Bài giảng R, số 5
-A/B testing với R-

TS.Tô Đức Khánh

20/03/2024

1 A/B testing

Trong phần này ta sẽ thực hành A/B testing với R.

Ví dụ 1: (Web stickiness) Ta xét dữ liệu đo thời gian phiên tương tác (session time, đơn vị giây) của khách hàng với hai trang web A và B. Thời gian này được dùng như một thước đo về sự thu hút của trang web đối với khách hàng. Về mặt ý tưởng, nếu thời gian phiên tương tác càng dài thì sức thu hút của trang web với khách hàng càng lớn. Dữ liệu được ghi nhận bởi Google Analytics, và được tổng hợp trong file web_page_data.csv

```
data_web <- read_csv(file = "datasets/web_page_data.csv")
data_web <- data_web |> clean_names()
glimpse(data_web)
```

```
## Rows: 36
## Columns: 2
## $ page <chr> "Page A", "Page B", "Page A", "Page B", "Page A", "Page B", "Page A", ~
## $ time <dbl> 0.21, 2.53, 0.35, 0.71, 0.67, 0.85, 2.11, 2.46, 1.32, 1.49, 0.68, 0.75~
```

```
data_web |> group_by(page) |>
  summarise(n = n(), mean = mean(time), sd = sd(time))
```

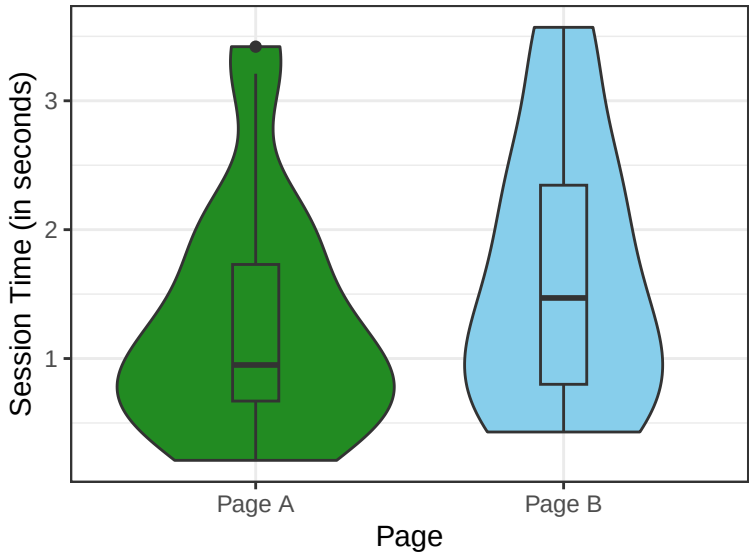
```
## # A tibble: 2 x 4
##   page       n    mean       sd
##   <chr>  <int>   <dbl>    <dbl>
## 1 Page A    21 1.26333 0.884632
## 2 Page B    15 1.62    1.01136
```

Kết quả cho thấy có 21 người tương tác với trang web A và 15 người tương tác với web B. Trung bình thời gian phiên làm việc với web A là thấp hơn web B. Tuy nhiên, độ lệch chuẩn của thời gian làm việc với web B lớn hơn, cho thấy độ biến động thời gian sử dụng giữa các người dùng web B là lớn hơn so với web A.

Biểu đồ violin dưới đây giúp ta khẳng định các nhận định trên, đồng thời cung cấp thêm thông tin về phân phối thời gian phiên làm việc, ở đây, dữ liệu của cả hai trang web đều cho thấy phân phối bất đối xứng và lệch phải của thời gian phiên làm việc.

```
ggplot(data_web, aes(x = page, y = time, fill = page)) +
  geom_violin() +
  geom_boxplot(width = 0.15) +
  scale_fill_manual(breaks = c("Page A", "Page B"),
                    values = c("forestgreen", "skyblue")) +
  labs(x = "Page", y = "Session Time (in seconds)") +
```

```
  theme_bw() +
  theme(legend.position = "none")
```



Thông qua bảng tổng hợp và biểu đổ violin, một giả định có thể là "thời gian phiên làm việc của web B là dài hơn web A". Do đó, ta cần kiểm chứng giả thuyết và đối thuyết sau:

Giả thuyết: $\mu_A = \mu_B$

Đối thuyết: $\mu_A < \mu_B$

Nếu Giả thuyết là đúng thì có nghĩa là sự dài hơn trong thời gian phiên làm việc của web B đối với web A chỉ là kết quả của sự ngẫu nhiên (không có ý nghĩa thống kê). Để kiểm định Giả thuyết này, ta áp dụng Permutation test, và p-value sẽ được tính cho kiểm định bên trái.

Đầu tiên ta viết hàm để xáo trộn dữ liệu trong hai nhóm, và tính sự khác biệt giữa trung bình của hai nhóm mới.

```
perm_fun <- function(x, nA, nB, R) {
  n <- nA + nB
  mean_diff <- numeric(R)
  for (i in 1:R){
    idx_a <- sample(x = 1:n, size = nA)
    idx_b <- setdiff(x = 1:n, y = idx_a)
    mean_diff[i] <- mean(x[idx_a]) - mean(x[idx_b])
  }
  return(mean_diff)
}
```

Chú ý:

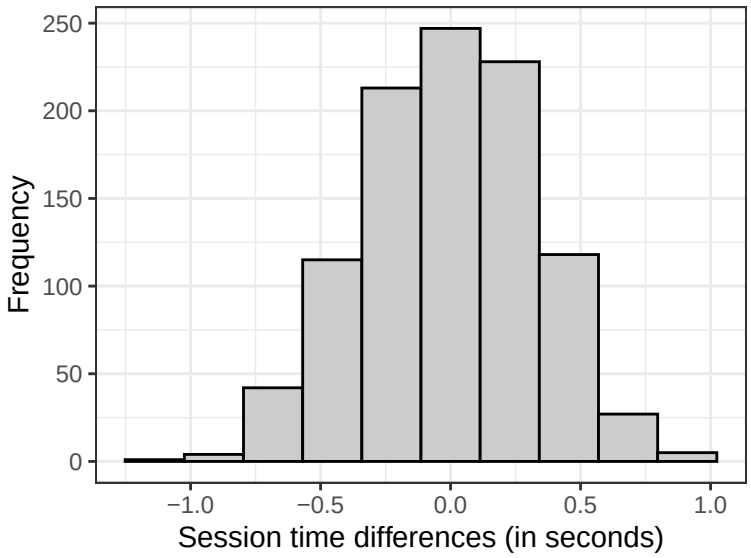
- hàm `sample()` để lấy một mẫu ngẫu nhiên từ tập gốc `x` với cỡ `size`;
- hàm `setdiff()` để in ra các phần tử khác nhau khi so sánh hai mẫu `x` và `y`.

Bây giờ ta sẽ sử dụng hàm này:

```
set.seed(21)
diff_mean_perm <- perm_fun(data_web$time, nA = 21, nB = 15, R = 1000)
```

Ta vẽ histogram để biểu diễn phân phối của kết quả từ 1000 lần hoán vị.

```
ggplot(data = tibble(perm_diffs = diff_mean_perm), aes(x = perm_diffs)) +
  geom_histogram(bins = 10, fill = "gray80", color = "black") +
  labs(x = "Session time differences (in seconds)", y = "Frequency") +
  theme_bw()
```



Giá trị p-value được tính bởi:

```
mean_a <- mean(data_web$time[data_web$page == 'Page A'])
mean_b <- mean(data_web$time[data_web$page == 'Page B'])
mean(diff_mean_perm < (mean_a - mean_b))
```

```
## [1] 0.15
```

Với mức ý nghĩa $\alpha = 0.05$, kết quả cho thấy Giả thuyết là không thể bị bác bỏ. Do đó, việc thời gian phiên làm việc với web B dài hơn so với web A là không có ý nghĩa thống kê, hay chỉ là kết quả của sự ngẫu nhiên.

Bài tập 1.1: Viết hàm thực hiện các bước xử lý của một permutation test cho hai trung bình, với

- input:

  `x`: vector chứa giá trị của biến định lượng cần kiểm định;

  `y`: vector chứa tên (nhãn) của hai nhóm;

  `R`: số lần hoán vị;

  `alter`: vector chứa thông tin về chiều của Đối thuyết (left, right, two-sided).

- output: một `list` bao gồm

  `res_perm`: vector chứa giá trị khác biệt giữa trung bình của hai nhóm;

  `mean_A`: trung bình của nhóm A (tính trên mẫu gốc);

  `mean_B`: trung bình của nhóm B (tính trên mẫu gốc);

  `p-value`: giá trị p-value của kiểm định.

Sau đó, hãy thử nghiệm với dữ liệu Web stickiness. *Gợi ý:* sử dụng hàm `split()` để tách dữ liệu thành hai nhóm theo vector chứa thông tin nhãn.

Bài tập 1.2: Xét dữ liệu `Beerwings.csv` cung cấp số lượng cánh gà dán cay (hotwings) được tiêu thụ với khách hàng là nam giới hoặc nữ giới.

(a) Tạo bảng tổng hợp và biểu đồ để khám phá lượng thu tiệu cánh gà theo hai nhóm giới tính.

(b) Dựa vào kết quả tìm thấy ở câu (a), hãy nêu ra Giả thuyết liên quan tới sự ảnh hưởng của giới tính tới số lượng cánh gà được tiêu thụ.

Bài tập 1.3: Dữ liệu `Verizon.csv` cung cấp thời gian sửa chữa được tiến hành của công ty viễn thông Verizon (Mỹ), dành cho hai nhóm: khách hàng thuộc Verizon (`ILEC`), và khách hàng thuộc các công ty đối thủ (`CLEC`).

(a) Tạo bảng tổng hợp và biểu đồ để khám phá thời gian sửa chữa dịch vụ của Verizon theo hai nhóm khách hàng.

(b) Ủy ban Tiện ích Công cộng New York (PUC) nghi ngờ rằng Verizon đang thực hiện việc sửa chữa chậm hơn đối với khách hàng của đối thủ cạnh tranh. Hàng nghìn cuộc thử nghiệm được thực hiện để so sánh tốc độ của các loại sửa chữa khác nhau, trong những khoảng thời gian khác nhau, so với các đối thủ cạnh tranh khác nhau. Nếu về cơ bản hơn 1% số thử nghiệm cho p-value dưới 1% thì Verizon được coi là đang phân biệt đối xử. Dựa vào dữ liệu này, hãy kiểm chứng Giả thuyết của PUC.

Bài tập 1.4: Viết hàm thực hiện bootstrap permuation test, và áp dụng cho các bài tập trên.

2 A/B testing cho nhiều nhóm

Trong phần này, ta áp dụng các thuật toán Permutation ANOVA vào trong việc so sánh trung bình của nhiều nhóm.

Ví dụ 2: Dữ liệu `four_sessions.csv` chứa dữ liệu về thời gian phiên làm việc của 4 trang web. Ta mong muốn kiểm tra xem liệu thời gian phiên làm việc của 4 trang web này có bằng nhau hay không?

```
data_web4 <- read_csv(file = "datasets/four_sessions.csv")
data_web4 <- data_web4 |> clean_names()
glimpse(data_web4)
```

```
## Rows: 20
## Columns: 2
## $ page <chr> "Page 1", "Page 2", "Page 3", "Page 4", "Page 1", "Page 2", "Page 3", ~
## $ time <dbl> 164, 178, 175, 155, 172, 191, 193, 166, 177, 182, 171, 164, 156, 185, ~
```

Bảng tổng hợp cho thấy trung bình thời gian phiên làm việc của 4 trang web là khác nhau (về mặt giá trị). Trong khi đó, độ lệch chuẩn là cho thấy độ biến động trong thời gian phiên làm việc là khác biệt giữa các nhóm.

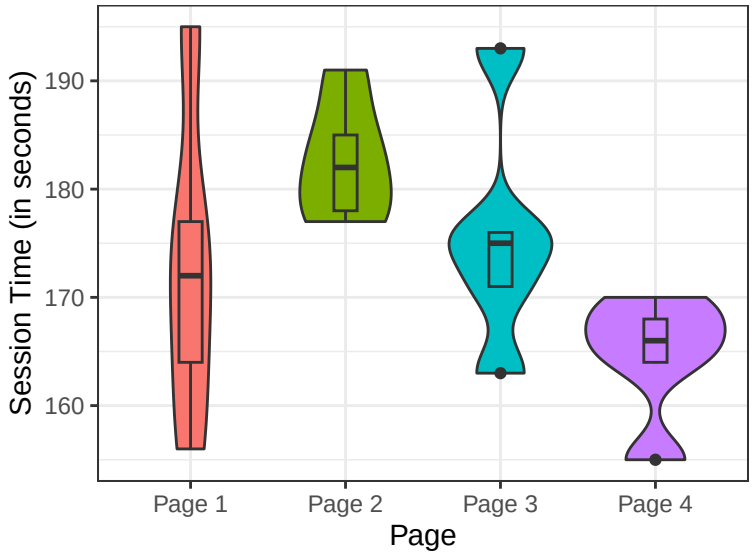
```
data_web4 |> group_by(page) |>
  summarise(n = n(), mean = mean(time), sd = sd(time))
```

```
## # A tibble: 4 x 4
##   page       n  mean       sd
##   <chr>  <int> <dbl>    <dbl>
## 1 Page 1     5 172.8 14.7547
## 2 Page 2     5 182.6  5.68331
## 3 Page 3     5 175.6 10.9909
## 4 Page 4     5 164.6  5.81378
```

Biểu đồ violin làm rõ hơn cho nhận xét ở trên. Chú ý rằng, vì số lượng quan sát trong mỗi nhóm là nhỏ (5 mỗi nhóm) nên ước lượng hàm mật độ xác suất của dữ liệu trong mỗi nhóm là không đủ tin cậy để đưa ra

nhận xét.

```
ggplot(data_web4, aes(x = page, y = time, fill = page)) +
  geom_violin() +
  geom_boxplot(width = 0.15) +
  labs(x = "Page", y = "Session Time (in seconds)") +
  theme_bw() +
  theme(legend.position = "none")
```



Để thực hiện Permutation ANOVA, trong `R`, ta có hàm `aovp()` được cung cấp bởi thư viện `lmPerm`. Cấu trúc của hàm này như sau:

```
out_aovp <- aovp(formula = x ~ y, data = data_name, perm = ...)
```

Trong đó,

- `formula` là công thức xác định mô hình phân tích ANOVA;
- `x` là tên của biến định lượng cần phân tích;
- `y` là tên của biến định tính, chứa nhãn của các nhóm;
- `data` là tên của dữ liệu chứa `x` và `y`;
- `perm` là tên của các phương pháp Permutation:
  - `"Exact"` tương ứng với Exhaustive Permutation ANOVA, chỉ dùng khi cỡ mẫu của dữ liệu nhỏ hơn 10;
  - `"Prob"` tương ứng với Permutation ANOVA, áp dụng cho các trường hợp cỡ mẫu lớn hơn 10;
  - `"SPR"` tương ứng với Permutation ANOVA.

```
library(lmPerm)
set.seed(56)
out_aov_1 <- aovp(formula = time ~ page, data = data_web4, perm = "Prob")
```

```
## [1] "Settings:  unique SS "
```

```
summary(out_aov_1)
```

```
## Component 1 :
```

```
##             Df R Sum Sq R Mean Sq Iter Pr(Prob)
## page1        3    831.4    277.13 4720  0.06886 .
## Residuals   16   1618.4    101.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Giá trị p-value được cho cung cấp bởi cột `Pr(Prob)` là 0.06886. Số lần lấy mẫu lặp lại là 4720, được cung cấp bởi cột `Iter`.

Với kết quả p-value này, ta có thể kết luận rằng, sự khác biệt trong thời gian phiên làm việc của 4 trang web là không có ý nghĩa thống kê, tại mức ý nghĩa 5%. Chú ý rằng, ta có cỡ mẫu nhỏ, nên có thể nếu dữ liệu được thu thập thêm, sự khác biệt có thể có ý nghĩa thống kê.

Bài tập 2.1 Dữ liệu `ChiMarathonMen.csv` chứa dữ liệu về một mẫu nam giới, ở độ tuổi từ 20 đến 39, đã hoàn thành Chicago Marathon năm 2015. Các vận động viên chạy được chia thành các nhóm tuổi 20–24, 25–29, 30–34 và 34–39. Thời gian hoàn thành của họ được tính bằng phút.

(a) Tạo bảng tổng hợp và biểu đồ để khám phá thời gian hoàn thành cuộc đua của các nam vận động viên theo nhóm tuổi.

(b) Tiến hành kiểm tra ANOVA để xem liệu thời gian hoàn thành trung bình giữa các nhóm tuổi có giống nhau hay không.

Bài tập 2.2 Starcraft là một trò chơi điện tử chiến lược nổi tiếng có chủ đề quân sự khoa học viễn tưởng. Người chơi chọn trở thành một trong ba chủng tộc - Terran, Zergs hoặc Protoss - và cạnh tranh để giành quyền thống trị ở một phần xa của thiên hà Milky Way. Tệp `Starcraft.csv` chứa thông tin về mẫu các tuyển thủ hàng đầu Hàn Quốc từ cơ sở dữ liệu http://www.teamliquid.net/tlpd/players (J. Evans). Ngoài chủng tộc mà người chơi đã chọn, tệp còn chứa tuổi của anh ta cũng như số trận thắng (trong số 40 trận gần đây nhất của anh ta).

(a) Tạo bảng tổng hợp và biểu đồ để khám phá tuổi của người chơi theo nhóm chủng tộc đã được chọn.

(b) Sử dụng Permutation ANOVA để xác định xem tuổi trung bình của các game thủ có giống nhau giữa các chủng tộc hay không.

(c) Lặp lại (a) và (b) cho số trận thắng trung bình của ba chủng tộc.

Bài tập 2.3 Xét các chuyến bay của hãng United Airlines trong dữ liệu `Fligts`.

(a) Tạo bảng tổng hợp và biểu đồ để khám phá thời gian cất cánh trễ của các chuyến bay theo các ngày trong tuần (từ thứ 2 tới chủ nhật).

(b) Tiến hành kiểm tra ANOVA để xem liệu thời gian cất cánh trễ của các chuyến bay theo các ngày trong tuần có giống nhau hay không.