Bioinformática

Métodos de computação intensiva– o *Bootstrap*

Manuela Neves

ISA/ULisboa

6 de Maio 2020

Plano das aulas

1. Métodos de reamostragem
2. A Abordagem Tradicional em Estatística
3. A Abordagem Tradicional em Estatística
4. 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 do parâmetro de interesse, em amostras geradas a partir da amostra 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 propriedades, 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 uma amostra aleatória:

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

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

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

$$\text{Seja} \quad T_n := T_n(X_1, X_2, ..., X_n) \quad \text{uma estatística.}$$

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

A Abordagem Tradicional em Estatística

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

$$T_n = \overline{X}_n \Rightarrow Var[T_n] = \frac{\sigma^2}{n} \text{ (função de } \sigma^2\text{, desconhecido)}$$

Então, como sabemos, uma estimativa de σ^2 é $\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2$, portanto

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

A Abordagem Tradicional em Estatística

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

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.

$\downarrow$

Anos 40 /50 $\longrightarrow$ 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 de reamostragem

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

Nestas aulas vamos apenas abordar a metodologia *bootstrap*, mas só uduas pavrasa sobre a história do *jackknife*

Métodos de reamostragem

Trabalho pioneiro $\rightarrow$ Quenouille (1949)

$\downarrow$

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, Quenouille generalizou esta ideia dividindo a amostra em várias subamostras. A forma mais frequentemente usada consiste na divisão da amostra em n sub-amostras de dimensão $n-1$, obtidas por eliminação consecutiva de uma observação.

Métodos de reamostragem

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)

$\longrightarrow$ técnica não paramétrica para a estimação do desvio padrão

$\longrightarrow$ recorrendo a métodos de computação intensiva.

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

Vejamos a ideia básica da metodologia *bootstrap não paramétrico*

A metodologia *Bootstrap*

Dada a amostra aleatória $\underline{X}_n = (X_1, X_2, \cdots X_n) \overset{i.i.d.}{\sim} F$, onde F tem um parâmetro desconhecido θ

Seja T_n um estimador de θ

A ideia é construir a *amostra bootstrap*, que vamos representar como

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

$$\uparrow$$

são variáveis reamostradas da amostra pbservada $\underline{x}_n = (x_1, \ldots, x_n)$.
Como?

A metodologia *Bootstrap*

Considera-se $\underline{x}_n = (x_1, x_2, \cdots x_n) \sim \widehat{F}_n$, valores observados de

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

A função de distribuição empírica,

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

atribui massa $1/n$ a cada um dos pontos da amostra.

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

A metodologia *Bootstrap*

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

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

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

A metodologia *Bootstrap*

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

$\longrightarrow$ 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.

Vejamos o algoritmo

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

- dada uma amostra observada $\underline{x}_n = (x_1, x_2, \cdots x_n)$, tem-se $\widehat{F}_n$ que atribui a cada x_i peso $1/n$;
- gera-se uma amostra *bootstrap* $\underline{x}^*_n = (x^*_1, x^*_2, \cdots x^*_n)$, de variáveis $X^*_i \overset{i.i.d.}{\sim} \widehat{F}_n$ e calcula-se $t^*_n = t_n(x^*_1, x^*_2, \cdots, x^*_n)$;
- repete-se, independentemente, B vezes o passo anterior, obtendo assim B réplicas $(t^{*,1}_n, t^{*,2}_n, \cdots t^{*,B}_n)$
- calcula-se as estimativas do viés, erro padrão e distribuição de amostragem *bootstrap*

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

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

A metodologia Bootstrap

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

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 — $\left(t^*_{\alpha/2}, t^*_{1-\alpha/2}\right)$, que são os percentis empíricos dos valores bootstrap $t^{*,i}_n$ – é só este que vamos usar
- intervalo *t-bootstrap* — $\left(t_n - t^*_{\alpha/2}\sigma^*[T_n]; t_n + t^*_{\alpha/2}\sigma^*[T_n]\right)$, 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 $\frac{t^{*,i}_n - t_n}{s^{*,i}[T_n]}$
- 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 –> (16,43,13,7,12,14,6,25,0,54).
Pretende-se determinar um IC para σ^2.

Vamos analisar os seguintes comandos

```
rm(list=ls())
x<-c(16,43,13,7,12,14,6,25,0,54)
hist(x);mean(x);var(x);sd(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

[1] 138.2554 973.9336
```

Exemplo–aplicação do *Bootstrap*

Vamo executar nós o *Bootstrap*

```
> n<-length(x);n
[1] 10
> var.star<-vector()
> for(i in 1:1000)
+
+ amostraboot<-sample(x,n,rep=TRUE)
+ var.star[i]<-var(amostraboot)
+
> hist(var.star)
> mean(var.star)
[1] 265.5348
> quantile(var.star,prob=c(0.025,0.975))
     2.5%     97.5%
 32.28111 490.62222
>
```

Exemplo–aplicação do *Bootstrap*

Vamos executar o *Bootstrap* no 

```
> rm(list=ls())
> x<-c(16,43,13,7,12,14,6,25,0,54)
> library(bootstrap)
> n<-length(x)
>
> theta<-function(x)var(x)
> result<-bootstrap(x,1000,theta,func=NULL)
> hist(result$thetastar)
> quantile(result$thetastar,prob=c(0.025,0.975))
     2.5%      97.5%
 31.82333 477.13972
```

A metodologia *Bootstrap* no R

Exemplo de uso do *bootstrap* no estudo da mediana

Considere-se o seguinte conjunto de dados:

```
x1<-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(x1) ## mediana da amostra -- estima a verdadeira mediana
[1] 6.33
> boot <-vector()
> for (i in 1:1000) boot[i] <- median(sample(x1,replace=T))
> mean(boot) ## esta é a estimativa boostrap da  mediana
[1] 6.36845
> bias_est<-mean(boot)-median(x1);bias_est
[1] 0.03845
> quantile(boot,prob=c(0.025,0.975))
 2.5% 97.5%
 5.80  7.02
```