Inferência estatística para ciência de dados

Paulo Justiniano Ribeiro Jr

Curso de Especialização em
Data Science & Big Data
Universidade Federal do Paraná

30 de junho de 2018

Recap: Verossimilhança

Em uma população (considerada *infinita*) uma proporção θ de indivíduos apresenta determinada característica.
Deseja-se (inferências):

- estimar θ,
- expressar a incerteza sobre esta estimativa,
- verificar se θ (e portanto a população) está dentro de normas/referências (proporção max. de 15%) ou se há evidências de um desvio "relevante" (significativo).

Dados de *uma* amostra (considerada aleatória):
$n = 80$ e $y = 17$
Como proceder?

Paradigmas e métodos de inferência

Objetivos:
Estimativa de θ,
expressão da incerteza,
opinião em relação a valor de interesse $\theta = 0.15$
Abordagens:

- frequentista,
- verossimilhança,
- bayesiano.

Abordagem frequentista

Inferência se baseia na distribuição amostral

Os textos "usuais" nos ensinam:

$$X \sim \mathrm{B}(n, \theta)$$

$$p = \hat{\theta} \sim \mathrm{N}(\mu = \theta, \sigma^2 = \frac{\theta(1-\theta)}{n})$$

esta é a distribuição amostral

$$\mathrm{IC}: p \pm z_{1-\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}}$$

$$\mathrm{TH}: (\theta > \theta_0): z = \frac{p - \theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} \sim \mathrm{N}(0, 1)$$

Obtenção do IC: θ e $\hat{\theta} = p$:

Usa-se $\theta = 0,5$ ou $\theta = \hat{\theta}$

Inferência para proporção - frequentista

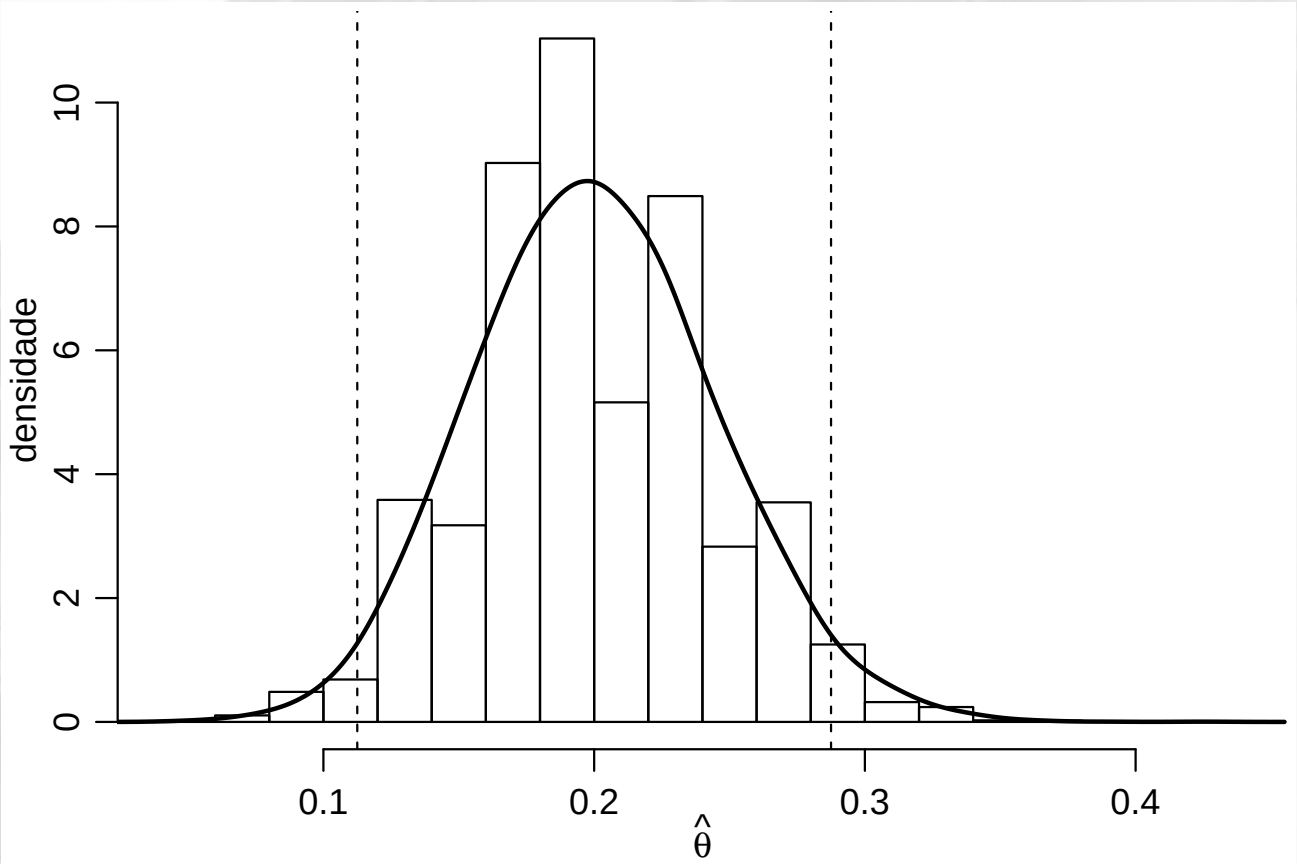
Na prática, com recursos computacionais

```
prop.test(17, 80)$conf


## [1] 0.1320616 0.3210584
## attr(,"conf.level")
## [1] 0.95


prop.test(17, 80, p=0.15, alt="greater")


##
##  1-sample proportions test with continuity correction
##
## data:  17 out of 80, null probability 0.15
## X-squared = 1.9853, df = 1, p-value = 0.07942
## alternative hypothesis: true p is greater than 0.15
## 95 percent confidence interval:
##  0.1420501 1.0000000
## sample estimates:
##      p
## 0.2125
```

Distribuição amostral obtida por simulação

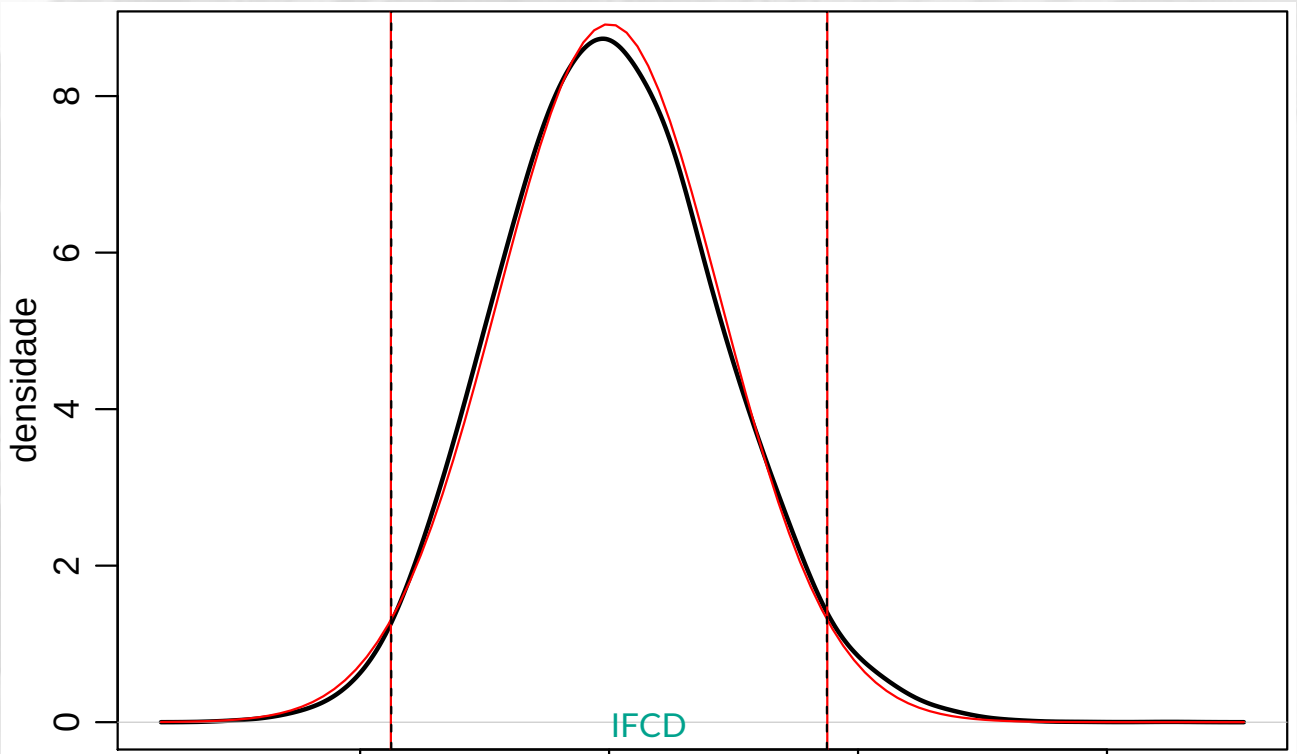
(código completo em arquivo aula07.R)

```
summary(ps)


##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0500  0.1750  0.2000  0.2005  0.2250  0.4250
```



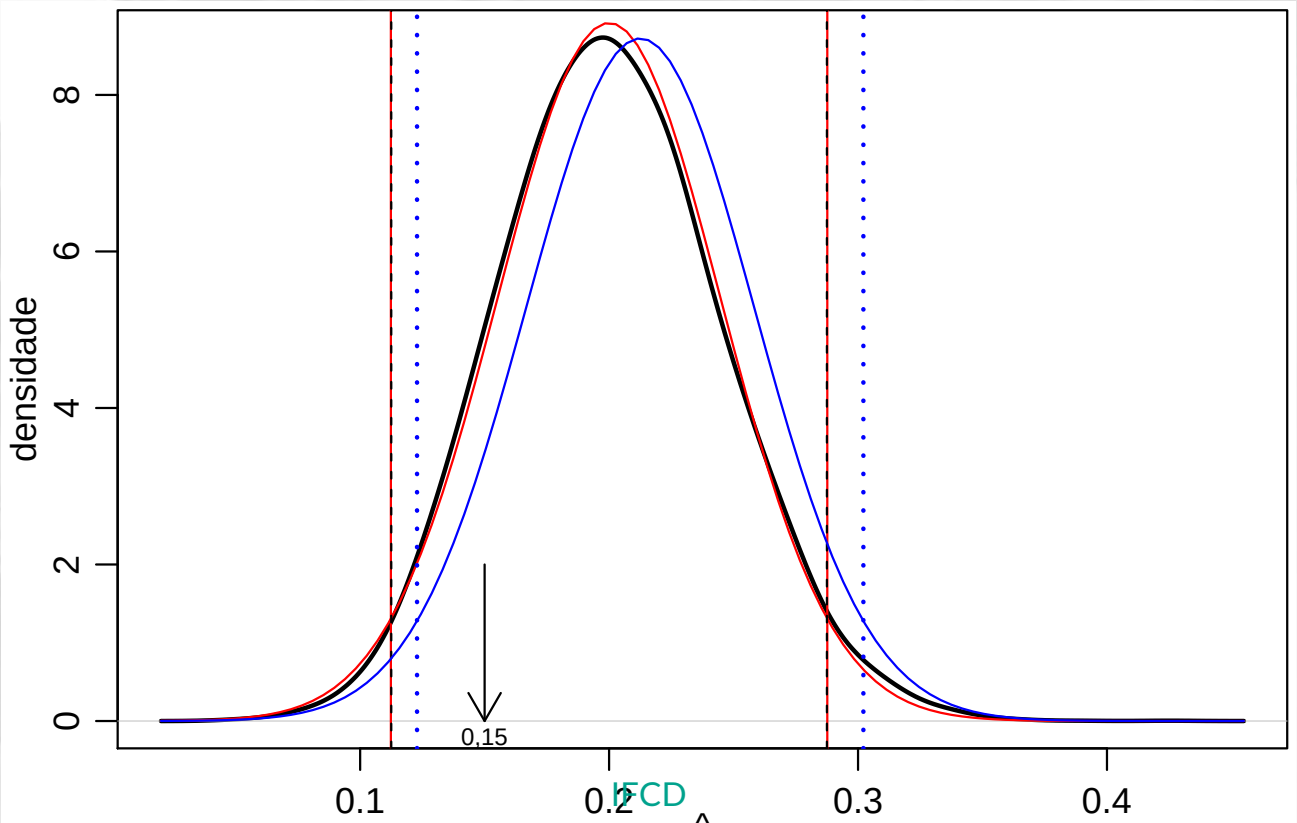
Inferência frequentista

- a distribuição amostral pode ter uma ”aparência” de alguma distribuição conhecida,
- a distribuição amostral pode ser deduzida em alguns casos chegando à distribuição conhecida,
- se tal distribuição é identificada, obtemos a distribuição amostral (e portanto pode-se fazer inferências) mesmo sem obter as diversas amostras da população.



Inferência frequentista

- Mas ainda temos um problema: não conhecemos θ.
- Usamos uma distribuição estimada com $p = \hat{\theta}$.
- Obtemos o IC nesta distribuição, que tem uma certa probabilidade (nível de confiança) de conter o valor verdadeiro do parâmetro.
- Notar diferentes $P[\hat{\theta} < 0.15]$.



Uma alternativa frequentista: Teste aleatorizado

Ideia básica:
Reproduzir a essência da idéia frequentista porém obtendo a distribuição amostral por simulação sob H_0

Algorítmo:

- Simular amostras da população sob H_0
- Calcular o valor de interesse ou estatística de teste para cada amostra simulada
- valor-p proporção destes que são mais “extremos” do que o valor observado na amostra

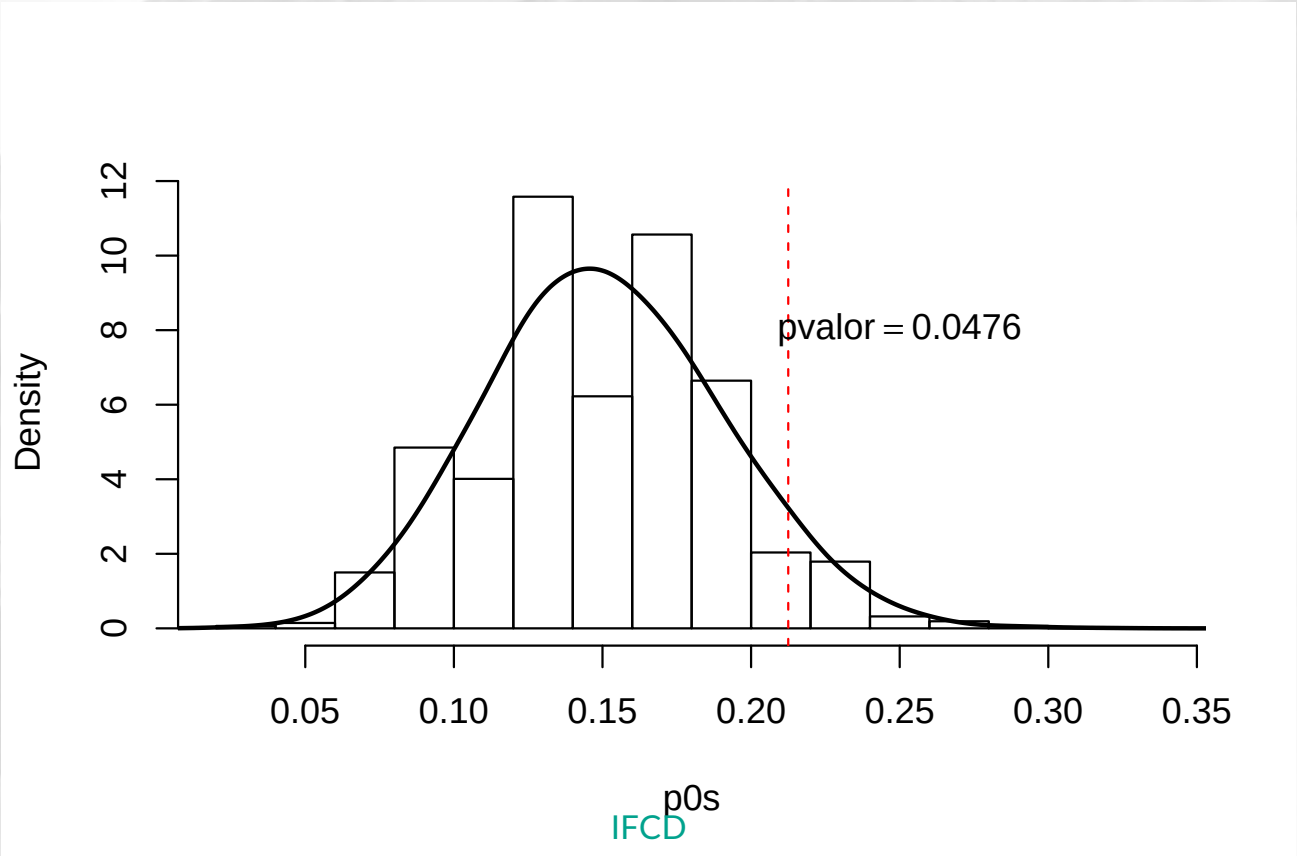
Teste aleatorizado

```
summary(p0s)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0250  0.1250  0.1500  0.1509  0.1750  0.3250
```

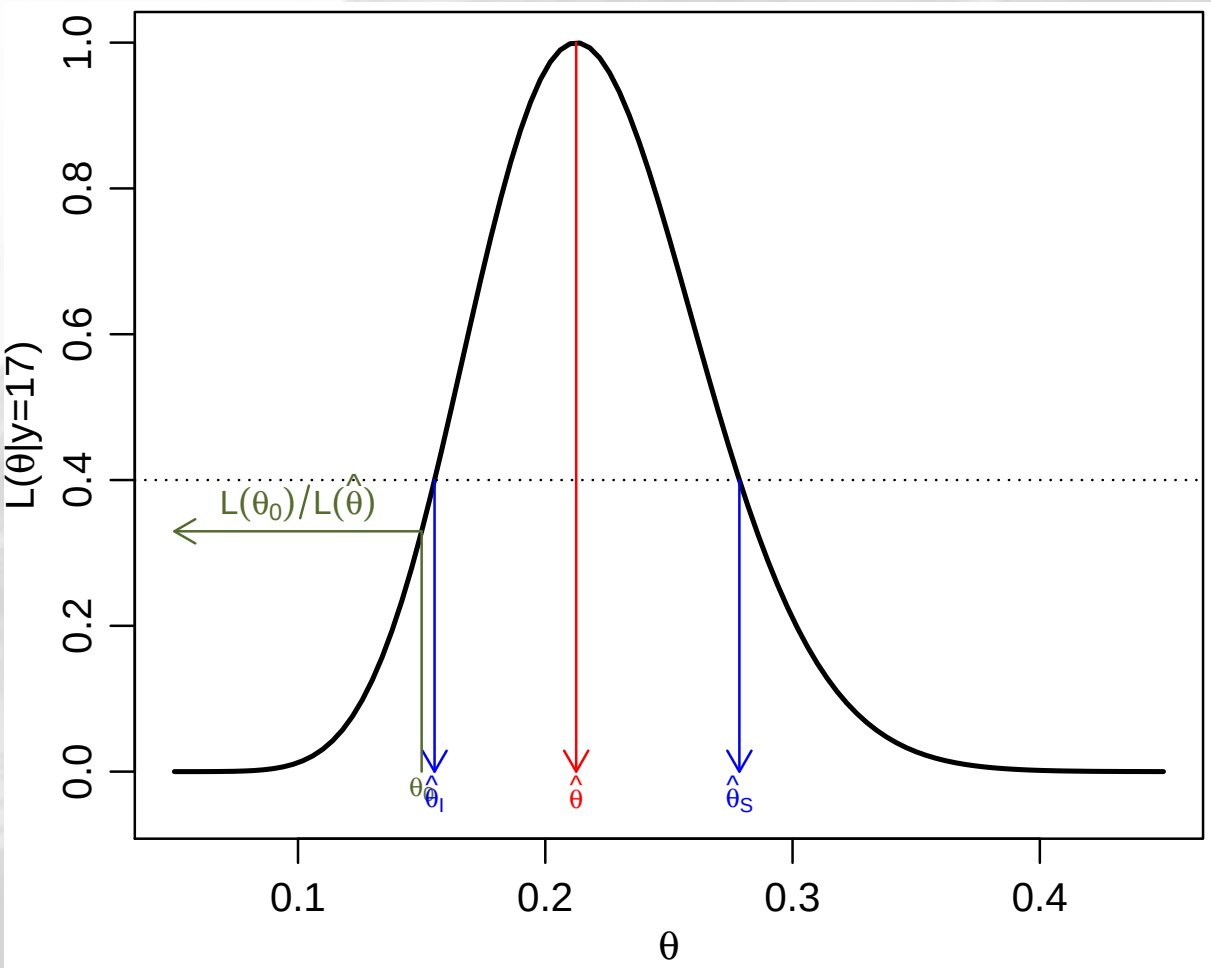
```
(pvalor <- mean(p0s > 17/80))
```

```
## [1] 0.0476
```



Abordagem pela verossimilhança

Inferência é baseada nas características da função de verossimilhança



Inferência frequentista

Em resumo:

- Estimativa de θ: fornecido por algum método de estimação
- expressão da incerteza: variabilidade da distribuição amostral
- opinião em relação a valor de interesse $\theta = 0.15$: probabilidade na distribuição amostral

Inferência pela verossimilhança

- Estimativa de θ: máximo (supremo) da função
- expressão da incerteza: faixa de valores dentro de um limite de compatibilidade com a amostra, curvatura da função
- opinião em relação a valor de interesse $\theta = 0.15$: comparação da verossimilhança deste valor com a do máximo

Inferência pela verossimilhança

Necessidade de critérios:

- definir o valor para corte da função para obter intervalos de confiança (IC's) ?
- definir limiar para o valor de verossimilhança (relativa ao máximo) para θ_0?

Possíveis soluções:

- critérios de razoabilidade e comparação (e.g. moedas ou lembre da família do Sr. João!)
- argumento frequentista (comportamento "médio" da verossimilhança) estabelece relações:

r	$P[\lvert Z\rvert < \sqrt{c^*}]$
50%	0,761
26%	0,899
15%	0,942
3,6%	0,990

Inferência Bayesiana

O objeto de inferência é a distribuição à posteriori

- A incerteza inicial sobre θ é expressa na forma de uma distribuição priori para θ
- Com amostra atualizamos opinião θ com a informação contida na verossimilhança
- O conhecimento/incerteza atualizados sobre θ é expresso pela distribuição posteriori

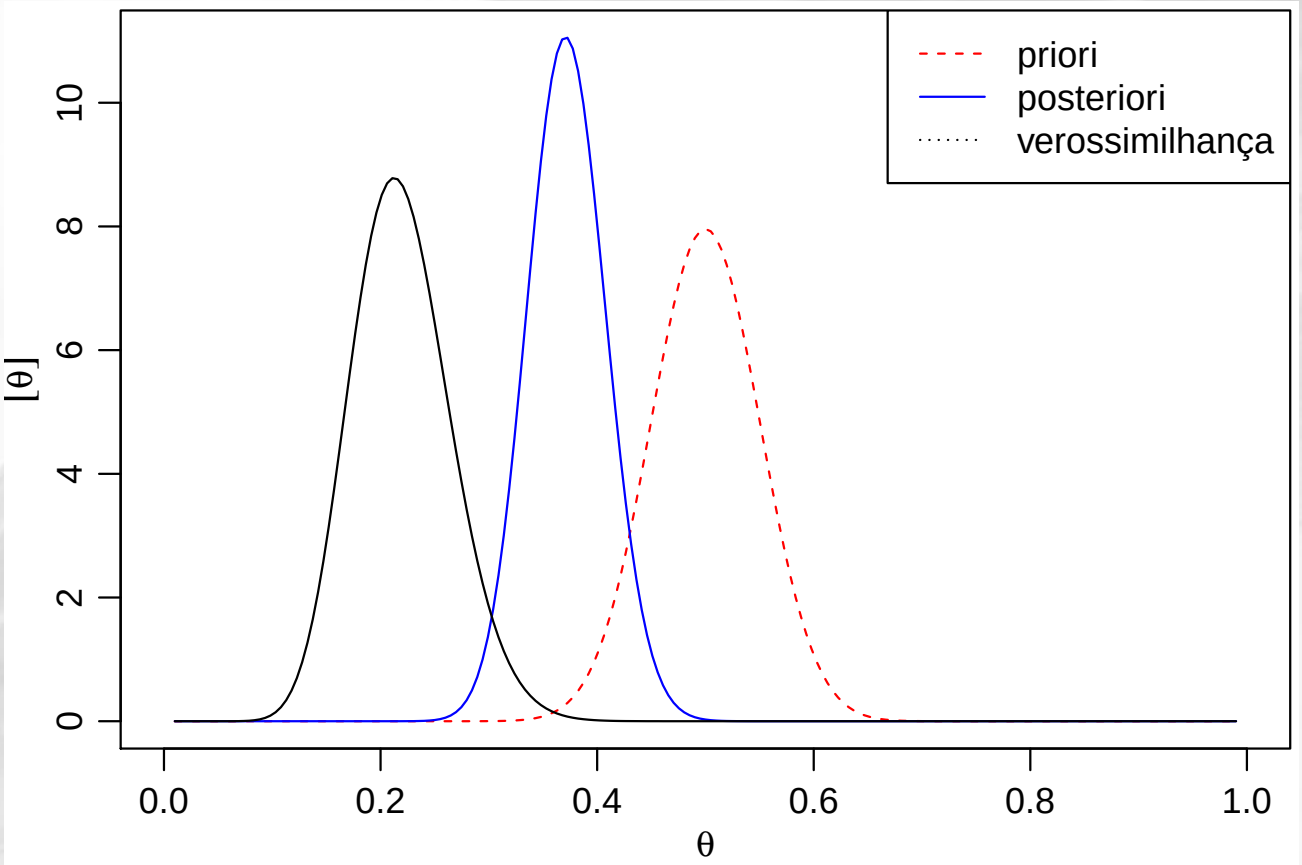
Formalmente:

$$f(\theta|y) \propto f(\theta) \cdot L(\theta|y)$$

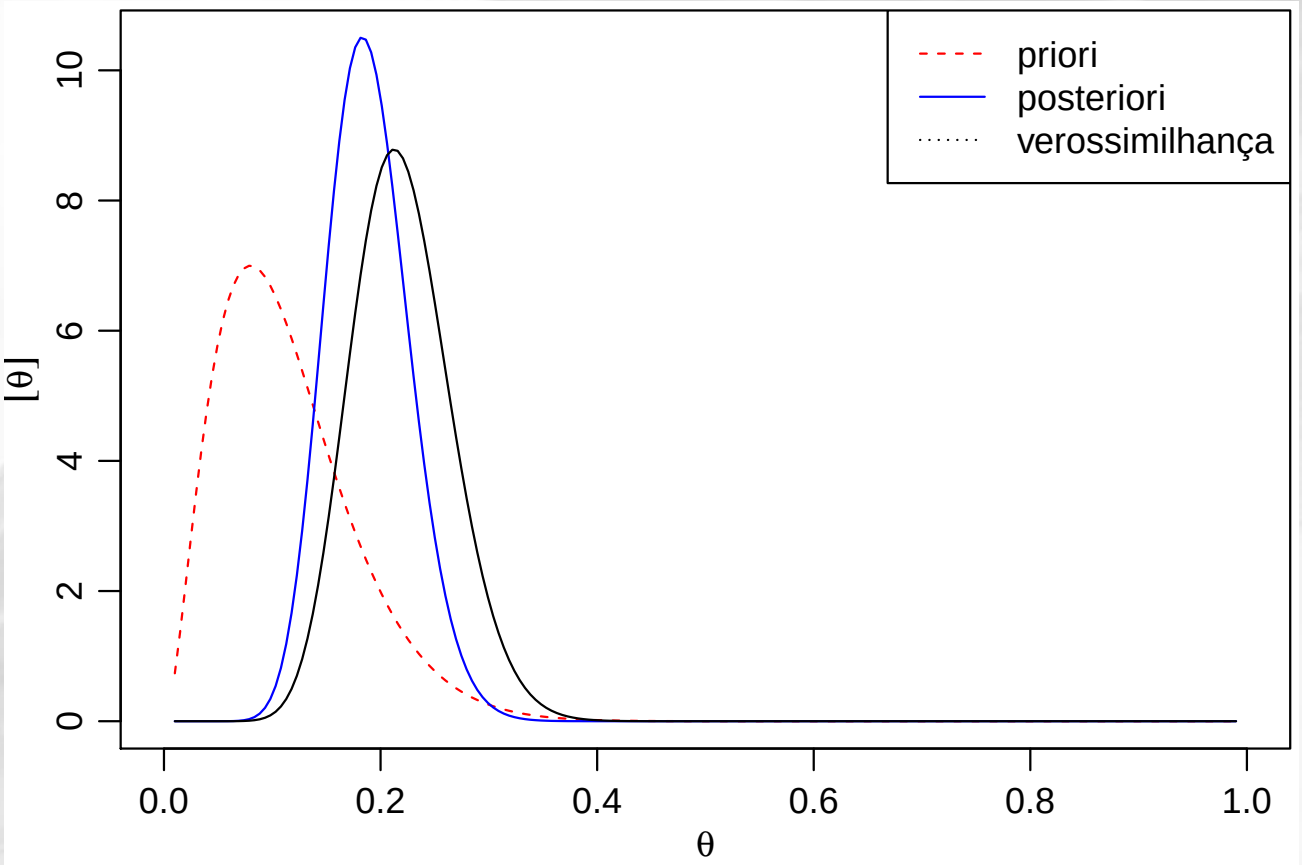
ou, usando jargão técnico:

$$\textit{posteriori} \propto \textit{priori} \cdot \textit{verossimilhança}$$

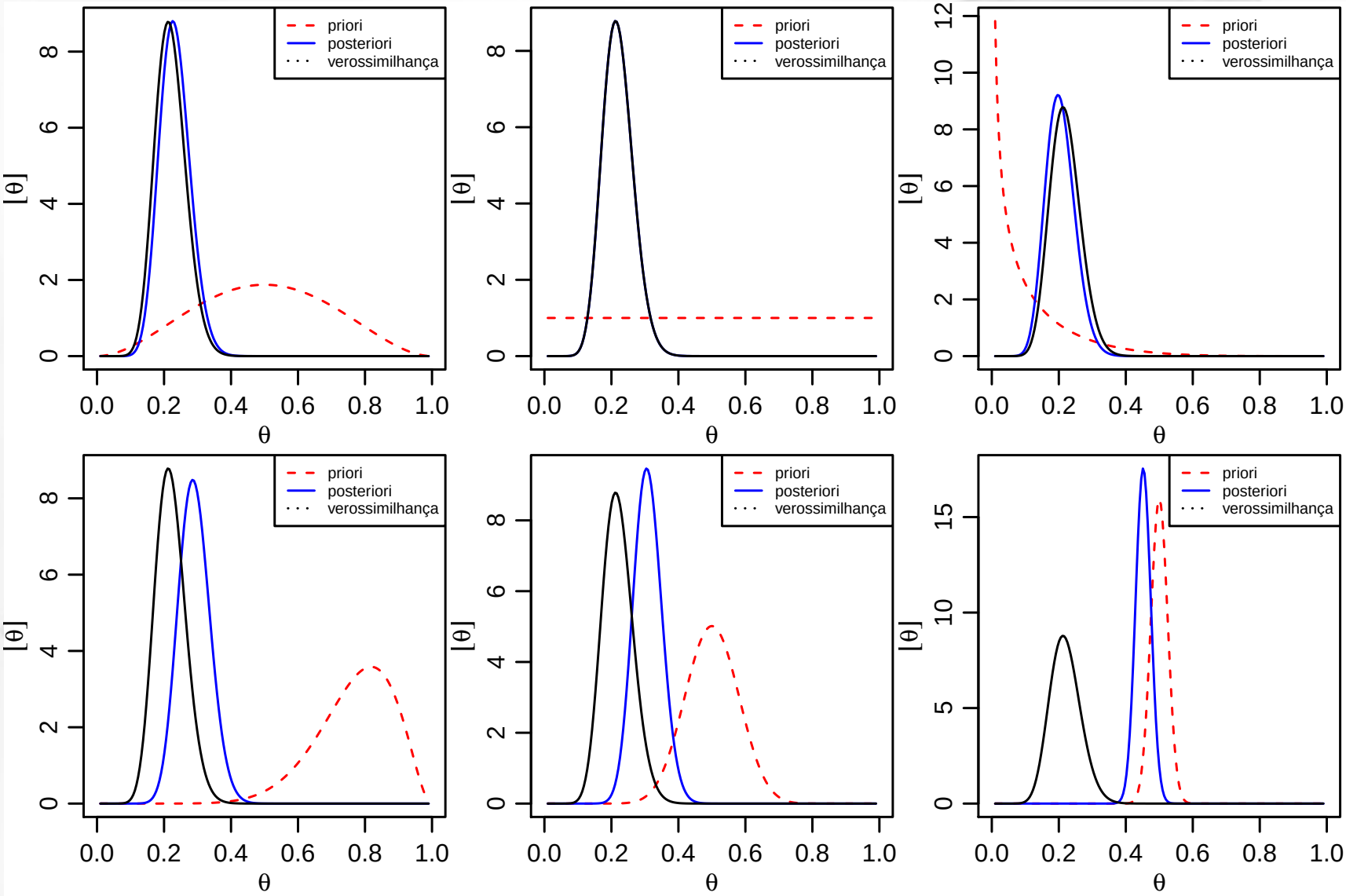
A essência de Bayes ilustrada (I)



A essência de Bayes ilustrada (II)



A essência de Bayes ilustrada (III)



Comentários

- Expressão da opinião “a priori” é necessária e sua especificação é um desafio,
- as interpretações de intervalo de confiança são agora probabilísticas, por exemplo pode-se falar em:

$$P[a < \theta < b] = 0.95$$

- bem como, no contexto do exemplo, pode-se falar em

$$P[\theta \leq 0,15]$$

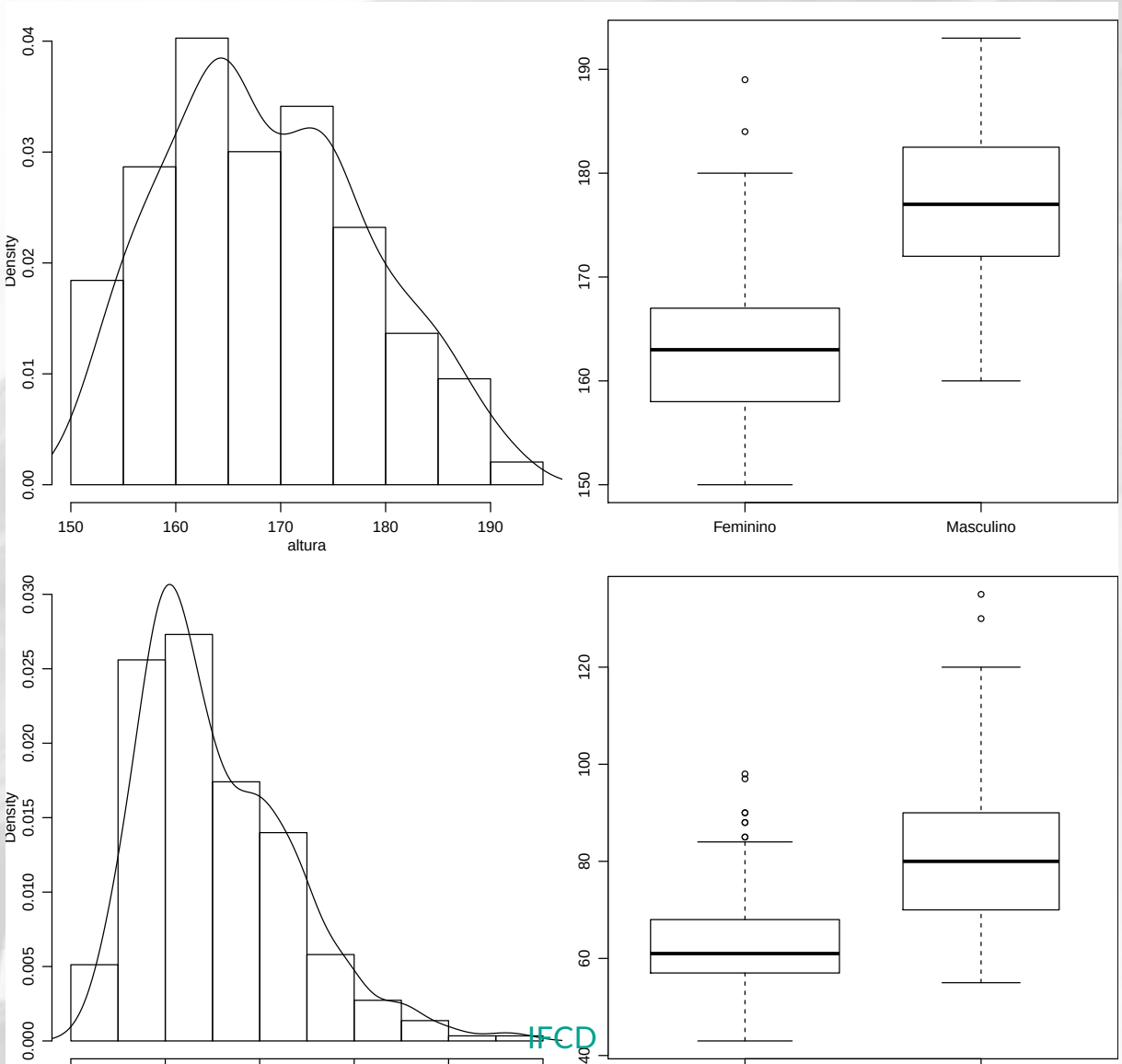
Inferência Bayesiana

Em resumo:

- Estimativa de θ: alguma medida resumo da posteriori (média, moda, mediana, ...
- expressão da incerteza: variabilidade da distribuição posteriori
- opinião em relação a valor de interesse $\theta = 0.15$: probabilidade na posteriori

Comparando dois grupos

Comparando dois grupos: alturas e pesos de homens e mulheres (AULA 6)

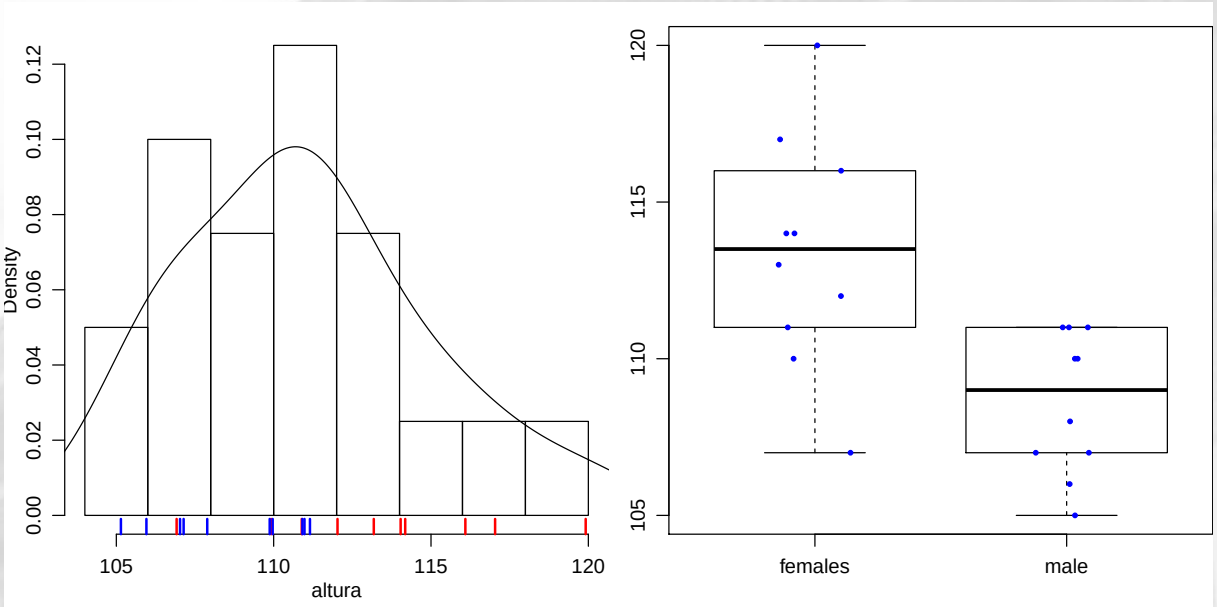


Um outro exemplo

Comparando dois grupos: comprimento da mandíbula (chacal dourado)

```
##    females male
## 1      120  110
## 2      107  111
## 3      110  107
## 4      116  108
## 5      114  110
## 6      111  105
## 7      113  107
## 8      117  106
## 9      114  111
## 10     112  111
```

Um outro exemplo



Um outro exemplo

- Frequentista: teste-*t*, opções e limitações (formulário)
- Teste aleatorizado – algorítmo
- Verossimilhança (perfilhada) para quantidade de interesse
- Bayesiana – distribuição marginal para parâmetro de interesse

O teste-*t* (1)

```
with(mandible, t.test(females, male))
```

```
##
##  Welch Two Sample t-test
##
## data:  females and male
## t = 3.4843, df = 14.894, p-value = 0.00336
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.861895 7.738105
## sample estimates:
## mean of x mean of y
##     113.4     108.6
```

O teste-*t* - opções

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
         mu = 0, paired = FALSE, var.equal = FALSE,
         conf.level = 0.95, ...)
```

O teste-*t* (2)

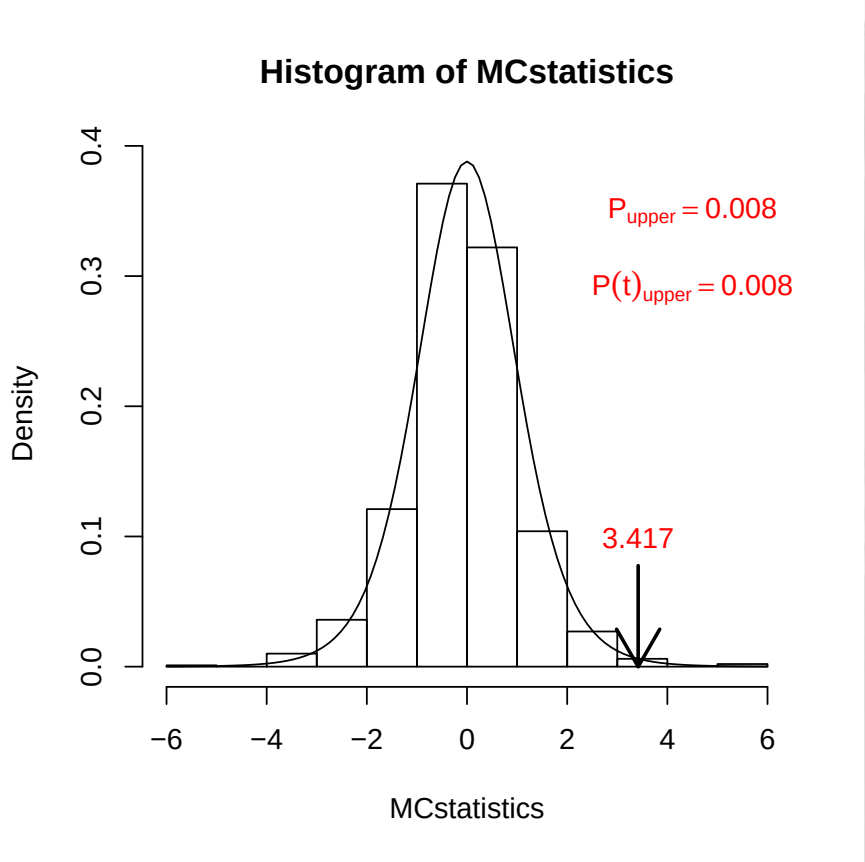
```
with(mandible, t.test(females, male, var.equal=TRUE))


##
##  Two Sample t-test
##
## data:  females and male
## t = 3.4843, df = 18, p-value = 0.002647
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.905773 7.694227
## sample estimates:
## mean of x mean of y
##     113.4     108.6
```

O teste-*t* (3)

```
with(mandible, t.test(females, male, paired=TRUE))
```

```
##
##  Paired t-test
##
## data:  females and male
## t = 3.417, df = 9, p-value = 0.007665
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.622226 7.977774
## sample estimates:
## mean of the differences
##                     4.8
```

Um outro exemplo




```
## Paired data
## data statistics =  3.416968784709
##
## probabilities based on Monte Carlo simulations:
## upper.tail lower.tail
##   0.007992   0.992008
##
## probabilities based on the "t" distribution:
## upper.tail lower.tail
##   0.003832   0.996168
```

Comparando modelos e verossimilhança

Comparando modelos para os dados

$$Y_{ij} \sim \mathrm{N}(\mu, \sigma^2)$$

$$Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2)$$

$$Y_{ij} \sim \mathrm{N}(\mu_i, \sigma_i^2)$$

```
L1 <- lm(mandible~1, data=mand)
L2 <- lm(mandible~sex, data=mand)
c(logLik(L1), logLik(L2))


## [1] -54.98137 -49.82638


anova(L1, L2)


## Analysis of Variance Table
##
## Model 1: mandible ~ 1
## Model 2: mandible ~ sex
##   Res.Df   RSS Df Sum of Sq     F   Pr(>F)
## 1     19 286.0
## 2     18 170.8  1     115.2 12.14 0.002647 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Inferência Bayesiana

Obtenção de distribuição posteriori para a diferença de médias
(Não será detalhado aqui)

Abordagem geral e generalizando

Generalizações . . . muitas possíveis
Vamos começar reescrevendo

$$Y_{ij}^{(\lambda)} \sim \mathrm{N}(\mu_{ij}, \sigma_{ij}^2)$$
$$g(\mu_{ij}) = f(x_{ij}, \beta)$$
$$g(\sigma_{ij}^2) = f(z_{ij}, \varphi)$$

Abordagem geral e generalizando

- Teste-*t* (amostras independentes, variâncias iguais)

$$
\begin{aligned}
Y_{ij} &\sim \mathrm{N}(\mu_{ij}, \sigma^2_{ij}) \\
\mu_{ij} &= \beta_o + \beta_1 x_{sex} \\
\sigma^2_{ij} &= \sigma^2
\end{aligned}
$$

- Mudando a distribuição

$$
\begin{aligned}
Y_{ij} &\sim \mathrm{G}(\mu_{ij}, \phi_{ij}) \\
\log(\mu_{ij}) &= \beta_o + \beta_1 x_{sex} \\
\phi_{ij} &= \phi
\end{aligned}
$$

Abordagem geral e generalizando

Modelos de regressão (linear) : exemplo do início do curso:
$Y = t$ (tempo) e $x = \sqrt{d}$ (distância)

$$Y_{ij} \sim \mathrm{N}(\mu_{ij}, \sigma^2_{ij})$$
$$\mu_{ij} = \beta_o + \beta_1 x \quad (\mathrm{ou} \mathrm{X}\beta)$$
$$\sigma^2_{ij} = \sigma^2$$

Generalizações

- Regressão linear múltipla
- Regressão para variável transformada
- Regressão heterocedástica
- Modelo linear generalizado
- *splines*
- Regressão não linear
- Modelo aditivo generalizado
- ...