Bài giảng R, số 1
-Phương pháp Bootstrap-

TS.Tô Đức Khánh

04/05/2024

Trong bài học này, ta tìm hiểu cách sử dụng phương pháp bootstrap để khảo sát phân phối mẫu của các ước lượng của dữ liệu.

1 Một số thao tác cơ bản

Cho bộ dữ liệu ngẫu nhiên $Y_1, Y_2, \ldots, Y_n$ cuả biến ngẫu nhiên Y. Quy trình bootstrap cho trung bình mẫu được trình bày như sau:

1. Tạo một mẫu ngẫu nhiên $Y_1^*, Y_2^*, \ldots, Y_n^*$ (cùng với cỡ của dữ liệu gốc) từ dữ liệu gốc, có lặp lại (with replacement).
2. Tính trung bình $\overline{Y}^*$ của mẫu vừa tạo.
3. Lặp lại bước 1 và 2 trong R lần (ít nhất 1000 lần), và lưu kết quả lại.

Trong `R`, để tạo mẫu ngẫu nhiên, ta có thể sử dụng hàm

```
sample(x, size, replace = FALSE)
```

trong đó,

- `x` ta tập hợp gốc mà ta muốn lấy dữ liệu từ đó;
- `size` là cỡ mẫu ngẫu nhiên cần lấy;
- `replace` là lựa chọn lấy mẫu có lặp lại hoặc không, nếu `FALSE` thì có nghĩa là không lặp lại, trong khi `TRUE` tương ứng là lặp lại.

Để thực hiện lặp, ta có thể sử dụng `for()` hoặc `sapply()`.

Ví dụ 1: Xét dữ liệu `birthwt.txt`. Ta muốn khảo sát phân phối mẫu của ước lượng trung bình của cân nặng của trẻ sơ sinh `bwt`.

```
data_birth <- read_table(file = "datasets/birthwt.txt")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   low = col_double(),
##   age = col_double(),
##   lwt = col_double(),
##   race = col_double(),
##   smoke = col_double(),
##   ptl = col_double(),
##   ht = col_double(),
##   ui = col_double(),
```

```
##   ftv = col_double(),
##   bwt = col_double()
## )
```

```
glimpse(data_birth)
```

```
## Rows: 189
## Columns: 10
## $ low   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ age   <dbl> 19, 33, 20, 21, 18, 21, 22, 17, 29, 26, 19, 19, 22, 30, 18, 18, 15, 2~
## $ lwt   <dbl> 182, 155, 105, 108, 107, 124, 118, 103, 123, 113, 95, 150, 95, 107, 1~
## $ race  <dbl> 2, 3, 1, 1, 1, 3, 1, 3, 1, 1, 3, 3, 3, 3, 1, 1, 2, 1, 3, 1, 3, 1, 1, ~
## $ smoke <dbl> 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, ~
## $ ptl   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ ht    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ ui    <dbl> 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, ~
## $ ftv   <dbl> 0, 3, 1, 2, 0, 0, 1, 1, 1, 0, 0, 1, 0, 2, 0, 0, 0, 3, 0, 1, 2, 3, 1, ~
## $ bwt   <dbl> 2523, 2551, 2557, 2594, 2600, 2622, 2637, 2637, 2663, 2665, 2722, 273~
```

Ta có cỡ mẫu $n = 189$, biến cần quan tâm là `bwt`, trung bình mẫu:

```
mean(data_birth$bwt)
```

```
## [1] 2944.656
```

Đoạn code cho quá trình bootstrap cho trung bình mẫu

```
n_bwt <- nrow(data_birth)
nR <- 1000
mu_bwt_boot <- numeric(nR)
for (i in 1:nR) {
  id_boot <- sample(1:n_bwt, size = n_bwt, replace = TRUE)
  bwt_boot <- data_birth$bwt[id_boot]
  mu_bwt_boot[i] <- mean(bwt_boot)
}
```

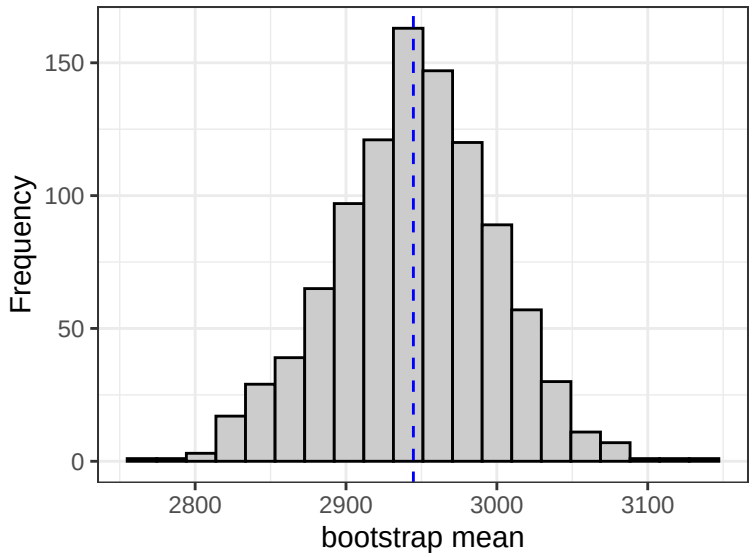
Trung bình của 1000 trung bình bootstrap là

```
mean(mu_bwt_boot)
```

```
## [1] 2945.493
```

Ta vẽ histogram của 1000 trung bình bootstrap, bằng sử dụng thư viện `ggplot2` như sau:

```
ggplot(data = data.frame(t = mu_bwt_boot), mapping = aes(x = t)) +
  geom_histogram(fill = "gray80", color = "black", bins = 20) +
  geom_vline(xintercept = mean(data_birth$bwt), color = "blue",
             linetype = "dashed") +
  xlab("bootstrap mean") + ylab("Frequency") +
  theme_bw()
```



Thực hành 1: hãy áp dụng thuật toán bootstrap cho trung bình với hàm `sapply()` thay cho vòng lặp `for()`.

Thực hành 2: Áp dụng thuật toán bootstrap cho trung vị mẫu.

2 Thư viện `boot`

Trong `R`, phương pháp bootstrap được thi hành bởi hàm `boot()` trong thư viện `boot`:

```
library(boot)
out_boot <- boot(data, statistic, R, sim = "ordinary", ...)
```

trong đó,

- `data` là tên của dữ liệu phân tích;
- `statistic` là tên hàm dùng để ước lượng một hoặc nhiều tham số;
- `R` là số lần lấy lại mẫu bootstrap;
- `sim` là lựa chọn phương pháp bootstrap, trong đó, `"ordinary"` tương ứng với phương pháp bootstrap cơ bản.

Ví dụ 2: Xét lại dữ liệu `birthwt.csv`. Ta muốn khảo sát phân phối mẫu của ước lượng trung bình của cân nặng của trẻ sơ sinh `bwt`.

Đầu tiên, ta cần viết hàm `statistic` để ước lượng trung bình trong mỗi lần lặp mẫu.

```
boot_mu_fun <- function(data, ind){
  data_new <- data[ind]
  out <- mean(data_new)
  return(out)
}
```

trong đó,

- `data` là vector chứa giá trị quan sát của mẫu;
- `ind` là vector chứa vị trị của dữ liệu được lựa chọn ngẫu nhiên.

Bên trong thân hàm, ta tính giá trị trung bình, và trả về bằng hàm `return()`. Bây giờ, ta sẽ nhúng hàm vừa viết `boot_mu_fun()` vào hàm `boot()`, và thực hiện 1000 lần lặp:

```
set.seed(34)
out_1 <- boot(data = data_birth$bwt, statistic = boot_mu_fun, R = 1000)
```

Chú ý, ở đây ta dùng `set.seed()` gieo "hạt mầm" nhằm giữa kết quả lặp bootstrap không thay đổi khi chạy lại đoạn code. Số ở trong `set.seed()` có thể được thay đổi theo ý thích. Kết quả thu được

```
out_1
```

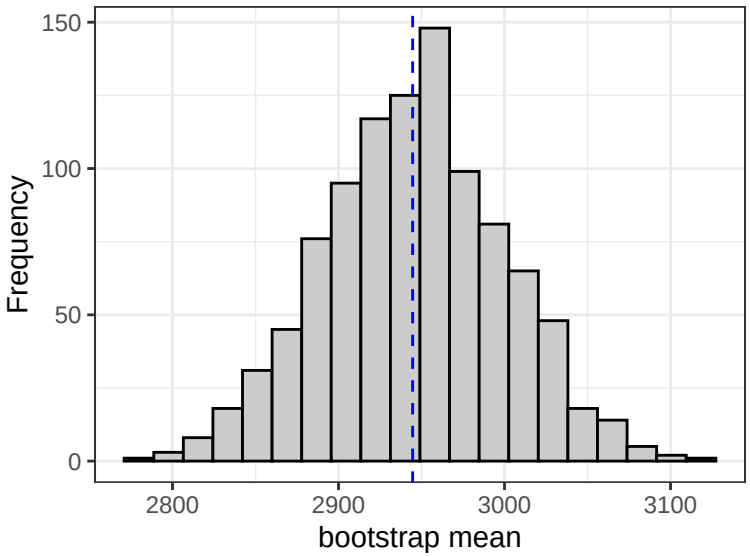
```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = data_birth$bwt, statistic = boot_mu_fun, R = 1000)
##
##
## Bootstrap Statistics :
##     original    bias    std. error
## t1* 2944.656 0.6391534    54.12104
```

Cột

- `original` là giá trị trung bình của `bwt`;
- `bias` là giá trị của độ chệch giữa giá trị trung bình (mẫu gốc) và trung bình của ước lượng bootstrap;
- `std. error` là sai số chuẩn của ước lượng bootstrap.

Các giá trị ước lượng bootstrap được lưu trong `out_1$t`. Ta có thể vẽ histogram để xác định phân phối mẫu của ước lượng.

```
ggplot(data = data.frame(t = out_1$t), mapping = aes(x = t)) +
  geom_histogram(fill = "gray80", color = "black", bins = 20) +
  geom_vline(xintercept = out_1$t0, color = "blue", linetype = "dashed") +
  xlab("bootstrap mean") + ylab("Frequency") +
  theme_bw()
```



Thực hành 3: Áp dụng hàm `boot` để thu được kết quả của quá trình bootstrap cho trung vị của `bwt`.

3 Bài tập

Bài tập 1: Xét dữ liệu `birthwt.txt`. Sử dụng phương pháp bootstrap để xác định phân phối mẫu và sai số chuẩn cho các tham số sau:

(a) trung bình của `lwt`;

(b) trung vị của `age`;

(c) tương quan giữa `lwt` và `bwt`.

Vẽ histogram cho từng trường hợp.

Bài tập 2: Xét dữ liệu `state.csv`. Hãy xác định phân phối bootstrap cho trung bình có trọng số của tỷ lệ vụ án giết người.

Bài tập 3: Dữ liệu dưới đây được Bradley Efron sử đụng để minh họa quá trình bootstrap. Dữ liệu chứa kết quả điểm thi LSAT (điểm xét tuyển vào trường luật) và GPA của học sinh.

`LSAT`	576	635	558	578	666	580	555	661
	651	605	653	575	545	572	594	
`GPA`	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43
	3.36	3.13	3.12	2.74	2.76	2.88	3.96	

(a) vẽ biểu đồ phân tán mô tả sự tương quan giữa hai biến `LSAT` và `GPA`;

(b) tính hệ số tương quan giữa hai biến;

(c) áp dụng phương pháp bootstrap để xác định phân phối mẫu của hệ số tương quan.

Bài tập 4: Xét dữ liệu `birthwt.txt`, xét mô hình hồi quy tuyến tính

$$\texttt{bwt} = \beta_0 + \beta_1 \texttt{age} + \varepsilon,$$

(a) ước lượng mô hình hồi quy tuyến tính trên;

(b) áp dụng phương pháp bootstrap để xác định phân phối mẫu của các hệ số của mô hình;

(c) lặp lại các thao tác trong câu (a) và (b) cho mô hình

$$\texttt{bwt} = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{smoke} + \varepsilon,$$

Bài tập 5: Xem xét một biến ngẫu nhiên có phân phối chuẩn $\mathcal{N}(36, 8^2)$.

(a) Tạo ngẫu nhiên một mẫu với cỡ $n = 200$ từ phân phối này, hãy tính trung bình và độ lệch chuẩn mẫu, từ đây, suy ra phân phối mẫu cho trung bình mẫu;

(b) áp dụng phương pháp bootstrap để xác định phân phối mẫu của mẫu vừa tạo, và so sánh với phân phối mẫu được xác định bằng lý thuyết;

(c) lặp lại các thao tác trong câu (a) và (b) cho trường hợp cỡ mẫu $n = 50$ và $n = 10$, nhận xét kết quả.

Bài tập 6: Xét dữ liệu `brucellosis.csv` chứa dữ liệu của 35 bệnh nhân bị bệnh Brucellosis (là một bệnh do vi khuẩn nội bào lây truyền sang người chủ yếu do tiếp xúc với động vật bị nhiễm bệnh hoặc do ăn phải các sản phẩm sữa chưa tiệt trùng) và 15 người khỏe mạnh.

```
## # A tibble: 6 x 2
##   scores group
##    <dbl> <dbl>
## 1     75     0
## 2     70     0
## 3     78     0
## 4     68     0
## 5     62     0
## # i 1 more row
```

Trong đó,

- `scores` là % tế bào CD3-positive lymphocytes (một protein trên bề mặt của tế bào lympho T, phản ứng lại sự ngoại xâm tế bào) của các bệnh nhân trong nghiên cứu;
- `group` cung cấp thông tin của nhóm bệnh nhân, với `0` là nhóm không bị bệnh, và `1` là nhóm bị bệnh brucellosis.

Khi một người bị nhiễm vi khuẩn, % tế bào CD3-positive lymphocytes sẽ tăng lên, và do đó, có thể dùng % tế bào CD3-positive lymphocytes như là một chỉ số để chuẩn đoán bệnh brucellosis.

Gọi X và Y lần lượt là giá trị của `scores` trong nhóm `0` và `1`, ta quan tâm tới đại lượng $\theta = \Pr(X < Y)$ như một thước đo độ chính xác trong chuẩn đoán bệnh nhân brucellosis bằng CD3.

(a) Hãy vẽ một biểu đồ mô tả sự so sánh % tế bào CD3-positive lymphocytes giữa hai nhóm bệnh nhân.

(b) Viết đoạn chương trình tính $\widehat{\theta}$

$$\widehat{\theta} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{I}(X_i < Y_j).$$

(c) Áp dụng phương pháp bootstrap để xác định phân phối mẫu của $\widehat{\theta}$.