Aplicações do Método Bootstrap

ESTAT0090 – Estatística Computacional
Prof. Dr. Sadraque E. F. Lucena
sadraquelucena@academico.ufs.br

Bootstrap

- Bootstrap é um método que resolve um problema clássico: muitas vezes temos um **estimador** (média, mediana, correlação, coeficiente de regressão) e queremos saber **quão confiável ele é**. Para isso, precisamos do **erro padrão** ou de um **intervalo de confiança**.
- Só que às vezes:
  - não existe fórmula teórica (mediana, quantis).
  - a fórmula é complicada ou exige suposições fortes (normalidade, independência).
  - temos amostras pequenas e não dá pra confiar muito em aproximações assintóticas.
- É aí que entra o bootstrap: a gente **simula** a distribuição do estimador **a partir da própria amostra**.
  - Como? Reamostrando com reposição.
- Nos Exercícios a seguir use sempre a semente `set.seed(123)`.

Atividade 18.1 – Erro padrão e viés da média

Gere amostras de tamanho $n = 10, 30, 50$ de:

1. $X \sim N(0, 1)$ (distribuição simétrica)
2. $X \sim Gama(shape = 2, scale = 5)$ (distribuição assimétrica)

Para cada amostra:

a. estime, por bootstrap ($R = 10.000$), o erro padrão e o viés da média amostral;

b. compare o erro padrão bootstrap com o erro padrão teórico ($\sqrt{1/n}$ e $\sqrt{50/n}$);

c. O que acontece com o erro padrão quando o tamanho da amostra aumenta?

d. O erro padrão se comporta diferente para a Normal e para a Gama?

Atividade 18.1 – Erro padrão e viés da média

```
# Fazendo para a Normal usando a função replicate()

# função que gera dados da normal e calcula a média
media.norm <- function(n, media=0, desvpad=1){
  mean(rnorm(n, mean=media, sd=desvpad))
}

set.seed(123) # semente

n <- 10 # tamanho da amostra

# Normal
media.norm.b <- replicate(10000, media.norm(n)) # médias boostrap
mean(media.norm.b) - 0        # viés
```

```
[1] 0.0009767488
```

```
sd(media.norm.b); 1/sqrt(n)  # erro padrão
```

```
[1] 0.3139096
```

```
[1] 0.3162278
```

Atividade 18.2 – Mediana e 3º Quartil em dados reais

Faça

```
library(boot)
x <- faithful$waiting
```

A variável `x` contém o tempo (em minutos) entre eupções do gêiser Old Faithful no Parque Nacional Yellowstone, Wyoming, EUA.

Calcule a mediana e o terceiro quartil do tempo entre as erupções e apresente erro padrão e o intervalo de confiança bootstrap BCa (que constuma ser o mais confiável, pois corrige viés e assimetria).

- Dica: use a função `boot.ci()` para obter o intervalo bootstrap BCa.

Atividade 18.2 – Mediana e 3º Quartil em dados reais

```
library(boot)
x <- faithful$waiting

mediana <- function(x, i) median(x[i]) # função parausar em boot

set.seed(123)
mediana.b <- boot(x, mediana, R = 10000)
```

Atividade 18.2 – Mediana e 3º Quartil em dados reais

```r
# erro padrão
sd(mediana.b$t)
```

```
[1] 1.001243
```

```r
# intervalo de confiança
boot.ci(mediana.b, type = c("perc","bca"), conf = .95)
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = mediana.b, conf = 0.95, type = c("perc", "bca"))

Intervals :
Level     Percentile            BCa
95%   (73.5, 77.0 )   (73.0, 77.0 )
Calculations and Intervals on Original Scale
```

Atividade 18.3 – Teste de hipóteses para diferença de médias

Verifique se carros manuais fazem mais milhas por galão do que carros automáticos usando o data frame `mtcars`. O consumo dos veículos foi registrado na variável `mpg`.

```
library(dplyr)
dados <- mtcars |>
  mutate(am = factor(am, labels = c("auto","manual"))) |>
  select(mpg, am)

difmedia <- function(bd){
  n.auto <- length(bd$mpg[bd$am == "auto"])
  n.manual <- length(bd$mpg[bd$am == "manual"])
  amostra.auto <- sample(bd$mpg[bd$am == "auto"],
                         n.auto,
                         replace = T)
  amostra.manual <- sample(bd$mpg[bd$am == "manual"],
                           n.manual,
                           replace = T)
  return( mean(amostra.manual) - mean(amostra.auto))
}
```

Atividade 18.3 – Teste de hipóteses para diferença de médias

```
( dif <- mean(dados$mpg[dados$am == "manual"]) -
    mean(dados$mpg[dados$am == "auto"]) )
```

```
[1] 7.244939
```

```
set.seed(123)
difmedia.b <- replicate(10000, difmedia(dados))

( pvalor <- (sum(dif >= difmedia.b)+1)/(10000+1) )
```

```
[1] 0.5007499
```

Atividade 18.4 – Regressão

Ajuste om modelo de regressão linear para `mpg` vs. `wt`. Obtenha o erro padrão do parâmetro associado a `wt` e obtenha um intervalo de confiança de 95% para ele.

```
fit <- lm(mpg ~ wt, data = mtcars)

regboot <- function(dados){
  n <- nrow(dados)
  id_amostrado <- sample(1:n, n, replace = TRUE)
  coef(lm(mpg ~ wt, data = dados[id_amostrado, ]))[2]
}

beta1.b <- replicate(10000, regboot(mtcars))
```

Atividade 18.4 – Regressão

```
# beta1
fit$coefficients[2]
```

```
       wt
-5.344472
```

```
# erro pdarão
sd(beta1.b)
```

```
[1] 0.7055303
```

```
# intervalo de confiança de 95%
quantile(beta1.b, c(.025,.975))
```

```
     2.5%     97.5%
-6.947418 -4.145405
```

Fim