Package ‘SPECIES’

April 23, 2011

Type Package

Title Statistical package for species richness estimation

Version 1.0

Date 2010-01-24

Author Ji-Ping Wang, <`jzwang@northwestern.edu`>

Maintainer Ji-Ping Wang <`jzwang@northwestern.edu`>

Description SPECIES is an R package for estimation of species richness or diversity.

License GPL-2

Archs i386, ppc, x86_64

R topics documented:

`SPECIES-package` *An R pacakge for species richness estimation*

Description

`SPECIES` provides multiple functions to compute popular estimators for species richness.These estimators include: (1) jackknife estimator by Burnham and Overton 1978, 1979; (2) lower-bound estimator by Chao 1984; (3) coverage-base estimators ACE, ACE-1 by Chao and Lee 1992; (4) coverage-duplication estimator from Poisson-Gamma model by Chao and Bunge 2002; (5) unconditional nonparametric maximum likelihood estimator by Norris and Pollock 1996, 1998; (6) penalized nonparametric maximum likelihood estimator by Wang and Lindsay 2005; and (7) Poisson-compound Gamma model with smooth nonparametric maximum likelihood estimation by Wang 2010.

Details

Package:	SPECIES
Type:	Package
Version:	1.0
Date:	2011-1-22
License:	GPL-2

functions: chao1984, ChaoBunge, ChaoLee1992, jackknife, pcg ,pnpmle, unpmle; data: butterfly, cottontail,EST, insect, microbial, traffic

Author(s)

Ji-Ping Wang, Department of Statistics, Northwestern University

Maintainer: jzwang@northwestern.edu

References

Acinas, S., Klepac-Ceraj, V., Hunt, D., Pharino, C., Ceraj, I., Distel, D., and Polz, M. (2004), Fine-scale phylogenetic architecture of a complex bacterial community. Nature, 430, 551-554.

Bohning, D. and Schon, D., Nonparametric maximum likelihood estimation of population size based on the counting distribution, Journal of the Royal Statistical Society, Series C: Applied Statistics, 54, 721-737.

Burnham, K. P., and Overton,W. S. (1978), Estimation of the Size of a Closed Population When Capture Probabilities Vary Among Animals, Biometrika, 65, 625-633.

Burnham, K. P., and Overton,W. S. (1979), Robust Estimation of Population Size When Capture Probabilities Vary Among Animals, Ecology, 60, 927-936.

Chao, A. (1984), Nonparametric Estimation of the Number of Classes in a Population, Scandinavian Journal of Statistics, 11, 265-270.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43, 783-791.

Chao, A., and Lee, S.-M. (1992), Estimating the Number of Classes via Sample Coverage, Journal of the American Statistical Association, 87, 210-217.

Chao, A., and Bunge, J. (2002), Estimating the Number of Species in a Stochastic Abundance Model, Biometrics, 58, 531-539.

Fisher, R. A., Corbet, A. S., and Williams, C. B. ,(1943), The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population, Journal of Animal Ecology, 12, 42-58.

Hong, S. H., and Bunge, J. and Jeon, S.O. and Epstein, S. (2006), Predicting microbial species richness, Proc. Natl. Acad. Sci, 103, 117-122.

Norris, J. L. I., and Pollock, K. H. (1996), Nonparametric MLE Under Two Closed Capture-Recapture Models With Heterogeneity, Biometrics, 52,639-649.

Norris, J. L. I., and Pollock, K. H.(1998), Non-Parametric MLE for Poisson Species Abundance Models Allowing for Heterogeneity Between Species, Environmental and Ecological Statistics, 5, 391-402.

Simar, L. (1976), Maximum likelihood estimation of a compound Poisson process, Annals of Statistics, 4, 1200-1209.

Wang, J.-P. Z. and Lindsay, B. G. (2005), A penalized nonparametric maximum likelihood approach to species richness estimation. Journal of American Statistical Association, 100(471):942-959.

Wang, J.-P., and Lindsay, B.G. (2008), An exponential partial prior for improving NPML estimation for mixtures, Statistical Methodology, 5:30-45.

Wang, J.-P. (2010), Estimating the species richness by a Poisson-Compound Gamma model, Biometrika, 97(3): 727-740.

Wang, J.-P. (2011), SPECIES: An R Package for Species Richness Estimation, Journal of Statistical Software, 40(9), 1-15, URL: http://www.jstatsoft.org/v40/i09/.

Examples

```
##load library
library(SPECIES)

## "butterfly" is the famous butterfly data by Fisher 1943.

data(butterfly)

##jackknife method
jackknife(butterfly,k=5)

##using only 'ACE' coverage method
ChaoLee1992(butterfly,t=10, method="all")

##using chao1984 lower bound estimator
chao1984(butterfly)

##using Chao and Bunge coverage-duplication method
ChaoBunge(butterfly,t=10)

##penalized NPMLE method
#pnpmle(butterfly,t=15,C=1,b=200)

##unconditonal NPMLE method
#unpmle(butterfly,t=10,C=1,b=200)

##Poisson-compound Gamma method
#pcg(butterfly,t=20,C=1,b=200)
```

`butterfly` *Fisher's butterfly data*

Description

The famous Fisher's butterfly data originally appeared in Fisher 1943. It has been re-analyzed in many publications in the literature.

References

Fisher, R. A., Corbet, A. S., andWilliams, C. B. ,1943, The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population, Journal of Animal Ecology, 12, 42-58.

Examples

```
##load library
library(SPECIES)

##load data that coming with the package.
data(butterfly)
```

`chao1984` *Lower-bound estimator for species richness*

Description

This function calculates the lower-bound estimator by Chao 1984.

Usage

```
chao1984(n,conf=0.95)
```

Arguments

`n`	a matrix or a numerical data frame of two columns. It is also called the "frequency of frequencies" data in literature. The first column is the frequency $j = 1, 2 \ldots$; and the second column is n_j, the number of species observed with j individuals in the sample.
`conf`	a positive number ≤ 1. `conf` specifies the confidence level for confidence interval. The default is 0.95.

Value

The function `chao1984` returns a list of: `Nhat`, `SE` and `CI`.

`Nhat`	point estimate.
`SE`	standard error of the point estimate.
`CI`	confidence interval using a log transformation explained in Chao 1987.

Author(s)

Ji-Ping Wang, Department of Statistics, Northwestern University

References

Chao, A. (1984), Nonparametric Estimation of the Number of Classes in a Population, Scandinavian Journal of Statistics, 11, 265-270.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43, 783-791.

Examples

```
library(SPECIES)

##load data from the package,
## \dQuote{butterfly}is the famous butterfly data by Fisher 1943.

data(butterfly)
chao1984(butterfly)
```

`ChaoBunge` *Coverage-duplication estimator for species richness*

Description

This function calculates coverage-duplication based estimator from a Poisson-Gamma model by Chao and Bunge 2002.

Usage

```
ChaoBunge(n, t = 10,conf = 0.95)
```

Arguments

`n`	a matrix or a numerical data frame of two columns. It is also called the "frequency of frequencies" data in literature. The first column is the frequency $j = 1, 2 \ldots$; and the second column is n_j, the number of species observed with j individuals in the sample.
`t`	a positive integer. `t` is the cutoff value to define the relatively less abundant species to be used in estimation. The frequencies n_j of $j > t$ will not be used in estimating the sample coverage. The default value is `t`=10.
`conf`	a positive number $\leq$ 1. `conf` specifies the confidence level for confidence interval. The default is 0.95.

Value

The function `ChaoBunge` returns a list of: `Nhat`, `SE` and `CI`.

`Nhat`	point estimate.
`SE`	standard error(s) of the point estimate.
`CI`	confidence interval using a log transformation explained in Chao 1987.

Author(s)

Ji-Ping Wang, Department of Statistics, Northwestern University

References

Chao, A. (1984), Nonparametric Estimation of the Number of Classes in a Population, Scandinavian Journal of Statistics, 11, 265-270.

Chao, A., and Bunge, J. (2002), Estimating the Number of Species in a Stochastic Abundance Model, Biometrics, 58, 531-539.

Examples

```
library(SPECIES)

##load data from the package,
##"butterfly" is the famous butterfly data by Fisher 1943.

data(butterfly)

##output estimates from all 4 methods using cutoff t=10
ChaoBunge(butterfly,t=10)
```

`ChaoLee1992` *Coverage-based estimators for species richness*

Description

This function calculates `ACE` and `ACE-1` estimators by Chao and Lee 1992 (`ACE-1` provides further bias correction based on `ACE`).

Usage

```
ChaoLee1992(n, t = 10, method = "all",conf = 0.95)
```

Arguments

`n`	a matrix or a numerical data frame of two columns. It is also called the "frequency of frequencies" data in literature. The first column is the frequency $j = 1, 2 \ldots$; and the second column is n_j, the number of species observed with j individuals in the sample.
`t`	a positive integer. `t` is the cutoff value to define the relatively less abundant species to be used in estimation. The frequencies n_j of $j > t$ will not be used in estimating the sample coverage. The default value is `t`=10.
`method`	a string. It can be any one of "ACE", "ACE-1", or "all". The default is "all".
`conf`	a positive number ≤ 1. `conf` specifies the confidence level for confidence interval. The default is 0.95.

Value

The function `ChaoLee1992` returns a list of: `Nhat`, `SE` and `CI`.

`Nhat`	point estimate of the specified method. If the default `method`="all" is used, the function returns an estimate vector including `ACE, ACE-1`.
`SE`	standard error(s) of the point estimate(s).
`CI`	confidence interval using a log transformation explained in Chao 1987.

Author(s)

Ji-Ping Wang, Department of Statistics, Northwestern University

References

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43, 783-791.

Chao, A., and Lee, S.-M. (1992), Estimating the Number of Classes via Sample Coverage, Journal of the American Statistical Association, 87, 210-217.

Examples

```
library(SPECIES)

##load data from the package,
## "butterfly" is the famous butterfly data by Fisher 1943.

data(butterfly)

##output estimates from all 4 methods using cutoff t=10
ChaoLee1992(butterfly,t=10,method="all")

##output estimates from ACE method using cutoff t=10
ChaoLee1992(butterfly,t=10,method="ACE")
```

`cottontail` *Cottontail data*

Description

The cottontail data was analyzed in Chao 1987

References

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43, 783-791.

Examples

```
##load library
library(SPECIES)

##load data that coming with the package.
data(cottontail)
```

`EST` *EST data*

Description

The Arabidopsis thaliana expressed sequence tag (EST) data originally appeared in Wang and Lindsay 2005. It was recently reanalyzed in Wang 2010. For convenience, the frequency at $j = 17$ is used to denote the total count of species with $j \geq 17$.

References

Wang, J.-P. Z. and Lindsay, B. G. ,(2005), A penalized nonparametric maximum likelilhood approach to species richness estimation. Journal of American Statistical Association, 2005,100(471):942-959

Examples

```
##load library
library(SPECIES)

##load data that coming with the package.
data(EST)
```

`insects` *Insects data*

Description

The insects data was analyzed in Burnham and Overton 1979. The frequency at $j = 6$ is used to denote the total count of species with $j \geq 6$.

References

Burnham, K. P., and Overton,W. S. (1979), Robust Estimation of Population Size When Capture Probabilities Vary Among Animals, Ecology, 60, 927-936.

Examples

```
##load library
library(SPECIES)

##load data that coming with the package.
data(insects)
```

`jackknife` *Jackknife estimator for the species richness*

Description

A function implementing the jackknife estimator of the species number by Burnham and Overton 1978 and 1979.

Usage

```
jackknife(n, k = 5, conf = 0.95)
```

Arguments

`n`	a matrix or a numerical data frame of two columns. It is also called the "frequency of frequencies" data in literature. The first column is the frequency $j = 1, 2 \ldots$; and the second column is n_j, the number of species observed with j individuals in the sample.
`k`	a positive integer. `k` is the specified Jackknife order. The default is `k`=5. Burnham and Overton 1978 and 1979 provided a testing procedure for the maximum order to be used in this estimator. If the specified order `k` or default is greater than the order obtained from the testing procedure, the function will automatically use the determined order rather than `k`. Currently this function only provide jackknife estimate up to order 10.
`conf`	a positive number ≤ 1. `conf` specifies the confidence level for confidence interval. The default is 0.95. `conf` also specifies the critical value in the sequential test for jackknife order.

Value

The function `jackknife` returns a list of: `JackknifeOrder`, `Nhat`, `SE` and `CI`.

`JackknifeOrder`	the jackknife estimator order specified order by the user or determined by the testing procedure.
`Nhat`	jackknife estimate.
`SE`	standard error of the jackknife estimate.
`CI`	confidence interval of the jackknife estimate.

Author(s)

Ji-Ping Wang, Department of Statistics, Northwestern University

References

Burnham, K. P., and Overton,W. S. (1978), Estimation of the Size of a Closed Population When Capture Probabilities Vary Among Animals, Biometrika, 65, 625-633.

Burnham, K. P., and Overton,W. S. (1979), Robust Estimation of Population Size When Capture Probabilities Vary Among Animals, Ecology, 60, 927-936.

Examples

```
library(SPECIES)

##load data from the package,
## "butterfly" is the famous tterfly data by Fisher 1943.

data(butterfly)
jackknife(butterfly,k=5)
```

`microbial` *Microbial species data*

Description

The microbial species data originally appeared in Acinas et al 2004. Recently it was re-analyzed by Bohning and Schon 2005, and Wang 2009.

References

Acinas, S., Klepac-Ceraj, V., Hunt, D., Pharino, C., Ceraj, I., Distel, D., and Polz, M. (2004), Fine-scale phylogenetic architecture of a complex bacterial community. Nature, 430, 551-554.

Hong, S. H., and Bunge, J. and Jeon, S.O. and Epstein, S. (2006), Predicting microbial species richness, Proc. Natl. Acad. Sci, 103, 117-122.

Examples

```
##load library
library(SPECIES)

##load data that coming with the package.
data(microbial)
```

`pcg` *Poisson-compound Gamma estimator for the species richness*

Description

Function to calculate the Poisson-compound Gamma estimators of the species number by Wang 2010. This method is essentially a conditional NPMLE method. The species abundance here is assumed to follow a compound Gamma model. The confidence interval is obtained based on a bootstrap procedure. A Fortran function is called to for the computing. This function requires Fortran compiler installed.

Usage

```
pcg(n,t=35,C=0,alpha=c(1:10),b=200,seed=NULL,conf=0.95,dis=1)
```

Arguments

`n`	a matrix or a numerical data frame of two columns. It is also called the "frequency of frequencies" data in literature. The first column is the frequency $j = 1, 2 \ldots$; and the second column is n_j, the number of species observed with j individuals in the sample.
`t`	a positive integer. `t` is the cutoff value defining the relatively less abundant species to be used in estimation. The default value for `t`=35. The estimator is more sensitive to t compared with pnpmle or unpmle estimators. We recommend to use $t \geq 20$ if the maximum frequency (j) is greater than 20. Otherwise use the maximum frequency of j for `t`.
`C`	integer either 0 or 1. It specifies whether bootstrap confidence interval should be calculated. "`C`=1" for YES and "`C`=0" for NO.The default of `C` is set as 0.
`b`	integer. `b` specifies the number of bootstrap samples to be generated for confidence interval. It is ignored if "`C`=0".
`alpha`	a positive grid for Gamma shape parameter. `alpha` must be a numerical vector for positive numbers. A cross-validation will be used to select a unified shape parameter value for the compound Gamma from the specified "`alpha`" grid. The default "`alpha`" grid is $1, 2, \ldots, 10$.
`conf`	a positive number $\leq$ 1. `conf` specifies the confidence level for confidence interval. The default is 0.95.
`seed`	a single value, interpreted as an integer. Seed for random number generation
`dis`	0 or 1. 1 for on-screen display of the mixture output, and 0 for none.

Details

The `pcg` estimator is computing intensive. The computing of bootstrap confidence interval may take up to a few hours.

Value

The function `pcg` returns a list of: `Nhat`, `CI` (if "`C`=1") and `AlphaModel`.

`Nhat`	point estimate of `N`.
`CI`	bootstrap confidence interval.
`AlphaModel`	unified shape parameter of compound Gamma selected from cross-validation.

Author(s)

Ji-Ping Wang, Department of Statistics, Northwestern University

References

Wang, J.-P. (2010), Estimating the species richness by a Poisson-Compound Gamma model, 97(3): 727-740

Examples

```
library(SPECIES)
##load data from the package,
## \dQuote{butterfly} is the famous butterfly data by Fisher 1943.
```

```
data(butterfly)

##output estimate without confidence interval using cutoff t=15
##pcg(butterfly,t=20,C=0,alpha=c(1:10))

##output estimate with confidence interval using cutoff t=15
#pcg(butterfly,t=20,C=1,alpha=c(1:10),b=200)
```

`pnpmle` *Penalized conditional NPML estimator for species richness*

Description

This function calculate the penalized conditional NPML estimator of the species number by Wang and Lindsay 2005. This estimator was based on the conditional likelihood of a Poisson mixture model. A penalty term was introduced into the model to prevent the boundary problem discussed in Wang and Lindsay 2008. The confidence interval is calculated based on a bootstrap procedure. A Fortran function is called to for the computing.

Usage

```
pnpmle(n,t=15,C=0,b=200,seed=NULL,conf=0.95,dis=1)
```

Arguments

`n`	a matrix or a numerical data frame of two columns. It is also called the "frequency of frequencies" data in literature. The first column is the frequency $j = 1, 2\ldots$; and the second column is n_j, the number of species observed with j individuals in the sample.
`t`	a positive integer. `t` is the cutoff value to define the relatively less abundant species to be used in estimation of the Poisson mixture. The default value is `t`=15. The recommendation is to use `t` ≥ 10.
`C`	integer either 0 or 1. It specifies whether bootstrap confidence interval should be calculated. "`C`=1" for YES and "`C`=0" for NO.The default of `C` is set as 0.
`b`	integer. `b` specifies the number of bootstrap samples to be generated for confidence interval. It is ignored if "`C`=0".
`conf`	a positive number ≤ 1. `conf` specifies the confidence level for confidence interval. The default is 0.95.
`seed`	a single value, interpreted as an integer. Seed for random number generation
`dis`	0 or 1. 1 for on-screen display of the mixture output, and 0 for none.

Value

The function `pnpmle` returns a list of: `Nhat`, `CI` (if "`C`=1").

`Nhat`	Point estimate of `N`
`CI`	bootstrap confidence interval

Author(s)

Ji-Ping Wang,Department of Statistics, Northwestern University

References

Wang, J.-P. Z. and Lindsay, B. G. ,2005, A penalized nonparametric maximum likelihood approach to species richness estimation. Journal of American Statistical Association, 2005,100(471):942-959

Wang, J.-P., and Lindsay, B.G., 2008, An exponential partial prior for improving NPML estimation for mixtures, Statistical Methodology, 2008,5:30-45

Examples

```
library(SPECIES)

##load data from the package,
## \dQuote{butterfly} is the famous butterfly data by Fisher 1943.
#data(butterfly)

##output estimate without confidence interval using cutoff t=15
#pnpmle(butterfly,t=15,C=0)

##output estimate with confidence interval using cutoff t=15
#pnpmle(butterfly,t=15,C=1, b=200)
```

`traffic` *Traffic data*

Description

The traffice data originally appeared in Simar 1976 where the total number of N is knowns as 9461. Recently it was re-analyzed by Bohning and Schon 2005.

References

Simar, L. (1976), Maximum likelihood estimation of a compound Poisson process, Annals of Statistics, 4, 1200-1209. Bohning, D., and Schon, D. (2005), Nonparametric maximum likelihood estimation of population size based on the counting distribution, Journal of the Royal Statistical Society, Series C: Applied Statistics, 54, 721-737.

Examples

```
##load library
library(SPECIES)

##load data that coming with the package.
data(traffic)
chao1984(traffic)
```

`unpmle` *Unconditional NPML estimator for the SPECIES number*

Description

This function calculate the unconditional NPML estimator of the species number by Norris and Pollock 1996, 1998. This estimator was obtained from the full likelihood based on a Poisson mixture model. The confidence interval is calculated based on a bootstrap procedure.

Usage

```
unpmle(n,t=15,C=0,method="W-L",b=200,conf=.95,seed=NULL,dis=1)
```

Arguments

`n`	a matrix or a numerical data frame of two columns. It is also called the "frequency of frequencies" data in literature. The first column is the frequency $j = 1, 2\ldots$; and the second column is n_j, the number of species observed with j individuals in the sample.
`t`	a positive integer. `t` specifies the cutoff value to define the relatively less abundant species to be used in estimation. The default value for t=15. The estimator is fairly insensitive to the choice of `t`. The recommendation is to use $t \geq 10$.
`C`	integer either 0 or 1. It specifies whether bootstrap confidence interval should be calculated. "`C`=1" for YES and "`C`=0" for NO.The default of `C` is set as 0.
`method`	string either "N-P" or "W-L"(default). If `method`="N-P", unconditional NPMLE will be used using an algorithm by Bonhing and Schon (2005). Sometimes this method can be extremely slow. Alternatively one can use method "W-L", an approximate method (but with high precision and much faster) by Wang and Lindsay 2005.
`b`	integer. `b` specifies the number of bootstrap samples for confidence interval. It is ignored if "`C`=0".
`conf`	a positive number ≤ 1. `conf` specifies the confidence level for confidence interval. The default is 0.95.
`seed`	a single value, interpreted as an integer. Seed for random number generation
`dis`	0 or 1. 1 for on-screen display of the mixture output, and 0 for none.

Details

The computing is intensive if `method`="N-P" is used particularly when extrapolation is large. It may takes hours to compute the bootstrap confidence interval. If `method`="W-L" is used, computing usually is much much faster. Estimates from both methods are often identical.

Value

The function `unpmle` returns a list of: `Nhat`, `CI` (if "`C`=1")

`Nhat`	point estimate of N
`CI`	bootstrap confidence interval.

Note

The unconditional NPML estimator is unstable from either `method='N-P'` or `method='W-L'`. Extremely large estimates may occur. This is also reflected in that the upper confidence bound often greatly vary from different runs of bootstrap procedure. In contrast the penalized NPMLE by `pnpmle` function is much more stable.

Author(s)

Ji-Ping Wang, Department of Statistics, Northwestern University

References

Norris, J. L. I., and Pollock, K. H. (1996), Nonparametric MLE Under Two Closed Capture-Recapture Models With Heterogeneity, Biometrics, 52,639-649.

Norris, J. L. I., and Pollock, K. H.(1998), Non-Parametric MLE for Poisson Species Abundance Models Allowing for Heterogeneity Between Species, Environmental and Ecological Statistics, 5, 391-402.

Bonhing, D. and Schon, D., (2005), Nonparametric maximum likelihood estimation of population size based on the counting distribution, Journal of the Royal Statistical Society, Series C: Applied Statistics, 54, 721-737.

Wang, J.-P. Z. and Lindsay, B. G. ,(2005), A penalized nonparametric maximum likelihood approach to species richness estimation. Journal of American Statistical Association, 2005,100(471):942-959

Examples

```
library(SPECIES)

##load data from the package,
## "butterfly" is the famous butterfly data by Fisher 1943.

data(butterfly)


##output estimate without confidence interval using cutoff t=15
#unpmle(butterfly,t=15,C=0)

##output estimate with confidence interval using cutoff t=15
#unpmle(butterfly,t=15,C=1,b=200)
```

Index