Package 'vacuum'

October 12, 2022

Type Package

Title Tukey's Vacuum Cleaner

Version 0.1.0

Author Ron Sielinski

Maintainer Ron Sielinski <sielinski@hotmail.com>

Description An implementation of three procedures developed by John Tukey: FUNOP (FUll NOrmal Plot), FUNOR-FUNOM (FUll NOrmal Rejection-FUll NOrmal Modification), and vacuum cleaner. Combined, they provide a way to identify, treat, and analyze outliers in two-way (i.e., contingency) tables, as described in his landmark paper ``The Future of Data Analysis'', Tukey, John W. (1962) <https://www.jstor.org/stable/2237638>.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Suggests testthat, ggplot2, knitr, tidyr, rmarkdown

Imports dplyr, magrittr

Depends R (>= 2.10)

URL https://github.com/Sielinski/vacuum

BugReports https://github.com/Sielinski/vacuum/issues

RoxygenNote 7.1.1

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2020-09-08 08:50:02 UTC

R topics documented:

`a_qnorm` *Returns the typical value from a unit-normal distribution*

Description

Returns the typical value from a unit-normal distribution for the *i*th ordered observation in an *n*-sized sample.

This is a helper function for FUNOP, which uses the output of this function as the denominator for its slope calculation.

Usage

```
a_qnorm(i, n)
```

Arguments

`i`	Non-zero index of an array
`n`	Non-zero length of the array

Value

Quantile of `i` from a unit-normal distribution

References

Tukey, John W. "The Future of Data Analysis." *The Annals of Mathematical Statistics*, *33*(1), 1962, pp 1-67. *JSTOR*, https://www.jstor.org/stable/2237638.

See Also

`funop()`

Examples

```
a_qnorm(i = 25, n = 42)
a_qnorm(21.5, 42)
```

`funop` *Identifies outliers in a numeric vector*

Description

FUNOP stands for FUll NOrmal Plot.

The procedure identifies outliers by calculating their slope (`z`), relative to the vector's median.

The procedure ignores values in the middle third of the *ordered* vector. The remaining values are all candidates for consideration. The slopes of all candidates are calculated, and the median of their slopes is used as the primary basis for identifying outliers.

Any value whose slope is `B` times larger than the median slope is identified as an outlier. Additionally, any value whose *magnitude* is larger than that of the slope-based outliers is also identified as an outlier.

However, the procedure will *not* identify as outliers any values within `A` standard deviations of the vector's median (i.e., not the median of candidate slopes).

Usage

```
funop(x, A = 0, B = 1.5)
```

Arguments

`x`	Numeric vector to inspect for outliers (does not need to be ordered)
`A`	Number of standard deviations beyond the median of `x`
`B`	Multiples beyond the median slope of candidate values

Value

A data frame containing one row for every member of `x` (in the same order as `x`) and the following columns:

- `y`: Original values of vector `x`
- `i`: Ordinal position of value `y` in the sorted vector `x`
- `middle`: Boolean indicating whether ordinal position `i` is in the middle third of the vector
- `a`: Result of `a_qnorm(i, length(x))`
- `z`: Slope of `y` relative to `median(y)`
- `special`: Boolean indicating whether `y` is an outlier

References

Tukey, John W. "The Future of Data Analysis." *The Annals of Mathematical Statistics*, *33*(1), 1962, pp 1-67. *JSTOR*, https://www.jstor.org/stable/2237638.

See Also

a_qnorm()

Examples

```
funop(c(1, 2, 3, 11))
funop(table_1)

attr(funop(table_1), 'z_split')
```

`funor_funom` *Identifies and treats outliers in a two-way table*

Description

FUNOR-FUNOM stands for FUll NOrmal Rejection-FUll NOrmal Modification.

The procedure treats a two-way (contingency) table for outliers by isolating residuals from the table's likely systemic effects, which are calculated from the table's grand, row, and column means.

The residuals are passed to separate *rejection* (FUNOR) and *modification* (FUNOM) procedures, which both depend upon FUNOP to identify outliers. As such, this procedure requires two sets of `A` and `B` parameters.

The procedure treats outliers by reducing their residuals, resulting in values that are much closer to their expected values (i.e., combined grand, row, and column effects).

Usage

```
funor_funom(x, A_r = 10, B_r = 1.5, A_m = 0, B_m = 1.5)
```

Arguments

`x`	Two-way table to treat for outliers
`A_r`	A for the FUNOR phase (see FUNOP for details)
`B_r`	B for the FUNOR phase slope
`A_m`	A for the FUNOM phase (`A_m` is usually 0)
`B_m`	B for the FUNOM phase

Value

A two-way table of the same size as `x`, treated for outliers.

References

Tukey, John W. "The Future of Data Analysis." *The Annals of Mathematical Statistics*, *33*(1), 1962, pp 1-67. *JSTOR*, https://www.jstor.org/stable/2237638.

See Also

funop()

Examples

```
funor_funom(table_2)
which(funor_funom(table_2) != table_2)
```

---

table_1 *Table 1*

Description

Example data taken from Table 1 of John Tukey's "Future of Data Analysis."

Usage

```
table_1
```

Format

A numeric vector containing 14 elements.

Source

Tukey, John W. "The Future of Data Analysis." *The Annals of Mathematical Statistics*, *33*(1), 1962, pp 1-67. *JSTOR*, https://www.jstor.org/stable/2237638.

Examples

```
table_1
funop(table_1)
```

---

table_2 *Table 2*

Description

Example data taken from Table 2 of John Tukey's "Future of Data Analysis."

Usage

```
table_2
```

Format

A 36x15 numeric matrix.

Source

Tukey, John W. "The Future of Data Analysis." *The Annals of Mathematical Statistics*, *33*(1), 1962, pp 1-67. *JSTOR*, https://www.jstor.org/stable/2237638.

Examples

```
table_2
funor_funom(table_2)
```

`table_8` *Table 8*

Description

Example data taken from Table 8 of John Tukey's "Future of Data Analysis." Note that this dataset fixes a typo in the original document: Column 23, row 10 contains 0.100, corrected from the original -0.100.

Usage

```
table_8
```

Format

A 36x15 numeric matrix.

Source

Tukey, John W. "The Future of Data Analysis." *The Annals of Mathematical Statistics*, *33*(1), 1962, pp 1-67. *JSTOR*, https://www.jstor.org/stable/2237638.

Examples

```
table_8
vacuum_cleaner(table_8)
```

`vacuum_cleaner` *Returns the residuals of a two-way table after removing systemic effects*

Description

To remove systemic effects from values in a contingency table, vacuum cleaner uses regression to identify the table's main effect (dual regression), row effect (deviations of row regression from dual regression), and column effect (deviations of column regression from dual regression).

Regression is performed twice: First on the table's original values, then on the resulting residuals. The output is a table of residuals "vacuum cleaned" of likely systemic effects.

Usage

```
vacuum_cleaner(x)
```

Arguments

`x`	Two-way table to analyze (must be 3x3 or greater).

Value

Residuals of `x`

References

Tukey, John W. "The Future of Data Analysis." *The Annals of Mathematical Statistics*, *33*(1), 1962, pp 1-67. *JSTOR*, https://www.jstor.org/stable/2237638.

See Also

`funop()`, `funor_funom()`

Examples

```
vacuum_cleaner(table_8)
```

Index