

# Bioinformática

## Métodos de computação intensiva– o *Bootstrap*

Manuela Neves

ISA/ULisboa

26 e 27 de Abril 2021

1 Métodos de reamostragem

2 A Abordagem Tradicional em Estatística

3 A metodologia *Bootstrap*

- O *Bootstrap não paramétrico* – simulação de Monte Carlo
- A metodologia *Bootstrap* no 

# Métodos de reamostragem

Os métodos de reamostragem, baseiam-se no cálculo repetido, um grande número de vezes, de estimativas da característica de interesse, em amostras geradas a partir da original.

Estes métodos, são em geral não paramétricos, e têm como objectivo estimar a distribuição amostral do estimador de interesse ou de algumas das suas características, quando não temos conhecimento do modelo estatístico subjacente.

A reamostragem é portanto, simplesmente, um processo de estimação de probabilidades realizando um número elevado de experiências numéricas.

# A abordagem Tradicional em Estatística

Na Abordagem Tradicional em Estatística o procedimento habitual consiste em dispormos de:

$$X_1, X_2, \dots, X_n \sim F \text{ (desconhecida)}$$

As análises são baseadas em **estatísticas** (funções da amostra aleatória). Conhecendo a distribuição de amostragem da **estatística**, pelo menos aproximadamente, podemos fazer Inferência.

Mas . . . a distribuição de amostragem de uma estatística depende, geralmente, da distribuição da população subjacente à amostra, que é desconhecida.

Seja  $T_n := T_n(X_1, X_2, \dots, X_n)$  uma estatística.

e a característica de interesse, por exemplo, **variância desta estatística**,  $Var[T_n]$

# A Abordagem Tradicional em Estatística

Se, por exemplo,  $T_n$  é uma média,  $T_n = \bar{X}_n$ , (ou uma função simples da média) e para facilitar suponhamos  $(X_1, X_2, \dots, X_n)$  i.i.d.

$$T_n = \bar{X}_n \Rightarrow \text{Var}[T_n] = \frac{\sigma^2}{n} \text{ (função de uma quantidade desconhecida)}$$

$$\text{Var}[\bar{X}_n] \longrightarrow \text{estimada por } \frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Mas para estimadores de outros parâmetros, a expressão regra geral é muito complicada e o que se costuma fazer é simplificar o problema considerando aproximações ou desenvolvimentos de  $\text{Var}[T_n]$ .

**Mas:** { há necessidade de valores muito elevados de  $n$   
dependência da fórmula teórica do modelo postulado  $\rightarrow$  validade(??)  
dificuldade na obtenção da fórmula teórica.  
...

# Métodos de reamostragem

As ideias iniciais de reamostragem e a orientação para uma linha de trabalho de análise estatística baseada em simulações.



Anos 40 /50 → o computador começou a ser usado para efectuar simulações porque permitiu:

- substituir complicadas e por vezes “grosseiras” aproximações por simulações
- despertar fortemente a atenção, quer de investigadores teóricos, quer de utilizadores de métodos estatísticos.

Métodos baseados em procedimentos repetidos sobre muitos conjuntos de réplicas dos próprios dados são os chamados **métodos de reamostragem**

***O jackknife e o bootstrap* são os métodos de reamostragem mais populares usados na análise estatística**

# Métodos de reamostragem

## Trabalho pioneiro → Quenouille (1949)



Metodologia desenvolvida para **estimar** e portanto controlar o **viés de estimadores** e para construir **intervalos de confiança robustos**, baseado na divisão da amostra em duas partes. Mais tarde, em 1956, generalizou esta ideia dividindo a amostra em várias subamostras. A forma mais frequentemente usada consiste na **divisão da amostra em  $n$  grupos de dimensão unitária**.

O termo **jackknife** foi introduzido por Tukey (1958), considerando que esta metodologia permitia testar hipóteses e calcular intervalos de confiança em situações em que não há melhores métodos que possam ser utilizados.



# Métodos de reamostragem

**Finais dos anos 70, Efron** unificou as ideias existentes e introduziu a metodologia *bootstrap não paramétrico simples*.

A origem do termo **bootstrap** deriva da obra de Rudolph Raspe, autor do século XVIII, a quem se deve as *Aventuras do Barão Munchausen*.

Numa das suas obras ele diz:

*“ The baron had fallen to the bottom of a deep lake. Just when he looked like all was lost, he thought to pick himself up by his own bootstraps. ”*



# A metodologia *Bootstrap*

Em “*Bootstrap methods. Another look at the jackknife*”, Efron (1979)  
→ técnica não paramétrica para a estimação do desvio padrão  
→ recorrendo a métodos de computação intensiva.

Genericamente → estimar viés, variância, quantis, distribuição de amostragem do estimador (ou melhorar estimadores existentes.)

## Ideia básica da metodologia ***bootstrap não paramétrico***

$\underline{X}_n = (X_1, X_2, \dots, X_n) \stackrel{i.i.d.}{\sim} F$        $T_n$  um estimador de  $\theta$

A ideia é construir a ***amostra bootstrap***

$$\underline{X}_n^* = (X_1^*, X_2^*, \dots, X_n^*)$$

são variáveis reamostradas de  $\underline{x}_n$  de acordo com a função de distribuição empírica,

$$\hat{F}_n(x) = \frac{\#\{i : x_i \leq x, 1 \leq i \leq n\}}{n}.$$

# A metodologia *Bootstrap*

Considera-se  $\underline{x}_n = (x_1, x_2, \dots, x_n) \sim \hat{F}_n$

$$\underline{X}_n^* = (X_1^*, X_2^*, \dots, X_n^*).$$

Nota: esta extracção é feita com reposição enquanto o *jackknife* faz extracções de amostras de dimensão  $n - 1$ , sem reposição.

Dada  $\underline{x}_n$  a distribuição da amostra *bootstrap* associada,  $\underline{X}_n^*$ , é

$$P(X_i^* = X_j | \underline{x}_n) = \frac{1}{n} \quad i, j = 1, \dots, n$$

$T_n^* := T_n(\underline{X}_n^*)$  versão *bootstrap* do estimador  $T_n$

# A metodologia *Bootstrap*

O comportamento da *versão bootstrap* deverá simular o comportamento de  $T_n$ .

→ A distribuição de  $T_n^*$ , é usada para aproximar a distribuição de amostragem (desconhecida), de  $T_n$ .

A forma mais usual de usar o *bootstrap* é o chamado *bootstrap não paramétrico*, com recurso à simulação de Monte Carlo

# A metodologia *Bootstrap* não paramétrico – simulação de Monte Carlo

- dada uma amostra observada  $\underline{x}_n = (x_1, x_2, \dots, x_n)$ , constrói-se  $\widehat{F}_n$  atribuindo a cada  $x_i$  peso  $1/n$ ;
- gera-se uma amostra *bootstrap*  $\underline{x}_n^* = (x_1^*, x_2^*, \dots, x_n^*)$ , de variáveis  $X_i^* \stackrel{i.i.d.}{\sim} \widehat{F}_n$  e calcula-se  $t_n^* = t_n(x_1^*, x_2^*, \dots, x_n^*)$ ;
- repete-se, independentemente,  $B$  vezes o passo anterior, obtendo assim  $B$  réplicas  $(t_n^{*,1}, t_n^{*,2}, \dots, t_n^{*,B})$
- calcula-se as estimativas do viés, erro padrão e distribuição de amostragem *bootstrap*

$$\widehat{Vies}_B^*[T_n] = \sum_{i=1}^B t_n^{*,i} / B - \theta(\widehat{F}_n) \quad \widehat{\sigma}_B^*[T_n] = \sqrt{\frac{1}{B-1} \sum_{i=1}^B (t_n^{*,i} - \overline{t_n^{*,i}})^2}$$

$$\widehat{F}_B^*(t) = \frac{\#\{i : t_n^{*,i} \leq t, 1 \leq i \leq B\}}{B}, \quad -\infty < t < +\infty$$

# A metodologia Bootstrap

**Nota:**  $\lim_{B \rightarrow \infty} \widehat{Vies}_B^* = Vies^*$      $\lim_{B \rightarrow \infty} \widehat{\sigma}_B^* = \sigma^*$      $\lim_{B \rightarrow \infty} \widehat{F}_B^* = F^*$

**Desvio padrão** — estimativa *bootstrap* é regra geral bastante boa, com um número pequeno de réplicas, Efron (1993) considera suficiente  $B = 200$ .

**Viés** — convergência é mais difícil de atingir.

**Intervalos de confiança *bootstrap*** — há **vários procedimentos**:

- o métodos dos percentis —  $(t_{\alpha/2}^*, t_{1-\alpha/2}^*)$ , que são os percentis empíricos dos valores bootstrap  $t_n^{*,i}$  — **é só este que vamos usar**
- intervalo *t-bootstrap* —  $(t_n - t_{\alpha/2}^* \sigma^*[T_n]; t_n + t_{\alpha/2}^* \sigma^*[T_n])$ , com  $t_{\alpha/2}^*$  quantis *bootstrap* da variável  $T_n^*$  estandardizada, i.e., de  $\frac{T_n^* - t_n}{\sigma^*[T_n]}$ , ou seja consideram-se os percentis da amostra de valores  $\frac{t_n^{*,i} - t_n}{s^{*,i}[T_n]}$
- há ainda outras formas de construir intervalos ...

# A metodologia *Bootstrap* no R

## Exemplo

De uma população exponencial de parâmetro  $\lambda = 1/10$  foi gerada a seguinte amostra  $\rightarrow (16,43,13,7,12,14,6,25,0,54)$ .

Pretende-se determinar um IC para  $\sigma^2$ .

## Vamos analisar os seguintes comandos

```
rm(list=ls())
library(bootstrap)
par(mfrow=c(2,2))
x<-c(16,43,13,7,12,14,6,25,0,54)
hist(x);mean(x);var(x);sd(x)
shapiro.test(x)
n<-length(x)
##I.C. para sigma2, com base na amostra dada
IC1<-c((n-1)*var(x)/qchisq(0.975,n-1),
      (n-1)*var(x)/qchisq(0.025,n-1))
IC1
```

# Exemplo—continuação

```
#####bootstrap para os valores x
theta<-function(x)var(x)
result1<-bootstrap(x,1000,theta,func=NULL)
hist(result1$thetastar)
quantile(result1$thetastar,prob=c(0.025,0.975))

## E se agora tentássemos gerar valores nossos?
lambda<-1/10 #sabe-se que a amostra foi obtida de uma
#população exponencial com lamda=1/10
x2<-c(rexp(n,lambda));x2
hist(x2)
mean(x2);var(x2);sd(x2);shapiro.test(x2)

##I.C. para sigma2, com base na amostra que agora retirámos
IC2<-c((n-1)*var(x2)/qchisq(0.975,n-1),
      (n-1)*var(x2)/qchisq(0.025,n-1))
##### experimentar o bootstrap para os valores x2
```

# A metodologia *Bootstrap* no R

## Exemplo de uso do *bootstrap* no estudo da mediana

Considere-se o seguinte conjunto de dados:

```
x<-c(8.26, 6.33, 10.4, 5.27, 5.35, 5.61, 6.12, 6.19, 5.2,  
7.01, 8.74, 7.78, 7.02, 6, 6.5, 5.8, 5.12, 7.41, 6.52, 6.21,  
12.28, 5.6, 5.38, 6.6, 8.74)
```

```
median(x)  
#queremos estudar a mediana. Vamos gerar 1000 amostras  
boot <-numeric(1000)  
  
for (i in 1:1000) boot[i] <- median(sample(x,replace=T))  
boot  
mean(boot); var(boot)  
hist(boot) # histograma dos valores  
bias_est<-mean(boot)-median(x);bias_est  
quantile(boot,prob=c(0.025,0.975))
```