Package 'AnthropMMD'

July 18, 2019

Type Package

Title An R Package for Smith's Mean Measure of Divergence (MMD)

Description Offers a graphical user interface for the calculation of the mean measure of divergence, with facilities for trait selection and graphical representations <doi:10.1002/ajpa.23336>.

Version 3.0.1

Depends R (>= 3.5.0)

Imports plotrix, scatterplot3d, shiny, smacof

Suggests cluster, covr, knitr, rmarkdown, testthat (>= 2.1.0)

License CeCILL-2 | file LICENSE

Encoding UTF-8

URL `https://gitlab.com/f.santos/anthropmmd/`

BugReports `https://gitlab.com/f.santos/anthropmmd/issues/new`

VignetteBuilder knitr

NeedsCompilation no

Author Frédéric Santos [aut, cre] (<https://orcid.org/0000-0003-1445-3871>)

Maintainer Frédéric Santos `<frederic.santos@u-bordeaux.fr>`

Repository CRAN

Date/Publication 2019-07-18 06:36:19 UTC

R topics documented:

`AnthropMMD-package` *An R package for Smith's Mean Measure of Divergence (MMD)*

Description

Offers a graphical user interface for the calculation of the mean measure of divergence, with facilities for trait selection and graphical representations.

Details

Package:	AnthropMMD
Type:	Package
Version:	3.0.1
Date:	2019-07-16
License:	CeCILL 2.1

Author(s)

Frédéric Santos, <frederic.santos@u-bordeaux.fr>

References

Harris, E. F. and Sjøvold, T. (2004) Calculation of Smith's mean measure of divergence for intergroup comparisons using nonmetric data. *Dental Anthropology*, **17**(3), 83–93.

Irish, J. (2010) The mean measure of divergence: Its utility in model-free and model-bound analyses relative to the Mahalanobis D2 distance for nonmetric traits. *American Journal of Human Biology*, **22**, 378–395. doi: 10.1002/ajhb.21010

Nikita, E. (2015) A critical review of the mean measure of divergence and Mahalanobis distances using artificial data and new approaches to the estimation of biodistances employing nonmetric traits. *American Journal of Physical Anthropology*, **157**, 284–294. doi: 10.1002/ajpa.22708

Santos, F. (2018) AnthropMMD: an R package with a graphical user interface for the mean measure of divergence. *American Journal of Physical Anthropology*, **165**(1), 200–205. doi: 10.1002/ajpa.23336

Examples

```
## Not run:  start_mmd()
```

absolute_freqs — *A toy example dataset for mean measures of divergence, in a table format*

Description

This artifical dataset includes 200 individuals described by 9 binary traits and splitted into 5 groups. To fit with commonly observed datasets in past sciences, a substantial amount of missing values have been added at random on this dataset.

Usage

```
data(absolute_freqs)
```

Format

A matrix with 10 rows and 9 columns:

`Trait1` summary statistics for this trait

`Trait2` summary statistics for this trait

`Trait3` summary statistics for this trait

`Trait4` summary statistics for this trait

`Trait5` summary statistics for this trait

`Trait6` summary statistics for this trait

`Trait7` summary statistics for this trait

`Trait8` summary statistics for this trait

`Trait9` summary statistics for this trait

binary_to_table — *Converts a data frame of binary (i.e., presence/absence) trait information into a table of sample sizes and frequencies.*

Description

This function allows to get a summary of sample sizes and frequencies for each trait in each group. It is also mandatory to apply this function before using the `mmd` function, since the latter only accepts table of frequencies, and cannot work with raw binary data.

Usage

```
binary_to_table(data, relative = FALSE)
```

Arguments

`data`	A binary (0/1 for presence/absence of traits) data frame with n rows (one per individual) and $p + 1$ columns (one for each of the p traits, plus one column provided as a group indicator).
`relative`	Boolean. Indicates if the last rows of the table must contain frequencies (i.e., number of individuals having a given trait) or relative frequencies (i.e., proportions).

Value

A matrix with $2 * K$ rows (K being the number of groups in the dataset) and p columns (one per trait). The first K rows are the sample sizes, the last K rows are trait frequencies.

Author(s)

Frédéric Santos, <frederic.santos@u-bordeaux.fr>

References

Santos, F. (2018) AnthropMMD: an R package with a graphical user interface for the mean measure of divergence. *American Journal of Physical Anthropology*, **165**(1), 200–205. doi: 10.1002/ajpa.23336

See Also

`start_mmd`

Examples

```
# Load and visualize a binary dataset:
data(toyMMD)
head(toyMMD)
# Convert this dataframe into a table of sample sizes and relative frequencies:
binary_to_table(toyMMD, relative = TRUE)
```

`mmd` *Compute MMD values from a table of sample sizes and relative frequencies*

Description

Compute various MMD results, typically using a table returned by the function `binary_to_table` with the argument `relative = TRUE`.

Usage

```
mmd(data, angular = c("Anscombe", "Freeman"))
```

Arguments

data	A table of sample sizes and frequencies
angular	Choice of a formula for angular transformation: either Anscombe or Freeman-Tukey transformation.

Value

A list with three elements:

MMDMatrix	Following the presentation adopted in many research articles, a matrix filled with MMD values above the diagonal, and standard deviations of MMD below the diagonal.
MMDSym	A symmetrical matrix of MMD values, where negative values are replaced by zeroes.
MMDSignif	A matrix where any pair of traits having a significant MMD value is indicated by a star, '*'.

Author(s)

Frédéric Santos, <frederic.santos@u-bordeaux.fr>

References

Harris, E. F. and Sjøvold, T. (2004) Calculation of Smith's mean measure of divergence for intergroup comparisons using nonmetric data. *Dental Anthropology*, **17**(3), 83–93.

Nikita, E. (2015) A critical review of the mean measure of divergence and Mahalanobis distances using artificial data and new approaches to the estimation of biodistances employing nonmetric traits. *American Journal of Physical Anthropology*, **157**, 284–294. doi: 10.1002/ajpa.22708

See Also

start_mmd

Examples

```
# Load and visualize a binary dataset:
data(toyMMD)
head(toyMMD)
# Convert this dataframe into a table of sample sizes and relative
# frequencies:
tab <- binary_to_table(toyMMD, relative = TRUE)
tab
# Compute and display a symmetrical matrix of MMD values:
mmd.out <- mmd(tab, angular = "Anscombe")
mmd.out$MMDSym
# Significant MMD values are indicated by a star:
mmd.out$MMDSignif
```

`plot_mmd` *Display a multidimensional scaling (MDS) plot with the MMD dissimilarities as input*

Description

This function plots a 2D or 3D MDS to represent the MMD dissimilarities among the groups compared. Various MDS methods are proposed, and most of them are based on the R package `smacof`.

Usage

```
plot_mmd(data, method = c("classical", "interval", "ratio", "ordinal"),
axes = FALSE, gof = FALSE, dim = 2, asp = TRUE, xlim = NULL)
```

Arguments

`data`	A symmetrical matrix of MMD values; typically, it will be the component `$MMDSym` of the result returned by the function `mmd`.
`method`	Specification of MDS type. `classical` uses the metric MDS implemented in `stats::cmdscale`; the three other values are passed to the R function `smacof::smacofSym` (see its help page for more details).
`axes`	Boolean: should the axes be displayed on the plot?
`gof`	Boolean: should goodness of fit statistics be displayed on the topleft corner of the plot? More details below.
`dim`	Numeric value, 2 or 3. Indicates the maximal dimension desired for the MDS plot. It should be noted that, even with `dim = 3`, the final solution may include only two axes.
`asp`	Boolean. If `TRUE`, the same scale is used for all axes. More details below.
`xlim`	Parameter passed to `plot`, can be `NULL`.

Details

- **Axes and scale**. Making all axes use the same scale is strongly recommended in all cases (Borg et al., 2013). For a 3D-plot, since the third axis carries generally only a very small percentage of the total variability, you might want to uncheck this option to better visualize the distances along the third axis. In this case, the axes scales must be displayed on the plot, otherwise the plot would be misleading.
- **Goodness of fit values**. (i) For classical metric MDS, a common statistic is given: the sum of the eigenvalues of the first two axes, divided by the sum of all eigenvalues. It indicates the fraction of the total variance of the data represented in the MDS plot. This statistic comes from the `$GOF` value returned by the function `stats::cmdscale`. (ii) For SMACOF methods, the statistic given is the `$stress` value returned by the function `smacof::smacofSym` It indicates the final stress-1 value. A value very close to 0 corresponds to a perfect fit. (iii) For both approaches, a 'rho' value is also given, which is the Spearman's correlation coefficient between real dissimilarities (i.e., MMD values) and distances observed on the MDS plot (Dzemyda et al.,2013). A value very close to 1 indicates a perfect fit.

Value

This function returns no value by itself, and only plots a MDS in a new device.

Author(s)

Frédéric Santos, <frederic.santos@u-bordeaux.fr>

References

G. Dzemyda, O. Kurasova and J. Zilinskas (2013) *Multidimensional Data Visualization*, Springer, chap. 2, p. 39–40.

I. Borg, P. Groenen and P. Mair (2013) *Applied Multidimensional Scaling*, Springer, chap. 7, p. 79.

See Also

`start_mmd`, `stats::cmdscale`, `smacof::smacofSym`

Examples

```
# Load and visualize a binary dataset:
data(toyMMD)
head(toyMMD)
# Convert this dataframe into a table of sample sizes and relative
# frequencies:
tab <- binary_to_table(toyMMD, relative = TRUE)
tab
# Compute and display a symmetrical matrix of MMD values:
mmd.out <- mmd(tab, angular = "Freeman")
# Plot a classical metric MDS in two dimensions:
plot_mmd(data = mmd.out$MMDSym, method = "classical",
         axes = TRUE, gof = TRUE, dim = 2)
```

`select_traits` *Select a subset of traits meeting certain criteria*

Description

This function provides several strategies to discard some useless traits (non-polymorphic, non-discriminatory, etc.) upstream the MMD analysis.

Usage

```
select_traits(tab, k = 10, strategy = c("none", "excludeNPT",
"excludeQNPT", "excludeNOMD", "keepFisher"), OMDvalue = NULL, groups,
angular = c("Anscombe", "Freeman"))
```

Arguments

tab	A table of sample sizes and frequencies, typically returned by the function binary_to_table with the argument relative = TRUE.
k	Numeric value: the required minimal number of individuals per group. Any trait that could be taken on fewer individuals in at least one group will be removed from the dataset. This allows to select only the traits with a sufficient amount of information in each group.
strategy	Strategy for trait selection, i.e. for the removal of non-polymorphic traits. The four options are fully described in Santos (2018) and in the help page of StartMMD.
OMDvalue	To be specified if and only if strategy = "excludeNOMD". Set the desired threshold for the "overall measure of divergence" that must be reached for a trait to be kept.
groups	A factor or character vector, indicating the group to be considered in the analysis. Since some groups can have a very low sample size, this will allow to discard those groups in order to facilitate the trait selection via the argument k. (Otherwise, almost all traits would be removed.)
angular	Formula for angular transformation, see Harris and Sjøvold (2004). Useful only for the calculation of overall measure of divergence.

Value

A list with two components:

filtered	The dataset filtered according to the user-defined criteria.
OMD	The "overall measure of divergence" for each trait.

Author(s)

Frédéric Santos, <frederic.santos@u-bordeaux.fr>

References

Harris, E. F. and Sjøvold, T. (2004) Calculation of Smith's mean measure of divergence for intergroup comparisons using nonmetric data. *Dental Anthropology*, **17**(3), 83–93.

Santos, F. (2018) AnthropMMD: an R package with a graphical user interface for the mean measure of divergence. *American Journal of Physical Anthropology*, **165**(1), 200–205. doi: 10.1002/ajpa.23336

See Also

start_mmd

Examples

```
# Load and visualize a binary dataset:
data(toyMMD)
head(toyMMD)
# Convert this dataframe into a table of sample sizes and
# relative frequencies:
```

```
tab <- binary_to_table(toyMMD, relative = TRUE)
tab

# Filter this dataset to keep only those traits that have at
# least k=10 individuals in each group:
select_traits(tab, k = 10)
# Only Trait1 is excluded.

# Filter this dataset to keep only those traits that have at
# least k=11 individuals in each group, and show significant
# differences at Fisher's exact test:
select_traits(tab, k = 11, strategy = "keepFisher")
# Traits 1, 5 and 8 are excluded.
```

`start_mmd` *An R-Shiny application for the mean measure of divergence*

Description

Launches a graphical user interface (GUI) for the calculation of the mean measure of divergence.

Usage

```
start_mmd()
StartMMD()
```

Details

The GUI of AnthropMMD is completely autonomous: reading the data file and specifying the parameters of the analysis are done through the interface. Once the dataset is loaded, the output reacts dynamically to any change in the analysis settings.

- AnthropMMD accepts .CSV or .TXT data files, but does not support .ODS or .XLS(X) files. Two types of data input formats can be used:
  - A 'Raw binary dataset' (one row for each individual, one column for each variable). The first column must be the group indicator, and the other columns are binary data for the traits studied, where 1 indicates the presence of a trait, and 0 its absence. Row names are optional for this type of file. An example of valid data file can be found as Supporting Information online in Santos (2018).
  - A 'Table of n's and absolute frequencies for each group', i.e. a dataset of sample sizes and absolute frequencies. This type of dataset has $2 \times K$ rows (K being the number of groups compared) and p columns (p being the number of traits studied). The first K lines must be the group n's for each trait, and the last K lines are absolute frequencies for each trait (i.e. the number of times the trait is present). Row names are mandatory for this type of file. The first K rows must be labelled with names beginning with 'N_', such as: N_GroupA, N_GroupB, ..., N_GroupK. The last K rows should be labelled with names beginning with 'Freq_', such as: Freq_GroupA, ..., Freq_GroupK. An example of valid data file can be found as Supporting Information online in Santos (2018).

For both data types, column names are strongly recommended for better interpretability of the results.

- One can choose between Anscombe or Freeman-Tukey formula for angular transformation (cf. Harris and Sjøvold 2004; Irish 2010).
- 'Only retain the traits with this minimal number of individuals per group': the traits with fewer individuals in at least one active group will not be considered in the analysis.
- 'Exclusion strategy': a careful selection of traits is crucial when using MMD (cf. Harris and Sjøvold 2004 for a complete explanation), and the user should probably "exclude the traits that are nondiscriminatory across groups" (Irish 2010).
  - 'Exclude nonpolymorphic traits' removes all the traits showing no variability at all, i.e. with the same value ('0' or '1') for all individuals.
  - 'Exclude quasi-nonpolymorphic traits' also removes the traits whose variability is only due to a single individual: for example, a trait with only one positive observation in the whole dataset.
  - 'Use Fisher's exact test' implements the advice given by Harris and Sjøvold (2004) to select contributory traits, defined as those "showing a statistically significant difference between at least one pair of the groups being evaluated". Fisher's exact tests are performed for each pair of groups, and the traits showing no intergroup difference at all are excluded. Note that if you have a large number of groups (say, 10 groups), a trait with strictly equal frequencies for the last 8 groups may be considered as useful according to this criterion if there is a significant difference for the first two groups. This criterion will select all traits that can be useful for a given pair of groups, even if they are nondiscriminatory for all the other ones.
  - 'Exclude traits with overall MD' lower than a given threshold: it is a simple way of removing the traits with quite similar frequencies across groups (the 'overall MD' is defined as the sum of the variable's measures of divergence over all pairs of groups). This criterion aims to select the traits whose frequency differs substantially across most or all groups.

  These four options are designed to avoid negative MMD values.
- Some groups/populations can be manually excluded from the analysis. This may be useful if very few individuals belonging to a given population could be recorded for the variables retained by the criteria described above.
- A MDS plot and a hierarchical clustering, done using MMD dissimilarities as inputs, are displayed in the last two tabs. As MMD can sometimes be negative, those negatives values are replaced by zeros, so that the MMD matrix can be seen as a symmetrical distance matrix. Please note that the classical two-dimensional metric MDS plot cannot be displayed if there is only one positive eigenvalue. Several MDS options are proposed, cf. the help page of the `smacofSym` function from the R package `smacof` for detailed technical information.

Value

The function returns no value by itself, but all results can be individually downloaded through the graphical interface.

- The 'true' MMD values (i.e., which can be negative in the case of small samples with similar traits frequencies, cf. Irish 2010) and their standard deviations are presented in the matrix labelled 'MMD values (upper triangular part) and associated SD values (lower triangular part)'.

- A MMD value can be considered as significant if it is greater than twice its standard deviation. Significance is assessed in another ad-hoc table of results.
- The negative MMD values, if any, are replaced by zeros in the 'Symmetrical matrix of MMD values'.

Note

The R console is not available when the GUI is active. To exit the GUI, type Echap (on MS Windows systems) or Ctrl+C (on Linux systems) in the R console.

On 14-inch (or smaller) screens, for convenience, it may be necessary to decrease the zoom level of your web browser and/or to turn on fullscreen mode.

Author(s)

Frédéric Santos, <frederic.santos@u-bordeaux.fr>

References

Harris, E. F. and Sjøvold, T. (2004) Calculation of Smith's mean measure of divergence for intergroup comparisons using nonmetric data. *Dental Anthropology*, **17**(3), 83–93.

Irish, J. (2010) The mean measure of divergence: Its utility in model-free and model-bound analyses relative to the Mahalanobis D2 distance for nonmetric traits. *American Journal of Human Biology*, **22**, 378–395. doi: 10.1002/ajhb.21010

Nikita, E. (2015) A critical review of the mean measure of divergence and Mahalanobis distances using artificial data and new approaches to the estimation of biodistances employing nonmetric traits. *American Journal of Physical Anthropology*, **157**, 284–294. doi: 10.1002/ajpa.22708

Santos, F. (2018) AnthropMMD: an R package with a graphical user interface for the mean measure of divergence. *American Journal of Physical Anthropology*, **165**(1), 200–205. doi: 10.1002/ajpa.23336

Examples

```
# An example of valid binary dataset:
data(toyMMD)
head(toyMMD)
# An example of valid table:
data(absolute_freqs)
absolute_freqs
# Launch the UI:
## Not run:  start_mmd()
```

`table_relfreq` *Converts a table of sample sizes and frequencies into a table of sample sizes and relative frequencies.*

Description

Mostly used as an internal function, but could also be convenient to transform frequencies (i.e., number of individuals having a given trait) into relative frequencies (i.e., proportions).

Usage

```
table_relfreq(tab)
```

Arguments

`tab`	A table of sample sizes and frequencies, such as the tables returned by the function `binary_to_table`.

Value

The last K rows (K being the number of groups) of `tab` are simply transformed to relative frequencies.

Author(s)

Frédéric Santos, <frederic.santos@u-bordeaux.fr>

See Also

`binary_to_table`, `start_mmd`

Examples

```
# Load and visualize a binary dataset:
data(toyMMD)
head(toyMMD)
# Convert this dataframe into a table of sample sizes and frequencies:
tab <- binary_to_table(toyMMD, relative = FALSE)
tab
# Convert this table into relative frequencies:
table_relfreq(tab)
```

`toyMMD` *A toy example dataset for mean measures of divergence, in a binary format*

Description

This artifical dataset includes 200 individuals described by 9 binary traits and splitted into 5 groups. To fit with commonly observed datasets in past sciences, a substantial amount of missing values have been added at random on this dataset.

Usage

```
data(toyMMD)
```

Format

A data frame with 200 observations on the following 10 variables:

`Group` a factor with 5 levels (group indicator)

`Trait1` a numeric vector of zeroes and ones

`Trait2` a numeric vector of zeroes and ones

`Trait3` a numeric vector of zeroes and ones

`Trait4` a numeric vector of zeroes and ones

`Trait5` a numeric vector of zeroes and ones

`Trait6` a numeric vector of zeroes and ones

`Trait7` a numeric vector of zeroes and ones

`Trait8` a numeric vector of zeroes and ones

`Trait9` a numeric vector of zeroes and ones