Package ‘univOutl’

March 17, 2021

Type Package

Title Detection of Univariate Outliers

Version 0.3

Date 2021-03-16

Author Marcello D'Orazio

Maintainer Marcello D'Orazio <mdo.statmatch@gmail.com>

Depends robustbase, Hmisc

Description Well known outlier detection techniques in the univariate case. Methods to deal with skewed distribution are included too. The Hidiroglou-Berthelot (1986) method to search for outliers in ratios of historical data is implemented as well. When available, survey weights can be used in outliers detection.

License GPL (>= 2)

URL https://github.com/marcellodo/univOutl

NeedsCompilation no

Repository CRAN

Date/Publication 2021-03-17 09:10:02 UTC

R topics documented:

univOutl-package *Detection of Univariate Outliers.*

Description

Well known outlier detection techniques in the univariate case. Ratios of two variables are covered too. When available, survey weights can be considered.

Details

The package provides few simple functions implementing well known outlier detection techniques in the univariate case. Methods to deal with skewed distributions are included. The Hidiroglou-Berthelot (1986) method to search for outliers in ratios of historical data is implemented as well. When available, the survey weights can be used in outliers' detection.

Author(s)

Author and Maintainer: Marcello D'Orazio <mdo.statmatch@gmail.com>

References

Hidiroglou, M.A. and Berthelot, J.-M. (1986) 'Statistical editing and Imputation for Periodic Business Surveys'. *Survey Methodology*, Vol 12, pp. 73-83.

McGill, R., Tukey, J. W. and Larsen, W. A. (1978) 'Variations of box plots'. *The American Statistician*, 32, pp. 12-16.

Rousseeuw, P.J. and Croux, C. (1993) 'Alternatives to the Median Absolute Deviation', *Journal of the American Statistical Association* 88, pp. 1273-1283.

Hubert, M., and Vandervieren, E. (2008) 'An Adjusted Boxplot for Skewed Distributions', *Computational Statistics & Data Analysis*, 52, pp. 5186-5201

boxB *BoxPlot based outlier detection*

Description

Identifies univariate outliers by using methods based on BoxPlots

Usage

```
boxB(x, k=1.5, method='asymmetric', weights=NULL, id=NULL,
     exclude=NA, logt=FALSE)
```

Arguments

`x` Numeric vector that will be searched for outliers.

`k` Nonnegative constant that determines the extension of the 'whiskers'. Commonly used values are 1.5 (default), 2, or 3. Note that when `method="adjbox"` then `k` is set automatically equal to 1.5

`method` Character, identifies the method to be used: `method="resistant"` provides the 'standard' boxplot fences; `method="asymmetric"` is a modification of standard method to deal with (moderately) skewed data; `method="adjbox"` uses Hubert and Vandervieren (2008) adjusted boxplot for skewed distributions.

`weights` Optional numeric vector with units' weights associated to the observations in `x`. Only nonnegative weights are allowed. Weights are used in estimating the quartiles (see Details).

`id` Optional vector with identifiers of units in `x`. If missing (`id=NULL`, default) the identifiers will be set equal to the positions in the vector (i.e. `id=1:length(x)`).

`exclude` Values of `x` that will be excluded by the analysis. By default missing values are excluded (`exclude=NA`).

`logt` Logical, if `TRUE`, before searching outliers the `x` variable is log-transformed (`log(x+1)` is considered). In this case the summary outputs (bounds, etc.) will refer to the log-transformed `x`

Details

When `method="resistant"` the outlying observations are those outside the interval:

$$[Q_1 - k \times IQR; \quad Q_3 + k \times IQR]$$

where Q_1 and Q_3 are respectively the 1st and the 3rd quartile of `x`, while $IQR = (Q_3 - Q_1)$ is the Inter-Quartile Range. The value $k = 1.5$ (said 'inner fences') is commonly used when drawing a boxplot. Values $k = 2$ and $k = 3$ provide middle and outer fences, respectively.

When `method="asymmetric"` the outlying observations are those outside the interval:

$$[Q_1 - 2k \times (Q_2 - Q_1); \quad Q_3 + 2k \times (Q_3 - Q_2)]$$

being Q_2 the median; such a modification allows to account for slight skewness of the distribution.

Finally, when `method="adjbox"` the outlying observations are identified using the method proposed by Hubert and Vandervieren (2008) and based on the Medcouple measure of skewness; in practice the bounds are:

$$[Q_1 - 1.5 \times e^{aM} \times IQR; \quad Q_3 + 1.5 \times e^{bM} \times IQR]$$

Where M is the medcouple; when $M > 0$ (positive skewness) then $a = -4$ and $b = 3$; on the contrary $a = -3$ and $b = 4$ for negative skewness ($M < 0$). This adjustment of the boxplot, according to Hubert and Vandervieren (2008), works with moderate skewness ($-0.6 \leq M \leq 0.6$). The bounds of the adjusted boxplot are derived by applying the function `adjboxStats` in the package **robustbase**.

When weights are available (passed via the argument weights) then they are used in the computation of the quartiles. In particular, the quartiles are derived using the function wtd.quantile in the package **Hmisc**.

Remember that when asking a log transformation (argument logt=TRUE) all the estimates (quartiles, etc.) will refer to $log(x+1)$.

Value

The output is a list containing the following components:

quartiles The quartiles of x after discarding the values in the exclude argument. When weights are provided they are used in quartiles estimation trough the function wtd.quantile in the package **Hmisc**.

fences The bounds of the interval, values outside the interval are detected as outliers.

excluded The identifiers or positions (when id=NULL) of units in x excluded by the computations, according to the argument exclude.

outliers The identifiers or positions (when id=NULL) of units in x detected as outliers.

lowOutl The identifiers or positions (when id=NULL) of units in x detected as outliers in the lower tail of the distribution.

upOutl The identifiers or positions (when id=NULL) of units in x detected as outliers in the upper tail of the distribution.

Author(s)

Marcello D'Orazio <mdo.statmatch@gmail.com>

References

McGill, R., Tukey, J. W. and Larsen, W. A. (1978) 'Variations of box plots'. *The American Statistician*, 32, pp. 12-16.

Hubert, M., and Vandervieren, E. (2008) 'An Adjusted Boxplot for Skewed Distributions', *Computational Statistics and Data Analysis*, 52, pp. 5186-5201.

See Also

adjboxStats, wtd.quantile

Examples

```
set.seed(321)
x <- rnorm(30, 50, 10)
x[10] <- 1
x[20] <- 100

out <- boxB(x = x, k = 1.5, method = 'asymmetric')
out$fences
out$outliers
```

```
x[out$outliers]

out <- boxB(x = x, k = 1.5, method = 'adjbox')
out$fences
out$outliers
x[out$outliers]

x[24] <- NA
x.ids <- paste0('obs',1:30)
out <- boxB(x = x, k = 1.5, method = 'adjbox', id = x.ids)
out$excluded
out$fences
out$outliers

set.seed(111)
w <- round(runif(n = 30, min=1, max=10))
out <- boxB(x = x, k = 1.5, method = 'adjbox', id = x.ids, weights = w)
out$excluded
out$fences
out$outliers
```

`HBmethod` *Hidiroglou-Berthelot procedure for detecting outliers with periodic data*

Description

This function implements the method proposed by Hidiroglou and Berthelot (1986) to identify outliers in periodic data, i.e. when the same variable is measured at two time points.

Usage

```
HBmethod(yt1, yt2, U=0.5, A=0.05, C=4, pct=0.25,
         id=NULL, std.score=FALSE, return.dataframe=FALSE, adjboxE=FALSE)
```

Arguments

`yt1` Numeric vector providing the values observed at time t1.

`yt2` Numeric vector providing the values observed at time t2 (t2 > t1).

`U` Numeric, parameter needed to determine the ‘importance’ of a ratio. The value should lie in [0, 1] interval; commonly used values are 0.3, 0.4, or 0.5 (default) (see Details for further information).

`A` Numeric, parameter needed when computing the scale measure used to derive the bounds. Hidiroglou and Berthelot (1986) suggest setting $A = 0.05$ (default) (see Details for further information).

`C` Numeric, parameter determining the extension of the interval; greater values will provide larger intervals, i.e. fewer expected outliers. Values commonly used are 4 (default) or 7, but also values close or grater than 40 can be used in some particular cases. Note that two C values can be provided instead of one, the first one will be used to determine the left tail bound, while the second determines the right tail bound; this setting can help in improving outlier detection in skewed distributions (see Details for further information).

`pct` Numeric, the percentage point of the scores that will be used to calculate the lower and upper bounds. By default, `pct = 0.25`, i.e. quartiles Q1 and Q3 are considered. In some cases, as suggested by Hidiroglou and Emond (2018), using `pct = 0.10`, i.e. percentiles P10 and P90, may be a better choice. Se Details for further information.

`id` Optional numeric or character vector, with identifiers of units. If `id=NULL` units identifiers will be set equal to their position.

`std.score` Logical, if `TRUE` the output will include a standardized score variable (see Details, for further information)

`return.dataframe`

Logical, if `TRUE` the output will save all the relevant information for outlier detection in a dataframe with the following columns: 'id' (units' identifiers), ‘yt1’, ‘yt2’, ‘ratio’ (= yt1/yt2), ‘sizeU’ (=max(yt1, yt2)^U),‘Escore’ (the E scores, see Details), ‘std.Escore’ (the standardized E scores when `std.score=TRUE`, see Details) and finally ‘outliers’, where value 1 indicates observations detected as an outlier, 0 otherwise.

`adjboxE` Logical (default `FALSE`), if `TRUE` an additional search of outliers will be done on the E-scores using the boxplot adjusted for skewness as implemented in the function `boxB` when run with with argument `method = "adjbox"`.

Details

The method proposed by Hidiroglou and Berthelot (1986) to identify outliers in periodic data consists in deriving a score variable based on the ratios $r_i = y_{i,t2}/y_{i,t1}$ (`yt2/yt1`) with $i = 1, 2, \dots, n$ being n the number of observations after discarding NAs and 0s in both `yt1` and `yt2`.

At first the ratios are centered around their median r_M:

$$s_i = 1 - r_M/r_i \qquad \text{if} \qquad 0 < r_i < r_M$$

$$s_i = r_i/r_M - 1 \qquad \text{if} \qquad r_i \geq r_M$$

Then, in order to account for the magnitude of data, the following score is derived:

$$E_i = s_i \times [max(y_{i,t1}, y_{i,t2})]^U$$

Finally, the interval is calculated as:

$$(E_M - C \times d_{Q1}, E_M + C \times d_{Q3})$$

where

$d_{Q1} = max(E_M - E_{Q1}, |A \times E_M|)$ and $d_{Q3} = max(E_{Q3} - E_M, |A \times E_M|)$

being d_{Q1}, E_M and d_{Q3} the quartiles of E scores when `pct = 0.25` (default). Recently Hidiroglou and Emond (2018) suggest using percentiles P10 and P90 of the E scores in replacement of respectively Q1 and Q3 to avoid the drawback of many units identified as outliers; this is likely to occur when a large proportion of units (>1/4) has the same ratio. P10 and P90 are achieved by setting `pct = 0.10` when running the function.

In practice, all the units with an E score outside the interval are considered as outliers. Notice that when two `C` values are provided, then the first is used to derive the left bound while the second determines the right bound.

When `std.score=TRUE` a standardized score is derived in the following manner:

$$z_{E,i} = g \times \frac{E_i - E_M}{d_{Q1}} \qquad \text{if} \qquad E_i < E_M$$

$$z_{E,i} = g \times \frac{E_i - E_M}{d_{Q3}} \qquad \text{if} \qquad E_i \geq E_M$$

The constant g is set equal to `qnorm(1-pct)` and makes d_{Q1} and d_{Q3} approximately unbiased estimators when the E scores follow the normal distribution.

When `adjboxE = TRUE` outliers on the E scores will all be searched using the boxplot adjusted for skewness as implemented in the function `boxB` when run with with argument `method = "adjbox"`.

Value

A list whose components depend on the `return.dataframe` argument. When `return.dataframe=FALSE` just the following components are provided:

`median.r`	the median of the ratios
`quartiles.E`	Quartiles of the E score
`bounds.E`	Bounds of the interval of the E score, values outside are considered outliers.
`excluded`	The identifiers or positions (when `id=NULL`) of units in both `yt1` and `yt2` that are excluded by the outliers detection, i.e. NAs and 0s.
`outliers`	The identifiers or positions (when `id=NULL`) of units in `yt1` or `yt2` identified as outliers.
`outliersBB`	The identifiers or positions (when `id=NULL`) of units in `yt1` or `yt2` identified as outliers by applying the boxplot adjusted for skewness to the E scores. This component appears in the output only when `adjboxE = TRUE`.

When `return.dataframe=TRUE`, the first three components remain the same with, in addition, two dataframes:

`excluded`	A dataframe with the subset of observations excluded. The data frame has the following columns: 'id' (units' identifiers), 'yt1' columns 'yt2'

`data` A dataframe with the the not excluded observations and the following columns: ‘id’ (units’ identifiers), ‘yt1’, ‘yt2’, ‘ratio’ (= yt1/yt2), ‘sizeU’ (=max(yt1, yt2)^U),‘Escore’ (the E scores, see Details), ‘std.Escore’ (the standardized E scores when `std.score=TRUE`, see Details) and ‘outliers’, where value 1 indicates observations detected as an outlier, 0 otherwise. in addition the column ‘outliersBB’ will also be included when `adjboxE = TRUE`.

Author(s)

Marcello D’Orazio <`mdo.statmatch@gmail.com`>

References

Hidiroglou, M.A. and Berthelot, J.-M. (1986) ‘Statistical editing and Imputation for Periodic Business Surveys’. *Survey Methodology*, Vol 12, pp. 73-83.

Hidiroglou, M.A. and Emond, N. (2018) ‘Modifying the Hidiroglou-Berthelot (HB) method’. *Unpublished note*, Business Survey Methods Division, Statistics Canada, May 18 2018.

See Also

`plot4ratios`, `ratioSize`

Examples

```
set.seed(222)
x0 <- rnorm(30, 50, 5)
x0[1] <- NA
set.seed(333)
rr <- runif(30, 0.9, 1.2)
rr[10] <- 2
x1 <- x0 * rr
x1[20] <- 0

out <- HBmethod(yt1 = x0, yt2 = x1)
out$excluded
out$median.r
out$bounds.E
out$outliers
cbind(x0[out$outliers], x1[out$outliers])

out <- HBmethod(yt1 = x0, yt2 = x1,
                return.dataframe = TRUE)
out$excluded
head(out$data)
```

LocScaleB *Univariate outlier detection with bounds based on robust location and scale estimates*

Description

This function identifies outliers in the tails of a distribution by detecting the observations outside the bounds built using a robust estimate of both location and scale parameters.

Usage

```
LocScaleB(x, k=3, method='MAD',  weights=NULL, id=NULL,
          exclude=NA, logt=FALSE, return.dataframe=FALSE)
```

Arguments

`x` Numeric vector that will be searched for outliers.

`k` Nonnegative constant that determines the extension of bounds. Commonly used values are 2, 2.5 and 3 (default).

`method` character identifying how to estimate the scale of the distribution. Available choices are:

`method='IQR'` for using the Inter-Quartile Range, i.e. Q3-Q1;

`method='IDR'` for using the Inter-Decile Range; i.e. P90-P10

`method='MAD'` for using the Median Absolute Deviation;

`method='Gini'` robust scale estimate based on Gini's Mean Difference (see `GiniMd`);

`method='ScaleTau2'` robust tau-estimate of univariate scale, as proposed by Maronna and Zamar (2002) (see also`scaleTau2`);

`method='Qn'` for using the Qn estimator proposed by Rousseeuw and Croux (1993) (see also `Qn`);

`method='Sn'` for using the Sn estimator proposed by Rousseeuw and Croux (1993) (see also `Sn`).

When `method='dQ'` the estimated scale for the left tail is (Q2-Q1)/0.6745, while for the right tail it is considered (Q3-Q2)/0.6745 (Q2 is the median); this double estimate should be able to account for slight skewness.

When `method='dD'` the estimated scale for the left tail is (P50-P10)/1.2816, while for the right tail it is considered (P90-P50)/1.2816 (P50 is the median); this double estimate should be able to account for skewness.

Finally, when `method='AdjOut'`, bounds are based on the adjusted outlyingness method as proposed by Hubert and Van der Veeken (2008).

`weights` Optional numeric vector that provides weights associated to observations. Only nonnegative weights are allowed. Note that weights can only be used when `method='MAD'`, `method='IQR'`, `method='IDR'`, `method='dQ'` or `method='dD'`.

`id` Optional numeric or character vector, with identifiers of units in `x`. If `id=NULL` (default) units' identifiers will be set equal to their position in `x`.

`exclude` Values of `x` that will be excluded by the analysis. By default missing values (`exclude = NA`)

`logt` Logical, if `TRUE`, before searching outliers the `x` variable is log-transformed (`log(x+1)` is considered). Note that in this case that summary output (bounds, etc.) will refer to log-transformed variable.

`return.dataframe` Logical, if `TRUE` the output will save all the relevant information for outlier detection in a dataframe with the following columns: 'id' (units' identifiers), 'x', 'log.x' (only if `logt=TRUE`), 'weight' (only when argument `weights` is provided), 'score' (the standardized scores, see Details) and, finally, 'outliers', where value 1 indicates observations detected as an outlier, 0 otherwise.

Details

The intervals are derived by considering the median Q_2 as a robust location estimate while different robust scale estimators are considered:

$$[Q_2 - k \times \tilde{s}_L; \quad Q_2 + k \times \tilde{s}_R]$$

where $\tilde{s}_L$ and $\tilde{s}_R$ are robust scale estimates. With most of the methods $\tilde{s}_L = \tilde{s}_R$ with exception of `method='dQ'` and `method='dD'` where respectively:

$$\tilde{s}_L = (Q_2 - Q_1)/0.6745 \qquad \text{and} \qquad \tilde{s}_R = (Q_3 - Q_2)/0.6745$$

and

$$\tilde{s}_L = (P_{50} - P_{10})/1.2816 \qquad \text{and} \qquad \tilde{s}_R = (P_{90} - P_{50})/1.2816$$

Note that when `method='dQ'` or `method='dD'` the function calculates and prints a the Bowley's coefficient of skewness, that uses Q1, Q2 and Q3 (they are replaced by respectively P10, P50 and P90 when `method='dD'`).

With `method='AdjOut'` the following estimates are considered:

$$\tilde{s}_L = (Q_2 - f_L) \qquad \text{and} \qquad \tilde{s}_R = (f_R - Q_2)$$

being f_R and f_L derived starting from the fences of the adjusted boxplot (Hubert and Vandervieren, 2008; see `adjboxStats`). In addition the medcouple (`mc`) measure of skewness is calculated and printed on the screen.

When weights are available (passed via the argument `weights`) then they are used in the computation of the quartiles. In particular, the quartiles are derived using the function `wtd.quantile` in the package **Hmisc**. Note that their use is allowed just with `method='IQR'`, `method='IDR'`, `method='dQ'`, `method='dD'` or `method='AdjOut'`.

The 'score' variable reported in the the `data` dataframe when `return.dataframe=TRUE` is the standardized score derived as (x - Median)/scale.

Value

A list whose components depend on the `return.dataframe` argument. When `return.dataframe = FALSE` just the following components are provided:

`pars` Vector with estimated median and scale parameters

`bounds` The bounds of the interval, values outside the interval are considered outliers.

`excluded` The position or identifiers of `x` values excluded by outlier detection, according to the argument `exclude`

`outliers` The position or identifiers of `x` values detected as outliers (outside bounds).

`lowOutl` The identifiers or positions (when `id=NULL`) of units in `x` detected as outliers in the lower tail of the distribution.

`upOutl` The identifiers or positions (when `id=NULL`) of units in `x` detected as outliers in the upper tail of the distribution.

When `return.dataframe=TRUE` the latter two components are substituted with two dataframes:

`excluded` A dataframe with the subset of observations excluded.

`data` A dataframe with the the not excluded observations and the following columns: ‘id’ (units’ identifiers), ‘x’, ‘log.x’ (only if `logt=TRUE`), ‘weight’ (only when argument `weights` is provided), ‘score’ (the standardized scores, see Details) and, finally, ‘outliers’, where value 1 indicates observations detected as an outlier and 0 otherwise.

Author(s)

Marcello D’Orazio <`mdo.statmatch@gmail.com`>

References

Hubert, M. and Van der Veeken, S. (2008) ‘Outlier Detection for Skewed Data’. *Journal of Chemometrics*, 22, pp. 235-246.

Maronna, R.A. and Zamar, R.H. (2002) ‘Robust estimates of location and dispersion of high-dimensional datasets’ *Technometrics*, 44, pp. 307-317.

Rousseeuw, P.J. and Croux, C. (1993) ‘Alternatives to the Median Absolute Deviation’, *Journal of the American Statistical Association* 88, pp. 1273-1283.

Vanderviere, E. and Huber, M. (2008) ‘An Adjusted Boxplot for Skewed Distributions’, *Computational Statistics & Data Analysis*, 52, pp. 5186-5201

See Also

`mad`, `scaleTau2`, `Qn`, `Sn`, `GiniMd`

Examples

```
set.seed(333)
x <- rnorm(30, 50, 1)
x[10] <- 1
x[20] <- 100

out <- LocScaleB(x = x,  k = 3, method='MAD')
out$pars
out$bounds
out$outliers
x[out$outliers]

out <- LocScaleB(x = x,  k = 3, method='MAD',
                 return.dataframe = TRUE)
head(out$data)

out <- LocScaleB(x = x, k = 3, method='AdjOut')
out$outliers
```

---

plot4ratios *Draws a scatter-plot that summarizes the findings of the Hidiroglou-Berthelot outliers' or the ratioSize detection method.*

Description

The function gets the output of the function of HBmethod or ratioSize when they are ran with the argument return.dataframe = TRUE) to draw a scatter-plot of ratios vs. the corresponding importance measures.

Usage

```
plot4ratios(out)
```

Arguments

out	Is the output of HBmethod or ratioSize when they are launched with the argument return.dataframe = TRUE.

Details

This function draws a scatter-plot. With the output of HBmethod the ratios (=yt2/yt1) are on Y axis while their importance measure (max(yt1, yt2)^U) are represented on the X axis. With the output of ratioSize on the Y axis the centered ratios are reported. In addition the acceptance bounds are drawn (blue lines); the dots (in red color) outside the bounds are the outliers. This is considered a useful diagnostic plot to understand how the procedure identifies the outliers.

Value

A scatter-plot is drawn and, in addition, the output includes a list with the needed data, i.e. the ratios (or centered ratios) (y), the importance measure (x) and the lower and upper bounds, respectively (y.low and y.up)

Author(s)

Marcello D'Orazio <mdo.statmatch@gmail.com>

References

Hidiroglou, M.A. and Berthelot, J.-M. (1986) 'Statistical editing and Imputation for Periodic Business Surveys'. *Survey Methodology*, Vol 12, pp. 73-83.

Hidiroglou, M.A. and Emond, N. (2018) 'Modifying the Hidiroglou-Berthelot (HB) method'. *Unpublished note*, Business Survey Methods Division, Statistics Canada, May 18 2018.

See Also

HBmethod, ratioSize

Examples

```
# generate some data
set.seed(222)
x0 <- rnorm(30, 50, 5)
set.seed(333)
rr <- runif(30, 0.9, 1.2)
rr[10] <- 2
x1 <- x0 * rr

# run HBmethod with argument return.dataframe = TRUE
out <- HBmethod(yt1 = x0, yt2 = x1,
                return.dataframe = TRUE)
# draw the scatterplot
plot4ratios(out)
```

ratioSize *Identifies outliers on ratios and filter them by a size measure*

Description

Identifies outliers on transformed ratios (centering with respect to their median) using the adjusted boxplot for skewed distributions. Outliers can be sorted/filtered according to a size measure.

Usage

```
ratioSize(numerator, denominator, id=NULL,
          size=NULL, U=1, size.th=NULL, return.dataframe=FALSE)
```

Arguments

`numerator` Numeric vector with the values that go at numerator of the ratio

`denominator` Numeric vector with the values that go at denominator of the ratio

`id` Optional numeric or character vector, with identifiers of units. If `id=NULL` units identifiers will be set equal to their positions in `x`.

`size` Optional numeric vector providing a measure of the importance of a ratio. If `size = NULL` the size measure is the maximum value between the numerator and the denominator of each ratio (makes sense if both the variables are observed using the same unit of measure). Observations' importance is also controlled by the argument `U`.

`U` Numeric, constant with $0 < U \leq 1$ controlling importance of each unit, in practice the final size measure is derived as (size^U). Commonly used values are 0.4, 0.5 or 1 (default).

`size.th` Numeric, size threshold. Can be specified when a size measure is used. In such a case just outliers with a size greater than the threshold will be returned. Note that when argument `U` is not set equal to 1, then the final threshold will be `size.th^U`.

`return.dataframe` Logical, if `TRUE` the output will save all the relevant information for outlier detection in a dataframe with the following columns: 'id' (units' identifiers), 'numerator', 'denominator', 'ratio' (= numerator/denominator), 'c.ratio' (centered ratios, see Details), 'sizeU' (size^U values) and finally 'outliers', where value 1 indicates observations detected as an outlier and 0 otherwise.

Details

This function searches for outliers starting from ratios `r=numerator/denominator`. At first the ratios are centered around their median, as in Hidiroglou Berthelot (1986) procedure (see `HBmethod`), then the outlier identification is based on the adjusted boxplot for skewed distribution (Hubert and Vandervieren 2008) (see `adjboxStats`). The subset of outliers is sorted in decreasing order according the size measure. If a size threshold is provided then just outliers with (size^U) > (size.th^U) will be returned.

Value

A list whose components depend on the `return.dataframe` argument. When `return.dataframe = FALSE` just the following components are returned:

`median.r` the median of the ratios

`bounds` The bounds of the interval for centered ratios `excluded` The position or the identifiers of the units with values excluded by the computations because of 0s or NAs.

`outliers` The position or the identifiers of the units detected as outliers. Remember that when `size.th` is set, just outliers with (size^U) > (size.th^U) will be returned.

When `return.dataframe=TRUE` the latter two components are substituted with two dataframes:

`excluded` A dataframe with the subset of observations excluded

`data` A dataframe with the not excluded observations with the following columns: 'id' (units' identifiers), 'numerator', 'denominator', 'ratio' (= numerator/denominator), 'c.ratio' (centered ratios, see Details), 'sizeU' (size^U values) and finally 'outliers', where value 1 indicates observations detected as an outlier and 0 otherwise. The data frame will be sorted in decreasing manner according to size^U. Note that when a size threshold is provided then ONLY outliers with (size^U) > (size.th^U) will be returned.

Author(s)

Marcello D'Orazio <`mdo.statmatch@gmail.com`>

References

Hidiroglou, M.A. and Berthelot, J.-M. (1986) 'Statistical editing and Imputation for Periodic Business Surveys'. *Survey Methodology*, Vol 12, pp. 73-83.

Hubert, M., and Vandervieren, E. (2008) 'An Adjusted Boxplot for Skewed Distributions', *Computational Statistics and Data Analysis*, 52, pp. 5186-5201.

See Also

`HBmethod`, `plot4ratios`, `boxB`,`adjboxStats`

Examples

```
set.seed(444)
x1 <- rnorm(30, 50, 5)
set.seed(555)
rr <- runif(30, 0.9, 1.2)
rr[10] <- 2
x2 <- x1 * rr

out <- ratioSize(numerator = x2, denominator = x1)
out

out <- ratioSize(numerator = x2, denominator = x1,
                 return.dataframe = TRUE)
head(out$data)

out <- ratioSize(numerator = x2, denominator = x1,
                 size.th = 65, return.dataframe = TRUE)
```

```
head(out$data)
```

skew.misc *Calculates some skewness measures.*

Description

The function calculates some skewness measures for the input vector data.

Usage

```
skew.misc(x, weights=NULL)
```

Arguments

x	Input vector containing data for which skewness will be calculated.
weights	Optional vector with eventual non-negative weights associated to the units in x

Details

This function calculates Pearson's skewness coefficient, the MedCouple measure of skewness and the non-parametric Bowley's measure of symmetry. The Bowley's skewness measure uses quartiles:

$$b = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

It ranges between -1 and +1, where positive (negative) values denote right (left) skewness. A value equal to 0 indicates symmetry. A crude measure of skewness can be obtained with a monotonic increasing function of b:

$$g = \frac{1 + b}{1 - b}$$

It ranges from 0 to Inf, g=1 indicates symmetry.

A measure of skewness similar to the Bowley's one is achieved by replacing Q3 and Q1 with respectively P90 and P10 percentiles:

$$b_P = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{P_{90} - P_{10}}$$

Similarly

$$g_P = \frac{1 + b_P}{1 - b_P}$$

For major details see Kotz at al. (2006, vol. 12, pp. 7771-7772).

The medCouple measure of skewness, M, ranges from -1 to +1 and is equal to 0 in case of symmetry, while $M > 0$ indicates positive skewness. For major details see mc.

Note that eventual weights, passed through the argument weights, are used ONLY in the calculation of the Bowley's type measures.

Value

A vector with the estimated measures of skewness.

Author(s)

Marcello D'Orazio <mdo.statmatch@gmail.com>

References

Kotz S. et al. (2006) *Encyclopedia of Statistical Sciences, Volume 12*. John Wiley and Sons.

See Also

mc

Examples

```
set.seed(112233)
y <- rnorm(n = 30, mean = 50, sd = 10)
y[20] <- 100

skew.misc(x = y, weights=NULL)

# use weights
ww <- runif(n = 30, min = 1, max = 10)
skew.misc(x = y, weights=ww)
```

Index