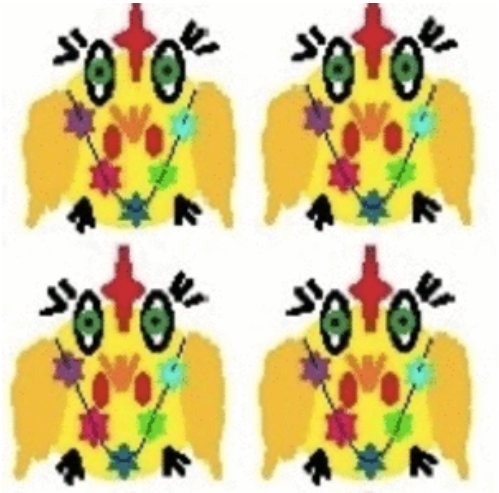
Reamostragem e Permutação

PlanECO 2017

Alexandre Adalardo de Oliveira

Rsampling



Reamostragem e Permutação

2/52

Introdução

Técnicas de simulações aleatórias, baseadas em dados ou distribuições teóricas, buscando soluções numéricas.

- teste de hipóteses
- medidas de precisão de estimativas
- otimizadores
- integração numérica
- algoritmos de amostragem

Definições

Monte Carlo

Técnicas de simulação buscando resultado numérico

- simulações aleatórias (numérico)
- distribuição conhecida (normal, poisson)
- MCMC "Markov Chain Monte Carlo"

Reamostragem

Técnicas de reamostragem de dados

- precisão de estimativa (bootstrap, jackknife)
- teste de significância (reordenação, permutação)
- validação de modelos (subconjuntos)

Definições

Teste de Permutação

- reordenamento (rótulos) em todas as combinações possíveis (teste exato de Fisher)
- Teste de Monte Carlo: uma amostra das combinações

Reposição

Bootstrap
Técnica de reamostragem de dados com reposição que permite a inferência sobre a precisão de uma estimativa.

Implicações

Não assumem uma distribuição de probabilidade teórica

Vantagens

- poucas restrições (dados)
- intuitiva (não necessita conhecimento matemático profundo)
- poucos pressupostos
- assintótica (p-value)

Implicações

Desvantagens

- dificuldade computacional (?)
- resultado pode variar (estatística de interesse)
- domínio de inferência restrito (?)
- necessidade de cenário nulo adequado (complexo)

Exemplo Simulação

Uma amostra

- Normal Padronizada (distribuição z)

```
set.seed(22)
amostra <- rnorm(20, mean=0, sd=1)
amostra
```

```
##  [1] -0.51213909  2.48518368  1.00782615  0.29281457 -0.20895936
##  [6]  1.85809239 -0.06602641 -0.16276495 -0.19986068  0.30056173
## [11] -0.76390728  0.08196190  0.74302828 -0.08402219 -0.79289452
## [16] -0.92215363  0.86156238  2.00294219  0.93655101 -1.61573487
```

Distribuição teórica (Normal Padronizada)

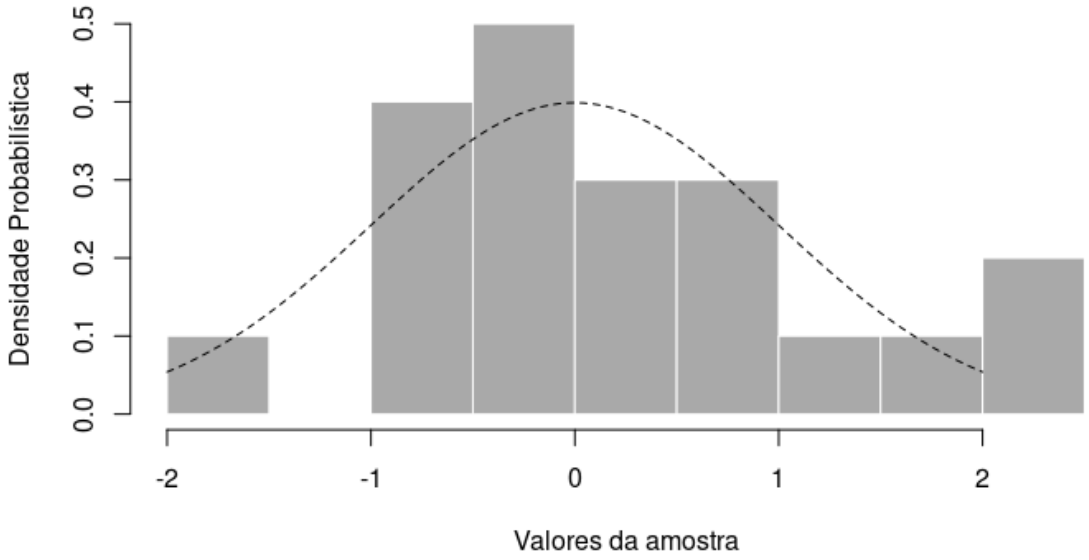
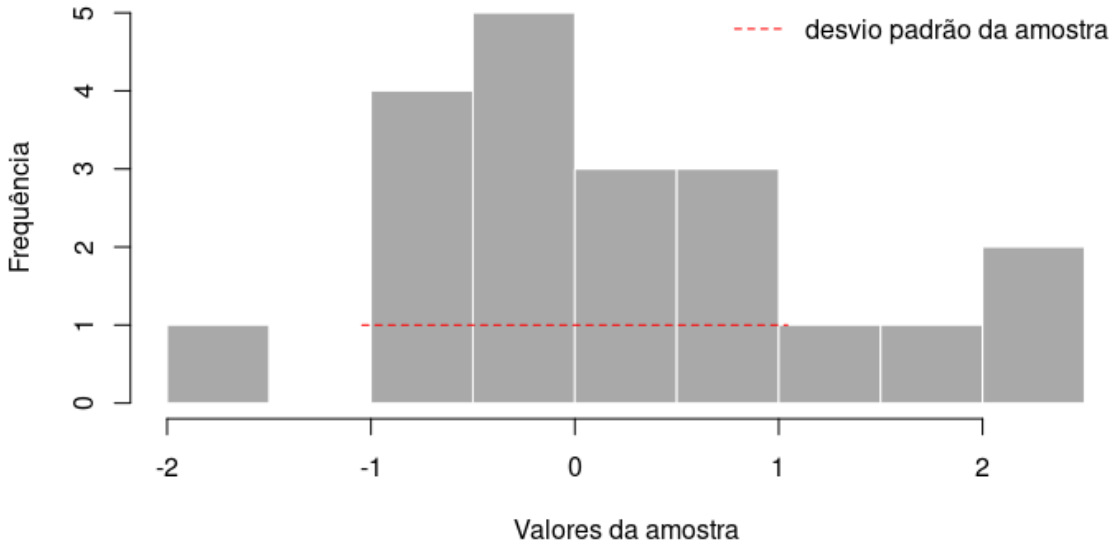


Gráfico da amostra



10/52

Simula desvio padrão da amostra

Numero de simulações:

10 5,000

10 2,006 4,002

Tamanho Amostral

3

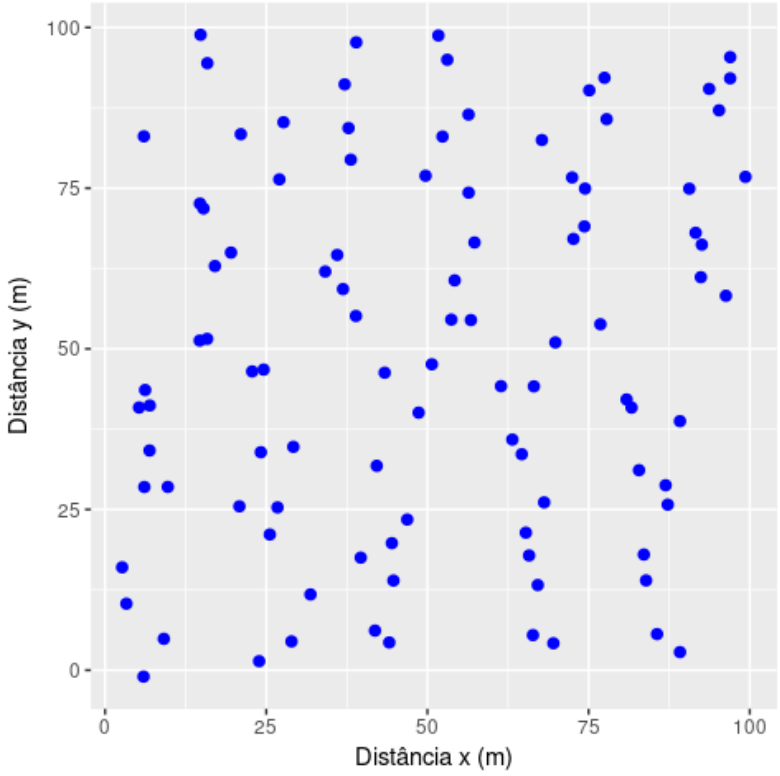
11/52

Teste de Hipóteses

1. Definir a estatística de interesse (EI)
2. Reamostrar dados (reordenar ou simular) no cenário nulo
3. Recalcular a EI (pseudovalores)
4. Criar a distribuição dos pseudovalores
5. Posicionar o observado na distribuição
6. Calcular o p-valor

Examplo de Monte Carlo

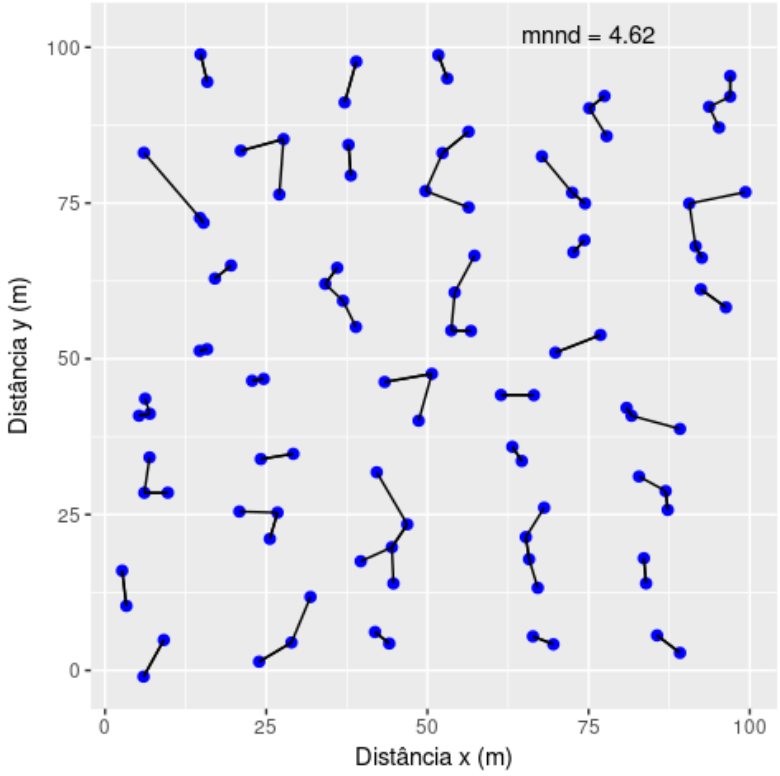
Plantas estão distribuídas aleatoriamente em uma parcela?



13/52

1a. Estatística de interesse

- média da distância ao vizinho mais próximo



14/52

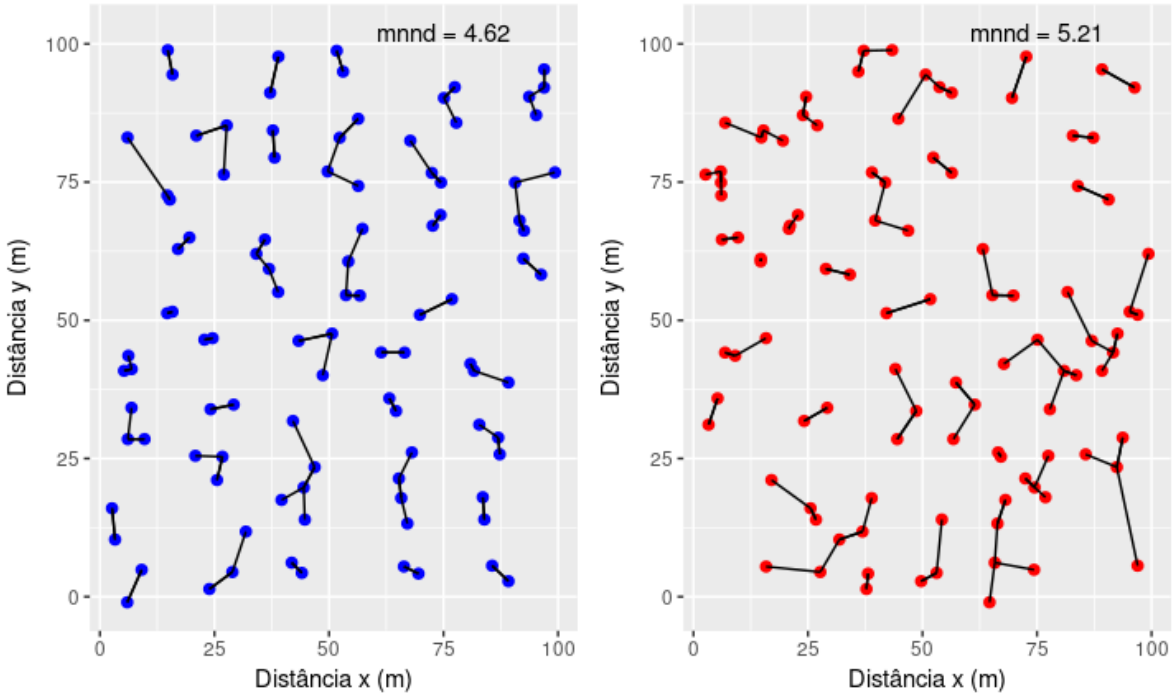
1b. Estatística de interesse observada

- média da distância ao vizinho mais próximo

```
mean(nnd)
```

```
## [1] 4.620198
```

2. Simula cenário nulo



3-4. Cria a distribuição de pseudovalores

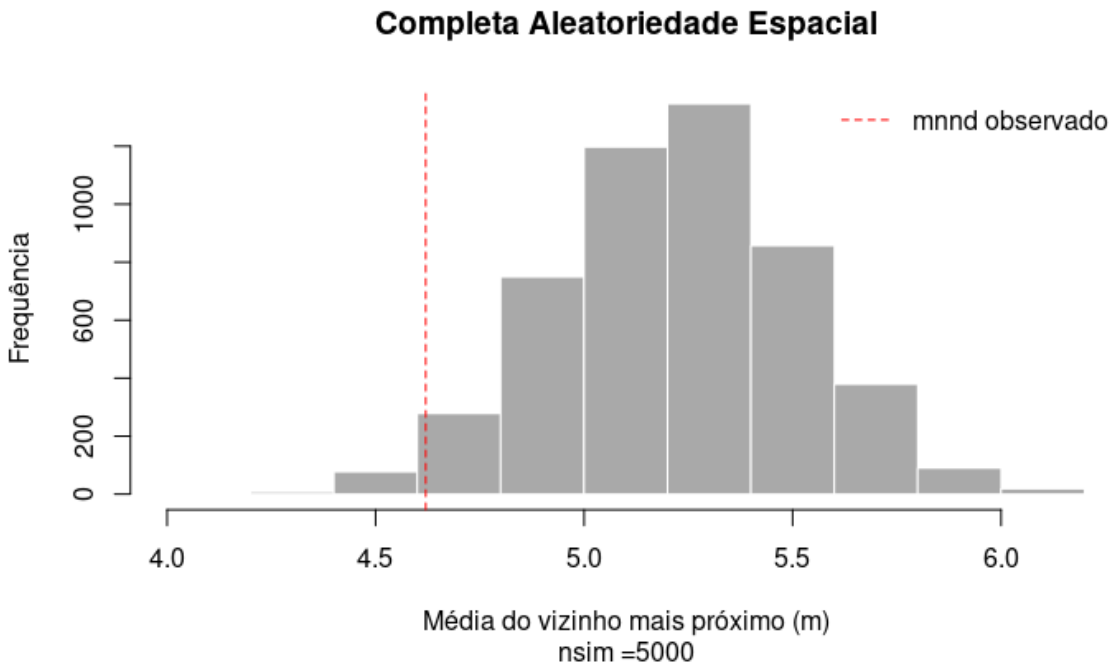
Numero de simulações:

10 5,000

10 2,006 4,002

17/52

5. Posiciona o observado na distribuição nula



18/52

6. Calcula o p-valor

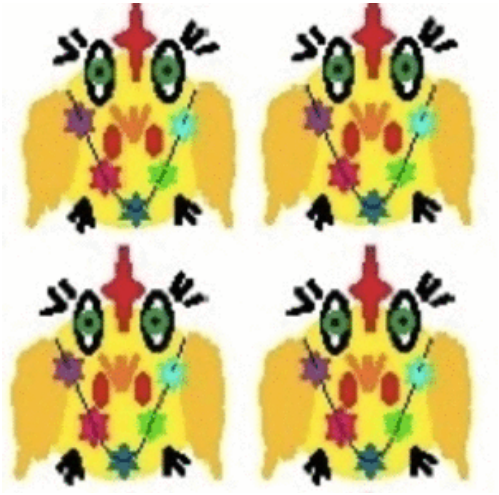
O p-valor é a proporção de simulações com valores iguais ou mais extremos do que o observado.

- centralizar os valores
- contar valores extremos (bi ou unicaudal)
- dividir pelo número de simulações

```
icentra <-  resim00 - mean(resim00)
## os pontos estão mais próximos do que o esperado para o acaso
p.valueUni <- sum(icentra <= icentra[1])/length(icentra)
## a distribuição dos pontos não pode ser considerada diferente do acaso
p.valueBi <-  sum(abs(icentra) >= abs(icentra[1]))/length(icentra)
c(p.valueUni, p.valueBi)
```

```
## [1] 0.0206 0.0384
```

Atividade



ANOVA de um fator por reamostragem

- Estatística de interesse: somatória das diferenças em módulo, estatística F, desvios quadráticos
- Exemplo: produtividade agrícola em diferentes solos
- 10 áreas de cultivos para cada solo (escolha aleatorizada)

```
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)
cbind(are, arg, hum)[1:5,]
```

```
##      are arg hum
## [1,]   6  17  13
## [2,]  10  15  16
## [3,]   8   3   9
## [4,]   6  11  12
## [5,]  14  14  15
```

ANOVA de um fator: estrutura dos dados

```
crop <- data.frame(solo = rep(c("are", "arg", "hum"), each=10), colhe = c(are, arg, hum))
head(crop)
```

```
##   solo colhe
## 1  are     6
## 2  are    10
## 3  are     8
## 4  are     6
## 5  are    14
## 6  are    17
```

```
tail(crop)
```

```
##    solo colhe
## 25  hum    15
## 26  hum    16
## 27  hum    17
## 28  hum    13
## 29  hum    18
## 30  hum    14
```

ANOVA um fator: estatística de interesse

- diferenças das médias de cada grupo em relação média geral (grande média)
- módulo da soma das diferenças

```
medTotal = mean(crop$colhe)
medTotal
```

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

```
medGrupo = tapply(crop$colhe, crop$solo, mean)
medGrupo
```

```
##  are  arg  hum
##  9.9 11.5 14.3
```

ANOVA um fator: estatística de interesse

```
(diffGr <- medGrupo - medTotal)
```

```
##  are  arg  hum
## -2.0 -0.4  2.4
```

```
sum(diffGr)
```

```
## [1] 0
```

```
(anovaEI <- sum(abs(diffGr)))
```

```
## [1] 4.8
```

ANOVA um fator: simular o cenário nulo

- As diferenças entre grupo são geradas pelo acaso?

```
crop$solo
```

```
##  [1] are are are are are are are are are are are arg arg arg arg arg arg arg
## [18] arg arg arg hum hum hum hum hum hum hum hum hum hum
## Levels: are arg hum
```

```
(scrop <- sample(crop$solo))
```

```
##  [1] are hum arg are arg are hum hum hum hum arg arg arg are hum are arg
## [18] are are hum hum arg are arg are hum arg are arg hum
## Levels: are arg hum
```

```
sum(abs(tapply(crop$colhe, scrop, mean)-medTotal ))
```

```
## [1] 2.6
```

```
anovaEI
```

ANOVA um fator: distribuição dos pseudovalores

Número de reamostragens:

10 5,000

10 2,006 4,002

ANOVA um fator: calculo do p

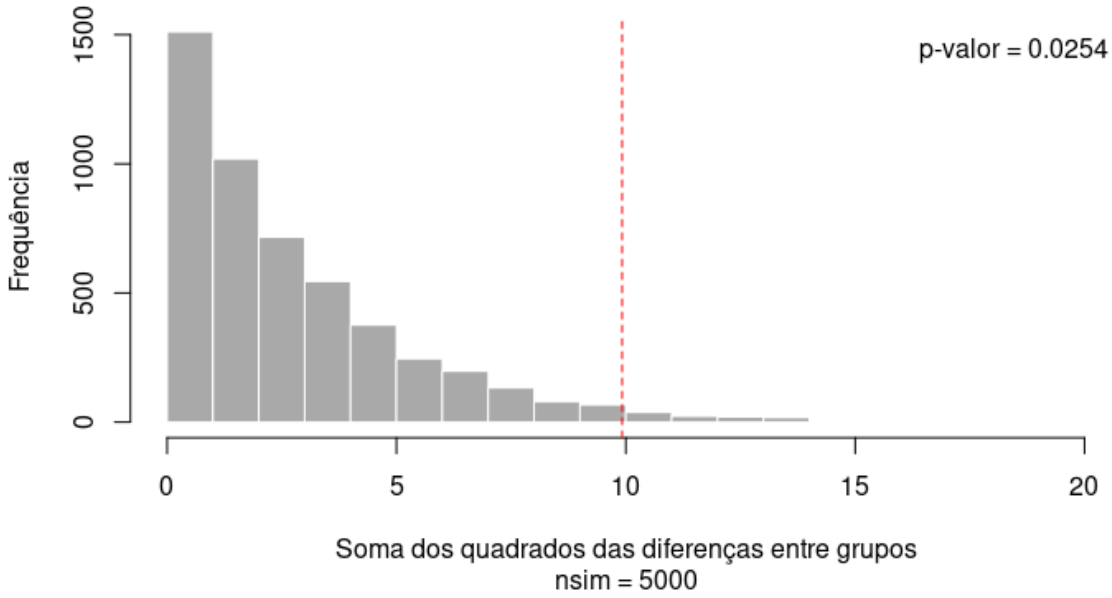
```
set.seed(42)
    resimAnova <- rep(NA, 5000)
    resimAnova[1] <- anovaEI
    for (i in 2:5000 )
    {
     resimAnova[i] <- sum(abs(tapply(crop$colhe, sample(crop$solo), mean) - medTotal))
    }
(p.anova <- sum(resimAnova >= resimAnova[1])/length(resimAnova))
```

```
## [1] 0.0324
```

ANOVA um fator: soma dos desvios quadrados

$$\sum_{i}^{n}(mediaGrupo_i - mediaTotal)^2$$

ANOVA um fator: soma dos desvios quadrados



29/52

ANOVA um fator: razão das variâncias (F)

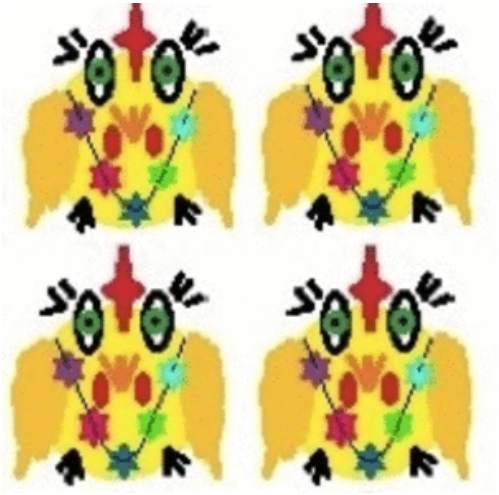
$$varEntre = \frac{\sum_i^n (mediaGrupo_i - mediaTotal)^2}{ngrupos - 1}$$

$$varIntra = \frac{\sum_1^n (x_{i,j} - mediaGrupo_j)^2}{n - ngrupos}$$

$$F = \frac{varEntre}{varIntra}$$

30/52

Atividade



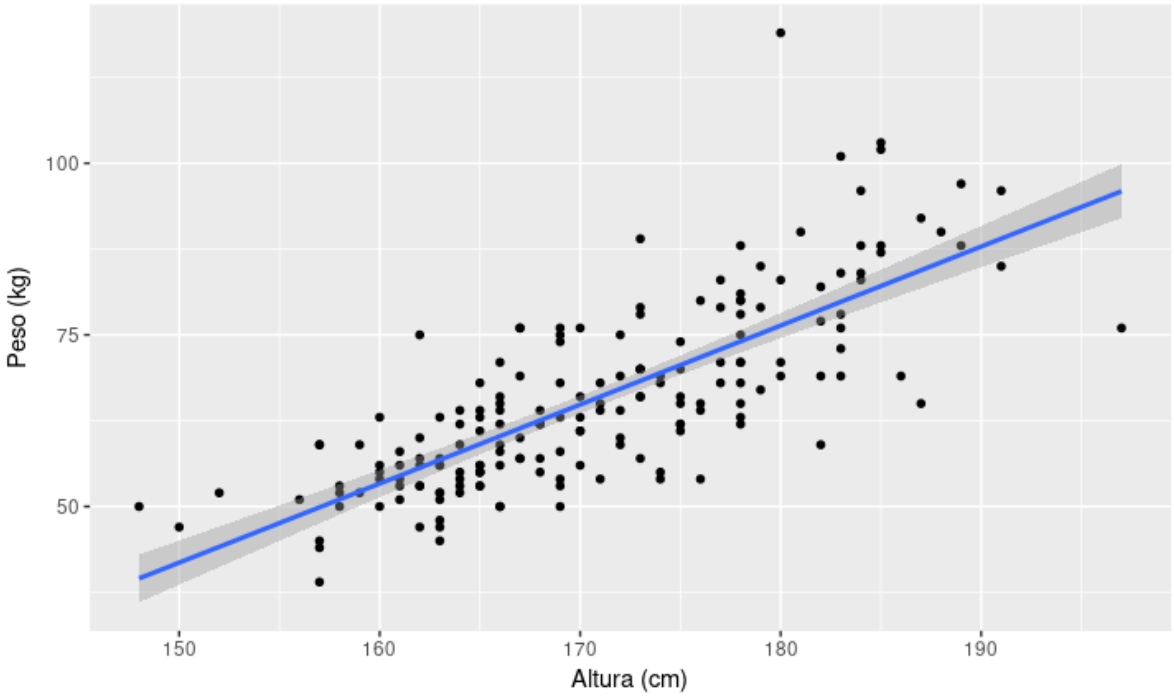
Anova no Rsampling

REGRESSÃO por reamostragem

- Davis (1990). Appetite (15)13-21

```
## 'data.frame':    180 obs. of  5 variables:
##  $ sex   : Factor w/ 2 levels "F","M": 2 1 1 2 1 2 2 2 2 2 ...
##  $ weight: int  77 58 53 68 59 76 76 69 71 65 ...
##  $ height: int  182 161 161 177 157 170 167 186 178 171 ...
##  $ repwt : int  77 51 54 70 59 76 77 73 71 64 ...
##  $ repht : int  180 159 158 175 155 165 165 180 175 170 ...
```

REGRESSÃO: gráfico



33/52

REGRESSÃO: estatística de interesse

- inclinação da reta (β)

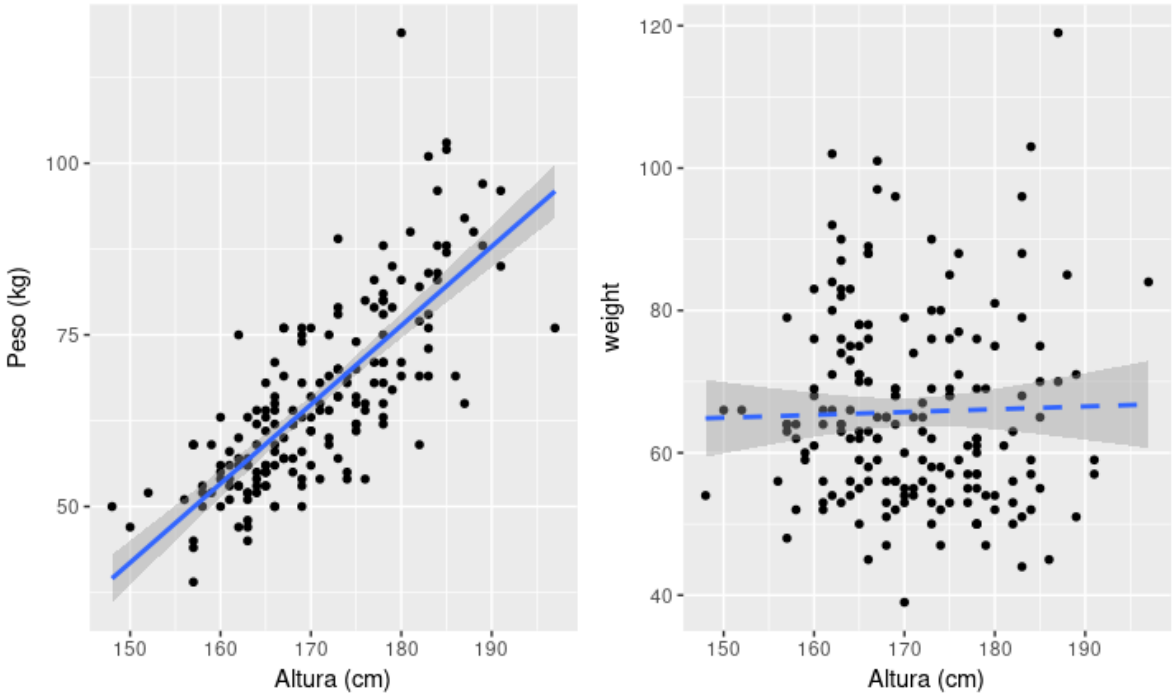
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

REGRESSÃO: cenário nulo

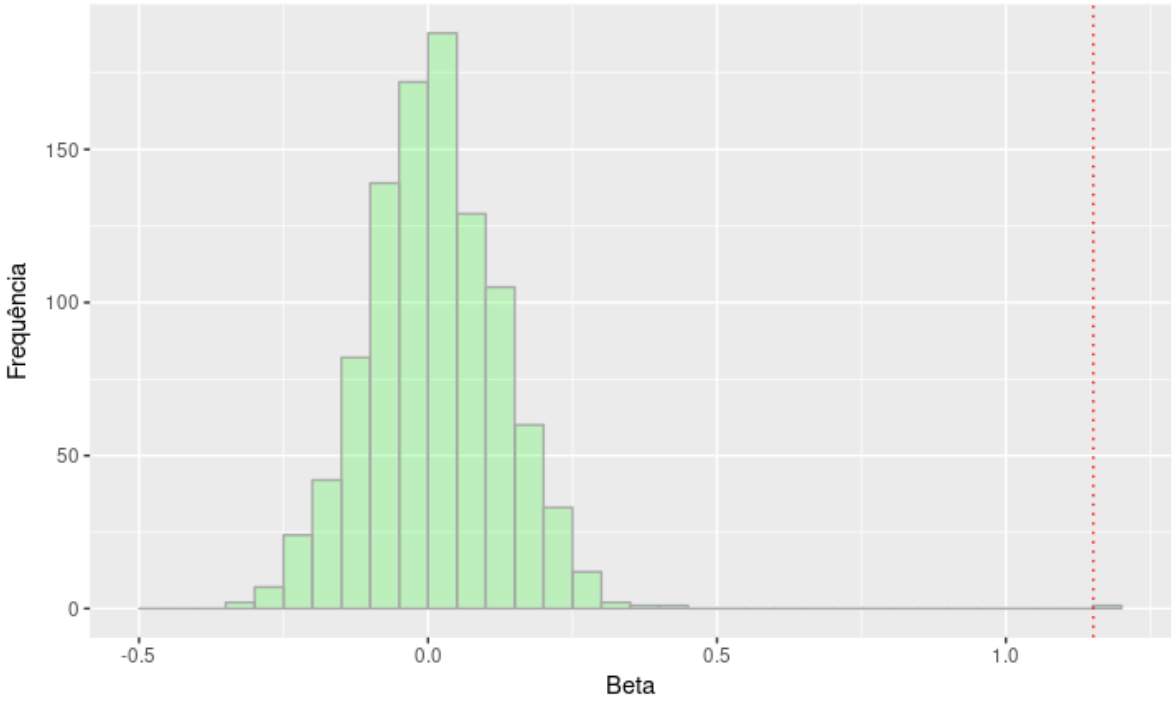
```
head(cbind( Davis[,c(2,3)], sHeight= sample(Davis$height)))
```

```
##   weight height sHeight
## 1     77    182     162
## 2     58    161     170
## 3     53    161     163
## 4     68    177     173
## 5     59    157     170
## 6     76    170     166
```

Cenário Nulo



REGRESSÃO: distribuição de pseudovalores



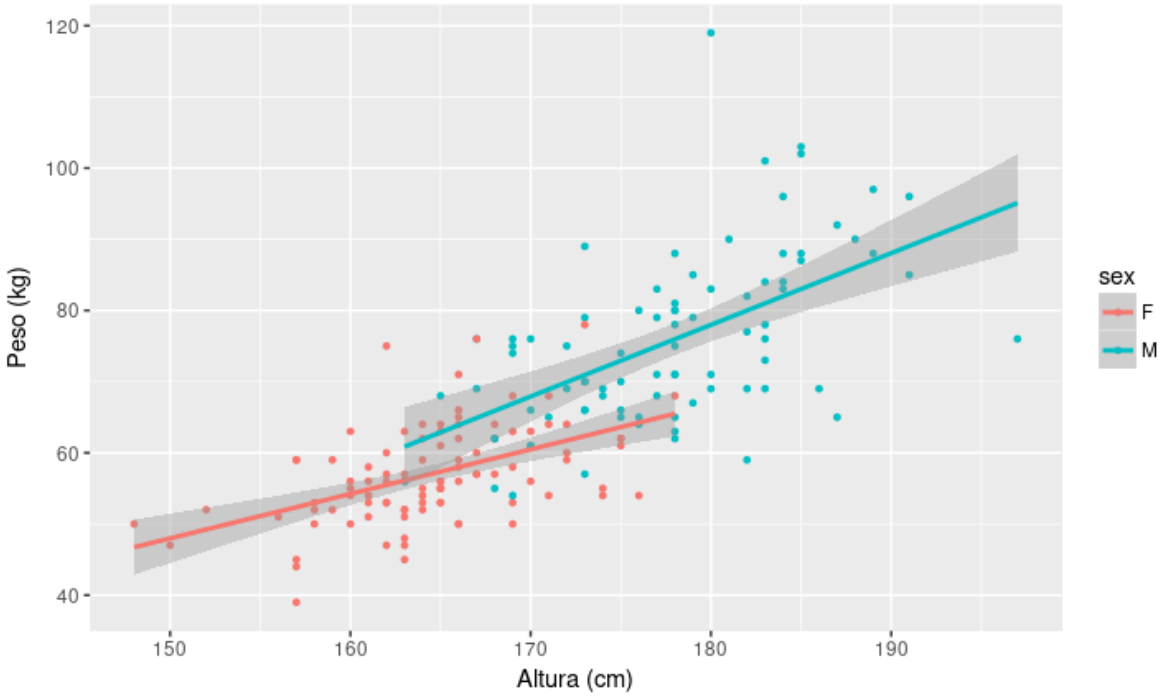
37/52

REGRESSÃO: p-valor

```
pvalor = sum(resReg$psvalue >= resReg[1,1])/nrow(resReg)
pvalor
```

```
## [1] 0.001
```

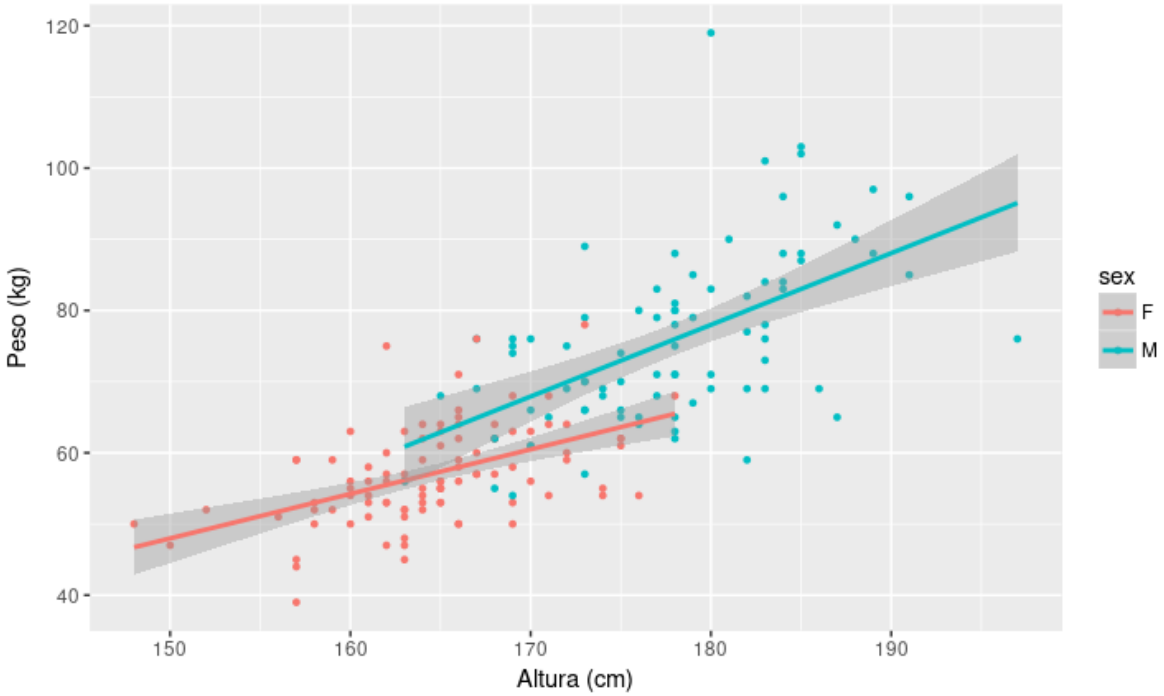
ANCOVA por reamostragem



39/52

ANCOVA por reamostragem

Que perguntas podemos fazer?

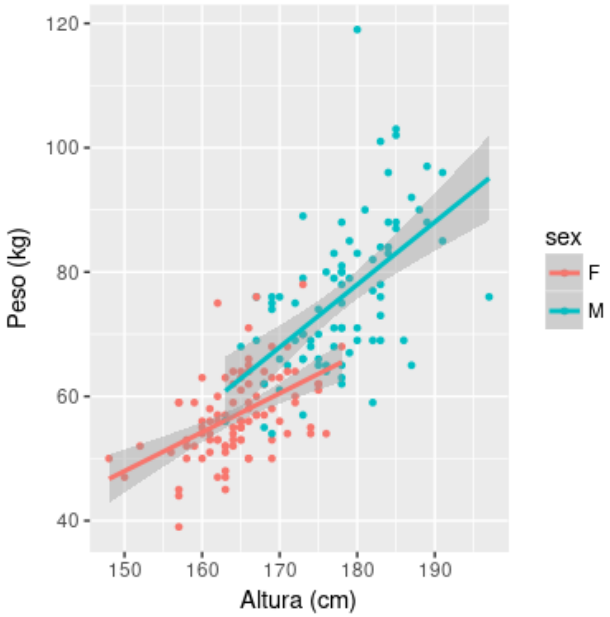


40/52

ANCOVA por reamostragem

Que perguntas podemos fazer?

1. a relação entre os sexos é a mesma, mas há um efeito de ser macho?
2. os sexos apresentam relações diferentes?

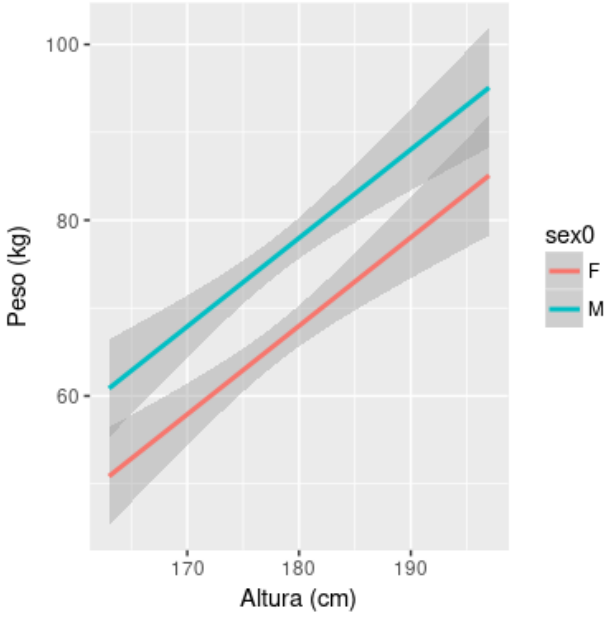


41/52

ANCOVA por reamostragem

a relação é a mesma, mas há um efeito de ser macho:

$$\hat{\alpha_m} \neq \hat{\alpha_f}$$

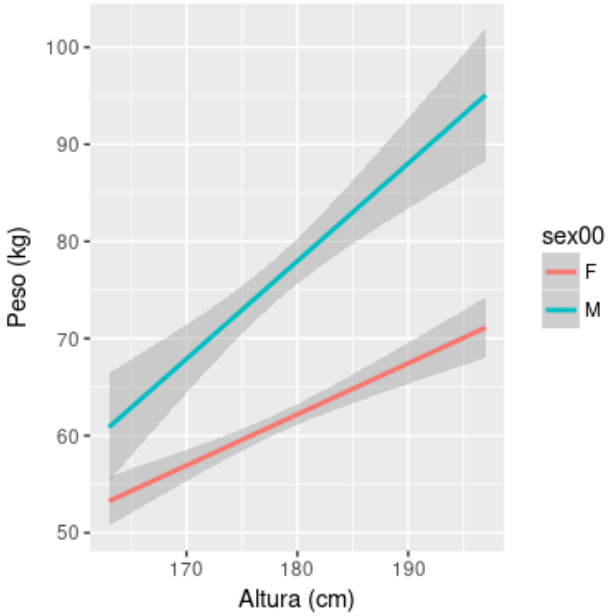


42/52

ANCOVA por reamostragem

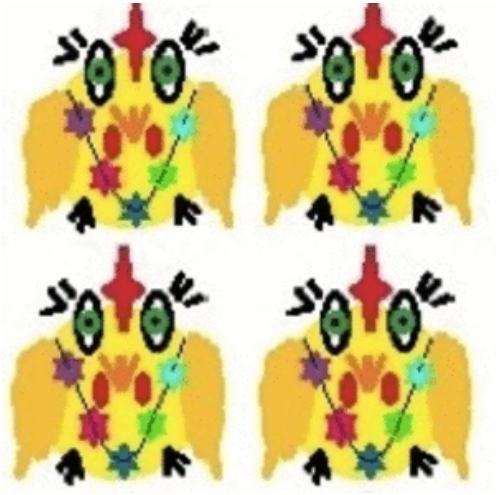
Os sexos apresentam relações diferentes:

$$\hat{\beta}_m \neq \hat{\beta}_f$$



43/52

Atividade

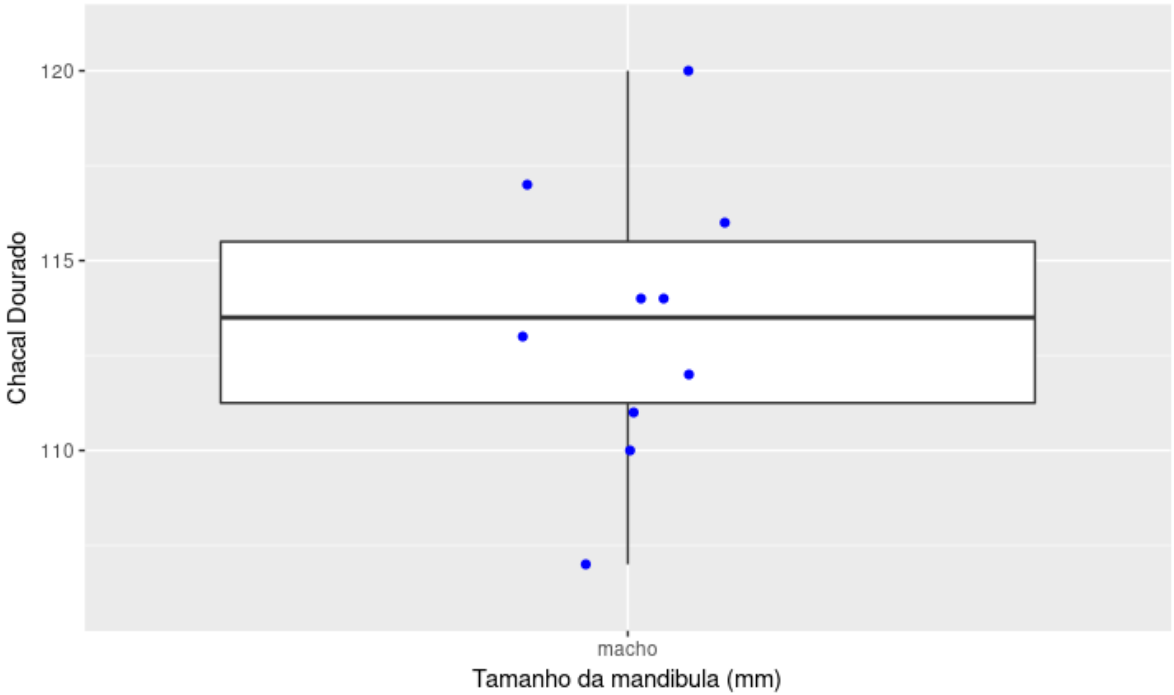


Bootstrap

Intervalo de confiança por percentil

```
##    sexo tam.mand
## 1 macho      120
## 2 macho      107
## 3 macho      110
## 4 macho      116
## 5 macho      114
## 6 macho      111
```

Chacal macho: resultados



Qual a minha confiança sobre uma estimativa?

intervalo de confiança da média

```
mchacal$tam.mand;
```

```
##  [1] 120 107 110 116 114 111 113 117 114 112
```

```
sample(mchacal$tam.mand, replace = TRUE)
```

```
##  [1] 114 114 116 111 111 111 113 112 117 117
```

```
sample(mchacal$tam.mand, replace = TRUE)
```

```
##  [1] 113 112 114 113 114 111 114 113 120 117
```

```
sample(mchacal$tam.mand, replace = TRUE)
```

```
##  [1] 111 110 114 107 120 116 114 107 120 111
```

47/52

Qual a minha confiança sobre uma estimativa?

```
mean(mchacal$tam.mand)
```

```
## [1] 113.4
```

```
mean(sample(mchacal$tam.mand, replace = TRUE))
```

```
## [1] 113.9
```

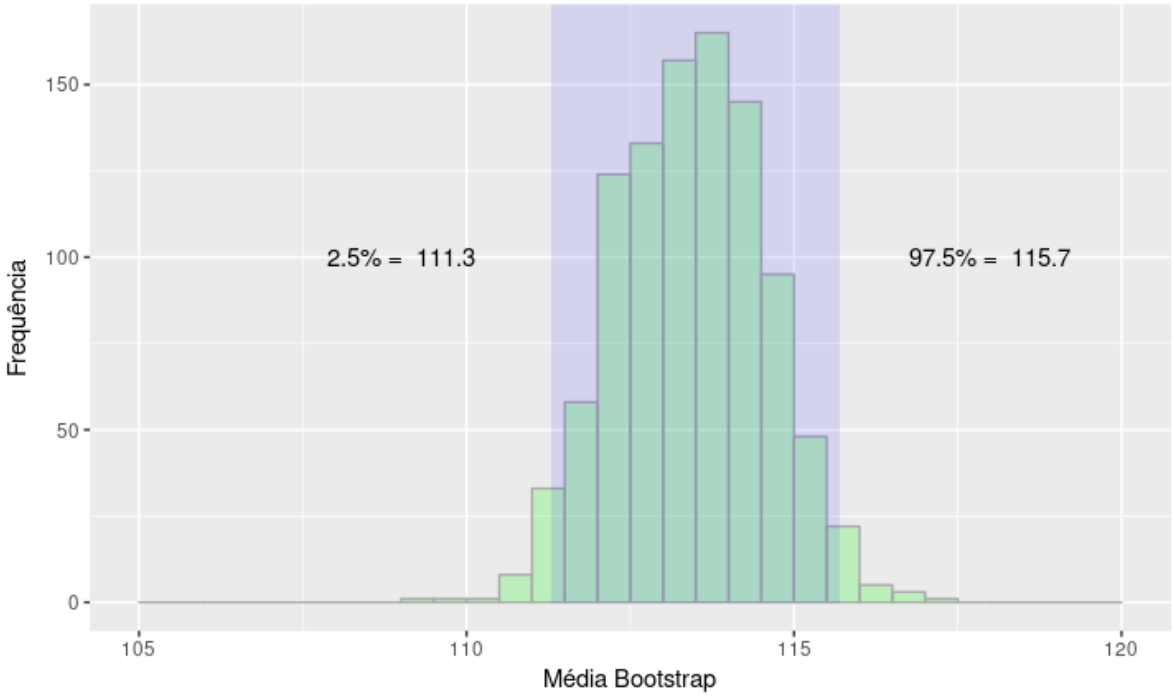
```
mean(sample(mchacal$tam.mand, replace = TRUE))
```

```
## [1] 113.8
```

```
mean(sample(mchacal$tam.mand, replace = TRUE))
```

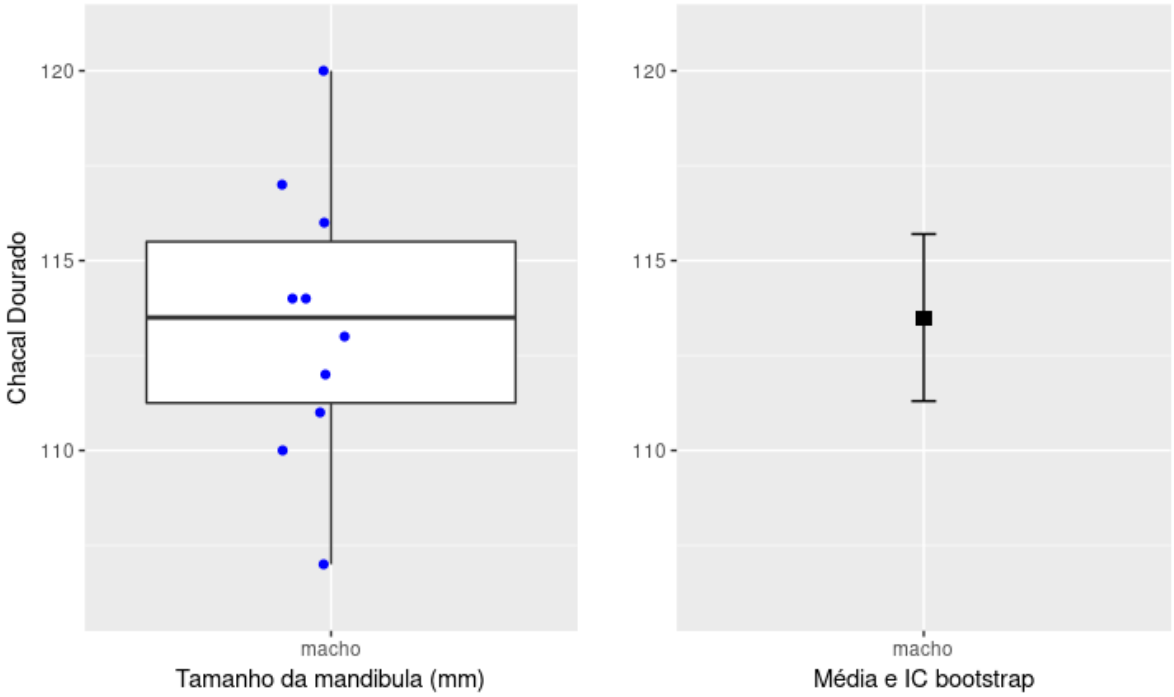
```
## [1] 113.2
```

Bootstrap



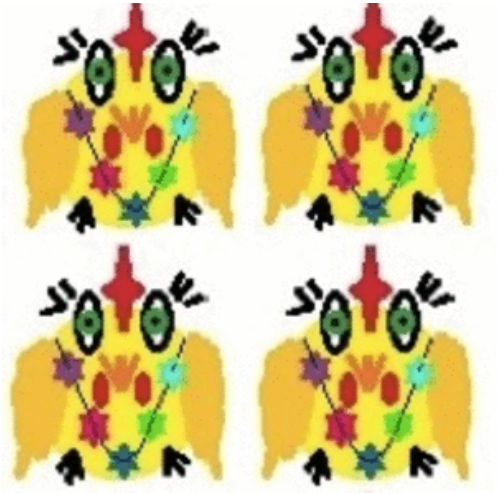
49/52

Chacal macho: Resultado

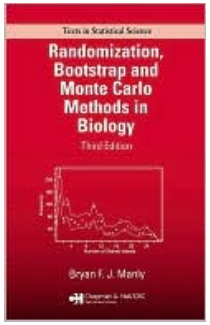


50/52

Atividade



Bibliografia



52/52