Package 'OnomasticDiversity'

July 21, 2025

Type Package

Title Onomastic Diversity Measures

Version 0.1

Date 2024-02-07

Author Maria Jose Ginzo Villamayor [aut, cre]

Maintainer Maria Jose Ginzo Villamayor <mariajose.ginzo@usc.es>

Depends R(>= 4.2.0)

Imports sqldf

Description
> Different measures which can be used to quantify similarities between regions. These measures are isonymy, isonymy between, Lasker distance, coefficients of Hedrick and Nei. In addition, it calculates biodiversity indices such as Margalef, Menhinick, Simpson, Shannon, Shannon-Wiener, Sheldon, Heip, Hill Numbers, Geometric Mean and Cressie and Read statistics.

License GPL-2

LazyLoad yes

Repository CRAN

Encoding UTF-8

NeedsCompilation no

RoxygenNote 7.2.3

Date/Publication 2024-02-08 21:10:09 UTC

Contents

OnomasticDiversity-package

Onomastic Diversity Measures

Description

Different measures which can be used to quantify similarities between regions. These measures are isonymy, isonymy between, Lasker distance, coefficients of Hedrick and Nei. In addition, it calculates biodiversity indices such as Margalef, Menhinick, Simpson, Shannon, Shannon-Wiener, Sheldon, Heip, Hill Numbers, Geometric Mean and Cressie and Read statistics.

Details

The DESCRIPTION file:

Package:	OnomasticDiversity
Type:	Package
Title:	Onomastic Diversity Measures
Version:	0.1
Date:	2024-02-07
Authors@R:	c(person("Maria Jose", "Ginzo Villamayor", role = c("aut", "cre"),email="mariajose.ginzo@usc.es"))
Author:	Maria Jose Ginzo Villamayor [aut, cre]
Maintainer:	Maria Jose Ginzo Villamayor <mariajose.ginzo@usc.es>
Depends:	R(>= 4.2.0)
Imports:	sqldf
Description:	Different measures which can be used to quantify similarities between regions. These measures are ison
License:	GPL-2
LazyLoad:	yes
Packaged:	2024-02-07 11:46:22 UTC;
Repository:	CRAN
Encoding:	UTF-8
NeedsCompilation:	no
RoxygenNote:	7.2.3

Index of help topics:

```
OnomasticDiversity-package
                        Onomastic Diversity Measures
fCressieRead            Cressie and Read
fGeneralisedMean        Calculate the Generalised Mean
fGeometricMean          Calculate the Geometric Mean
fHeip                   Calculate the Heip's diversity index
fHill                   Calculate the Hill's diversity numbers
fIsonymy                Calculate the Isonymy within a region
fIsonymyAll             Calculate the Isonymy, Isonymy between regions,
                        Lasker distances, Euclidean distance and Nei's
                        distances
fMargalef               Calculate the Margalef's diversity index
fMenhinick              Calculate the Menhinick's diversity index
fPielou                 Calculate the Pielou's diversity index
fShannon                Calculate the Shannon-Weaver diversity index
fSheldon                Calculate the Sheldon's diversity index
fSimpson                Calculate the Simpson's diversity index
fSimpsonInf             Calculate the Simpson's diversity index and the
                        inverse
namesmengal16           namesmengal16 data
nameswomengal16         nameswomengal16 data
surnamesgal14           surnamesgal14 data
```

This package computes the different measures which can be used to quantify similarities between regions. These measures are isonymy, isonymy between, Lasker distance, coefficients of Hedrick and Nei. A diversity index is a numerical measure of how many different types (such as species) are present in a dataset (a community), as well as the evolutionary relationships among the individuals distributed throughout those types, such as richness, divergence, and evenness. These indicators are numerical representations of biodiversity in several dimensions (richness, evenness, and dominance). Then, this package calculates biodiversity indices such as Margalef, Menhinick, Simpson, Shannon, Shannon-Wiener Sheldon, Heip, Hill Numbers, Geometric Mean and Cressie and Read statistics.

Author(s)

Maria Jose Ginzo Villamayor [aut, cre]

Maintainer: Maria Jose Ginzo Villamayor <mariajose.ginzo@usc.es>

References

Buckland, S.T., Studeny, A.C., Magurran, A.E., Illian, J.B., & Newson, S.E. (2011). The geometric mean of relative abundance indices: a biodiversity measure with a difference. *Ecosphere*, **2**(9), art.100. <https://doi.org/10.1890/ES11-00186.1>

Cressie, Noel and Read, Timothy RC (1984) Multinomial goodness-of-fit tests. *Computational Statistics and Data Analysis*, **46**(3), 440–464. <http://www.jstor.org/stable/2345686>

Sheldon, A. L. (1969). Equitability indices: dependence on the species count. *Ecology*, **50**, 466–467. <https://doi.org/10.2307/1933900>

Simpson (1949) Measurement of diversity. *Nature*, **163**. <https://doi.org/10.1038/163688a0>

Studeny, A.C. (2012). *Quantifying Biodiversity Trends in Time and Space*. PhD thesis, University of St Andrews. <https://research-repository.st-andrews.ac.uk/bitstream/handle/10023/3414/AngelikaStudenyPhDThesis.pdf?se

van Strien, A.J., Soldaat, L.L., & Gregory, R.D. (2012). Desirable mathematical properties of indicators for biodiversity change. *Ecological Indicators*, **14**, 202–208. <https://doi.org/10.1016/j.ecolind.2011.07.007>

See Also

[fCressieRead](), [fGeneralisedMean](), [fGeometricMean](), [fHeip](), [fHill](), [fIsonymy](), [fIsonymyAll](), [fMargalef](), [fMenhinick](), [fPielou](), [fShannon](), [fSheldon](), [fSimpson](), [fSimpsonInf](), [fGeneralisedMean](), [fHeip]()

---

fCressieRead   *Cressie and Read*

Description

This function obtains the Cressie and Read statistics introduced by Noel Cressie and Timothy Read. It is a method for quantifying species biodiversity that can be adapted to the context of onomastics.

Usage

```
fCressieRead(x, number, population, ni, location, lambda)
```

Arguments

x	dataframe of the data values.
number	name of a variable which represents number of individuals of each species.
population	name of variable which represents total number of individuals.
ni	name of variable which represent number of species.
location	name of variable which represent represents the grouping element.
lambda	free parameter.

Details

For a community i, Cressie and Read (1984) introduced the following parametric form for a generalised statistic $I_n(\lambda) = \frac{2}{\lambda(\lambda+1)} \sum_{k \in S_i} n_{ki} \left[\left(\frac{n_{ki}}{n/S_i} \right)^{\lambda} - 1 \right]$, where n_{ki} represents the number of individuals of species k in a sample (in the population is N_{ki}), S_i represents all species at the community, species richness, and λ is a free parameter.

Varying the value of λ gets different statistics. If $\lambda = -1$ and $\lambda = 0$, $I_n(\lambda)$ is not defined, but in any case, limits $\lambda = -1$ and $\lambda = 0$ can be taken.

In onomastic context, n_{ki} ($\approx N_{ki}$) denotes the absolute frequency of surname k in region i ($\approx$ community diversity context i).

Value

A dataframe containing the following components:

location   represents the grouping element, for example the communities / regions.

cressieRead   the value of Cressie and Read statistics.

Author(s)

Maria Jose Ginzo Villamayor

References

Cressie, Noel and Read, Timothy RC (1984) Multinomial goodness-of-fit tests. *Computational Statistics and Data Analysis*, **46**(3), 440–464.

See Also

[fHill](#)

Examples

```
data(surnamesgal14)
result = fCressieRead(x= surnamesgal14 , number="number",
population="population", location = "muni", ni="ni",
lambda = 2)
result
```

---

fGeneralisedMean   *Calculate the Generalised Mean*

Description

This function obtains the generalised mean of relative abundances for a collection of species introduced by Angelika C. Studeny. It is a method for quantifying species biodiversity that can be adapted to the context of onomastics.

Usage

```
fGeneralisedMean (x, pki, pki0, s, location, lambda)
```

Arguments

x   dataframe of the data values for each species not null (because if you have a sample, there might be species that are not represented).

pki   name of a variable which represents the relative frequency for each species.

pki0   variable which represents the relative frequency for each species not null (because if you have a sample, there might be species that are not represented).

location   name of a variable which represents the grouping element.

s   vector which represents total number of species.

lambda   free parameter.

Details

For a community i, the generalised mean of relative abundances is defined by $M_t(\lambda) = \left[\frac{1}{S_i} \sum_{k \in S_i} \left(\frac{N_{ki}^t}{N_{ki}^{t0}} \right)^\lambda \right]^{\frac{1}{\lambda}}$,

where N_{ki}^t denotes the number of individuals of species k at times t, $t0$ is the baseline year and S_i are all species at the community, species richness, and λ can be any non-zero real number.

In onomastic context, N_{ki}^t denotes the absolute frequency of surname k in region ($\approx$ community diversity context) i at times t.

Value

A dataframe containing the following components:

location   represents the grouping element, for example the communities / regions.

generalisedMean
                  the value of generalised mean.

Author(s)

Maria Jose Ginzo Villamayor

References

Studeny, A.C. (2012). *Quantifying Biodiversity Trends in Time and Space*. PhD thesis, University of St Andrews.

See Also

fMargalef, fMenhinick, fPielou, fShannon, fSheldon, fSimpson, fSimpsonInf, fGeometricMean, fHeip

Examples

```
library(sqldf)
data(surnamesgal14)

loc <- length(unique(surnamesgal14$muni))

apes2=sqldf('select  muni, count(surname) as ni,
sum(number) as population from surnamesgal14
group by muni;')

result = fGeneralisedMean(x= surnamesgal14[surnamesgal14$number != 0,],
pki="pki", pki0=surnamesgal14[surnamesgal14$number != 0,"pki"],
location  = "muni", s = apes2$ni[1:loc], lambda = 1 )
result
```

```
data(namesmengal16)

loc <- length(unique(namesmengal16$muni))

namesmengal16$pki <- (namesmengal16$number /
namesmengal16$population)

names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from namesmengal16
group by muni;')

result = fGeneralisedMean(x= namesmengal16[namesmengal16$number != 0,],
pki="pki", pki0=namesmengal16[namesmengal16$number != 0,"pki"],
location  = "muni", s = names2$ni[1:loc], lambda = 1 )
result

data(nameswomengal16)

loc <- length(unique(nameswomengal16$muni))

nameswomengal16$pki <- (nameswomengal16$number /
nameswomengal16$population)

names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from nameswomengal16
group by muni;')

result = fGeneralisedMean(x= nameswomengal16[nameswomengal16$number != 0,],
pki="pki", pki0=nameswomengal16[nameswomengal16$number != 0,"pki"],
location  = "muni", s = names2$ni[1:loc], lambda = 1 )
result
```

---

fGeometricMean *Calculate the Geometric Mean*

Description

This function obtains the geometric mean introduced by Stephen Terrence Buckland and coauthors. It is a method for quantifying species biodiversity that can be adapted to the context of onomastic.

Usage

```
fGeometricMean(x, pki, pki0, s, location)
```

Arguments

x	dataframe of the data values for each species not null (because if you have a sample, there might be species that are not represented).
pki	name of a variable which represents the relative frequency for each species.

pki0	name of a variable which represents the relative frequency for each species at initial time point.
s	vector which represents total number of species.
location	represents the grouping element.

Details

For a community i, the geometric mean of relative abundances is defined by $G_t = \exp\left(\frac{1}{S_i}\sum_{k \in S_i}\log\frac{N_{ki}^t}{N_{ki}^{t_0}}\right)$, where N_{ki}^t denotes the number of individuals of species k at times t, t_0 is the baseline year and S_i are all species at the community, species richness.

In onomastic context, N_{ki}^t denotes the absolute frequency of surname k in region ($\approx$ community diversity context) i at times t.

Value

A dataframe containing the following components:

location	represents the grouping element, for example the communities / regions.
geometricMean	the value of geometric mean.

Author(s)

Maria Jose Ginzo Villamayor

References

Buckland, S.T., Studeny, A.C., Magurran, A.E., Illian, J.B., & Newson, S.E. (2011). The geometric mean of relative abundance indices: a biodiversity measure with a difference. *Ecosphere*, **2**(9), art.100.

Studeny, A.C. (2012). *Quantifying Biodiversity Trends in Time and Space*. PhD thesis, University of St Andrews.

van Strien, A.J., Soldaat, L.L., & Gregory, R.D. (2012). Desirable mathematical properties of indicators for biodiversity change. *Ecological Indicators*, **14**, 202–208.

See Also

fMargalef, fMenhinick, fPielou, fShannon, fSheldon, fSimpson, fSimpsonInf, fGeneralisedMean, fHeip

Examples

```
library(sqldf)
data(surnamesgal14)
loc <- length(unique(surnamesgal14$muni))

apes2=sqldf('select  muni, count(surname) as ni,
sum(number) as population from surnamesgal14
group by muni;')
```

```
surnamesgal14$pki0 <- surnamesgal14$pki

result = fGeometricMean (x= surnamesgal14[surnamesgal14$number != 0,],
pki="pki", pki0="pki0" , location  = "muni",
s = apes2$ni[1:loc])
result

data(namesmengal16)
loc <- length(unique(namesmengal16$muni))

names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from namesmengal16
group by muni;')

namesmengal16$pki <- (namesmengal16$number /
namesmengal16$population)
namesmengal16$pki0 <- namesmengal16$pki

result = fGeometricMean (x= namesmengal16[namesmengal16$number != 0,],
pki="pki", pki0="pki0" , location  = "muni",
s = names2$ni[1:loc])
result

data(nameswomengal16)
loc <- length(unique(nameswomengal16$muni))

names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from nameswomengal16
group by muni;')

nameswomengal16$pki <- (nameswomengal16$number /
nameswomengal16$population)
nameswomengal16$pki0 <- nameswomengal16$pki

result = fGeometricMean (x= nameswomengal16[nameswomengal16$number != 0,],
pki = "pki", pki0 = "pki0", location  = "muni",
s = names2$ni[1:loc])
result
```

---

fHeip	*Calculate the Heip's diversity index*

Description

This function obtains the Heip's diversity index introduced by Carlo H. R. Heip. It is a method for quantifying species biodiversity that can be adapted to the context of onomastic.

Usage

```
fHeip (x, k, n, location, s)
```

Arguments

x	dataframe of the data values for each species not null (because if you have a sample, there might be species that are not represented).
k	name of a variable which represents absolute frequency for each species.
n	name of a variable which represents total number of individuals.
location	represents the grouping element.
s	vector which represents total number of species.

Details

For a community i, the Heip's diversity index is defined by $E_{He} = \frac{2^{H'}-1}{S_i-1}$ where H' is the Shannon diversity index and S_i are all species at the community, species richness. This index varies from 0 to 1 and measures how equally the species richness contributes to the total abundance of the community.

In onomastic context, S_i are all surnames in region ($\approx$ community diversity context) i.

Value

A dataframe containing the following components:

location	represents the grouping element, for example the communities / regions.
heip	the value of the Heip's diversity index.

Author(s)

Maria Jose Ginzo Villamayor

References

Heip, C. (1974). A New Index Measuring Evenness. *Journal of the Marine Biological Association of the United Kingdom*, **54**(3), 555–557.

See Also

fMargalef, fMenhinick, fPielou, fShannon, fSheldon, fSimpson, fSimpsonInf, fGeneralisedMean, fGeometricMean.

Examples

```
library(sqldf)
data(surnamesgal14)
loc <- length(unique(surnamesgal14$muni))


apes2=sqldf('select  muni, count(surname) as ni,
sum(number) as population from surnamesgal14
group by muni;')
```

```
result = fHeip (x= surnamesgal14[surnamesgal14$number != 0,],
k="number", n="population", location = "muni",
s = apes2$ni[1:loc] )
result

data(namesmengal16)
loc <- length(unique(namesmengal16$muni))


names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from namesmengal16
group by muni;')


result = fHeip (x= namesmengal16[namesmengal16$number != 0,],
k="number", n="population", location  = "muni",
s = names2$ni[1:loc] )
result


data(nameswomengal16)
loc <- length(unique(nameswomengal16$muni))


names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from nameswomengal16
group by muni;')


result = fHeip (x= nameswomengal16[nameswomengal16$number != 0,],
k="number", n="population", location  = "muni",
s = names2$ni[1:loc] )
result
```

---

fHill   *Calculate the Hill's diversity numbers*

Description

This function obtains the Hill's diversity numbers introduced by M. O. Hill. It is a method for quantifying species biodiversity that can be adapted to the context of onomastic.

Usage

```
fHill(x, k, n, location, lambda)
```

Arguments

x	dataframe of the data values for each species.
k	name of a variable which represents absolute frequency for each species.

n	name of a variable which represents total number of individuals.
location	represents the grouping element.
lambda	free parameter.

Details

For a community i, the Hill's diversity numbers are defined by the expression $J(\lambda) = \left(\sum\limits_{k \in S_i} p_{ki}^{\lambda} \right)^{\frac{1}{1-\lambda}}$ with the restriction $\lambda \geq 0$ where p_{ki} represents the relative frequency of species k and S_i are all species at the community, species richness, and λ is a free parameter. (This is equivalent to the exponential of Renyi's generalised entropy). The Renyi entropy of order λ, where $\lambda \geq 0$ and $\lambda \neq 1$, is defined as $\mathrm{H}_\lambda(X) = \frac{1}{1-\lambda} \log \left(\sum\limits_{i=1}^{n} p_i^{\lambda} \right)$ Here, X is a discrete random variable with possible outcomes in the set $\mathcal{A} = \{x_1, x_2, \ldots, x_n\}$ and corresponding probabilities $p_i \doteq \Pr(X = x_i)$ for $i = 1, \ldots, n$. The logarithm is conventionally taken to be base 2, especially in the context of information theory where bits are used. If the probabilities are $p_i = 1/n$ for all $i = 1, \ldots, n$, then all the Renyi entropies of the distribution are equal: $\mathrm{H}_\lambda(X) = \log n$. In general, for all discrete random variables X, $\mathrm{H}_\lambda(X)$ is a non-increasing function in λ..

Particular cases of λ values: $\lambda = 0$, $J(0) = S_i$, it corresponds species richness; $\lambda = 1$, $J(1) = e^{H_t}$, it corresponds the exponential of Shannon's entropy; and $\lambda = 2$, $J(2) = D_{S_i}$, it corresponds the 'inverse' Simpson index.

In onomastic context, p_{ki} denotes the relative frequency of surname k in region ($\approx$ community diversity context) i and S_i are all surnames in region i.

Value

A dataframe containing the following components:

location	represents the grouping element, for example the communities / regions.
hill	the value of the Hill's diversity index.

Author(s)

Maria Jose Ginzo Villamayor

References

Hill, M. O. (1973). Diversity and Evenness: a unifying notation and its consequences. *Ecology*, **54**, 427–32.

See Also

fCressieRead.

Examples

```
data(surnamesgal14)
result = fHill (x= surnamesgal14, k="number", n="population",
location  = "muni", lambda= 0)
result

data(namesmengal16)
result = fHill (x= namesmengal16, k="number", n="population",
location  = "muni", lambda= 0)
result

data(nameswomengal16)
result = fHill (x= nameswomengal16, k="number", n="population",
location  = "muni", lambda= 0)
result
```

---

fIsonymy   *Calculate the Isonymy within a region*

Description

This function obtains the isonymy within a region i which has an associated collection S_i of surnames.

Usage

```
fIsonymy(x, category)
```

Arguments

x   a vector of relative frequency squared for each surname.

category   represents the grouping element, for example the regions.

Details

Isonymy is defined as $I_i = \sum\limits_{k \in S_i} p_{ki}^2$ where p_{ki} denotes the relative frequency of surname k in region i.

In diversity context, p_{ki} denotes the relative frequency of species k in community ($\approx$ region onomastic context) i and S_i are all species in community i.

Value

A dataframe containing the following components:

category   represents the grouping element, for example the regions / communities.

x   the value of isonymy.

Author(s)

Maria Jose Ginzo Villamayor

References

Crow J.F. and Mange A.P., (1965). Measurement of inbreeding from the frequency of marriages between persons of the same surname. *Eugenics Quarterly*, **12**(4), 199–203.

Barrai, I., Scapoli, C., Beretta, M., Nesti, C., Mamolini, E., and Rodriguez–Larralde, A., (1996). Isonymy and the genetic structure of Switzerland. I: The distributions of surnames. *Annals of Human Biology*, **23**, 431–455.

See Also

fIsonymyAll.

Examples

```
data(surnamesgal14)
surnamesgal14$pki2 <- (surnamesgal14$number / surnamesgal14$population)^2
result = fIsonymy(surnamesgal14$pki2, surnamesgal14$namuni)
result

data(namesmengal16)
namesmengal16$pki2 <- (namesmengal16$number / namesmengal16$population)^2
result = fIsonymy(namesmengal16$pki2, namesmengal16$namuni)
result

data(nameswomengal16)
nameswomengal16$pki2 <- (nameswomengal16$number / nameswomengal16$population)^2
result = fIsonymy(nameswomengal16$pki2, nameswomengal16$namuni)
result
```

---

fIsonymyAll	*Calculate the Isonymy, Isonymy between regions, Lasker distances, Euclidean distance and Nei's distances*

Description

This function obtains the Isonymy, Isonymy between regions, Lasker distance, Euclidean distance and Nei's distances and Hedrick's coefficient.

Usage

```
fIsonymyAll (x, n, location, union, measure)
```

Arguments

x   data frame with the data.

n   number of the locations in the data frame.

location   name of a variable which represents the location in the data.

union   variable to be used to search for matching surnames in two locations.

measure   name of a variable which represents the relative frequency for each surname.

Details

Values of Isonymy, Isonymy between regions, Lasker distance, Euclidean distance and Nei's distances and Hedrick's coefficient.

Surname (dis)similarity among regions can be quantified by different measures. Consider index $i = 1, \ldots, n$ for denoting a certain geographical region (for two regions, (i, j)). Each region has an associated collection S_i of surnames, and for a pair of regions, the collection of all the surnames in them is denoted by $S_{ij}(S_{ij} = S_i \cup S_j)$. The total number of surnames in a certain region i is denoted by n_i. Surnames will be denoted by indices k and l.

Isonymy is defined as $I_i = \sum_{k \in S_i} p_{ki}^2$ where p_{ki} denotes the relative frequency of surname k in region i. Isonymy can be also extended as a measure of population similarities between groups. Under the assumption of a common origin, isonymy between two regions i and j is defined as $I_{ij} = \sum_{k \in S_{ij}} p_{k_i} p_{k_j}$.

Other different measures of the isonymic distance between a pair of locations can be derived from isonymy between. For instance, the Lasker distance is given by $L = -\log(I_{ij})$.

Lasker distance can be interpreted as a measure of similarity between to areas, where large distance indicate less similarity in surname composition. Nevertheless, Lasker distance is not the only option to quantify surname similarity. Other common coefficients are the Euclidean distance and Nei's distance, both of them given by $E = \sqrt{1 - \sum_{k \in S_{ij}} \sqrt{p_{ki} p_{kj}}}$ and $N = -\log\left(\frac{I_{ij}}{\sqrt{I_i I_j}}\right)$, respectively. Finally, Hedrick's coefficient gives a standardized measure of isonymy using a procedure similar to that utilized in the calculation of a correlation coefficient. Specifically: $H_{ij} = \dfrac{2 \sum_{k \in S_{ij}} p_{ki} p_{kj}}{\left(\sum_{k \in S_{ij}} p_{ki}^2 + \sum_{k \in S_{ij}} p_{kj}^2\right)}$, with $i, j = 1 \ldots, n$.

In diversity context, p_{ki} denotes the relative frequency of species k in community ($\approx$ region onomastic context) i and S_i are all species in community i.

Value

A list containing the following components:

isonymy   data frame with two columns and number of rows the number of regions / communities (n). For each location, it returns the value of the isonymy.

isonymy.btw   the value of isonymy between. Matrix, $n \times n$.

hedrick   the value of Hedrick's coefficient. Matrix, $n \times n$.

nei	the value of Nei's distance. Matrix, $n \times n$.
lasker	the value of Lasker distance. Matrix, $n \times n$.
distE	the value of Euclidean distance. Matrix, $n \times n$.

Author(s)

Maria Jose Ginzo Villamayor

References

Barrai, I., Scapoli, C., Beretta, M., Nesti, C., Mamolini, E., and Rodriguez–Larralde, A., (1996) Isonymy and the genetic structure of Switzerland. I: The distributions of surnames. *Annals of Human Biology*, **23**, 431–455.

Cavalli-Sforza, L. L., and Edwards, A. W. F., (1967), Phylogenetic analysis models and estimation procedures. *American Journal of Human Genetics*, **19**, 233 257.

Hedrick, P. W. (1971), A new approach to measuring genetic similarity. *Evolution*, **25**: 276–280.

Lasker, G. W. (1977) A coefficicnt of relationship by isonymy: a method for estimating the genetic relationship between populations. *Human Biology*, **49**, 489–493.

Mikerezi, I., Shina, E. Scapoli, C., Barbujani, G. Mamolini, E., Sandri, M., Carrieri, A., Rodriguez–Larralde, A. and Barrai, I. (2013). Surnames in Albania: a study of the population of Albania through isonymy. *Annals of Human Genetics*, **77**, 232–243.

Nei, M.(1973). The theory and estimation of genetic distance. In *Genetic Structure of Populations*, edited by N. E. Morton, (Honolulu: University Press of Hawaii), 45–54.

Weiss, V. 1980. Inbreeding and genetic distance between hierarchically structured populations measured by surname frequencies. *Mankind Quarterly*, **21**, 135–149.

See Also

[fIsonymy](#).

Examples

```
data(surnamesgal14)
result = fIsonymyAll (x= surnamesgal14, n= 314, location = 'muni',
union = 'surname', measure = 'pki')
result

data(namesmengal16)
namesmengal16$pki <- (namesmengal16$number /
namesmengal16$population)
result = fIsonymyAll (x= namesmengal16, n= 313, location = 'muni',
union = 'name', measure = 'pki')
result

data(nameswomengal16)
nameswomengal16$pki <- (nameswomengal16$number /
nameswomengal16$population)
result = fIsonymyAll (x= nameswomengal16, n= 313, location = 'muni',
```

```
  union = 'name', measure = 'pki')
  result
```

---

fMargalef   *Calculate the Margalef's diversity index*

Description

This function obtains the Margalef's diversity index which is a species diversity index developed by Ramon Margalef Lopez during the 1950s. It is a method for quantifying species biodiversity that can be adapted to the context of onomastic.

Usage

```
fMargalef(x, s, n, location)
```

Arguments

x	dataframe which contains the number of species and population for each location.
s	name of a variable which represents number of species.
n	name of a variable which represents total number of individuals.
location	name of a variable which represents represents the grouping element.

Details

For a community i, the Margalef's diversity index is defined by $R_1 = \frac{S_i - 1}{\ln(N_i)}$, where S_i represents the number of species (richness) and N_i represents the total number of individuals in all S_i.

In onomastic context, N_i denotes the number of individuals in region ($\approx$ community diversity context) i and S_i represents the total number of surnames.

Value

A dataframe containing the following components:

location	represents the grouping element, for example the communities / regions.
margalef	the value of the Margalef's diversity index.

Author(s)

Maria Jose Ginzo Villamayor

References

Margalef D.R., (1958), Information theory in ecology. *International Journal of General Systems*, **3**, 36–71.

See Also

fMenhinick, fPielou, fShannon, fSheldon, fSimpson, fSimpsonInf, fGeneralisedMean, fGeometricMean, fHeip.

Examples

```
library(sqldf)
data(surnamesgal14)

apes2=sqldf('select  muni, count(surname) as ni,
sum(number) as population from surnamesgal14
group by muni;')

result = fMargalef (x= apes2, s="ni", n="population", location  = "muni")
result

data(namesmengal16)

names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from namesmengal16
group by muni;')

result = fMargalef (x= names2, s="ni", n="population", location  = "muni")
result

data(nameswomengal16)

names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from nameswomengal16
group by muni;')

result = fMargalef (x= names2, s="ni", n="population", location  = "muni")
result
```

---

fMenhinick   *Calculate the Menhinick's diversity index*

Description

This function obtains the Menhinick's diversity index introduced by Edward F. Menhinick. It is a method for quantifying species biodiversity that can be adapted to the context of onomastic.

Usage

```
fMenhinick(x, s, n, location)
```

Arguments

x	dataframe which contains the number of species and population for each location.
s	name of a variable which represents number of species.
n	name of a variable which represents total number of individuals.
location	name of a variable which represents represents the grouping element.

Details

For a community i, the Menhinick's diversity index is defined by $R_2 = \frac{s_i}{\sqrt{N_i}}$, where s_i represents the number of species (richness) and N_i represents the total number of individuals in all s_i.

In onomastic context, N_i denotes the number of individuals in region ($\approx$ community diversity context) i and s_i represents the total number of surnames.

Value

A dataframe containing the following components:

location	represents the grouping element, for example the communities / regions.
menhinick	the value of the Menhinick's diversity index.

Author(s)

Maria Jose Ginzo Villamayor

References

Menhinick E.F. (1964) A comparison of some species-individuals diversity indices applied to samples of field insects. *Ecology*, **45**, 859–861.

See Also

[fMargalef](), [fPielou](), [fShannon](), [fSheldon](), [fSimpson](), [fSimpsonInf](), [fGeneralisedMean](), [fGeometricMean](), [fHeip]().

Examples

```
library(sqldf)
data(surnamesgal14)

apes2=sqldf('select  muni, count(surname) as ni,
sum(number) as population from surnamesgal14
group by muni;')

result = fMenhinick(x= apes2, s="ni", n="population",
location  = "muni")
result

data(namesmengal16)
```

```
names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from namesmengal16
group by muni;')

result = fMenhinick(x= names2, s="ni", n="population",
location  = "muni")
result

data(nameswomengal16)

names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from nameswomengal16
group by muni;')

result = fMenhinick(x= names2, s="ni", n="population",
location  = "muni")
result
```

---

fPielou   *Calculate the Pielou's diversity index*

Description

This function obtains the Pielou's diversity index which is an index that measures diversity along with species richness introduced by Evelyn Chrystalla Pielou. It is a method for quantifying species biodiversity that can be adapted to the context of onomastic.

Usage

```
fPielou(x, k, n, location, s)
```

Arguments

x	dataframe of the data values for each species not null (because if you have a sample, there might be species that are not represented).
k	name of a variable which represents absolute frequency for each species.
n	name of a variable which represents total number of individuals.
location	represents the grouping element.
s	vector which represents number of species.

Details

For a community i, the Pielou's diversity index is defined by $J' = \frac{H'}{\log_2 S_i}$, where H' denotes the Shannon-Wiener index and $\log_2 S_i$ denotes the maximum diversity $H'_{\max}$. Pielou's index is the Shannon-Weiner index computed for the sample S_i and represents a measure of Evenness of the community. If all species are represented in equal numbers in the sample, then $J' = 1$. If one species strongly dominates J' is close to zero.

In onomastic context, S_i are all surnames in region ($\approx$ community diversity context) i.

Value

A dataframe containing the following components:

location   represents the grouping element, for example the communities / regions.

pielou   the value of the Pielou's diversity index.

Author(s)

Maria Jose Ginzo Villamayor

References

Pielou, E. C. (1966) The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, **13**, 131-144.

See Also

fMargalef, fMenhinick, fShannon, fSheldon, fSimpson, fSimpsonInf, fGeneralisedMean, fGeometricMean, fHeip.

Examples

```
library(sqldf)
data(surnamesgal14)

apes2=sqldf('select  muni, count(surname) as ni,
sum(number) as population from surnamesgal14
group by muni;')

result = fPielou (x= surnamesgal14[surnamesgal14$number != 0,],
k="number", n="population", location  = "muni", s = apes2$ni )
result

data(namesmengal16)

names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from namesmengal16
group by muni;')

result = fPielou (x= namesmengal16[namesmengal16$number != 0,],
k="number", n="population", location  = "muni", s = names2$ni )
result

data(nameswomengal16)

names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from nameswomengal16
group by muni;')

result = fPielou (x= nameswomengal16[nameswomengal16$number != 0,],
k="number", n="population", location  = "muni", s = names2$ni )
result
```

---

fShannon   *Calculate the Shannon-Weaver diversity index*

Description

This function obtains the Shannon-Weaver diversity index introduced by Claude Elwood Shannon. This diversity measure came from information theory and measures the order (or disorder) observed within a particular system. It is a method for quantifying species biodiversity that can be adapted to the context of onomastics.

Usage

```
fShannon(x, k, n, location)
```

Arguments

x	dataframe of the data values for each species not null (because if you have a sample, there might be species that are not represented)..
k	name of a variable which represents absolute frequency for each species.
n	name of a variable which represents total number of individuals.
location	represents the grouping element.

Details

For a community i, the index of Shannon-Weaver is defined by the expression $H' = -\sum_{k \in S_i} (p_{ki} \log_2 p_{ki})$, where p_{ki} represents the relative frequency of species k, because $p_{ki} = \frac{N_{ki}}{N_i}$, (where N_{ki} denotes the number of individuals of species k and N_i total number of individuals in all S_i species at the community, species richness. This index is related to the weighted geometric mean of the proportional abundances of the types.

In onomastic context, p_{ki} denotes the relative frequency of surname k in region ($\approx$ community diversity context) i and S_i are all surnames in region i.

Value

A dataframe containing the following components:

location	represents the grouping element, for example the communities / regions.
shannon	the value of the Shannon-Weaver diversity index.

Author(s)

Maria Jose Ginzo Villamayor

References

Shannon C.E. (1948). *A mathematical theory of communication*. Bell System Technical Journal, **27**, 379–423.

Shannon C.E., Weaver W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press. USA, **96**. pp. 117.

See Also

fMargalef, fMenhinick, fPielou, fSheldon, fSimpson, fSimpsonInf, fGeneralisedMean, fGeometricMean, fHeip.

Examples

```
data(surnamesgal14)
result = fShannon (x= surnamesgal14[surnamesgal14$number != 0,],
k="number", n="population", location  = "muni" )
result

data(namesmengal16)
result = fShannon (x= namesmengal16[namesmengal16$number != 0,],
k="number", n="population", location  = "muni" )
result

data(nameswomengal16)
result = fShannon (x= nameswomengal16[nameswomengal16$number != 0,],
k="number", n="population", location  = "muni" )
result
```

---

fSheldon   *Calculate the Sheldon's diversity index*

Description

This function obtains the Sheldon's diversity index introduced by A. L. Sheldon. It is a method for quantifying species biodiversity that can be adapted to the context of onomastics.

Usage

```
fSheldon (x, k, n, location, s)
```

Arguments

x	dataframe of the data values for each species not null (because if you have a sample, there might be species that are not represented)..
k	name of a variable which represents absolute frequency for each species.
n	name of a variable which represents total number of individuals.
location	represents the grouping element.
s	vector which represents number of species.

Details

For a community i, the Sheldon's diversity index is defined by $E_{She} = \frac{2^{H'}}{S_i}$, where H' denotes the Shannon-Wiener index and S_i represents the number of species (richness).

In onomastic context, S_i are all surnames in region ($\approx$ community diversity context) i.

Value

A dataframe containing the following components:

location   represents the grouping element, for example the communities / regions.

sheldon   the value of the Pielou's diversity index.

Author(s)

Maria Jose Ginzo Villamayor

References

Sheldon, A. L. (1969). Equitability indices: dependence on the species count. *Ecology*, **50**, 466–467.

See Also

fMargalef, fMenhinick, fPielou, fShannon, fSimpson, fSimpsonInf, fGeneralisedMean, fGeometricMean, fHeip.

Examples

```
library(sqldf)
data(surnamesgal14)
apes2=sqldf('select  muni, count(surname) as ni,
sum(number) as population from surnamesgal14
group by muni;')

result = fSheldon (x= surnamesgal14[surnamesgal14$number != 0,],
k="number", n="population", location  = "muni",
s = apes2$ni)
result

data(namesmengal16)
names2=sqldf('select  muni, count(name) as ni,
sum(number) as population from namesmengal16
group by muni;')

result = fSheldon (x= namesmengal16[namesmengal16$number != 0,],
k="number", n="population", location  = "muni",
s = names2$ni)
result

data(nameswomengal16)
names2=sqldf('select  muni, count(name) as ni,
```

```
sum(number) as population from nameswomengal16
group by muni;')

result = fSheldon (x= nameswomengal16[nameswomengal16$number != 0,],
k="number", n="population", location  = "muni",
s = names2$ni)
result
```

---

fSimpson   *Calculate the Simpson's diversity index*

Description

This function obtains the Simpson's diversity index and the inverse introduced by Edward Hugh Simpson. It was the first index used in ecology. It is a method for quantifying species biodiversity that can be adapted to the context of onomastics.

Usage

```
fSimpson(x, k, n, location)
```

Arguments

x	dataframe of the data values for each species.
k	name of a variable which represents absolute frequency for each species
n	name of a variable which represents total number of individuals.
location	represents the grouping element.

Details

For a community i, the Simpson's diversity index is defined by $D_{S_i} = \sum\limits_{k \in S_i} p_{ki}^2$, where p_{ki} represents the relative frequency of species k, because $p_{ki} = \frac{N_{ki}}{N_i}$, (where N_{ki} denotes the number of individuals of species k and N_i total number of individuals in all S_i species at the community, species richness. The Simpson index tends to be smaller when the community is more diverse.

In onomastic context, p_{ki} denotes the relative frequency of surname k in region ($\approx$ community diversity context) i, i.e., Simpson's diversity index is equivalent to the concept of isonymy..

Value

A dataframe containing the following components:

location	represents the grouping element, for example the communities / regions.
simpson	the value of the Simpson's diversity index.
divSimpson	the value of the inverse Simpson's diversity index.

Author(s)

Maria Jose Ginzo Villamayor

References

Simpson (1949) Measurement of diversity. *Nature*, **163**.

See Also

fMargalef, fMenhinick, fPielou, fShannon, fSheldon, fSimpsonInf, fGeneralisedMean, fGeometricMean, fHeip.

Examples

```
data(surnamesgal14)
result = fSimpson (x= surnamesgal14, k="number",
n="population", location  = "muni" )
result

data(namesmengal16)
result = fSimpson (x= namesmengal16, k="number",
n="population", location  = "muni" )
result

data(nameswomengal16)
result = fSimpson (x= nameswomengal16, k="number",
n="population", location  = "muni" )
result
```

---

fSimpsonInf   *Calculate the Simpson's diversity index and the inverse*

Description

This function obtains the Simpson's diversity index and the inverse introduced by Edward Hugh Simpson. It is a method for quantifying species biodiversity that can be adapted to the context of onomastics.

Usage

```
fSimpsonInf(x, k, n, location)
```

Arguments

x	dataframe of the data values for each species.
k	name of a variable which represents absolute frequency for each species.
n	name of a variable which represents total number of individuals.
location	represents the grouping element.

Details

For a community i, the Simpson (when N_i is not finite, data are assumed to come from a sample of size N_i) diversity index is defined by $D'_{S_i} = \sum\limits_{k \in S_i} \frac{n_{ki}(n_{ki}-1)}{n_i(n_i-1)}$, where n_{ki} represents the number of individuals of species k in a sample (in the total is N_{ki}) and S_i represents all species at the community, species richness.

In onomastic context, n_{ki} ($\approx N_{ki}$) denotes the absolute frequency of surname k in region i and S_i are all surnames in region ($\approx$ community diversity context) i.

Value

A dataframe containing the following components:

location	represents the grouping element, for example the communities / regions.
simpson	the value of the Simpson's Diversity Index.

Author(s)

Maria Jose Ginzo Villamayor

References

Simpson (1949) Measurement of diversity. *Nature*, **163**.

See Also

[fMargalef](), [fMenhinick](), [fPielou](), [fShannon](), [fSheldon](), [fSimpson](), [fGeneralisedMean](), [fGeometricMean](), [fHeip]().

Examples

```
data(surnamesgal14)
result = fSimpsonInf (x= surnamesgal14, k="number",
n="population", location  = "muni" )
result

data(namesmengal16)
result = fSimpsonInf (x= namesmengal16, k="number",
n="population", location  = "muni" )
result

data(nameswomengal16)
result = fSimpsonInf (x= nameswomengal16, k="number",
n="population", location  = "muni" )
result
```

---

namesmengal16 *namesmengal16 data*

Description

This dataset corresponds to 25 most frequent men's names by municipality in Galicia in 2016.

Usage

```
data(namesmengal16)
```

Format

namesmengal16 is a data frame with men's names from Galicia in 2016

Source

The data corresponds to 25 most frequent men's names by municipality in Galicia in 2016. The dataset contains 6 columns, prov: the province, muni: the municipality, namuni: the name of the municipality, name: the name, number: the number of people with that name and population: the total population considered by municipality.

These data have been extracted from the website of the Galician Institute of Statistics (IGE). The IGE offers information on the surnames and names of the population whose residence is in the Autonomous Community of Galicia. The base information for the elaboration data is the file of the Municipal Register of inhabitants of 2014 that the National Institute of Statistics (INE) provides to the IGE.

References

Galician Institute of Statistics (IGE), https://www.ige.eu/

Examples

```
data(namesmengal16)
```

---

nameswomengal16 *nameswomengal16 data*

Description

This dataset corresponds to 25 most frequent women's names by municipality in Galicia in 2016.

Usage

```
data(nameswomengal16)
```

Format

`nameswomengal16` is a data frame with women's names from Galicia in 2016.

Source

The data corresponds to 25 most frequent women's names by municipality in Galicia in 2016. The dataset contains 6 columns, `prov`: the province, `muni`: the municipality, `namuni`: the name of the municipality, `name`: the name, `number`: the number of people with that name and `population`: the total population considered by municipality.

These data have been extracted from the website of the Galician Institute of Statistics (IGE). The IGE offers information on the surnames and names of the population whose residence is in the Autonomous Community of Galicia. The base information for the elaboration data is the file of the Municipal Register of inhabitants of 2014 that the National Institute of Statistics (INE) provides to the IGE.

References

Galician Institute of Statistics (IGE), https://www.ige.eu/

Examples

`data(nameswomengal16)`

---

surnamesgal14 *surnamesgal14 data*

Description

This dataset corresponds to 25 most frequent surnames by municipality in Galicia in 2014.

Usage

`data(surnamesgal14)`

Format

`surnamesgal14` is a data frame with surnames from Galicia in 2014.

Source

The data corresponds to 25 most frequent surnames by municipality in Galicia in 2014. The dataset contains 8 columns, `prov`: the province, `muni`: the municipality, `namuni`: the name of the municipality, `surname`: the surname, `number`: the number of people with that surname, `population`: the total population considered by municipality, `ni`: the number of surnames considered and p_{ki} which is the frequency of surname k in municipality i.

These data have been extracted from the website of the Galician Institute of Statistics (IGE). The IGE offers information on the surnames and names of the population whose residence is in the

Autonomous Community of Galicia. The base information for the elaboration data is the file of the Municipal Register of inhabitants of 2014 that the National Institute of Statistics (INE) provides to the IGE.

References

Galician Institute of Statistics (IGE), https://www.ige.eu/

Examples

```
data(surnamesgal14)
```

Index