

Package ‘fairsubset’

July 22, 2025

Type Package

Title Choose Representative Subsets

Version 1.0

Date 2020-08-14

Author Joe Delaney

Maintainer Joe Delaney <delaneyj@musc.edu>

Description Allows user to obtain subsets of columns of data or vectors within a list. These subsets will match the original data in terms of average and variation, but have a consistent length of data per column. It is intended for use on automated data generation which may not always output the same N per replicate or sample.

URL <https://pubmed.ncbi.nlm.nih.gov/31583263/>

License GPL-3

Imports matrixStats, stats

RoxygenNote 7.1.1

Encoding UTF-8

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2020-09-17 08:50:08 UTC

Contents

fairsubset	2
Index	4

 fairsubset

fairsubset

Description

Allows user to obtain subsets of columns of data or vectors within a list. These subsets will match the original data in terms of average and variation, but have a consistent length of data per column. It is intended for use on automated data generation which may not always output the same N per replicate or sample.

Usage

```
fairSubset(
  input_list,
  subset_setting = "mean",
  manual_N = NULL,
  random_subsets = 1000
)
```

Arguments

<code>input_list</code>	A list, data frame, or matrix. If matrix or data frame, columns should represent each sample's data.
<code>subset_setting</code>	Choose from <code>c("mean", "median", "ks")</code> . Mean or median will use these averages to choose the best subset. "ks" will use the Kolmogorov Smirnov test to choose the best subset. Defaults to "mean".
<code>manual_N</code>	To manually choose how many data points should be in each sample, enter an integer value here. Otherwise, <code>fairSubset</code> chooses the length of the sample with the most data. Defaults to <code>NULL</code> .
<code>random_subsets</code>	To manually choose how many random subsets should be used to choose the best subset, enter an integer value here. Defaults to 1000.

Value

Returns a list.

`$best_subset` is a data.frame containing data best representative of original data, given the parameters chosen for `fairsubset`

`$worst_subset` is a data.frame containing data as far from the original as observed in all randomly chosen subsets. It is used solely as a comparator for the worst case scenario from randomly choosing subsets

`$report` is a data.frame of averages and variation regarding original data, best subset, and worst subset

`$warning` is a character string. If `!= ""`, it represents known errors

Author(s)

Joe Delaney

Examples

```
input_list <- list(a= stats::rnorm(100, mean = 3, sd = 2),  
b = stats::rnorm(50, mean = 5, sd = 5),  
c= stats::rnorm(75, mean = 2, sd = 0.5))  
fairSubset(input_list, subset_setting = "mean", manual_N = 10, random_subsets = 1000)$report
```

Index

* **array**

[fairsubset](#), 2

* **manip**

[fairsubset](#), 2

[fairSubset \(fairsubset\)](#), 2

[fairsubset](#), 2