BIE5782

Aula 4:
ANÁLISE EXPLORATÓRIA

32º mandamento

Use R!

OTEIRO

1. Definição e importância de AED
2. Conferência e correção dos dados
3. AED univariadas
4. AED bivariadas e relações entre variáveis
5. AED multivariadas: definição

Objetivos da AED (ou EDA)

- Controle de qualidade dos dados
- Sugerir hipóteses para os padrões observados
- Apoia a escolha dos procedimentos estatísticos de testes de hipótese
- Avaliar se os dados atendem às premissas dos procedimentos estatísticos escolhidos
- Indica novos estudos e hipóteses

John W. Tukey
(1915-2000)

Um alerta inicial

A análise exploratória não é “dragagem” de dados!

Assume-se que o pesquisador formulou *a priori* hipóteses biológicas plausíveis amparadas pela teoria ecológica.

Conheça seus dados!

Análise exploratória de dados

Pode levar entre 20 e 50% do tempo das análises.

Deve ser iniciada ainda durante a coleta de dados.

Utiliza-se largamente técnicas visuais (gráficos) nesta fase da pesquisa.

Quarteto de Anscombe

- Demonstrou a importância dos gráficos para conhecer a estrutura dos dados e relações entre variáveis

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Francis Anscombe (1918-2001)

Conferência dos Dados

`summary()` , `str()` ,
`head()`, `tail()`

Um protocolo de AED

Perguntas que devemos fazer:

1) Existem valores faltantes (NAs)? Eles são mesmo faltantes?

2) Existem muitos zeros?

Teste lógico para valores perdidos

`is.na()`

```
> a
 [1]  1  2  3  4  5 NA 6  7  8  9  10 NA

> is.na(a)
 [1] FALSE FALSE FALSE FALSE FALSE  TRUE
 [7] FALSE FALSE FALSE FALSE FALSE  TRUE

> a[!is.na(a)]
 [1]  1  2  3  4  5  6  7  8  9 10

> a[is.na(a)] <- 0

> a
 [1]  1  2  3  4  5  0  6  7  8  9 10  0
```

Teste lógico para presença de zeros

```
> b
 [1]  1  0  3  0  5 NA  6  0  8  0 10 NA

> b==0
 [1] FALSE  TRUE FALSE  TRUE FALSE    NA
 [7] FALSE  TRUE FALSE  TRUE FALSE    NA

> sum(b==0, na.rm=T)
 [1] 4
```

Um protocolo de AED

Perguntas que devemos fazer:

3) Onde os dados estão centrados? Como eles estão espalhados? São simétricos, enviesados, bi-modais?

4) Existem valores extremos (outliers)?

5) As variáveis têm distribuição normal?

Uma Variável

- Estatísticas descritivas
- Contagens de valores e tabelas
- Gráficos de distribuição
- Gráfico quantil-quantil

Medidas de Tendência Central

mean(), median()

```
> mean( c(0,1,2,3,4,5))
[1] 2.5

> median( c(0,1,2,3,4,5))
[1] 2.5

> mean( c(0,1,2,3,4,100))
[1] 18.33333

> median( c(0,1,2,3,4,100))
[1] 2.5
```

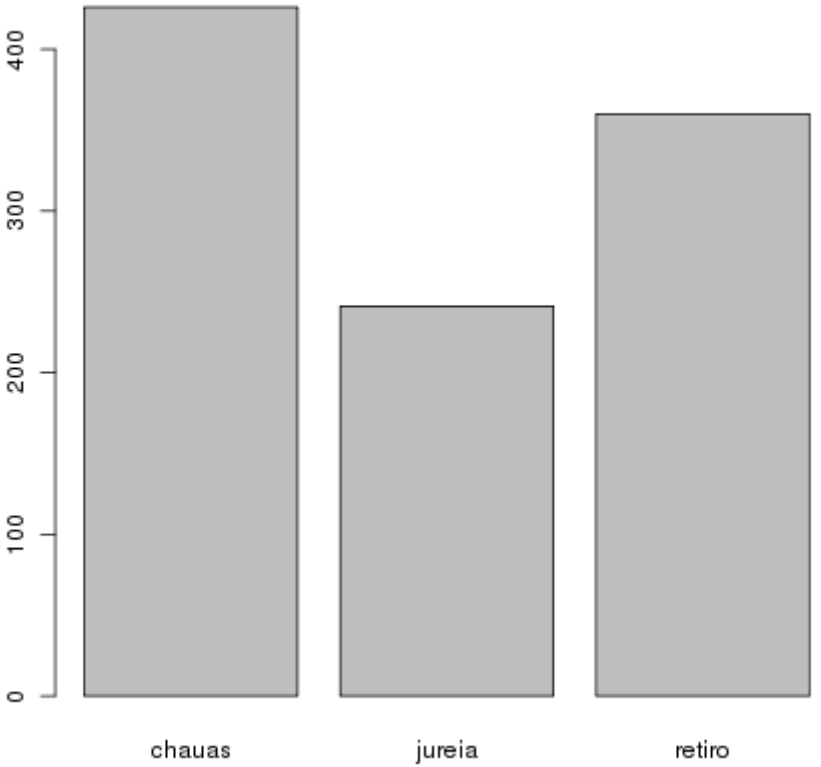
Média (normal e truncada) mediana, quantis: o pacote básico.

`mean(trim= ), mean() , median() , quantile()`

Vamos ao R!

Contagens de Fatores
`table()`, `barplot()`

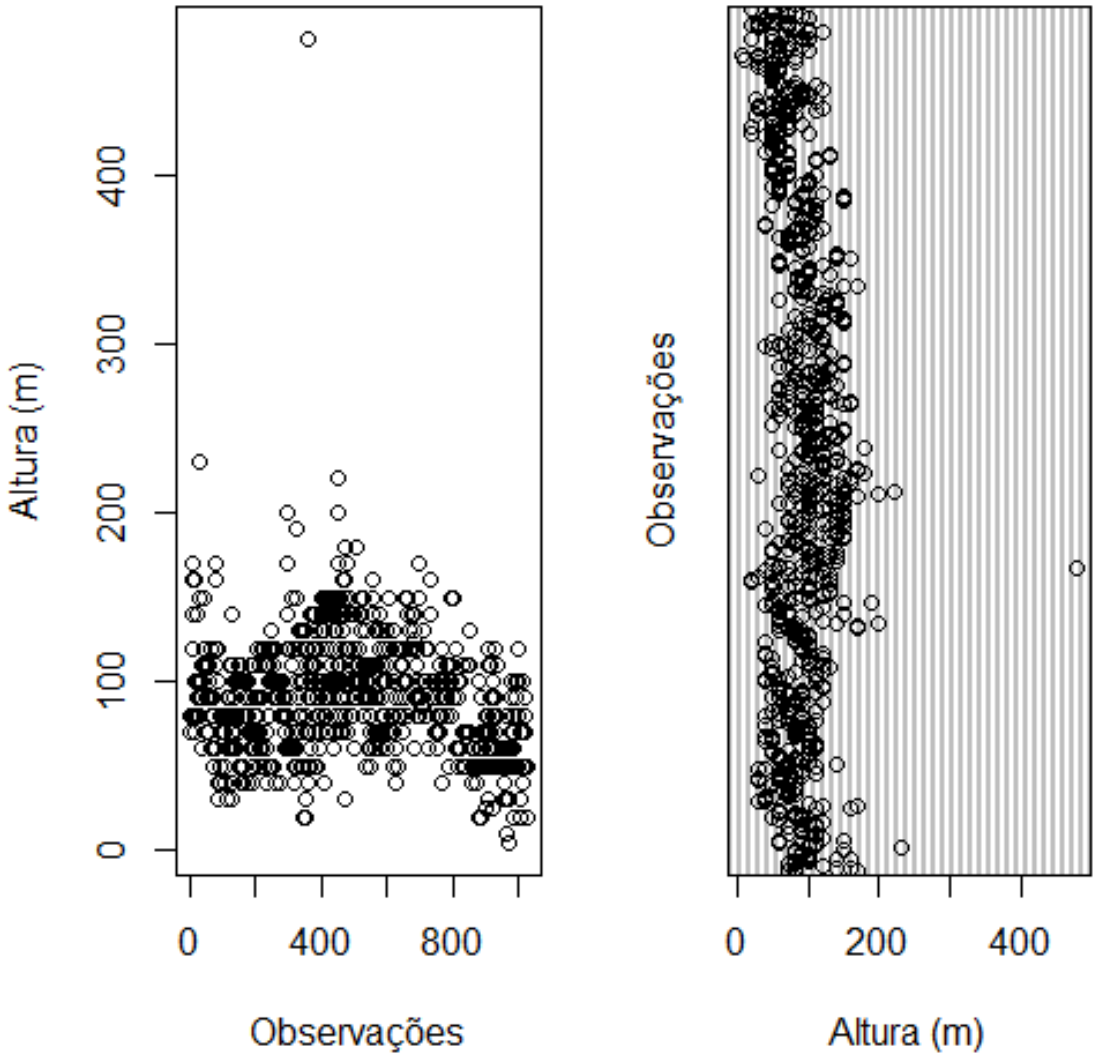
```
> table(caixeta$local)

chauas jureia retiro
   426     241     360
```




```
> barplot(table(caixeta$local))
```

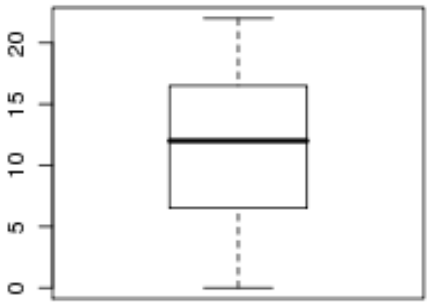
Gráficos univariados básicos

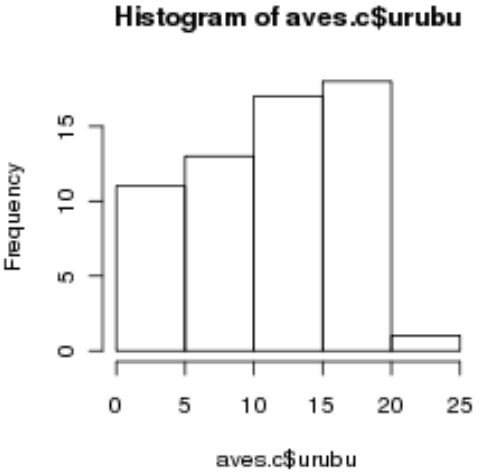
`plot()`, `dotchart()`

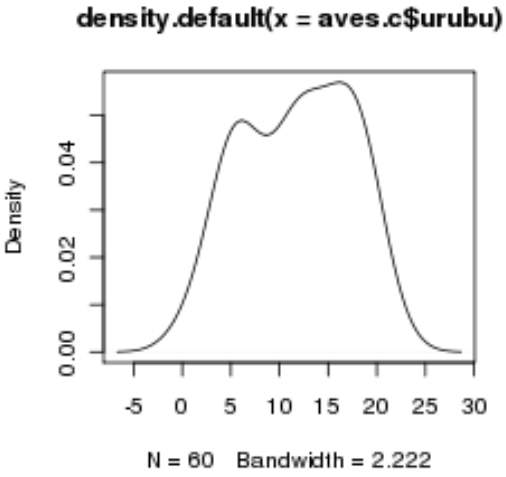


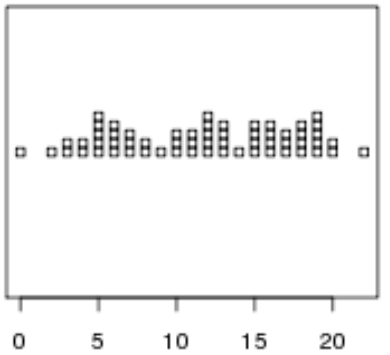
Gráficos univariados básicos

`boxplot()`, `hist()`, `density()`, `stripchart()`



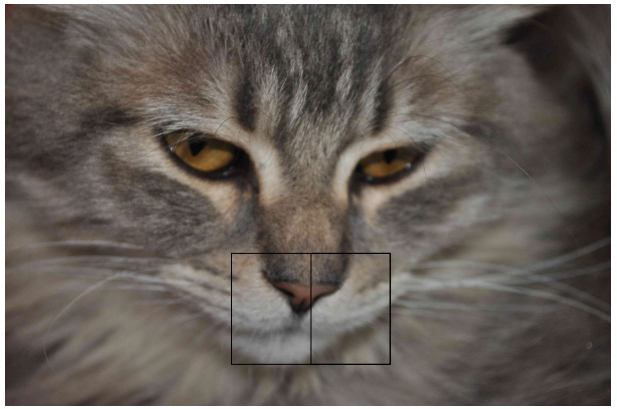


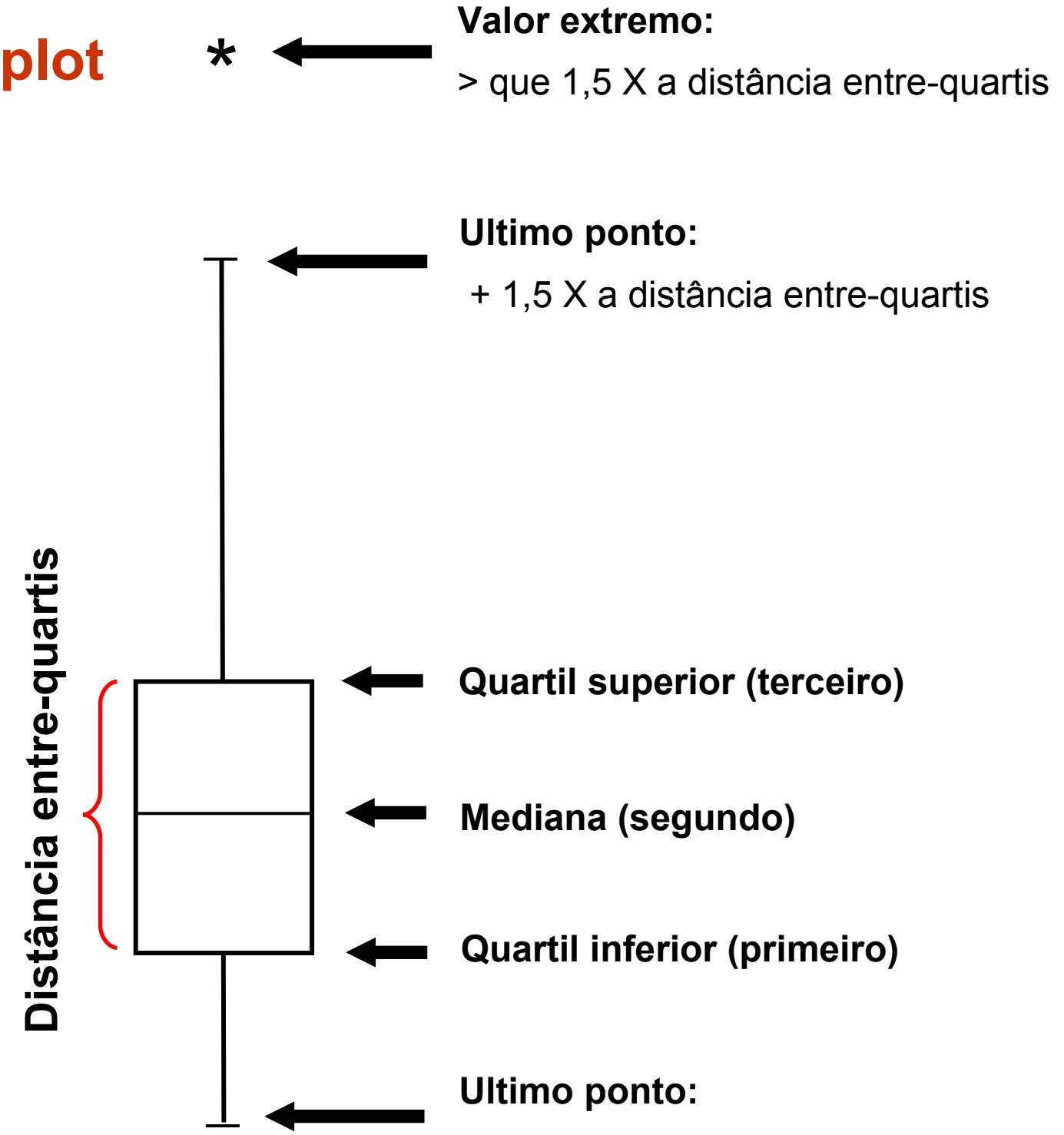




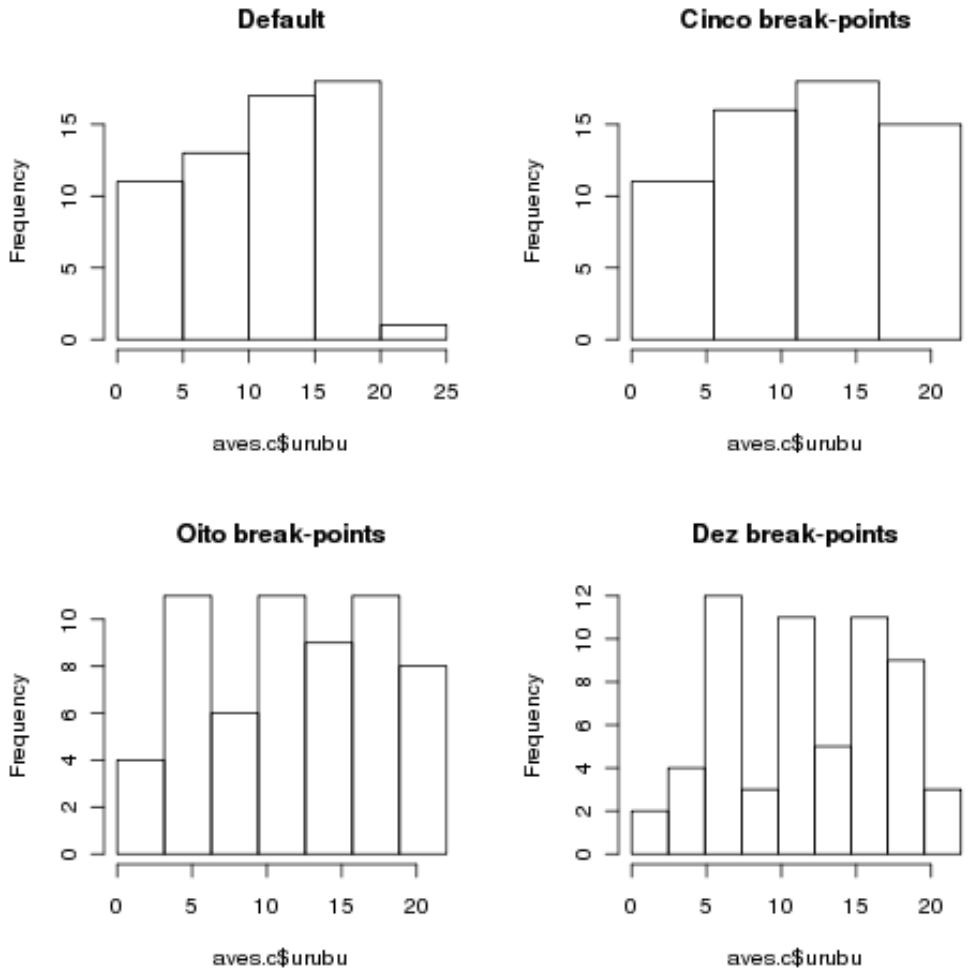
Box-and-whisker plot

ou box-plot





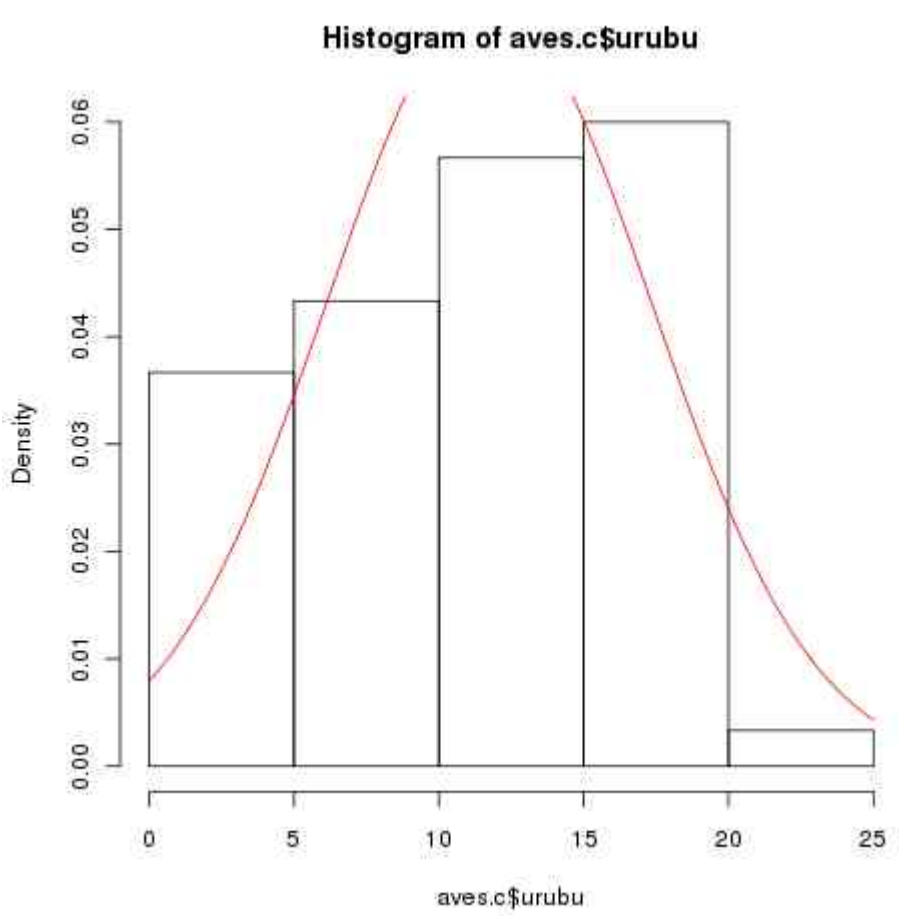
O problema do n de classes do histograma

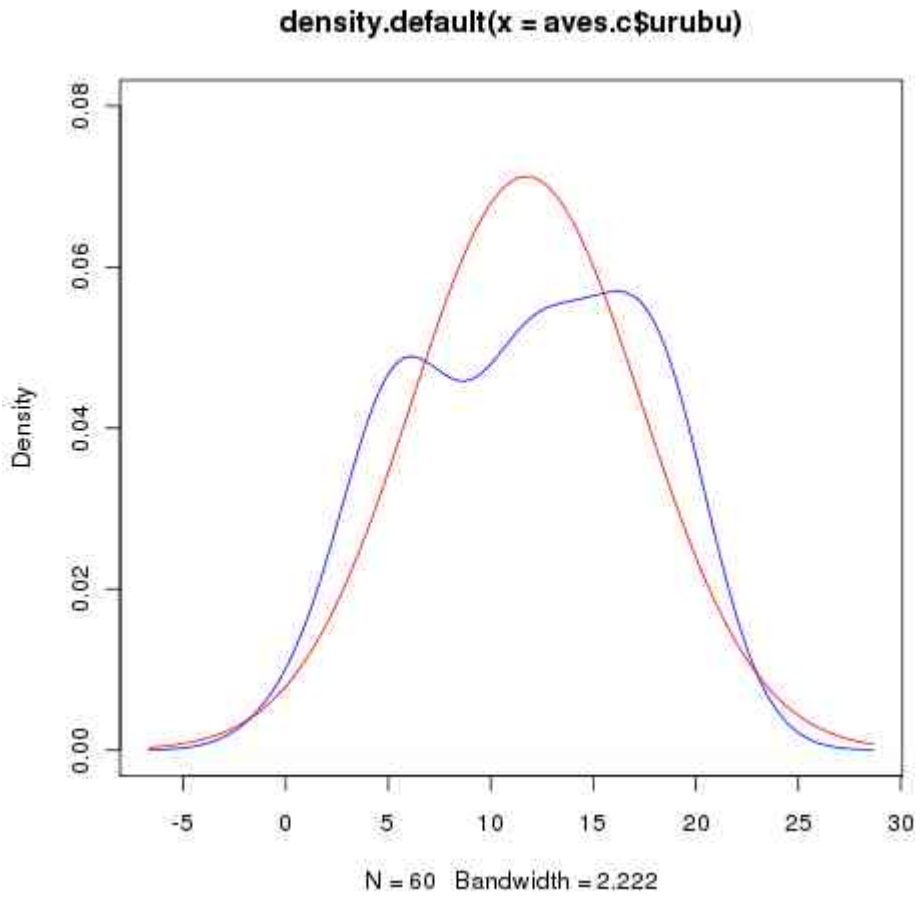



```
hist(aves.c$urubu)
hist(aves.c$urubu,breaks=seq(0,max(aves.c$urubu),length=5))
hist(aves.c$urubu,breaks=seq(0,max(aves.c$urubu),length=8))
hist(aves.c$urubu,breaks=seq(0,max(aves.c$urubu),length=10)
```

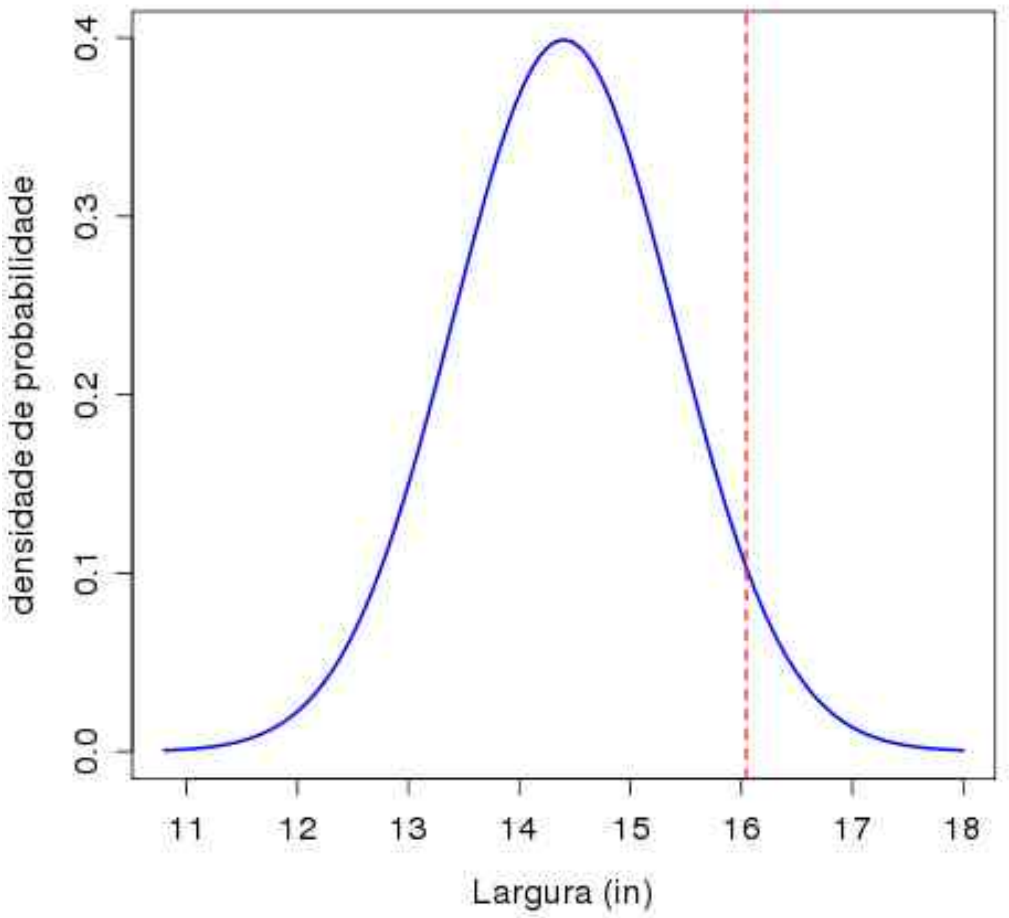
Curvas Empíricas e Teóricas de Densidade

`hist(prob=T)`, `density()`, `curve()`





Quantil teórico da distribuição normal




```
> qnorm(p=0.95, mean=14.4, sd=1)
[1] 16.04485
```

Gráfico Quantil-Quantil

```
      x percentil   q.norm
1 23.83      0.01 23.05859
2 24.07      0.02 23.86540
3 24.08      0.03 24.37730
4 24.09      0.04 24.76238
5 24.43      0.05 25.07561

...

95  35.03       0.95 34.81219
96  35.32       0.96 35.12542
97  35.35       0.97 35.51050
98  36.04       0.98 36.02240
99  36.35       0.99 36.82921
100 36.82       1.00      Inf
```

Gráfico Quantil-Quantil (Q-Q plot)

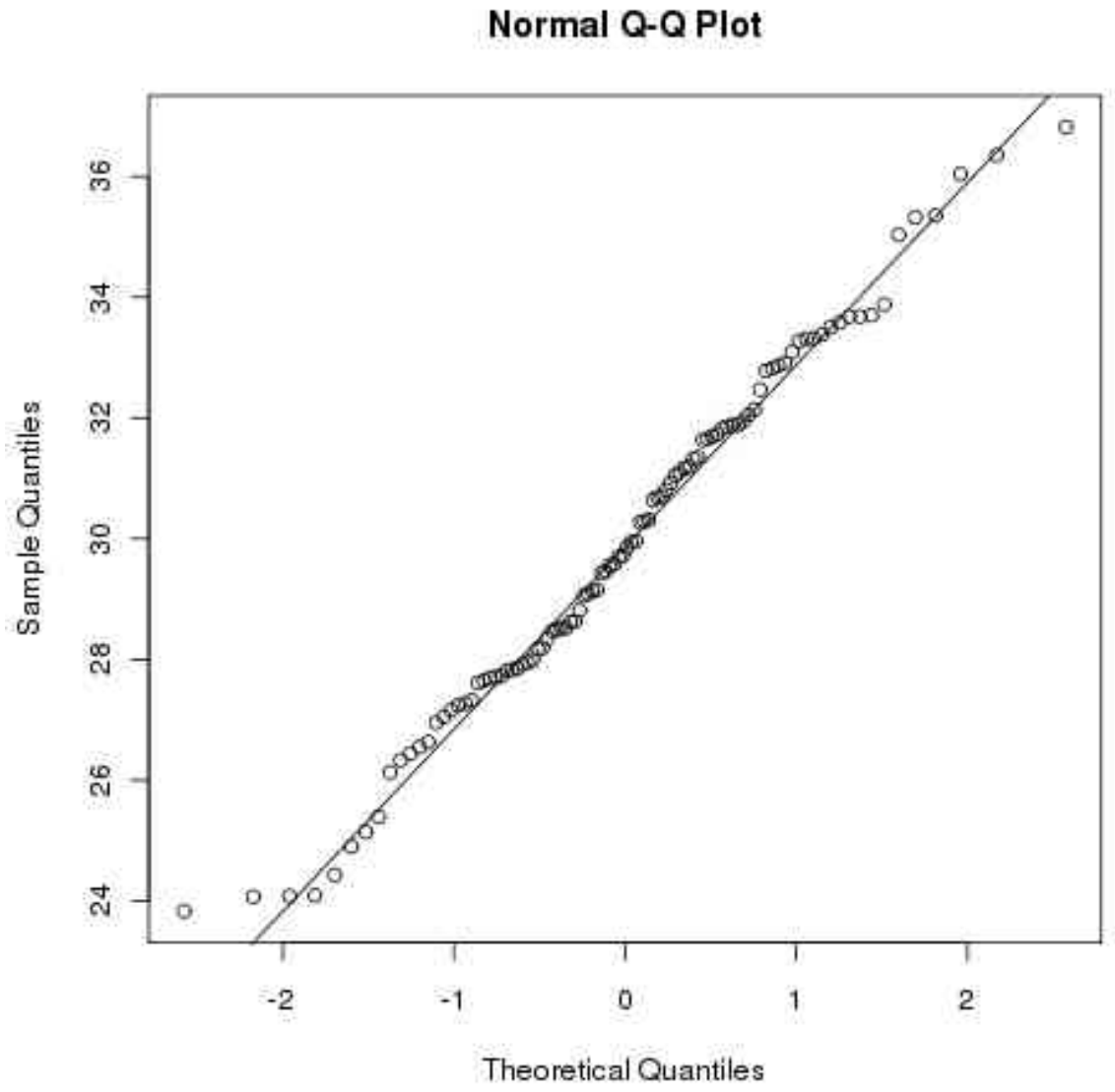
`qqnorm()`, `qqline()`

- Melhor teste de normalidade

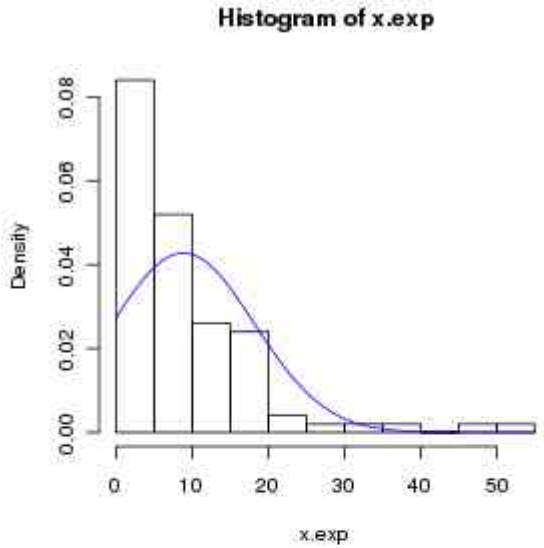
- Quantil empírico (dados) vs. Quantil teórico de acordo com alguma distribuição (como a normal)

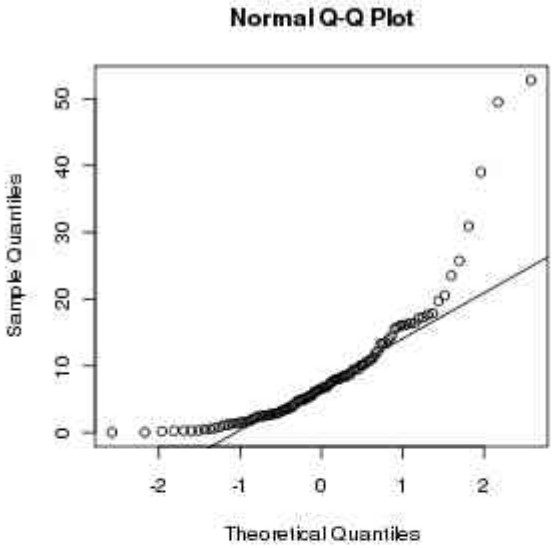
O melhor teste de normalidade

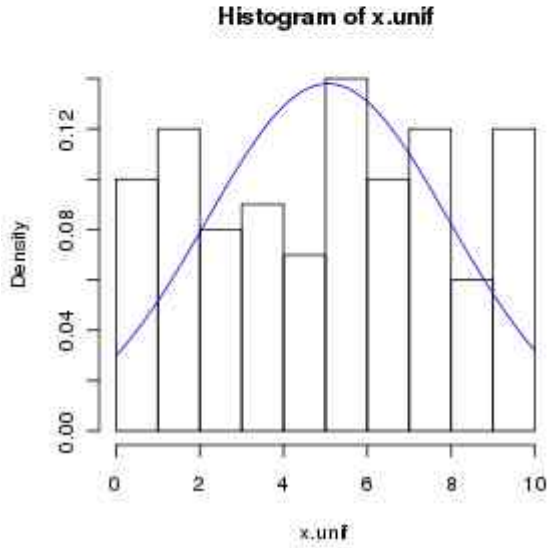
`qqnorm()`, `qqline()`

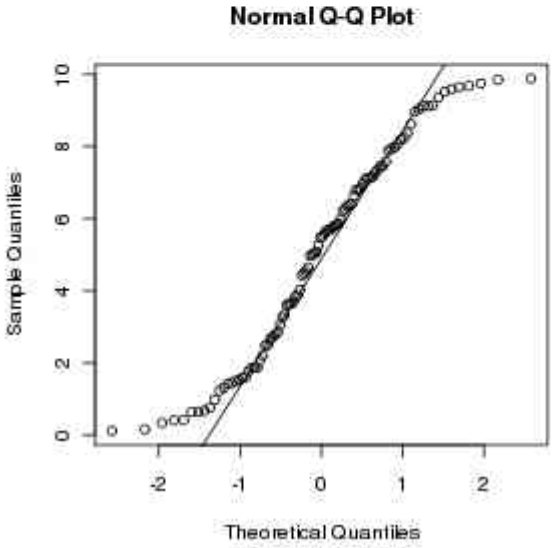


O melhor teste de normalidade









Vai para o R!

Um protocolo de AED

Perguntas que devemos fazer:

6) Existe alguma relação entre as variáveis?

7) A relação é linear?

DUAS VARIÁVEIS

- Fatores e contagens:
  - Tabelas de contingência
  - Tabelas de frequência
  - Estatísticas agregadas por fatores
- Gráficos
  - Dispersão
  - Linhas de tendência
  - Box-plot por classes
  - Gráficos condicionais

Tabelas de Contingência

table()

```
> table(caixeta$especie,caixeta$local)

                           chauas jureia retiro
  Alchornea triplinervia        0       3      12
  Andira fraxinifolia           0       4       0
  bombacaceae                   0       1       0
  Cabralea canjerana            0       4       0
  Callophyllum brasiliensis     7       0       0
  Calophyllum brasiliensis      0       4       0
  Cecropia sp                   0       0       1
  Coussapoa macrocarpa          0       3       0
  Coussapoa micropoda           2       0       7
  Cryptocaria moschata          0       2       0
  Cyathea sp                    0       0       2
```

Tabulação de Frequências

xtabs()

```
> head(Titanic.df)
  Class     Sex   Age Survived Freq
1   1st    Male Child       No    0
2   2nd    Male Child       No    0
3   3rd    Male Child       No   35
4  Crew    Male Child       No    0
5   1st Female Child       No    0
6   2nd Female Child       No    0
> xtabs(Freq~Sex+Survived, data=Titanic.df)

        Survived
Sex        No  Yes
  Male   1364  367
  Female  126  344
```

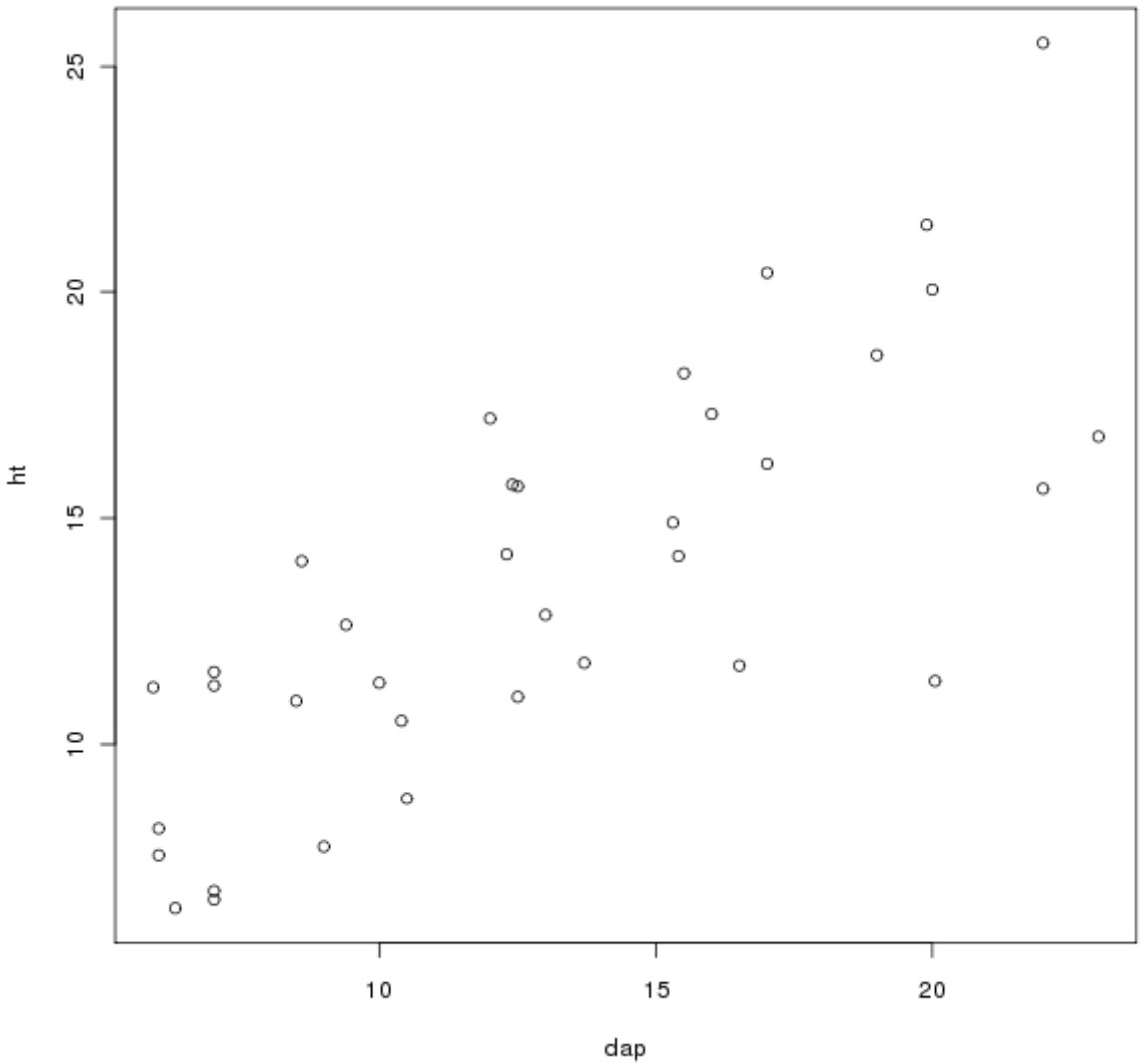
"Tabelas Dinâmicas"

aggregate()

```
> names(caixeta)
[1] "local"   "parcela" "arvore"  "fuste"   "cap"
[5] "h"       "especie" "ab"

> caixeta.alt <- aggregate(caixeta$h,
+ by=list(local=caixeta$local,
+ especie=caixeta$especie),FUN=max)

> head(caixeta.alt)
   local                    especie   x
1 jureia    Alchornea triplinervia 140
2 retiro    Alchornea triplinervia 100
3 jureia       Andira fraxinifolia  90
4 jureia                bombacaceae 150
5 jureia        Cabralea canjerana 150
6 chauas Callophyllum brasiliensis 200
```

Diagrama de dispersão (espalhagramas)

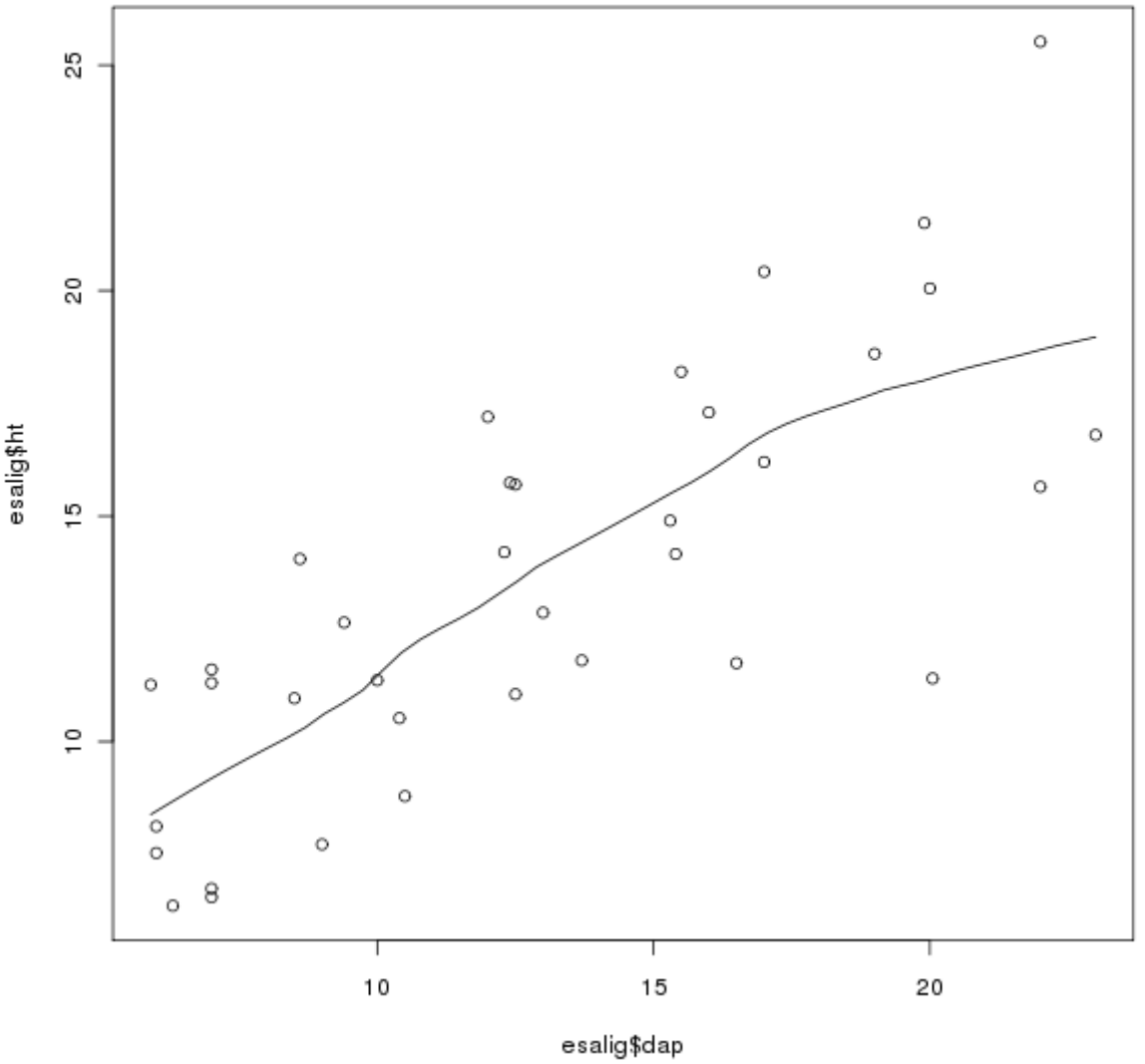
`plot(y~x)`




```
> plot(ht~dap, data=esalig)
```

Espalhagramas com Linha de Tendência

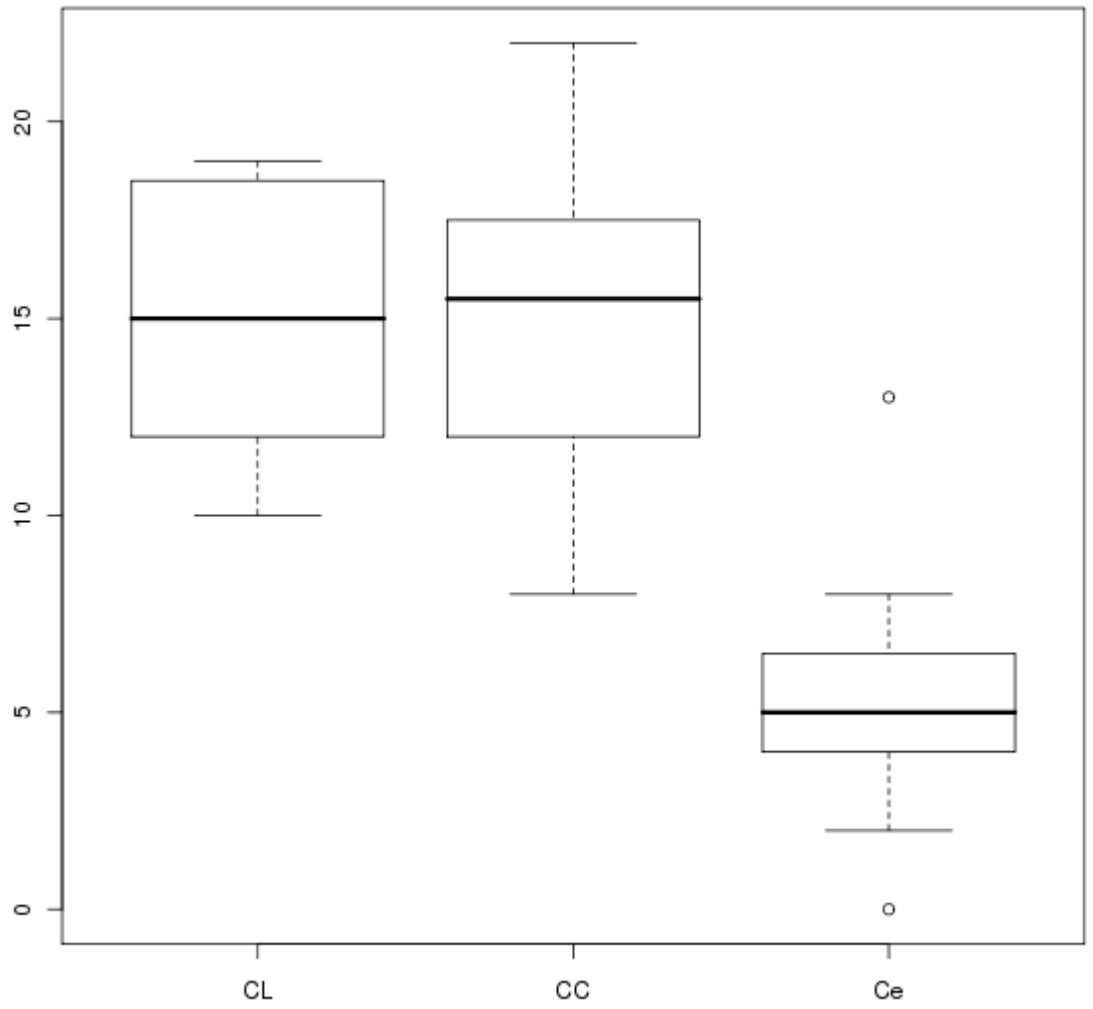
`scatter.smooth(y~x)`




```
> scatter.smooth(esalig$ht~esalig$dap, span=1/2)
```

Boxplot por Classes

`boxplot(y~x)`




```
> boxplot(urubu~fisionomia, data=aves.c)
```

Vai para o R!

MAIS DE DUAS VARIÁVEIS

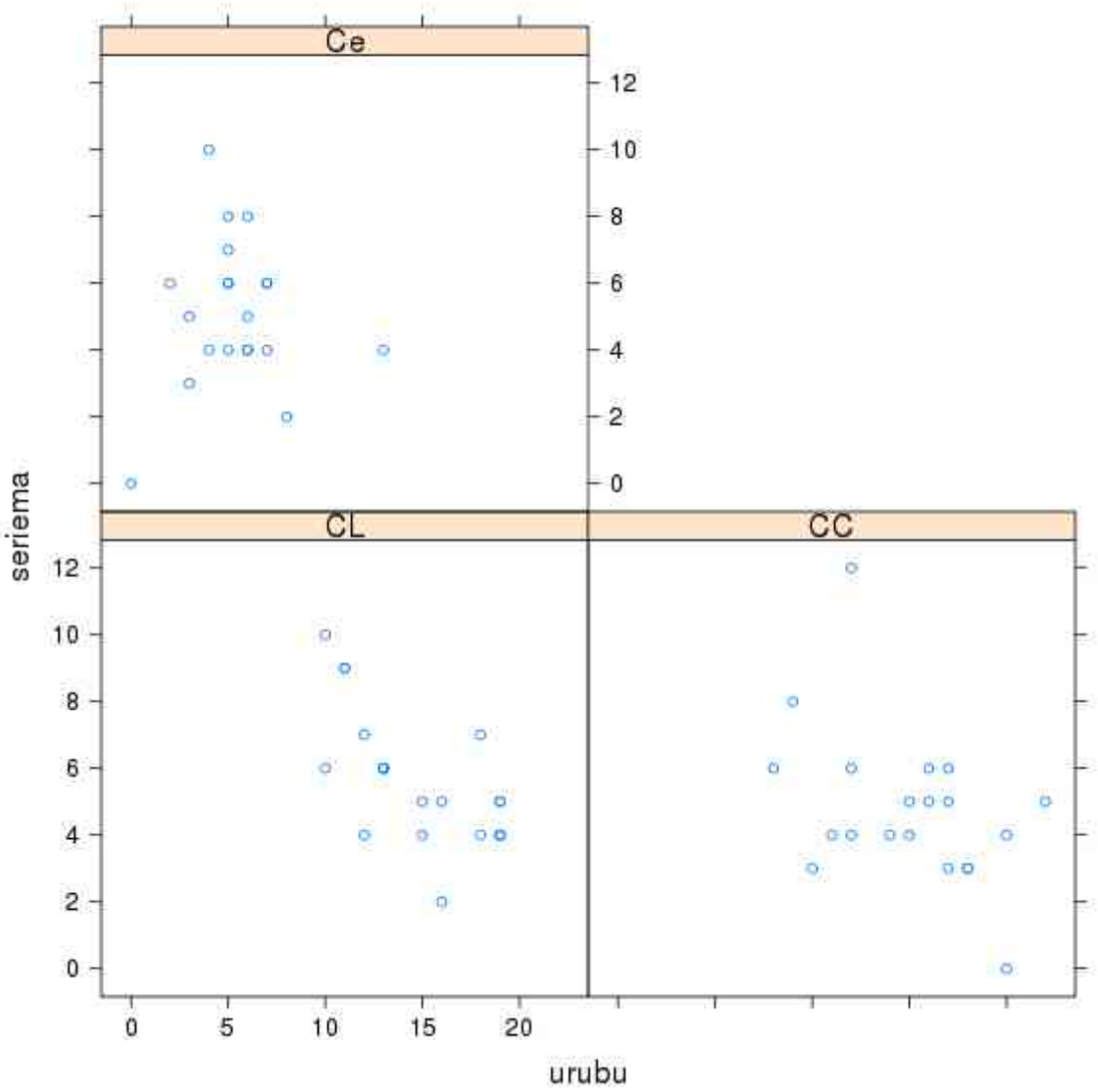
- Fatores e contagens:
  - Tabelas multidimensionais
  - Matrizes de correlação e distância
  - Estatísticas agregadas por fatores
- Gráficos
  - Gráficos condicionados
  - Matrizes de gráficos
  - Ordenação e classificação

Tabelas Multidimensionais

```
> xtabs(Freq~Class+Survived+Sex, data=Titanic.df)
, , Sex = Male

       Survived
Class    No Yes
  1st   118  62
  2nd   154  25
  3rd   422  88
  Crew  670 192

, , Sex = Female

       Survived
Class    No Yes
  1st     4 141
  2nd    13  93
  3rd   106  90
  Crew    3  20
```

Pacote lattice: gráficos condicionados

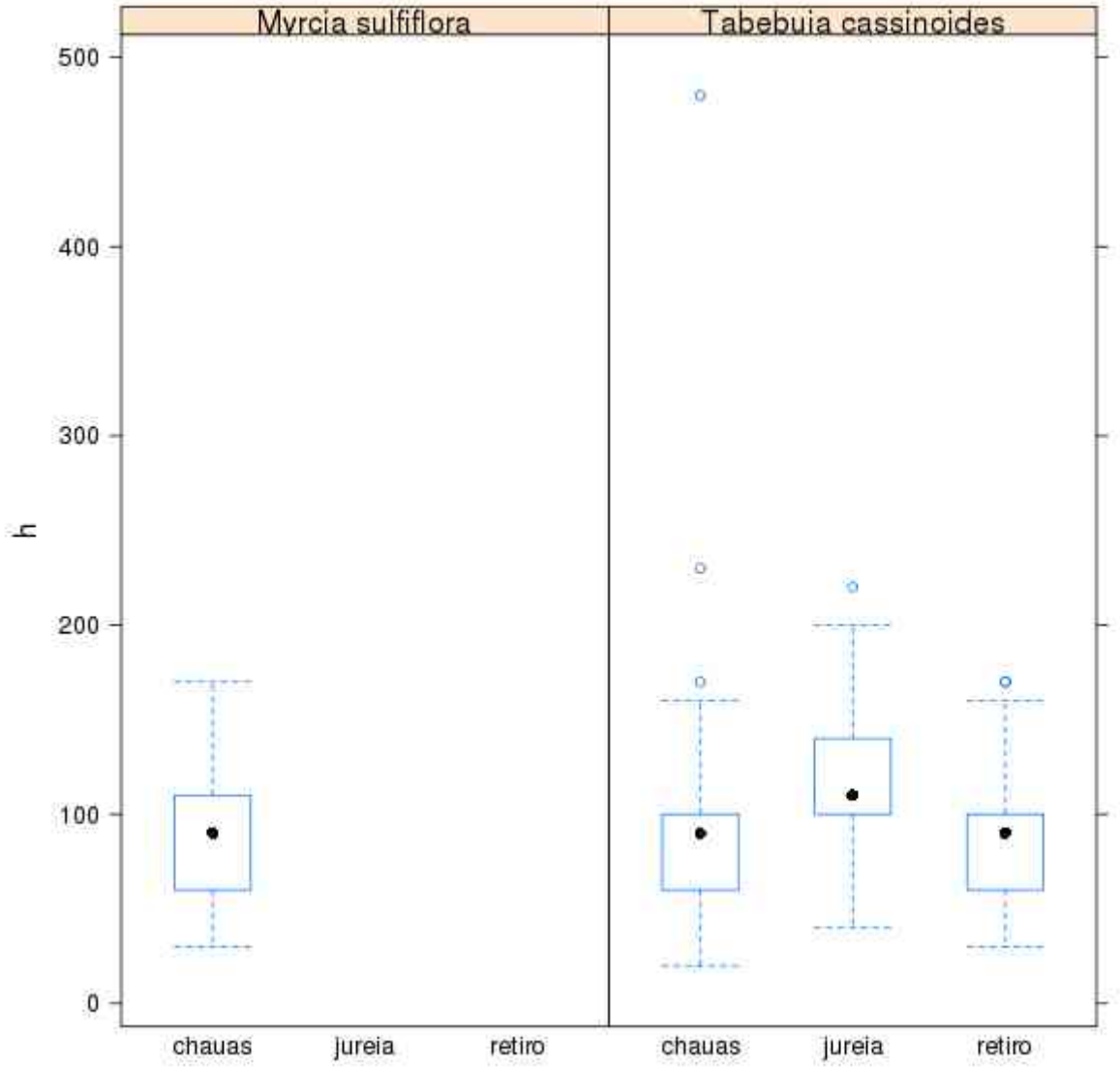
`xyplot(y~x|z)`




```
> xyplot(seriema~urubu|fisionomia, data= aves.c)
```

Box-plot no lattice

`bwplot(y~x|z)`




```
>  bwplot(h~local|especie, data=caixeta.abund)
```

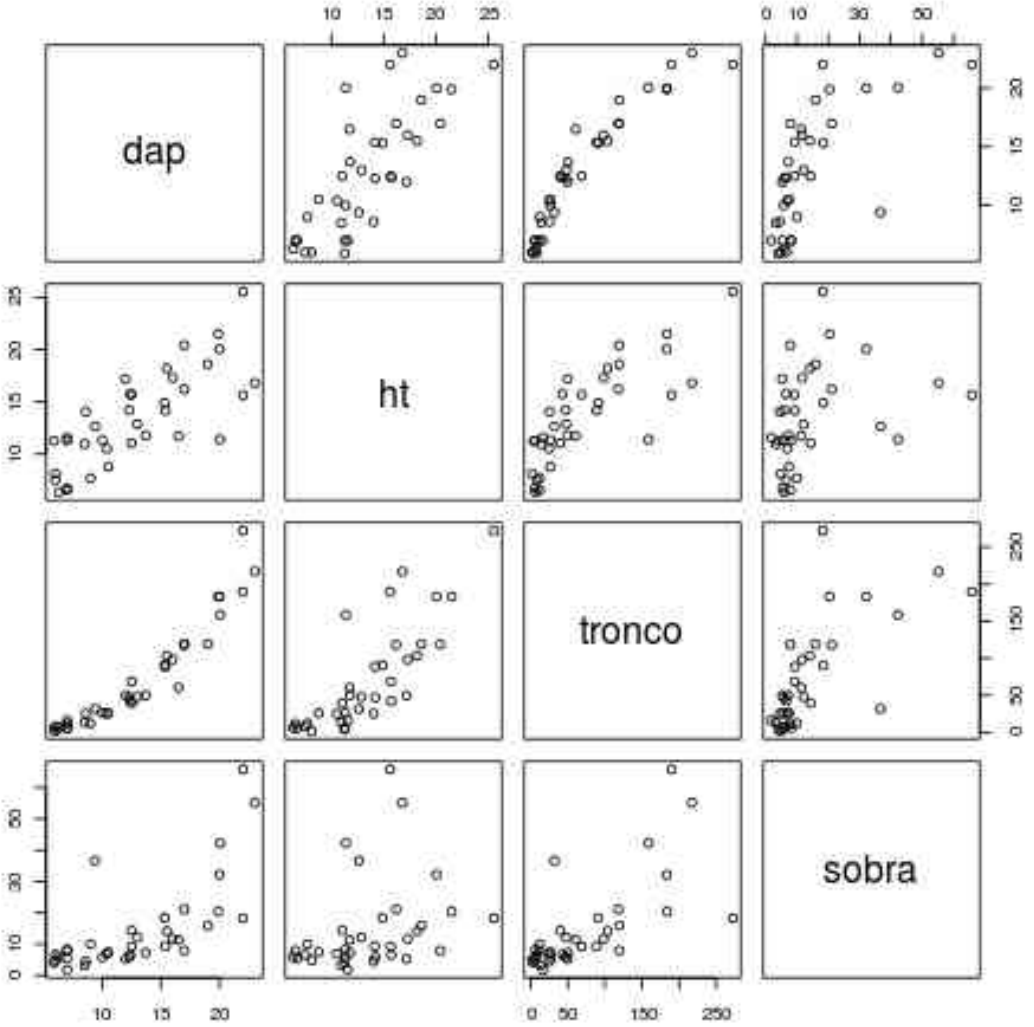
Matrizes de correlação

cor()

```
> cor(esaligna[,4:7])
              dap        ht    tronco     sobra
dap     1.0000000 0.7745167 0.9407805 0.6863613
ht      0.7745167 1.0000000 0.8054810 0.3204422
tronco 0.9407805 0.8054810 1.0000000 0.6933458
Sobra   0.6863613 0.3204422 0.6933458 1.0000000
```

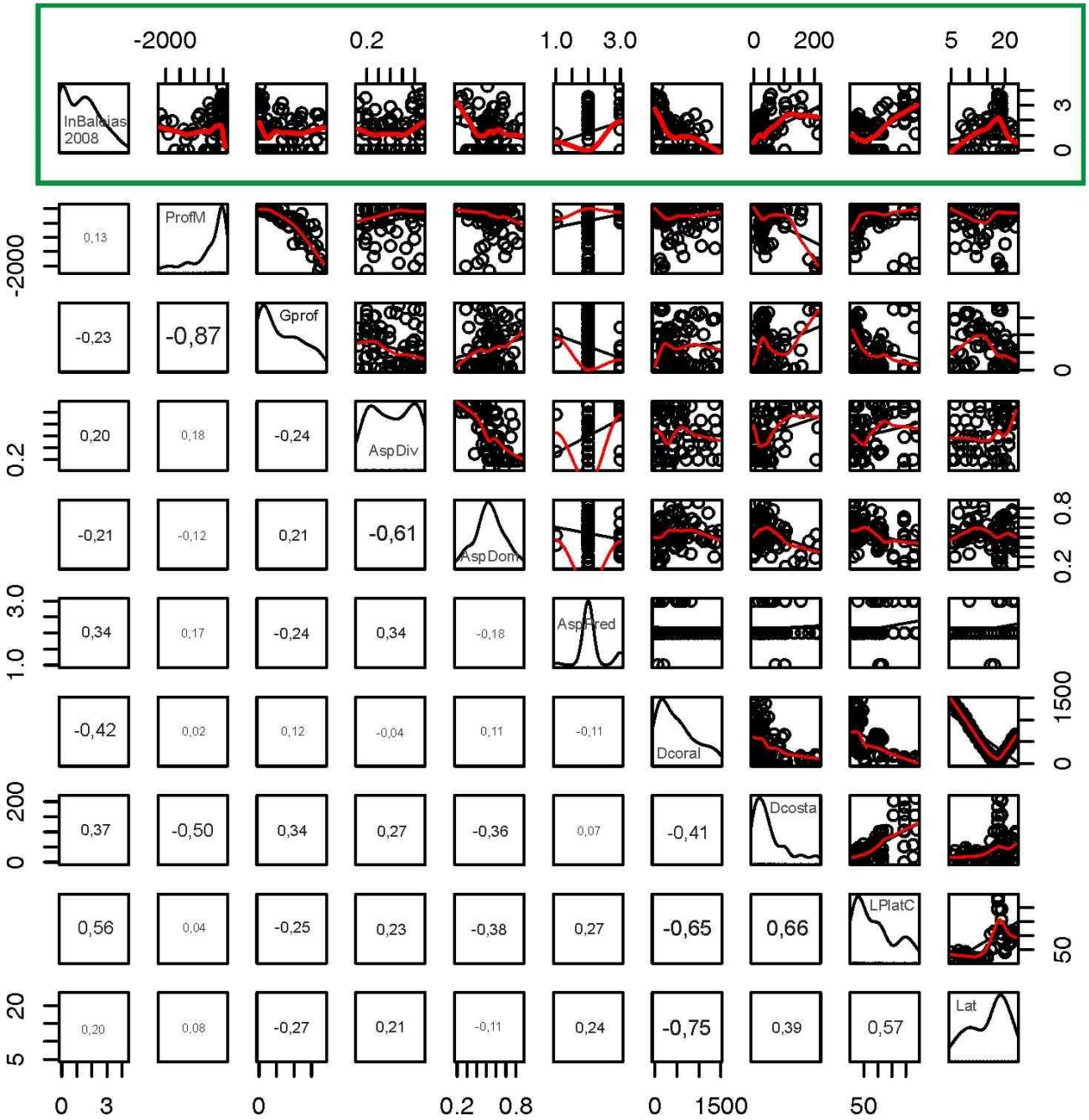
```
> cor(esaligna[,4:7], method="spearman")
              dap        ht    tronco     sobra
dap     1.0000000 0.7795958 0.9773287 0.7850061
ht      0.7795958 1.0000000 0.8512227 0.4857143
tronco 0.9773287 0.8512227 1.0000000 0.7534106
sobra   0.7850061 0.4857143 0.7534106 1.0000000
```

Matriz de diagramas de dispersão
`pairs()`




```
> pairs(esaligna[,4:7])
```

E ainda mais complexa...



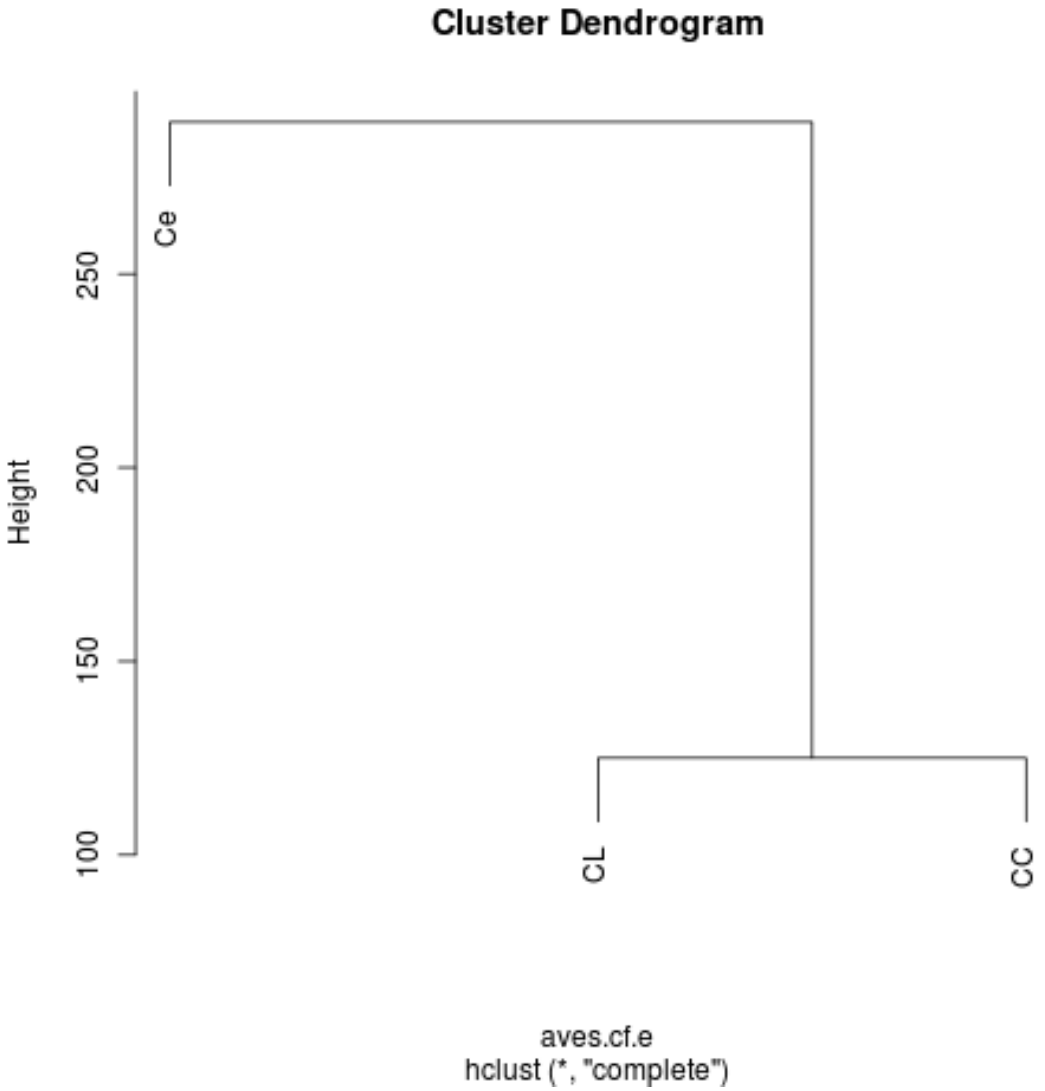
Matrizes de distância

dist()

```
> aves.cf
   fisio urubu carcara seriema
CL    CL   298      88     112
CC    CC   299     212      96
Ce    Ce   107     305     102

> aves.cf.e <- dist(aves.cf[,2:4])

> aves.cf.e
          CL       CC
CC 125.0320
Ce 289.2577 213.4221
```

Análise de aglomerados

`hclust()`

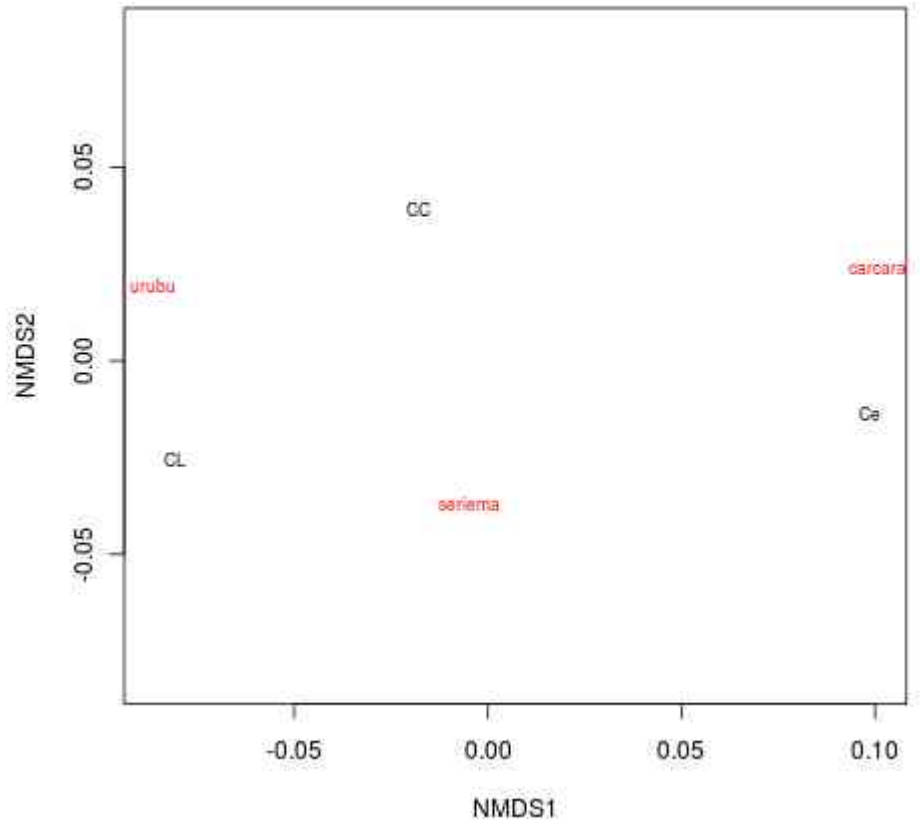


Esta é a função básica, ver pacotes vegan e ADE4 para análises multivariadas em Ecologia

```
>  plot(hclust(aves.cf.e))
```

Ordenação (um exemplo)

`metaMDS()`




```
> require(vegan)
> plot(metaMDS(aves.cf[,2:4]),type="t")
```

Vai para o R!

Resumo:

1) Conferência dos dados (NAs e erros de digitação)

2) Valores extremos e muitos zeros

3) Distribuição das variáveis (simetria, normal)

4) Relação entre variáveis (linearidade, colinearidade)

Veja Zuur *et al.* (2010):

1 Formulate biological hypothesis
Carry out experiment & collect data

Data exploration

1. Outliers Y & X	*boxplot & Cleveland dotplot*
2. Homogeneity Y	*conditional boxplot*
3. Normality Y	*histogram or QQ-plot*
4. Zero trouble Y	*frequency plot or corrgram*
5. Collinearity X	*VIF & scatterplots correlations & PCA*
6. Relationships Y & X	*(multi-panel) scatterplots conditional boxplots*
7. Interactions	*coplots*
8. Independence Y	*ACF & variogram plot Y versus time/space*

2

3 Apply statistical model

Funções principais que vimos na aula:

- `summary`
- `str`
- `head, tail`
- `is.na`
- `mean, median, quantile`
- `plot`
- `scatter.smooth`
- `barplot`
- `boxplot`
- `hist`
- `density`
- `stripchart`
- `dotchart`
- `table, xtabs`
- `qqnorm, qqline`
- `aggregate`
- `xyplot, bwplot`
- `pairs`
- `cor`
- `dist`
- `Hclust, metaMDS`

Sugestões de leitura

Cleveland, W. 1993. **Visualizing data**. Hobart Press.

Ellison, A. M. 1993. Exploratory data analysis and graphic display. In: Scheiner, S. M. (ed.), ***Design and analysis of ecological experiments***. Chapman & Hall, pp. 14-45.

Zuur, A., Ieno, E. N. & Smith G. M. 2007. **Analysing ecological data**. Springer. *** Capítulo 4.

Zuur, A., Ieno, E. N. & Elphick, C. S. 2010. A protocol for data exploration to avoid common statistical problems. **Methods in Ecology & Evolution**, 1: 3-14.

Outros tópicos importantes

- Transformação de variáveis

- Independência dos dados (autocorrelação espacial e temporal)

FIM DA UNIDADE 4

Para a tarde:

Plantão Tutoriais e exercícios EDA

Até segunda:

Lista 4 de Exercícios:
http://ecologia.ib.usp.br/bie5782/doku.php?id=bie5782:01_curso_atual:exercicios4