Estatística Computacional

Patrícia de Siqueira Ramos

PPGEAB
UNIFAL-MG

4 de Julho de 2018

Introdução

- Métodos de reamostragem são aplicados quando há dados mas não há um modelo adequado para descrevê-los
- Nessas situações, não há como aplicar simulação Monte Carlo para gerar amostras (pois é necessário um modelo estatístico para gerá-las)
- Ideia: reamostrar os valores da amostra disponível de forma aleatória
- Esse procedimento de reutilizar os dados para amostrar é chamado reamostragem
- Métodos de reamostragem são menos precisos do que os métodos Monte Carlo, mas apresentam a vantagem de serem simples e possibilitarem várias aplicações

Introdução

- Métodos de reamostragem
  - *bootstrap*
  - *jackniffe*
- A partir de uma só amostra geram-se novas amostras à imagem da original
- Os métodos de reamostragem tratam a amostra observada como uma população finita, então são geradas amostras partir da original para estimar características populacionais e fazer inferência sobre a população.

Bootstrap

- Efron (1979)
- extensão do *jackknife*
- $\pm$1990: ferramenta mais geral
- Google acadêmico "*bootstrap method*": 672.000 resultados (06/2018)

Bootstrap

- Método de computação intensiva que permite a simulação da distribuição de alguma estatística
- Ideia: reamostrar os dados observados repetidamente e, para cada amostra gerada, um novo valor da estatística é calculado
- A coleção dos valores obtidos provê uma estimativa da distribuição amostral da estatítica de interesse
- É um método não paramétrico por natureza

Bootstrap

- Estudo amostral: inferências (extrapolação)
- Algumas estatísticas: média, mediana, desvio padrão, correlação etc.
- Uma estatística $\widehat{\theta}$ (estimador) varia de amostra para amostra e há interesse em saber a magnitude dessas flutuações em torno de θ (parâmetro) que se deseja estimar
- Distribuição amostral: conjunto de todos os valores possíveis da estatística amostral apresentado na forma de uma distribuição de probabilidade

Bootstrap

- Os métodos *bootstrap* são uma classe de métodos de Monte Carlo não paramétricos que estimam a distribuição da população por reamostragem
- O termo "*bootstrap*" pode ser usado para o *bootstrap* não paramétrico (mais usado) ou paramétrico
- A distribuição da população finita representada pela amostra pode ser encarada como uma pseudo população, com características análogas às da verdadeira população
- Por meio da geração repetida de amostras aleatórias desta pseudo população (reamostragem), a distribuição de amostragem de uma estatística pode ser estimada
- O *bootstrap* gera amostras aleatoriamente a partir da distribuição empírica da amostra

Bootstrap

- *Bootstrap*: método de reamostragem para inferência estatística usado, geralmente, para estimar IC, viés e variância de um estimador $\widehat{\theta}$

- Dois atributos interessantes:
  - provê abordagem automática para inferência (usando computador)

  - provê uma maneira de lidar com a inferência em situações em que abordagens padrão que precisam de pressuposições fortes são inadequadas

Ideia do *bootstrap*

- Reamostragem: sortear com reposição dados pertencentes a uma amostra retirada anteriormente de modo a formar novas amostras
- Ideia básica: na ausência de qualquer conhecimento sobre a população, é realizar reamostragem com reposição de tamanho n da amostra original milhares de vezes
- A distribuição *bootstrap* de algum estimador de interesse é utilizada no lugar da "distribuição teórica" deste mesmo estimador (dificuldade de desenvolvê-la ou do desconhecimento da distribuição da população de onde foi obtida a amostra aleatória)

Ideia do *bootstrap*

Dado um conjunto de observações X_i, $i = 1, \ldots, n$,

- Supomos que os dados foram obtidos da a.a. de alguma distribuição F
- A reamostragem de $x_1, \ldots, x_n$ gera um conjunto de valores que são a amostra *bootstrap*
- Estimamos θ com $\widehat{\theta}$ em cada amostra *bootstrap*
- Os valores obtidos de $\{\widehat{\theta}_i^*\}$ para B reamostragens são a distribuição empírica amostral de $\widehat{\theta}$
- A distribuição empírica obtida é então usada para estimar viés, desvio padrão ou construir IC da estatística de interesse $\widehat{\theta}$

Passos do *bootstrap*

Para estimar θ por meio do estimador $\widehat{\theta}$, gerando amostras *bootstrap* a partir de $x_1, \ldots, x_n$, fazer:

- Para cada valor de b, sendo $b = 1, \ldots, B$:
  a) gerar amostra *bootstrap* $x^* = x_1^*, x_2^* \ldots, x_n^*$ por meio da amostragem com reposição da amostra observada $x_1, \ldots, x_n$
  b) calcular a b-ésima estimativa $\widehat{\theta}_b^*$ na amostra *bootstrap* x^*
- A estimativa *bootstrap* de $F_{\widehat{\theta}}(\cdot)$ é a função distribuição empírica das estimativas $\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$ dada por

$$F_n^*(x) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\{\widehat{\theta}_b^* \leq x\}}$$

Exemplo 1 - *bootstrap*

```
# exemplo de amostra
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)

# estatística escolhida: CV (theta) - função para calculá-la
CV = function(x) sqrt(var(x))/mean(x)*100

# estimativa (valor de theta chapéu)
CV(x)

# gerar apenas uma amostra bootstrap
bb = sample(x,replace=T)

# obter o CV usando a amostra bootstrap
CV(bb)
```

Exemplo 1 - *bootstrap*

```
# gerar B = 1000 amostras bootstrap
B = 1000
(boo = numeric(B))

# gerar B amostras e calcular o CV para cada amostra
for (b in 1:B) boo[b] = CV(sample(x,replace=T))
# boot é theta chapéu*

# obter média e variância da coleção de amostras
mean(boo); var(boo)

# histograma
hist(boo, prob = T)

# valor dos quantis 0,975 e 0,025
quantile(boo, 0.025); quantile(boo, 0.975)
```

Observações

- A amostra original representa a população de onde ela foi retirada

  - Reamostras dessa amostra representam o que obteríamos ao retirar muitas amostras da população

- A distribuição *bootstrap* é centrada próxima da média da amostra original

  - Assim, a média da distribuição *bootstrap* tem pequeno viés

  (Sabemos que a distribuição amostral de $\bar{X}$ é centrada na média populacional μ, ou seja, a distribuição *bootstrap* se comporta como esperamos de uma distribuição amostral)

Viés de $\widehat{\theta}$

O viés de um estimador $\widehat{\theta}$ de θ é definido como

$$\text{viés}(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

Estimativa bootstrap do viés:

Viés de $\widehat{\theta}$

O viés de um estimador $\widehat{\theta}$ de θ é definido como

$$\text{viés}(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

Estimativa bootstrap do viés:

$$\widehat{viés}^*(\widehat{\theta}) = \overline{\widehat{\theta}^*} - \widehat{\theta},$$

em que

$$\overline{\widehat{\theta}^*} = \frac{\sum_{b=1}^{B} \widehat{\theta}_b^*}{B}$$

e

$$\widehat{\theta} = \widehat{\theta}\mathbf{x} = \widehat{\theta}(x_1, x_2, \ldots, x_n)$$

Exemplo 1 (cont.)

```
# estimativa do viés = diferença entre a média
# dos valores e a estimativa inicial do estimador
(viesb = mean(boo) - CV(x))
```

Exemplo 2

- O conjunto de dados `law` de Direito do pacote *bootstrap* é de Efron e Tibshirani
- O *data frame* contém dados referentes ao LSAT (Law School Average Test) e GPA (Grade-Point Average) para 15 faculdades de Direito.
- Esses dados são uma amostra aleatória do universo de 82 faculdades de Direito em `law82`

Exemplo 2 (cont.)

```
library(bootstrap)
data(law)
law

# estatística escolhida: coeficiente de correlação amostral
# estimativa (theta chapéu) para a amostra law de dimensão n = 15
(corr = cor(law$LSAT, law$GPA))

# gerar B amostras bootstrap
B = 2000; n = nrow(law); x = law
(boo = numeric(B))
# gerar B amostras e calcular a correlação para cada amostra
for (b in 1:B){
  i = sample(1:n, size = n, replace = TRUE) # i é o vetor dos índices
  LSAT = x$LSAT[i]
  GPA = x$GPA[i]
  boo[b] = cor(LSAT,GPA)
}
boo
```

Exemplo 2 (cont.)

```
# obter média e variância da coleção de amostras
mean(boo)
var(boo)

# histograma
hist(boo, prob = T)
```

Desvio padrão de $\widehat{\theta}$

A estimação *bootstrap* do desvio padrão de um estimador $\widehat{\theta}$ é o desvio padrão empírico da reamostragens *bootstrap* $\{\widehat{\theta}_i^*\} = \widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$

Estimativa bootstrap do desvio padrão de $\widehat{\theta}$*:*

$$\widehat{S}^*(\widehat{\theta}) = \sqrt{\frac{\sum_{b=1}^{B} \left(\widehat{\theta}_b^* - \overline{\widehat{\theta}^*}\right)^2}{B-1}},$$

em que

$$\overline{\widehat{\theta}^*} = \frac{\sum_{b=1}^{B} \widehat{\theta}_b^*}{B}$$

Exemplos 2 (cont.) - viés e desvio padrão *bootstrap*

```
# estimativa do viés:
# média das amostras bootstrap - estimativa inicial
(viesb = mean(boo) - corr)

# desvio padrão bootstrap
(SB = sd(boo))
# ou
(SB = sqrt(sum((boo - mean(boo)) ** 2)/(B-1)))
```

- Então, a estimativa *bootstrap* do desvio padrão do coeficiente de correlação amostral r é dada por $S(r) = \widehat{S}^*(r) = sd(boo)$
- Obter o desvio padrão *bootstrap* do CV, $\widehat{S}^*(\widehat{CV})$, para os dados do Exemplo 1

Alternativa: uso da função `boot` para o Exemplo 2

```
# usando a função boot
# escrever a função da correlação usando
# índice como argumento
r = function(x, i) cor(x[i,1], x[i,2])

library(boot)
# função boot
obj = boot(data = law, statistic = r, R = 2000)
obj
```

Jackknife

- Outra forma de reamostragem é o *jackknife*, proposta por Quenouille (1949, 1956) para estimar o viés e por Tukey (1958) para estimar o desvio padrão (algumas décadas antes do *bootstrap*)
- Para um estimador

$$\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n),$$

- o método vai deixando de fora uma observação em cada reamostra de tamanho $n - 1$

$$X^{(j)} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n), \qquad j = 1, \ldots, n.$$

- As novas amostras são denominadas amostras *jackknife*

Jackknife

- Com base nas amostras *jackknife* calculam-se as estimativas *jackknife*

$$\widehat{\theta}^{(j)} = \widehat{\theta}_{n-1}(X^{(j)}), \qquad j = 1, \ldots, n.$$

Estimativa jackknife do viés:

$$\widehat{viés}_{jack}(\widehat{\theta}) = (n-1)(\overline{\widehat{\theta}^{(\cdot)}} - \widehat{\theta}),$$

em que

$$\overline{\widehat{\theta}^{(\cdot)}} = \frac{\sum_{j=1}^{n} \widehat{\theta}^{(j)}}{n}$$

Exemplo 3

- Os dados *patch* do pacote `bootstrap` de Efron e Tibshirani contêm medidas de um certo hormônio na corrente sanguínea de oito sujeitos depois de usarem um medicamento
- O parâmetro de interesse é

$$\theta = \frac{E(\text{novo}) - E(\text{antigo})}{E(\text{antigo}) - E(\text{placebo})}.$$

- Assim, a estatística de interesse é

$$\widehat{\theta} = \frac{\bar{Y}}{\bar{Z}},$$

em que $Y = \text{novo} - \text{antigo}$ e $Z = \text{antigo} - \text{placebo}$.

Exemplo 3

```
data(patch, package = "bootstrap")
patch
n = nrow(patch)
y = patch$y
z = patch$z
# estimativa theta chapéu
theta_ch = mean(y)/mean(z)
theta_ch

# obter as reamostragens jackknife,
# deixando uma observação de fora
theta_jack = numeric(n)
    for (i in 1:n){
    theta_jack[i] = mean(y[-i]) / mean(z[-i])
}
theta_jack
mean(theta_jack) # estimativa jackknife

vies = (n-1) * (mean(theta_jack) - theta_ch)
vies # estimativa jackknife do viés
```

Exercícios

1 Utilizar a função `boot` para os dados do Exemplo 1 e avaliar se os resultados foram consistentes com os obtidos para as estimativas vistas. Para utilizar a função `boot`, altere a função `CV` para:

```
CV = function(x, ind){
    X = x[ind]
    return (sqrt(var(X))/mean(X)*100)
}
# uso
CV(x, 1:length(x))  # estimativa (theta chapéu)
```

Exercícios

2 Os dados a seguir referem-se ao tempo de vida, em anos, de $n = 40$ pessoas:

```
y = c(33.23, 11.81, 23.67, 19.28, 3.92, 4.01, 43.84, 0.19, 1.31, 2.20,
    1.48, 7.82, 3.04, 81.09, 11.18, 3.83, 7.46, 23.23, 22.77, 33.57,
    16.75, 2.69, 39.31, 15.72, 10.99, 33.99, 2.72, 26.31, 14.66, 63.40,
    52.51, 18.22, 11.54, 29.45, 6.27, 28.32, 12.14, 54.62, 10.24, 17.52)
```

Considerando que a distribuição do tempo de vida é a exponencial e que o parâmetro λ pode ser estimado pelo estimador de máxima verossimilhança $\hat{\lambda} = 1/\bar{X}$, em que $\bar{X} = \sum_i^n X_i/n$, faça o que se pede:

Exercícios

a) Obter o intervalo de 95% de confiança para a média da exponencial utilizando o seguinte procedimento:

i) gerar uma amostra da exponencial de tamanho $n = 40$, utilizando a função `rexp` e considerando o parâmetro igual à estimativa obtida;
ii) determinar a estimativa da média $\mu = 1/\lambda$ por $\bar{X}$ nesta amostra simulada de tamanho $n = 40$;
iii) repetir 1.000 vezes os passos (i) e (ii) e armazenar os valores obtidos;
iv) ordenar as estimativas e tomar os quantis $2,5\%$ e $97,5\%$. Os valores obtidos são o intervalo de confiança pedido, considerando como verdadeira a densidade exponencial para modelar o tempo de vida das pessoas. Este procedimento é denominado de *bootstrap* paramétrico.

Exercícios

b) Repetir esse processo, gerando 100.000 amostras de tamanho $n = 40$. Comparar os resultados e verificar se o custo adicional de ter aumentado o número de simulações compensou a possível maior precisão obtida.

Exercícios

3 Obter a estimativa e o viés *jackknife* do exemplo 1

4 Obter a estimativa e o viés *bootstrap* dos dados do exemplo 3

5 Para os dados do exemplo 1 usar a média amostral como estatística e obter o *IC bootstrap* a 95%. Comparar com o *IC* para a média assumindo normalidade (a função `t.test` faz isso).