Package 'ClustMC'

August 27, 2024

Title Cluster-Based Multiple Comparisons

Version 0.1.1

Description Multiple comparison techniques are typically applied following an F test from an ANOVA to decide which means are significantly different from one another. As an alternative to traditional methods, cluster analysis can be performed to group the means of different treatments into non-overlapping clusters. Treatments in different groups are considered statistically different. Several approaches have been proposed, with varying clustering methods and cut-off criteria. This package implements cluster-based multiple comparisons tests and also provides a visual representation in the form of a dendrogram. Di Rienzo, J. A., Guzman, A. W., & Casanoves, F. (2002) <jstor.org/stable/1400690>. Bautista, M. G., Smith, D. W., & Steiner, R. L. (1997) <doi:10.2307/1400402>.

License MIT + file LICENSE

Depends R (>= 2.10)

Imports cli, dplyr, graphics, lifecycle, magrittr, procs, psych, stats, usedist

Encoding UTF-8

RoxygenNote 7.3.2

Suggests ggdendro, ggplot2, knitr, rmarkdown, testthat (>= 3.0.0)

Config/testthat/edition 3

URL https://github.com/SGS2000/ClustMC, https://sgs2000.github.io/ClustMC/

BugReports https://github.com/SGS2000/ClustMC/issues

LazyData true

VignetteBuilder knitr

NeedsCompilation no

Author Santiago Garcia Sanchez [aut, cre, cph]

Maintainer Santiago Garcia Sanchez <santiagoesquel@gmail.com>

Repository CRAN

Date/Publication 2024-08-27 16:40:07 UTC

Contents

`bread` *Loaf volumes from a bread-baking experiment*

Description

Includes the volumes (ml) of 85 loaves of bread made under controlled conditions from 100-gram batches of dough made with 17 different varieties of wheat flour and 5 levels of potassium bromate (mg).

Usage

```
bread
```

Format

A tibble with 85 rows and 3 columns:

variety a factor indicating the variety of flour used.

bromate a number denoting the amount of potassium bromate used (milligrams).

volume a number denoting the volume of the loaf made under each condition (milliliters).

Details

Data from a bread-baking experiment by Larmour (1941). Later reproduced by Scheffe (1959) and then used by Duncan (1965) to contrast different multiple comparison methods. Jolliffe (1975) applies this dataset to illustrate his cluster-based test.

Source

Larmour, R. K. (1941). A comparison of hard red spring and hard red winter wheats. *Cereal Chemistry, 18*(6), 778-789. Available at: https://archive.org/details/sim_cereal-chemistry_1941-11_18_6

References

Duncan, D. B. (1965). A bayesian approach to multiple comparisons. *Technometrics, 7*(2), 171-222. doi:10.2307/1266670

Jolliffe, I. T. (1975). Cluster analysis as a multiple comparison method. *Applied Statistics: Proceedings of Conference at Dalhousie University, Halifax*, 159-168.

Scheffe, H. (1950).*The analysis of variance*. Wiley-Interscience Publication.

Examples

```
data(bread)
summary(bread)
```

`bss_test` *Bautista, Smith and Steiner test for multiple comparisons*

Description

Bautista, Smith and Steiner (BSS) test for multiple comparisons. Implements a procedure for grouping treatments following the determination of differences among them. First, a cluster analysis of the treatment means is performed and the two closest means are grouped. A nested analysis of variance from the original ANOVA is then constructed with the treatment source now partitioned into "groups" and "treatments within groups". This process is repeated until there are no differences among the group means or there are differences among the treatments within groups.

Usage

```
bss_test(
  y,
  trt,
  alpha = 0.05,
  show_plot = TRUE,
  console = TRUE,
  abline_options,
  ...
)
```

Arguments

`y`	Either a model (created with `lm()` or `aov()`) or a numerical vector with the values of the response variable for each unit.
`trt`	If `y` is a model, a string with the name of the column containing the treatments. If `y` is a vector, a vector of the same length as `y` with the treatments for each unit.
`alpha`	Numeric value corresponding to the significance level of the test. The default value is 0.05.
`show_plot`	Logical value indicating whether the constructed dendrogram should be plotted or not.
`console`	Logical value indicating whether the results should be printed on the console or not.
`abline_options`	`list` with optional arguments for the line in the dendrogram.
`...`	Optional arguments for the `plot()` function.

Value

A list with three `data.frame` and one `hclust`:

`stats`	`data.frame` containing summary statistics by treatment.
`groups`	`data.frame` indicating the group to which each treatment is assigned.
`parameters`	`data.frame` with the values used for the test. `treatments` is the total number of treatments and `alpha` is the significance level used.
`dendrogram_data`	object of class `hclust` with data used to build the dendrogram.

Author(s)

Santiago Garcia Sanchez

References

Bautista, M. G., Smith, D. W., & Steiner, R. L. (1997). A Cluster-Based Approach to Means Separation. *Journal of Agricultural, Biological, and Environmental Statistics, 2*(2), 179-197. doi:10.2307/1400402

Examples

```
data("PlantGrowth")
# Using vectors ------------------------------------------------------
weights <- PlantGrowth$weight
treatments <- PlantGrowth$group
bss_test(y = weights, trt = treatments, show_plot = FALSE)
# Using a model ------------------------------------------------------
model <- lm(weights ~ treatments)
bss_test(y = model, trt = "treatments", show_plot = FALSE)
```

---

clover *Nitrogen content of red clover plants*

Description

Includes the nitrogen content (mg) of 30 red clover plants inoculated with one of four single-strain cultures of *Rhizobium trifolii* or a composite of five *Rhizobium meliloti* strains, resulting in six treatments in total.

Usage

```
clover
```

Format

A tibble with 30 rows and 2 columns:

treatment a factor denoting the treatment applied to each plant.

nitrogen a number denoting the nitrogen content of each plant (milligrams).

Details

Data originally from an experiment by Erdman (1946), conducted in a greenhouse using a completely random design. The current dataset was presented by Steel and Torrie (1980) and later used by Bautista et al. (1997) to illustrate their proposed procedure.

Source

Steel, R., & Torrie, J. (1980). *Principles and procedures of statistics: A biometrical approach (2nd ed.)*. San Francisco: McGraw-Hill. Available at: https://archive.org/details/principlesproce00stee

References

Bautista, M. G., Smith, D. W., & Steiner, R. L. (1997). A Cluster-Based Approach to Means Separation. *Journal of Agricultural, Biological, and Environmental Statistics, 2*(2), 179-197. doi:10.2307/1400402

Erdman, L. W. (1946). Studies to determine if antibiosis occurs among rhizobia. *Journal of the American Society of Agronomy, 38*, 251-258. doi:10.2134/agronj1946.00021962003800030005x

Examples

```
data(clover)
summary(clover)
```

---

dgc_test *Di Rienzo, Guzman and Casanoves test for multiple comparisons*

Description

Di Rienzo, Guzman and Casanoves (DGC) test for multiple comparisons. Implements a cluster-based method for identifying groups of nonhomogeneous means. Average linkage clustering is applied to a distance matrix obtained from the sample means. The distribution of Q (distance between the source and the root node of the tree) is used to build a test with a significance level of α. Groups whose means join above c (the α-level cut-off criterion) are statistically different.

Usage

```
dgc_test(
  y,
  trt,
  alpha = 0.05,
  show_plot = TRUE,
  console = TRUE,
  abline_options,
  ...
)
```

Arguments

y	Either a model (created with lm() or aov()) or a numerical vector with the values of the response variable for each unit.
trt	If y is a model, a string with the name of the column containing the treatments. If y is a vector, a vector of the same length as y with the treatments for each unit.
alpha	Value equivalent to 0.05 or 0.01, corresponding to the significance level of the test. The default value is 0.05.
show_plot	Logical value indicating whether the constructed dendrogram should be plotted or not.
console	Logical value indicating whether the results should be printed on the console or not.
abline_options	list with optional arguments for the line in the dendrogram.
...	Optional arguments for the plot() function.

Value

A list with three data.frame and one hclust:

stats	data.frame containing summary statistics by treatment.
groups	data.frame indicating the group to which each treatment is assigned.
parameters	data.frame with the values used for the test. treatments is the total number of treatments, alpha is the significance level used, c is the cut-off criterion for the dendrogram (the height of the horizontal line on the dendrogram), q is the $1 - \alpha$ quantile of the distribution of Q (distance from the root node) under the null hypothesis and SEM is an estimate of the standard error of the mean.
dendrogram_data	object of class hclust with data used to build the dendrogram.

Author(s)

Santiago Garcia Sanchez

References

Di Rienzo, J. A., Guzman, A. W., & Casanoves, F. (2002). A Multiple-Comparisons Method Based on the Distribution of the Root Node Distance of a Binary Tree. *Journal of Agricultural, Biological, and Environmental Statistics, 7*(2), 129-142. <jstor.org/stable/1400690>

Examples

```
data("PlantGrowth")
# Using vectors ------------------------------------------------------
weights <- PlantGrowth$weight
treatments <- PlantGrowth$group
dgc_test(y = weights, trt = treatments, show_plot = FALSE)
# Using a model ------------------------------------------------------
model <- lm(weights ~ treatments)
dgc_test(y = model, trt = "treatments", show_plot = FALSE)
```

---

jolliffe_test *Jolliffe test for multiple comparisons*

Description

I.T. Jolliffe test for multiple comparisons. Implements a cluster-based alternative closely linked to the Student-Newman-Keuls multiple comparison method. Single-linkage cluster analysis is applied, using the p-values obtained with the SNK test for pairwise mean comparison as a similarity measure. Groups whose means join beyond $1 - \alpha$ are statistically different. Alternatively, complete linkage cluster analysis can also be applied.

Usage

```
jolliffe_test(
  y,
  trt,
  alpha = 0.05,
  method = "single",
  show_plot = TRUE,
  console = TRUE,
  abline_options,
  ...
)
```

Arguments

y	Either a model (created with `lm()` or `aov()`) or a numerical vector with the values of the response variable for each unit.
trt	If `y` is a model, a string with the name of the column containing the treatments. If `y` is a vector, a vector of the same length as `y` with the treatments for each unit.

alpha	Numeric value corresponding to the significance level of the test. The default value is 0.05.
method	string indicating the clustering method to be used. For single linkage (the default method) either "single" or "slca". For complete linkage, either "complete" or "clca".
show_plot	Logical value indicating whether the constructed dendrogram should be plotted or not.
console	Logical value indicating whether the results should be printed on the console or not.
abline_options	list with optional arguments for the line in the dendrogram.
...	Optional arguments for the plot() function.

Value

A list with three data.frame and one hclust:

stats	data.frame containing summary statistics by treatment.
groups	data.frame indicating the group to which each treatment is assigned.
parameters	data.frame with the values used for the test. treatments is the total number of treatments, alpha is the significance level used, n is either the number of repetitions for all treatments or the harmonic mean of said repetitions, MSE is the mean standard error from the ANOVA table and SEM is an estimate of the standard error of the mean.
dendrogram_data	object of class hclust with data used to build the dendrogram.

Author(s)

Santiago Garcia Sanchez

References

Jolliffe, I. T. (1975). Cluster analysis as a multiple comparison method. *Applied Statistics: Proceedings of Conference at Dalhousie University, Halifax*, 159-168.

Examples

```
data("PlantGrowth")
# Using vectors ------------------------------------------------------
weights <- PlantGrowth$weight
treatments <- PlantGrowth$group
jolliffe_test(y = weights, trt = treatments, alpha = 0.1, show_plot = FALSE)
# Using a model ------------------------------------------------------
model <- lm(weights ~ treatments)
jolliffe_test(y = model, trt = "treatments", alpha = 0.1, show_plot = FALSE)
```

Index