Package 'hotspots'

May 31, 2018

Type Package

Title Hot Spots

Version 1.0.3

Date 2018-05-18

Author Anthony Darrouzet-Nardi

Maintainer Anthony Darrouzet-Nardi <anthony@darrouzet-nardi.net>

Description The hotspots package is designed to look within a set of measured values of a variable and identify values that are disproportionately high based on both the deviance of any given value from a statistical distribution and its similarity to other values. Because this relative magnitude of each value is taken into account, a value that is a statistical outlier may not always be a hot spot if other values are similarly large.

Depends lattice, ineq

License GPL-2

LazyLoad yes

NeedsCompilation no

Repository CRAN

Date/Publication 2018-05-30 22:28:52 UTC

R topics documented:

disprop *Disproportionality*

Description

Calculates the magnitude of disproportionality for values within a dataset.

Usage

```
disprop(z)
```

Arguments

z	"hotspots" object

Details

Calculates the magnitude of disproportionality for each value within the data by dividing the difference between each value and the median by the difference between the hot spot cutoff, (Ch, as calculated by the function hotspots), and the median:

```
disproportionality = (x - med(x)) / (Ch - med(x))
```

Using this equation, all hot spots have a magnitude of disproportionality of > 1. Increasingly skewed distributions (for example, lognormal distributions with higher standard deviation) will have higher magnitudes of disproportionality for some of their values.

Value

A list containing the objects positive, negative, or both, depending on the which tails were calculated in the hotspots object. These objects are numeric vectors of the magnitudes of disproportionality. NA values are preserved.

Author(s)

Anthony Darrouzet-Nardi

See Also

hotspots

Examples

```
rln30 <- sort(c(rlnorm(15),rlnorm(15)*-1,NA), na.last = TRUE)
rln30
disprop(hotspots(rln30, tail = "both"))

#higher levels of disproportionality
rln30sd2 <- sort(c(rlnorm(15,sd = 3),rlnorm(15,sd = 3)*-1,NA), na.last = TRUE)
rln30sd2
disprop(hotspots(rln30sd2, tail = "both"))
```

hotspots *Hot spots and outliers*

Description

Calculates a hot spot or outlier cutoff for a statistical population based on deviance from the normal or t distribution. In the case of the hot spot cutoff, the relative magnitude of the values is also taken into account to determine if values are disproportionately large relative to other values. Thus, a value that is a statistical outlier may not always be a hot spot if other values are similarly large.

Usage

```
hotspots(x, p = 0.99, tail = "positive", distribution = "t", var.est = "mad")

outliers(x, p = 0.99, tail = "positive", distribution = "t", var.est = "mad",
  center.est = "mean")
```

Arguments

x	a numeric vector
p	probability level of chosen distribution used for calculation of cutoff (between 0 and 1)
tail	determines whether cutoffs are calculated for positive numbers within x, negative numbers, or both. Defaults to "positive" but can also be "negative" or "both".
distribution	statistical distribution used to calculate the hot spot or outlier cutoff. Defaults to "t" but can also be "normal". Other distributions could be implemented through simple modifications to the source code.
var.est	character vector indicating the function to be used to estimate the level of variation within the data. Defaults to the robust measure "mad". Non-robust measures such as "sd" may also be used, but result in greater variation in cutoff location.
center.est	character vector indicating the function to be used to center the data for identification of outliers. Defaults to "mean".

Details

This function first scales the data by dividing them by a robust version of the root mean square. The robust root mean square (rrms) is calculated as:

```
rrms = sqrt(med(x)^2 + var.est(x)^2)
```

where var.est is the user-specified function for estimating the level of variation within the data. This scaling of the data allows for the comparison of scaled values with a statistical distribution, which in turn allows discrimination between outliers that do not substantially influence the data from those that do. For the outlier function, the data are scaled after centering the data using the user-specified center.est function, which defaults to the mean. The hotspot or outlier cutoff (for positive values, negative values, or both) is then calculated as:

`cutoff = (med(x/rrms) + F^-1(p))*rrms`

where `F` is a cumulative distribution function for the t or normal distribution (its inverse `F^-1` being a quantile function; e.g., `qt`), and `p` is a user-defined parameter indicating the probability of `F^-1` beyond which we wish to define the cutoff.

Value

Returns an object of class "`hotspots`". The functions `summary` and `plot`, can be used to examine the properties of the cutoff. The function `disprop` can be used to calculate the level of disproportionality for each value in the data. An object of class "`hotspots`" is a list containing some or all of the following components:

`x`	numeric input vector
`data`	vector with missing values (`NA`) removed
`distribution`	statistical distribution used to calculate the hot spot or outlier cutoff.
`var.est`	function used to estimate the level of variation within the data
`p`	probability level of chosen distribution used for calculation of cutoff
`tail`	tail(s) of data for which cutoffs were calculated
`dataset_name`	character vector with name of input data
`rrms`	robust root mean square
`positive.cut`	calculated hot spot or outlier cutoff for positive values
`negative.cut`	calculated hot spot or outlier cutoff for negative values
`center.est`	function to be used to center the data for identification of outliers (only for `outliers` function

Author(s)

Anthony Darrouzet-Nardi

See Also

`summary.hotspots`, `plot.hotspots`, `disprop`

Examples

```
#basic operation on lognormal data
rln100 <- hotspots(rlnorm(100))
summary(rln100)
plot(rln100)

#greater skew in data
rln100sd2 <- hotspots(rlnorm(100,sd=2))
print(summary(rln100sd2),top = 5)
plot(rln100sd2)

#both tails on normally distributed data
n100 <- hotspots(rnorm(100), tail = "both")
```

```
summary(n100)
plot(n100)

#both tails on skewed data
rln100pn <- hotspots(c(rlnorm(50),rlnorm(50)*-1),tail = "both")
summary(rln100pn)
plot(rln100pn)

#importance of disproportionality on normally distributed data
#contrast with n100
n100p3 <- hotspots(n100$x+3, tail = "both")
summary(n100p3)
plot(n100p3)

#importance of disproportionality on skewed data
#contrast with rln100
rln100p10 <- hotspots(rlnorm(100)+10)
summary(rln100p10)
plot(rln100p10)

#outliers function ignores disproportionality
rln100p10o <- outliers(rlnorm(100)+10)
summary(rln100p10o)
plot(rln100p10o)

#some alternative parameters
rln100a <- hotspots(rlnorm(100), p = 0.9, distribution = "normal",  var.est = "sd")
summary(rln100a)
plot(rln100a)
```

Lchs *Lorenz curve with hot spot cutoff*

Description

Plot a Lorenz curve with a hot spot cutoff on it.

Usage

```
Lchs(x, ...)
```

Arguments

x	"hotspots" object
...	further plotting parameters to pass to plot.Lc

Details

Uses the function plot.Lc from the ineq package to plot a Lorenz curve based on the data in a hotspots object. The location of the hot spot cutoff on the Lorenz curve is then drawn as a filled black circle.

Author(s)

Anthony Darrouzet-Nardi

See Also

hotspots, Lc, plot.Lc

Examples

```
Lchs(hotspots(rlnorm(100)))
```

---

plot.hotspots *Plotting hot spot and outlier cutoffs*

Description

plot method for class "hotspots".

Usage

```
## S3 method for "hotspots" objects
## S3 method for class 'hotspots'
plot(x, pch = par("pch"), ...)
```

Arguments

x	"hotspots" object
pch	plotting character. See par
...	further plotting parameters to pass to densityplot

Details

Uses the function densityplot from the lattice package to show the distribution of the data and the position of the positive and/or negative hot spot or outlier cutoffs.

Value

An object of class "trellis".

Author(s)

Anthony Darrouzet-Nardi

See Also

hotspots, summary.hotspots, densityplot

Examples

```
#both tails on skewed data
rln100pn <- hotspots(c(rlnorm(50),rlnorm(50)*-1),tail = "both")
plot(rln100pn)

#modify graphical parameters
plot(rln100pn, pch = 16, cex = 1.5)
```

---

summary.hotspots *Summarizing hot spot and outlier cutoffs*

Description

summary method for class "hotspots".

Usage

```
## S3 method for "hotspots" objects
## S3 method for class 'hotspots'
summary(object, ...)

## S3 method for "summary.hotspots" objects
## S3 method for class 'summary.hotspots'
print(x, digits = max(3, getOption("digits") - 3), p_round = 1, top = 0, ...)
```

Arguments

object	"hotspots" object
x	"summary.hotspots" object
digits	the number of significant digits to use when printing
p_round	the number of decimal places to print for percentages when printing
top	the number of the most disproportionate (highest or lowest) data values to print with their percent contributions to the total
...	further arguments passed to or from other methods

Details

The importance of hot spots within the data is evaluated by reporting the number of hot spots, the percentage of values that are hot spots, and the percent of the sum of values attributable to hot spots. The percent of the sum of values is likely only relevant if the data are either all positive or all negative. A warning is given if they are not.

Value

A summary.hotspots object is a list containing all of the objects in a `hotspots` object as well as the following:

`num_phs`	number of positive hot spots or outliers in data
`percent_phs`	percent of values identified as positive hot spots or outliers
`percent_phs_sum`	percent of the sum of the values attributable to positive hot spots or outliers
`num_nhs`	number of negative hot spots or outliers in data
`percent_nhs`	percent of values identified as negative hot spots or outliers
`percent_nhs_sum`	percent of the sum of the values attributable to negative hot spots or outliers
`m`	A list of summary statistics pertaining to the data (mean, median, min, max, scale (determined by the argument '`var.est` in the `hotspots` function), and coefficient of variation (scale/median)
)	
`disprop`	vector of levels of disproportionality as calculated by `disprop`

Author(s)

Anthony Darrouzet-Nardi

See Also

`hotspots`, `plot.hotspots`, `disprop`

Examples

```
rln100.sum <- summary(hotspots(rlnorm(101), tail = "both"))
rln100.sum
print(rln100.sum, top = 10, p_round = 0)
```

Index