Package 'preseqR'

May 22, 2017

Type Package

Title Predicting the Number of Species in a Random Sample

Version 3.1.0

Date 2017-05-21

Author Chao Deng, Timothy Daley and Andrew D. Smith

Maintainer Chao Deng <chaodeng@usc.edu>

Description The relation between the number of species and the number of individuals in a random sample is a classic problem back to Fisher (1943) <doi:10.2307/1411>. We generalize this problem to estimate the number of species represented at least r times in a random sample. In particular when r=1, it becomes the classic problem. We use a mixture of Poisson processes to model sampling procedures and apply a nonparametric empirical Bayes approach to obtain an estimator. For more information on preseqR, see Deng C, Daley T and Smith AD (2015) <doi:10.1007/s40484-015-0049-7> and Deng C and Smith AD (2016) <arXiv:1607.02804v2>.

License GPL-3

Imports polynom, graphics, stats

NeedsCompilation no

Repository CRAN

Date/Publication 2017-05-21 23:25:35 UTC

R topics documented:

`preseqR-package` *An R package for estimating the number of species represented at least r times*

Description

preseqR provides functions to estimate the number of species represented at least r times in a random sample based on an initial sample. Functions work through rational function approximations to a modified Good and Toulmin's (1956) non-parametric empirical Bayes power series estimator. The rational function approximation is then boosted to an estiamtor for the number species represented at least r times, based on a relation bewteen the number of species represented at least r times and the number of species represented at least once.

Details

functions:

ds.mincount.bootstrap

ds.mincount

ztpois.mincount

ztnb.mincount

boneh.mincount

chao.mincount

pois.mincount

nb.mincount

preseqR.ztnb.em

preseqR.simu.hist

preseqR.nonreplace.sampling

preseqR.interpolate.mincount

data: FisherButterflyHist, ShakespeareWordHist

Author(s)

Chao Deng, Timothy Daley, and Andrew D. Smith

Maintainer: Chao Deng <chaodeng@usc.edu>

References

Baker, G. A., & Graves-Morris, P. (1996). Pade approximants (Encyclopedia of Mathematics and its Applications vol 59).

Boneh, S., Boneh, A., & Caron, R. J. (1998). Estimating the prediction function and the number of unseen species in sampling with replacement. Journal of the American Statistical Association, 93(441), 372-379.

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

Chao, A., & Shen, T. J. (2004). Nonparametric prediction in species sampling. Journal of agricultural, biological, and environmental statistics, 9(3), 253-269.

Cohen Jr, A. C. (1960). Estimating the parameters of a modified Poisson distribution. Journal of the American Statistical Association, 55(289), 139-143.

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. Nature methods, 10(4), 325-327.

Deng C, Daley T and Smith AD (2015). Applications of species accumulation curves in large-scale biological data analysis. Quantitative Biology, 3(3), 135-144. URL http://dx.doi.org/10.1007/s40484-015-0049-7.

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know?. Biometrika, 63(3), 435-447.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. The annals of Statistics, 1-26.

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

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.

Good, I. J., & Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. Biometrika, 43(1-2), 45-63.

Heck Jr, K. L., van Belle, G., & Simberloff, D. (1975). Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. Ecology, 1459-1461.

Kalinin V (1965). Functionals related to the poisson distribution and statistical structure of a text. Articles on Mathematical Statistics and the Theory of Probability pp. 202-220.

`boneh.mincount` *Estimating the expected number of species represented r or more times*

Description

The function estimates the expected number of species represented at least r times in a random sample based on the initial sample using a nonparametric approach by Boneh et al. (1998).

Usage

```
boneh.mincount(n, r=1)
```

Arguments

n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of species with each species represented j times in the initial sample. The first column must be sorted in an ascending order.
r	A vector of positive integers. Default is 1.

Value

The constructed estimator for the number of species represneted at least r times in a sample. The input of the estimator is a vector of sampling efforts t, i.e. the relative sample sizes comparing with the initial sample. For example, t = 2 means the sample is twice the size of the initial sample.

Author(s)

Chao Deng

References

Boneh, S., Boneh, A., & Caron, R. J. (1998). Estimating the prediction function and the number of unseen species in sampling with replacement. Journal of the American Statistical Association, 93(441), 372-379.

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterflyHist)

## construct the estimator for the number of species
## represented at least once, twice or three times in a sample
boneh.estimator <- boneh.mincount(FisherButterflyHist, r=1:3)

## The number of species represented at least once, twice or three times
## when the sample size is 10 or 20 times of the initial sample
boneh.estimator(c(10, 20))
```

---

chao.mincount *Estimating the expected number of species represented r or more times*

Description

The function estimates the expected number of species represented at least r times in a random sample based on the initial sample using a nonparametric approach by Chao and Shen (2004).

Usage

```
chao.mincount(n, r=1, k=10)
```

Arguments

n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of species with each species represented j times in the initial sample. The first column must be sorted in an ascending order.
r	A vector of positive integers. Default is 1.
k	A cutoff for common species. Default is 10.

Value

The constructed estimator for the number of species represneted at least r times in a sample. The input of the estimator is a vector of sampling efforts t, i.e. the relative sample sizes comparing with the initial sample. For example, t = 2 means the sample is twice the size of the initial sample.

Author(s)

Chao Deng

References

Chao, A., & Shen, T. J. (2004). Nonparametric prediction in species sampling. Journal of agricultural, biological, and environmental statistics, 9(3), 253-269.

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterflyHist)

## construct the estimator for the number of species
## represented at least once, twice or three times in a sample
chao.estimator <- chao.mincount(FisherButterflyHist, r=1:3)

## The number of species represented at least once, twice or three times
## when the sample size is 10 or 20 times of the initial sample
chao.estimator(c(10, 20))
```

Dickens *Dickens' vocabulary*

Description

Words frequencies of a collection of Charles Dickens from Project Gutenberg

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of unique words appeared j times in a collection of Charles Dickens.

References

http://zipfr.r-forge.r-project.org/

Examples

```
##load library
library(preseqR)

##load data
data(Dickens)
```

ds.mincount *Estimating the expected number of species represented r or more times*

Description

The function estimates the expected number of species represented at least r times in a random sample based on the initial sample.

Usage

```
ds.mincount(n, r=1, mt=100)
```

Arguments

n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of species with each species represented j times in the initial sample. The first column must be sorted in an ascending order.
mt	An positive integer constraining possible rational function approximations. Default is 100.
r	A vector of positive integers. Default is 1.

Details

The difference between this function and ds.mincount.bootstrap is that no bootstrapping for the initial sample. Therefore the function could be less stable than estimates by bootstrap. However, this function is much faster. In general, we recommand ds.mincount.bootstrap for estimating the expected number of species represented at least r times in a sample.

See ds.mincount.bootstrap for more information.

Value

FUN	The constructed estimator for the number of species represneted at least r times in a sample. The input of the estimator is a vector of sampling efforts t, i.e. the relative sample sizes comparing with the initial sample. For example, t = 2 means the sample is twice the size of the initial sample.
FUN.elements	A list of two components for the estimator. The estimator can be expressed as $$\hat{E}(S_r(t)) = \sum_{i=1}^{l} c_i \left(\frac{t}{t - x_i}\right)^r.$$ PF.elements contains both coefficients c_i and roots x_i.
M	The number of terms used when applying rational function approximation to the power series of the average discovery rate.
M.adjust	The number of terms in the estiamtor, equal to l

Author(s)

Chao Deng

References

Kalinin V (1965). Functionals related to the poisson distribution and statistical structure of a text. Articles on Mathematical Statistics and the Theory of Probability pp. 202-220.

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. Nature methods, 10(4), 325-327.

Examples

```
## load library
library(preseqR)

## import data
data(ShakespeareWordHist)

## construct the estimator for the number of unique word
## represented at least once, twice or twenty times in a sample
estimator = ds.mincount(ShakespeareWordHist, r=c(1,2,20))

## print the elements of the estimator
estimator$FUN.elements
```

```
## The number of unique words represented at least once, twice or twenty times
## when the sample size is 10 or 20 times of the initial sample
estimator$FUN(c(10, 20))
```

`ds.mincount.bootstrap` *Estimating the number of species represented r or more times*

Description

The function estimates the expected number of species represented at least r times in a random sample based on the initial sample. The initial sample is bootstrapped to improve the stability of estimates.

Usage

```
ds.mincount.bootstrap(n, r=1, mt=100, times=100)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of species with each species represented j times in the initial sample. The first column must be sorted in an ascending order.
`mt`	An positive integer constraining possible rational function approximations. Default is 100.
`r`	A vector of positive integers. Default is 1.
`times`	An positive integer representing the minimum required number of successful estimation. Default is 100. See detail below.

Details

Under a mixture of Poisson models, the expected number of species represented at least r times in a random sample can be expressed as higher derivatives of the expected number of species represented at least once. We first use rational function approximations to the modified Good and Toulmin's (1956) non-parametric empirical Bayes power series to estimate the average discovery rate. By differentiating the rational function approximation, we obtain an estimator for the number of species represented at least r times in a random sample.

Value

`FUN.nobootstrap`	The estimator constructed based on the initial sample by the function. No bootstrap procesure is involved.
`FUN.bootstrap`	The bootstrap samples from an initial sample are used to construct estimators. The median value of these estimators are estimates of the number of species represented at least r times in a sample.
`var`	The estimated variance for the estimator `FUN.nobootstrap` by bootstrap.

Author(s)

Chao Deng

References

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

Kalinin V (1965). Functionals related to the poisson distribution and statistical structure of a text. Articles on Mathematical Statistics and the Theory of Probability pp. 202-220.

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. Nature methods, 10(4), 325-327.

Examples

```
## load library
#library(preseqR)

## import data
#data(FisherButterflyHist)

## estimate the number of species captured at least once, twice or 20 times
## as a function of the number of individuals

# result = ds.mincount.bootstrap(FisherButterflyHist, r=c(1,2, 20), times=10)

## estimates of the number of unique words appeared at least once, twice or three
## times when the sample size 10 times the size of the initial sample

## estimates by the function ds.mincount
# result$FUN.nobootstrap$FUN(10)

## estimates by the bootstrapped estimator
# result$FUN.bootstrap(10)
```

FisherButterflyHist *Fisher's butterfly data*

Description

Frequencies data of butterflies collected in the Malay peninsula was from Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943).

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of butterflies captured j times in the sample.

References

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, Table 1,2.

Examples

```
##load library
library(preseqR)

##load data
data(FisherButterflyHist)
```

`preseqR.interpolate.mincount`

Interpolating the number of species represented r or more times

---

Description

Interpolating the expected number of species represented at least r times in a random sample based on an initial sample.

Usage

```
preseqR.interpolate.mincount(ss, n, r=1)
```

Arguments

`ss`	An positive double equal to the step size between samples.
`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of species with each species represented j times in the initial sample. The first column must be sorted in an ascending order.
`r`	A positive integer.

Details

Assume that a random sample (subsample) follows a multivariant hypergeometric distribution given an initial sample. The expected number of unique species represented at least r times in the subsample is then calculated by an expended version of the formula in Heck Jr, KL. et al. (1975).

Value

A two-column matrix for the number of species represented at least r times in a random sample. The first column is the size of the random sample; the second column is the expected number of species represented at least r times in the sample.

NULL if failed.

Author(s)

Chao Deng

References

Heck Jr, K. L., van Belle, G., & Simberloff, D. (1975). Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. Ecology, 1459-1461.

Examples

```
## load library
library(preseqR)

## import data
data(ShakespeareWordHist)

## The expected number of species represented twice or more in a random sample
## The step size is 1e5; the initial sample is "ShakespeareWordHist"
preseqR.interpolate.mincount(n=ShakespeareWordHist, ss=1e5, r=2)
```

---

preseqR.nonreplace.sampling

Sampling without replacement

Description

Generating a histogram by subsampling without replacement.

Usage

```
preseqR.nonreplace.sampling(size, n)
```

Arguments

size	An positive integer representing the size of the subsample.
n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of species with each species represented j times in the initial sample. The first column must be sorted in an ascending order.

Details

The function sample() in R is used to implement the function. We wrap the sample() function in a way that both input and output are histograms.

Value

A two-column matrix as a subsample. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of species with each species represented j times in the subsample.

Author(s)

Chao Deng

References

https://stat.ethz.ch/R-manual/R-patched/library/base/html/sample.html

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterflyHist)

## generate a subsample of size 1000.
preseqR.nonreplace.sampling(size=1000, FisherButterflyHist)
```

---

preseqR.simu.hist *Simulating a histogram*

Description

Generating a histogram based on a Poisson mixture model.

Usage

```
preseqR.simu.hist(L=1e8, N, FUN)
```

Arguments

L	A positive integer, the number of species in a population.
N	A positive interger, the simulated sample size.
FUN	An RNG generating non negative real number.

Details

The function uses a compound Poisson model to generate a sample of size n. It assumes for each species the number of individuals captured in a sample follows a Poisson process. The Poisson rates among species are generated by a given function FUN per unit of sampling effort. Under this statistical assumption, for a given sample size N, the number of individuals in the sample for each species follow a multinomial distributions.

The function FUN must take an argument indicating the number of random numbers generated and return a vector of generated numbers.

Value

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of species with each species represented j times in the initial sample. The first column must be sorted in an ascending order.

Author(s)

Chao Deng

Examples

```
## load library
library(preseqR)
## construct a RNG
f <- function(n) {
  rgamma(n, shape=0.5, scale=1)
}

preseqR.simu.hist(L=1e5, N=1, f)
```

---

preseqR.ztnb.em *Fitting a zero-truncated negative binomial distribution*

Description

This function fits a zero-truncated negative binomial (ZTNB) distribution to the initial sample. Since the species with zero observations are missed in the sample, an EM algorithm is used to estimate the parameters assuming the number of individuals for each species follows a Negative Binomial distribution with the zero counts as a missing latent data.

Usage

```
preseqR.ztnb.em(n, size = SIZE.INIT, mu = MU.INIT)
```

Arguments

n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of species with each species represented j times in the initial sample. The first column must be sorted in an ascending order.
size	A positive double setting the initial value of the parameter size in a negative binomial distribution for the EM algorithm. Default value is 1.
mu	A positive double setting the initial value of the parameter mu in a negative binomial distribution for the EM algorithm. Default value is 0.5.

Details

See the supplement of Daley and Smith (2013).

Value

size	The estimate of the parameter size in the negative binomial.
mu	The estimate of the parameter mu in the negative binomial.
loglik	Log-likelihood under estimated ZTNB.

Author(s)

Chao Deng

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterflyHist)

## print the parameters of a fitting negative binomial distribution
preseqR.ztnb.em(FisherButterflyHist)
```

ShakespeareWordHist *Shakespeare's word type frequencies*

Description

The Shakespeare's word type frequencies data was from Efron, B., & Thisted, R. (1976).

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of unique words appeared j times in Shakespeare's work.

References

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know?. Biometrika, 63(3), 435-447.

Examples

```
##load library
library(preseqR)

##load data
data(ShakespeareWordHist)
```

Twitter *Social network*

Description

Following relationships of Twitter's social network

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of users with j followers.

References

Zafarani R, Liu H (2009) Social computing data repository at ASU.

Examples

```
##load library
library(preseqR)

##load data
data(Twitter)
```

ztnb.mincount *Estimating the expected number of species represented r or more times*

Description

The function estimates the expected number of species represented at least r times in a random sample based on the initial sample using zero truncated negative binomial model.

Usage

```
ztnb.mincount(n, r=1, size=SIZE.INIT, mu=MU.INIT)
```

Arguments

n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of species with each species represented j times in the initial sample. The first column must be sorted in an ascending order.
r	A vector of positive integers. Default is 1.
size	A positive double setting the initial value of the parameter size in a negative binomial distribution for the EM algorithm. Default value is 1.
mu	A positive double setting the initial value of the parameter mu in a negative binomial distribution for the EM algorithm. Default value is 0.5.

Details

The statistical assumption is that for each species the number of individuals in a sample follows a Poisson distribution. The Poisson rate `lambda` obeys a latent gamma distribution. So the random variable `X`, which is the number of species represented x (x > 0) times, follows a zero-truncated negative binomial distribution. The unknown parameters are estimated by the function `preseqR.ztnb.em`. Based on the estimated distribution, we calculate the expected number of species in a random sample. Details of the estimation procedure see supplement of Daley T. and Smith AD. (2013).

Value

The constructed estimator for the number of species represneted at least r times in a sample. The input of the estimator is a vector of sampling efforts t, i.e. the relative sample sizes comparing with the initial sample. For example, t = 2 means the sample is twice the size of the initial sample.

Author(s)

Chao Deng

References

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. Nature methods, 10(4), 325-327.

See Also

`preseqR.ztnb.em`

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterflyHist)

## construct the estimator for the number of species
## represented at least once, twice or three times in a sample
ztnb.estimator <- ztnb.mincount(FisherButterflyHist, r=1:3)

## The number of species represented at least once, twice or three times
## when the sample size is 10 or 20 times of the initial sample
ztnb.estimator(c(10, 20))
```

`ztpois.mincount` *Estimating the expected number of species represented r or more times*

Description

The function estimates the expected number of species represented at least r times in a random sample based on the initial sample using zero truncated Poisson distributino.

Usage

```
ztpois.mincount(n, r=1)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of species with each species represented j times in the initial sample. The first column must be sorted in an ascending order.
`r`	A vector of positive integers. Default is 1.

Details

The statistical assumption is that for each species the number of individuals in a sample follows a Poisson distribution. The Poisson rate `lambda` is the same among all species. So the random variable `X`, which is the number of species represented x (x > 0) times, follows a zero-truncated Poisson distribution. The unknown parameters are estimated by Cohen (1960). Based on the estimated distribution, we calculate the expected number of species in a random sample.

Value

The constructed estimator for the number of species represneted at least r times in a sample. The input of the estimator is a vector of sampling efforts t, i.e. the relative sample sizes comparing with the initial sample. For example, t = 2 means the sample is twice the size of the initial sample.

Author(s)

Chao Deng

References

Cohen Jr, A. C. (1960). Estimating the parameters of a modified Poisson distribution. Journal of the American Statistical Association, 55(289), 139-143.

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterflyHist)

## construct the estimator for the number of species
## represented at least once, twice or three times in a sample
ztpois.estimator <- ztpois.mincount(FisherButterflyHist, r=1:3)

## The number of species represented at least once, twice or three times
## when the sample size is 10 or 20 times of the initial sample
ztpois.estimator(c(10, 20))
```

Index