

# OUTLIERS

## Função

*id.outliers()*

**Entrada de dados:** Será realizado por meio de uma tabela de classe `data.frame` no R. Essa tabela terá linhas correspondendo às observações medidas e as colunas corresponderão às variáveis. Somente as variáveis quantitativas (numéricas) poderão ser usadas nessa função.

**Argumentos:** a função terá 7 argumentos;

```
id.outliers (x, quant=, group= , id= , visual= , res= , csv= FALSE)
```

Onde:

- *x*: é o data frame, a tabela com todas as colunas e linhas;
- *quant*: um vetor contendo as posições das colunas que representem variáveis quantitativas/numéricas;
- *group*: argumento que indica apenas uma coluna do data frame que contém a informação dos subgrupos de interesses (pode ser espécies, áreas, experimentos, idades, sexo, localidades etc.). Com esse argumento indicado, a função transformará a tal coluna em um objeto da classe *factor* e os testes serão realizados para todos os níveis desse fator separadamente. Se não for definido, o teste será feito considerando todas as observações do data frame como pertencentes ao mesmo grupo;
- *id*: indica o algoritmo utilizado para a identificação dos outliers. Existem quatro opções, o *default* será **box**:
  1. **id="box"**: utilizará a função *boxplot()* para a identificação dos outliers;
  2. **id="z"**: utilizará o teste *modified Z-Score* (Iglewicz & Hoaglin, 1993);
  3. **id="ESD"**: utilizará o teste *generalized ESD test* (Rosner 1983) para a identificação do número de outliers, e não quais são eles. Para esse teste os dados devem estar distribuídos próximo a curva normal;
  4. **id="ALL"**: utilizará todos os três testes acima;
- *visual*: determina a forma de visualização dos outliers, em qual dos gráficos eles serão visualizados, o *default* será **box**:
  1. **visual="box"**: utilizará a função *boxplot()* para a construção do gráfico;
  2. **visual="pontos"**: utilizará um gráfico de pontos com a função *dotchart()*;
  3. **visual="biplot"**: um gráfico de dispersão com duas variáveis será construído. Uma das variáveis do gráfico será aquela com os outliers identificados e a outra variável será retirada da lista de variáveis quantitativas informada pelo argumento *quant*;
  4. **visual="ALL"**: criará todos os três gráficos acima;
- *res*: determina a resolução dos gráficos. Terá três opções, o *default* será **LOW**:
  1. **res="LOW"**: qualidade alta, 150 dpi;

2. **res="MED"**: qualidade média, 300 dpi;
3. **res="HIGH"** qualidade alta, 600 dpi;

- **csv**= indica se a função criará um *output* com os *outliers* no formato *.csv* para a utilização em outros programas. O *default* será **FALSE**, a função retornará apenas uma lista com os outliers.

**Objetos de Saída:** Ao término da função serão geradas no máximo 3 tipos de objetos: tabelas em formato *.csv*, gráficos salvos em formato *.jpeg* e um objeto tipo *list* contendo os *outliers*.

**Gráficos Gerados:** Os gráficos para todas as variáveis indicadas pelo argumento *quant* serão apresentados em uma única prancha, e uma prancha para cada tipo de gráfico, no fim, se o argumento *visual*="ALL" for escolhido serão gerados três pranchas.

**Passos da Função:** Para cada um dos grupos indicados pelo argumento *group* será criada uma pasta de diretório separada, onde serão salvos os objetos de saída. Essa pasta receberá o nome do grupo e dentro dela haverá um pasta para cada tipo de gráfico escolhido e para cada algoritmo utilizado se o argumento *csv*=*TRUE* for designado.

**Utilidade da Função:** A função poderá ser utilizada por qualquer pessoa que trabalhe com dados quantitativos e tenha o interesse de verificar a existência e identificar possível *outliers*.

### Referências:

IGLWICZ, Boris; HOAGLIN, David. **Volume 16: How to Detect and Handle Outliers**. IN: MYKYTKA, Edward F.(ed.), The ASQC Basic References in Quality Control: Statistical Techniques, 1993.

ROSNER, Bernard. Percentage Points for a Generalized ESD Many-Outlier Procedure. **Technometrics**, 25(2), pp. 165-172, 1983.

ZUUR, Alain F.; IENO, Elena N.; ELPHICK, Chris S. A protocol for data exploration to avoid common statistical problems. **Methods in Ecology and Evolution**, v. 1, n. 1, p. 3-14, 2010.

### Fórmulas para os Algoritmos:

[Generalized ESD test](#)

[Modified Z-Score test](#)

From:  
<http://ecor.ib.usp.br/> - **ecoR**

Permanent link:  
[http://ecor.ib.usp.br/doku.php?id=05\\_curso\\_antigo:r2016:alunos:trabalho\\_final:jdalapicolla:proposta\\_a\\_reformulada\\_-\\_analise\\_exploratoria\\_dos\\_dados](http://ecor.ib.usp.br/doku.php?id=05_curso_antigo:r2016:alunos:trabalho_final:jdalapicolla:proposta_a_reformulada_-_analise_exploratoria_dos_dados)

Last update: 2020/08/12 06:04