Capítulo 4: técnicas de reamostragem

Matheus Victal Cerqueira

26/05/2021

Introdução

O conceito de reamostragem está ligado ao método de obter-se amostras de uma amostra inicial, sem que seja necessária a obtenção de novas amostras da população original.

Função Distribuição Empírica (fde)

A fde ($\hat{F}(x)$) se trata de uma função de distribuição acumulada obtida a partir dos valores observados de uma amostra da váriavel aleatória de interesse X. Sendo $X_1, \cdots, X_n$ uma a.a. de X, tem-se:

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x),\ x \in \mathbb{R}$$
$$\mathbb{I}(X_i \leq x) = \begin{cases} 1, X_1 > x \\ 0, X_i \leq x \end{cases}$$

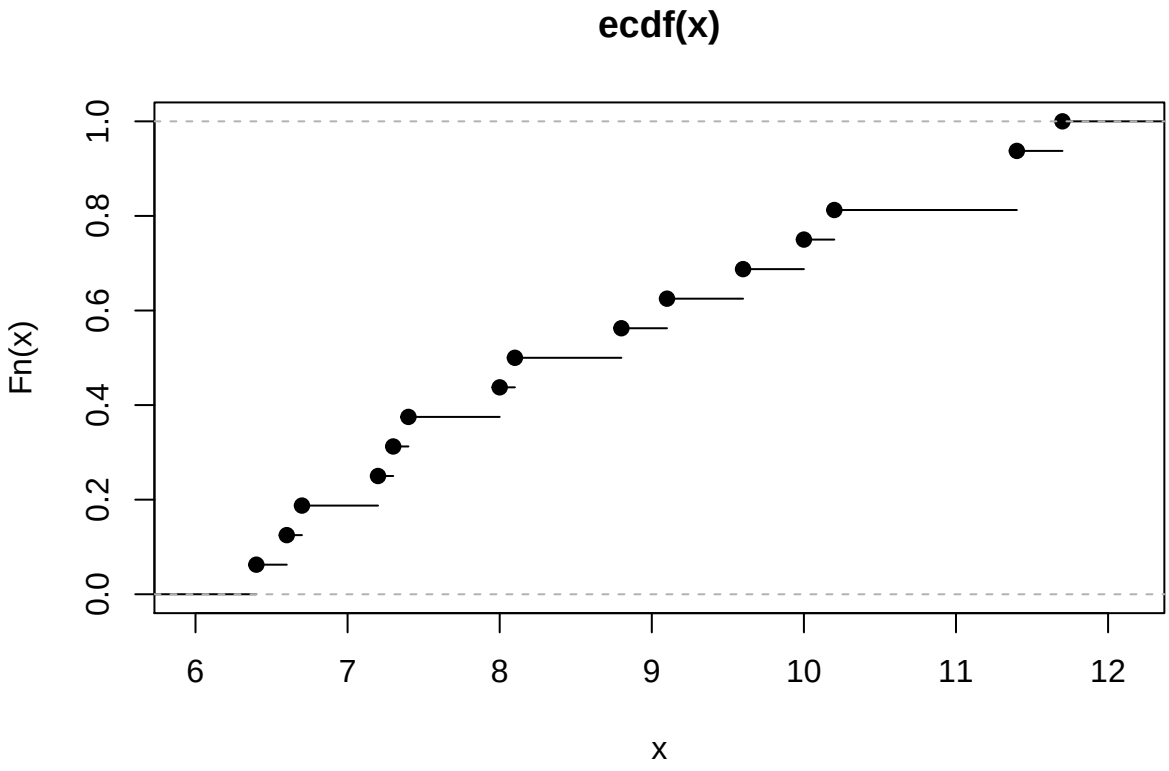
Tal função não paramétrica é utilizada para processos de reamostragem, já que ela representa o comportamento de X através de uma a.a.. $\hat{F}(x)$ possui todas as propriedades de uma fda. Os seus pontos de descontinuidade são exatamente os pontos observados da reta real na amostra, sendo sempre descreta. Pode-se mostrar que $\hat{F}(x)$ é um estimador não viesado e consistente de $F(x)$, mostrando-se como um bom estimador para a verdadeira fda. $\hat{F}(x) \longrightarrow F(x)$ pela lei forte dos grandes números, mostrando que tal estimador é uma função adequada para ser utilizada no lugar da verdadeira fda quando necessário.

Abaixo segue um exemplo de obtenção de uma fde a partir de uma amostra.

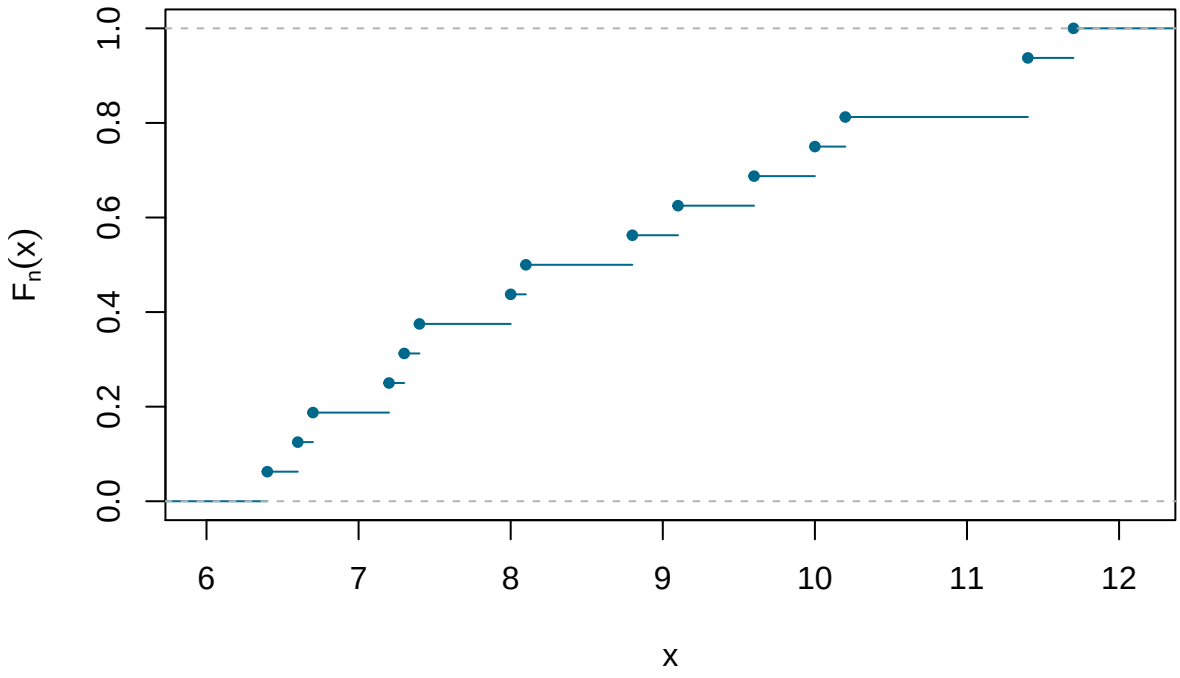
```
rm(list=ls(all=TRUE))

x <- c(6.7,  6.6,  6.4,  8.8,  7.4,  8.0, 10.0,  7.3, 11.7,
       10.2, 11.4, 8.1,  7.2, 11.4,  9.6,  9.1)

length(x)
```

```
## [1] 16
```

```
# De forma simples:
plot(ecdf(x))
```




```
# Um pouco mais elaborado:
plot(ecdf(x) , main = "", ylab = expression(hat(F)[n](x)), pch = 20,
     col = "deepskyblue4")
```




```r
#Atribuição da fde à um objeto
Fn <- ecdf(x)

# Valores de x onde ocorre descontinuidade
knots(Fn)
```

```
##  [1]  6.4  6.6  6.7  7.2  7.3  7.4  8.0  8.1  8.8  9.1  9.6 10.0 10.2 11.4 11.7
```

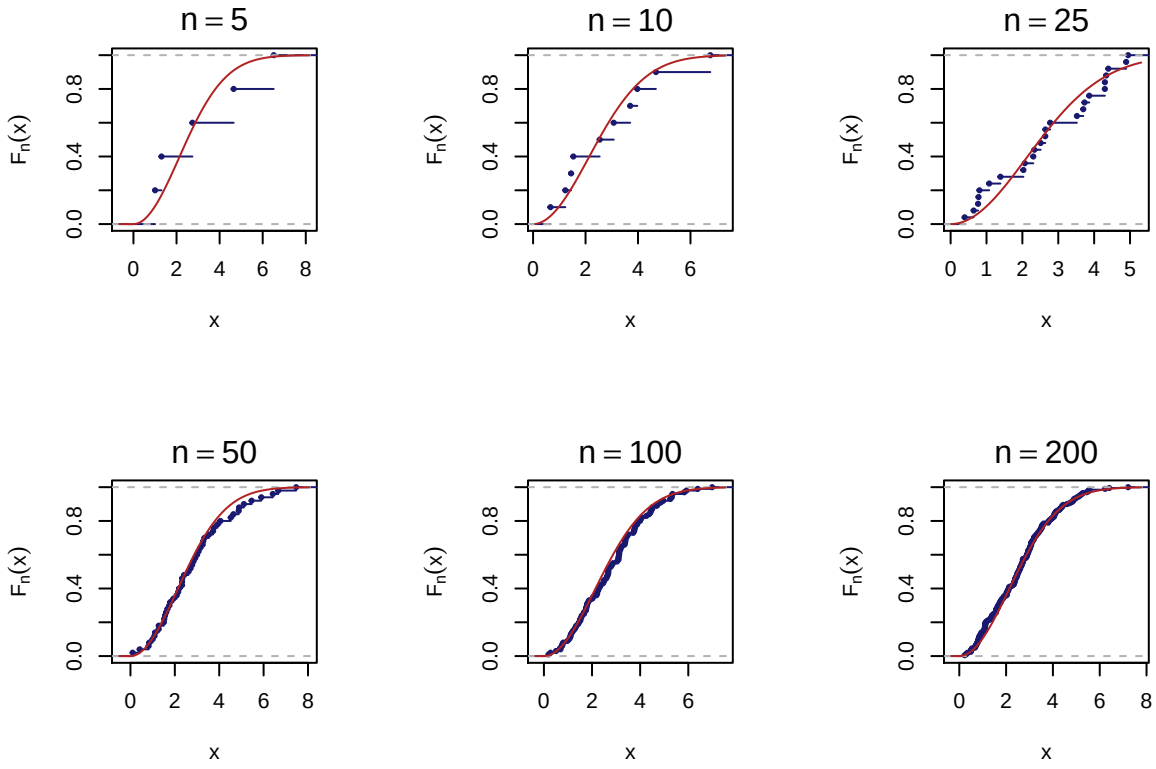
```r
# Função Fn calculda em alguns pontos
Fn(c(-1, 7.5, 11, 20))
```

```
## [1] 0.0000 0.3750 0.8125 1.0000
```

Vejamos um exemplo com uma variável aleatória Weibull.

```r
# Exemplo com diferentes tamanhos de amostra
# Distribuição Weibull(forma = 2, escala = 3)
n <- c(5, 10, 25, 50, 100, 200)
par(mfrow = c(2, 3))
par(mai = c(1, 1, 0.3, 0.1))
for (tamanho in n) {
  # Geração de observações da distribuição Weibull
  dados <- rweibull(tamanho, shape = 2, scale = 3)
```

```r
  # Plotar a fde criada a patir da amostra
  plot(ecdf(dados) , main = bquote(n == .(tamanho)), col = "midnightblue",
       ylab = expression(F[n](x)), pch = 20, cex.main = 1.5)

  # Plotar a curva referente à verdadeira F(x)
  curve(pweibull(x, shape = 2, scale = 3), add = TRUE, col = "firebrick")
  }
```



É notório que as propriedades da função de distribuição acumulada empírica permitem que ela seja um bom substituto para a função de distribuição acumuada teórica, podendo ser utilizada para a obtenção de pseudo-amostras.

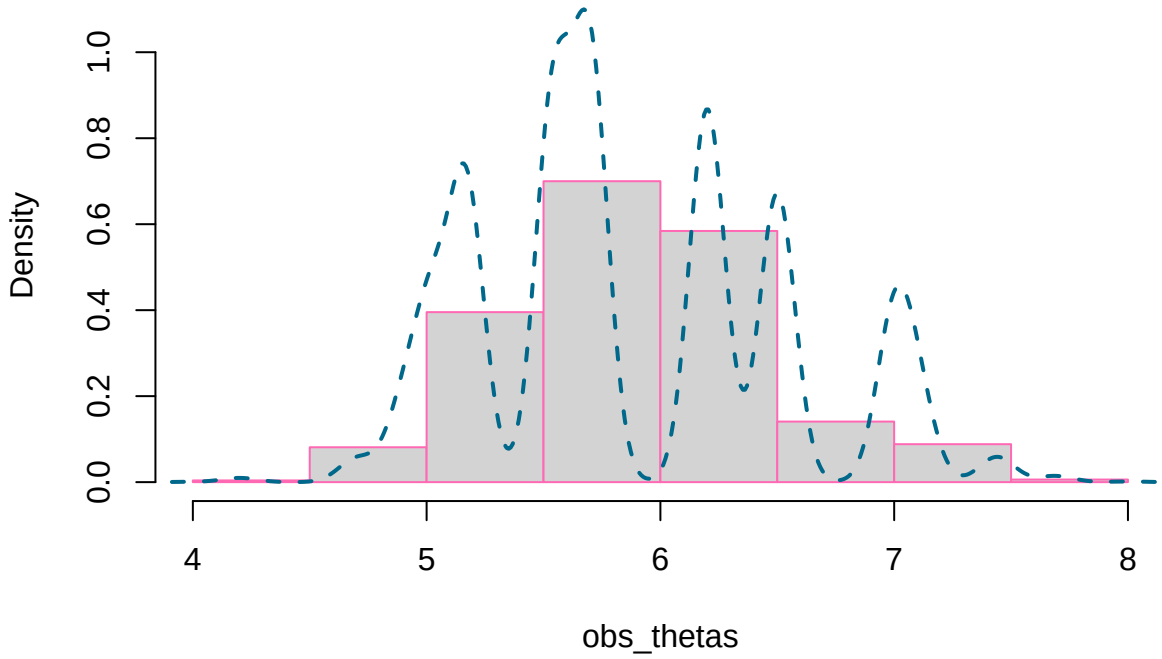
Método Bootstrap

O método bootstrap consiste na obtenção de pseudo-amostras a partir de uma amostra aleatória inicial do problema de interesse.

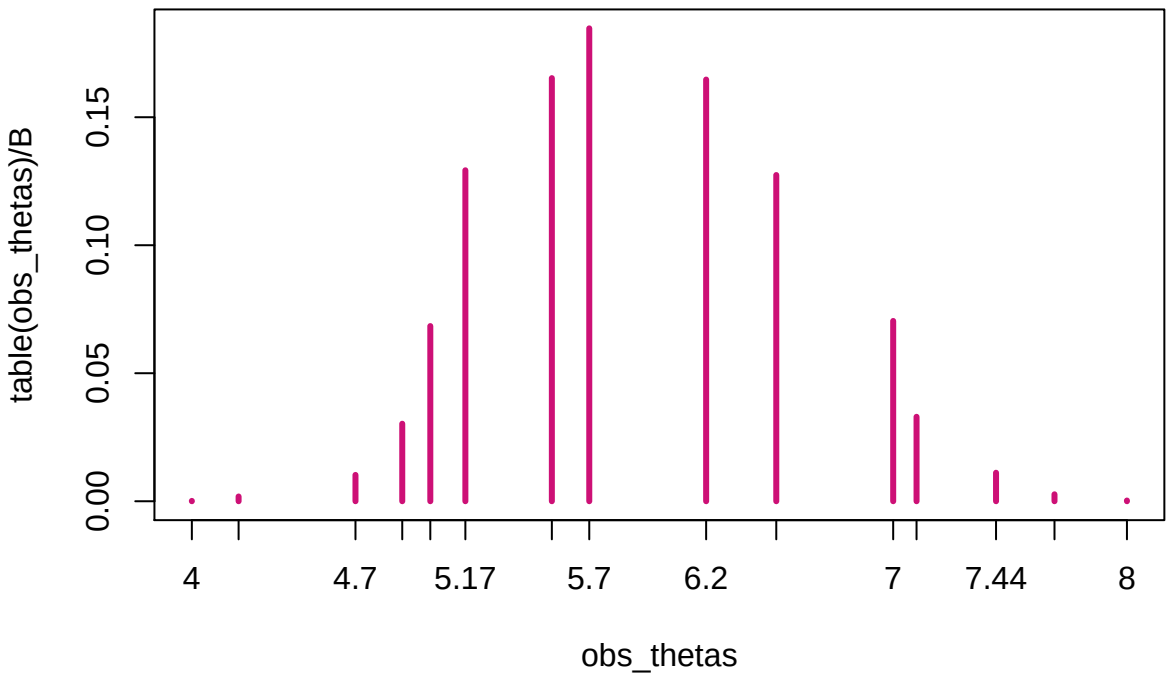
```r
# x é uma amostra aleatória obtida de uma população original de interesse

set.seed(2112)
x <- c(3.4, 3.5, 4, 4.2, 4.7, 4.9, 5.02, 5.17, 5.54, 5.7, 6.2, 6.5, 7, 7.1,
       7.44, 7.69, 8, 8.88, 8.89)

length(x)
```

```
## [1] 19
```

```r
# quer-se analisar o comportamento de um estimador theta dado pela mediana
theta_c <- median(x)

# Número de amostras bootstrap
B <- 7000

# vetor de the
obs_thetas <- c()

# laço de obtenção de amostras bootstrap
for(b in 1:B){
  obs_thetas[b] <- median(sample(x, replace = T))
}

hist(obs_thetas, freq = F, main = "", ylim = c(0,1.1), xlim = c(4,8), border = "hotpink")
lines(density(obs_thetas),  col = "deepskyblue4", lwd = 2, lty = 2)
```




```r
# visualização mais adequada
plot(table(obs_thetas)/B, col = "deeppink3", lwd = 3)
```




```
sd(obs_thetas)
```

```
## [1] 0.656174
```