Análise de Dados e Simulação

Márcia Branco

Universidade de Sao Paulo
Instituto de Matematica e Estatistica
http:www.ime.usp.br/m̃branco

Exercícios - bootstrap -

Ideia do Método de *bootstrap*

1. Mundo ideal

Suponha uma população $X \sim F$. Retira-se B amostras de tamanho n dessa população, obtendo-se as estimativas pontuais $T(x^1), T(x^2), \ldots, T(x^B)$ do parâmetro θ . Com esses valores é possível estimar o $EQM[T(X)] = E_F[(T(X) - \theta)^2]$ fazendo

$$\frac{1}{B} \sum_{j=1}^{B} \left[T(x^j) - \bar{T} \right]^2$$

em que $\bar{T} = \frac{1}{B} \sum_{i=1}^{B} T(x^i)$.

Se $T(X)$ for não viesado para θ.

2. Mundo da reamostragem (*bootstrap*)

Só existe uma amostra observada x_0 e vamos considera-la como sendo a nossa população. Retira-se B amostras de *Boot* usando F_e. Calcula-se $T(x_j^*)$ para $j = 1, 2, \ldots, B$. O $EQM_{Boot} = E_{Fe}[(T(X^*) - \theta)^2]$ é estimado por

$$\frac{1}{B} \sum_{j=1}^{B} \left[T(x_j^*) - \theta(F_e) \right]^2 .$$

em que $\theta(F_e)$ o valor do parâmetro obtido usando a amostra x_0.

Exercício 1: Prove que

$$EQM_{Boot}(\bar{X}_n) = \frac{(n-1)}{n}\frac{S_x^2}{n},$$

em que $\bar{X}_n$ e S_x^2 são as usuais média e variâncias amostrais.

Solução: Por definição o erro quadrático médio de *bootstrap* é

$$EQM_{Boot}(\hat{\theta}) = E_{F_e}[(\hat{\theta}(X) - \theta)^2],$$

em que F_e é a função distribuição empírica.

No nosso caso $\hat{\theta}(X) = \bar{X}_n$ e $\theta = E[X]$.

Seja $(x_1, x_2, \ldots, x_n)$ a amostra observada, a função empírica atribui probabilidades iguais a cada um desses elementos.
Além disso, $\bar{X}_n$ é um estimador não viesado e portanto

$$EQM_{Boot}(\bar{X}_n) = Var_{Fe}[\bar{X}_n].$$

Mas

$$Var_{Fe}[\bar{X}_n] = \frac{1}{n^2} Var_{Fe}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var_{Fe}[X_i] =$$

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

Exercício 2: Considere uma amostra $x = (x_1, x_2, \ldots, x_n)$.
(a) Obtenha a probabilidade de que o j-ésimo elemento da amostra não esteja na amostra de *Boot*.

(b) Para $n = 5, 100, 10000$ obtenha as probabilidades do j-ésimo elemento não estar na amostra de *Boot*.

(c) Faça um gráfico ilustrando o comportamento dessa probabilidade como função de n. O que pode ser observado?

(d) Considere que a sua amostra é composta pelos 100 primeiros números naturais ($n = 100$). Implemente o algoritmo de *bootstrap* e obtenha a proporção de amostras que contêm o número 4. Considere $B = 10000$ réplicas.

Solução:

(a) Primeiro note que a probabilidade do j-ésimo elemento ser escolhido para estar na primeira (ou em qualquer outra posição) da amostra de *Boot* é $\frac{1}{n}$. Portanto, dele não ser escolhido é $1 - \frac{1}{n}$. Como as seleções são realizadas de forma independente e com mesma distribuição, resulta que a probabilidade solicitada é $\left(1 - \frac{1}{n}\right)^n$.

(b)

n	5	100	10000
prob	0.3277	0.3660	0.3679

Gráfico da probabilidade de um elemento *j* não pertencer a amostra de *Boot* - n=1 a 10

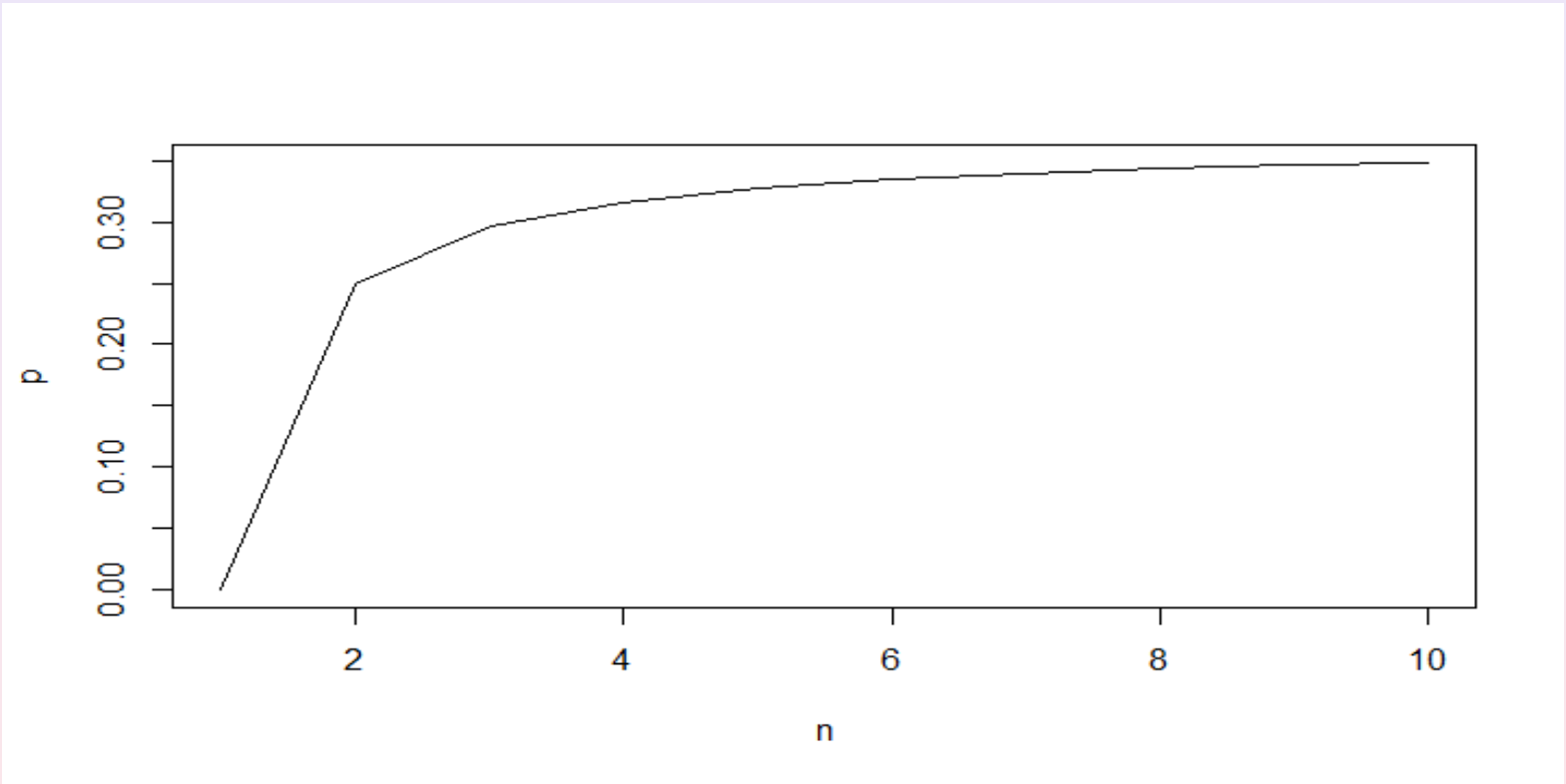
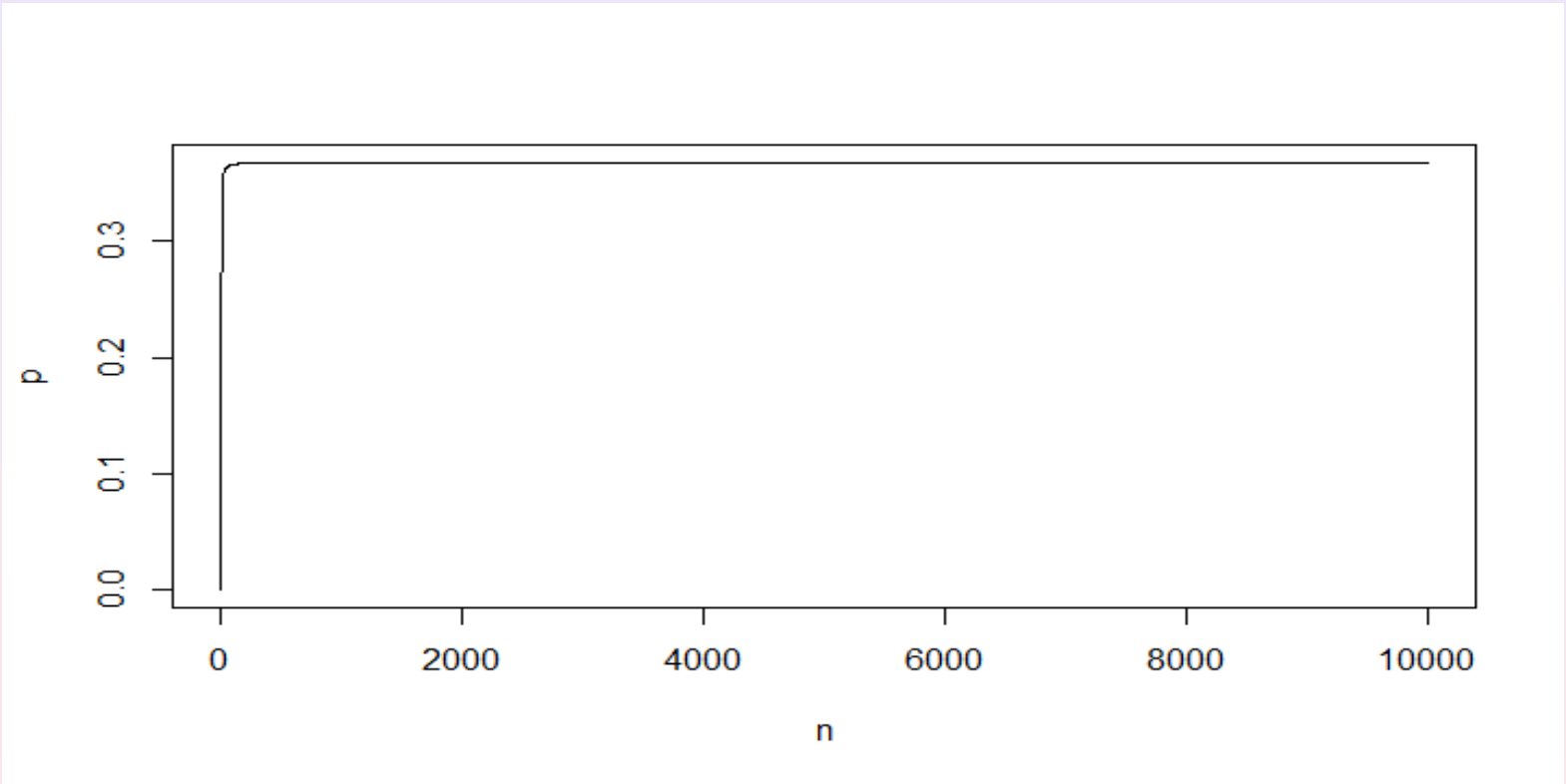


Gráfico da probabilidade de um elemento *j* não pertencer a amostra de *Boot* - n=1 a 10000



Nota-se que o valor dessa probabilidade converge para uma constante quando n cresce. Qual será essa constante?

O limite de $n \to \infty$ de $\left(1 - \frac{1}{n}\right)^n$ que é igual a $e^{-1} = 0.367879$.

Portanto, a probabilidade de um valor qualquer j pertencer a uma amostra de *Boot* é aproximadamente $1 - 0.368 = 0.632$ para n grande.

Método de *bootstrap*

(d) Programa no **R** para implementar o *bootstrap*.

```
I = rep(NA,10000)
for (i in 1:10000) {
I[i] = sum(sample(1:100, rep=TRUE)==4) > 0 }
mean(I)
```

Saída: 0.6285

Exercício 3: Para $n = 2$ e $(x_1, x_2) = (1, 3)$, determine o estimador de *bootstrap* de $Var[S_x^2]$.

Solução: Lembre que

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

Primeiro vamos obter a variância da amostra, pois esse será nosso valor de referência, i.e, aquele que irá substituir o parâmetro.

$$\theta(F_e) = \frac{1}{2}[(1-2)^2 + (3-2)^2] = 1$$

Vamos agora obter as amostras de *boot* e a variância amostral em cada uma delas.

Método de *bootstrap*

Amostra	s^2
$(1,1)$	0
$(3,3)$	0
$(1,3)$	2
$(3,1)$	2

Estimativa do erro quadrático médio

$$E\hat{Q}M = \frac{1}{4}[2(0-1)^2 + 2(2-1)^2] = 1$$