AULA 5: ANÁLISE GRÁFICA E EXPLORATÓRIA

Análise Quantitativa de Dados Ambientais

Thiago S. F. Silva - tsfsilva@rc.unesp.br

28 de Agosto de 2015

Programa de Pós Graduação em Geografia - IGCE/UNESP

OUTLINE

PRA QUE SERVE ANÁLISE EXPLORATÓRIA?

Você iria a um encontro surpresa, sem saber nada sobre seu par?

Você iria a um encontro surpresa, sem saber nada sobre seu par?

Análise exploratória de dados (AED)

- É essencial ficar "íntimo" dos dados antes de qualquer análise
- Você já possui um modelo conceitual (**Ou deveria...**)
- Será que seus dados se conformam a esse modelo?
- Será que seus dados foram coletados corretamente?
- Será que seus dados foram *registrados* corretamente?
- Será ...?

Análise exploratória de dados (AED)

- É essencial ficar "íntimo" dos dados antes de qualquer análise
- Você já possui um modelo conceitual (**Ou deveria...**)
- Será que seus dados se conformam a esse modelo?
- Será que seus dados foram coletados corretamente?
- Será que seus dados foram *registrados* corretamente?
- Será ...?

Análise exploratória de dados (AED)

- É essencial ficar "íntimo" dos dados antes de qualquer análise
- Você já possui um modelo conceitual (**Ou deveria...**)
- Será que seus dados se conformam a esse modelo?
- Será que seus dados foram coletados corretamente?
- Será que seus dados foram *registrados* corretamente?
- Será ...?

Análise exploratória de dados (AED)

- É essencial ficar "íntimo" dos dados antes de qualquer análise
- Você já possui um modelo conceitual (**Ou deveria...**)
- Será que seus dados se conformam a esse modelo?
- Será que seus dados foram coletados corretamente?
- Será que seus dados foram *registrados* corretamente?
- Será ...?

Análise exploratória de dados (AED)

- É essencial ficar "íntimo" dos dados antes de qualquer análise
- Você já possui um modelo conceitual (**Ou deveria...**)
- Será que seus dados se conformam a esse modelo?
- Será que seus dados foram coletados corretamente?
- Será que seus dados foram *registrados* corretamente?
- Será ...?

Análise exploratória de dados (AED)

- É essencial ficar "íntimo" dos dados antes de qualquer análise
- Você já possui um modelo conceitual (**Ou deveria...**)
- Será que seus dados se conformam a esse modelo?
- Será que seus dados foram coletados corretamente?
- Será que seus dados foram *registrados* corretamente?
- Será ...?

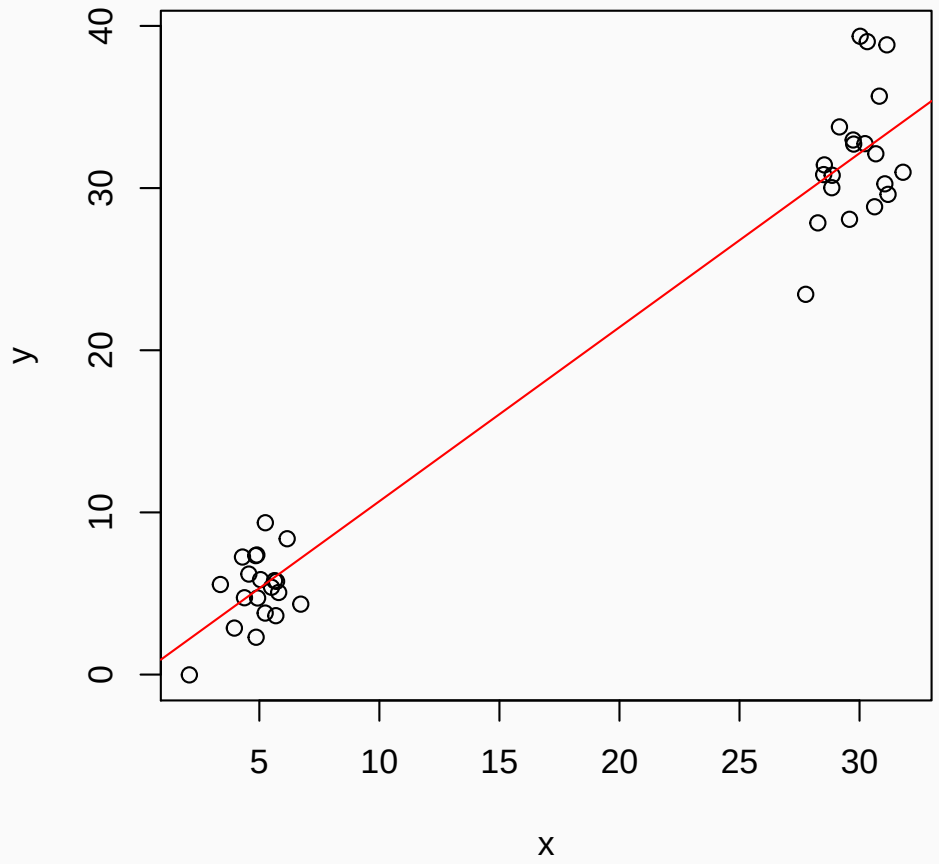
ANÁLISE EXPLORATÓRIA DE DADOS (AED)

```r
summary(m1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2944 -2.2683 -0.1736  1.8512  7.1844
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04236    0.79101  -0.054    0.958
## x            1.07306    0.03693  29.054   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.918 on 38 degrees of freedom
## Multiple R-squared:  0.9569,^^IAdjusted R-squared:  0.9558
## F-statistic: 844.2 on 1 and 38 DF,  p-value: < 2.2e-16
```

ANÁLISE EXPLORATÓRIA DE DADOS (AED)

```r
x <- c(rnorm(20,5,1),rnorm(20,30,1))
y <- x + rnorm(40,0,3)
```

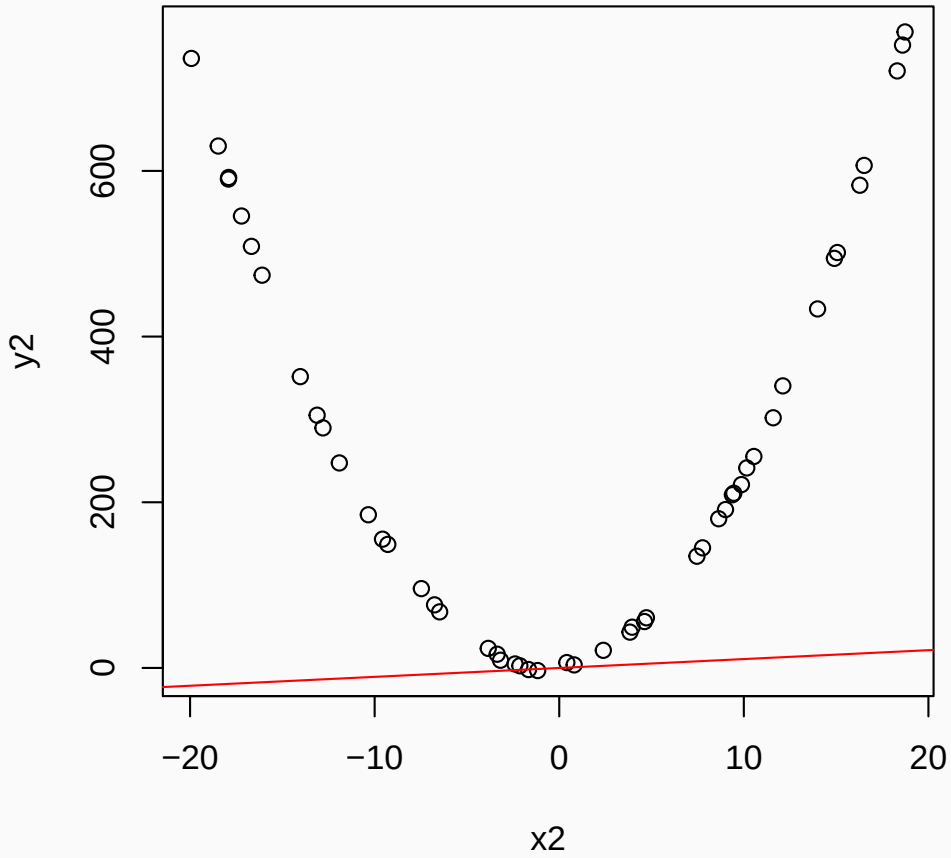


ANÁLISE EXPLORATÓRIA DE DADOS (AED)

```r
summary(m2)
```

```
##
## Call:
## lm(formula = y2 ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -273.08 -219.65  -72.45  210.62  488.11
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  271.285     34.386   7.889 3.24e-10 ***
## x2             1.185      2.969   0.399    0.692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 243.1 on 48 degrees of freedom
## Multiple R-squared:  0.003305,^^IAdjusted R-squared:  -0.01746
## F-statistic: 0.1592 on 1 and 48 DF,  p-value: 0.6917
```

```
x2 <- runif(50,-20,20)
y2 <- 2 + 3*x2 + (2*x22)+rnorm(50,0,3)
```



”O Quarteto de Anscombe”

Anscombe, F.J., 1973. Graphs in Statistical Analysis. The American Statistician 27, 17–21.

```r
m1 <- lm(y1 ~ x1, data = ans)

m1$coefficients
```

```
## (Intercept)          x1
##   3.0000909   0.5000909
```

```r
m2 <- lm(y2 ~ x2, data = ans)

m2$coefficients
```

```
## (Intercept)          x2
##    3.000909    0.500000
```

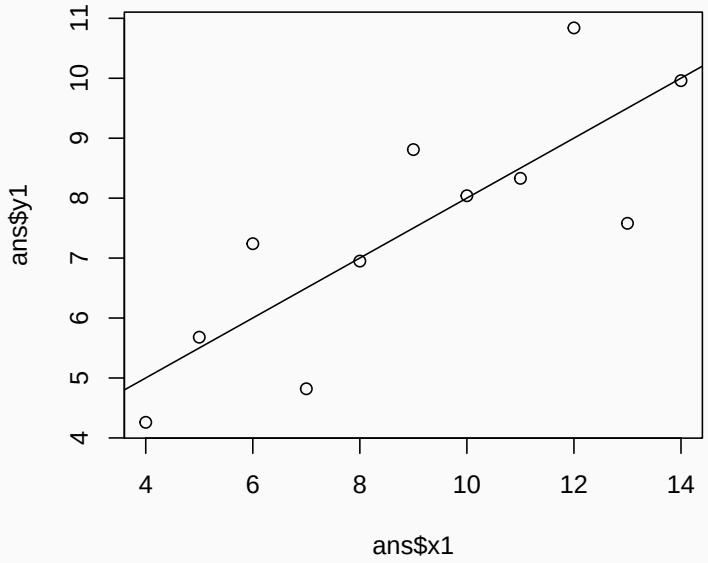
```r
m3 <- lm(y3 ~ x3, data = ans)

m3$coefficients
```

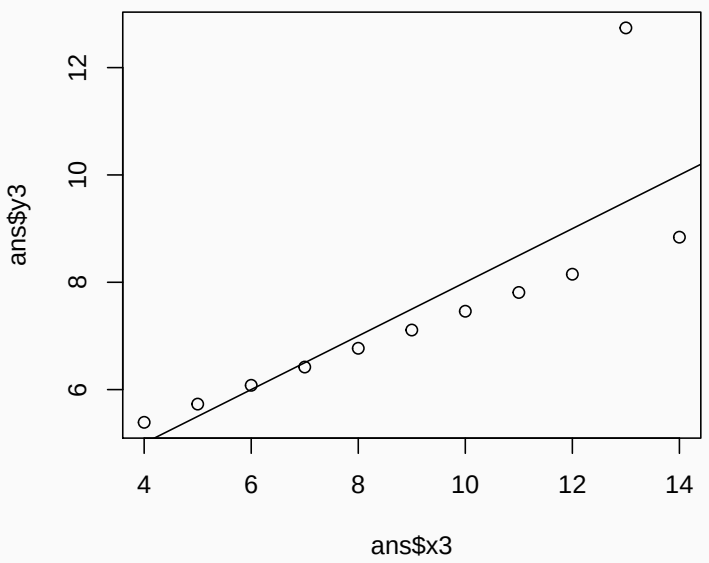
```
## (Intercept)          x3
##   3.0024545   0.4997273
```

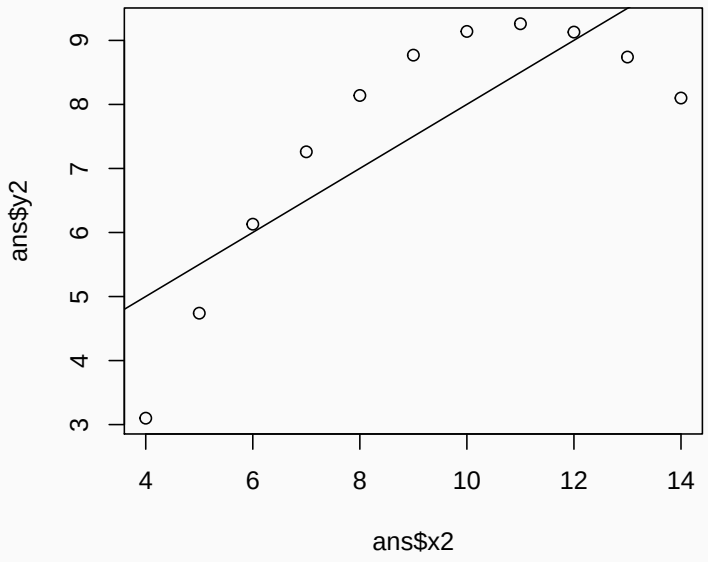
```r
m4 <- lm(y4 ~ x4, data = ans)

m4$coefficients
```

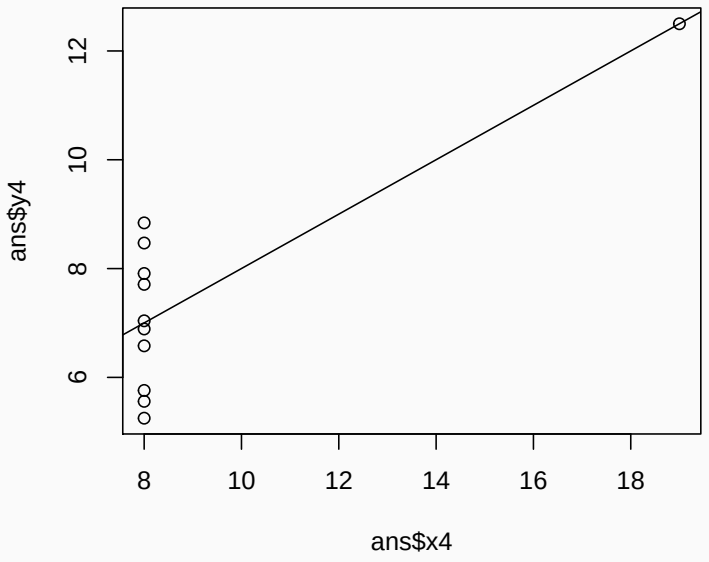
```
## (Intercept)          x4
##   3.0017273   0.4999091
```

O QUARTETO DE ANSCOMBE









Análise Exploratória de Dados (AED)

A Análise Exploratória é normalmente composta por:

- **Estatísticas Descritivas**
- **Análise Gráfica**
- Aderência à distribuição
- Análise de Relações

CODIFICAÇÃO E ORGANIZAÇÃO DE DADOS

- Na maioria das vezes, recebemos ou tabulamos nossos dados no formato *wide* (largo)
- Mas a maioria dos pacotes de análise requer tabelas no formato *long* (longo)
- Exemplo: um estudo de lagos em ambientes litorâneos, interiores e montanhosos

	Lago	Tinv	Tver	Phinv	Phver
1	L1	20.32	28.35	6.50	7.07
2	L2	19.63	27.33	6.82	7.26
3	L3	20.99	26.05	7.11	7.08
4	L4	20.31	26.13	8.00	5.28
5	L5	20.13	26.54	6.56	6.58
6	L6	19.93	26.53	7.46	8.05
7	L7	20.95	24.63	7.25	7.15
8	L8	21.13	27.25	7.06	6.77
9	L9	19.70	26.82	6.70	6.59
10	L10	20.04	27.90	5.92	6.91

Tabela 1: Lagos litorâneos

	Lago	Tinv	Tver	Phinv	Phver
1	L11	14.99	23.41	6.37	6.60
2	L12	14.35	22.82	5.50	7.40
3	L13	14.22	23.39	6.39	7.71
4	L14	15.44	22.59	6.74	6.81
5	L15	14.23	23.70	6.63	7.45
6	L16	14.83	22.29	6.11	6.77
7	L17	15.93	23.06	8.07	7.28
8	L18	14.17	22.11	7.03	6.65
9	L19	14.78	23.94	6.55	7.15
10	L20	14.73	23.68	6.93	5.93

Tabela 2: Lagos interiores

	Lago	Tinv	Tver	Phinv	Phver
1	L21	9.83	22.70	6.78	7.40
2	L22	10.13	22.95	7.19	6.76
3	L23	10.65	22.89	6.93	6.27
4	L24	8.99	22.56	7.10	6.55
5	L25	9.81	21.31	6.56	7.26
6	L26	8.86	22.76	6.69	7.26
7	L27	9.84	22.08	6.69	7.07
8	L28	9.16	23.34	7.38	7.71
9	L29	10.90	20.33	7.98	7.15
10	L30	10.43	21.62	7.30	6.75

Tabela 3: Lagos de montanha

Quantas variáveis possui esse estudo?

QUANTAS VARIÁVEIS?

- Lago (L1, L2,...- poderia ser analisada se tomássemos várias observações por lago)

QUANTAS VARIÁVEIS?

- Lago (L1, L2,...- poderia ser analisada se tomássemos várias observações por lago)
- Região (Litoral, Interior, Montanha)

QUANTAS VARIÁVEIS?

- Lago (L1, L2,...- poderia ser analisada se tomássemos várias observações por lago)
- Região (Litoral, Interior, Montanha)
- Estação (Inverno, Verão)

QUANTAS VARIÁVEIS?

- Lago (L1, L2,...- poderia ser analisada se tomássemos várias observações por lago)
- Região (Litoral, Interior, Montanha)
- Estação (Inverno, Verão)
- Temperatura

QUANTAS VARIÁVEIS?

- Lago (L1, L2,...- poderia ser analisada se tomássemos várias observações por lago)
- Região (Litoral, Interior, Montanha)
- Estação (Inverno, Verão)
- Temperatura
- pH

- No formato longo, cada coluna descreve uma variável, e cada linha representa uma observação:

- No formato longo, cada coluna descreve uma variável, e cada linha representa uma observação:

```
lago.long <- read.csv('lakedata_long.csv')
```

	Lago	Regiao	Estacao	Temp	pH
1	L1	Litoral	Inverno	20.32	6.50
2	L2	Litoral	Inverno	19.63	6.82
3	L3	Litoral	Inverno	20.99	7.11
4	L4	Litoral	Inverno	20.31	8.00
5	L5	Litoral	Inverno	20.13	6.56
6	L6	Litoral	Inverno	19.93	7.46
7	L7	Litoral	Inverno	20.95	7.25
8	L8	Litoral	Inverno	21.13	7.06
9	L9	Litoral	Inverno	19.70	6.70
10	L10	Litoral	Inverno	20.04	5.92
11	L11	Interior	Inverno	14.99	6.37
12	L12	Interior	Inverno	14.35	5.50
...	...	...	...	...	...
54	L24	Montanha	Verao	22.56	6.55
55	L25	Montanha	Verao	21.31	7.26
56	L26	Montanha	Verao	22.76	7.26
57	L27	Montanha	Verao	22.08	7.07
58	L28	Montanha	Verao	23.34	7.71
59	L29	Montanha	Verao	20.33	7.15
60	L30	Montanha	Verao	21.62	6.75

ESTATÍSTICA DESCRITIVA

ESTATÍSTICA DESCRITIVA: TENDÊNCIA CENTRAL E DISPERSÃO

Através da estatística descritiva, buscamos:

- *Localizar* nossos dados no espaço (numérico)
  - Quais os valores *esperados* para estes dados?

- Quantificar a *dispersão* destes dados em torno desta localidade
  - Qual a *variância* dos meus dados?

Tendência central para dados contínuos?

Tendência central para dados contínuos?

- Média

$$\bar{X}_{(arit)} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Dispersão para dados contínuos?

Dispersão para dados contínuos?

- Variância e Desvio Padrão

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s = \sqrt{s^2}$$

Dispersão para dados contínuos?

- Variância e Desvio Padrão
- Amplitude

$$A = max(x) - min(x)$$

Dispersão para dados contínuos?

- Variância e Desvio Padrão
- Amplitude
- Coeficiente de Variação

$$CV = \frac{s}{\bar{x}} * 100(\%)$$

```r
set.seed(1979)
x <- rnorm(500,50,10)
cv <- function(x) sd(x)/mean(x) * 100
mean(x)
```

```
## [1] 50.36405
```

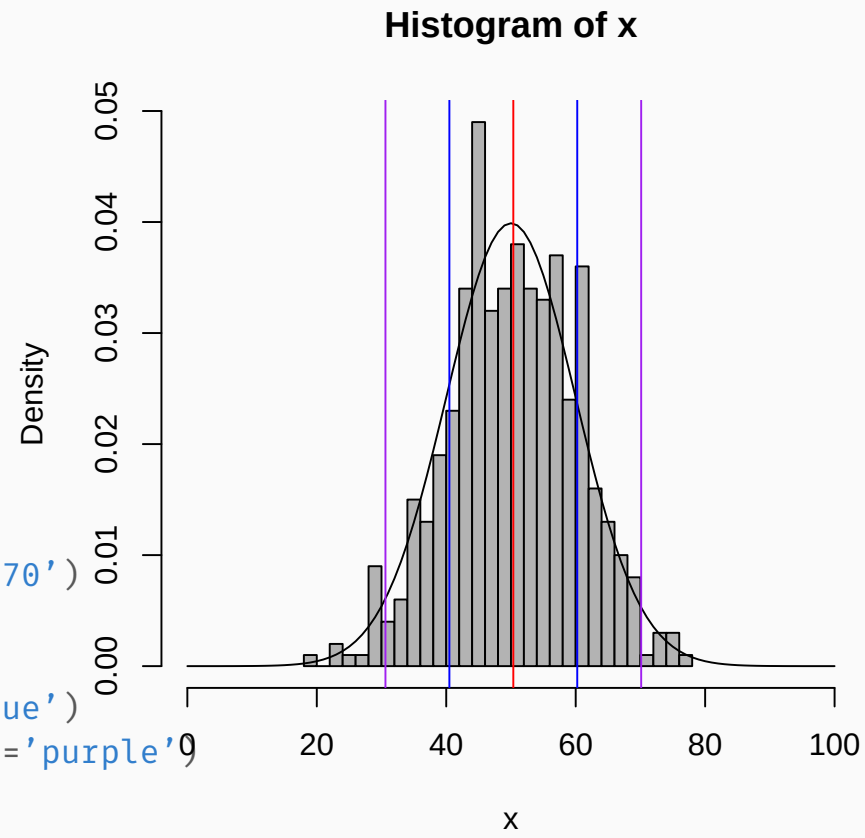
```r
sd(x)
```

```
## [1] 9.877513
```

```r
cv(x)
```

```
## [1] 19.61223
```

```r
hist(x,breaks=40,prob=T,xlim=c(0,100),col='gray70')
curve(dnorm(x,mean=50, sd=10), add=T)
abline(v=mean(x),col='red')
abline(v=c(mean(x)+sd(x),mean(x)-sd(x)),col='blue')
abline(v=c(mean(x)+2*sd(x),mean(x)-2*sd(x)),col='purple')
```



```r
set.seed(1979)
x <- rnorm(500,50,20)
cv <- function(x) sd(x)/mean(x) * 100
mean(x)
```

```
## [1] 50.7281
```

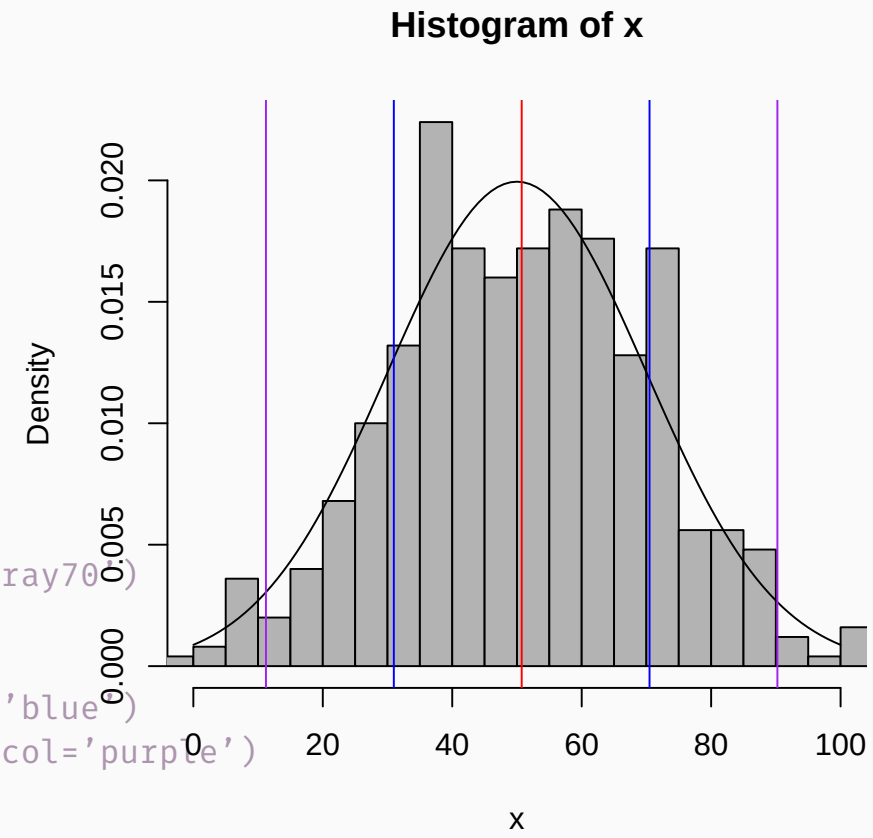
```r
sd(x)
```

```
## [1] 19.75503
```

```r
cv(x)
```

```
## [1] 38.94297
```

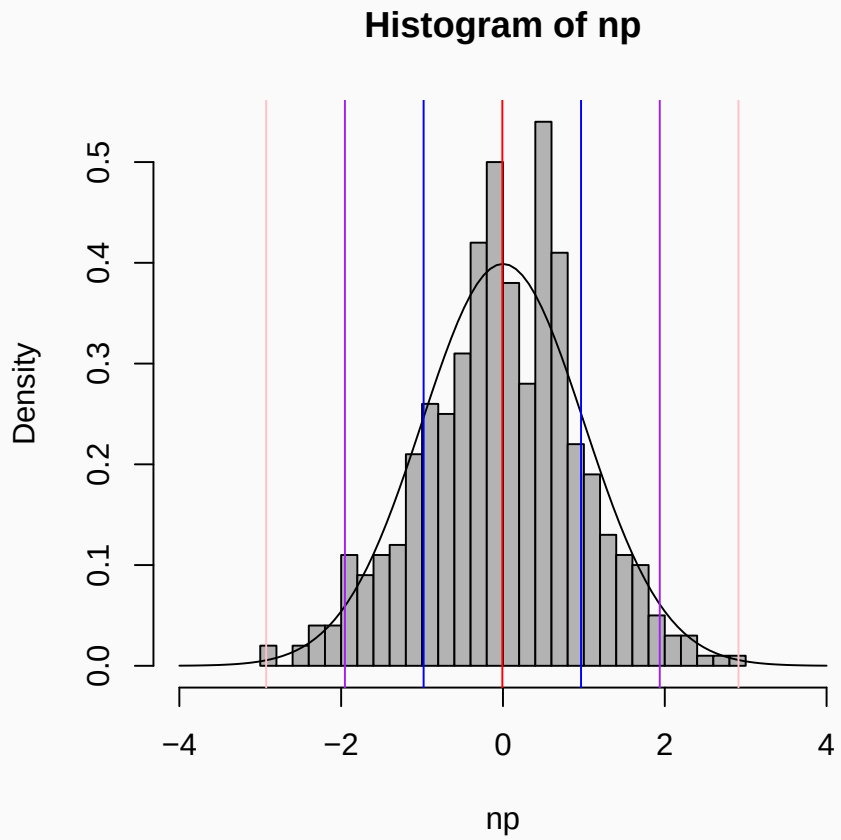
```r
## hist(x,breaks=40,prob=T,xlim=c(0,100),col='gray70')
## curve(dnorm(x,mean=50, sd=20), add=T)
## abline(v=mean(x),col='red')
## abline(v=c(mean(x)+sd(x),mean(x)-sd(x)),col='blue')
## abline(v=c(mean(x)+2*sd(x),mean(x)-2*sd(x)),col='purple')
```



ESTATÍSTICA DESCRITIVA: DESVIO PADRÃO

A regra do 68-95-99.7
Para uma distribuição normal

- $1 \times s \approx 68$ % dos dados
- $2 \times s \approx 95$ % dos dados
- $3 \times s \approx 99.7$ % dos dados



Percentis (*Percentiles*): medidas robustas de tendência central e dispersão

k-ésimo percentil = k% de n está abaixo desse valor, $(100 - k)$% de n está acima desse valor

- Mediana: $k = 50\%$, $100 - k = 50\%$

Percentis (*Percentiles*): medidas robustas de tendência central e dispersão

k-ésimo percentil = k% de n está abaixo desse valor, $(100 - k)$% de n está acima desse valor

- **Mediana**: $k = 50\%$, $100 - k = 50\%$

Problema: nem sempre existe um valor que satisfaça a condição, para um dado k

$x = [1, 2, 4, 33, 200]$

$k(0\%) = 1$

$k(25\%) = 2$

$k(50\%) = 4$

$k(75\%) = 33$

$k(100\%) = 200$

$k(95\%) =?$

Quantis (*Quantiles*): podem ser calculados para qualquer amostra

$$Q(p) = x : P(X < x) \leq p; P(X > x) \geq (1 - p)$$

Se não há valores que satisfazem essa condição, o valor é interpolado proporcionalmente à distância entre os percentis dos valores x_i e x_{i+1} que enquadram o quantil desejado.

- Mediana: $Q(0.5)$
- Quartil Superior: $Q(0.75)$
- Quartil Inferior: $Q(0.25)$

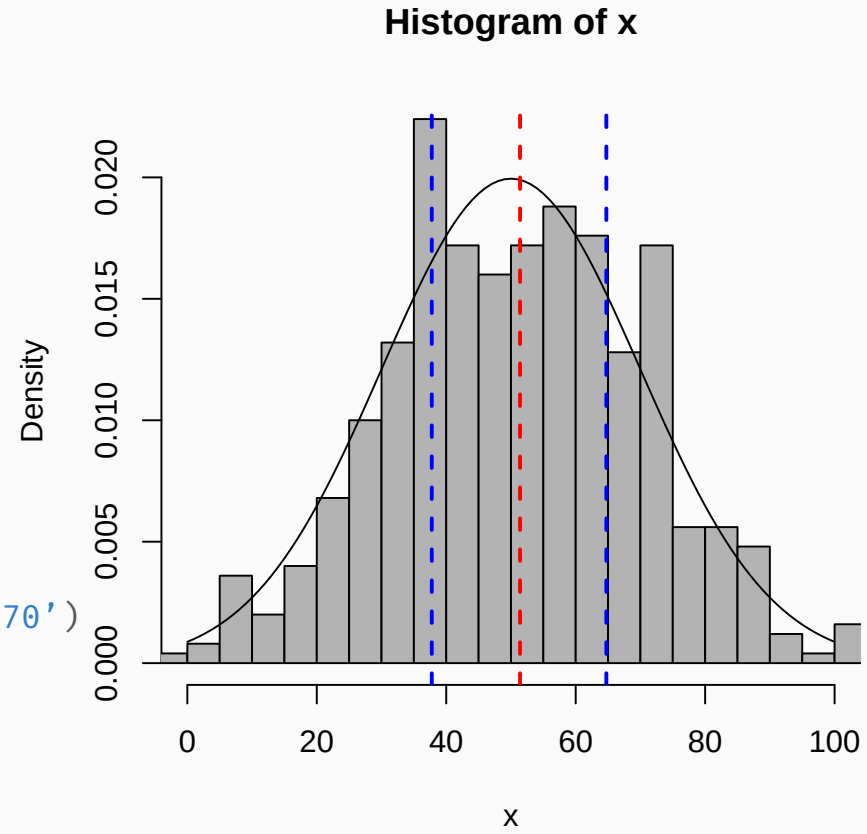
Quantis (*Quantiles*): podem ser calculados para qualquer amostra

$$Q(p) = x : P(X < x) \leq p; P(X > x) \geq (1 - p)$$

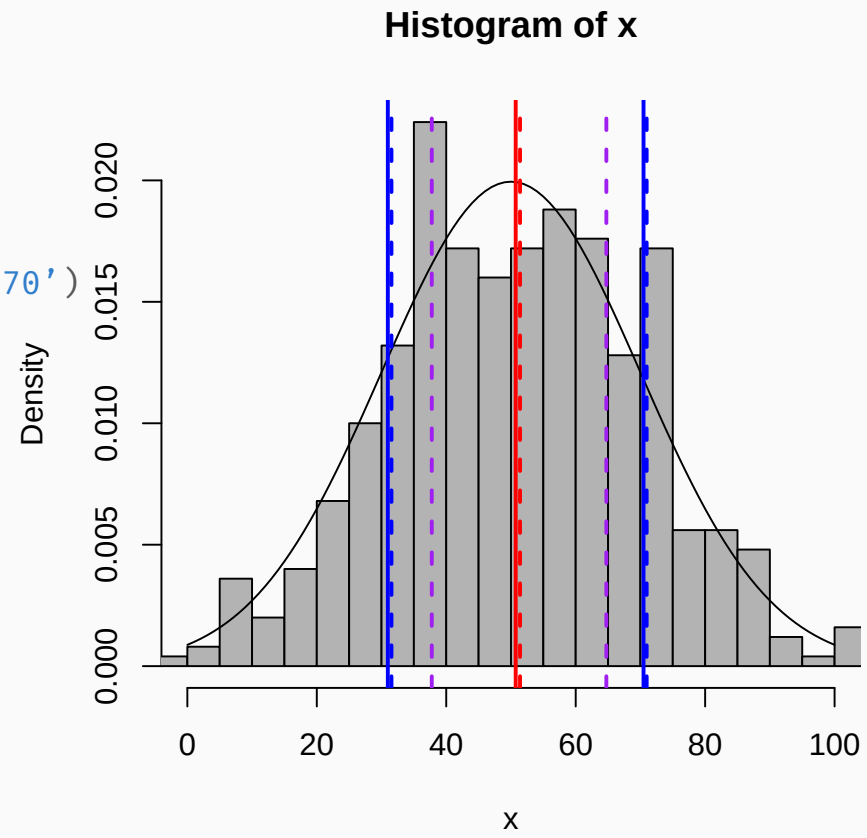
Se não há valores que satisfazem essa condição, o valor é interpolado proporcionalmente à distância entre os percentis dos valores x_i e x_{i+1} que enquadram o quantil desejado.

- Mediana: $Q(0.5)$
- Quartil Superior: $Q(0.75)$
- Quartil Inferior: $Q(0.25)$

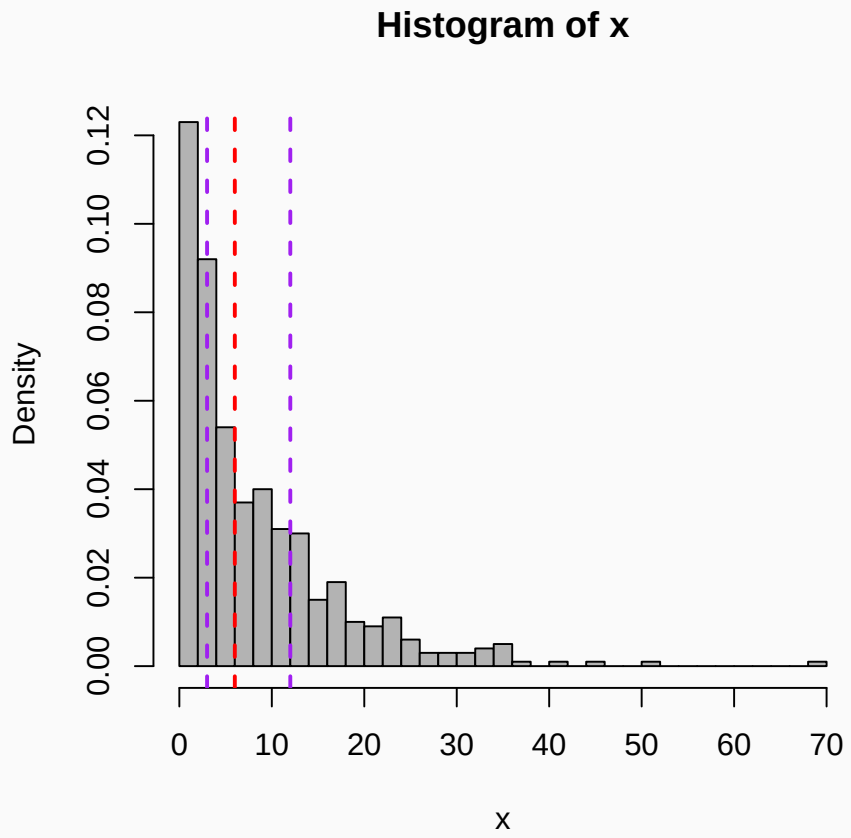
```r
set.seed(1979)
x <- rnorm(500,50,20)

median(x)
## [1] 51.39951
quantile(x,prob=0.5)
##      50%
## 51.39951
quantile(x,prob=c(0.25,0.75))
##      25%      75%
## 37.76660 64.73483
hist(x,breaks=40,prob=T,xlim=c(0,100),col='gray70')
curve(dnorm(x,mean=50, sd=20), add=T)
abline(v=median(x),col='red',lwd=2,lty=2)
abline(v=quantile(x,prob=c(0.25,0.75)),#
       col='blue',lwd=2,lty=2)
```



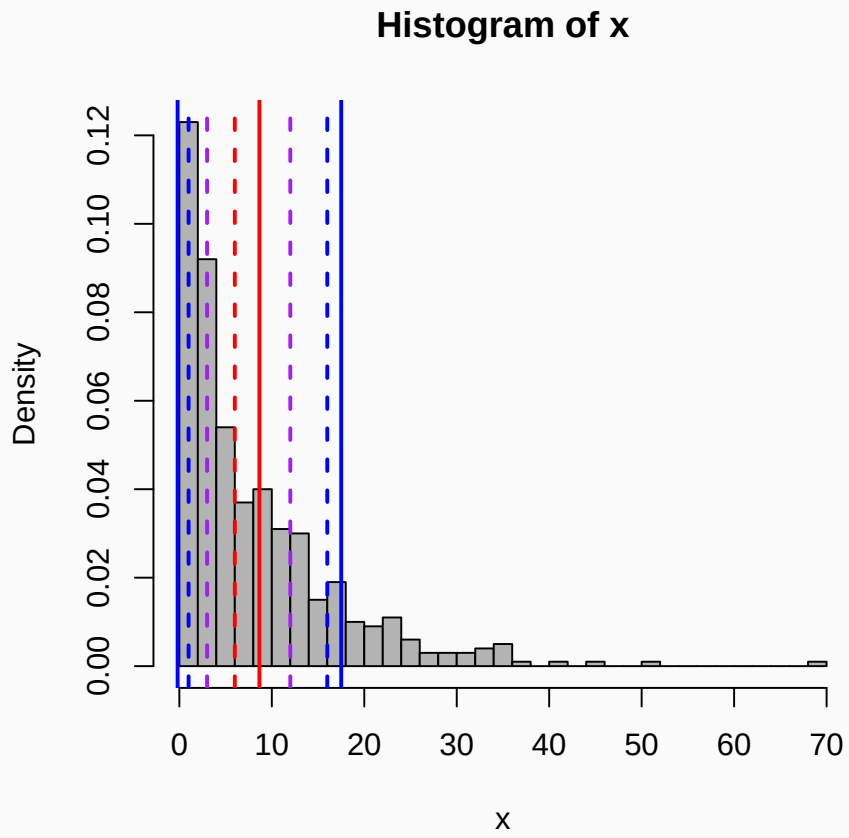
```r
set.seed(1979)
x <- rnorm(500,50,20)

hist(x,breaks=40,prob=T,xlim=c(0,100),col='gray70')
curve(dnorm(x,mean=50, sd=20), add=T)
abline(v=mean(x),col='red',lwd=2,lty=1)
abline(v=median(x),col='red',lwd=2,lty=2)
abline(v=c(mean(x)+sd(x),mean(x)-sd(x)),#
       col='blue',lwd=2,lty=1)
abline(v=quantile(x,prob=c(0.25,0.75)),#
       col='purple',lwd=2,lty=2)
abline(v=quantile(x,prob=c(0.16,0.84)),#
       col='blue',lwd=2,lty=2)
```



```r
set.seed(1979)
x <- rgeom(500,0.1)

mean(x)
## [1] 8.66
median(x)
## [1] 6
sd(x)
## [1] 8.846973
quantile(x,probs=c(0.25,0.50))
## 25% 50%
##   3   6
```

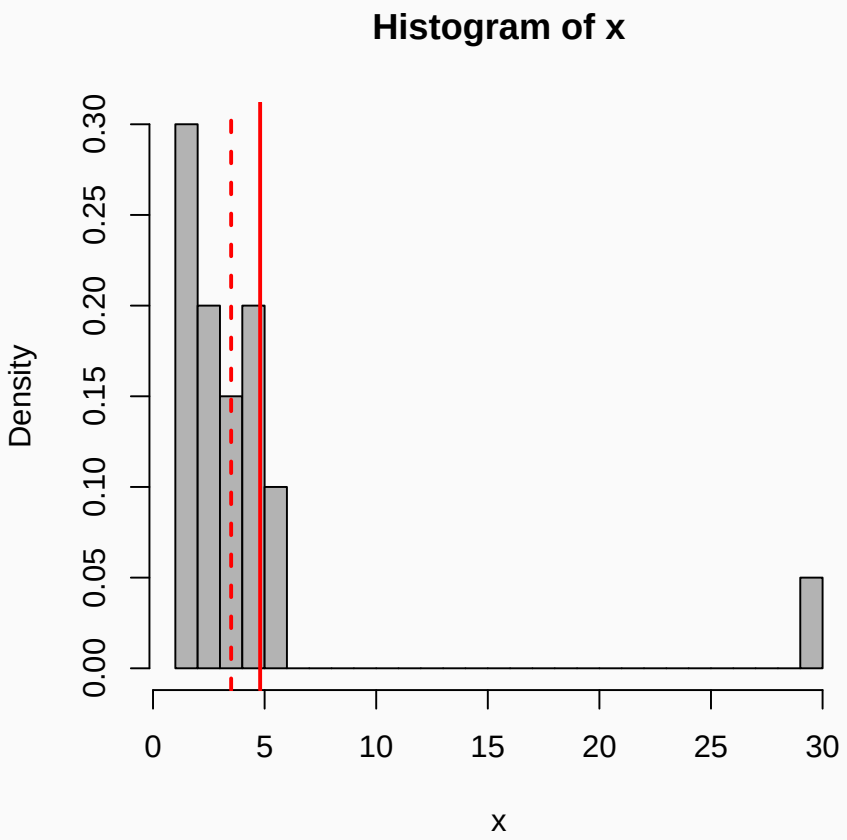


```r
set.seed(1979)
x <- rgeom(500,0.1)
##
## hist(x,breaks=30,prob=T,col='gray70')
## abline(v=mean(x),col='red',lwd=2,lty=1)
## abline(v=median(x),col='red',lwd=2,lty=2)
## abline(v=c(mean(x)+sd(x),mean(x)-sd(x)),#
##        col='blue',lwd=2,lty=1)
## abline(v=quantile(x,prob=c(0.25,0.75)),#
##        col='purple',lwd=2,lty=2)
## abline(v=quantile(x,prob=c(0.16,0.84)),#
##        col='blue',lwd=2,lty=2)
```



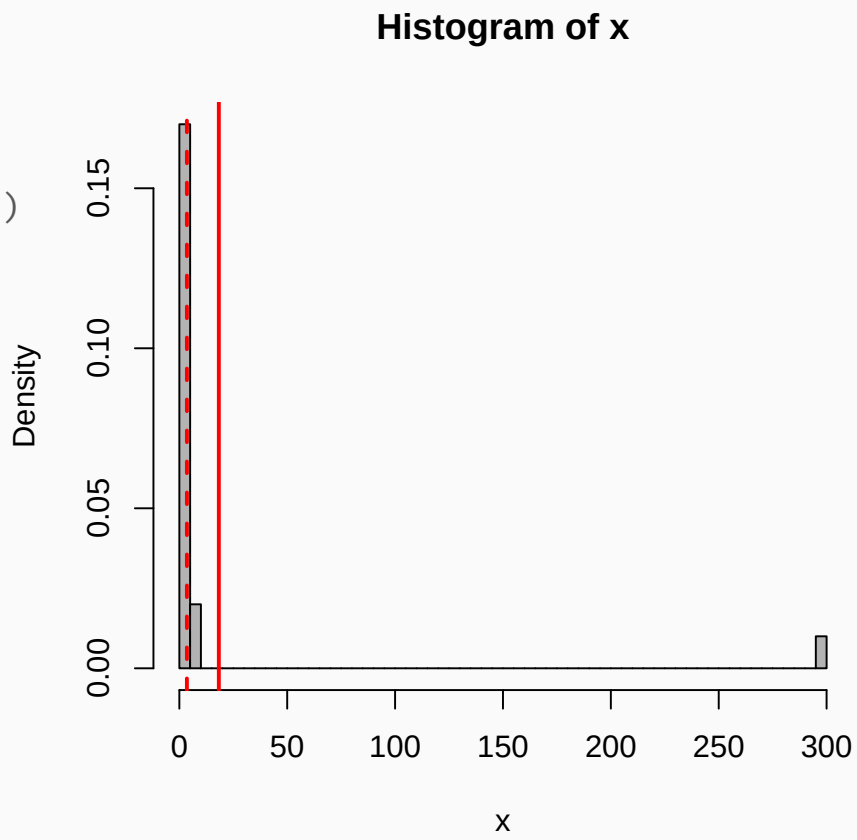
- Os quantis são muito mais robustos com relação a valores extremos (*outliers*)

```r
x <- c(1,2,3,2,3,4,5,6,2,3,5,4,1,2,3,4,5,5,6,30)

mean(x)

## [1] 4.8

median(x)

## [1] 3.5

hist(x,breaks=40,prob=T,,col='gray70')
abline(v=mean(x),col='red',lwd=2,lty=1)
abline(v=median(x),col='red',lwd=2,lty=2)
```



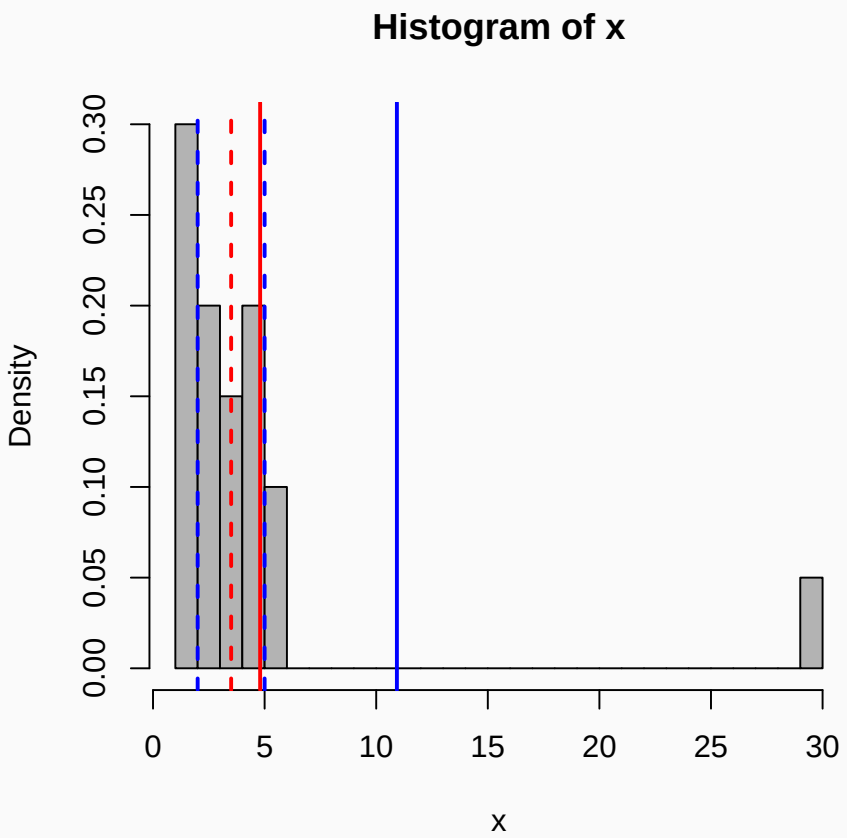
- Os quantis são muito mais robustos com relação a valores extremos (*outliers*)

```r
x <- c(1,2,3,2,3,4,5,6,2,3,5,4,1,2,3,4,5,5,6,300)

mean(x)

## [1] 18.3

median(x)

## [1] 3.5

hist(x,breaks=80,prob=T,,col='gray70')
abline(v=mean(x),col='red',lwd=2,lty=1)
abline(v=median(x),col='red',lwd=2,lty=2)
```



- Os quantis são muito mais robustos com relação a valores extremos (*outliers*)

```
x <- c(1,2,3,2,3,4,5,6,2,3,5,4,1,2,3,4,5,5,6,30)

sd(x)

## [1] 6.126732

quantile(x,prob=c(0.16,0.84))

## 16% 84%
##   2   5

hist(x,breaks=40,prob=T,,col='gray70')
abline(v=mean(x),col='red',lwd=2,lty=1)
abline(v=median(x),col='red',lwd=2,lty=2)
abline(v=c(mean(x)-sd(x),mean(x)+sd(x)),#
       col='blue',lwd=2,lty=1)
abline(v=quantile(x,prob=c(0.16,0.84)),#
       col='blue',lwd=2,lty=2)
```



Tendência e dispersão para dados *categóricos*?

- Qual a média de (Floresta, Campo, Cidade)?

Tendência e dispersão para dados *categóricos*?

- Qual a média de (Floresta, Campo, Cidade)?
- Solução: contagens, frequência, porcentagem, *odds*

Tendência e dispersão para dados *categóricos*?

- Qual a média de (Floresta, Campo, Cidade)?
- Solução: contagens, frequência, porcentagem, *odds*
- Exemplo: Você gosta de estatística?

Tendência e dispersão para dados *categóricos*?

- Qual a média de (Floresta, Campo, Cidade)?
- Solução: contagens, frequência, porcentagem, *odds*
- Exemplo: Você gosta de estatística?

Obs.	Gosto
1	Sim
2	Não
3	Não
4	Não
5	Não
6	Não
7	Não
8	Sim
9	Sim

Tendência e dispersão para dados *categóricos*?

- Qual a média de (Floresta, Campo, Cidade)?
- Solução: contagens, frequência, porcentagem, *odds*
- Exemplo: Você gosta de estatística?

Obs.	Gosto
1	Sim
2	Não
3	Não
4	Não
5	Não
6	Não
7	Não
8	Sim
9	Sim

Variável	Contagem	Frequência	Porcentagem	*odds*
Sim	3	0.33	33%	0.5
Não	6	0.66	66%	2

Tendência e dispersão para dados *categóricos*?

- Exemplo: O quanto você gosta de estatística? (1-Abomino, 2-Odeio, 3-Não Gosto, 4-Tolero,5-Adoro)

Obs.	Gosto
1	5
2	1
3	1
4	1
5	2
6	2
7	3
8	3
9	4

Variável	Contagem	Frequência	Porcentagem	*odds*
Abomino	3	0.33	33%	0.5
Odeio	2	0.25	25%	0.29
Não Gosto	2	0.25	25%	0.29
Tolero	1	0.1	10%	0.125
Adoro	1	0.1	10%	0.125

ANÁLISE GRÁFICA

ANÁLISE GRÁFICA

- O ser humano tem uma capacidade incrível de processar informações visuais
- A análise gráfica pode ser considerada uma das partes mais importantes do processo
- Muitas questões podem ser respondidas sem a necessidade de (*mindless*) testes
- Métodos de visualição tem sido um *hot topic* em análise de dados atualmente

ANÁLISE GRÁFICA

- O ser humano tem uma capacidade incrível de processar informações visuais
- A análise gráfica pode ser considerada uma das partes mais importantes do processo
- Muitas questões podem ser respondidas sem a necessidade de (*mindless*) testes
- Métodos de visualição tem sido um *hot topic* em análise de dados atualmente

ANÁLISE GRÁFICA

- O ser humano tem uma capacidade incrível de processar informações visuais
- A análise gráfica pode ser considerada uma das partes mais importantes do processo
- Muitas questões podem ser respondidas sem a necessidade de (*mindless*) testes
- Métodos de visualição tem sido um *hot topic* em análise de dados atualmente

ANÁLISE GRÁFICA

- O ser humano tem uma capacidade incrível de processar informações visuais
- A análise gráfica pode ser considerada uma das partes mais importantes do processo
- Muitas questões podem ser respondidas sem a necessidade de (*mindless*) testes
- Métodos de visualição tem sido um *hot topic* em análise de dados atualmente

Histograma

- Adequado para mostrar distribuições, pode ser usado tanto para dados categóricos quanto contínuos

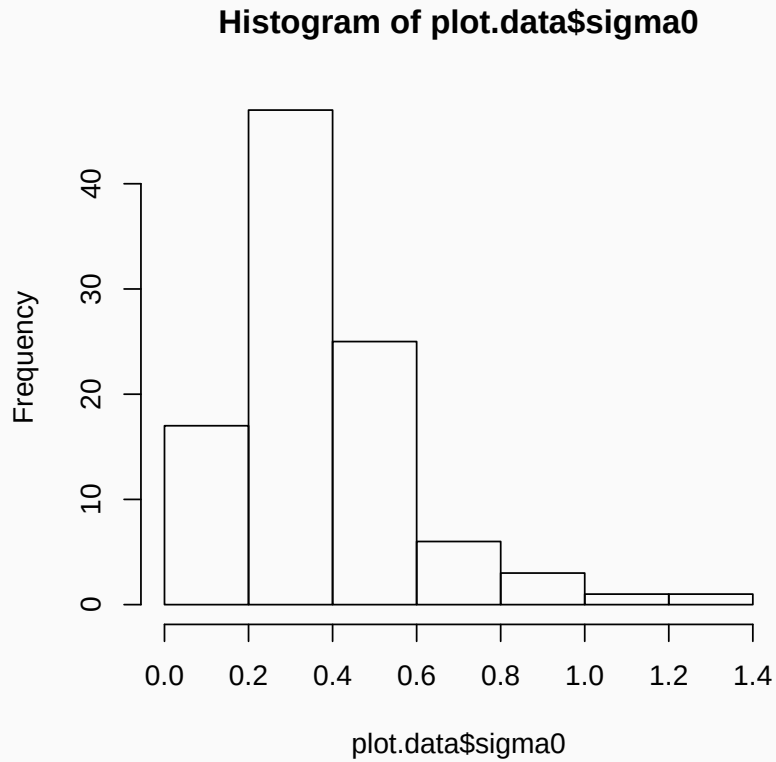
Histograma

- Adequado para mostrar distribuições, pode ser usado tanto para dados categóricos quanto contínuos

- É importante definirem-se bem as subdivisões (*bins*)

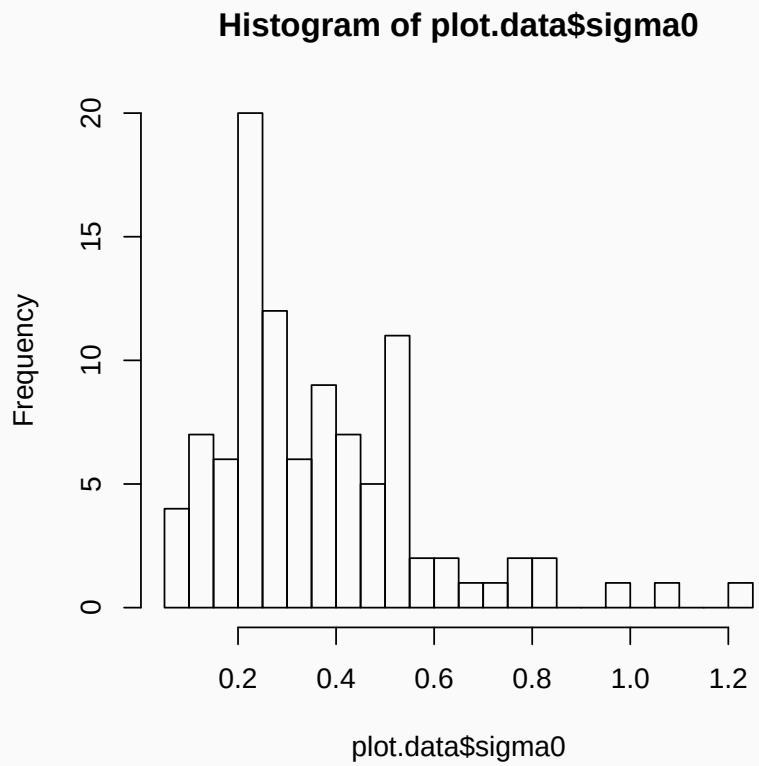
TIPOS DE GRÁFICOS - 1 VARIÁVEL: HISTOGRAMA

```
hist(plot.data$sigma0)
```



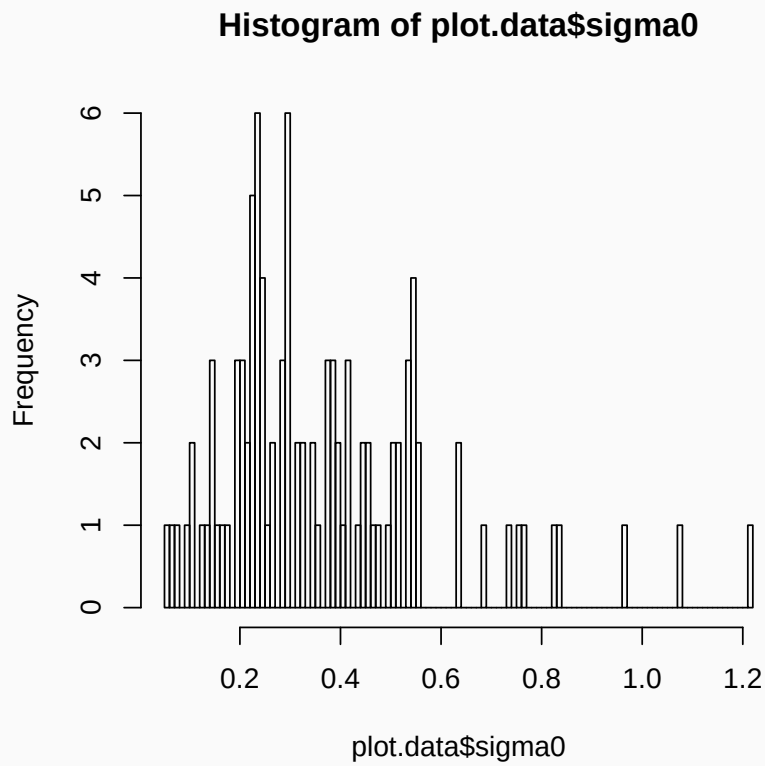
TIPOS DE GRÁFICOS - 1 VARIÁVEL: HISTOGRAMA

```r
hist(plot.data$sigma0)
hist(plot.data$sigma0,breaks=40)
```



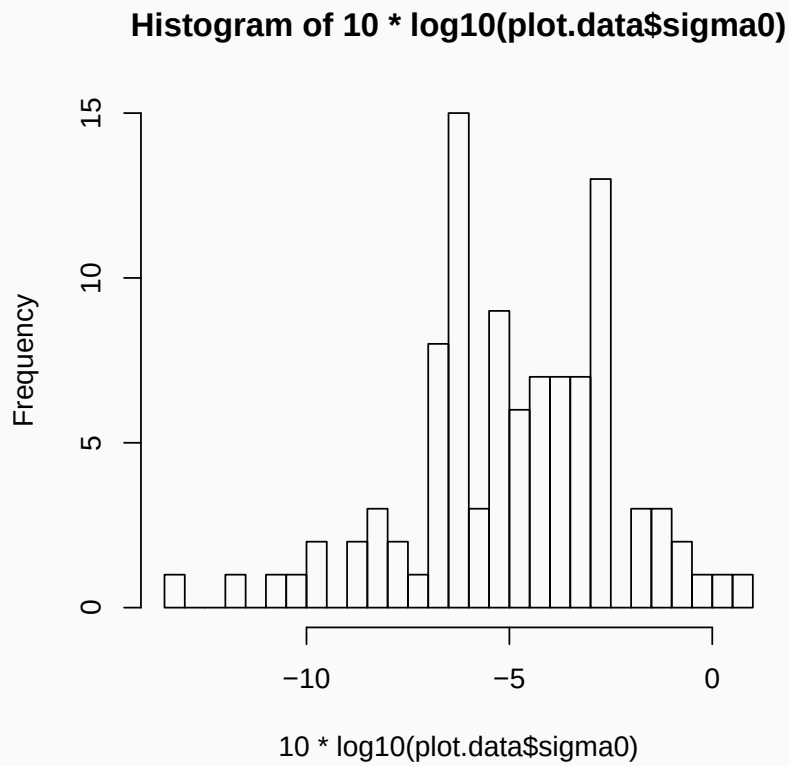
TIPOS DE GRÁFICOS - 1 VARIÁVEL: HISTOGRAMA

```
hist(plot.data$sigma0,breaks=40)
hist(plot.data$sigma0,breaks=100)
```



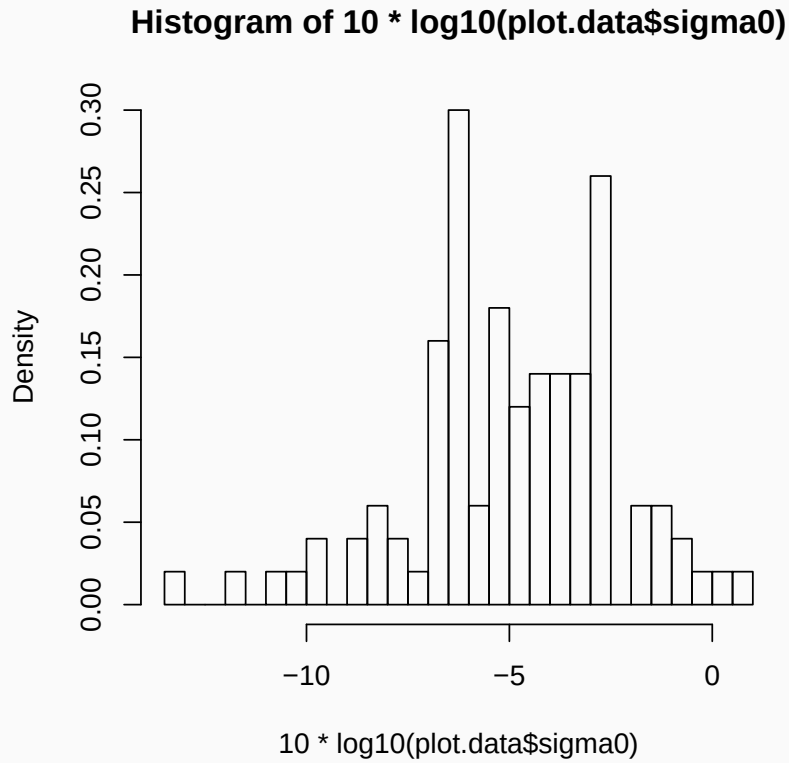
TIPOS DE GRÁFICOS - 1 VARIÁVEL: HISTOGRAMA

```r
hist(plot.data$sigma0)
hist(plot.data$sigma0,breaks=40)
hist(10*log10(plot.data$sigma0),breaks=30)
```



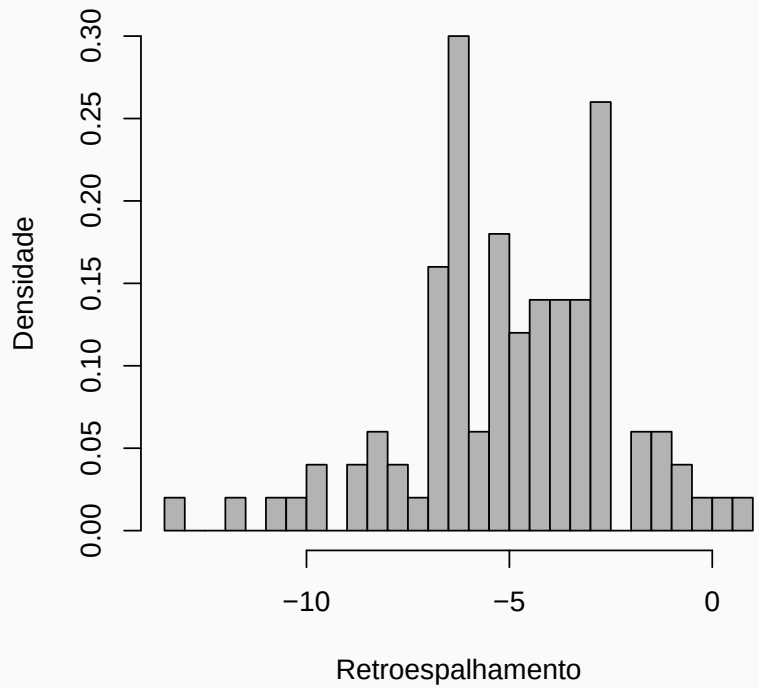
TIPOS DE GRÁFICOS - 1 VARIÁVEL: HISTOGRAMA

```r
hist(plot.data$sigma0)
hist(plot.data$sigma0,breaks=40)
hist(plot.data$sigma0,breaks=400)
hist(10*log10(plot.data$sigma0),breaks=30)
hist(10*log10(plot.data$sigma0),breaks=30,prob=T)
```



TIPOS DE GRÁFICOS - 1 VARIÁVEL: HISTOGRAMA

```
hist(plot.data$sigma0)
hist(plot.data$sigma0,breaks=40)
hist(plot.data$sigma0,breaks=400)
hist(10*log10(plot.data$sigma0),breaks=30)
hist(10*log10(plot.data$sigma0),breaks=30,prob=T)
hist(10*log10(plot.data$sigma0),breaks=30,prob=T, #
     col='gray70',xlab="Retroespalhamento",#
     ylab="densidade",main=NA)
```



Dot Plot

- Similar ao histograma, mas mostra todos as observações

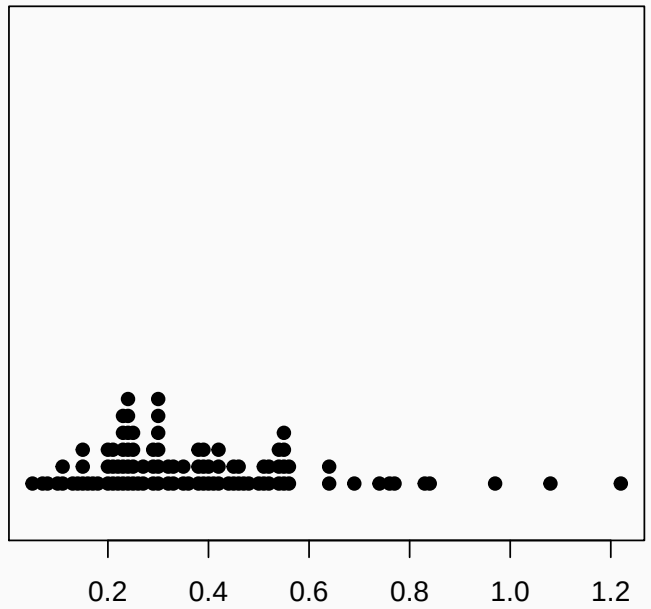
Dot Plot

- Similar ao histograma, mas mostra todos as observações

- Para amostras com poucas observações, ou se a escolha dos *bins* não for adequada, o histograma pode distorcer a forma real da distribuição

TIPOS DE GRÁFICOS - 1 VARIÁVEL: DOT PLOT

```
stripchart(plot.data$sigma0, method = "stack", offset = .5, at = .15, pch = 19)
```



Densidade *Kernel*

- Similar ao histograma, mas ajusta uma linha suavizada à distribuição

Densidade *Kernel*

- Similar ao histograma, mas ajusta uma linha suavizada à distribuição

- Assim como o histograma depende das subdivisões, este gráfico depende da largura do kernel (bandwidth)

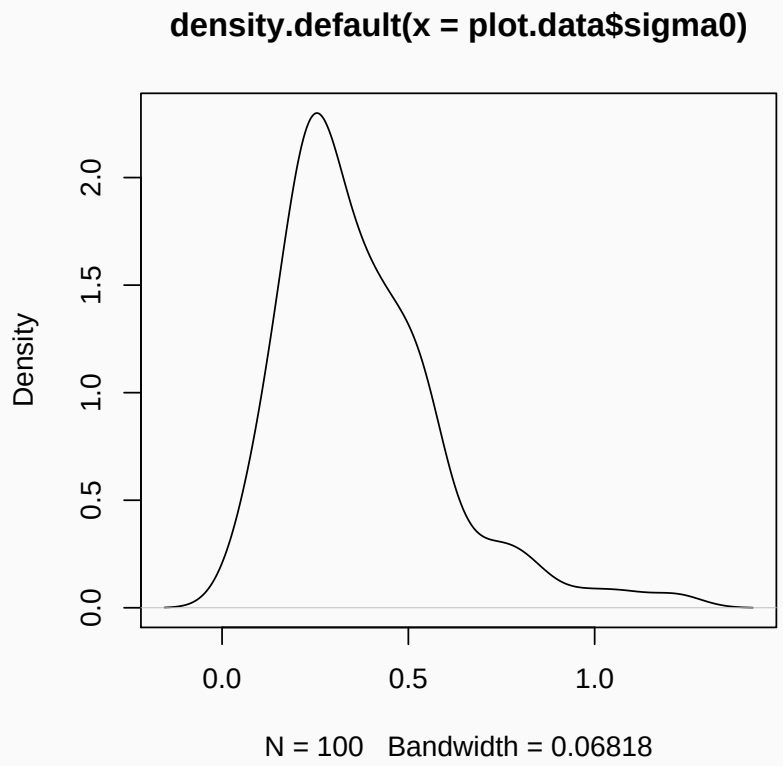
Densidade *Kernel*

- Similar ao histograma, mas ajusta uma linha suavizada à distribuição

- Assim como o histograma depende das subdivisões, este gráfico depende da largura do kernel (bandwidth)

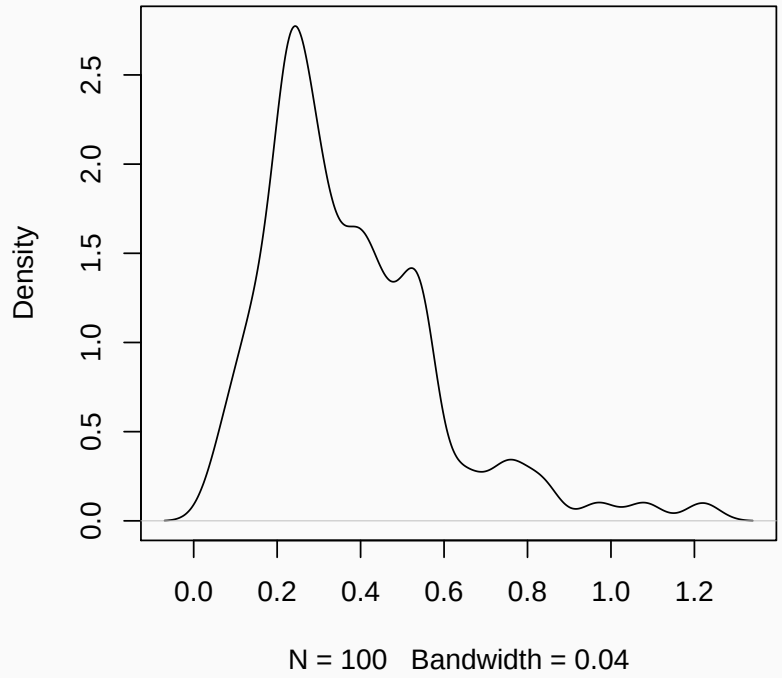
TIPOS DE GRÁFICOS - 1 VARIÁVEL:DENSIDADE *KERNEL*

```
plot(density(plot.data$sigma0))
```



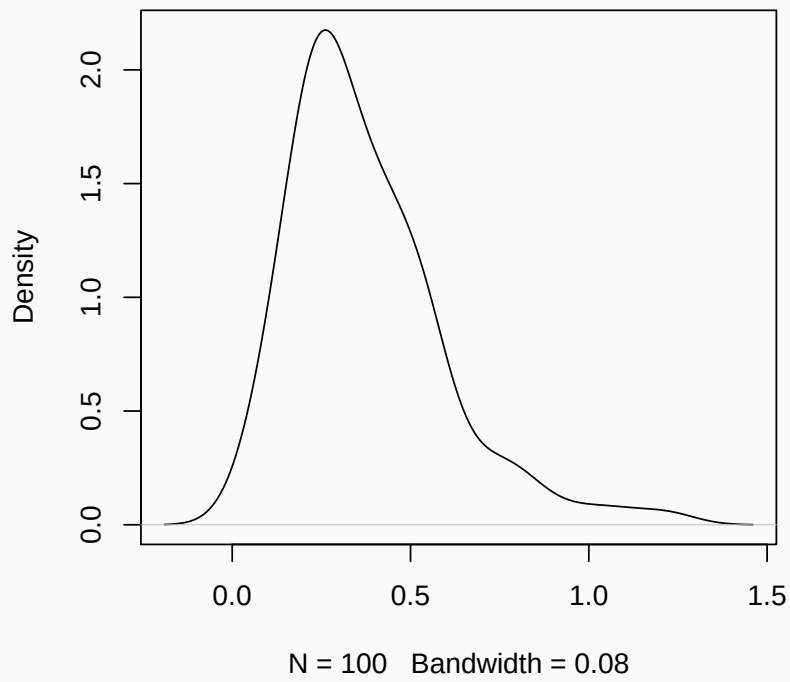
TIPOS DE GRÁFICOS - 1 VARIÁVEL:DENSIDADE *KERNEL*

```
plot(density(plot.data$sigma0))
plot(density(plot.data$sigma0, bw = 0.04),main=NA)
```



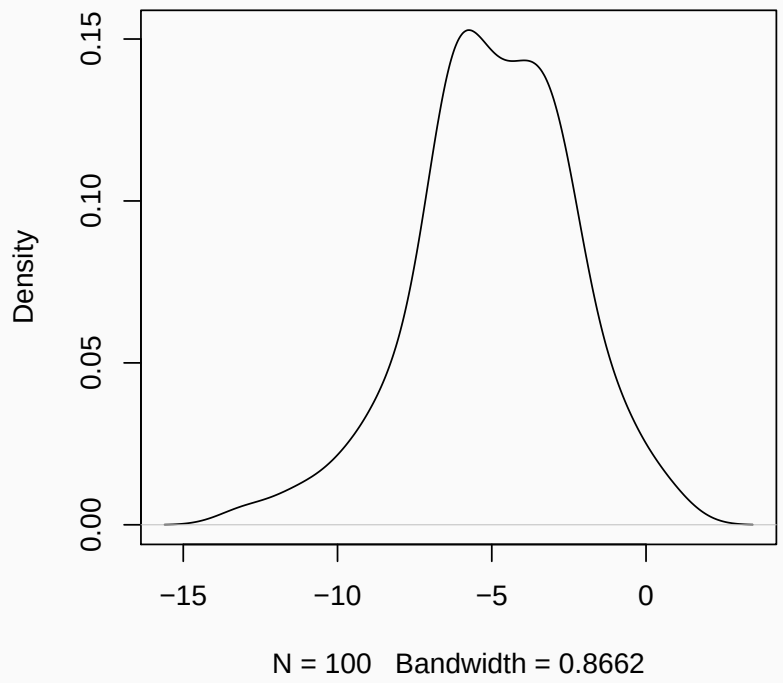
TIPOS DE GRÁFICOS - 1 VARIÁVEL:DENSIDADE *KERNEL*

```
plot(density(plot.data$sigma0))
plot(density(plot.data$sigma0, bw = 0.04),main=NA)
plot(density(plot.data$sigma0, bw = 0.08),main=NA)
```



TIPOS DE GRÁFICOS - 1 VARIÁVEL:DENSIDADE *KERNEL*

```
plot(density(plot.data$sigma0))
plot(density(plot.data$sigma0, bw = 0.04),main=NA)
plot(density(plot.data$sigma0, bw = 0.08),main=NA)
plot(density(10*log10(plot.data$sigma0)),main=NA)
```



TIPOS DE GRÁFICOS - 1 VARIÁVEL:DENSIDADE *KERNEL*

```
plot(density(plot.data$sigma0))
plot(density(plot.data$sigma0, bw = 0.04),main=NA)
plot(density(plot.data$sigma0, bw = 0.08),main=NA)
plot(density(10*log10(plot.data$sigma0)),main=NA)
plot(density(10*log10(plot.data$sigma0), bw=1),main=NA)
```

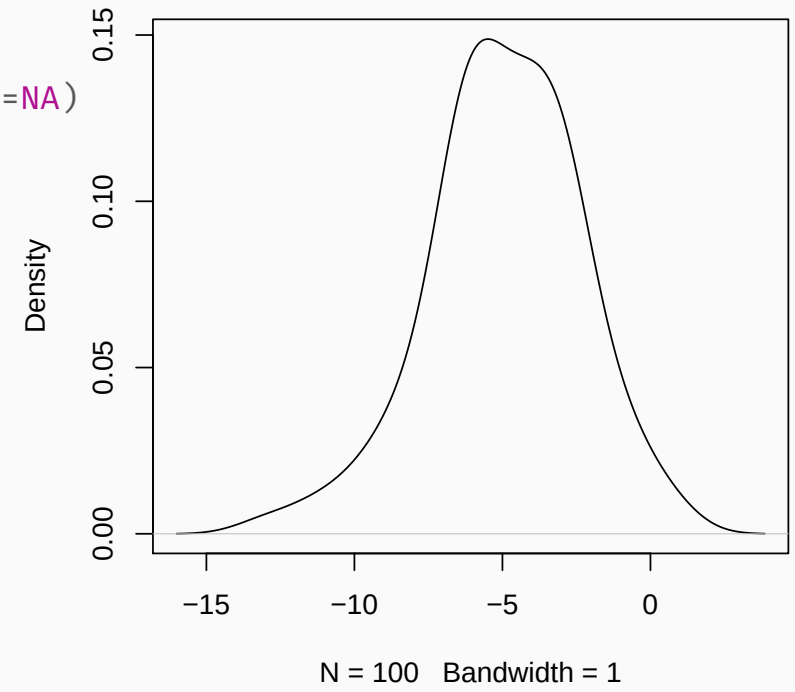


Gráfico de Barras

- Adequado para mostrar proporções, especialmente apropriado para contagens de variáveis categóricas

Gráfico de Barras

- Adequado para mostrar proporções, especialmente apropriado para contagens de variáveis categóricas

- Pode ser mostrado lado a lado ou empilhado

Gráfico de Barras

- Adequado para mostrar proporções, especialmente apropriado para contagens de variáveis categóricas
- Pode ser mostrado lado a lado ou empilhado
- Transmite a impressão de um dado **cumulativo**

Gráfico de Barras

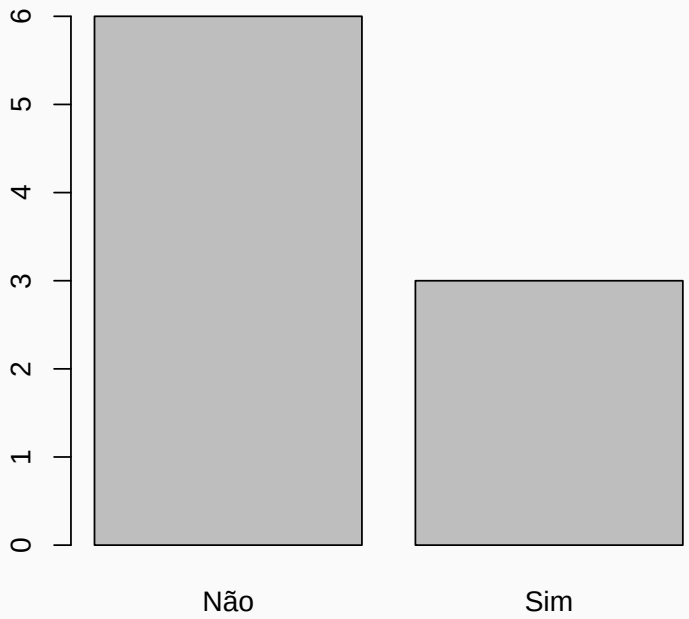
- Adequado para mostrar proporções, especialmente apropriado para contagens de variáveis categóricas
- Pode ser mostrado lado a lado ou empilhado
- Transmite a impressão de um dado **cumulativo**
- Não é recomendado para valores pontuais (ex: média)

TIPOS DE GRÁFICOS - 1 VARIÁVEL: GRÁFICO DE BARRAS

```
stats <- factor(c("Sim","Não","Não","Não","Não",
                  "Não","Não","Sim","Sim"))
summ <- table(stats)
summ
```

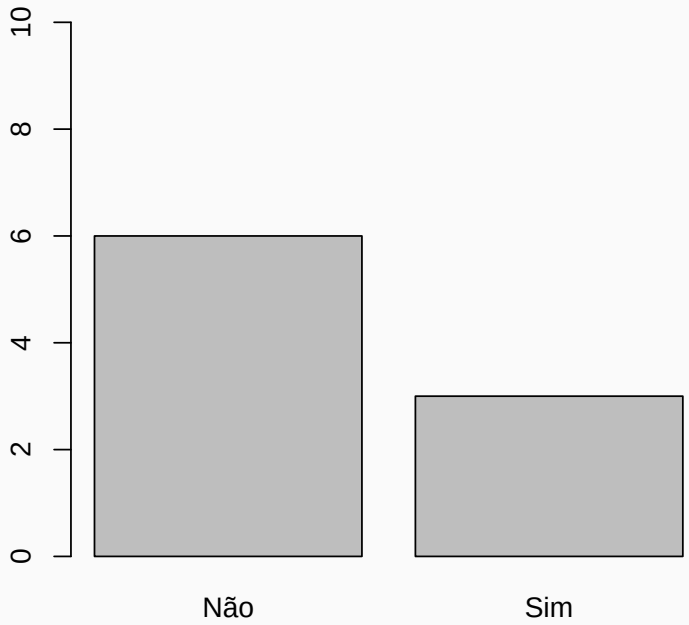
```
## stats
## Não Sim
##   6   3
```

```
barplot(summ)
```



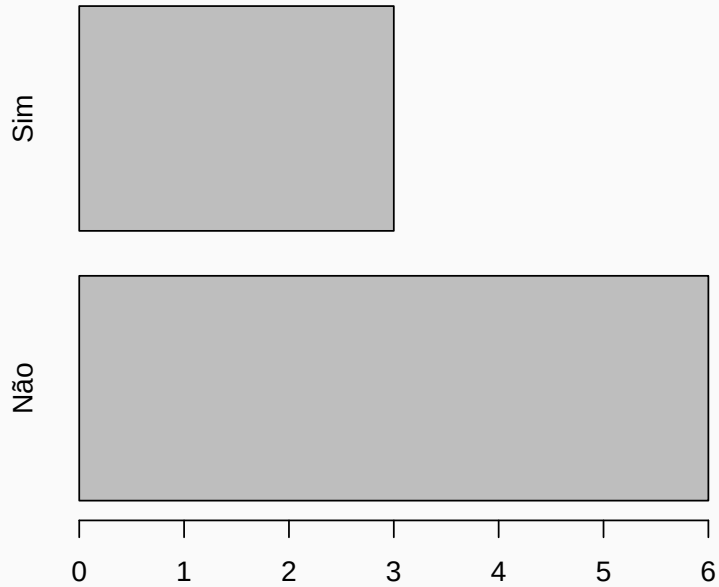
TIPOS DE GRÁFICOS - 1 VARIÁVEL: GRÁFICO DE BARRAS

```
stats <- factor(c("Sim","Não","Não","Não","Não",
                  "Não","Não","Sim","Sim"))
summ <- table(stats)
summ
barplot(summ)
barplot(summ,ylim=c(0,10))
```



TIPOS DE GRÁFICOS - 1 VARIÁVEL: GRÁFICO DE BARRAS

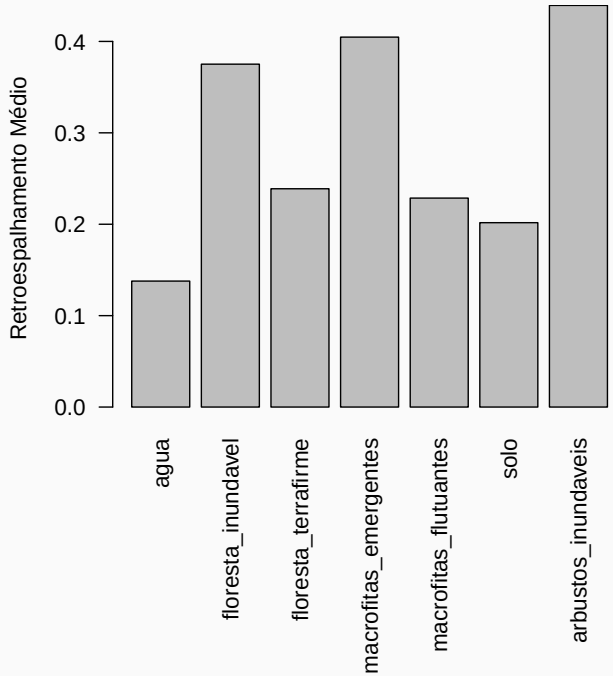
```
stats <- factor(c("Sim","Não","Não","Não","Não",
                  "Não","Não","Sim","Sim"))
summ <- table(stats)
summ
barplot(summ)
barplot(summ,ylim=c(0,10))
barplot(summ,horiz=T)
```



TIPOS DE GRÁFICOS - 1 VARIÁVEL: GRÁFICO DE BARRAS

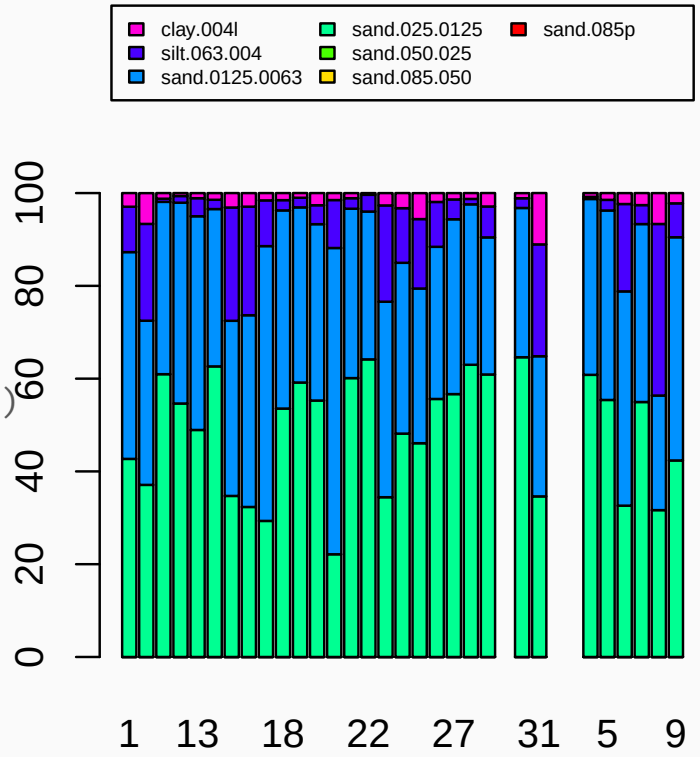
- Inapropriado, pois as médias são valores pontuais, e centrais.

```
barplot(mean.bs,las=2,ylab="Retroespalhamento Médio")
```



TIPOS DE GRÁFICOS - 1+ VARIÁVEL: GRÁFICO DE BARRAS

```
load('rich+env_jun.Rdata')
granulo <- rich.env.jun[,34:40]
granulo <- t(as.matrix(granulo))
barplot(granulo,col=rainbow(7),legend=T,#
   args.legend = list(x="top", inset=c(0,-0.7),ncol=3))
```



Tipos de gráficos - 1 variável: gráfico de pizza

- Gráficos de pizza servem para...

Tipos de gráficos - 1 variável: gráfico de pizza

- Gráficos de pizza servem para...
- ...nada!
- Nosso cérebro é muito mais apto em julgar distâncias do que áreas ou ângulos.

Tipos de gráficos - 1 variável: gráfico de pizza

- Gráficos de pizza servem para...
- ...nada!
- Nosso cérebro é muito mais apto em julgar distâncias do que áreas ou ângulos.
- A partir de hoje, podem abolir gráficos de pizza do seu repertório.

Diagrama de Dispersão (*scatterplot*)

- Um dos gráficos mais úteis em estatística ...

Diagrama de Dispersão (*scatterplot*)

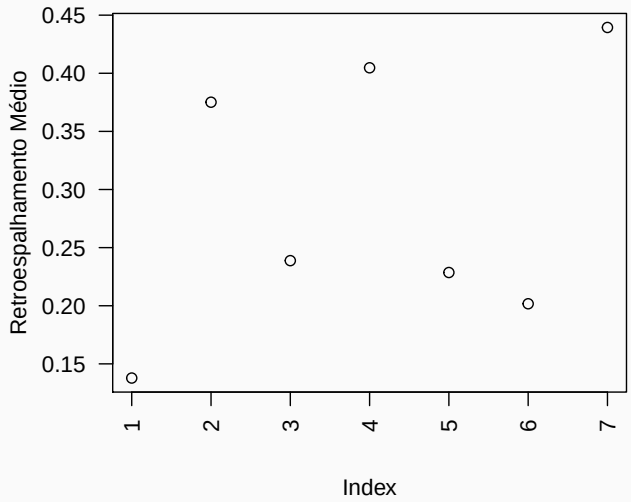
- Um dos gráficos mais úteis em estatística ...
- Pode servir para visualizar duas variáveis contínuas, ou uma variável contínua vs. uma categórica
  - Desde que a variável categórica seja codificada
- Pode ser complementado por barras de erro

Diagrama de Dispersão (*scatterplot*)

- Um dos gráficos mais úteis em estatística ...
- Pode servir para visualizar duas variáveis contínuas, ou uma variável contínua vs. uma categórica
  - Desde que a variável categórica seja codificada
- Pode ser complementado por barras de erro
- Cuidado ao unir os pontos com linhas, pois isso passa uma noção de continuidade!

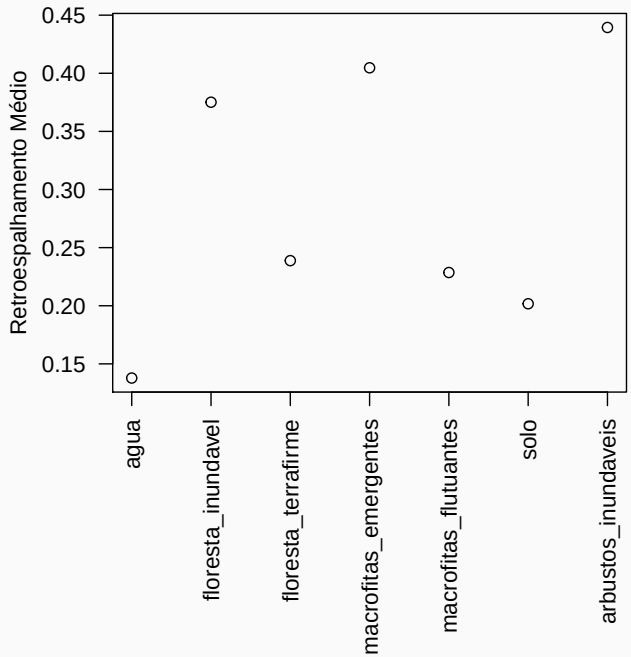
TIPOS DE GRÁFICOS - 2 VARIÁVEIS: DISPERSÃO (*SCATTERPLOT*)

```
plot(mean.bs,las=2,
     ylab="Retroespalhamento Médio",type="p")
```



TIPOS DE GRÁFICOS - 2 VARIÁVEIS: DISPERSÃO (*SCATTERPLOT*)

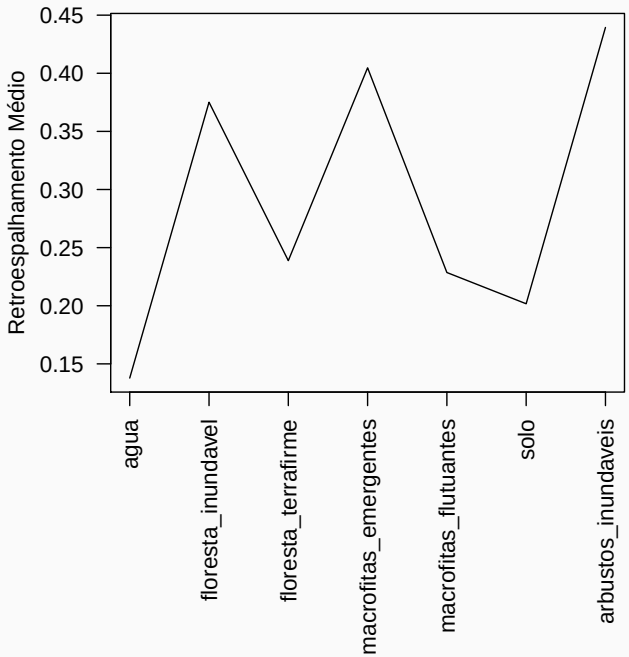
```
plot(mean.bs,las=2,ylab="Retroespalhamento Médio",#
     type="p",xaxt="n",xlab=NA)
axis(1,c(1:7),labels=names(mean.bs),las=2)
```



TIPOS DE GRÁFICOS - 2 VARIÁVEIS: DISPERSÃO (*SCATTERPLOT*)

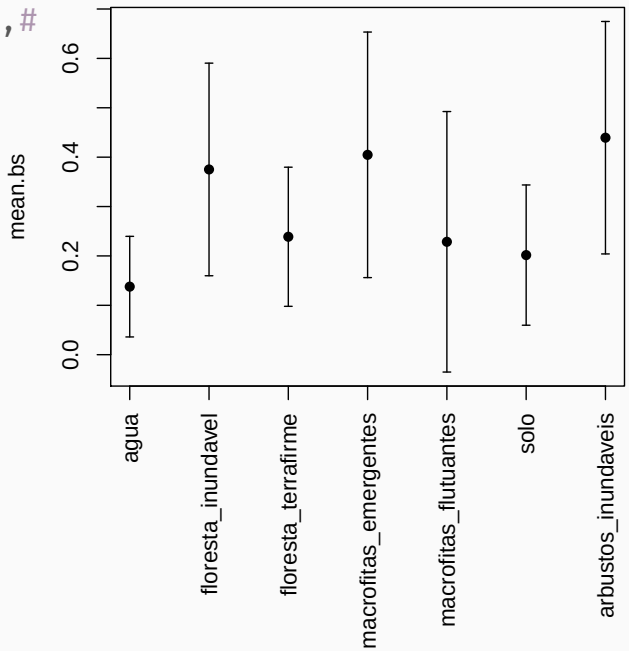
Incorreto, não existe continuidade entre as categorias.

```
plot(mean.bs,las=2,ylab="Retroespalhamento Médio",#
     type="l",xaxt="n",xlab=NA)
axis(1,c(1:7),labels=names(mean.bs),las=2)
```



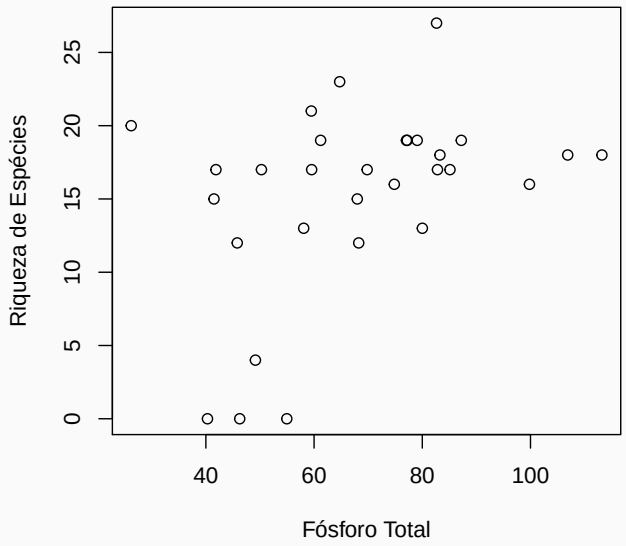
TIPOS DE GRÁFICOS - 2 VARIÁVEIS: DISPERSÃO (*SCATTERPLOT*)

```
plot(mean.bs,las=2,ylab="Retroespalhamento Médio",type="l",#
     xaxt="n",xlab=NA)
axis(1,c(1:7),labels=names(mean.bs),las=2)
library(Hmisc)
errbar(c(1:7),mean.bs,yplus=mean.bs+sd.bs,#
       yminus=mean.bs-sd.bs,lty=1,xaxt="n",xlab=NA)
axis(1,c(1:7),labels=names(mean.bs),las=2)
```



TIPOS DE GRÁFICOS - 2 VARIÁVEIS: DISPERSÃO (*SCATTERPLOT*)

```
load('rich+env_jun.Rdata')
plot(rich.env.jun$p.tot,rich.env.jun$rich,
    xlab="Fósforo Total", ylab = "Riqueza de Espécies")
# A sintaxe de "em função de" (~) também pode ser usada
plot(rich ~ p.tot, data=rich.env.jun,
    xlab="Fósforo Total", ylab = "Riqueza de Espécies")
```



Área

- Pode ser visto como uma versão contínua do gráfico de barras ...

Área

- Pode ser visto como uma versão contínua do gráfico de barras ...
- Mostra diferenças ponto a ponto, e cumulativas
- Não deve ser usado se a área sob a curva não fizer sentido para os dados plotados

Área

- Pode ser visto como uma versão contínua do gráfico de barras ...
- Mostra diferenças ponto a ponto, e cumulativas
- Não deve ser usado se a área sob a curva não fizer sentido para os dados plotados
- A ordem do empilhamento pode afetar a percepção

Área

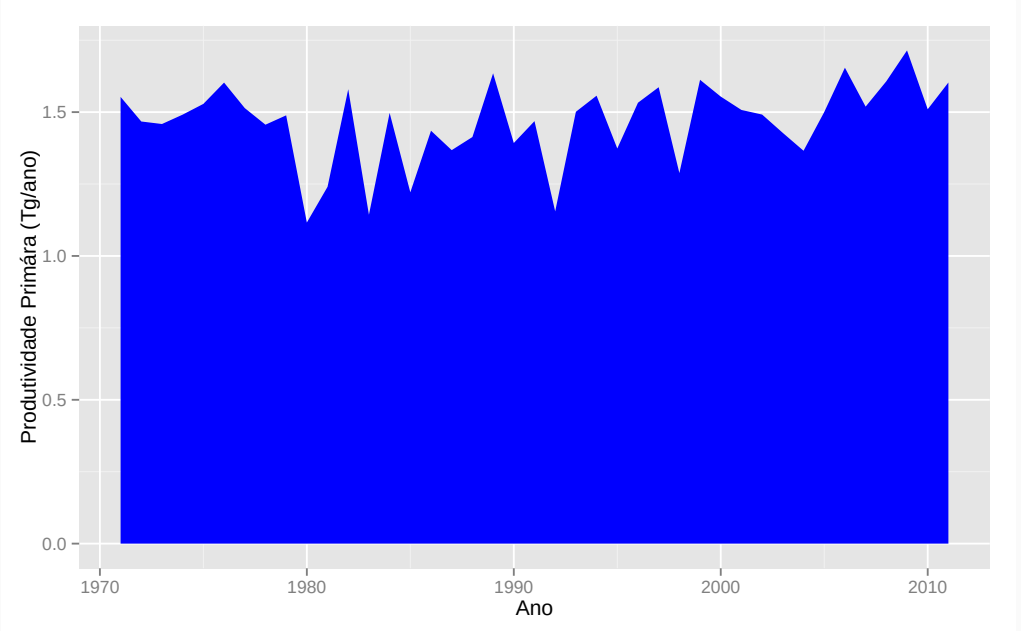
- Pode ser visto como uma versão contínua do gráfico de barras ...
- Mostra diferenças ponto a ponto, e cumulativas
- Não deve ser usado se a área sob a curva não fizer sentido para os dados plotados
- A ordem do empilhamento pode afetar a percepção
- se a variável x não for contínua, melhor usar barras empilhadas

Área

- Pode ser visto como uma versão contínua do gráfico de barras ...
- Mostra diferenças ponto a ponto, e cumulativas
- Não deve ser usado se a área sob a curva não fizer sentido para os dados plotados
- A ordem do empilhamento pode afetar a percepção
- se a variável x não for contínua, melhor usar barras empilhadas

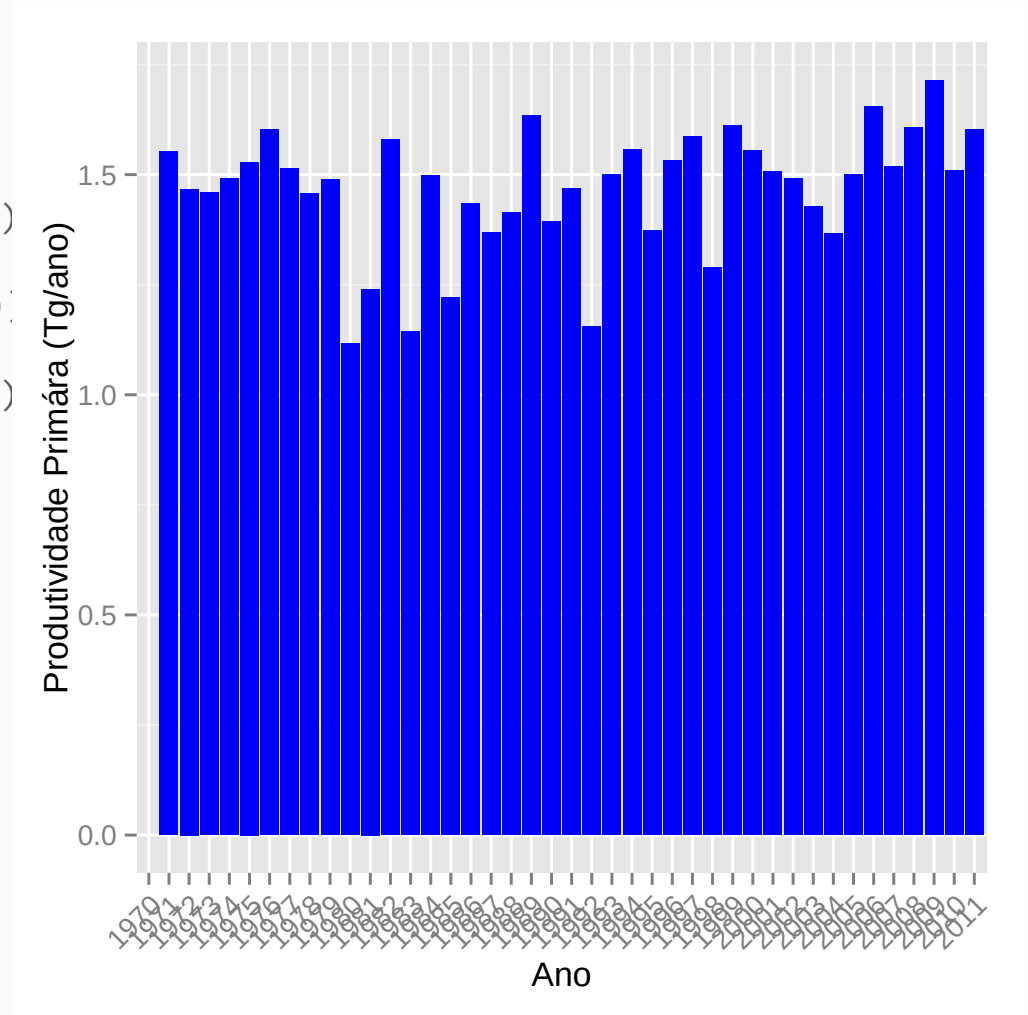
TIPOS DE GRÁFICOS - 2 VARIÁVEIS: GRÁFICO DE ÁREA

```
load('npp_summary.Rdata')
library(ggplot2)
ggplot(npp.df,aes(year,mean)) +#
  geom_area(fill='gray50') +#
  xlab("Ano") +#
 ylab("Produtividade Primára (Tg/ano)")
```



TIPOS DE GRÁFICOS - 2 VARIÁVEIS: GRÁFICO DE ÁREA

```
library(ggplot2)
ggplot(npp.df,aes(year,mean)) + #
  geom_area(fill='blue') +#
  xlab("Ano") + #
 ylab("Produtividade Primária (Tg/ano)")

ggplot(npp.df,aes(as.factor(year),mean)
 geom_bar(fill='blue') +xlab("Ano") +
 ylab("Produtividade Primária (Tg/ano)")
      theme(axis.text.x   =   ele-
ment_text(angle = 90, hjust = 1))
```



Tipos de gráficos - 2 variáveis: gráfico de área

http://www.leancrew.com/all-this/2011/11/i-hate-stacked-area-charts/

Boxplot

- Usado para combinações entre variáveis contínuas e categóricas ...

Boxplot

- Usado para combinações entre variáveis contínuas e categóricas ...
- Na opinião de muitos, um dos gráficos mais informativos que existem ...

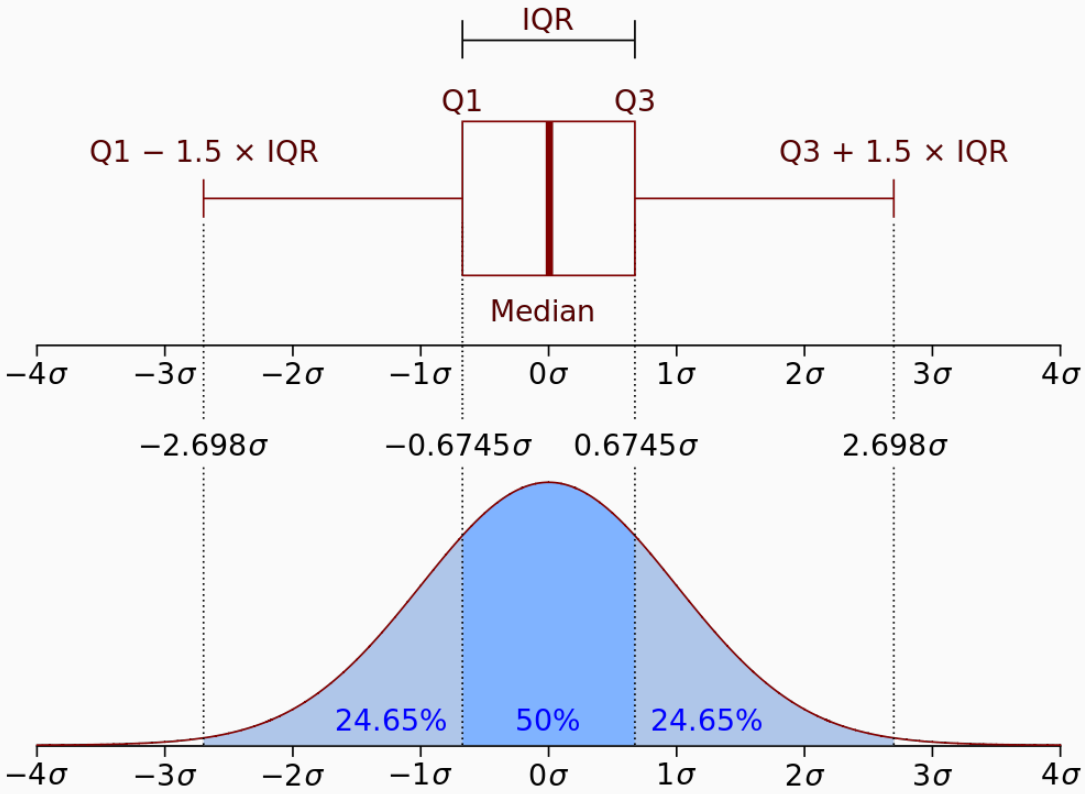
Boxplot

- Usado para combinações entre variáveis contínuas e categóricas ...
- Na opinião de muitos, um dos gráficos mais informativos que existem ...
- Combina as propriedades de um histograma e de um scatterplot

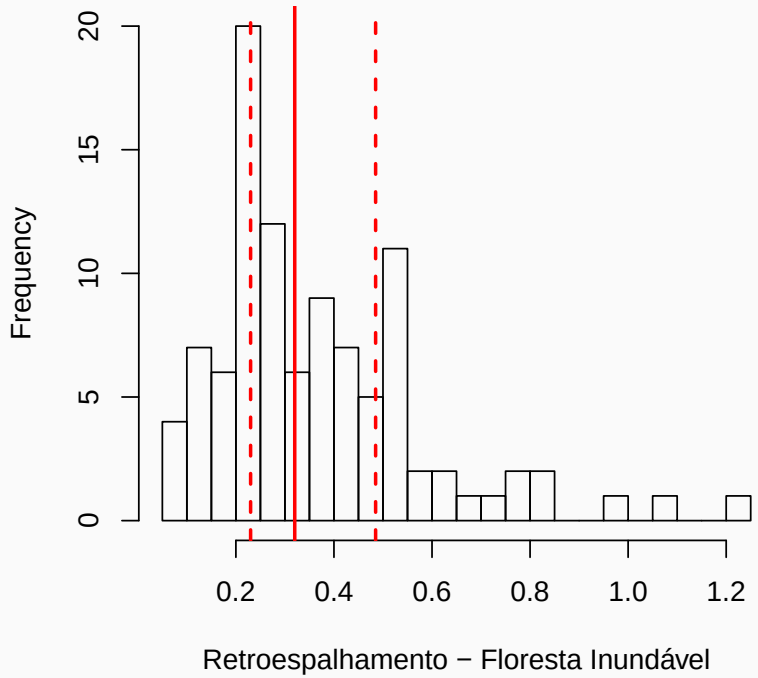
Boxplot

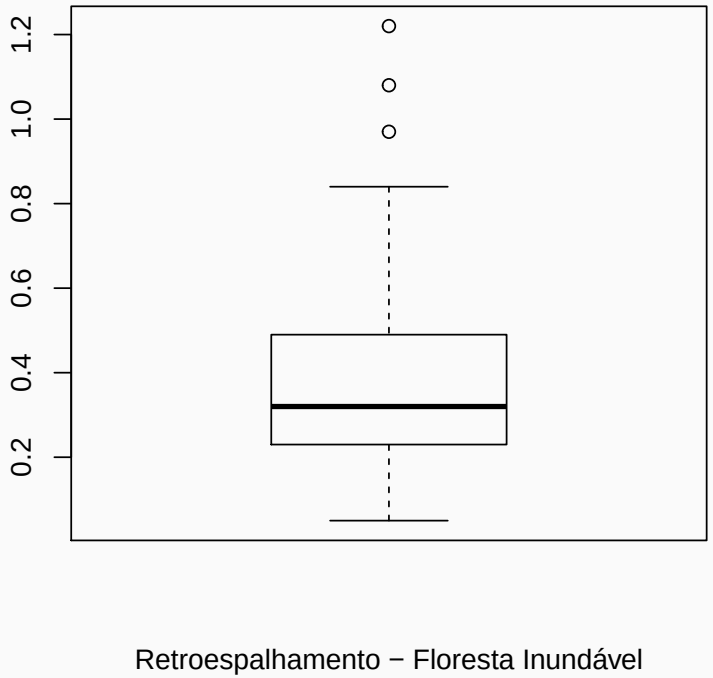
- Usado para combinações entre variáveis contínuas e categóricas ...
- Na opinião de muitos, um dos gráficos mais informativos que existem ...
- Combina as propriedades de um histograma e de um scatterplot
- Faz uso dos quantis para uma descrição robusta dos dados

TIPOS DE GRÁFICOS - 2 VARIÁVEIS: *BOXPLOT*



TIPOS DE GRÁFICOS - 2 VARIÁVEIS: *BOXPLOT*

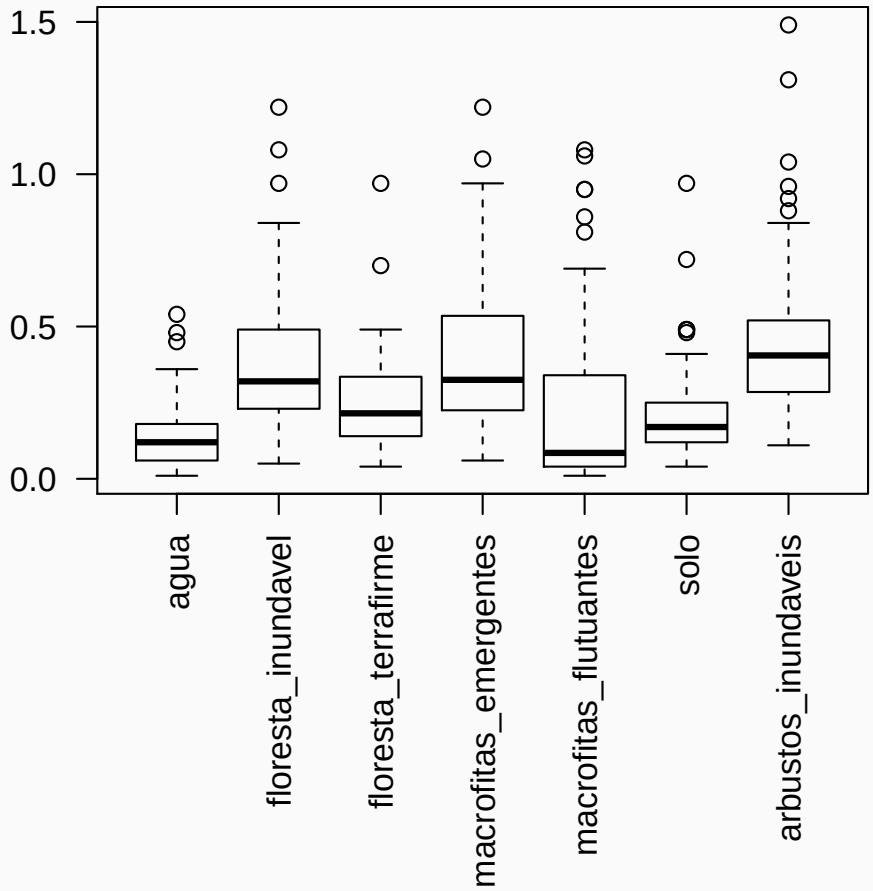




TIPOS DE GRÁFICOS - 2 VARIÁVEIS: *BOXPLOT*

```r
boxplot(sigma0 ~ classe, data=plot.data, las=2)

#linha central: mediana
#
#caixa : quartis
#
#linhas verticais: valor mais alto/baixo dentro da
# distância quartil+/-1.5*distancia interquartil
#
# pontos: outliers, tudo que for maior
# do que quartil +/- 1.5 quartil
```

Superposição da distância interquartil é um indício de diferença/separabilidade



TIPOS DE GRÁFICOS - 2 VARIÁVEIS: *VIOLIN PLOT*

- Tentativa de ir além do boxplot ...
- Combina as propriedades de um gráfico de densidades e de um scatterplot
- Pode ficar estranho se as distribuições não forem bem-comportadas

TIPOS DE GRÁFICOS - 2 VARIÁVEIS: *VIOLIN PLOT*

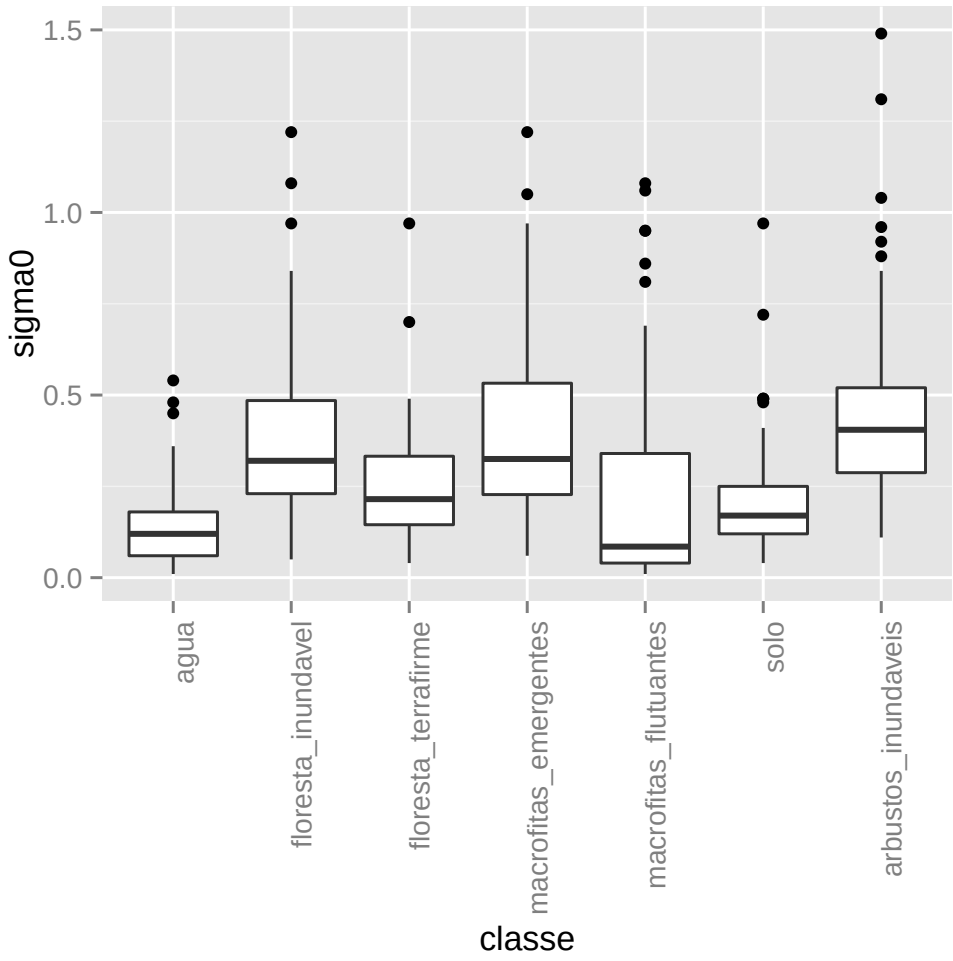
- Tentativa de ir além do boxplot ...
- Combina as propriedades de um gráfico de densidades e de um scatterplot
- Pode ficar estranho se as distribuições não forem bem-comportadas

TIPOS DE GRÁFICOS - 2 VARIÁVEIS: *VIOLIN PLOT*

- Tentativa de ir além do boxplot ...
- Combina as propriedades de um gráfico de densidades e de um scatterplot
- Pode ficar estranho se as distribuições não forem bem-comportadas

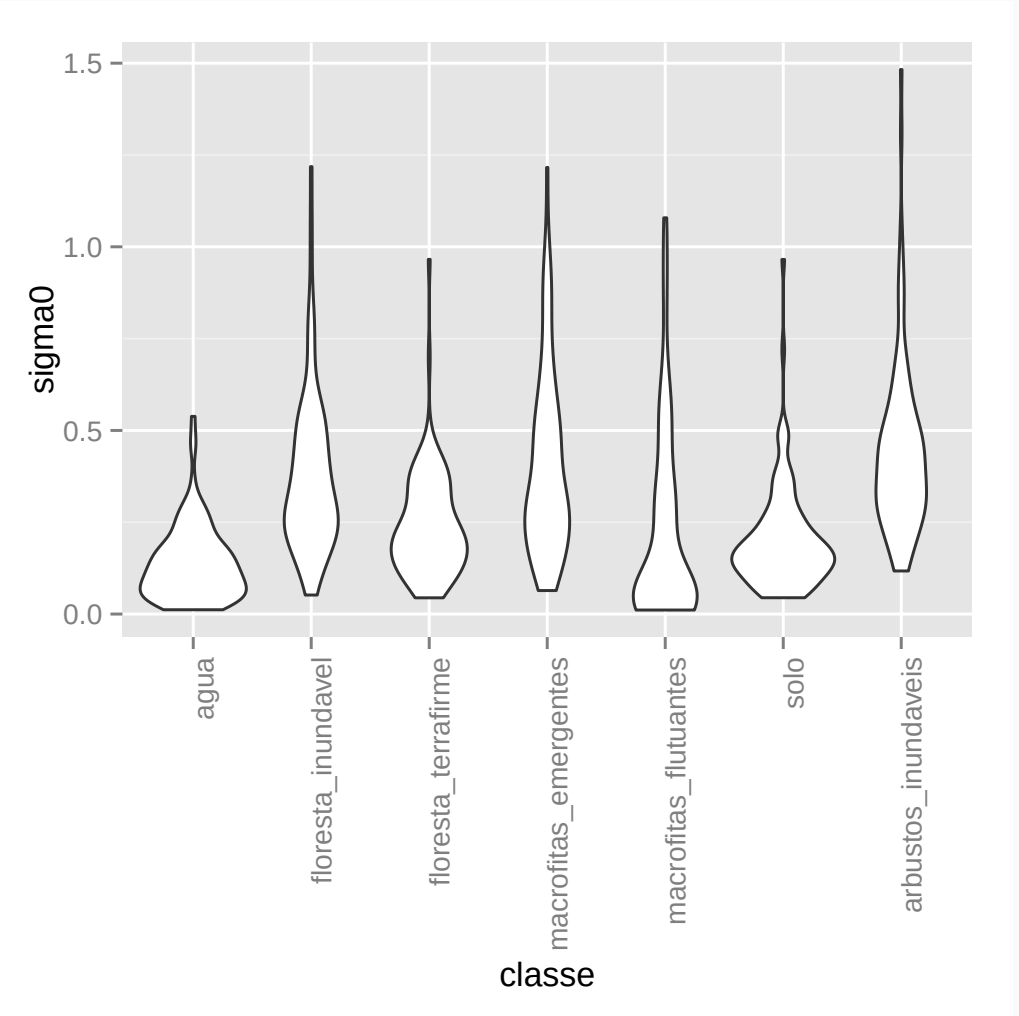
TIPOS DE GRÁFICOS - 2 VARIÁVEIS: *VIOLIN PLOT*

```
ggplot(plot.data,aes(classe,sigma0)) +#
  geom_boxplot() + theme(axis.text.x = ele-
ment_text(angle = 90, hjust = 1))
```



TIPOS DE GRÁFICOS - 2 VARIÁVEIS: *VIOLIN PLOT*

```
ggplot(plot.data,aes(classe,sigma0)) +#
  geom_boxplot() +#
          theme(axis.text.x      =      ele-
ment_text(angle = 90, hjust = 1))
ggplot(plot.data,aes(classe,sigma0)) +#
  geom_violin() +#
          theme(axis.text.x      =      ele-
ment_text(angle = 90, hjust = 1))
```



COMBINANDO MAIS DE DUAS VARIÁVEIS ...SEM USAR “3D”

- Gráficos “3D” são dependentes de perspectiva, e não enfatizam bem as diferenças
- São uma boa ferramenta de visualização se puderem ser manipulados
- Mas para exibição em papel, dificultam a interpretação
- Ao invés de usar múltiplos eixos, podemos explorar as relações entre cor, forma e tamanho dos objetos plotados

COMBINANDO MAIS DE DUAS VARIÁVEIS ...SEM USAR “3D”

- Gráficos “3D” são dependentes de perspectiva, e não enfatizam bem as diferenças
- São uma boa ferramenta de visualização se puderem ser manipulados
- Mas para exibição em papel, dificultam a interpretação
- Ao invés de usar múltiplos eixos, podemos explorar as relações entre cor, forma e tamanho dos objetos plotados

COMBINANDO MAIS DE DUAS VARIÁVEIS ...SEM USAR “3D”

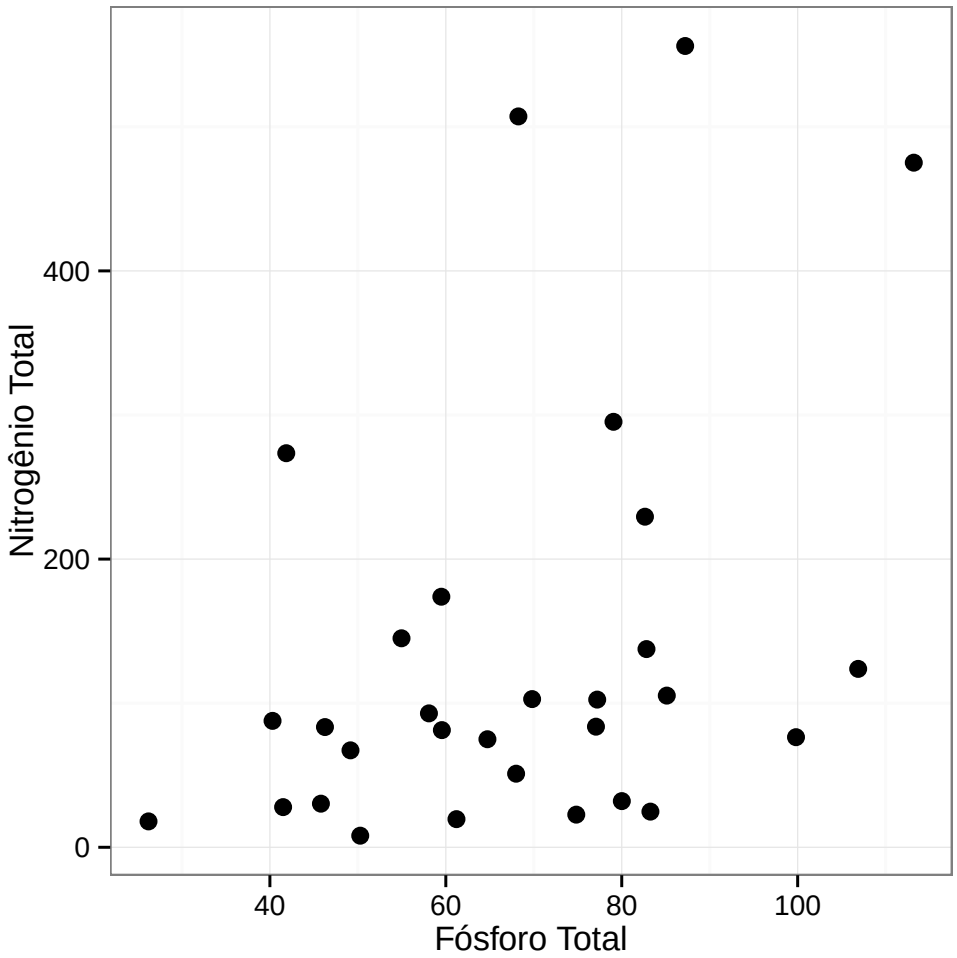
- Gráficos “3D” são dependentes de perspectiva, e não enfatizam bem as diferenças
- São uma boa ferramenta de visualização se puderem ser manipulados
- Mas para exibição em papel, dificultam a interpretação
- Ao invés de usar múltiplos eixos, podemos explorar as relações entre cor, forma e tamanho dos objetos plotados

COMBINANDO MAIS DE DUAS VARIÁVEIS ...SEM USAR “3D”

- Gráficos “3D” são dependentes de perspectiva, e não enfatizam bem as diferenças
- São uma boa ferramenta de visualização se puderem ser manipulados
- Mas para exibição em papel, dificultam a interpretação
- Ao invés de usar múltiplos eixos, podemos explorar as relações entre cor, forma e tamanho dos objetos plotados

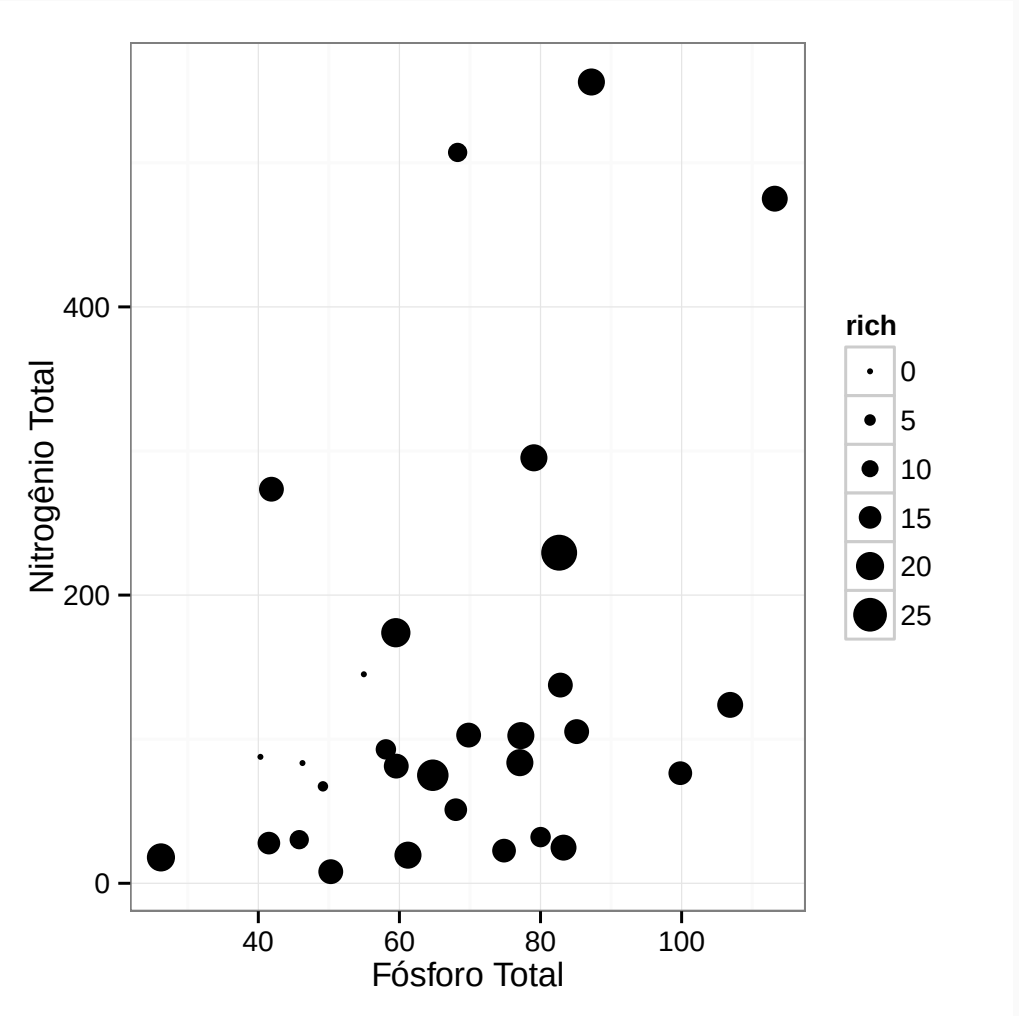
TIPOS DE GRÁFICOS - 2 VARIÁVEIS

```
ggplot(rich.env.jun,aes(p.tot,n.tot)) +#
  geom_point(size=3) +#
  ylab("Nitrogênio Total") +#
  xlab("Fósforo Total") +#
  theme_bw()
```



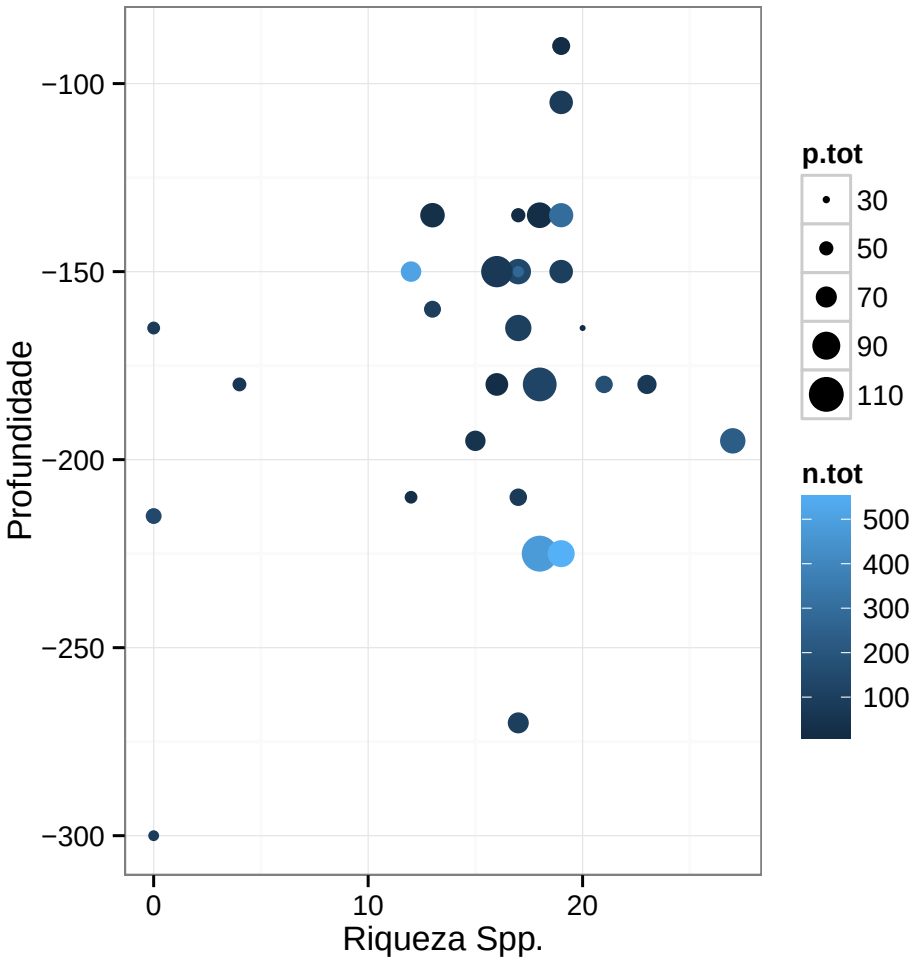
TIPOS DE GRÁFICOS - 3 VARIÁVEIS

```
ggplot(rich.env.jun,aes(p.tot,n.tot)) +#
  geom_point(size=3) +#
  ylab("Nitrogênio Total") +#
  xlab("Fósforo Total") +#
  theme_bw()
ggplot(rich.env.jun,aes(p.tot,n.tot)) +#
  geom_point(aes(size=rich)) +#
  ylab("Nitrogênio Total") +#
  xlab("Fósforo Total") +#
  theme_bw()
```



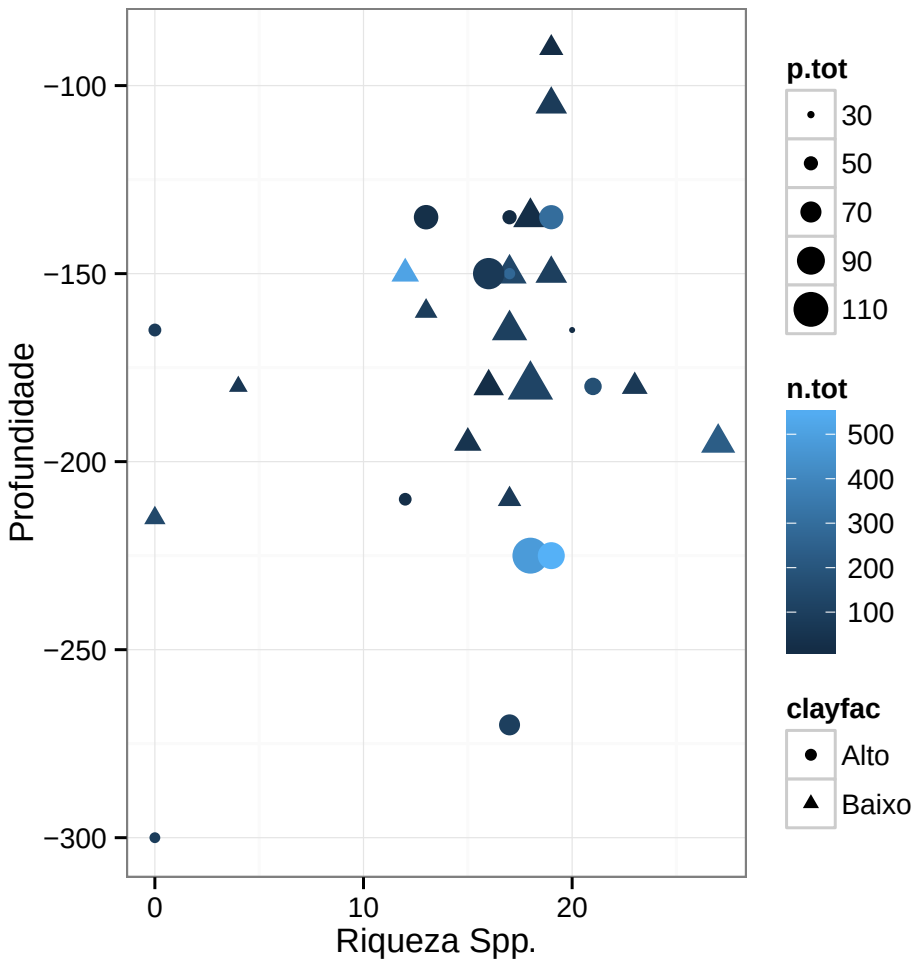
TIPOS DE GRÁFICOS - 4 VARIÁVEIS

```
ggplot(rich.env.jun,aes(p.tot,n.tot)) +#
  geom_point(size=3) +#
  ylab("Nitrogênio Total") +#
  xlab("Fósforo Total") +#
  theme_bw()
ggplot(rich.env.jun,aes(p.tot,n.tot)) +#
  geom_point(aes(size=rich)) +#
  ylab("Nitrogênio Total") +#
  xlab("Fósforo Total") +#
  theme_bw()
ggplot(rich.env.jun,aes(p.tot,n.tot)) +#
  geom_point(aes(color=rich, size=prof*-1)) +#
  ylab("Nitrogênio Total") +#
  xlab("Fósforo Total") +#
  theme_bw()
```



TIPOS DE GRÁFICOS - 5 VARIÁVEIS

```
ggplot(rich.env.jun,aes(p.tot,n.tot)) +#
  geom_point(size=3) +#
  ylab("Nitrogênio Total") +#
  xlab("Fósforo Total") +#
  theme_bw()
ggplot(rich.env.jun,aes(p.tot,n.tot)) +#
  geom_point(aes(size=rich)) +#
  ylab("Nitrogênio Total") +#
  xlab("Fósforo Total") +#
  theme_bw()
ggplot(rich.env.jun,aes(rich,prof)) +#
 geom_point(aes(color=n.tot, size=p.tot)) +#
  ylab("Profundidade") +#
  xlab("Riqueza Spp.") +#
  theme_bw()
ggplot(rich.env.jun,aes(rich,prof)) +#
 geom_point(aes(color=n.tot, size=p.tot, shap
  ylab("Profundidade") +#
  xlab("Riqueza Spp.") +#
  theme_bw()
```



ELEMENTOS DE UM BOM GRÁFICO

- Foco na informação que se quer enfatizar

ELEMENTOS DE UM BOM GRÁFICO

- Foco na informação que se quer enfatizar
- Quanto menor a razão tinta/papel, melhor

ELEMENTOS DE UM BOM GRÁFICO

- Foco na informação que se quer enfatizar
- Quanto menor a razão tinta/papel, melhor
- Selecione e ordene suas variáveis de acordo com a pergunta a ser respondida

ELEMENTOS DE UM BOM GRÁFICO

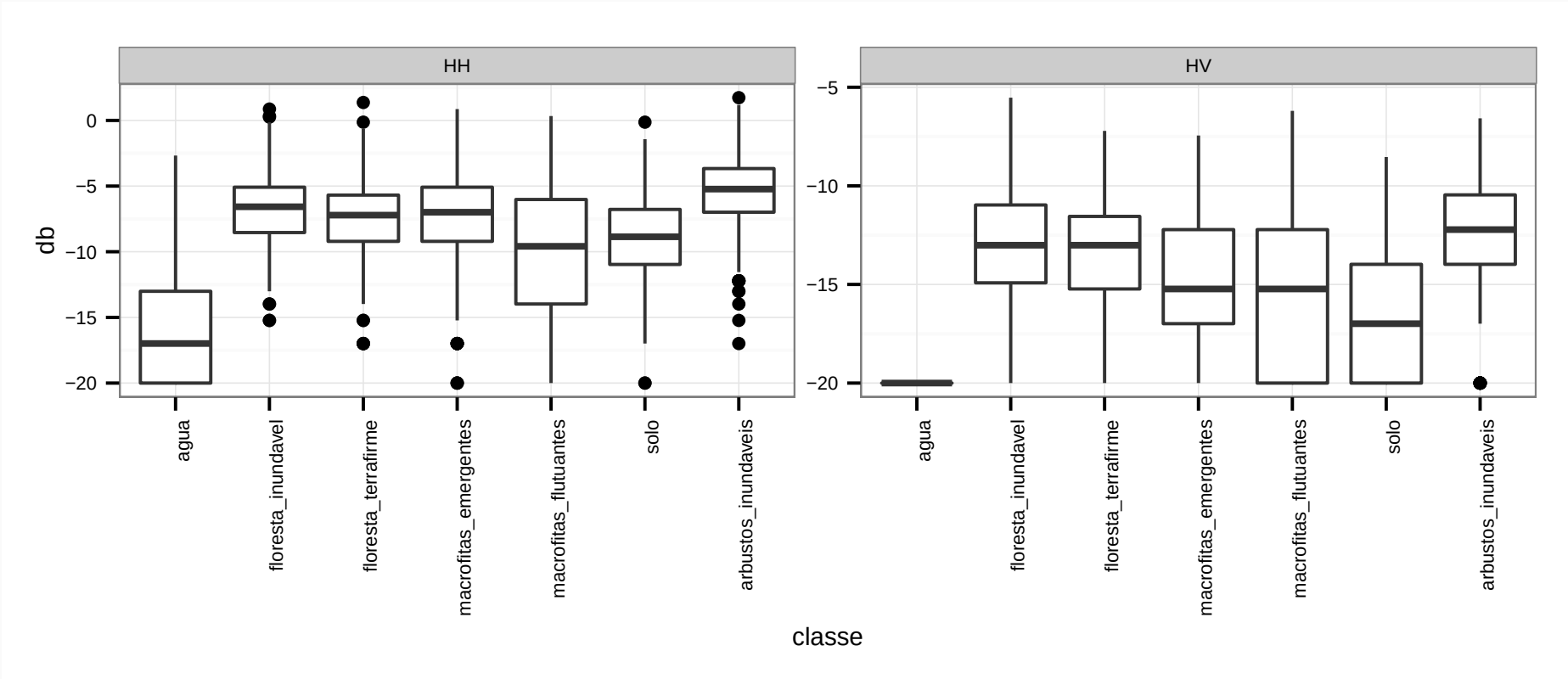
- Foco na informação que se quer enfatizar
- Quanto menor a razão tinta/papel, melhor
- Selecione e ordene suas variáveis de acordo com a pergunta a ser respondida
- Cores e formas só devem ser usadas se também trouxerem informação!

ELEMENTOS DE UM BOM GRÁFICO

- Foco na informação que se quer enfatizar
- Quanto menor a razão tinta/papel, melhor
- Selecione e ordene suas variáveis de acordo com a pergunta a ser respondida
- Cores e formas só devem ser usadas se também trouxerem informação!

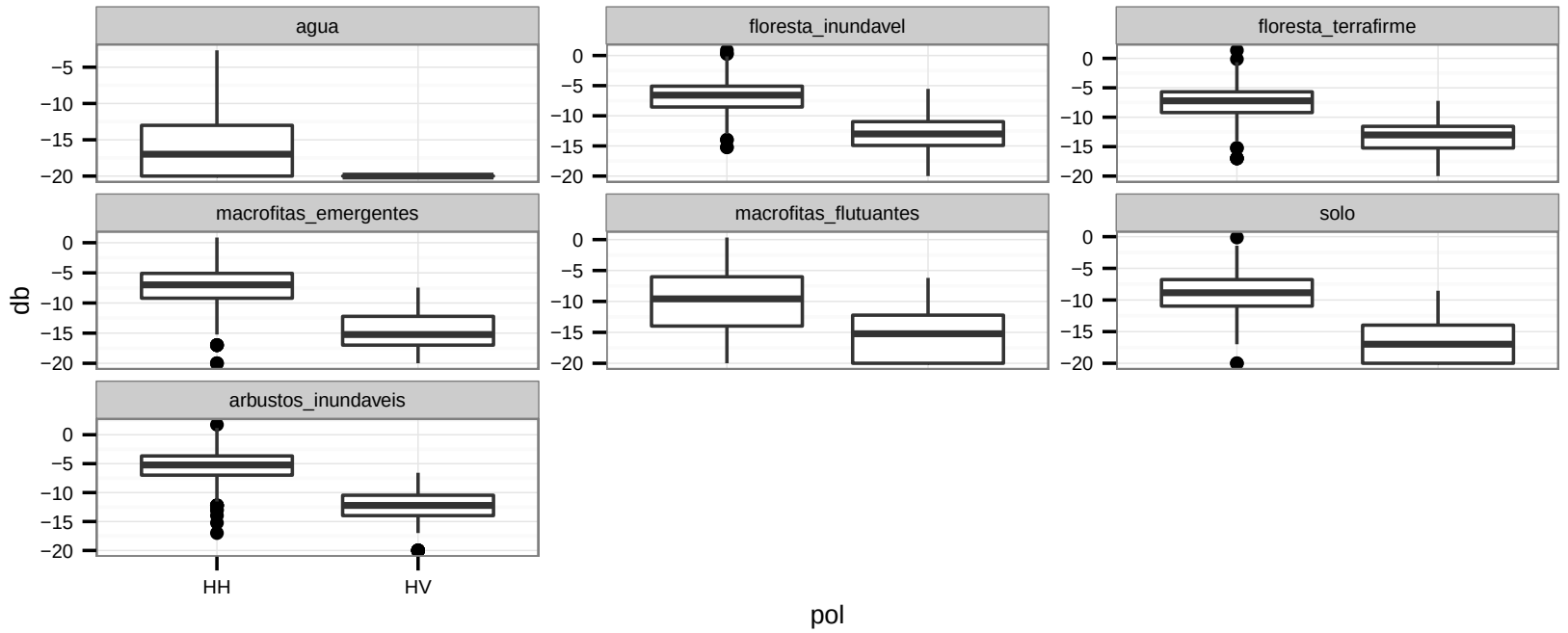
QUAL A PERGUNTA A SER RESPONDIDA?

Diferença entre classes, para cada polarização?



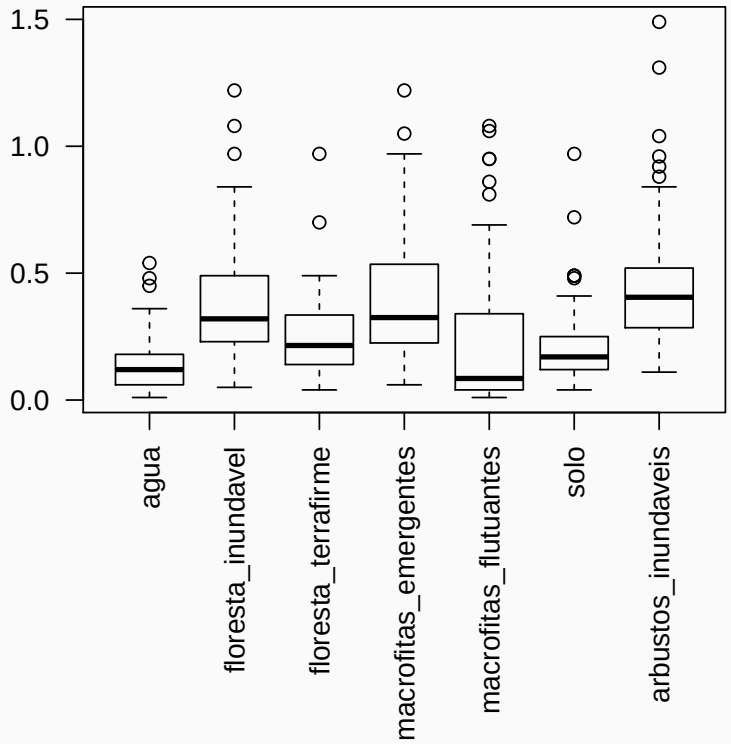
QUAL A PERGUNTA A SER RESPONDIDA?

Ou a diferença entre polarizações, para cada classe?



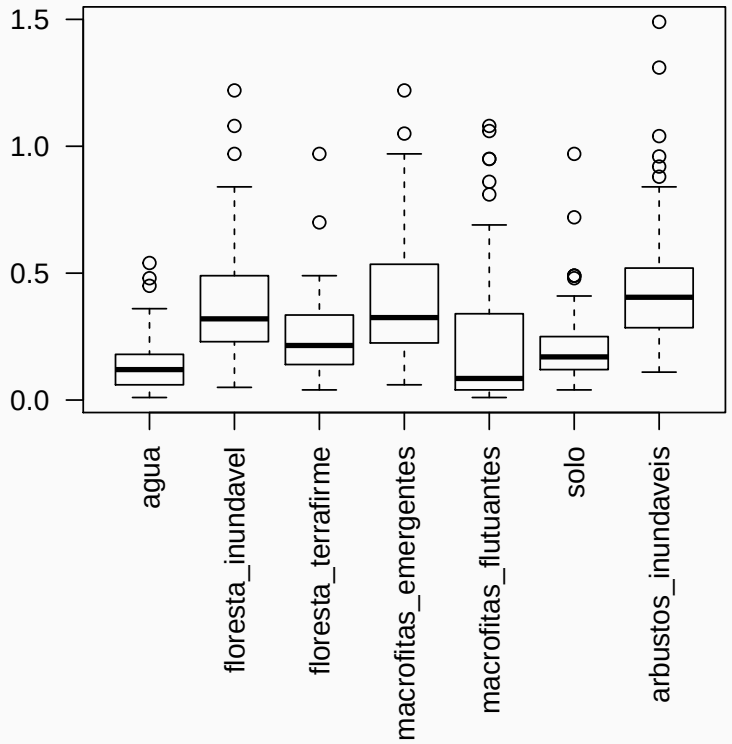
TIPOS DE GRÁFICOS

Isso é melhor ...



TIPOS DE GRÁFICOS

Isso é melhor ...



...do que isso!