IDENTIFICAÇÃO DE OUTLIERS EM DADOS NÚMERICOS

DESCRIPTION

Função para a identificação de possíveis *outliers* em um *data frame* com dados numéricos (quantitativos). São gerados até três objetos de saídas: tabelas em formato *.csv* contento os possíveis *outliers*, gráficos em formato *.jpeg* para visualização e um objeto tipo *list* contendo os possíveis *outliers* para manipulação dentro do ambiente R.

USAGE

```
id.outliers(x=data, quant=6:30, group=0, id="box", NUMBER=10, visual="boxplot", res="LOW", csv=FALSE)
```

ARGUMENTS

`x:` o *data frame*, a tabela com todas as colunas e linhas;

`quant:` um vetor contendo as posições das colunas que representem as variáveis quantitativas/numéricas de `x`;

`group:` 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 dessa coluna 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* é "box":

`id="box":` utilizará a função *boxplot()* para a identificação dos *outliers*;

`id="z":` utilizará o teste *modified Z-Score* (Iglewicz & Hoaglin, 1993);

`id="ESD":` utilizará o *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 à curva normal;

`id="ALL":` ulitizará todos os três algoritmos citados acimas;

`NUMBER:` Argumento necessário só se o método `"ESD"` ou `"ALL"` forem selecionados. Indica o número máximo de *outliers* que as variáveis podem ter. O *default* é `10`, significa que o teste analisará a possibilidade de `"quant"` ter entre 1 e 10 *outliers*. Não há limite máximo, mas é recomendado testar no máximo a metade do número amostral. A função corrige o `NUMBER` para a metade do número amostral se número escolhido for muito grande.

`visual:` determina a o tipo de gráfico para a visualização dos possíveis *outliers*, o *default* é `"boxplot".` 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:

`visual="boxplot":` utilizará a função *boxplot()* para a construção do gráfico;

`visual="pontos"`: utilizará um gráfico de pontos com a função *dotchart();*

`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 retira da lista de variáveis quantitativas informada pelo argumento `"quant"`;

`visual="ALL":` criará todos os três gráficos acima;

`res:` determina a resolução dos gráficos. Terá três opções, o *default* é `LOW`:

`res="LOW"`: qualidade alta, 150 dpi;

`res="MED"`: qualidade média, 300 dpi;

`res="HIGH"` qualidade alta, 600 dpi;

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

DETAILS

Para cada `group` da análise será criada uma pasta no diretório atual da área de trabalho do R. Cada pasta poderá ter até 6 subpastas se o argumento **"ALL"** for escolhido para `visual` e `id`, 3 para cada método e 3 para cada tipo de gráfico. Se apenas um gráfico e um método for escolhido haverá apenas duas subpastas, uma para cada. Se não for identificado *outliers* em um dos métodos escolhidos, a subpasta referente a esse método não será criada.

Para a realização das análises, cada variável quantitativa deve apresentar cinco ou mais observações. Se esse número não for alcançado a função retorna uma mensagem no console informando que não foi possível realizar tal teste, para tal variável, pelo número de amostras ser insuficiente.

Para a realização do *generalized ESD test* (Rosner 1983), `id="ESD"`, o autor recomenda que haja mais de 25 observações para que o teste tenha poder máximo, mas que 15 amostras são suficientes para uma boa aproximação dos resultados. Entre cinco e 14 amostras o teste pode ser realizado, mas seus resultados devem ser analisados com cautela. Menos de cinco amostras o teste não pode ser feito.

Para a realização do *generalized ESD test* (Rosner 1983), `id="ESD"`, o autor recomenda que as observações tenham uma distribuição normal. É recomendado antes de usar o teste, verificar se as amostras se aproximam da distribuição normal.

VALUE

A função retorna uma lista com outras três listas, cada uma das sub-listas é referente a um método de análise do argumento `"id"`. Quando um dos métodos não for usado, ou não apresentar *outliers* a lista referente estará vazia.

`list1`: contém os resultados do método `"box"` e terá o número de elementos igual ao comprimento de `"quant"`, nomeados com o nome da variável. Quando uma variável não tiver *outliers* identificados, a posição na lista ficará vazia.

`list2`: contém os resultados do método `"z"` e terá o número de elementos igual ao comprimento de `"quant"`, nomeados com o nome da variável. Quando uma variável não tiver *outliers* identificados, a posição na lista ficará vazia.

`list3`: contém o número de elementos igual ao comprimento de `"quant"`, nomeados com o nome da variável. Dentro de cada elemento da lista, há uma tabela gerada pelo método `"ESD"` que terá os números de *outliers*, o valor estatítico e valor crítico do teste. Quando o valor estatístico supera o valor crítico significa que há *outliers*. A função retorna no console o número de possíveis *outliers*, a lista é apenas para conferência dos valores estatísticos e para o seu uso em publicações. Se uma variável não tiver *outliers* identificados, a posição na lista ficará vazia.

WARNING

Evite dar nome extensos às linhas, por exemplo, com mais de três dígitos. Não indique `rownames` (nomes de linhas) para o `x` (*data frame)* usado na função, ao invés, deixe o R usar a posição das linhas na tabela. A função funcionará com qualquer `rownames`, mas na hora da construção dos gráficos, nomes extensos podem causar muita poluição visual e atrapalhar a identificação visual dos *outliers*.

EXAMPLES

```
#lendo uma tabela (data frame) com os dados para análise, sem usar o argumento rownames.x=
read.table("dados.csv", header = T, sep=",", as.is = T)

#identificar outliers na tabela "x"; salvar o resultados no objeto "out"; as 25 colunas com dados quantitativos serão analisados, entre a 6ª e 30ª colunas; a coluna 4 informa os dados de subgrupos (espécies); todos os métodos serão usados, número máximo de 18 outliers no dados será testado para o método ESD; todos os tipos de gráficos serão produzidos como qualidade média, de 300 dpi; será gerado arquivos . csv;
out=id.outliers(x, 6:30, 4, "ALL", 18, "ALL", "MED", TRUE)

#identificar outliers na tabela "data"; salvar o resultados no objeto "out2"; 9 colunas com dados quantitativos serão analisados, entre a 9ª e 15ª colunas e a 2ª e 5ª; não há divisão de subgrupos e os outros argumentos estarão no default.
out2=id.outliers(data, c(2, 5, 9:15))
```

NOTE

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.

AUTHOR

DALAPICOLLA, J. 2016. **id.outliers**: função para identificação de outliers em dados numéricos. Disponível em: <http://ecologia.ib.usp.br/bie5782/doku.php?id=bie5782:01_curso_atual:alunos:trabalho_final:jdalapicolla:start>