Package 'HMP'

July 2, 2014

Type Package

Title Hypothesis Testing and Power Calculations for Comparing Metagenomic Samples from HMP

Version 1.3.1

Date 2013-05-08

Author Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Maintainer Berkley Shands <bshands@dom.wustl.edu>

Depends R (>= 2.13.0), MCMCpack, dirmult

Description This package provides several functions to perform formal
hypothesis testing, and power and sample size calculations for
human microbiome experiments. Dirichlet-Multinomial
distribution is proposed to model species abundance and ranked abundance data.

License GPL-2

LazyData yes

NeedsCompilation no

Repository CRAN

Date/Publication 2013-05-09 00:21:55

R topics documented:

HMP-package 2
Barchart.data 3
C.alpha.multinomial 4
Data.filter 5
Dirichlet.multinomial 6
DM.MoM 7
MC.Xdc.statistics 8
MC.Xmc.statistics 10

MC.Xmcupo.statistics . . . 11
MC.Xoc.statistics . . . 13
MC.Xsc.statistics . . . 14
MC.ZT.statistics . . . 16
Multinomial . . . 17
pioest . . . 18
saliva . . . 19
throat . . . 19
tongue . . . 20
tonsils . . . 20
Xdc.sevsample . . . 21
Xmc.sevsample . . . 22
Xmcupo.effectsize . . . 23
Xmcupo.sevsample . . . 24
Xoc.sevsample . . . 25
Xsc.onesample . . . 27

Index **28**

`HMP-package` *Hypothesis Testing and Power Calculations for Comparing Metagenomic Samples from HMP*

Description

Generating data matrices following Multinomial and Dirichlet-Multinomial distributions, Computing the following test-statistics and their corresponding p-values, and Computing the power and size of the tests described above using Monte-Carlo simulations.

Details

Hypothesis Test	**Test Statistics Function**	**Power Calculation Function**
2+ Sample Means w/ Reference Vector	`Xmc.sevsample`	`MC.Xmc.statistics`
1 Sample Mean w/ Reference Vector	`Xsc.onesample`	`MC.Xsc.statistics`
2+ Sample Means w/o Reference Vector	`Xmcupo.sevsample`	`MC.Xmcupo.statistics`
2+ Sample Overdispersions	`Xoc.sevsample`	`MC.Xoc.statistics`
2+ Sample DM-Distribution	`Xdc.sevsample`	`MC.Xdc.statistics`
Multinomial vs DM	`C.alpha.multinomial`	`MC.ZT.statistics`

In addition to hypothesis testing and power calculations you can:

1. Perform basic data management to exclude samples with fewer than pre-specified number of reads, collapse rare taxa and order the taxa by frequency. This is useful to exclude failed samples (i.e. samples with very few reads) - `Data.filter`
2. Plot your data - `Barchart.data`

3. Generate random sample of Dirichlet Multinomial data with pre-specified parameters - `Dirichlet.multinomial`

Note: Thought the description of the functions refer its application to ranked abundance distributions (RAD) data, every function is also applicable to model species abundance data. See references for a discussion and application to both type of ecological data.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

References

La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, et al. (2012) Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. PLoS ONE 7(12): e52078. doi:10.1371/journal.pone.0052078

`Barchart.data` *A Graphical Representation of Taxa Proportions*

Description

Creates a bar plot of taxonomic proportions.

Usage

```
Barchart.data(data, title = "Taxa Proportions")
```

Arguments

`data`	A matrix of taxonomic counts(columns) for each sample(rows).
`title`	A string to be used as the plots title. The default is "Taxa Proportions".

Value

A bar plot of taxonomic proportions for all samples at a given taxonomic level.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Examples

```
data(saliva)

Barchart.data(saliva)
```

`C.alpha.multinomial` *$C(\alpha)$ - Optimal Test for Assessing Multinomial Goodness of Fit Versus Dirichlet-Multinomial Alternative*

Description

A function to compute the $C(\alpha)$-optimal test statistics of Kim and Margolin (1992) for evaluating the Goodness-of-Fit of a Multinomial distribution (null hypothesis) versus a Dirichlet-Multinomial distribution (alternative hypothesis).

Usage

```
C.alpha.multinomial(data)
```

Arguments

`data`	A matrix of taxonomic counts(columns) for each sample(rows).

Details

In order to test if a set of ranked-abundance distribution(RAD) from microbiome samples can be modeled better using a multinomial or Dirichlet-Multinomial distribution, we test the hypothesis $\mathrm{H} : \theta = 0$ versus $\mathrm{H} : \theta \neq 0$, where the null hypothesis implies a multinomial distribution and the alternative hypothesis implies a DM distribution. Kim and Margolin (Kim and Margolin, 1992) proposed a $C(\alpha)$-optimal test- statistics given by,

$$T = \sum_{j=1}^{K} \sum_{i=1}^{P} \frac{1}{\sum_{i=1}^{P} x_{ij}} \left(x_{ij} - \frac{N_i \sum_{i=1}^{P} x_{ij}}{N_{\mathrm{g}}} \right)^2$$

Where K is the number of taxa, P is the number of samples, x_{ij} is the taxon j, $j = 1, \ldots, K$ from sample i, $i = 1, \ldots, P$, N_i is the number of reads in sample i, and N_{g} is the total number of reads across samples.

As the number of reads increases, the distribution of the T statistic converges to a Chi-square with degrees of freedom equal to $(P-1)(K-1)$, when the number of sequence reads is the same in all samples. When the number of reads is not the same in all samples, the distribution becomes a weighted Chi-square with a modified degree of freedom (see (Kim and Margolin, 1992) for more details).

Note: Each taxa in `data` should be present in at least 1 sample, a column with all 0's may result in errors and/or invalid results.

Value

A list containing the $C(\alpha)$-optimal test statistic and p-value.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

References

Kim, B. S., and Margolin, B. H. (1992). Testing Goodness of Fit of a Multinomial Model Against Overdispersed Alternatives. Biometrics 48, 711-719.

Examples

```
data(saliva)

### Change the number of display digits so our output looks better
defaultd <- .Options$digits
options(digits=5)

calpha_check <- C.alpha.multinomial(saliva)
calpha_check

options(digits=defaultd) ### Set the number of digits back for the user
```

`Data.filter` *A Data Filter*

Description

This function creates a new dataset from an existing one that collapses less-abundant taxa into one category as specified by the user and excludes samples with a total number of reads fewer than the user-specified value.

Usage

```
Data.filter(data, order.type, reads.crit, K)
```

Arguments

`data`	A matrix of taxonomic counts(columns) for each sample(rows).
`order.type`	If `"sample"`: Rank taxa based on its taxonomic frequency. If `"data"`: Rank taxa based on cumulative taxonomic counts across all samples.
`reads.crit`	Samples with a total number of reads less than read.crit's value will be deleted.
`K`	The K most abundant taxa.

Value

A data frame with K+1 ranked columns and the number of rows equal to number of samples with a total number of reads greater than the critical value. The (K+1)th taxon contains the sum of the remaining less abundant taxa equal to (number of columns-K).

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Examples

```
data(saliva)

### Excludes all samples with fewer than 1000 reads and collapses
### taxas with 11th or smaller abundance into one category
filter_data <- Data.filter(saliva, "sample", 1000, 10)
```

`Dirichlet.multinomial` *Generation of Dirichlet-Multinomial Random Samples*

Description

Random generation of Dirichlet-Multinomial samples.

Usage

```
Dirichlet.multinomial(Nrs, shape)
```

Arguments

`Nrs`	A vector specifying the number of reads or sequence depth for each sample.
`shape`	A vector of Dirichlet parameters for each taxa.

Details

The Dirichlet-Multinomial distribution is given by (Mosimann, J. E. (1962); Tvedebrink, T. (2010)),

$$\mathbf{P}\left(\mathbf{X}_i = x_i; \{\pi_j\}, \theta\right) = \frac{N_i!}{x_{i1}!, \ldots, x_{iK}!} \frac{\prod_{j=1}^{K} \prod_{r=1}^{x_{ij}} \{\pi_j (1-\theta) + (r-1)\theta\}}{\prod_{r=1}^{N_i} (1-\theta) + (r-1)\theta}$$

where $\mathbf{x}_i = [x_{i1}, \ldots, x_{iK}]$ is the random vector formed by K taxa (features) counts (RAD vector), $N_i = \sum_{j=1}^{K} x_{ij}$ is the total number of reads (sequence depth), $\{\pi_j\}$ are the mean of taxa-proportions (RAD-probability mean), and θ is the overdispersion parameter.

Note: Though the test statistic supports an unequal number of reads across samples, the performance has not yet been fully tested.

Value

A data matrix of taxa counts where the rows are samples and columns are the taxa.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

References

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika 49, 65-82.
Tvedebrink, T. (2010). Overdispersion in allelic counts and theta-correction in forensic genetics. Theor Popul Biol 78, 200-210.

Examples

```
data(saliva)

### Generate a random vector of number of reads per sample
Nrs <- rep(15000, 20)

### Get a list of dirichlet-multinomial parameters for the data
fit.saliva <- dirmult(saliva)

dirmult_data <- Dirichlet.multinomial(Nrs, fit.saliva$gamma)
dirmult_data
```

`DM.MoM` *Method-of-Moments (MoM) Estimators of the Dirichlet-Multinomial Parameters*

Description

Method-of-Moments (MoM) estimators of the Dirichlet-multinomial parameters: taxa proportions and overdispersion.

Usage

```
DM.MoM(data)
```

Arguments

`data`	A matrix of taxonomic counts(columns) for each sample(rows).

Details

Given a set of taxa-count vectors $\{\mathbf{x}_i, \ldots, \mathbf{x}_P\}$, the methods of moments (MoM) estimator of the set of parameters θ and $\{\pi_j\}_{j=1}^K$ is given as follows (Mosimann, 1962; Tvedebrink, 2010):

$$\hat{\pi}_j = \frac{\sum_{i=1}^P x_{ij}}{\sum_{i=1}^P N_i},$$

and

$$\hat{\theta} = \sum_{j=1}^K \frac{S_j - G_j}{\sum_{j=1}^K \left(S_j + (N_c - 1)\, G_j\right)},$$

where $N_c = (P-1)^{-1}\left(\sum_{i=1}^{P} N_i - \left(\sum_{i=1}^{P} N_i\right)^{-1} \sum_{i=1}^{P} N_i^2\right)$, and $S_j = (P-1)^{-1} \sum_{i=1}^{P} N_i \left(\hat{\pi}_{ij} - \hat{\pi}_j\right)^2$, and $G_j = \left(\sum_{i=1}^{P} (N_i - 1)\right)^{-1} \sum_{i=1}^{P} N_i \hat{\pi}_{ij} (1 - \hat{\pi}_{ij})$ with $\hat{\pi}_{ij} = \frac{x_{ij}}{N_i}$.

Value

A list providing the MoM estimator for overdispersion, the MoM estimator of the RAD-probability mean vector, and the corresponding loglikelihood value for the given data set and estimated parameters.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

References

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika 49, 65-82.
Tvedebrink, T. (2010). Overdispersion in allelic counts and theta-correction in forensic genetics. Theor Popul Biol 78, 200-210.

Examples

```
data(throat)

dm.mom_check <- DM.MoM(throat)
dm.mom_check
```

`MC.Xdc.statistics` *Size and Power for the Several-Sample DM Parameter Test Comparison*

Description

This Monte-Carlo simulation procedure provides the power and size of the several sample Dirichlet-Multinomial parameter test comparison, using the likelihood-ratio-test statistics.

Usage

```
MC.Xdc.statistics(Nrs, MC, alphap, n.groups, type = "ha", siglev = 0.05, est = "mom")
```

Arguments

`Nrs`	A list specifying the number of reads/sequence depth for each sample in a group with one group per list entry.
`MC`	Number of Monte-Carlo experiments. In practice this should be at least 1,000.

alphap	If "hnull": A matrix where rows are vectors of alpha parameters for each group. If "ha": A matrix consisting of vectors of alpha parameters for each taxa in each group.
n.groups	The number of groups to compare.
type	If "hnull": Computes the size of the test. If "ha": Computes the power of the test. (default)
siglev	Significance level for size of the test / power calculation. The default is 0.05.
est	The type of parameter estimator to be used with the Likelihood-ratio-test statistics, 'mle' or 'mom'. Default value is 'mom'. (See Note 2 in details)

Details

1. Note 1: Though the test statistic supports an unequal number of reads across samples, the performance has not yet been fully tested.
2. Note 2: 'mle' will take significantly longer time and may not be optimal for small sample sizes; 'mom' will provide a more conservative result in such a case.
3. Note 3: All components of alphap should be non-zero or it may result in errors and/or invalid results.

Value

Size of the test statistics (under "hnull") or power (under "ha") of the test.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Examples

```
data(saliva)
data(throat)
data(tonsils)

### Get a list of dirichlet-multinomial parameters for the data
fit.saliva <- DM.MoM(saliva)
fit.throat <- DM.MoM(throat)
fit.tonsils <- DM.MoM(tonsils)

### Set up the number of Monte-Carlo experiments
### We set MC=1 due to CRAN restrictions, Please set MC to be at least 1,000
MC <- 1

### Generate a random vector of number of reads per sample
Nrs1 <- rep(12000, 9)
Nrs2 <- rep(12000, 11)
Nrs3 <- rep(12000, 12)
group.Nrs <- list(Nrs1, Nrs2, Nrs3)

### Computing size of the test statistics (Type I error)
```

```
mc.xdc_check1 <- MC.Xdc.statistics(group.Nrs, MC, fit.saliva$gamma, 3, "hnull", .05, "mom")
mc.xdc_check1

### Computing Power of the test statistics (1 - Type II error)
group.alphap <- rbind(fit.saliva$gamma, fit.throat$gamma, fit.tonsils$gamma)
mc.xdc_check2 <- MC.Xdc.statistics(group.Nrs, MC, group.alphap, 3, "ha", 0.05, "mom")
mc.xdc_check2
```

`MC.Xmc.statistics` *Size and Power of Several Sample RAD-Probability Mean Test Comparison*

Description

This Monte-Carlo simulation procedure provides the power and size of the several sample RAD-probability mean test comparison with known reference vector of proportions, using the Generalized Wald-type statistics.

Usage

```
MC.Xmc.statistics(Nrs, MC, pi0, group.pi, group.theta, type = "ha", siglev = 0.05)
```

Arguments

`Nrs`	A list specifying the number of reads/sequence depth for each sample in a group with one group per list entry.
`MC`	Number of Monte-Carlo experiments. In practice this should be at least 1,000.
`pi0`	The RAD-probability mean vector.
`group.pi`	If `"hnull"`: This argument is ignored. If `"ha"`: A matrix where each row is a vector pi values for each group.
`group.theta`	A vector of overdispersion values for each group.
`type`	If `"hnull"`: Computes the size of the test. If `"ha"`: Computes the power of the test. (default)
`siglev`	Significance level for size of the test / power calculation. The default is 0.05.

Details

Note: Though the test statistic supports an unequal number of reads across samples, the performance has not yet been fully tested.

Value

Size of the test statistics (under `"hnull"`) or power (under `"ha"`) of the test.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Examples

```
data(saliva)
data(throat)
data(tonsils)

### Get a list of dirichlet-multinomial parameters for the data
fit.saliva <- DM.MoM(saliva)
fit.throat <- DM.MoM(throat)
fit.tonsils <- DM.MoM(tonsils)

### Set up the number of Monte-Carlo experiments
### We set MC=1 due to CRAN restrictions, Please set MC to be at least 1,000
MC <- 1

### Generate a random vector of number of reads per sample
Nrs1 <- rep(12000, 18)
Nrs2 <- rep(12000, 10)
Nrs3 <- rep(12000, 15)
group.Nrs <- list(Nrs1, Nrs2, Nrs3)

### Computing size of the test statistics (Type I error)
group.theta <- c(0.01, 0.05)
mc.xmc_check1 <- MC.Xmc.statistics(group.Nrs, MC, fit.saliva$pi, , group.theta, "hnull")
mc.xmc_check1

### Computing Power of the test statistics (1 - Type II error)
pi_2grp <- rbind(fit.throat$pi, fit.tonsils$pi)
mc.xmc_check2 <- MC.Xmc.statistics(group.Nrs, MC, fit.saliva$pi, pi_2grp, group.theta, "ha")
mc.xmc_check2
```

`MC.Xmcupo.statistics` *Size and Power of Several Sample RAD-Probability Mean Test Comparisons: Unknown Vector of Proportion*

Description

This Monte-Carlo simulation procedure provides the power and size of the several sample RAD-probability mean test comparisons without reference vector of proportions, using the Generalized Wald-type statistics.

Usage

```
MC.Xmcupo.statistics(Nrs, MC, pi0, group.pi, group.theta, type = "ha", siglev = 0.05)
```

Arguments

`Nrs`	A list specifying the number of reads/sequence depth for each sample in a group with one group per list entry.
`MC`	Number of Monte-Carlo experiments. In practice this should be at least 1,000.

pi0	The RAD-probability mean vector.
group.pi	If "hnull": This argument is ignored. If "ha": A matrix where each row is a vector pi values for each group.
group.theta	A vector of overdispersion values for each group.
type	If "hnull": Computes the size of the test. If "ha": Computes the power of the test. (default)
siglev	Significance level for size of the test / power calculation. The default is 0.05.

Details

Note: Though the test statistic supports an unequal number of reads across samples, the performance has not yet been fully tested.

Value

Size of the test statistics (under "hnull") or power (under "ha") of the test.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Examples

```
data(saliva)
data(throat)
data(tonsils)

### Get a list of dirichlet-multinomial parameters for the data
fit.saliva <- DM.MoM(saliva)
fit.throat <- DM.MoM(throat)
fit.tonsils <- DM.MoM(tonsils)

### Set up the number of Monte-Carlo experiments
### We set MC=1 due to CRAN restrictions, Please set MC to be at least 1,000
MC <- 1

### Generate a random vector of number of reads per sample
Nrs1 <- rep(12000, 10)
Nrs2 <- rep(12000, 19)
Nrs3 <- rep(12000, 19)
group.Nrs <- list(Nrs1, Nrs2, Nrs3)

### Computing size of the test statistics (Type I error)
group.theta <- c(fit.throat$theta, fit.tonsils$theta)
mc.xmcupo_check1 <- MC.Xmcupo.statistics(group.Nrs, MC, fit.saliva$pi, ,
group.theta, "hnull", 0.01)
mc.xmcupo_check1

### Computing Power of the test statistics (1 - Type II error)
pi_2grp <- rbind(fit.throat$pi, fit.tonsils$pi)
```

```
mc.xmcupo_check2 <- MC.Xmcupo.statistics(group.Nrs, MC, fit.saliva$pi,
pi_2grp, group.theta, "ha", 0.01)
mc.xmcupo_check2
```

`MC.Xoc.statistics` *Size and Power of Several Sample-Overdispersion Test Comparisons*

Description

This Monte-Carlo simulation procedure provides the power and size of the several sample-overdispersion test comparison, using the likelihood-ratio-test statistics.

Usage

```
MC.Xoc.statistics(Nrs, MC, group.alphap, n.groups, type = "ha", siglev = 0.05)
```

Arguments

`Nrs`	A list specifying the number of reads/sequence depth for each sample in a group with one group per list entry.
`MC`	Number of Monte-Carlo experiments. In practice this should be at least 1,000.
`group.alphap`	If `"hnull"`: A vector of alpha parameters for each taxa. If `"ha"`: A list consisting of vectors of alpha parameters for each taxa.
`n.groups`	The number of groups to compare.
`type`	If `"hnull"`: Computes the size of the test. If `"ha"`: Computes the power of the test. (default)
`siglev`	Significance level for size of the test / power calculation. The default is 0.05.

Details

1. Note 1: Though the test statistic supports an unequal number of reads across samples, the performance has not yet been fully tested.
2. Note 2: All components of `group.alphap` should be non-zero or it may result in errors and/or invalid results.

Value

Size of the test statistics (under `"hnull"`) or power (under `"ha"`) of the test.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Examples

```
data(saliva)
data(throat)
data(tonsils)

### Get a list of dirichlet-multinomial parameters for the data
fit.saliva <- DM.MoM(saliva)
fit.throat <- DM.MoM(throat)
fit.tonsils <- DM.MoM(tonsils)

### Set up the number of Monte-Carlo experiments
### We set MC=1 due to CRAN restrictions, Please set MC to be at least 1,000
MC <- 1

### Generate a random vector of number of reads per sample
Nrs1 <- rep(12000, 10)
Nrs2 <- rep(12000, 11)
Nrs3 <- rep(12000, 9)
group.Nrs <- list(Nrs1, Nrs2, Nrs3)

### Computing size of the test statistics (Type I error)
mc.xoc_check1 <- MC.Xoc.statistics(group.Nrs, MC, fit.tonsils$gamma, 3, "hnull")
mc.xoc_check1

## Not run:
### Computing Power of the test statistics (1 - Type II error)
group.alphap <- rbind(fit.saliva$gamma, fit.throat$gamma, fit.tonsils$gamma)
mc.xoc_check2 <- MC.Xoc.statistics(group.Nrs, MC, group.alphap, 3, "ha")
mc.xoc_check2

## End(Not run)
```

`MC.Xsc.statistics` *Size and Power for the One Sample RAD Probability-Mean Test Comparison*

Description

This Monte-Carlo simulation procedure provides the power and size of the one sample RAD probability-mean test, using the Generalized Wald-type statistic.

Usage

```
MC.Xsc.statistics(Nrs, MC, fit, pi0, type = "ha", siglev = 0.05)
```

Arguments

`Nrs`	A vector specifying the number of reads/sequence depth for each sample.
`MC`	Number of Monte-Carlo experiments. In practice this should be at least 1,000.

fit	A list (in the format of the output of dirmult function) containing the data parameters for evaluating either the size or power of the test.
pi0	The RAD-probability mean vector. If the type is set to "hnull" then pi0 is set by the sample in fit.
type	If "hnull": Computes the size of the test. If "ha": Computes the power of the test. (default)
siglev	Significance level for size of the test / power calculation. The default is 0.05.

Details

Note: Though the test statistic supports an unequal number of reads across samples, the performance has not yet been fully tested.

Value

Size of the test statistics (under "hnull") or power (under "ha") of the test.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Examples

```
data(saliva)
data(throat)
data(tonsils)

### Get a list of dirichlet-multinomial parameters for the data
fit.saliva <- DM.MoM(saliva)
fit.throat <- DM.MoM(throat)
fit.tonsils <- DM.MoM(tonsils)

### Set up the number of Monte-Carlo experiments
### We set MC=1 due to CRAN restrictions, Please set MC to be at least 1,000
MC <- 1

### Generate a random vector of number of reads per sample
Nrs <- rep(15000, 25)

### Computing size of the test statistics (Type I error)
mc.xsc_check1 <- MC.Xsc.statistics(Nrs, MC, fit.tonsils, fit.saliva$pi, "hnull", 0.05)
mc.xsc_check1

### Computing Power of the test statistics (1 - Type II error)
mc.xsc_check2 <- MC.Xsc.statistics(Nrs, MC, fit.throat,  fit.tonsils$pi, "ha", 0.01)
mc.xsc_check2
```

`MC.ZT.statistics` *Size and Power of Goodness of Fit Test: Multinomial vs. Dirichlet-Multinomial*

Description

This Monte-Carlo simulation procedure provides the power and size of the Multinomial vs. Dirichlet-Multinomial goodness of fit test, using the C(α)-optimal test statistics of Kim and Margolin (1992) (t statistics) and the C(α)-optimal test statistics of (Paul et al., 1989).

Usage

```
MC.ZT.statistics(Nrs, MC, fit, type = "ha", siglev = 0.05)
```

Arguments

`Nrs`	A vector specifying the number of reads/sequence depth for each sample.
`MC`	Number of Monte-Carlo experiments. In practice this should be at least 1,000.
`fit`	A list (in the format of the output of dirmult function) containing the data parameters for evaluating either the size or power of the test.
`type`	If `"hnull"`: Computes the size of the test. If `"ha"`: Computes the power of the test. (default)
`siglev`	Significance level for size of the test / power calculation. The default is 0.05.

Details

Note: Though the test statistic supports an unequal number of reads across samples, the performance has not yet been fully tested.

Value

A vector containing both the size of the test statistics (under `"hnull"`) or power (under `"ha"`) of the test for both the z and t statistics.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Examples

```
data(saliva)

### Get a list of dirichlet-multinomial parameters for the data
fit.saliva <- DM.MoM(saliva)

### Set up the number of Monte-Carlo experiments
### We set MC=1 due to CRAN restrictions, Please set MC to be at least 1,000
```

```
MC <- 1

### Generate a random vector of number of reads per sample
Nrs <- rep(15000, 25)

### Computing size of the test statistics (Type I error)
mc.zt_check1 <- MC.ZT.statistics(Nrs, MC, fit.saliva, "hnull")
mc.zt_check1

### Computing Power of the test statistics (1 - Type II error)
mc.zt_check2 <- MC.ZT.statistics(Nrs, MC, fit.saliva, "ha")
mc.zt_check2
```

`Multinomial` *Generation of Multinomial Random Samples*

Description

It generates a data matrix with random samples from a multinomial distribution where the rows are the samples and the columns are the taxa.

Usage

```
Multinomial(Nrs, probs)
```

Arguments

`Nrs`	A vector specifying the number of reads or sequence depth for each sample.
`probs`	A vector specifying taxa probabilities.

Details

Note: Though the test statistic supports an unequal number of reads across samples, the performance has not yet been fully tested.

Value

A data matrix of taxa counts where the rows are the samples and the columns are the taxa.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Examples

```
### Generate a random vector of number of reads per sample
Nrs <- rep(15000, 25)

mypi <- c(0.4, 0.3, 0.2, .05, 0.04, .01)

mult_data <- Multinomial(Nrs, mypi)
mult_data
```

---

pioest *Weighted Average of Taxa Frequency for Several Groups*

Description

This function computes a weighted-average of taxa frequency estimated from several groups.

Usage

```
pioest(group.data)
```

Arguments

group.data	A list where each element is a matrix of taxonomic counts(columns) for each sample(rows).

Details

Note: The matrices in group.data must contain the same taxa, in the same order.

Value

A vector containing the weighted-average of taxa frequency.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Examples

```
data(saliva)
data(throat)
group.data <- list(saliva, throat)

pio <- pioest(group.data)
pio
```

saliva *Saliva Data Set*

Description

The saliva data set formed by the Ranked-abundance distribution vectors of 24 subjects. The RAD vectors contains 21 elements formed by the 20 most abundant taxa at the genus level and additional taxa containing the sum of the remaining less abundant taxa per sample. Note that the incorporation of the additional taxon (taxon 21) in the analysis allows for estimating the RAD proportional-mean of taxa with respect to all the taxa within the sample.

Usage

```
data(saliva)
```

Format

The format is a matrix of 24 rows by 21 columns, with the each row being a separate subject and each column being a different taxa.

Examples

```
data(saliva)
```

throat *Throat Data Set*

Description

The throat data set formed by the Ranked-abundance distribution vectors of 24 subjects. The RAD vectors contains 21 elements formed by the 20 most abundant taxa at the genus level and additional taxa containing the sum of the remaining less abundant taxa per sample. Note that the incorporation of the additional taxon (taxon 21) in the analysis allows for estimating the RAD proportional-mean of taxa with respect to all the taxa within the sample.

Usage

```
data(throat)
```

Format

The format is a matrix of 24 rows by 21 columns, with the each row being a separate subject and each column being a different taxa.

Examples

```
data(throat)
```

tongue *Tongue Data Set*

Description

The tongue data set formed by the Ranked-abundance distribution vectors of 24 subjects. The RAD vectors contains 21 elements formed by the 20 most abundant taxa at the genus level and additional taxa containing the sum of the remaining less abundant taxa per sample. Note that the incorporation of the additional taxon (taxon 21) in the analysis allows for estimating the RAD proportional-mean of taxa with respect to all the taxa within the sample.

Usage

```
data(tongue)
```

Format

The format is a matrix of 24 rows by 21 columns, with the each row being a separate subject and each column being a different taxa.

Examples

```
data(tongue)
```

tonsils *Palatine Tonsil Data Set*

Description

The palatine tonsil data set formed by the Ranked-abundance distribution vectors of 24 subjects. The RAD vectors contains 21 elements formed by the 20 most abundant taxa at the genus level and additional taxa containing the sum of the remaining less abundant taxa per sample. Note that the incorporation of the additional taxon (taxon 21) in the analysis allows for estimating the RAD proportional-mean of taxa with respect to all the taxa within the sample.

Usage

```
data(tonsils)
```

Format

The format is a matrix of 24 rows by 21 columns, with the each row being a separate subject and each column being a different taxa.

Examples

```
data(tonsils)
```

`Xdc.sevsample` *Likelihood-Ratio-Test Statistics: Several Sample Dirichlet-Multinomial Test Comparison*

Description

This routine provides the value of the Likelihood-Ratio-Test Statistics and the corresponding p-value for evaluating the several sample Dirichlet-Multinomial parameter test comparison.

Usage

```
Xdc.sevsample(group.data, epsilon = 10^(-4), est = "mom")
```

Arguments

`group.data`	A list where each element is a matrix of taxonomic counts(columns) for each sample(rows). (See Notes 1 and 2 in details)
`epsilon`	Convergence tolerance. To terminate, the difference between two succeeding log-likelihoods must be smaller than epsilon. Default value is 10^(-4).
`est`	The type of parameter estimator to be used with the Likelihood-ratio-test statistics, 'mle' or 'mom'. Default value is 'mom'. (See Note 3 in details)

Details

To assess whether the Dirichlet parameter vector, $\alpha_{\mathrm{m}} = \pi_{\mathrm{m}} \frac{1-\theta_{\mathrm{m}}}{\theta_{\mathrm{m}}}$(a function of the RAD probability-mean vector and overdispersion), observed in J groups of microbiome samples are equal to each other, the following hypothesis $\mathrm{H_o} : \alpha_1 = \cdots = \alpha_{\mathrm{m}} = \cdots = \alpha_{\mathrm{J}} = \alpha_{\mathrm{o}}$ versus $\mathrm{H_a} : \alpha_{\mathrm{m}} \neq \alpha_{\mathrm{o}}, m = 1, \ldots, J$ can be tested. The null hypothesis implies that the HMP samples across groups have the same mean and overdispersion, indicating that the RAD models are identical. In particular, the likelihood-ratio test statistic is used, which is given by,

$$x_{\mathrm{dc}} = -2 \log \left\{ \frac{L\left(\alpha_{\mathrm{o}}; \mathbf{X}_1, \ldots, \mathbf{X}_{\mathrm{J}}\right)}{L\left(\alpha_1, \ldots, \alpha_{\mathrm{J}}; \mathbf{X}_1, \ldots, \mathbf{X}_{\mathrm{J}}\right)} \right\}.$$

The asymptotic null distribution of x_{dc} follows a Chi-square with degrees of freedom equal to (J-1)*K, where K is the number of taxa (Wilks, 1938).

1. Note 1: The matrices in `group.data` must contain the same taxa, in the same order.
2. Note 2: Each taxa should be present in at least 1 sample, a column with all 0's may result in errors and/or invalid results.
3. Note 3: 'mle' will take significantly longer time and may not be optimal for small sample sizes; 'mom' will provide more conservative results in such a case.

Value

A list containing the Xdc statistics and p-value.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

References

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. The Annals of Mathematical Statistics 9, 60-62.

Examples

```
data(saliva)
data(throat)

### Change the number of display digits so our output looks better
defaultd <- .Options$digits
options(digits=5)

xdc.sev_check <- Xdc.sevsample(list(saliva, throat))
xdc.sev_check

options(digits=defaultd) ### Set the number of digits back for the user
```

Xmc.sevsample — *Generalized Wald-type Statistics: Several Sample RAD Probability-Mean Test Comparison with a Known Common Vector*

Description

This function computes the Generalized Wald-type test statistic (Wilson and Koehler, 1984) and corresponding p-value to assess whether the sample RAD probability-means from multiple populations are the same or different. The statistics assumes that a common RAD probability-mean vector for comparison under the null hypothesis is known.

Usage

```
Xmc.sevsample(group.data, pi0)
```

Arguments

`group.data`	A list where each element is a matrix of taxonomic counts(columns) for each sample(rows).
`pi0`	The RAD-probability mean vector.

Details

Note: The matrices in `group.data` must contain the same taxa, in the same order.

Value

A list containing the Generalized Wald-type statistics and p-value.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

References

Wilson, J. R., and Koehler, K. J. (1984). Testing of equality of vectors of proportions for several cluster samples. Proceedings of Joint Statistical Association Meetings. Survey Research Methods.

Examples

```
data(saliva)
data(throat)
data(tonsils)

### Get a list of dirichlet-multinomial parameters for the data
fit.saliva <- dirmult(saliva)
mygroup <- list(throat[,1:15], tonsils[,1:15])

### Change the number of display digits so our output looks better
defaultd <- .Options$digits
options(digits=5)

xmc.sev_check <- Xmc.sevsample(mygroup, fit.saliva$pi)
xmc.sev_check

options(digits=defaultd) ### Set the number of digits back for the user
```

Xmcupo.effectsize *Effect Size for Xmcupo Statistic*

Description

This function computes the Cramer's Phi and Modified Cramer's Phi Criterion for the test statistic `Xmcupo.sevsample`.

Usage

```
Xmcupo.effectsize(group.data)
```

Arguments

`group.data`	A list where each element is a matrix of taxonomic counts(columns) for each sample(rows).

Details

Note: The matrices in `group.data` must contain the same taxa, in the same order.

Value

A vector containing the Chi-Squared statistic value, the Cramer's Phi Criterion, and the modified Cramer's Phi Criterion.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Examples

```
data(saliva)
data(throat)
group.data <- list(saliva, throat)

effect <- Xmcupo.effectsize(group.data)
effect
```

Xmcupo.sevsample *Generalized Wald-type Statistics: Several Sample RAD Probability-Mean Test Comparison with an Unknown Common Vector*

Description

This function computes the Generalized Wald-type test statistic (Wilson and Koehler, 1984) and corresponding p-value to assess whether the sample RAD probability-means from multiple populations are same or different. The statistics assumes that a common RAD probability-mean vector for comparison under the null hypothesis is unknown.

Usage

```
Xmcupo.sevsample(group.data, K)
```

Arguments

`group.data`	A list where each element is a matrix of taxonomic counts(columns) for each sample(rows).
`K`	The number of taxa.

Details

Note: The matrices in `group.data` must contain the same taxa, in the same order.

Value

A list containing the Generalized Wald-type statistics and p-value.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

References

Wilson, J. R., and Koehler, K. J. (1984). Testing of equality of vectors of proportions for several cluster samples. Proceedings of Joint Statistical Association Meetings. Survey Research Methods.

Examples

```
data(saliva)
data(tonsils)
data(throat)

mygroup <- list(throat[,1:15], tonsils[,1:15], saliva[,1:15])

### Change the number of display digits so our output looks better
defaultd <- .Options$digits
options(digits=5)

xmcupo.sev_check <- Xmcupo.sevsample(mygroup, 15)
xmcupo.sev_check

options(digits=defaultd) ### Set the number of digits back for the user
```

Xoc.sevsample *Likelihood-Ratio-Test Statistics: Several Sample Overdispersion Test Comparison*

Description

This routine provides the value of the likelihood-ratio-test statistic and the corresponding p-value to assess whether the overdispersion observed in multiple groups of microbiome samples are equal.

Usage

```
Xoc.sevsample(group.data, epsilon = 10^(-4))
```

Arguments

group.data	A list where each element is a matrix of taxonomic counts(columns) for each sample(rows). (See Notes 1 and 2 in details)
epsilon	Convergence tolerance. To terminate, the difference between two succeeding log-likelihoods must be smaller than epsilon. Default value is 10^(-4).

Details

To assess whether the over dispersion parameter vectors θ_{m} observed in J groups of microbiome samples are equal to each other, the following hypothesis $\mathrm{H_o} : \theta_1 = \cdots = \theta_{\mathrm{m}} = \cdots = \theta_{\mathrm{J}} = \theta_{\mathrm{o}}$ versus $\mathrm{H_a} : \theta_{\mathrm{m}} \neq \theta_{\mathrm{o}}, m = 1, \ldots, J$ can be tested. In particular, the likelihood-ratio test statistic is used (Tvedebrink, 2010), which is given by,

$$x_{\mathrm{oc}} = -2 \log \left\{ \frac{L\left(\theta_{\mathrm{o}}; \mathbf{X}_1, \ldots, \mathbf{X}_{\mathrm{J}}\right)}{L\left(\theta_1, \ldots, \theta_{\mathrm{J}}; \mathbf{X}_1, \ldots, \mathbf{X}_{\mathrm{J}}\right)} \right\}.$$

The asymptotic null distribution of x_{oc} follows a Chi-square with degrees of freedom equal to (J-1) (Wilks, 1938).

1. Note 1: The matrices in `group.data` must contain the same taxa, in the same order.
2. Note 2: Each taxa should be present in at least 1 sample, a column with all 0's may result in errors and/or invalid results.

Value

A list containing the Xoc statistics and p-value.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

References

Tvedebrink, T. (2010). Overdispersion in allelic counts and theta-correction in forensic genetics. Theor Popul Biol 78, 200-210.
Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. The Annals of Mathematical Statistics 9, 60-62.

Examples

```
data(saliva)
data(tonsils)

mygroup <- list(saliva[,1:10], tonsils[,1:10])

### Change the number of display digits so our output looks better
defaultd <- .Options$digits
options(digits=5)

xoc.sev_check <- Xoc.sevsample(mygroup)
xoc.sev_check

options(digits=defaultd) ### Set the number of digits back for the user
```

`Xsc.onesample` *Generalized Wald-Type Statistics: One Sample RAD Probability-Mean Test Comparison*

Description

This routine provides the value of the Generalized Wald-type statistic to assess whether the RAD probability-mean observed in one group of samples is equal to a known RAD probability-mean.

Usage

```
Xsc.onesample(data, pi0)
```

Arguments

`data`	A matrix of taxonomic counts(columns) for each sample(rows).
`pi0`	The RAD-probability mean vector.

Value

A list containing Generalized Wald-type statistics and p-value.

Author(s)

Patricio S. La Rosa, Elena Deych, Berkley Shands, William D. Shannon

Examples

```
data(saliva)
data(throat)

### Get a list of dirichlet-multinomial parameters for the data
fit.saliva <- dirmult(saliva)

### Change the number of display digits so our output looks better
defaultd <- .Options$digits
options(digits=5)

xsc.one_check <- Xsc.onesample(throat, fit.saliva$pi)
xsc.one_check

options(digits=defaultd) ### Set the number of digits back for the user
```

Index

∗Topic **datasets**
saliva, 19
throat, 19
tongue, 20
tonsils, 20
∗Topic **package**
HMP-package, 2

Barchart.data, 3

C.alpha.multinomial, 4

Data.filter, 5
Dirichlet.multinomial, 6
DM.MoM, 7

HMP *(*HMP-package*)*, 2
HMP-package, 2

MC.Xdc.statistics, 8
MC.Xmc.statistics, 10
MC.Xmcupo.statistics, 11
MC.Xoc.statistics, 13
MC.Xsc.statistics, 14
MC.ZT.statistics, 16
Multinomial, 17

pioest, 18

saliva, 19

throat, 19
tongue, 20
tonsils, 20

Xdc.sevsample, 21
Xmc.sevsample, 22
Xmcupo.effectsize, 23
Xmcupo.sevsample, 24
Xoc.sevsample, 25
Xsc.onesample, 27