

Curso R

Teste de Hipóteses

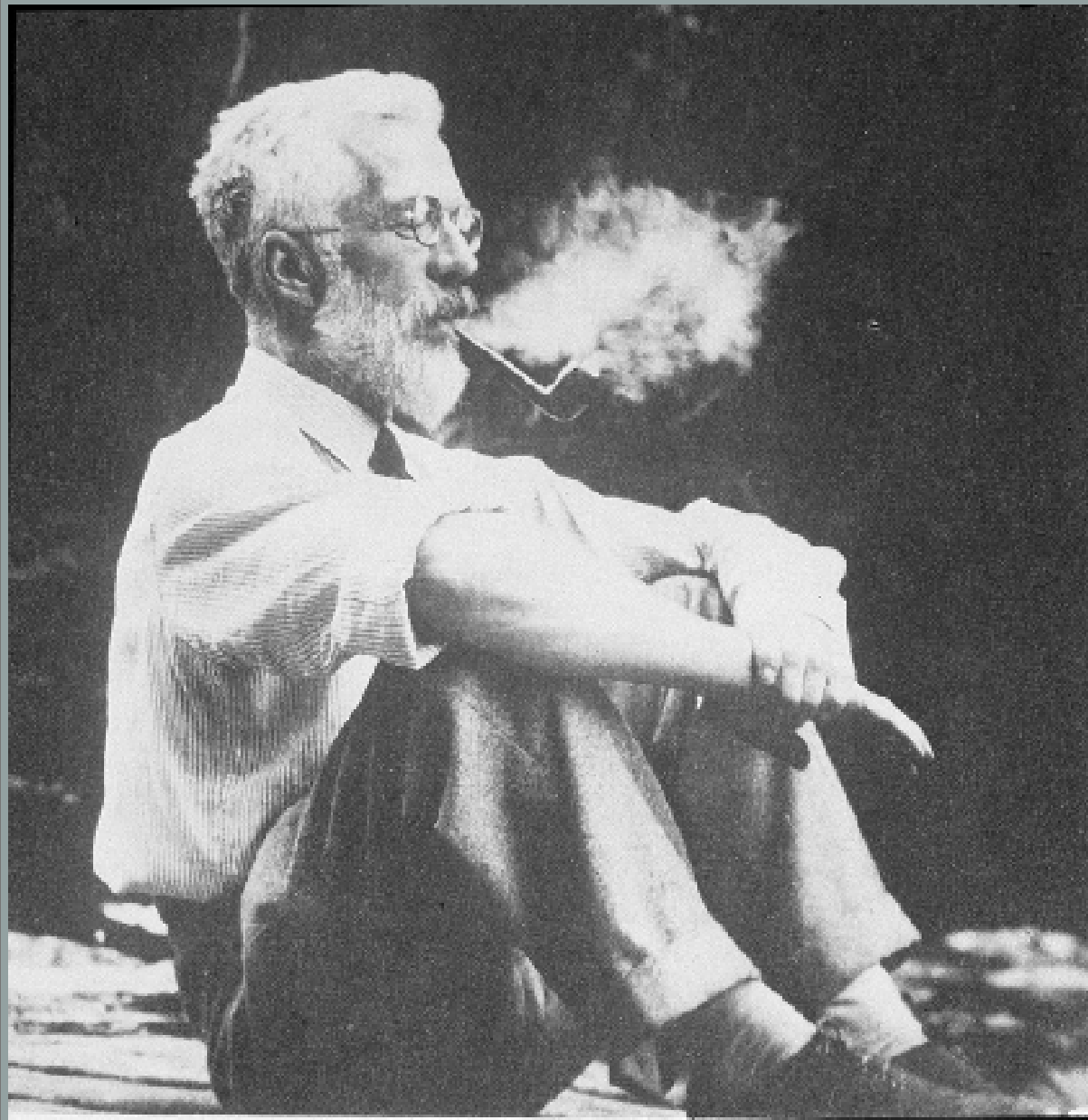
Alexandre Adalardo de Oliveira

Ecologia- IBUSP maio 2017

Uso da Linguagem R : Teste de Hipóteses

Uso da Linguagem R : Teste de Hipóteses

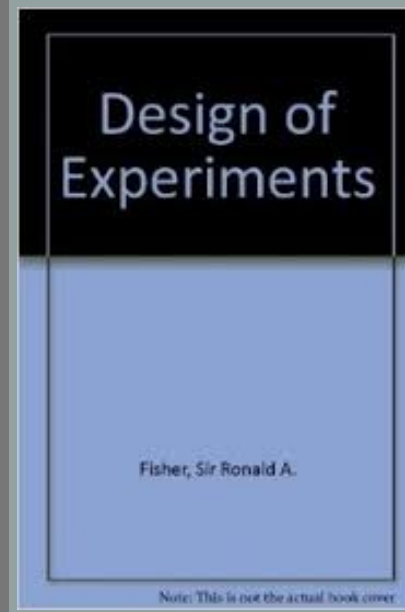
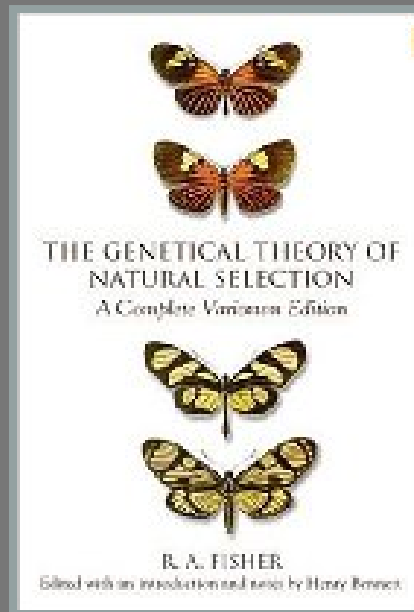
1. a lógica do teste
2. o significado do p-valor
3. a tabela de anova (partição da variância)
4. biologia \leftrightarrow estatística



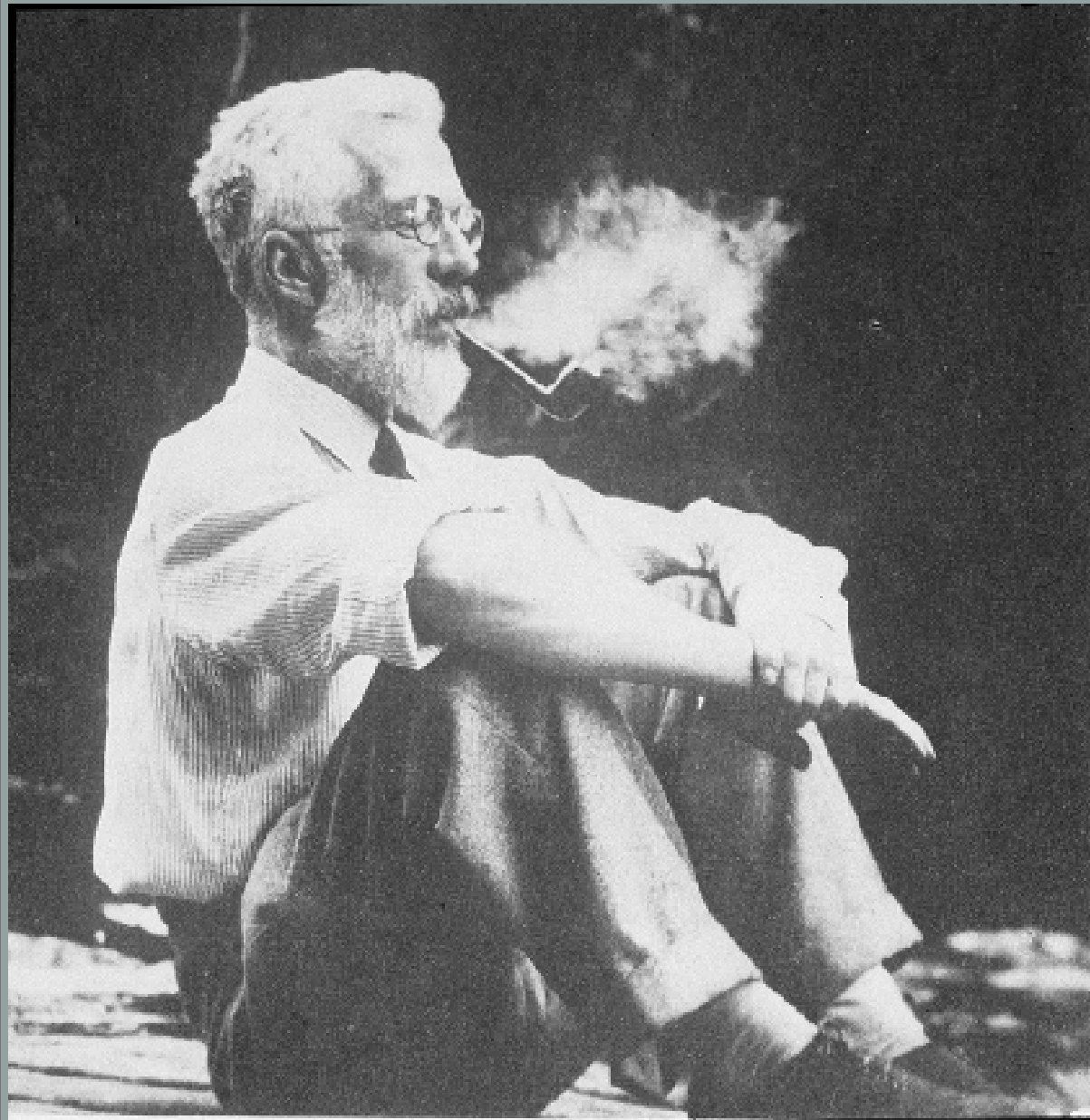
Desenho Experimental



Referência em diferentes áreas



Sir Ronald A. Fisher





Inferência Estatística



Inferência Estatística

“fazer afirmações sobre um universo a partir de um conjunto de valores representativo (amostra). Tal tipo de afirmação deve sempre vir acompanhada de uma medida de precisão sobre sua veracidade”

Inferência Estatística

fazer afirmações sobre um universo a partir de um conjunto de valores representativo (amostra). Tal tipo de afirmação deve sempre vir acompanhada de uma medida de precisão sobre sua veracidade

Amostra de Palmitos



IlhadosCardosoPlot

Image NASA
Image © 2008 TerraMetrics
© 2008 MapLink/Tele Atlas
© 2008 Europa Technologies

©2007 Google

nter 24°56'26.09" S 48°03'38.26" Welev 153 m

Streaming |||||100%

Eye alt 4.1 23.25 km

Palmito (*Euterpe edulis*)



Parque Estadual de Carlos Botelho

```
> load("data/palmcb.RData")
```

```
> ls()
```

```
## [1] "palmcb"
```

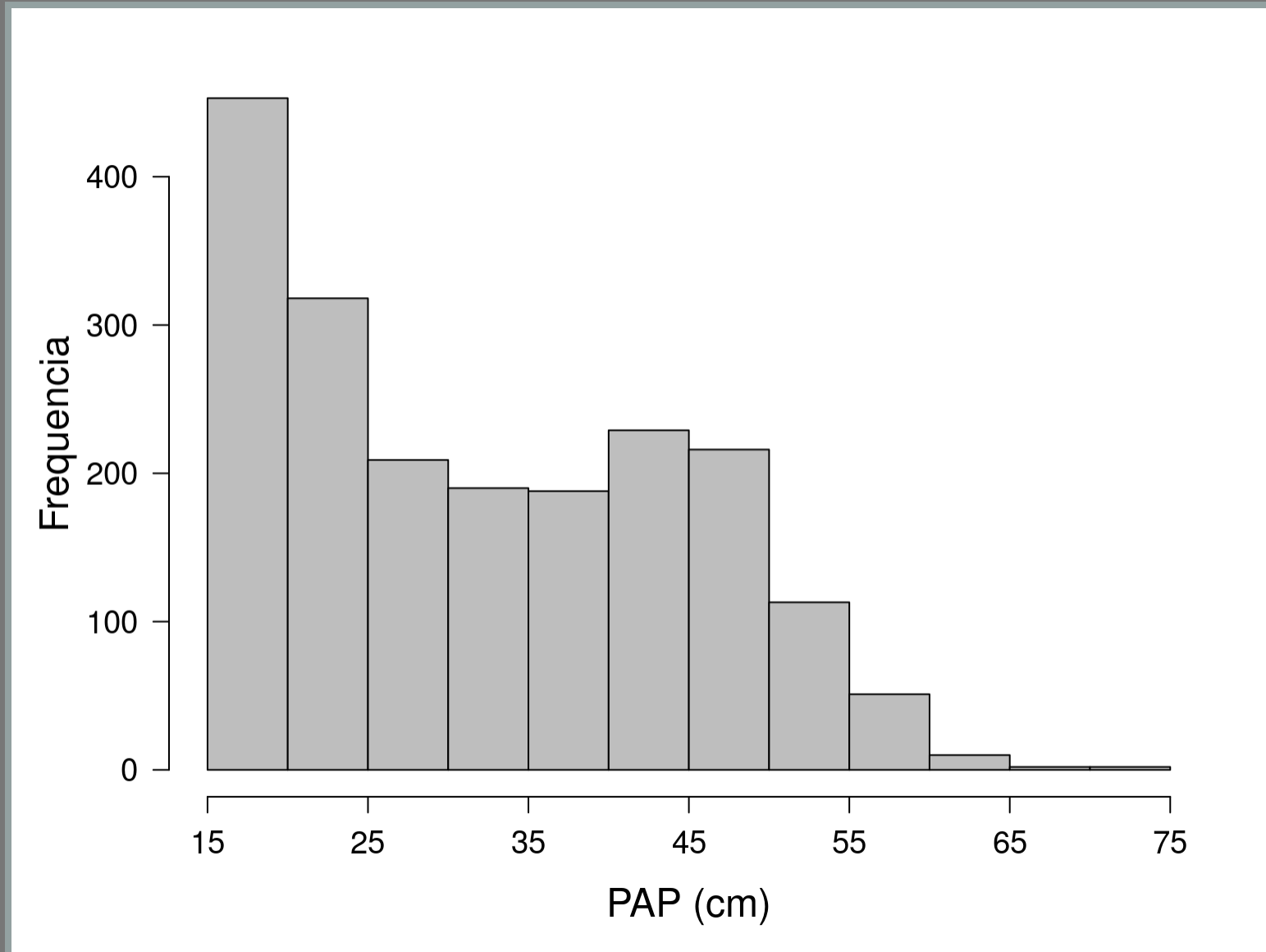
```
> str(palmcb)
```

```
## int [1:1981] 15 15 15 15 15 15 15 15
```

Palmitos de Carlos Botelho

```
> par(mar=c(5,5,2,2), cex.lab=1.5,  
+      cex.axis=1.2, las=1)  
> hist(palmcb, xlab="PAP (cm)",  
+       ylab="Frequencia", xlim=c(15,75),  
+       col="gray", main="", xaxt="n" )  
> axis(1, at=seq(15, 75, by=10))
```


Palmitos de Carlos Botelho



Palmitos de Carlos Botelho

```
> mean(palmcb)
```

```
## [1] 32.60071
```

```
> summary(palmcb)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Q  
##      15.0   21.0   31.0   32.6   43
```

```
> sample(palmcb, size=10)
```

```
## [1] 51 52 22 47 38 32 43 18 39 41
```

```
> mean(sample(palmcb, size=10))
```

```
## [1] 36.3
```

Amostra de Palmitos

```
> mean(sample(palmcb, size=10))
```

```
## [1] 37.9
```

```
> mean(sample(palmcb, size=10))
```

```
## [1] 33.8
```

```
> mean(sample(palmcb, size=10))
```

```
## [1] 38.6
```

```
> mean(sample(palmcb, size=10))
```

```
## [1] 28.9
```

Palmitos de Carlos Botelho

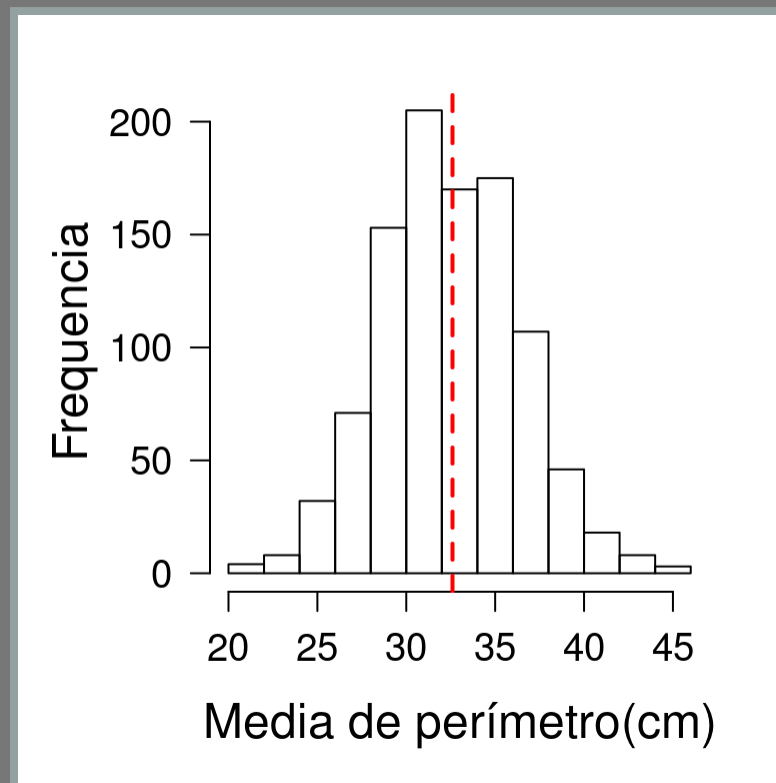
```
> resulta <- rep(NA, 1000)
> for(i in 1:1000)
+ {
+   resulta[i] <- mean(sample(palmcb,
+                             size=10))
+ }

> resulta[1:10]

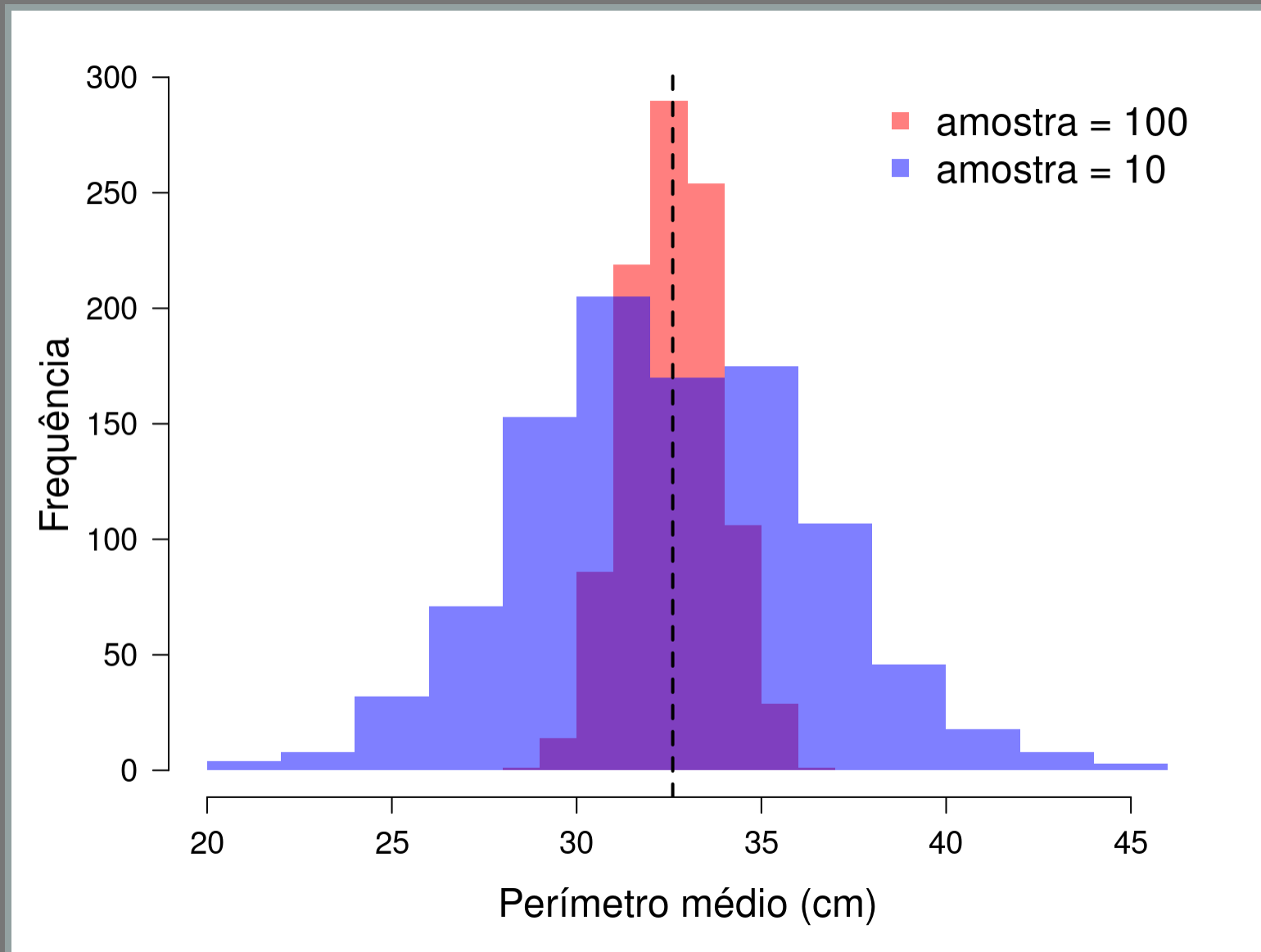
##      [1] 37.5 23.3 27.2 35.6 34.5 34.4 35
```

Palmitos de Carlos Botelho

- > `par(mar=c(5,5,2,2), cex.lab=1.5, cex.a`
- > `hist(resulta, xlab="Media de perímetro`
- > `abline(v=mean(palmcb), lty=2, lwd=2, c`



Tamanho da Amostra



Pergunta

Os palmitos em Carlos Botelho são, em média, maiores que os da restinga da Ilha do Cardoso?



IlhadoCardosoPlot



Image NASA
Image © 2008 TerraMetrics
© 2008 MapLink/Tele Atlas
© 2008 Europa Technologies

©2007 Google

nter 24°56'26.09" S 48°03'38.26" Welev 153 m

Streaming |||||100%

Eye alt 23.25 km

4.11

Hipótese Biológica



Hipótese Biológica

Pergunta:

Os palmitos em Carlos Botelho são maiores que os da restinga da Ilha do Cardoso?

Hipótese:

Palmito em solos mais argilosos e férteis crescem mais que os de solos arenosos e pobres

Predição:

A média do perímetro dos palmitos em Carlos Botelho é maior que a média da restinga da Ilha do Cardoso

Hipótese Biológica

Pergunta:

Os palmitos em Carlos Botelho são maiores que os da restinga da Ilha do Cardoso?

Hipótese:

Palmito em solos mais argilosos e férteis crescem mais que os de solos arenosos e pobres

Predição:

A média do perímetro dos palmitos em Carlos Botelho é maior que a média da restinga da Ilha do Cardoso

Hipótese Biológica

Pergunta:

Os palmitos em Carlos Botelho são maiores que os da restinga da Ilha do Cardoso?

Hipótese:

Palmito em solos mais argilosos e férteis crescem mais que os de solos arenosos e pobres

Predição:

A média do perímetro dos palmitos em Carlos Botelho é maior que a média da restinga da Ilha do Cardoso

Desenho amostral

```
> load("data/eutdata.RData")
> ls(pattern="palm")

## [1] "palmcb" "palmic"

> amoscb <- sample(palmcb, 100)
> amosic <- sample(palmic, 100)
```

Estatística de interesse

```
> mean(amoscb)
```

```
## [1] 33.74
```

```
> mean(amosic)
```

```
## [1] 28.805
```

```
> mean(amoscb) - mean(amosic)
```

```
## [1] 4.935
```

```
> difpalm <- mean(amoscb) - mean(amosic)
```

Afirmação

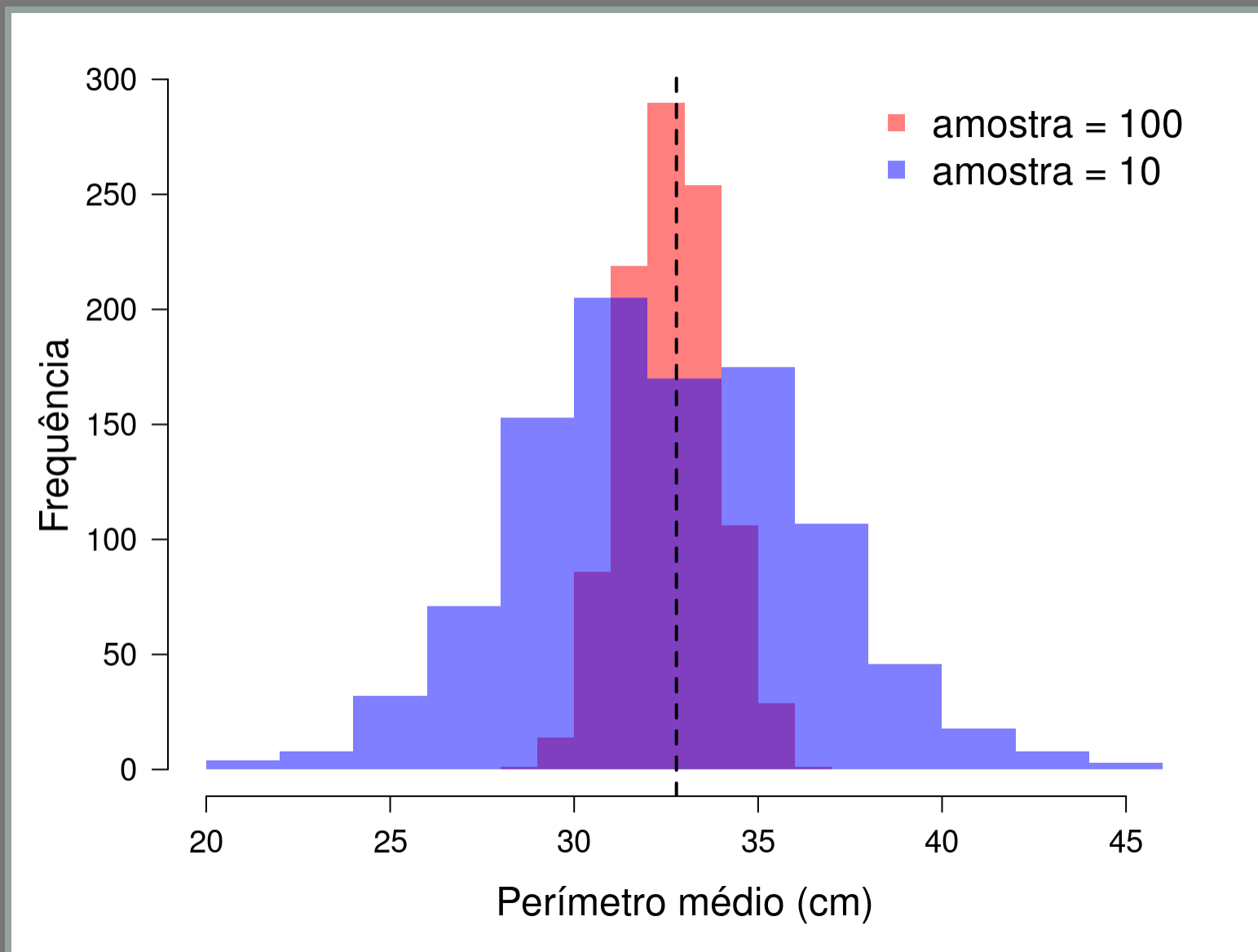
fazer afirmações sobre um universo a partir de um conjunto de valores representativo (amostra). Tal tipo de afirmação deve sempre vir acompanhada de uma medida de precisão sobre sua veracidade

```
> difpalm
```

```
## [1] 4.935
```

Os palmitos de CB são maiores que na IC?

Erro Amostral



Medida de Incerteza

fazer afirmações sobre um universo a partir de um conjunto de valores representativo (amostra). Tal tipo de afirmação deve sempre vir acompanhada de uma medida de precisão sobre sua veracidade

Os palmitos de CB são maiores que da IC.

Organizando nossos resultados

```
> amospalm <- c(amoscb, amosic)
> local <- rep(c("cb", "ic"), each =100)
> local <- factor(local, levels = c("ic"
> tapply(amospalm, local, mean)
```

```
##      ic      cb
## 28.805 33.740
```

```
> diff(tapply(amospalm, local, mean))
```

```
##      cb
## 4.935
```

Cenário Nulo

A diferença observada pode ser gerada pelo acaso

Cenário Nulo

```
> cenulo <- sample(amospalm)
> diff(tapply(cenulo, local, mean))

##      cb
## 0.285

> diff(tapply(sample(amospalm), local, r

##      cb
## 2.225

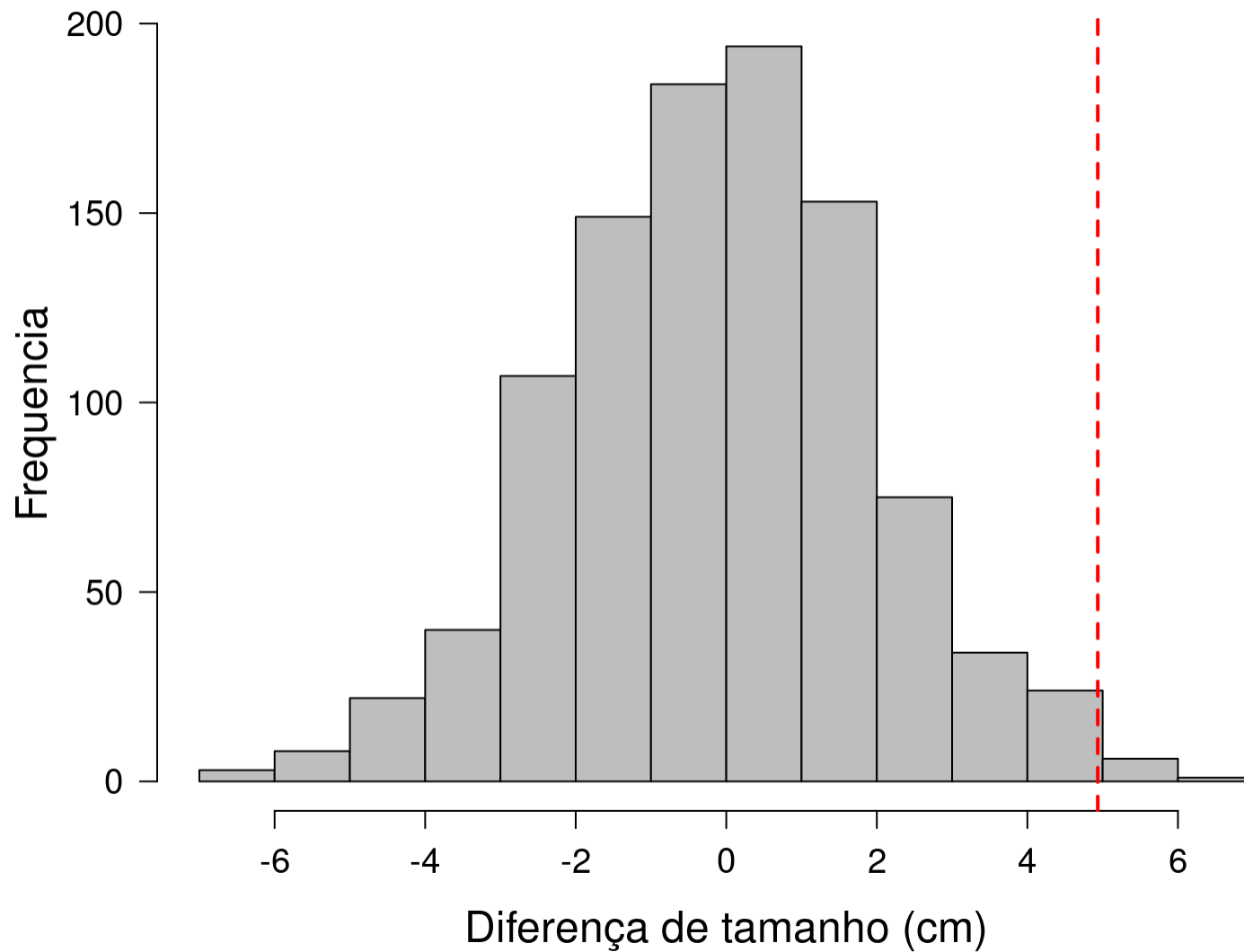
> diff(tapply(sample(amospalm), local, r

##      cb
## -2.875
```

Distribuição Nula

```
> dnulo <- rep(NA, 1000)
> dnulo[1] <- difpalm
> for(i in 2:1000)
+ {
+ dnulo[i] <- diff(tapply(sample(amospal
+ }
```

Distribuição Nula



Incerteza Sobre a Afirmação

A probabilidade da diferença observada ter sido gerada pelo acaso.

```
> sum(dnulo >= difpalm)
```

```
## [1] 9
```

```
> sum(dnulo >= difpalm) / length(dnulo)
```

```
## [1] 0.009
```

Incerteza Sobre a Afirmação

A probabilidade da diferença observada ter sido gerada pelo acaso.

```
> sum(dnulo >= difpalm)
```

```
## [1] 9
```

```
> sum(dnulo >= difpalm) / length(dnulo)
```

```
## [1] 0.009
```

p-valor

Teste Frequentista

Test T

```
##  
## Welch Two Sample t-test  
##  
## data: amoscb and amosic  
## t = 2.4187, df = 191.66, p-value = 0.  
## alternative hypothesis: true differen  
## 95 percent confidence interval:  
## 1.562573 Inf  
## sample estimates:  
## mean of .. mean of ..
```




Tabela de Anova



Anova: um exemplo

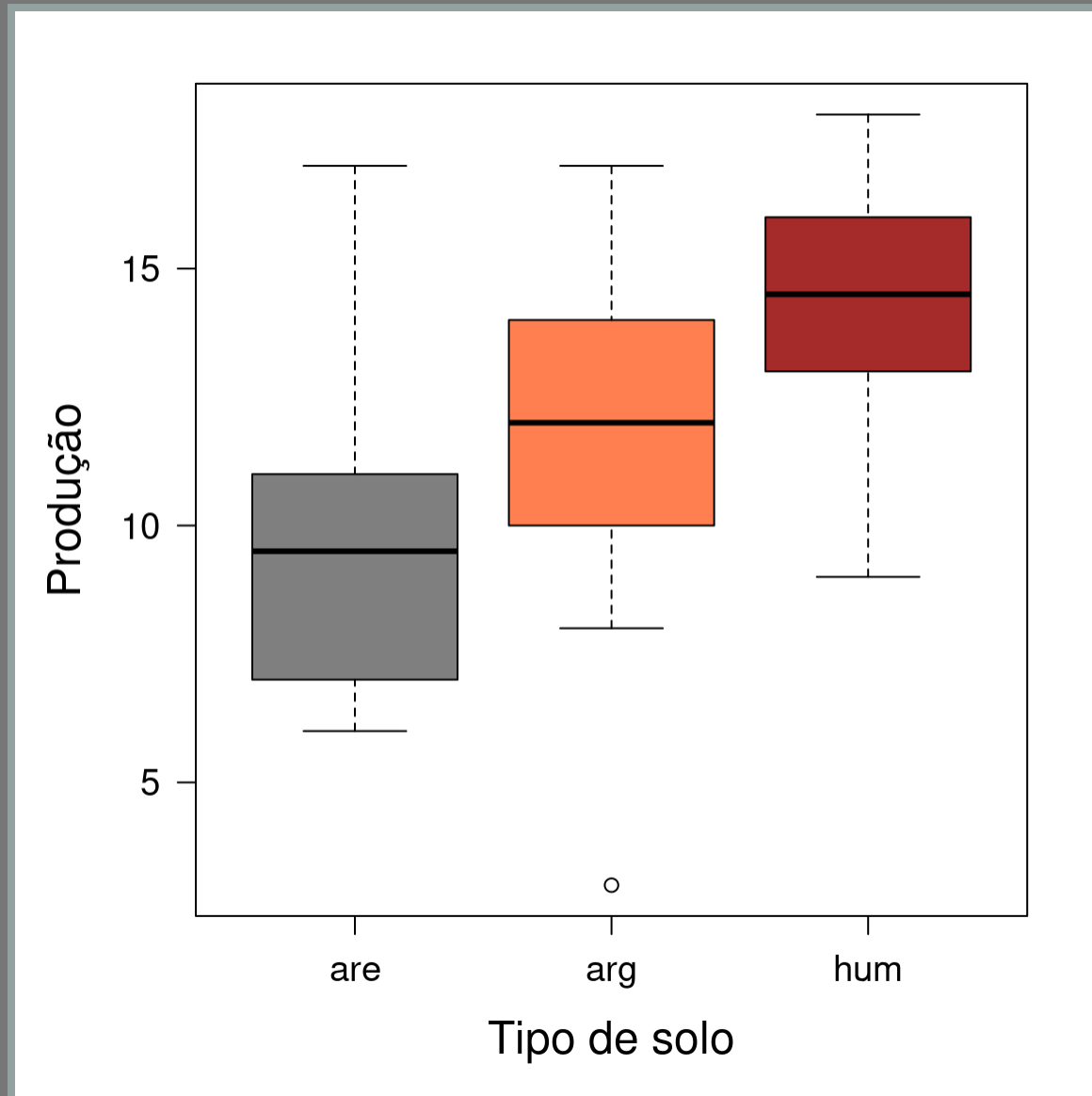
```
are=c(6,10,8,6,14,17,9,11,7,11)
arg=c(17,15,3,11,14,12,12,8,10,13)
hum=c(13,16,9,12,15,16,17,13,18,14)
prod <- c(are, arg, hum)
solos=data.frame(tipo=rep(c("are","arg",
str(solos)
```

```
## 'data.frame':   30 obs. of  2 variab
## $ tipo: Factor w/ 3 levels "are","ar
## $ prod: num   6 10 8 6 14 17 9 11 7 1
```

Anova: um gráfico

```
par(mar=c(5,5,2,2), cex.lab=1.5, cex.axis=1.5)  
boxplot(prod~tipo, data=solos, xlab="Tipo", ylab="Prod",  
        col=c("red", "green", "blue", "yellow", "purple"),  
        las=2)
```

Anova: um gráfico



Cálculos base

```
(media.geral=mean(solos$prod))
```

```
## [1] 11.9
```

```
(media.solos=tapply(solos$prod,list(solo
```

```
## are arg hum
```

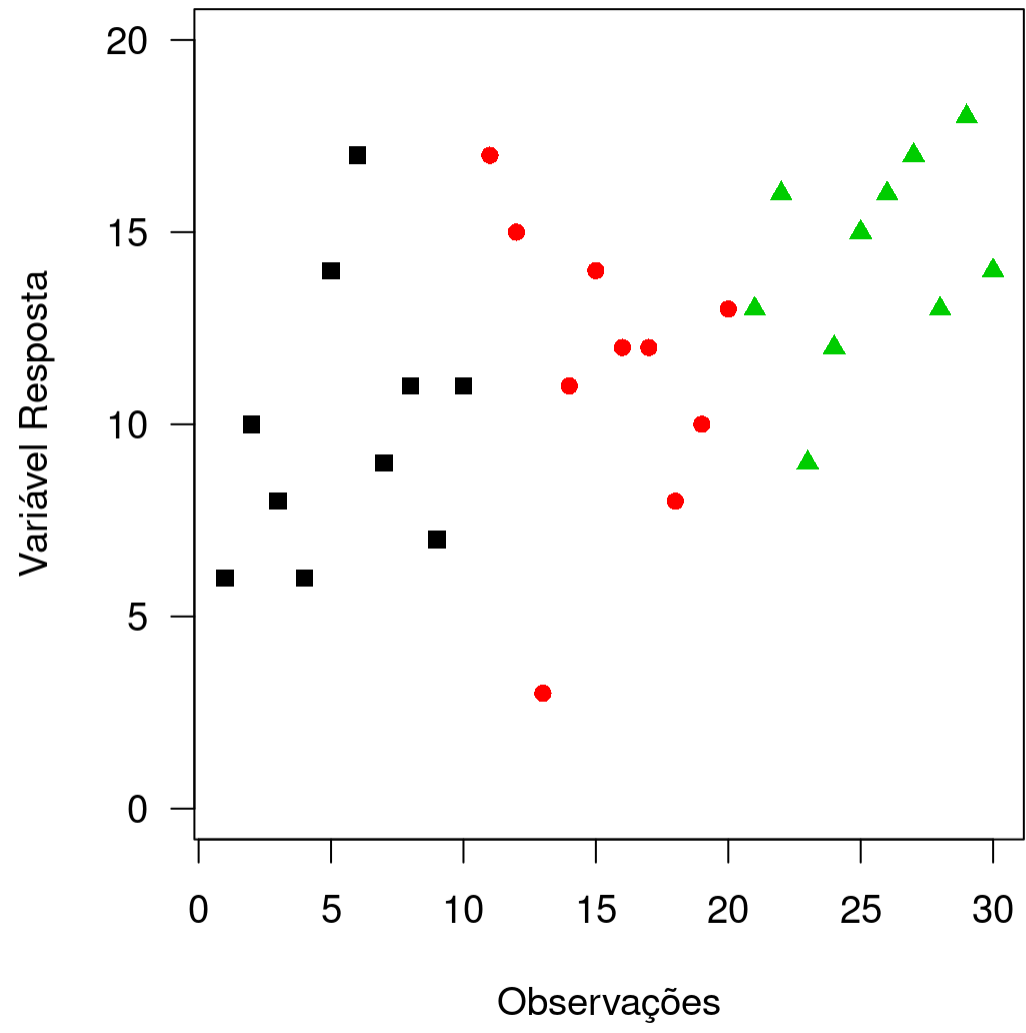
```
## 9.9 11.5 14.3
```

```
vetor.cor = rep(1:3,each=10)
```

```
vetor.medias=rep(media.solos, each=10)
```

Gráfico: desvios

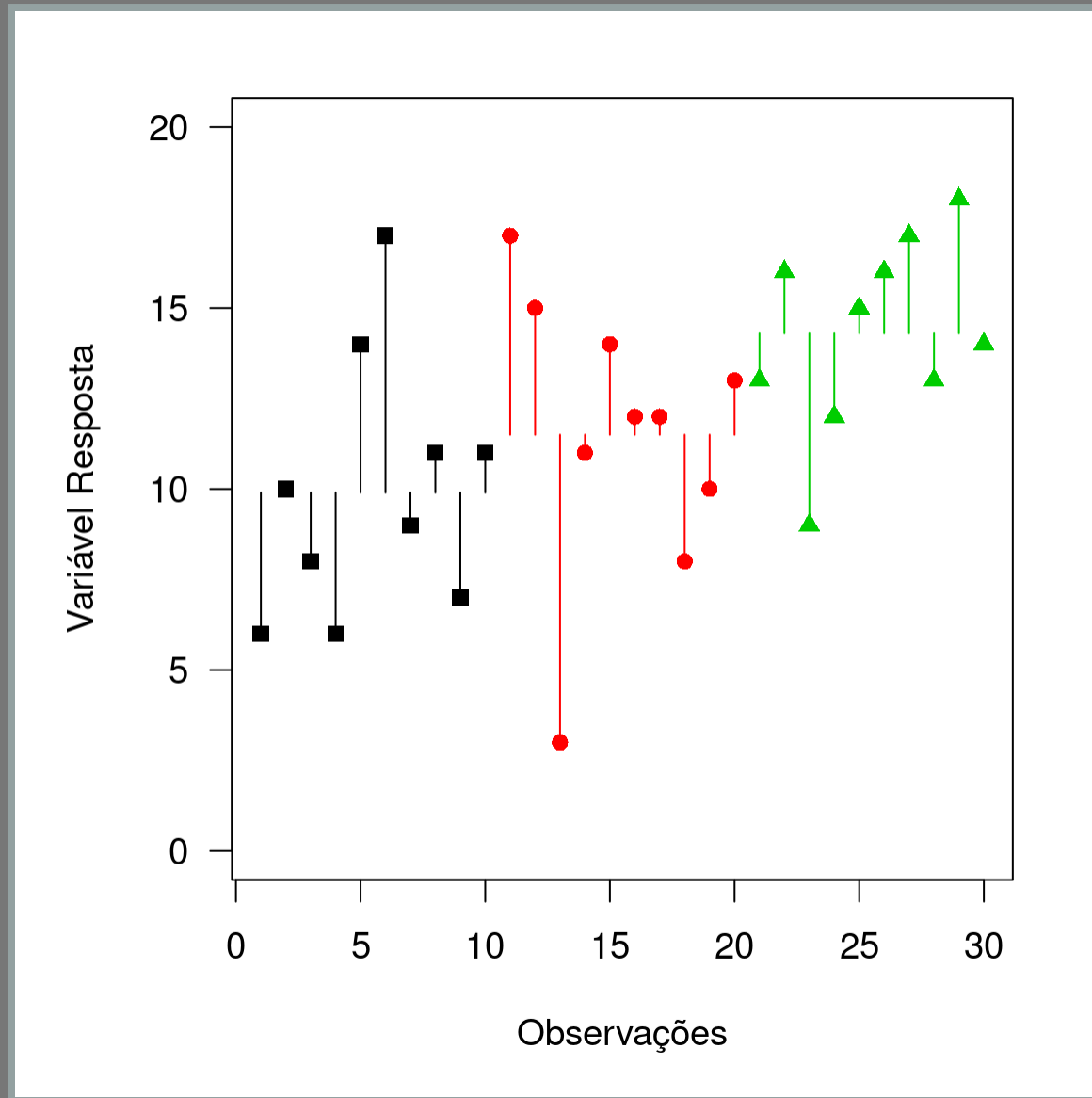
```
par(mar=c(5,5,2,2), las=1, cex=1.2)  
plot(1:30, solos$prod, ylim=c(0,20), pch=(r
```



Adicionando linhas: desvios

```
segments(x0 = 1:30, y0=prod, x1= 1:30,  
         y1=vetor.medias, col=vetor.cor)
```

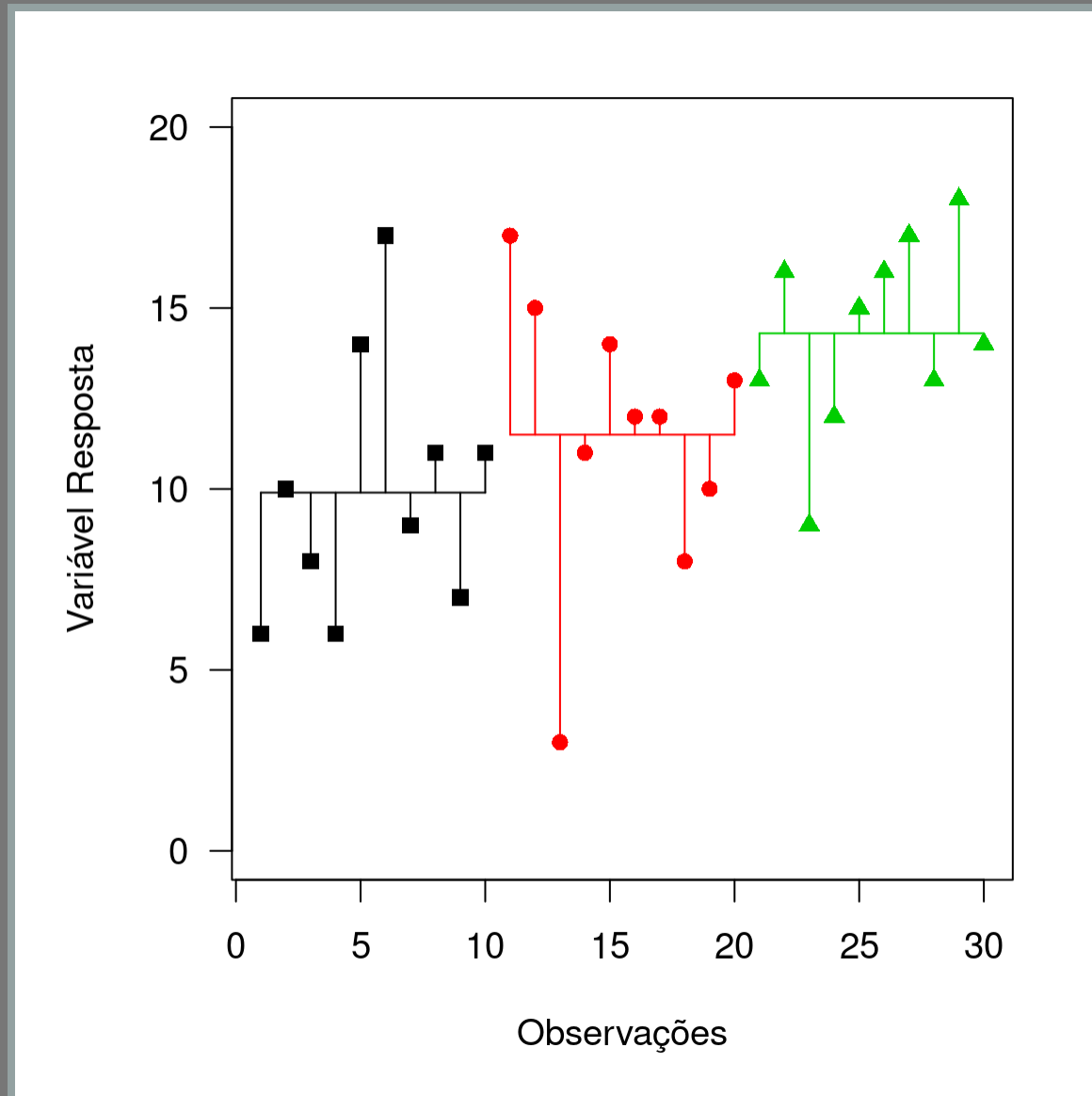
Adicionando linhas: desvios



Adicionando segmentos: médias

```
segments(x0=c(1,11,21), y0=media.solos,  
         x1=c(10,20,30), col=1:3)
```

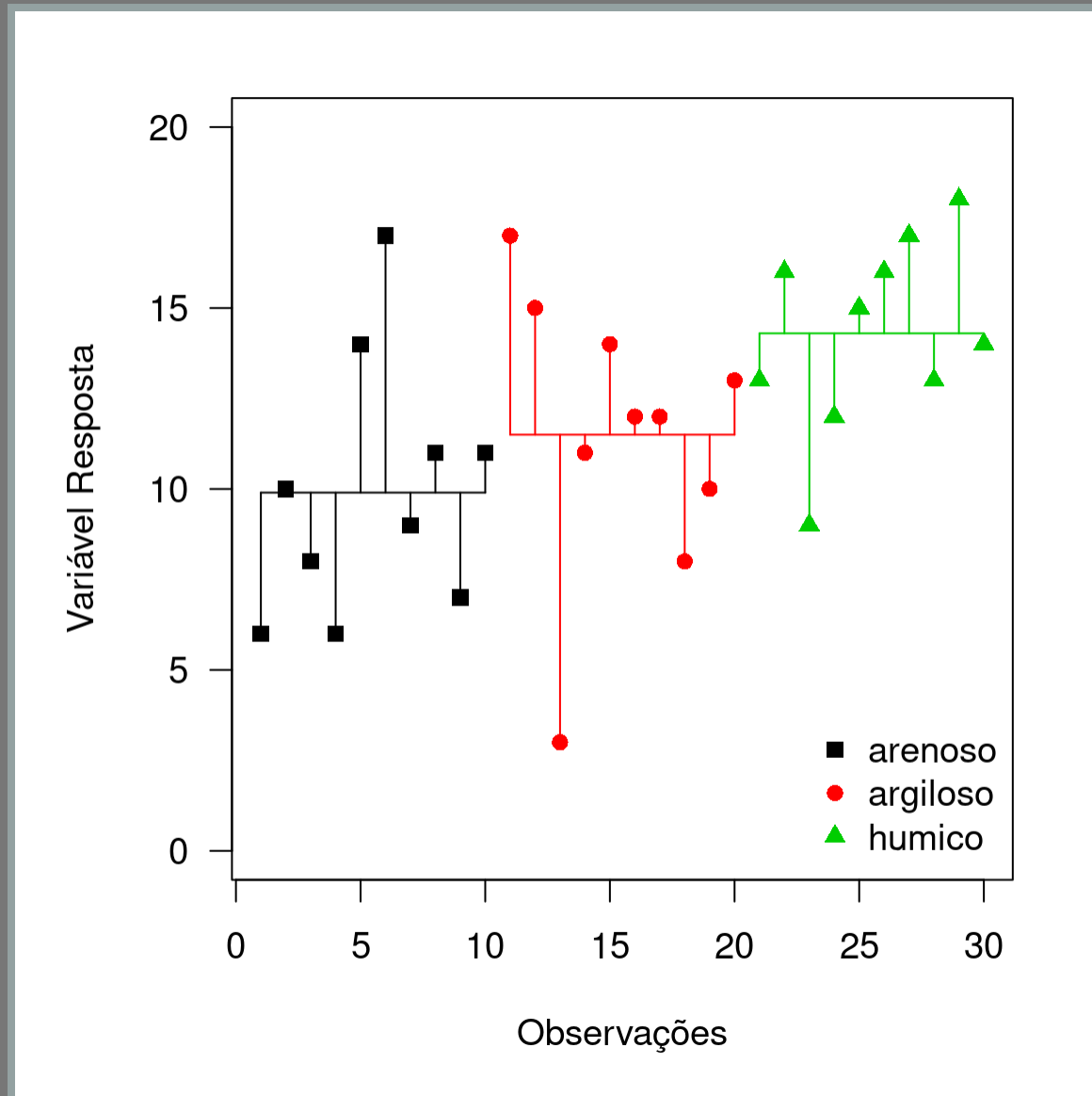
Adicionando segmentos: médias



Adicionando legenda

```
legend( "bottomright",  
        legend=c("arenoso", "argiloso", "h  
pch=c(15, 16, 17),  
col=1:3, bty="n")
```

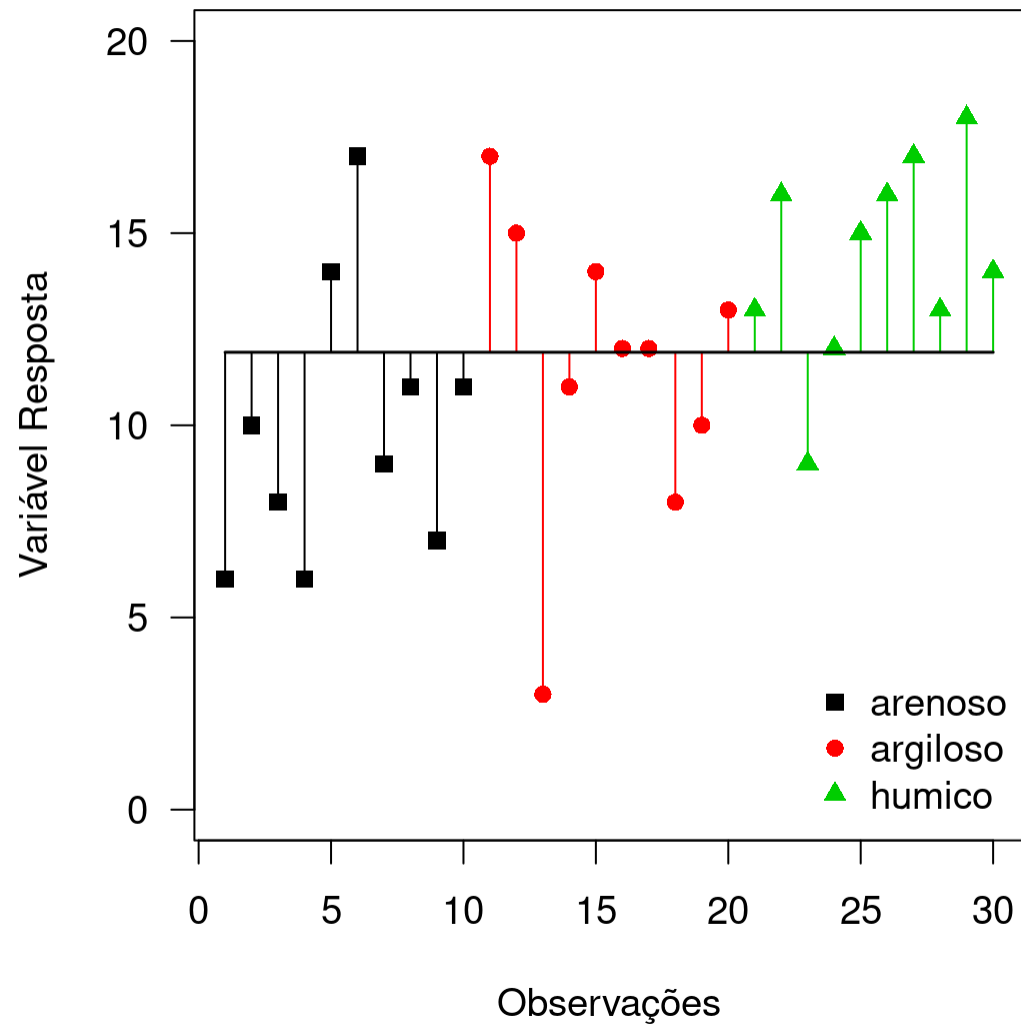
Gráficos Desvios: intra grupos



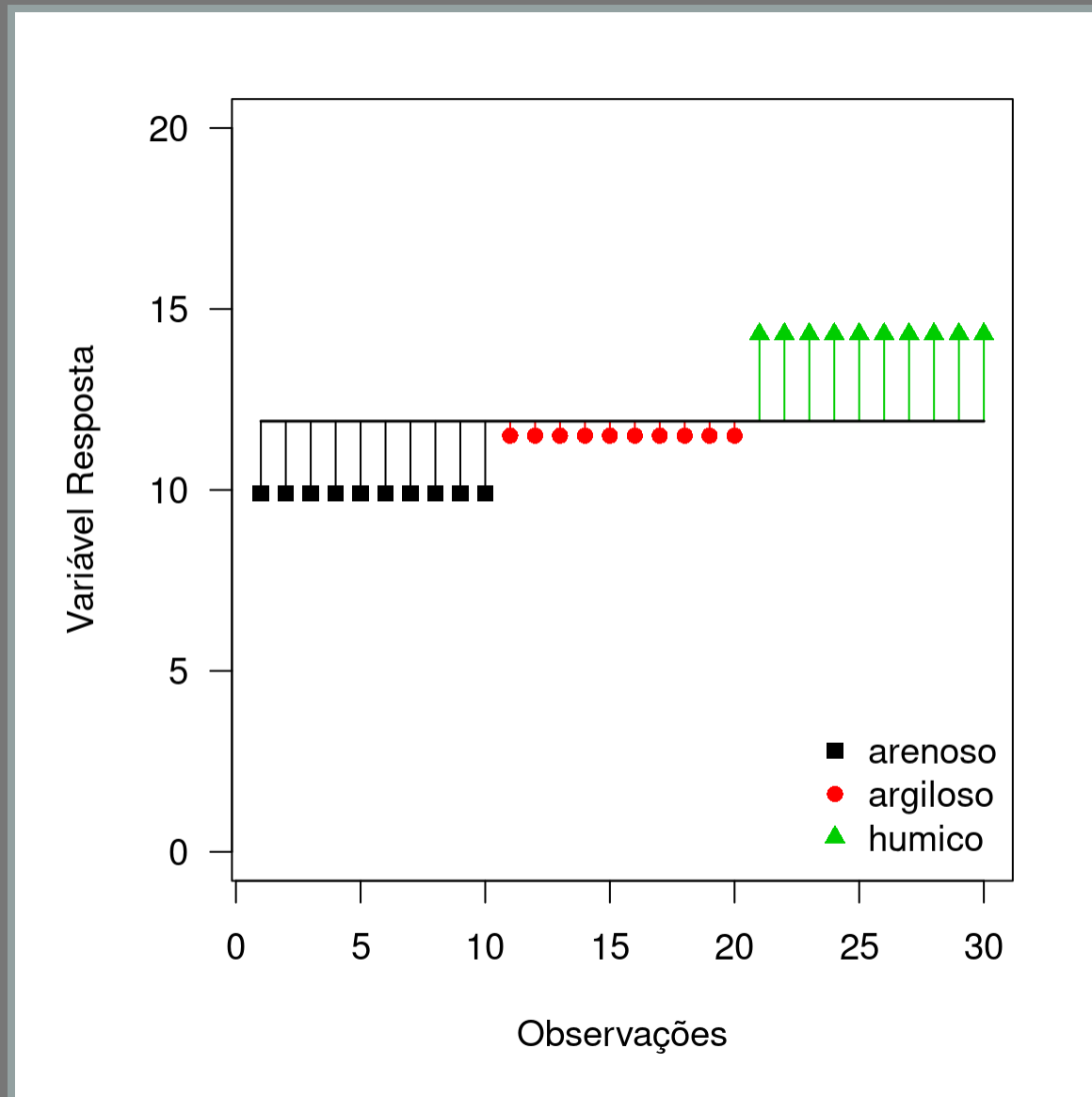
Gráficos Desvios: total

```
plot(1:30, solos$prod, ylim=c(0, 20), pch=(r
segments(x0 = 1:30, y0 = prod,
          x1= 1:30, y1=media.geral, col=v
segments(x0 = 1, y0= media.geral,
          x1 = 30, lty = 1, lwd=1.5)
legend( "bottomright", legend=c("arenosc
```

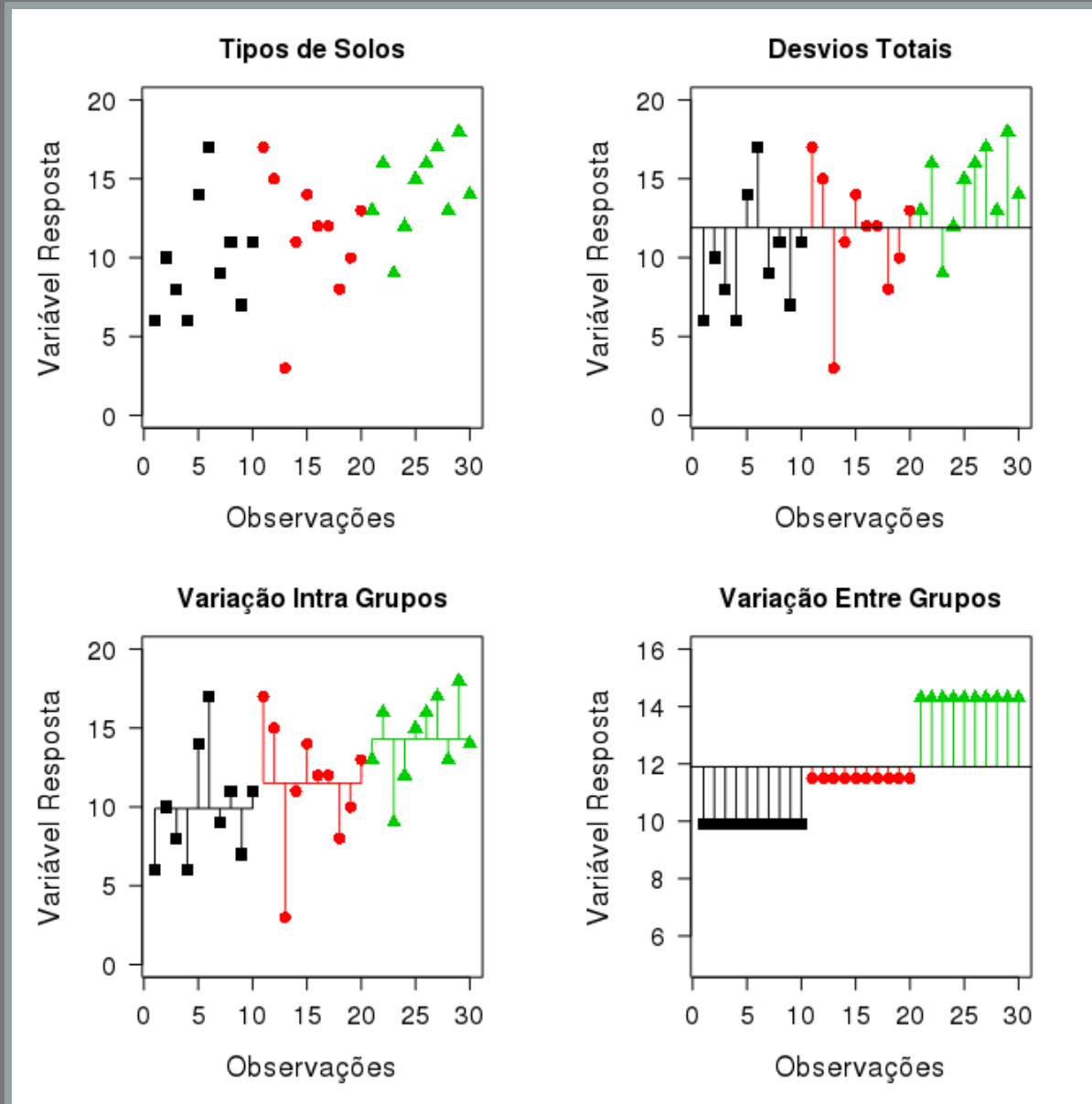
Gráficos Desvios: total



Gráficos Desvios: inter grupos



Anova: partição da variação



Anova: razão entre variâncias

$$F = \frac{\sigma_{entre}^2}{\sigma_{intra}^2}$$

Médias Quadráticas: intra

$$SS_{intra} = \sum_{i=1}^k \sum_{j=1}^n (y_{i,j} - \bar{y}_i)^2$$

$$MS_{intra} = \frac{SS_{intra}}{df_{intra}}$$

Médias Quadráticas: entre

$$SS_{entre} = \sum_{i=1}^k \sum_{j=1}^n (\bar{y}_i - \bar{\bar{y}})^2$$

$$MS_{entre} = \frac{SS_{entre}}{df_{intra}}$$

Estatística F

$$F = \frac{MS_{entre}}{MS_{intra}}$$

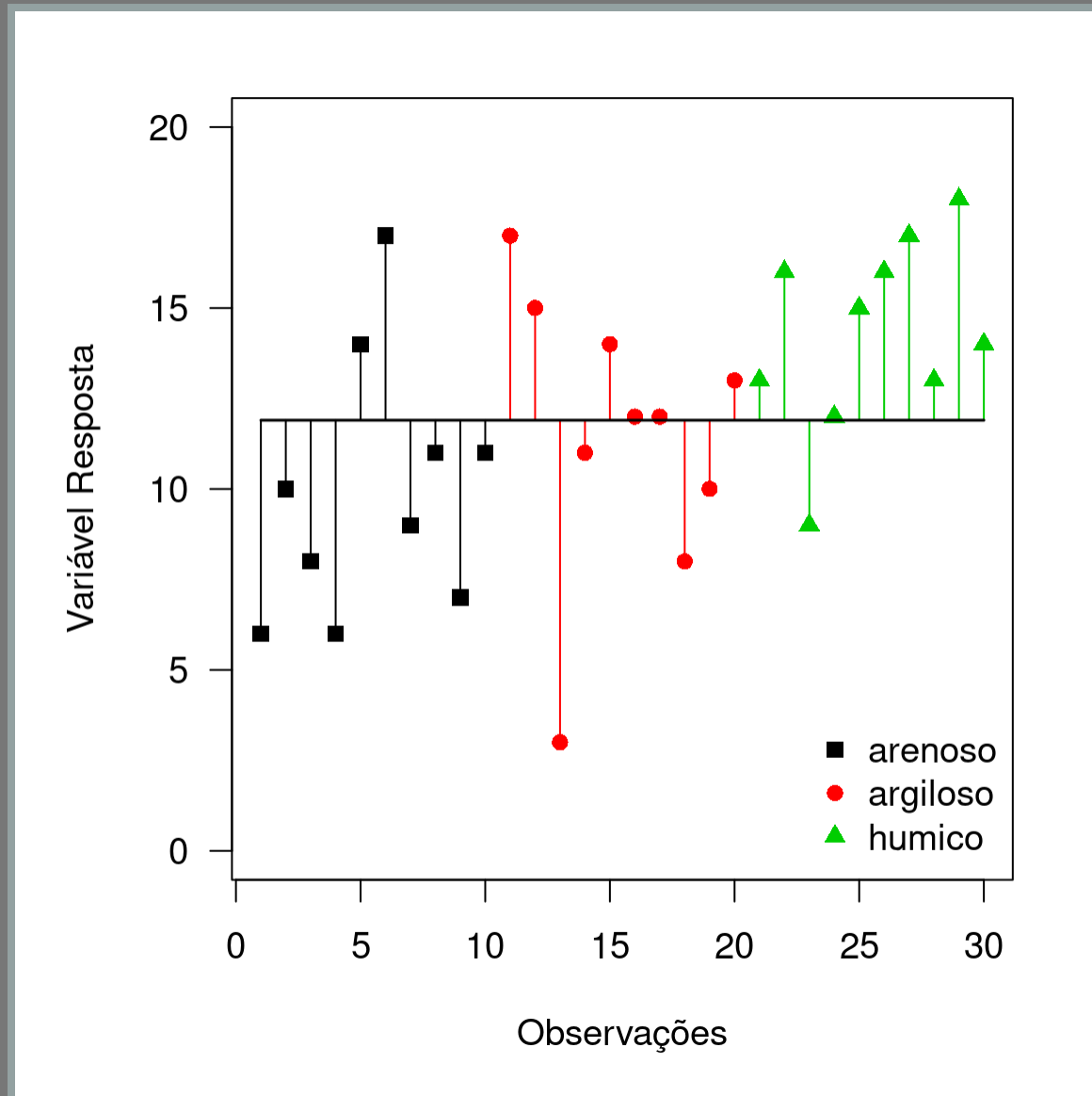
Mesmo que:

$$F = \frac{\sigma_{entre}^2}{\sigma_{intra}^2}$$

Tabela de Anova

Fonte	Desvio Quadrático	Graus de Liberdade	Desvio Médio	Razão das Variâncias	P-valor
Entre Grupos					
Intra Grupos					
TOTAL	x				

Tabela de Anova: SStotal



Desvios quadráticos totais

$$SS_{total} = \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y})^2$$

Desvios quadráticos: totais

```
prod
```

```
## [1] 6 10 8 6 14 17 9 11 7 11 17  
## [24] 12 15 16 17 13 18 14
```

```
mean(prod)
```

```
## [1] 11.9
```

```
(ss_total <- sum((prod - mean(prod))^2 )
```

```
## [1] 414.7
```

Tabela de Anova

Fonte	Desvio Quadrático	Graus de Liberdade	Desvio Médio	Razão das Variâncias	P-valor
Entre Grupos					
Intra Grupos	x				
TOTAL	414.7				

Tabela de Anova: SSintra

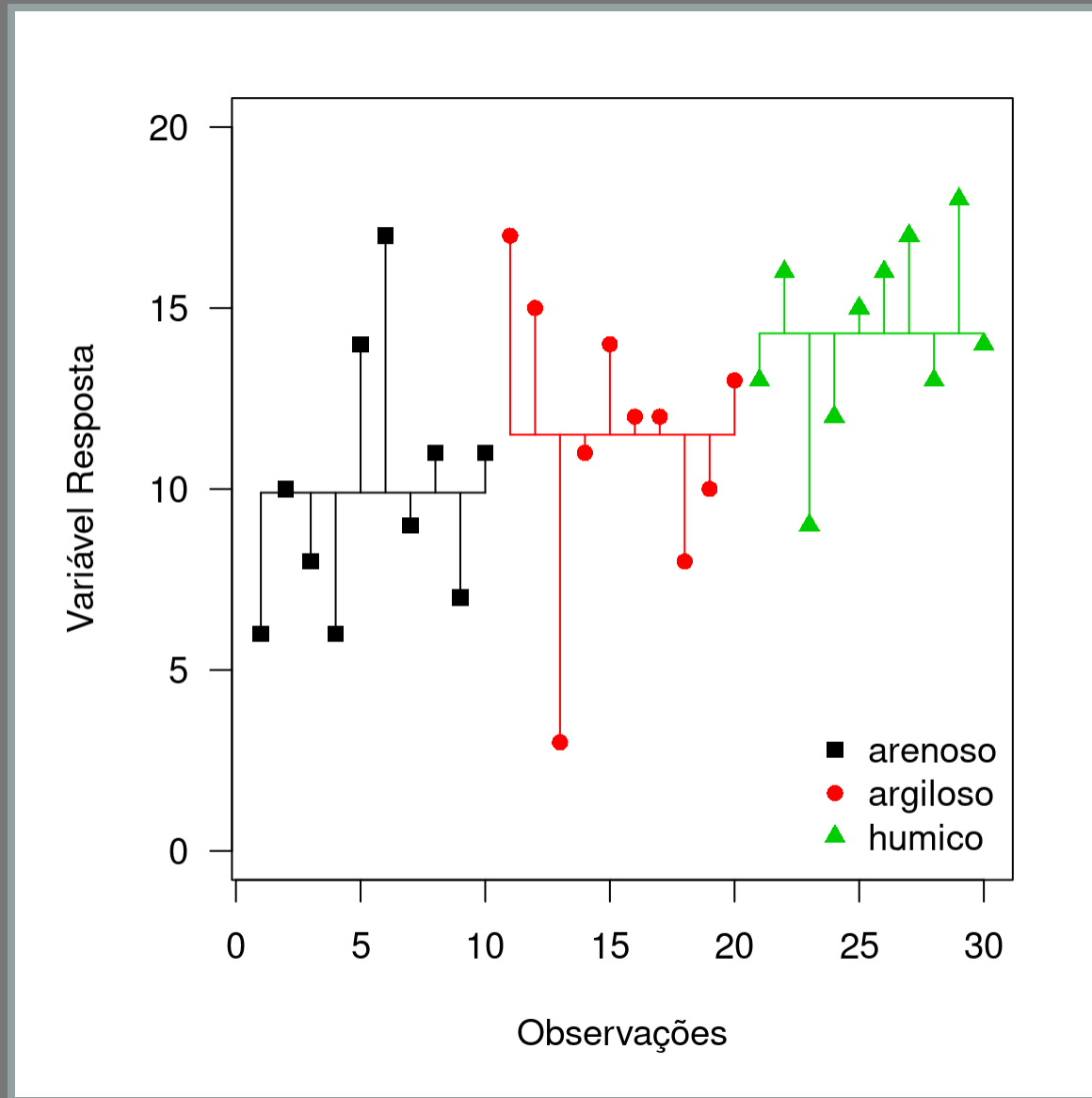


Tabela de Anova: SS_{intra}

$$SS_{intra} = \sum_{i=1}^k \sum_{j=1}^n (y_{i,j} - \bar{y}_i)^2$$

Atividades desta tarde

- Até as 16h:
 - tutorial 5a e 5b
 - apostila
 - dúvidas das unidades anteriores
- Após as 16h
 - dúvidas dos exercícios de gráficos