric vector, time from HIV infection to July 1, 1986 (in months).

AGE a numeric vector, age of the individual at infection (in years).

References

Kalbfleisch JD and Lawless JF (1989) Inference based on retrospective ascertainment: An analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association* **84**, 360–372.

Examples

```
data(AIDS.DT)
str(AIDS.DT)
```

 ChildCancer

 Childhood Cancer Data

Description

This dataset corresponds to all children diagnosed from cancer between January 1, 1999 and December 31, 2003 in the region of North Portugal. The database includes information of 406 children with complete records on the age at diagnosis. Because of the interval sampling, the age at diagnosis is doubly truncated by the time from birth to the end of the study, and time from birth to the beginning of the study.

Usage

```
data("ChildCancer")
```

Format

A data frame with 406 observations on the following 8 variables.

X a numeric vector, age at diagnosis (time in days).

U a numeric vector, time from birth to the beginning of the study (time in days).

V a numeric vector, time from birth to the end of the study (time in days).

ICCGroup a numeric vector, cancer types identified according to the International Classification of Childhood Cancer (ICCC). 1=Leukaemias, 2=Lymphoma, 3=Nervous System Tumour, 4=Neuroblastoma, 5=Retinoblastoma, 6=Renal, 7=Hepatic, 8=Bone, 9=Soft Tissues, 10=Germ Cell, 11=Melanoma and other epithelial tumours, 12=Other Tumours.

Status a numeric vector, the status indicator at the end of the study: 0=alive, 1=dead.

SurvTime a numeric vector, the survival time (time from birth to death or end of the study).

Residence a numeric vector, districts of residence. 1=Braga, 2=Bragança, 3=Porto, 4=Viana do Castelo, 5=Vila Real.

Sex a numeric vector, sex of the participants (1 = female, 2 = male).

Source

The childhood cancer data were gathered from the IPO (Registo Oncológico do Norte) service, kindly provided by Doctor Maria José Bento.

References

Mandel M, de Uña-Álvarez J, Simon DK and Betensky R (2018). Inverse Probability Weighted Cox Regression for Doubly Truncated Data. *Biometrics* **74**, 481-487.

Moreira C and de Uña-Álvarez J (2010) Bootstrapping the NPMLE for doubly truncated data. *Journal of Nonparametric Statistics* **22**, 567-583.

Examples

```
data(ChildCancer)
str(ChildCancer)
```

densityDT	<i>Estimation of a kernel density function under random double truncation</i>
-----------	---

Description

This function provides the nonparametric kernel density estimation of a doubly truncated random variable.

Usage

```
densityDT(X, U, V, bw = "DPI2", from, to, n, wg = NA)
```

Arguments

X	numeric vector with the values of the target variable.
U	numeric vector with the values of the left truncation variable.
V	numeric vector with the values of the right truncation variable.
bw	The smoothing bandwidth to be used, but can also be a character string giving a rule to choose the bandwidth. This must be one of "NR", "DPI1", "DPI2", "LSCV" or "SBoot" with default "DPI2".
from	the left point of the grid at which the density is to be estimated. The default is $\min(X)+1e-04$.
to	the right point of the grid at which the density is to be estimated. The default is $\max(X)-1e-04$.
n	number of evaluation points on an equally spaced grid.
wg	Numeric vector of random weights to correct for double truncation. Default weights correspond to the Efron-Petrosian NPMLE.

Details

The nonparametric kernel density estimation for a variable which is observed under random double truncation is computed as proposed in *Moreira and de Uña-Álvarez (2012)*. As usual in kernel smoothing, the estimator is obtained as a convolution between a kernel function and an appropriate estimator of the cumulative df. Gaussian kernel is used. The automatic bandwidth selection procedures for the kernel density estimator are those proposed in *Moreira and Van Keilegom (2013)*. The automatic bandwidth selection alternatives are appropriate modifications, i.e, taking into account the double truncation issue, of the normal reference rule, two types of plug-in procedures, the least squares cross-validation and a bootstrap based method proposed in *Cao et al. (1994)* and *Sheater and Jones (1991)* for the complete data.

Value

A list containing the following values:

x	the n coordinates of the points where the density is estimated.
y	the estimated density values.
bw	the bandwidth used.

Author(s)

Carla Moreira, de Uña-Álvarez and Rosa Crujeiras

References

Cao R, Cuevas A and González-Manteiga W (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analysis* **17**, 153-176.

Moreira C and de Uña-Álvarez J (2012) Kernel density estimation with doubly truncated data. *Electronic Journal of Statistics* **6**, 501-521.

Moreira C and Van Keilegom I (2013) Bandwidth selection for kernel density estimation with doubly truncated data. *Computational Statistics and Data Analysis* **61**, 107-123.

Sheather S and Jones M (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B* **53**, 683-690.

Silverman BW (1986) Density Estimation. London: Chapman and Hall.

See Also

[hazardDT](#)

Examples

```
set.seed(4321)

n<-50
X <- runif(n, 0, 1)
U <- runif(n,-1/3, 1)
V <- U + 1/3
for (i in 1:n){

while (U[i] > X[i] | V[i] < X[i]){
X[i] <- runif(1, 0, 1)
  U[i] <- runif(1, -1/3, 1)
V[i] <- U[i] + 1/3
}

}

vxDens1<-densityDT(X,U,V,bw="DPI1",0,1,500)
plot(vxDens1, type = "l")
vxDens2<-densityDT(X,U,V,bw="DPI2",0,1,500)
```



```

vxDens3<-densityDT(X,U,V, bw=0.5,0,1,500)
vxDens4<-densityDT(X,U,V,bw="LSCV",0,1,500)

data(Quasars)
densityDT(Quasars[,1],Quasars[,2],Quasars[,3],bw="DPI1",-2.5,2.2,500)
densityDT(Quasars[,1],Quasars[,2],Quasars[,3], bw=0.5,500)

```

efron.petrosian

NPMLE computation with the first Efron-Petrosian algorithm

Description

This function computes the NPMLE for the cumulative distribution function of X observed under one-sided (right or left) and two-sided (double) truncation. It provides simple bootstrap pointwise confidence limits too.

Usage

```

efron.petrosian(X, U = NA, V = NA, wt = NA, error = NA,
  nmaxit = NA, boot = TRUE, B = NA, alpha = NA,
  display.F = FALSE, display.S = FALSE)

```

Arguments

<code>X</code>	Numeric vector with the values of the target variable.
<code>U</code>	Numeric vector with the values of the left truncation variable. If there are no truncation values from the left, put <code>U=NA</code> .
<code>V</code>	Numeric vector with the values of the right truncation variable. If there are no truncation values from the right, put <code>V=NA</code> .
<code>wt</code>	Numeric vector of non-negative initial solution, with the same length as <code>X</code> . Default value is set to $1/n$, being n the length of <code>X</code> .
<code>error</code>	Numeric value. Maximum pointwise error when estimating the density associated to X (f) in two consecutive steps. If this is missing, it is $1e-06$.
<code>nmaxit</code>	Numeric value. Maximum number of iterations. If this is missing, it is set to <code>nmaxit=100</code> .
<code>boot</code>	Logical. If <code>TRUE</code> (default), the simple bootstrap method is applied to lifetime distribution estimation. Pointwise confidence bands are provided.
<code>B</code>	Numeric value. Number of bootstrap resamples. The default <code>NA</code> is equivalent to <code>B=500</code> .
<code>alpha</code>	Numeric value. $(1-\alpha)$ is the nominal coverage for the pointwise confidence intervals.
<code>display.F</code>	Logical. Default is <code>FALSE</code> . If <code>TRUE</code> , the estimated cumulative distribution function associated to X , (F) is plotted.
<code>display.S</code>	Logical. Default is <code>FALSE</code> . If <code>TRUE</code> , the estimated survival function associated to X , (S) is plotted.

Details

The NPMLE for the cumulative distribution function is computed by the first algorithm proposed in *Efron and Petrosian (1999)*. This is an iterative algorithm which converges to the NPMLE after a number of iterations. If the second (respectively third) argument is missing, computation of the Lynden-Bell estimator for right-truncated (respectively left-truncated) data is obtained. Note that individuals with NAs in the three first arguments will be automatically excluded.

Value

A list containing the following values:

time	The timepoint on the curve.
n.event	The number of events that occurred at time t.
events	The total number of events.
density	The estimated density values.
cumulative.df	The estimated cumulative distribution values.
truncation.probs	The probability of X falling into each truncation interval.
S0	error reached in the algorithm.
Survival	The estimated survival values.
n.iterations	The number of iterations used by this algorithm.
B	Number of bootstrap resamples computed.
alpha	The nominal level used to construct the confidence intervals.
upper.df	The estimated upper limits of the confidence intervals for F.
lower.df	The estimated lower limits of the confidence intervals for F.
upper.Sob	The estimated upper limits of the confidence intervals for S.
lower.Sob	The estimated lower limits of the confidence intervals for S.
sd.boot	The bootstrap standard deviation of F estimator.
boot.repeat	The number of resamples done in each bootstrap call to ensure the existence and uniqueness of the bootstrap NPMLE.

Author(s)

Carla Moreira, Jacobo de Uña-Álvarez and Rosa Crujeiras

References

- Efron B and Petrosian V (1999) Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association* **94**, 824-834.
- Lynden-Bell D (1971) A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monograph National Royal Astronomical Society* **155**, 95-118.
- Xiao J and Hudgens MG (2020) On nonparametric maximum likelihood estimation with double truncation. *Biometrika* **106**, 989-996.

See Also[lynden](#)**Examples**

```
## Generating data which are doubly truncated

set.seed(4321)

n<-25
X<-runif(n,0,1)
U<-runif(n,0,0.5)
V<-runif(n,0.5,1)

for (i in 1:n){
  while (X[i]<U[i]|X[i]>V[i]){
    U[i]<-runif(1,0,0.5)
    X[i]<-runif(1,0,1)
    V[i]<-runif(1,0.5,1)
  }
}
efron.petrosian(X=X,U=U,V=V,boot=FALSE,display.F=TRUE,display.S=TRUE)
```

EqSRounded

Equipment-S Rounded Failure Time Data

Description

Digitized data from Figure 2 in *Ye and Tang 2016*. The dataset contains (rounded) observations of 174 failure times of certain devices, observed under interval sampling. Right-runcation is years between installation and 2011 and left truncation corresponds to right-truncation time minus 15 years. The failure time is doubly truncated due to the interval sampling.

Usage

```
data("EqSRounded")
```

Format

A data frame with 174 observations on the following 3 variables.

X a numeric vector, time to failure in years.

U a numeric vector, the number of years between installation and 2011 minus 15 years.

V a numeric vector, the number of years between installation and 2011.

Source

Digitalization of the data plot in the original paper of *Ye and Tang 2016*.

References

Ye ZS and Tang LC (2016) Augmenting the unreturned for field data with information on returned failures only. *Technometrics* **58**, 513-523.

Examples

```
data(EqSRounded)
str(EqSRounded)
```

hazardDT	<i>Estimation of the kernel hazard function under random double truncation</i>
----------	--

Description

This function provides the nonparametric kernel hazard estimation for a variable which is observed under random double truncation, which is defined as a convolution of a kernel function with the NPMLE of the cumulative hazard. Least square cross validation bandwidth selection procedure is implemented too.

Usage

```
hazardDT(X, U, V, bw = "LSCV", from, to, n, wg = NA)
```

Arguments

X	Numeric vector with the values of the target variable.
U	Numeric vector with the values of the left truncation variable.
V	Numeric vector with the values of the righth truncation variable.
bw	The smoothing bandwidth to be used, but can also be a character string giving a rule to choose the bandwidth. This must be "LSCV".
from	the left point of the grid at which the density is to be estimated. The default is $\min(X)+1e-04$.
to	the righth point of the grid at which the density is to be estimated. The default is $\max(X)-1e-04$.
n	number of evaluation points on a equally spaced grid.
wg	Numeric vector of random weights to correct for double truncation. Default weights correspond to the Efron-Petrosian NPMLE.

Details

The nonparametric kernel density estimation for a variable which is observed under random double truncation is computed as proposed in *Moreira et al.(2021)*. As usual in kernel smoothing, the estimator is obtained as a convolution between a kernel function and an appropriate estimator of the cumulative hazard. Gaussian kernel is used. The automatic bandwidth selection procedures for the kernel hazard estimator is the least square cross validation, presented in *Moreira et al. (2021)*.

Value

A list containing the following values:

x	the n coordinates of the points where the hazard is estimated.
y	the estimated hazard values.
bw	the bandwidth used.

Author(s)

Carla Moreira, Jacobo de Uña-Álvarez and Rosa Crujeiras

References

Moreira C, de Uña-Álvarez J, Santos AC and Barros H (2021) Smoothing Methods to estimate the hazard rate under double truncation. <https://arxiv.org/abs/2103.14153>.

See Also

[densityDT](#)

Examples

```
set.seed(4321)

n<-100
X <- runif(n, 0, 1)
U <- runif(n,-1/3, 1)
V <- U + 1/3
for (i in 1:n){

while (U[i] > X[i] | V[i] < X[i]){
X[i] <- runif(1, 0, 1)
  U[i] <- runif(1, -1/3, 1)
V[i] <- U[i] + 1/3
}

}

vxhazard1<-hazardDT(X,U,V,bw=0.3,0,1,500)
vxhazard2<-hazardDT(X,U,V,bw="LSCV",0,1,500)
```

Description

This function computes the NPMLE for the cumulative distribution function of X observed under one-sided (right or left) and two-sided (double) truncation. It provides simple bootstrap pointwise confidence limits too. This function allows for ties in the samples of X , U and V .

Usage

```
lynden(X, U = NA, V = NA, error = NA, nmaxit = NA,
       boot = TRUE, B = NA, alpha = NA,
       display.F = FALSE, display.S = FALSE)
```

Arguments

<code>X</code>	Numeric vector with the values of the target variable.
<code>U</code>	Numeric vector with the values of the left truncation variable. If there are no truncation values from the left, put <code>U=NA</code> .
<code>V</code>	Numeric vector with the values of the right truncation variable. If there are no truncation values from the right, put <code>V=NA</code> .
<code>error</code>	Numeric value. Maximum pointwise error when estimating the density associated to X (f) in two consecutive steps. If this is missing, it is $1e-06$.
<code>nmaxit</code>	Numeric value. Maximum number of iterations. If this is missing, it is set to <code>nmaxit = 100</code> .
<code>boot</code>	Logical. If <code>TRUE</code> (default), the simple bootstrap method is applied to lifetime distribution estimation. Pointwise confidence bands are provided.
<code>B</code>	Numeric value. Number of bootstrap resamples. The default <code>NA</code> is equivalent to <code>B = 500</code> .
<code>alpha</code>	Numeric value. $(1-\alpha)$ is the nominal coverage for the pointwise confidence intervals.
<code>display.F</code>	Logical. Default is <code>FALSE</code> . If <code>TRUE</code> , the estimated cumulative distribution function associated to X , (F) is plotted.
<code>display.S</code>	Logical. Default is <code>FALSE</code> . If <code>TRUE</code> , the estimated survival function associated to X , (S) is plotted.

Details

The NPMLE for the cumulative distribution function is computed by the second algorithm proposed in *Efron and Petrosian (1999)*. This is an iterative algorithm which converges to the NPMLE after a number of iterations. If the second (respectively third) argument is missing, the Lynden-Bell estimator for right-truncated (respectively left-truncated) data is obtained. Note that individuals with NAs in the three first arguments will be automatically excluded.

Value

A list containing the following values:

time	The timepoint on the curve.
n.event	The number of events that occurred at time t.
events	The total number of events.
NJ	The number of individuals in risk considering the left truncation times.
density	The estimated density values.
cumulative.df	The estimated cumulative distribution values.
truncation.probs	The probability of X falling into each truncation interval.
hazard	The estimated hazard values.
S0	error reached in the algorithm.
Survival	The estimated survival values.
n.iterations	The number of iterations used by this algorithm.
B	Number of bootstrap resamples computed.
alpha	The nominal level used to construct the confidence intervals.
upper.df	The estimated upper limits of the confidence intervals for F.
lower.df	The estimated lower limits of the confidence intervals for F.
upper.Sob	The estimated upper limits of the confidence intervals for S.
lower.Sob	The estimated lower limits of the confidence intervals for S.
sd.boot	The bootstrap standard deviation of F estimator.
boot.repeat	The number of resamples done in each bootstrap call to ensure the existence and uniqueness of the bootstrap NPMLE.

Author(s)

Carla Moreira, Jacobo de Uña-Álvarez and Rosa Crujeiras

References

- Efron B and Petrosian V (1999) Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association* **94**, 824-834.
- Lynden-Bell D (1971) A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monograph National Royal Astronomical Society* **155**, 95-118.

See Also

[efron.petrosian](#)

Examples

```

# Generating data which are doubly truncated

set.seed(4321)

n<-25
X<-runif(n,0,1)
U<-runif(n,0,0.25)
V<-runif(n,0.75,1)

for (i in 1:n){
  while (X[i]<U[i]|X[i]>V[i]){
    U[i]<-runif(1,0,0.25)
    X[i]<-runif(1,0,1)
    V[i]<-runif(1,0.75,1)
  }
}
res<-lynden(X=X, U=U, V=V, boot=FALSE, display.F=TRUE, display.S=TRUE)

# Generating data which are right truncated

set.seed(4321)

n<-25
X<-runif(n,0,1)
V<-runif(n,0.75,1)

for (i in 1:n){
  while (X[i]>V[i]){
    X[i]<-runif(1,0,1)
    V[i]<-runif(1,0.75,1)
  }
}
res<-lynden(X=X,U=NA, V=V, boot=FALSE)

```

Description

The sample consists of DNA from 99 Caucasian Parkinson's Disease (PD) patients with earlier onset PD (age 35-55 years). To remove the selection bias related to survival, the study was limited to patients diagnosed from PD who had their DNA sample taken within eight years after onset. Consequently, the age of onset is doubly truncated by the age at blood sampling and this time minus 8 years. This is a situation of interval sampling, the sampling interval being subject-specific.

Usage

```
data("PDearly")
```

Format

A data frame with 99 observations on the following 5 variables.

X a numeric vector, age at onset of PD (in years).

U a numeric vector, age at blood sampling minus 8 years.

V a numeric vector, age at blood sampling.

SNP_A10398G a factor with alleles levels A and G.

SNP_PGC1a a factor with alleles levels A, AG and G.

Details

Clark et al., 2011 hypothesized that the rs8192678 PGC-1a single nucleotide polymorphism (SNP) and the A10398G mitochondrial SNP may influence risk or age of onset of PD. To test these hypotheses, genomic DNA samples from human blood samples were obtained from the National Institute of Neurological Disorders and Stroke (NINDS) Human Genetics DNA and Cell Line Repository at the Coriell Institute for Medical Research (Camden, New Jersey).

Source

Mandel M, de Uña-Álvarez J, Simon DK and Betensky R (2018). Inverse Probability Weighted Cox Regression for Doubly Truncated Data. *Biometrics* **74**, 481-487.

References

Clark J, Reddy S, Zheng K, Betensky RA and Simon DK (2011) Association of PGC-1a polymorphisms with age of onset and risk of Parkinson's disease. *BMC Medical Genetics* **12**, 69.

See Also

[PDlate](#)

Examples

```
data(PDearly)
str(PDearly)
```

PDlate

Parkinson's Disease Data: late onset

Description

The sample consists of DNA from 100 Caucasian Parkinson's Disease (PD) patients with late onset PD (age 63-87 years). To remove the selection bias related to survival, the study was limited to patients diagnosed from PD who had their DNA sample taken within eight years after onset. Consequently, the age of onset is doubly truncated by the age at blood sampling and this time minus 8 years. This is a situation of interval sampling, the sampling interval being subject-specific.

Usage

```
data("PDlate")
```

Format

A data frame with 100 observations on the following 5 variables.

X a numeric vector, age at onset of PD (in years).

U a numeric vector, age at blood sampling minus 8 years.

V a numeric vector, age at blood sampling.

SNP_A10398G a factor with alleles levels A and G.

SNP_PGC1a a factor with alleles levels A, AG and G.

Details

Clark et al., 2011 hypothesized that the rs8192678 PGC-1a single nucleotide polymorphism (SNP) and the A10398G mitochondrial SNP may influence risk or age of onset of PD. To test these hypotheses, genomic DNA samples from human blood samples were obtained from the National Institute of Neurological Disorders and Stroke (NINDS) Human Genetics DNA and Cell Line Repository at the Coriell Institute for Medical Research (Camden, New Jersey).

Source

Mandel M, de Uña-Álvarez J, Simon DK and Betensky R (2018). Inverse Probability Weighted Cox Regression for Doubly Truncated Data. *Biometrics* **74**, 481-487.

References

Clark J, Reddy S, Zheng K, Betensky RA and Simon DK (2011) Association of PGC-1α polymorphisms with age of onset and risk of Parkinson's disease. *BMC Medical Genetics* **12**, 69.

See Also

[PDearly](#)

Examples

```
data(PDlate)
str(PDlate)
```

 Quasars

Quasars Data

Description

The original dataset studied by Efron and Petrosian (1999) comprised independently collected quadruplets of the redshift and the apparent magnitude of a quasar object. Due to experimtnal constraints, the distribution of each luminosity in a log-scale is truncated to a known interval.

Usage

```
data(Quasars)
```

Format

A data frame with 210 observations on the following 3 variables.

y (adj lum) a numeric vector, the log lominosity values.

u (lower) a numeric vector, lower truncation limits.

v (upper) a numeric vector, upper truncation limits.

Details

Quadruplets in the original data set studied by *Efron and Petrosian (1999)* are of the form $(z_i; m_i; a_i; b_i)$, $i = 1, \dots, n$, where z_i is the redshift of the i th quasar and m_i is the apparent magnitude. Due to experimtnal constraints, the distribution of each luminosity in the log-scale ($y_i = t(z_i, m_i)$) is truncated to a known interval $[a_i; b_i]$, where t represents a transformation which depends on the cosmological model assumed (see *Efron and Petrosian (1999)* for details). Quasars with apparent magnitude above b_i were too dim to yield dependent redshifts, and hence they were excluded from the study. The lower limit a_i was used to avoid confusion with non quasar stellar objects.

Source

Vahé Petrosian and Bradley Efron.

References

Boyle BJ, Fong R, Shanks, T and Peterson, BA (1990) A catalogue of faint, UV-excess objects. *Monograph National Royal Astronomical Society* **243**, 1-56.

Efron B and Petrosian V (1999) Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association* **94**, 824-834.

Examples

```
data(Quasars)
str(Quasars)
```

rsim.DT

Random generation functions of doubly truncated data

Description

Random generation functions of doubly truncated data with two different patterns of observational bias.

Usage

```
rsim.DT(n, tau, model=NULL)
```

Arguments

n	number of observations to generate.
tau	length of the observational window.
model	model to be simulated. Number 1 or 2 corresponding to different patterns of observational bias.

Details

If model=1, $U \sim Unif(-\tau, 1)$ and $V = U + \tau$. If model=2, $U \sim Unif(0, 1)^2 \times (\tau + 1) - \tau$ and $V = U + \tau$. In model=1 there is no observational bias due double truncation while in model=2 double truncation induces observational bias.

Value

A matrix with n unit length rows representing the generated values from a doubly truncated data with triplets $(X, U \text{ and } V)$, in which $(U \leq X \leq V)$.

Author(s)

Carla Moreira, Jacobo de Uña-Álvarez and Rosa Crujeiras

Examples

```
set.seed(4321)
rsim.DT(500, 1/2, model=2)
```

shen

*NPMLE computation with Shen algorithm***Description**

This function computes the NPMLE for the cumulative distribution function of X observed under one-sided (right or left) and two-sided (double) truncation. The NPMLE of the joint distribution of the truncation times along with its marginal distributions are also computed. It provides bootstrap pointwise confidence limits too.

Usage

```
shen(X, U = NA, V = NA, wt = NA, error = NA,
      nmaxit = NA, boot = TRUE, boot.type = "simple",
      B = NA, alpha = NA, display.FS = FALSE,
      display.UV = FALSE, plot.joint = FALSE, plot.type = NULL)
```

Arguments

<code>X</code>	Numeric vector with the values of the target variable.
<code>U</code>	Numeric vector with the values of the left truncation variable. If there are no truncation values from the left, put <code>U=NA</code> .
<code>V</code>	Numeric vector with the values of the right truncation variable. If there are no truncation values from the right, put <code>V=NA</code> .
<code>wt</code>	Numeric vector of non-negative initial solution, with the same length as <code>X</code> . Default value is set to $1/n$, being n the length of <code>X</code> .
<code>error</code>	Numeric value. Maximum pointwise error when estimating the density associated to X (f) in two consecutive steps. If this is missing, it is $1e-06$.
<code>nmaxit</code>	Numeric value. Maximum number of iterations. If this is missing, it is set to <code>nmaxit=100</code> .
<code>boot</code>	Logical. If <code>TRUE</code> (default), the simple bootstrap method is applied to lifetime and truncation times distributions estimation. Pointwise confidence bands are provided.
<code>boot.type</code>	A character string giving the bootstrap type to be used. This must be one of "simple" or "obvious", with default "simple".
<code>B</code>	Numeric value. Number of bootstrap resamples. The default <code>NA</code> is equivalent to <code>B=500</code> .
<code>alpha</code>	Numeric value. $(1-\alpha)$ is the nominal coverage for the pointwise confidence intervals.
<code>display.FS</code>	Logical. Default is <code>FALSE</code> . If <code>TRUE</code> , the estimated cumulative distribution function and the estimated survival function associated to X , (F) and (S) respectively, are plotted.
<code>display.UV</code>	Logical. Default is <code>FALSE</code> . If <code>TRUE</code> , the marginal distributions of U (f_U) and V (f_V), are plotted.

<code>plot.joint</code>	Logical. Default is FALSE. If TRUE, the joint distribution of the truncation times is plotted.
<code>plot.type</code>	A character string giving the plot type to be used to represent the joint distribution of the truncation times. This must be one of "image" or "persp", with default NULL.

Details

The NPMLE for the cumulative distribution function is computed by a single algorithm proposed in *Shen (2010)*. This is an iterative algorithm which converges to the NPMLE after a number of iterations. Initial solutions are given by the ordinary empirical distribution functions. If the second (respectively third) argument is missing, computation of the Lynden-Bell estimator for right-truncated (respectively left-truncated) data is obtained. Note that individuals with NAs in the three first arguments will be automatically excluded.

Value

A list containing the following values:

<code>time</code>	The timepoint on the curve.
<code>n.event</code>	The number of events that occurred at time t .
<code>events</code>	The total number of events.
<code>density</code>	The estimated density values associated to X .
<code>cumulative.df</code>	The estimated cumulative distribution values of X .
<code>truncation.probs</code>	The probability of X falling into each truncation interval.
<code>S0</code>	error reached in the algorithm.
<code>Survival</code>	The estimated survival values.
<code>density.joint</code>	The estimated joint densities values associated to (U, V) .
<code>marginal.U</code>	The estimated cumulative univariate marginal values of the U .
<code>marginal.V</code>	The estimated cumulative univariate marginal values of the V .
<code>cumulative.joint</code>	The estimated joint cumulative distribution values.
<code>n.iterations</code>	The number of iterations used by this algorithm.
<code>biasf</code>	The estimated probabilities of observing the lifetimes.
<code>Boot</code>	The type of bootstrap method applied.
<code>B</code>	Number of bootstrap resamples computed.
<code>alpha</code>	The nominal level used to construct the confidence intervals.
<code>upper.df</code>	The estimated upper limits of the confidence intervals for F .
<code>lower.df</code>	The estimated lower limits of the confidence intervals for F .
<code>upper.Sob</code>	The estimated upper limits of the confidence intervals for S .
<code>lower.Sob</code>	The estimated lower limits of the confidence intervals for S .
<code>upper.fU</code>	The estimated upper limits of the confidence intervals for fU .

lower.fU	The estimated lower limits of the confidence intervals for fU.
upper.fV	The estimated upper limits of the confidence intervals for fV.
lower.fV	The estimated lower limits of the confidence intervals for fV.
sd.boot	The bootstrap standard deviation of F estimator.
boot.repeat	The number of resamples done in each bootstrap call to ensure the existence and uniqueness of the bootstrap NPMLE.

Author(s)

Carla Moreira, Jacobo de Uña-Álvarez and Rosa Crujeiras

References

- Lynden-Bell D (1971) A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monograph National Royal Astronomical Society* **155**, 95-118.
- Shen P-S (2010) Nonparametric analysis of doubly truncated data. *Annals of the Institute of Statistical Mathematics* **62**, 835-853.
- Xiao J, Hudgens MG (2020) On nonparametric maximum likelihood estimation with double truncation. *Biometrika* **106**, 989-996.

See Also

[lynden](#)

Examples

```
## Generating data which are doubly truncated

set.seed(4321)
n<-100
X<-runif(n,0,1)
U<-runif(n,0,0.67)
V<-runif(n,0.33,1)
for (i in 1:n){
  while (X[i]<U[i]|X[i]>V[i]){
    U[i]<-runif(1,0,0.67)
    X[i]<-runif(1,0,1)
    V[i]<-runif(1,0.33,1)
  }
}

res<-shen(X,U,V,boot=FALSE, plot.joint=TRUE, plot.type="persp")
```

Index

- * **Bandwidth selection**

- densityDT, [7](#)

- hazardDT, [12](#)

- * **Double truncation**

- densityDT, [7](#)

- hazardDT, [12](#)

- * **Kernel density estimation**

- densityDT, [7](#)

- * **Kernel hazard estimation**

- hazardDT, [12](#)

- * **datasets**

- ACS, [2](#)

- ACSred, [3](#)

- AIDS, [4](#)

- AIDS.DT, [5](#)

- ChildCancer, [6](#)

- EqSRounded, [11](#)

- PDearly, [16](#)

- PDlate, [18](#)

- Quasars, [19](#)

- * **double truncation**

- rsim.DT, [20](#)

- * **iteration**

- efron.petrosian, [9](#)

- lynden, [14](#)

- shen, [21](#)

- * **nonparametric**

- efron.petrosian, [9](#)

- lynden, [14](#)

- shen, [21](#)

- * **observational bias**

- rsim.DT, [20](#)

ACS, [2](#), [4](#)

ACSred, [3](#), [3](#)

AIDS, [4](#)

AIDS.DT, [5](#)

ChildCancer, [6](#)

densityDT, [7](#), [13](#)

efron.petrosian, [9](#), [15](#)

EqSRounded, [11](#)

hazardDT, [8](#), [12](#)

lynden, [11](#), [14](#), [23](#)

PDearly, [16](#), [18](#)

PDlate, [17](#), [18](#)

Quasars, [19](#)

rsim.DT, [20](#)

shen, [21](#)