Package ‘anscombiser’

October 12, 2022

Title  Create Datasets with Identical Summary Statistics
Version  1.1.0
Date  2022-10-03

Description  Anscombe’s quartet are a set of four two-variable datasets that have several common summary statistics but which have very different joint distributions. This becomes apparent when the data are plotted, which illustrates the importance of using graphical displays in Statistics. This package enables the creation of datasets that have identical marginal sample means and sample variances, sample correlation, least squares regression coefficients and coefficient of determination. The user supplies an initial dataset, which is shifted, scaled and rotated in order to achieve target summary statistics. The general shape of the initial dataset is retained. The target statistics can be supplied directly or calculated based on a user-supplied dataset. The ‘datasauRus’ package <https://cran.r-project.org/package=datasauRus> provides further examples of datasets that have markedly different scatter plots but share many sample summary statistics.

Imports  graphics, stats
License  GPL (>= 2)
LazyData  TRUE
Encoding  UTF-8
Depends  R (>= 3.3.0)
RoxygenNote  7.2.1
Suggests  datasauRus, datasets, gganimate, ggplot2, maps, testthat, knitr, rmarkdown
VignetteBuilder  knitr

BugReports  https://github.com/paulnorthrop/anscombiser/issues
Config/testthat/edition  3
Description

Anscombe’s quartet (Anscombe, 1973) are a set of four two-variable datasets that have several common summary statistics but which have very different joint distributions. This becomes apparent when the data are plotted, which illustrates the importance of using graphical displays in Statistics. This package enables the creation of datasets that have identical marginal sample means and sample variances, sample correlation, least squares regression coefficients and coefficient of determination. The user supplies an initial dataset, which is shifted, scaled and rotated in order to achieve target summary statistics. The general shape of the initial dataset is retained. The target statistics can be supplied directly or calculated based on a user-supplied dataset.

Details

The main functions in anscombiser are

- `anscombise`, which modifies a user-supplied dataset so that it shares sample summary statistics with Anscombe’s quartet.
- `mimic`, which modified a user-supplied dataset so that is shares sample summary statistics with another user-supplied dataset.
See vignette("intro-to-anscombiser", package = "anscombiser") for an overview of the package.

**Author(s)**

**Maintainer**: Paul J. Northrop <p.northrop@ucl.ac.uk> [copyright holder]

**References**


**See Also**

`anscombise` and `mimic`

---

### anscombe

**Anscombe’s Quartet Separated**

**Description**

Provides Anscombe’s Quartet as separate data frames.

**Usage**

```r
anscombe1
anscombe2
anscombe3
anscombe4
```

**Format**

All datasets are objects of class `data.frame` with 11 rows and 2 columns.

**Source**

Anscombe’s Quartet of ‘Identical’ Simple Linear Regressions: `datasets::anscombe` in the `datasets` package. The `i`th dataset is `datasets::anscombe[, c(i, i + 4)]`.

**References**

anscombise

Create new versions of Anscombe's quartet

Description

Modifies a dataset \( x \) so that it shares sample summary statistics with Anscombe’s quartet.

Usage

\[
\text{anscombise}(x, \text{which} = 1, \text{idempotent} = \text{TRUE})
\]

Arguments

- \( x \): A numeric matrix or data frame. Each column contains observations on a different variable. Missing observations are not allowed.
- \( \text{which} \): An integer in \( \{1, 2, 3, 4\} \). Which of Anscombe’s datasets to use as the target dataset. Obviously, this makes very little difference.
- \( \text{idempotent} \): A logical scalar. If \( \text{idempotent} = \text{TRUE} \) then applying \text{anscombise} to one of the datasets in Anscombe’s Quartet will return the dataset unchanged, apart from a change of \text{class}. If \( \text{idempotent} = \text{FALSE} \) then the returned dataset will be a rotated version of the original dataset, with the same summary statistics. See Details.

Details

The input dataset \( x \) is modified by shifting, scaling and rotating it so that its sample mean and covariance matrix match those of the Anscombe quartet.

The rotation is based on the square root of the sample correlation matrix. If \( \text{idempotent} = \text{FALSE} \) then this square root is based on the Cholesky decomposition this matrix, using \text{chol}. If \( \text{idempotent} = \text{TRUE} \) the square root is based on the spectral decomposition of this matrix, using the output from \text{eigen}. This is a minimal rotation square root, which means that if the input data \( x \) already have the exactly/approximately the required summary statistics then the returned dataset is exactly/approximately the same as the target dataset.

Value

An object of class \text{c("anscombe", "matrix", "array")} with \text{plot} and \text{print} methods. This returned dataset has the following summary statistics in common with Anscombe’s quartet.

- The sample means of each variable.
- The sample variances of each variable.
- The sample correlation matrix.
- The estimated regression coefficients from least squares linear regressions of each variable on each other variable.

The target and new summary statistics are returned as attributes \text{old_stats} and \text{new_stats} and the chosen Anscombe’s quartet dataset as an attribute \text{old_data}.
See Also

- **mimic** to modify a dataset to share sample summary statistics with another dataset.
- **datasets::anscombe** for Anscombe’s Quartet and **anscombe** for Anscombe’s Quartet as 4 separate datasets.

**input datasets**: input1 to input8 for some input dataset of the same size as those in Anscombe’s quartet.

Examples

```r
# Produce Anscombe-like datasets using input1 to input8
a1 <- anscombise(input1, idempotent = FALSE)
plot(a1)
a2 <- anscombise(input2)
plot(a2)
a3 <- anscombise(input3, idempotent = FALSE)
plot(a3)
a4 <- anscombise(input4, idempotent = FALSE)
plot(a4)
a5 <- anscombise(input5, idempotent = FALSE)
plot(a5)
a6 <- anscombise(input6)
plot(a6)
a7 <- anscombise(input7, idempotent = FALSE)
plot(a7)
a8 <- anscombise(input8, idempotent = FALSE)
plot(a8)

# Old faithful to new faithful
ew_faithful <- anscombise(datasets::faithful, which = 4)
plot(new_faithful)
# Then check that the sample summary statistics are the same
plot(new_faithful, input = TRUE)

# Map of Italy
got_maps <- requireNamespace("maps", quietly = TRUE)
if (got_maps) {
  italy <- mapdata("Italy")
  new_italy <- anscombise(italy, which = 4)
  plot(new_italy)
}
```

---

A **anscombise_gif** animation of several Anscombised datasets

Description

Create an animation to show datasets that share sample summary statistics with Anscombe’s quartet.
Usage

anscombise_gif(
  x,
  which = 1,
  idempotent = TRUE,
  theme_name = "classic",
  ease = "cubic-in-out",
  transition_length = 3,
  state_length = 1,
  wrap = TRUE
)

Arguments

x A list of input datasets. Each one must be a suitable argument for anscombise.
which, idempotent Vectors that provide the arguments of the same names to anscombise for each
dataset. If necessary, rep_len is used to replicate these arguments so that they
each have length length(x).
theme_name A character scalar used to set the ggtheme. One of "grey", "gray", "bw",
  "linedraw", "light", "dark", "minimal", "classic", "void" or "test".
ease A character scalar passed to ease_aes to control how the points move in transi-
tioning from one dataset to the next.
transition_length, state_length, wrap Arguments passed to transition_states.

Details

For this function to work the packages ggplot2 and gganimate must be installed.

Value

An object of class c("gganim", "gg", "ggplot") with an additional attribute new_data that is a
data frame with 3 variables, x, y and dataset containing the datasets output from anscombise.
The returned object may be displayed using by typing its name, e.g., anim or saved as a GIF file
using anim_save, e.g., gganimate::anim_save("anscombe.gif", anim).

See Also

anscombise modifies a dataset so that it shares sample summary statistics with Anscombe’s quartet.
input_datasets: input1 to input8 for some input datasets of the same size as those in Anscombe’s
quartet.

Examples

# Animate some Anscombe-like datasets produced using input1 to input8
x <- list(input1, input2, input3, input4, input5, input6, input7, input8)
idem <- c(FALSE, TRUE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE)
get_stats <- anscombise_gif(x, idempotent = idem)

get_stats

Calculate Anscombe's summary statistics

Description

Calculates a particular set of summary statistics for a dataset.

Usage

get_stats(x)

Arguments

x

a numeric matrix or data frame with at least 2 columns/variables. Each column contains observations on a different variable. Missing observations are not allowed.

Value

A named list of summary statistics containing

- n The sample size.
- means The sample means of each variable.
- variances The sample variances of each variable.
- correlation The sample correlation matrix.
- intercepts, slopes, rsquared Matrices whose (i,j)th entries are the estimated regression coefficients in a regression of x[, i] on x[, j] and the resulting coefficient of determination $R^2$.

Examples

get_stats(anscombe[, c(1, 5)])
input_datasets

Description

Provides input datasets from which anscombe will produce transformed datasets that behave like Anscombe’s quartet of datasets, that is, with the same traditional statistical properties but different general behaviours. Use plot(input1), for example, to see the behaviours of the datasets.

Usage

input1
input2
input3
input4
input5
input6
input7
input8

Format

All datasets are objects of class matrix (inherits from array) with 11 rows and 2 columns.

Source

None. Created for use in ‘anscombiser’.

References

mapdata

Extract longitude and latitude values

Description

Extracts longitude and latitude values for a particular region from the world map supplied by the maps package.

Usage

mapdata(region = ".", map = "world", exact = FALSE, ...)

Arguments

- region: Passed to `map` as the argument regions.
- map: Passed to `map` as the argument database
- exact: The argument exact passed to the `map` function.
- ...: Additional arguments to be passed to `map`.

Value

A dataframe with two columns: `long` and `lat` for longitude and latitude.

Examples

See the examples in `mimic`.

mimic

Modify a dataset to mimic another dataset

Description

Modifies a dataset x so that it shares sample summary statistics with a target dataset x2.

Usage

mimic(x, x2, idempotent = TRUE, ...)

Arguments

- **x, x2**
  Numeric matrices or data frames. Each column contains observations on a different variable. Missing observations are not allowed. `get_stats(x2)` sets the target summary statistics. If x2 is missing then `set_stats` is called with d = `ncol(x)` and any additional arguments supplied via `...`. This can be used to set target summary statistics (means, variances and/or correlations).

- **idempotent**
  A logical scalar. If idempotent = TRUE then `mimic(x, x)` returns x, apart from a change of `class`. If idempotent = FALSE then the returned dataset may be a rotated version of the original dataset, with the same summary statistics. See Details.

- **...**
  Additional arguments to be passed to `set_stats`.

Details

The input dataset x is modified by shifting, scaling and rotating it so that its sample mean and covariance matrix match those of the target dataset x2.

The rotation is based on the square root of the sample correlation matrix. If idempotent = FALSE then this square root is based on the Cholesky decomposition this matrix, using `chol`. If idempotent = TRUE the square root is based on the spectral decomposition of this matrix, using the output from `eigen`. This is a minimal rotation square root, which means that if the input data x already have the exactly/approximately the required summary statistics then the returned dataset is exactly/approximately the same as the target dataset x2.

Value

An object of class `c("anscombe", "matrix", "array")` with `plot` and `print` methods. This returned dataset has the following summary statistics in common with x2.

- The sample means of each variable.
- The sample variances of each variable.
- The sample correlation matrix.
- The estimated regression coefficients from least squares linear regressions of each variable on each other variable.

The target and new summary statistics are returned as attributes `old_stats` and `new_stats`. If x2 is supplied then it is returned as a attribute `old_data`.

See Also

`anscombise` modifies a dataset so that it shares sample summary statistics with Anscombe’s quartet.

Examples

```r
### 2D examples

# The UK and a dinosaur
got_maps <- requireNamespace("maps", quietly = TRUE)
got_datasauRus <- requireNamespace("datasauRus", quietly = TRUE)
```
mimic_gif

if (got_maps && got_datasauRus) {
  library(maps)
  library(datasauRus)
  dino <- datasaurus_dozen_wide[, c("dino_x", "dino_y")]
  UK <- mapdata("UK")
  new_UK <- mimic(UK, dino)
  plot(new_UK)
}

# Trump and a dinosaur
if (got_datasauRus) {
  library(datasauRus)
  dino <- datasaurus_dozen_wide[, c("dino_x", "dino_y")]
  new_dino <- mimic(dino, trump)
  plot(new_dino)
}

## Examples of passing summary statistics

# The default is zero mean, unit variance and no correlation
new_faithful <- mimic(faithful)
plot(new_faithful)

# Change the correlation
mat <- matrix(c(1, -0.9, -0.9, 1), 2, 2)
new_faithful <- mimic(faithful, correlation = mat)
plot(new_faithful)

### A 3D example

new_randu <- mimic(datasets::randu, datasets::trees)
# The samples summary statistics are equal
get_stats(new_randu)
get_stats(datasets::trees)

mimic_gif

Animation of several mimicking datasets

Description

Create an animation to show datasets that mimic a target dataset x2.

Usage

mimic_gif(
x, x2,
  idempotent = TRUE,
  theme_name = "classic",
  ease = "cubic-in-out",
...
transition_length = 3,
state_length = 1,
wrap = TRUE
)

Arguments

x A list of input datasets. Each one must be suitable argument x for for mimic.
x2 A suitable argument x2 for mimic.
idempotent A logical vector that provides the argument of the same names to mimic for each dataset. If necessary, rep_len is used to replicate this argument so that it has length length(x).
theme_name A character scalar used to set the ggtheme. One of "grey", "gray", "bw", "linedraw", "light", "dark", "minimal", "classic", "void" or "test".
ease A character scalar passed to ease_aes to control how the points move in transitioning from one dataset to the next.
transition_length, state_length, wrap Arguments passed to transition_states.

Details

For this function to work the packages ggplot2 and gganimate must be installed.

Value

An object of class c("gganim", "gg", "ggplot") with an additional attribute new_data that is a data frame with 3 variables, x, y and dataset containing the datasets output from mimic.

The returned object may be displayed using by typing its name, e.g., anim or saved as a GIF file using anim_save, e.g., gganimate::anim_save("anscombe.gif", anim).

See Also

mimic to modify a dataset to share sample summary statistics with another dataset.

input_datasets: input1 to input8 for some input datasets of the same size as those in Anscombe’s quartet.

Examples

# Create 8 datasets that mimic Anscombe's first dataset
x <- list(input1, input2, input3, input4, input5, input6, input7, input8)
anim <- mimic_gif(x, anscombel)
plot.anscombe

Description

plot method for objects inheriting from class "anscombe".

Usage

## S3 method for class 'anscombe'
plot(x, input = FALSE, stats = TRUE, digits = 3, legend_args = list(), ...)

Arguments

- **x**: an object of class 'anscombe', a result of a call to anscombise or mimic.
- **input**: A logical scalar. Should the old, input data, that is, the Anscombe’s dataset chosen for anscombise or the argument x2 to mimic, be plotted? If old = FALSE then the new, output data are plotted. If old = TRUE then the old data are plotted.
- **stats**: A logical scalar. Should the sample summary statistics n, means, variances and correlation be added to the plot?
- **digits**: An integer. The argument digits passed to signif to round the values of the statistics before adding them to the plot.
- **legend_args**: A list of arguments to be passed to legend when stats = TRUE, especially legend_args$x to control the position of the legend.
- **...**: Further arguments to be passed to plot.

Details

This function is only applicable in 2 dimensions, that is, when length(attr(x, "new_stats"))$means) = 2.

Value

Nothing is returned.

Examples

See the examples in anscombise and mimic.

See Also

anscombise and mimic.
print.anscombe

*Print method for objects of class "anscombe"*

**Description**

print method for class "anscombe".

**Usage**

```r
## S3 method for class 'anscombe'
print(x, ...)
```

**Arguments**

- `x`: an object of class "anscombe", a result of a call to `anscombise` or `mimic`.
- `...`: Additional optional arguments to be passed to `print`.

**Details**

Just extracts the new dataset from `x` and prints it using `print`.

**Value**

The argument `x`, invisibly.

**See Also**

`anscombise` and `mimic`

---

set_stats

*Create a list of summary statistics*

**Description**

Creates a list of summary statistics to pass to `mimic`.

**Usage**

```r
set_stats(d = 2, means = 0, variances = 1, correlation = diag(2))
```

**Arguments**

- `d`: An integer that is no smaller than 2.
- `means`: A numeric vector of sample means.
- `variances`: A numeric vector of positive sample variances.
- `correlation`: A numeric correlation matrix. None of the off-diagonal entries in `correlation` are allowed to be equal to 1 in absolute value.
Details
The vectors `means` and `variances` are recycled using `rep_len` to have length d.

Value
A list containing the following components.

- `means` a d-vector of sample means.
- `variances` a d-vector sample variances.
- `correlation` a d by d correlation matrix.

Examples

```r
# Uncorrelated with zero means and unit variances
set_stats()

# Sample correlation = 0.9
set_stats(correlation = matrix(c(1, 0.9, 0.9, 1), 2, 2))
```

trump

Description
A dataset that provides an image of Donald Trump’s face.

Usage
trump

Format
A matrix with 4885 rows and 2 columns: `x` and `y`.

Source
This image was created by Accentaur from the Noun Project. [https://thenounproject.com/term/donald-trump/727774/](https://thenounproject.com/term/donald-trump/727774/)

Examples

```r
plot(trump)
```
Index

* datasets
  * anscombe, 3
    * input_datasets, 8
  * trump, 15
  * PACKAGE(anscombiser-package), 2

* anim_save, 6, 12
  * anscombe, 3, 5, 8
  * Anscombe's quartet, 4–6, 8, 10
  * anscombe1(anscombe), 3
  * anscombe2(anscombe), 3
  * anscombe3(anscombe), 3
  * anscombe4(anscombe), 3
  * anscombise, 2, 3, 4, 6, 10, 13, 14
  * anscombise_gif, 5
  * anscombiser(anscombiser-package), 2
  * anscombiser-package, 2

* chol, 4, 10
* class, 4, 10

* datasets, 3
* datasets::anscombe, 3, 5

* ease_aes, 6, 12
* eigen, 4, 10

* get_stats, 7, 10
* gganimate, 6, 12
* ggplot2, 6, 12
* ggtheme, 6, 12

* input1(input_datasets), 8
* input2(input_datasets), 8
* input3(input_datasets), 8
* input4(input_datasets), 8
* input5(input_datasets), 8
* input6(input_datasets), 8
* input7(input_datasets), 8
* input8(input_datasets), 8
* input_datasets, 5, 6, 8, 12

* legend, 13

* map, 9
* mapdata, 9
* mimic, 2, 3, 5, 9, 9, 12–14
* mimic_gif, 11

* plot, 4, 10, 13
* plot.anscombe, 13
* print, 4, 10, 14
* print.anscombe, 14

* rep_len, 6, 12, 15

* set_stats, 10, 14
* signif, 13

* transition_states, 6, 12
* trump, 15