Bài giảng R, số 4
-Phương pháp Bootstrap-

TS.Tô Đức Khánh

18/03/2024

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, cũng như xác định khoảng tin cậy cho tham số được quan tâm.

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.

Ví dụ 1: Xét dữ liệu `birthwt.csv`. 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~
```

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()` treo "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
```

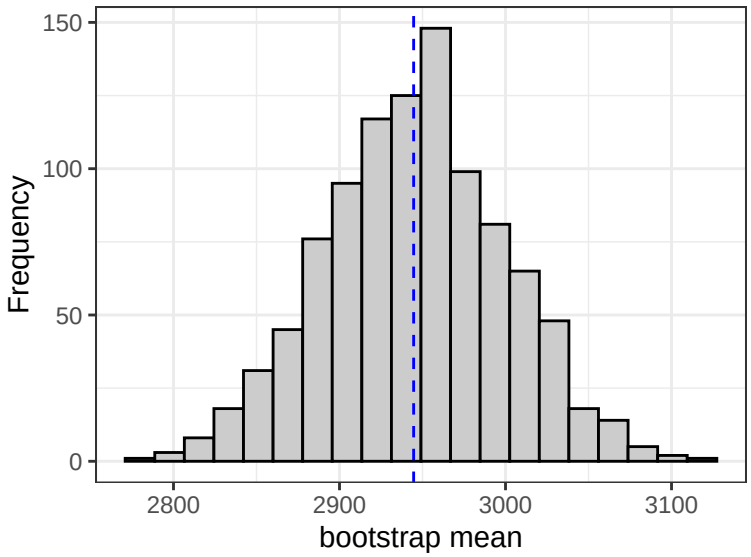
```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = data_birth$bwt, statistic = boot_mu_fun, R = 1000)
##
##
## Bootstrap Statistics :
##     original    bias    std. error
## t1* 2944.656 0.6391534    54.12104
```

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()
```



Để xác định khoảng tin cậy bootstrap percentile, ta dùng hàm:

```
boot.ci(boot.out, type = "perc", conf = 0.95)
```

với `conf` là độ tin cậy.

```
boot.ci(out_1, type = "perc", conf = 0.95)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = out_1, conf = 0.95, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   (2839, 3052 )
## Calculations and Intervals on Original Scale
```

Kết quả cho ta 95% khoảng tin cậy cho trung bình cân nặng của trẻ sơ sinh.

Bài tập 1: Xét dữ liệu `birthwt.csv`. Sử dụng phương pháp bootstrap để xác định sai số chuẩn và khoảng tin cậy 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 khoảng tin cậy bootstrap cho trung bình có trọng số của tỷ lệ vụ án giết người.

Bài tập 3: Xét dữ liệu `flights`. Hãy xác định khoảng tin cậy bootstrap cho trung bình cho số phút cất cánh trễ trong cách kịch bản sau:

(a) một mẫu ngẫu nhiên của 250 chuyến bay bất kỳ trong năm 2013;

(b) một mẫu ngẫu nhiên của 300 chuyến bay bất kỳ của hãng United Airline;

(c) một mẫu ngẫu nhiên của 250 chuyến bay bất kỳ của hãng Delta.

Bài tập 4: Xây dựng khoảng tin cậy bootstrap cho sự khác biệt giữa trung bình số phút cất cánh trễ của hai mẫu ngẫu nhiên được tạo trong câu (b) và (c) của bài tập 3.

Bài tập 5: Xét dữ liệu `flights`. Chọn ngẫu nhiên một mẫu với cỡ 1000 chuyến bay trong các tháng mùa đông (tháng 11, 12, 1).

(a) Tính tỷ lệ chuyến bay có số phút cất cánh trễ nhiều hơn 30 phút.

(b) Xác định khoảng tin cậy bootstrap cho tỷ lệ chuyến bay có số phút cất cánh trễ nhiều hơn 30 phút.

(c) Xác định khoảng tin cậy bootstrap cho chênh lệch tỷ lệ giữa tỷ lệ chuyến bay có số phút cất cánh trễ nhiều hơn 30 phút của hai hãng EV và DL.