runGESD.R usage

https://github.com/raunakms/GESD

last update: May 20, 2016

Generalized Extreme Studentized Deviate (GESD) test

The generalized (extreme Studentized deviate) ESD test (Rosner 1983) is used to detect one or more outliers in a univariate data set that follows an approximately normal distribution. The primary limitation of the Grubbs test and the Tietjen-Moore test is that the suspected number of outliers, k, must be specified exactly. If k is not specified correctly, this can distort the conclusions of these tests. On the other hand, the generalized ESD test (Rosner 1983) only requires that an upper bound for the suspected number of outliers be specified.

Given the upper bound, r, the generalized ESD test essentially performs r separate tests: a test for one outlier, a test for two outliers, and so on up to r outliers.

- The generalized ESD test is defined for the hypothesis:

  - H_0: There are no outliers in the data set
  - H_a: There are up to r outliers in the data set

Test Statistic: Compute $R_i = max_i(|x_i - x_{mean}|)/s$, with x_{mean} and s denoting the sample mean and sample standard deviation, respectively. Remove the observation that maximizes $|x_i - x_{mean}|$ and then recompute the above statistic with $n - 1$ observations. Repeat this process until r observations have been removed. This results in the 'r' test statistics $R_1, R_2, ..., R_r$.

Significance Level: alpha (α)

Critical Region: Corresponding to the 'r' test statistics, compute the following 'r' critical values

$$\lambda_i = [(n - i) * t(n - i - 1, p)]/[SQRT((n - i - 1 + t(n - i - 1, p)^2) * (n - i + 1))]$$

where $i = 1, 2, ..., r, t_p, v$ is the 100_p percentage point from the t distribution with v degrees of freedom.

$$p = 1 - \alpha/(2 * (n - i + 1))$$

The number of outliers is determined by finding the largest i such that $R_i > \lambda_i$ Simulation studies by Rosner indicate that this critical value approximation is very accurate for $n >= 25$ and reasonably accurate for $n >= 15$.

- Note that although the generalized ESD is essentially Grubbs test applied sequentially, there are a few important distinctions:

  - The generalized ESD test makes appropriate adjustments for the critical values based on the number of outliers being tested for that the sequential application of Grubbs test does not.
  - If there is significant masking, applying Grubbs test sequentially may stop too soon. The example below identifies three outliers at the 5 % level when using the generalized ESD test.
  - However, trying to use Grubbs test sequentially would stop at the first iteration and declare no outliers.

The generalized ESD test can be used to answer the following question: How many outliers does the data set contain?

Original Paper Citation:

B. Rosner (1983). Percentage Points for a Generalized ESD Many-Outlier Procedure. Technometrics 25(2), pp. 165-172. http://www.jstor.org/stable/1268549?seq=1

Decription:

Computes outliers for the given data using GESD statistics.

Usage

gesd(obs, alpha, value.zscore, r)

Arguments

- **obs** : a vector of observation
- **alpha** : significance Level
- **value.zscore** : if the observation value are already z-normalized. Takes values "YES" or "NO".
- **r** : upperbound to the number of observations to call as an outlier. If NA, computes GESD test statistic until values of half sample size have been removed from the sample.

Value

- **Total** : The first column gives the total number of outliers
- Rest of the columns are for each observation
- Value of the outlier indicates the outlier rank in the acending order where 0 inidicates not an outlier, 1 inidicate the highest ranked outlier (i.e. most extreme observation).

```
source("runGESD.R")

set.seed(1234)

# Create matrix with observation values
rnames <- paste("R",c(1:10),sep="")
cnames <- paste("C",c(1:20),sep="")
mat <- matrix(rexp(200), 10, dimnames=list(rnames,cnames))

# Get outliers
dat.output <- t(apply(mat, 1, function(x) gesd(x, alpha=0.1, value.zscore="NO", r=NA)))
```

GESD Output

```
dat.output
```

```
##      Total C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12 C13 C14 C15 C16 C17 C18
## R1       6  3  6  4  0  0  5  2  0  0   0   0   0   0   0   0   1   0   0
## R2       1  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   1   0   0
## R3       0  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0
## R4       1  0  1  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0
## R5       1  0  0  0  0  0  0  0  0  0   0   0   0   0   0   1   0   0   0
## R6       0  0  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0
## R7       2  0  0  0  1  0  2  0  0  0   0   0   0   0   0   0   0   0   0
## R8       3  0  0  0  0  0  0  0  0  3   0   0   0   0   0   0   1   0   0
## R9       1  0  0  0  0  0  0  0  0  0   0   0   0   0   0   1   0   0   0
## R10      5  0  0  0  2  0  1  3  0  0   0   0   5   0   0   4   0   0   0
##      C19 C20
## R1     0   0
## R2     0   0
## R3     0   0
## R4     0   0
## R5     0   0
## R6     0   0
## R7     0   0
## R8     0   2
## R9     0   0
## R10    0   0
```