

Báo cáo bài tập 6

1612174 - Phùng Tiến Hào - tienhaophung@gmail.com

05/05/2019

Contents

1	Khảo sát một biến	1
1.1	Biến định tính (Categorical variable)	1
1.1.1	DecisionMale (Yes/No)	1
1.1.2	RaceF (Caucasian, Asian,..., Other)	2
1.2	Biến định lượng (Quantative variable)	3
1.2.1	AttractiveM (0-10)	3
1.2.2	LikeM (0-10)	6
1.2.3	SincereM (0-10)	8
2	Khảo sát cặp biến	11
2.1	Biến định tính vs biến định tính	11
2.1.1	Kiểm định cho tỉ lệ khác biệt giữa nữ da trắng nhận phản hồi "Yes" và "No"	12
2.1.2	Kiểm định cho tỉ lệ khác biệt giữa nữ châu Á nhận phản hồi "Yes"/"No"	13
2.2	Biến định tính và biến định lượng	15
2.2.1	Kiểm định cho độ chênh lệch kì vọng giữa mức độ hấp dẫn của nữ trong phản hồi "Yes" và mức độ hấp dẫn của nữ trong phản hồi "No"	15
2.2.2	Kiểm định cho độ chênh lệch trung vị (median) giữa mức độ hấp dẫn của nữ trong phản hồi "Yes" và mức độ hấp dẫn của nữ trong phản hồi "No"	17
2.3	Biến định lượng và biến định lượng	19
2.3.1	Kiểm định cho hệ số tương quan giữa AttractiveM và LikeM	21
2.3.2	Kiểm định cho hệ số hồi qui β (regression slope) của regression line	22

Dữ liệu khảo sát: SpeedDating trong package Lock5withR

Load package và thêm các thư viện cần thiết trước khi đi vào xử lý:

```
1 require(Lock5withR) # Load package
2 library(Lock5withR)
3 library(mosaic)
4 head(SpeedDating)
5 attach(SpeedDating) # Avoid dollar sign before each variables name
6
```

1 Khảo sát một biến

1.1 Biến định tính (Categorical variable)

1.1.1 DecisionMale (Yes/No)

Giả sử, ta cần khảo sát tỉ lệ nam phản hồi (Yes/No) cho quần thể (population) là toàn bộ học sinh nam của trường Columbia. Từ tổng thể, ta thu thập được một mẫu dữ liệu ngẫu nhiên (random sample) gồm 276 quan sát trong đó có 146 phản hồi "Yes" và 130 phản hồi "No". Dựa vào mẫu dữ liệu này, ta kiểm định nghi vấn "tỉ lệ phản hồi Yes cao hơn phản hồi No" với mức ý nghĩa (significance level) 5%.

Gọi p là tỉ lệ nam phản hồi "Yes" trong trường, \hat{p} là tỉ lệ nam phản hồi "Yes" trong mẫu dữ liệu. Ta có

$$\hat{p} = \frac{146}{276} = 0.529$$

Để tính đến sự biến động của \hat{p} theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : p = p_0 = 0.5 \\ H_1 : p > 0.5 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

```
1 # TK can tính
2 stat <- function(data){
3   return (mean(data == "Yes")) # Tỉ lệ
4 }
5
6 randomization <- function(B){
7   return (replicate(B, stat(sample(nullsample, n, replace = TRUE))))
8 }
9
10 > # Mau du lieu ban dau
11 > sample <- DecisionMale
12 > # Kich thuc mau, tham so mac dinh, ti le mau, muc y nghĩa
13 > (n <- length(sample)); (p0 <- 0.5); (p_hat <- stat(sample)); (alpha <- 0.05)
14 [1] 276
15 [1] 0.5
16 [1] 0.5289855
17 [1] 0.05
18 > nullsample <- c(rep("Yes", n/2), rep("No", n/2)) # Mau du lieu tuong ung voi H0
19
```

```

20 > # Layphanphoi cua randomization
21 > rand_dist <- randomization(10000)
22 > # Tinh p-value trong kiem dinh mot phia (one-tailed)
23 > (p_value <- mean(rand_dist >= p_hat))
24 [1] 0.1847
25 > # Tinh gia tri toi han (critical value) voi muc phan vi 1-alpha
26 > (crit_val <- quantile(rand_dist, 1 - alpha, names = FALSE))
27 [1] 0.5507246
28 > # Kiem tra xem p_value co be hon alpha, neu co thi bac bo H0
29 > p_value < alpha; crit_val < p_hat
30 [1] FALSE
31 [1] FALSE
32

```

Vì $p\text{-value} > \alpha$ nên ta không bác bỏ H_0 . Tương tự, ta có giá trị tới hạn (critical value) $\text{crit_val} > \hat{p}$ nên ta không bác bỏ H_0 .

Như vậy, với mức ý nghĩa 5%, tỉ lệ phản hồi "Yes" bằng phản hồi "No" của trường.

1.1.2 RaceF (Caucasian, Asian,..., Other)

Giả sử, ta cần khảo sát tỉ lệ dân tộc nữ (Caucasian, Asian,..., Other) cho quần thể (population) là toàn bộ học sinh nữ của trường Columbia. Từ tổng thể, ta thu thập được một mẫu dữ liệu ngẫu nhiên (random sample) gồm 276 quan sát trong đó có 4 rỗng, 70 Asians, 15 Blacks, 148 Caucasians, 23 Latino và 16 Others. Dựa vào mẫu dữ liệu này, ta kiểm định nghi vấn "tỉ lệ nữ da trắng cao hơn tỉ lệ dân tộc nữ các nhóm còn lại" với mức ý nghĩa (significance level) 5%.

Gọi p là tỉ lệ nữ da trắng trong trường, \hat{p} là tỉ lệ nữ da trắng trong mẫu dữ liệu. Ta có

$$\hat{p} = \frac{148}{276} = 0.536$$

Để tính đến sự biến động của \hat{p} theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : p = p_0 = 0.5 \\ H_1 : p > 0.5 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

```

1 # TK can tinh
2 stat <- function(data){
3   return (mean(data)) # Ti le
4 }
5
6 randomization <- function(B){
7   return (replicate(B, stat(sample(nullsample, n, replace = TRUE))))
8 }
9
10 > # Mau du lieu ban dau
11 > sample <- RaceF
12 > # Kich thuoc mau, tham so mac dinh, ti le mau, muc y nghĩa
13 > (n <- length(sample)); (p0 <- 0.5); (p_hat <- mean(sample == "Caucasian")); (alpha <- 0.05)
14 [1] 276
15 [1] 0.5
16 [1] 0.5362319
17 [1] 0.05

```

```

18 > nullsample <- c(rep(1, n/2), rep(0, n/2)) # Mau du lieu tuong ung voi H0
19
20 > # Layphan phoi cua randomization
21 > rand_dist <- randomization(10000)
22 > # Tinh p-value trong kiem dinh mot phia (one-tailed)
23 > (p_value <- mean(rand_dist >= p_hat))
24 [1] 0.1295
25 > # Tinh gia tri toi han (critical value) voi muc phan vi 1-alpha
26 > (crit_val <- quantile(rand_dist, 1 - alpha, names = FALSE))
27 [1] 0.5507246
28 > # Kiem tra xem p_value co be hon alpha, neu co thi bac bo H0
29 > p_value < alpha; crit_val < p_hat
30 [1] FALSE
31 [1] FALSE
32

```

Vì $p\text{-value} > \alpha$ nên ta không bác bỏ H_0 . Tương tự, ta có giá trị tới hạn (critical value) $\text{crit_val} > \hat{p}$ nên ta không bác bỏ H_0 .

Như vậy, với mức ý nghĩa 5%, tỉ lệ nữ da trắng bằng tỉ lệ dân tộc nữ các nhóm còn lại của trường.

1.2 Biến định lượng (Quantative variable)

1.2.1 AttractiveM (0-10)

a) Kiểm định cho kì vọng

Giả sử, ta cần khảo sát kì vọng (mean) mức độ quyến rũ của nữ (0,1,...,10) cho quần thể (population) là toàn bộ học sinh nữ của trường Columbia. Từ tổng thể, ta thu thập được một mẫu dữ liệu ngẫu nhiên (random sample) gồm 276 quan sát. Dựa vào mẫu dữ liệu này, ta kiểm định nghi vấn "Kì vọng mức độ quyến rũ của sinh viên nữ trong trường là 6.6" với mức ý nghĩa (significance level) 5%.

Gọi μ là mức độ quyến rũ trung bình của sinh nữ trong trường, \bar{x} là mức độ quyến rũ trung bình của sinh nữ trong mẫu dữ liệu. Ta có

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = 6.687$$

Để tính đến sự biến động của \bar{x} theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : \mu = \mu_0 = 6.6 \\ H_1 : \mu \neq 6.6 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

```

1 # TK can tinh
2 stat <- function(data){
3   return (mean(data, na.rm = TRUE)) # tb mau
4 }
5
6 randomization <- function(B){
7   return (replicate(B, stat(sample(nullsample, n, replace = TRUE))))
8 }

```

```

9
10 > # Mau du lieu ban dau
11 > sample <- AttractiveM
12 > # Kich thuc mau, tham so mac dinh, trung binh mau, muc y nghia
13 > (n <- length(sample)); (mu0 <- 6.6); (x_bar <- stat(sample)); (alpha <- 0.05)
14 [1] 276
15 [1] 6.6
16 [1] 6.686813
17 [1] 0.05
18 > nullsample <- sample - (x_bar - mu0) # Mau du lieu tuong ung voi H0
19 > # Check lai mean cua nullsample co bang mu0 chua
20 > mean(nullsample, na.rm = TRUE)
21 [1] 6.6
22
23 > # Lay phan phoi cua randomization
24 > rand_dist <- randomization(10000)
25 > # Tinh p-value trong kiem dinh hai phia (two-tailed)
26 > (p_value <- mean(abs(rand_dist - mu0) >= abs(x_bar - mu0)))
27 [1] 0.4256
28 > # Tinh gia tri toi han (critical value) voi muc phan vi 1-alpha/2
29 > (crit_val <- quantile(rand_dist, 1 - alpha/2, names = FALSE))
30 [1] 6.81172
31 > # Kiem tra xem p_value co be hon alpha, neu co thi bac bo H0
32 > p_value < alpha; abs(crit_val - mu0) < abs(x_bar - mu0)
33 [1] FALSE
34 [1] FALSE
35

```

Vì $p\text{-value} > \alpha$ nên ta không bác bỏ H_0 . Tương tự, ta có giá trị tới hạn (critical value) $\text{crit_val} > \bar{x}$ nên ta không bác bỏ H_0 .

Như vậy, với mức ý nghĩa 5%, kì vọng mức độ quấy rĩ của sinh viên nữ là 6.6.

b) Kiểm định cho trung vị (median)

Giả sử cùng tổng thể và mẫu dữ liệu ở câu a) nhưng ta kiểm định cho trung vị mức độ quấy rĩ của sinh viên nữ trong trường với mức ý nghĩa (significance level) 5% thay vì trung bình của mức độ quấy rĩ. Mặc dù trung bình thường được sử dụng như là con số mô tả trọng tâm của phân phối nhưng nó lại rất nhạy cảm với ngoại lệ (outlier).

Gọi med là median mức độ quấy rĩ của sinh nữ trong trường, \hat{med} là median mức độ quấy rĩ của sinh nữ trong mẫu dữ liệu. Ta có

$$\hat{med} = 7.000$$

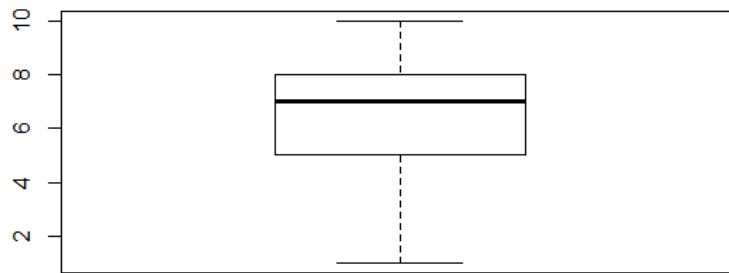


Figure 1: Boxplot của AttractiveM

Ta thấy rằng, dữ liệu này khá tốt khi không có ngoại lệ nhưng để chắc chắn thì ta sẽ kiểm định khoảng tin cậy cho trung vị của AttractiveM.

Để tính đến sự biến động của med theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : med = med_0 = 7.0 \\ H_1 : med \neq 7.0 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

1

```

1  # TK can tính
2  stat <- function(data){
3    return (median(data, na.rm = TRUE)) # tb mau
4  }
5
6  > # tham so mac dinh, trung vi mau, muc y nghia
7  > (med0 <- 7.0); (med_hat <- stat(sample)); (alpha <- 0.05)
8  [1] 7
9  [1] 7
10 [1] 0.05
11 > nullsample <- sample - (med_hat - med0) # Mau du lieu tuong ung voi H0
12 > stat(nullsample)
13 [1] 7
14
15 > # Lay phan phoi cua randomization
16 > rand_dist <- randomization(10000)
17 > # Tinh p-value trong kiem dinh hai phia (two-tailed)
18 > (p_value <- mean(abs(rand_dist - med0) >= abs(med_hat - med0)))
19 [1] 1
20 > # Tinh gia tri toi han (critical value) voi muc phan vi 1-alpha/2
21 > (crit_val <- quantile(rand_dist, 1 - alpha/2, names = FALSE))
22 [1] 7
23 > # Kiem tra xem p_value co be hon alpha, neu co thi bac bo H0

```

¹Hàm randomization vẫn y như ở câu a)

```

24 > p_value < alpha; abs(crit_val - med0) < abs(med_hat - med0)
25 [1] FALSE
26 [1] FALSE
27

```

Vì $p\text{-value} > \alpha$ nên ta không bác bỏ H_0 . Tương tự, ta có giá trị tới hạn (critical value) $\text{crit_val} > \hat{med}$ nên ta không bác bỏ H_0 .

Như vậy, với mức ý nghĩa 5%, trung vị mức độ quấy rối của sinh viên nữ trong trường là 7.0.

1.2.2 LikeM (0-10)

a) Kiểm định cho kì vọng

Giả sử, ta cần khảo sát kì vọng (mean) mức độ thích của nam (0,1,...,10) đối với nữ cho quần thể (population) là toàn bộ sinh viên nam của trường Columbia. Từ tổng thể, ta thu thập được một mẫu dữ liệu ngẫu nhiên (random sample) gồm 276 quan sát. Dựa vào mẫu dữ liệu này, ta kiểm định nghi vấn "Kì vọng mức độ thích của sinh viên nam đối với sinh viên nữ trong trường là 6.6" với mức ý nghĩa (significance level) 5%.

Gọi μ là mức độ thích trung bình của sinh viên nam trong trường, \bar{x} là mức độ thích trung bình của sinh nam trong mẫu dữ liệu. Ta có

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = 6.682$$

Để tính đến sự biến động của \bar{x} theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : \mu = \mu_0 = 6.6 \\ H_1 : \mu \neq 6.6 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

```

1  # TK can tinh
2  stat <- function(data){
3    return (mean(data, na.rm = TRUE)) # tb mau
4  }
5
6  randomization <- function(B){
7    return (replicate(B, stat(sample(nullsample, n, replace = TRUE))))
8  }
9
10 > # Mau du lieu ban dau
11 > sample <- LikeM
12 > # Kich thuoc mau, tham so mac dinh, trung binh mau, muc y nghia
13 > (n <- length(sample)); (mu0 <- 6.6); (x_bar <- stat(sample)); (alpha <- 0.05)
14 [1] 276
15 [1] 6.6
16 [1] 6.682482
17 [1] 0.05
18 > nullsample <- sample - (x_bar - mu0) # Mau du lieu tuong ung voi H0
19 > # Check lai mean cua nullsample co bang mu0 chua
20 > mean(nullsample, na.rm = TRUE)

```



```

21 [1] 6.6
22
23 > # Layphan phoi cua randomization
24 > rand_dist <- randomization(10000)
25 > # Tinh p-value trong kiem dinh hai phia (two-tailed)
26 > (p_value <- mean(abs(rand_dist - mu0) >= abs(x_bar - mu0)))
27 [1] 0.4384
28 > # Tinh gia tri toi han (critical value) voi muc phan vi 1-alpha/2
29 > (crit_val <- quantile(rand_dist, 1 - alpha/2, names = FALSE))
30 [1] 6.811291
31 > # Kiem tra xem p_value co be hon alpha, neu co thi bac bo H0
32 > p_value < alpha; abs(crit_val - mu0) < abs(x_bar - mu0)
33 [1] FALSE
34 [1] FALSE
35

```

Vì $p\text{-value} > \alpha$ nên ta không bác bỏ H_0 . Tương tự, ta có giá trị tới hạn (critical value) $\text{crit_val} > \bar{x}$ nên ta không bác bỏ H_0 .

Như vậy, với mức ý nghĩa 5%, kì vọng mức độ thích của sinh viên nam là 6.6.

b) Kiểm định cho trung vị (median)

Giả sử cùng tổng thể và mẫu dữ liệu ở câu a) nhưng ta kiểm định cho trung vị mức độ thích của sinh viên nam đối với sinh viên nữ trong trường với mức ý nghĩa (significance level) 5% thay vì trung bình của mức độ thích. Mặc dù trung bình thường được sử dụng như là con số mô tả trọng tâm của phân phối nhưng nó lại rất nhạy cảm với ngoại lệ (outlier).

Gọi med là median mức độ thích của nam đối với nữ trong trường, \hat{med} là median mức độ thích của nam đối với nữ trong mẫu dữ liệu. Ta có

$$\hat{med} = 7.000$$

Ta có thể thấy các ngoại lệ qua boxplot sau đây:

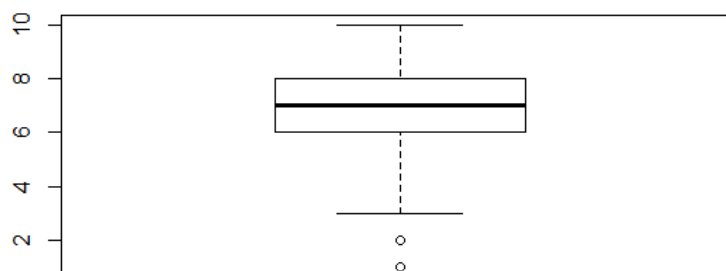


Figure 2: Boxplot của LikeM

Ta thấy rằng có một số outliers dưới 3 điểm. Trong những trường hợp như thế này ta có thể dùng trung vị là một thống kê ít bị ảnh hưởng bởi ngoại lệ.

Để tính đến sự biến động của \hat{med} theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : med = med_0 = 7.0 \\ H_1 : med \neq 7.0 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

2

```

1 # TK can tinh
2 stat <- function(data){
3   return (median(data, na.rm = TRUE)) # median
4 }
5
6 > # tham so mac dinh, trung vi mau, muc y nghia
7 > (med0 <- 7.0); (med_hat <- stat(sample)); (alpha <- 0.05)
8 [1] 7
9 [1] 7
10 [1] 0.05
11 > nullsample <- sample - (med_hat - med0) # Mau du lieu tuong ung voi H0
12 > stat(nullsample)
13 [1] 7
14
15 > # Lay phan phoi cua randomization
16 > rand_dist <- randomization(10000)
17 > # Tinh p-value trong kiem dinh hai phia (two-tailed)
18 > (p_value <- mean(abs(rand_dist - med0) >= abs(med_hat - med0)))
19 [1] 1
20 > # Tinh gia tri toi han (critical value) voi muc phan vi 1-alpha/2
21 > (crit_val <- quantile(rand_dist, 1 - alpha/2, names = FALSE))
22 [1] 7
23 > # Kiem tra xem p_value co be hon alpha, neu co thi bac bo H0
24 > p_value < alpha; abs(crit_val - med0) < abs(med_hat - med0)
25 [1] FALSE
26 [1] FALSE
27

```

Vì $p\text{-value} > \alpha$ nên ta không bác bỏ H_0 . Tương tự, ta có giá trị tới hạn (critical value) $crit_val > \hat{med}$ nên ta không bác bỏ H_0 .

Như vậy, với mức ý nghĩa 5%, trung vị mức độ thích của sinh viên nam đối với sinh viên nữ trong trường là 7.0.

1.2.3 SincereM (0-10)

a) Kiểm định cho kì vọng

Giả sử, ta cần khảo sát kì vọng (mean) mức độ chân thành (0,1,...,10) của nữ cho quần thể (population) là toàn bộ sinh viên nữ của trường Columbia. Từ tổng thể, ta thu thập được một mẫu dữ liệu ngẫu nhiên (random sample) gồm 276 quan sát. Dựa vào mẫu dữ liệu này, ta kiểm định nghi vấn "Kì vọng mức độ chân thành của sinh viên nữ trong trường là 7.8" với mức ý nghĩa (significance level) 5%.

²Hàm randomization vẫn y như ở câu a)

Gọi μ là mức độ chân thành trung bình của sinh viên nữ trong trường, \bar{x} là mức độ chân thành trung bình của sinh nữ trong mẫu dữ liệu. Ta có

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = 7.856$$

Để tính đến sự biến động của \bar{x} theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : \mu = \mu_0 = 7.8 \\ H_1 : \mu \neq 7.8 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

```

1  # TK can tinh
2  stat <- function(data){
3    return (mean(data, na.rm = TRUE)) # tb mau
4  }
5
6  randomization <- function(B){
7    return (replicate(B, stat(sample(nullsample, n, replace = TRUE))))
8  }
9
10 > # Mau du lieu ban dau
11 > sample <- SincereM
12 > # Kich thuc mau, tham so mac dinh, trung binh mau, muc y nghia
13 > (n <- length(sample)); (mu0 <- 7.8); (x_bar <- stat(sample)); (alpha <- 0.05)
14 [1] 276
15 [1] 7.8
16 [1] 7.856089
17 [1] 0.05
18 > nullsample <- sample - (x_bar - mu0) # Mau du lieu tuong ung voi H0
19 > # Check lai mean cua nullsample co bang mu0 chua
20 > mean(nullsample, na.rm = TRUE)
21 [1] 7.8
22
23 > # Lay phan phoi cua randomization
24 > rand_dist <- randomization(10000)
25 > # Tinh p-value trong kiem dinh hai phia (two-tailed)
26 > (p_value <- mean(abs(rand_dist - mu0) >= abs(x_bar - mu0)))
27 [1] 0.5422
28 > # Tinh gia tri toi han (critical value) voi muc phan vi 1-alpha/2
29 > (crit_val <- quantile(rand_dist, 1 - alpha/2, names = FALSE))
30 [1] 7.976758
31 > # Kiem tra xem p_value co be hon alpha, neu co thi bac bo H0
32 > p_value < alpha; abs(crit_val - mu0) < abs(x_bar - mu0)
33 [1] FALSE
34 [1] FALSE
35

```

Vì $p\text{-value} > \alpha$ nên ta không bác bỏ H_0 . Tương tự, ta có giá trị tới hạn (critical value) $\text{crit_val} > \bar{x}$ nên ta không bác bỏ H_0 .

Như vậy, với mức ý nghĩa 5%, kì vọng mức độ chân thành của sinh viên nữ là 7.8.

b) Kiểm định cho trung vị (median)

Giả sử cùng tổng thể và mẫu dữ liệu ở câu a) nhưng ta kiểm định cho trung vị mức độ chân thành của sinh viên nữ trong trường với mức ý nghĩa (significance level) 5% thay vì trung bình của mức độ chân thành. Mặc dù trung bình thường được sử dụng như là con số mô tả trọng tâm của phân phối nhưng nó lại rất nhạy cảm với ngoại lệ (outlier).

Gọi med là median mức độ chân thành của nữ trong trường, \hat{med} là median mức độ chân thành trong mẫu dữ liệu. Ta có

$$\hat{med} = 8.000$$

Ta có thể thấy các ngoại lệ qua boxplot sau đây:

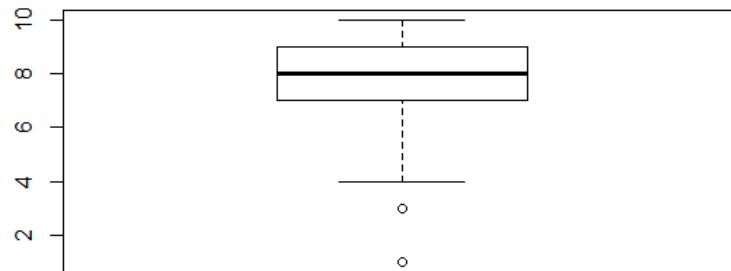


Figure 3: Boxplot của LikeM

Ta thấy rằng có một số outliers dưới 4 điểm. Trong những trường hợp như thế này ta có thể dùng trung vị là một thống kê ít bị ảnh hưởng bởi ngoại lệ.

Để tính đến sự biến động của \hat{med} theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : med = med_0 = 8.0 \\ H_1 : med \neq 8.0 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

3

```

1 # TK can tinh
2 stat <- function(data){
3   return (median(data, na.rm = TRUE)) # trung vi
4 }
5
6 > # tham so mac dinh, trung vi mau, muc y nghia
7 > (med0 <- 8.0); (med_hat <- stat(sample)); (alpha <- 0.05)
8 [1] 8
9 [1] 8
10 [1] 0.05
11 > nullsample <- sample - (med_hat - med0) # Mau du lieu tuong ung voi H0

```

³Hàm randomization vẫn y như ở câu a)

```

12 > stat(nullsample)
13 [1] 8
14
15 > # Layphanphoi cuarandomization
16 > rand_dist <- randomization(10000)
17 > # Tinh p-value trong kiem dinh hai phia (two-tailed)
18 > (p_value <- mean(abs(rand_dist - med0) >= abs(med_hat - med0)))
19 [1] 1
20 > # Tinh gia tri toi han (critical value) voi muc phan vi 1-alpha/2
21 > (crit_val <- quantile(rand_dist, 1 - alpha/2, names = FALSE))
22 [1] 8
23 > # Kiem tra xem p_value co be hon alpha, neu co thi bac bo H0
24 > p_value < alpha; abs(crit_val - med0) < abs(med_hat - med0)
25 [1] FALSE
26 [1] FALSE
27

```

Vì $p\text{-value} > \alpha$ nên ta không bác bỏ H_0 . Tương tự, ta có giá trị tới hạn (critical value) $\text{crit_val} > \hat{med}$ nên ta không bác bỏ H_0 .

Như vậy, với mức ý nghĩa 5%, trung vị mức độ chân thành của sinh viên nữ trong trường là 8.0.

2 Khảo sát cặp biến

2.1 Biến định tính vs biến định tính

Chọn 2 biến định tính: DecisionMale (Yes/No) và RaceF (Asian, Black, Caucasian, Latino, Other)

Khảo sát 2 biến định tính DecisionMale và RaceF

```

1 # 2 biến định tính
2 tab1 = table(DecisionMale, RaceF)
3 # Them margin
4 addmargins(tab1)
5 >
6 RaceF
7 DecisionMale Asian Black Caucasian Latino Other Sum
8 No 2 32 7 72 7 10 130
9 Yes 2 38 8 76 16 6 146
10 Sum 4 70 15 148 23 16 276
11
12 # 2-way table
13 # Ti le chung toc nu (Asian, Black, ...) nhan phan hoi
14 prop.table(tab1, margin = 1)
15 >
16 RaceF
17 DecisionMale Asian Black Caucasian Latino Other
18 No 0.01538462 0.24615385 0.05384615 0.55384615 0.05384615 0.07692308
19 Yes 0.01369863 0.26027397 0.05479452 0.52054795 0.10958904 0.04109589
20
21 barplot(tab1, legend = TRUE)
22

```

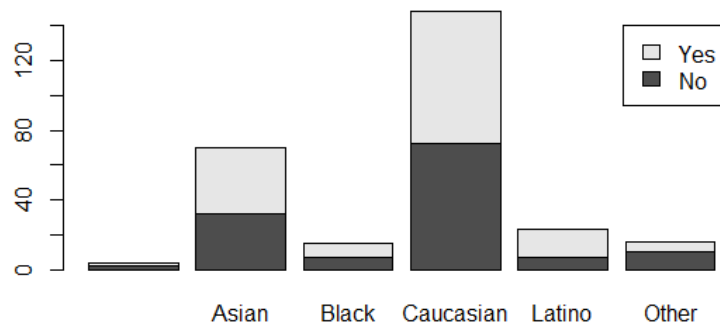


Figure 4: Segmented barchart của DecisionMale và RaceF

2.1.1 Kiểm định cho tỉ lệ khác biệt giữa nữ da trắng nhận phản hồi "Yes" và "No"

Giả sử, ta cần khảo sát tỉ lệ khác biệt giữa nữ da trắng được nam phản hồi (Yes/No) cho quần thể (population) là toàn bộ sinh viên nữ của trường Columbia bằng cách gom nhóm các dân tộc nữ khác còn lại thành 1 cụm.

Ta chỉ phân tích giữa tỉ lệ nữ da trắng và nhóm còn lại. Từ tổng thể, ta thu thập được một mẫu dữ liệu ngẫu nhiên (random sample) gồm 276 quan sát trong đó có 148 nữ da trắng (gồm 76 phản hồi "Yes", 72 phản hồi "No") và 128 dân tộc khác (gồm 70 phản hồi "Yes" và 58 phản hồi "No").

Dựa vào mẫu dữ liệu này, ta kiểm định nghi vấn "Tỉ lệ nữ da trắng nhận phản hồi Yes nhiều hơn phản hồi No" với mức ý nghĩa (significance level) 5%.

Gọi p_{Yes} là tỉ lệ nữ da trắng nhận phản hồi "Yes" và p_{No} là tỉ lệ nữ da trắng nhận phản hồi "No" trong trường. Từ đây, suy ra tỉ lệ khác biệt giữa phản hồi Yes và phản hồi No trong trường là Δp :

$$\Delta p = p_{Yes} - p_{No}$$

Gọi \hat{p}_{Yes} là tỉ lệ nữ da trắng nhận phản hồi "Yes" và \hat{p}_{No} là tỉ lệ nữ da trắng nhận phản hồi "No" trong mẫu dữ liệu. Từ đây, suy ra tỉ lệ khác biệt giữa phản hồi Yes và phản hồi No trong mẫu dữ liệu là $\Delta \hat{p}$:

$$\Delta \hat{p} = \hat{p}_{Yes} - \hat{p}_{No} = 0.01449275$$

Để tính đến sự biến động của $\Delta \hat{p}$ theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : \Delta p = \Delta p_0 = 0 \\ H_1 : \Delta p > 0 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

```
1 stat <- function(data){
2   return (mean(data$DecisionMale == 'Yes' & data$RaceF == 'Caucasian') - mean(data$DecisionMale == '
   No' & data$RaceF == 'Caucasian')) # Tỉ lệ khác biệt
```

```

3 }
4
5 randomization <- function(B){
6   return (replicate(B, stat(sample(nullsample, n, replace = TRUE))))
7 }
8
9 > # Mau du lieu ban dau
10 > sample <- data.frame(DecisionMale, RaceF)
11 > # tham so mac dinh, ti le mau, muc y nghia
12 > (p0 <- 0.5); (p_hat <- stat(sample)); (alpha <- 0.05)
13 [1] 0.5
14 [1] 0.01449275
15 [1] 0.05
16 > # Kich thuoc mau
17 > (n <- nrow(sample))
18 [1] 276
19 > nullsample <- data.frame("DecisionMale" = c(rep("Yes", n/2), rep("No", n/2)), "RaceF" = rep("Caucasian", n)) # Mau du lieu tuong ung voi H0
20
21 > # Layphan phoi cua randomization
22 > rand_dist <- randomization(10000)
23 > # Tinh p-value trong kiem dinh mot phia (one-tailed)
24 > (p_value <- mean(rand_dist >= p_hat))
25 [1] 0.4285
26 > # Tinh gia tri toi han (critical value) voi muc phan vi 1-alpha
27 > (crit_val <- quantile(rand_dist, 1 - alpha, names = FALSE))
28 [1] 0.1014493
29 > # Kiem tra xem p_value co be hon alpha, neu co thi bac bo H0
30 > p_value < alpha; crit_val < p_hat
31 [1] FALSE
32 [1] FALSE
33

```

Vì $p\text{-value} > \alpha$ nên ta không bác bỏ H_0 . Tương tự, ta có giá trị tới hạn (critical value) $\text{crit_val} > \Delta\hat{p}$ nên ta không bác bỏ H_0 .

Như vậy, với mức ý nghĩa 5%, tỉ lệ nữ da trắng nhận phản hồi "Yes" bằng với phản hồi "No".

2.1.2 Kiểm định cho tỉ lệ khác biệt giữa nữ châu Á nhận phản hồi "Yes"/"No"

Giả sử, ta cần khảo sát tỉ lệ khác biệt giữa nữ châu Á được nam phản hồi (Yes/No) cho quần thể (population) là toàn bộ sinh viên nữ của trường Columbia bằng cách gom nhóm các dân tộc nữ khác còn lại thành 1 cụm.

Ta chỉ phân tích giữa tỉ lệ nữ châu Á và nhóm còn lại. Từ tổng thể, ta thu thập được một mẫu dữ liệu ngẫu nhiên (random sample) gồm 276 quan sát trong đó có 70 nữ châu Á (gồm 32 phản hồi "Yes", 32 phản hồi "No") và 206 dân tộc khác (gồm 108 phản hồi "Yes" và 98 phản hồi "No").

Dựa vào mẫu dữ liệu này, ta kiểm định nghi vấn "Tỉ lệ nữ châu Á nhận phản hồi Yes bằng với phản hồi No" với mức ý nghĩa (significance level) 5%.

Gọi p_{Yes} là tỉ lệ nữ châu Á nhận phản hồi "Yes" và p_{No} là tỉ lệ nữ châu Á nhận phản hồi "No" trong trường. Từ đây, suy ra tỉ lệ khác biệt giữa phản hồi Yes và phản hồi No trong trường

là Δp :

$$\Delta p = p_{Yes} - p_{No}$$

Gọi \hat{p}_{Yes} là tỉ lệ nữ châu Á nhận phản hồi "Yes" và \hat{p}_{No} là tỉ lệ nữ châu Á nhận phản hồi "No" trong mẫu dữ liệu. Từ đây, suy ra tỉ lệ khác biệt giữa phản hồi Yes và phản hồi No trong mẫu dữ liệu là $\Delta \hat{p}$:

$$\Delta \hat{p} = \hat{p}_{Yes} - \hat{p}_{No} = 0.02173913$$

Để tính đến sự biến động của $\Delta \hat{p}$ theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : \Delta p = \Delta p_0 = 0 \\ H_1 : \Delta p \neq 0 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

```
1 # TK can tinh
2 stat <- function(data){
3   return (mean(data$DecisionMale == 'Yes' & data$RaceF == 'Asian') - mean(data$DecisionMale == 'No'
4     & data$RaceF == 'Asian')) # Ti le khac biet
5 }
6
7 randomization <- function(B){
8   return (replicate(B, stat(sample(nullsample, n, replace = TRUE))))
9 }
10
11 > # tham so mac dinh, ti le mau, muc y nghia
12 > (p0 <- 0); (p_hat <- stat(sample)); (alpha <- 0.05)
13 [1] 0
14 [1] 0.02173913
15 [1] 0.05
16 > # Kich thuc mau
17 > (n <- nrow(sample))
18 [1] 276
19 > nullsample <- data.frame("DecisionMale" = c(rep("Yes", n/2), rep("No", n/2)), "RaceF" = rep("Asian", n
20   )) # Mau du lieu tuong ung voi H0
21
22 > # Lay phan phoi cua randomization
23 > rand_dist <- randomization(10000)
24 > # Tinh p-value trong kiem dinh mot phia (one-tailed)
25 > (p_value <- mean(abs(rand_dist - p0) >= abs(p_hat - p0)))
26 [1] 0.7658
27 > # Tinh gia tri toi han (critical value) voi muc phan vi 1-alpha/2
28 > (crit_val <- quantile(rand_dist, 1 - alpha/2, names = FALSE))
29 [1] 0.115942
30 > # Kiem tra xem p_value co be hon alpha, neu co thi bac bo H0
31 > p_value < alpha; crit_val < p_hat
32 [1] FALSE
[1] FALSE
```

Vì $p\text{-value} > \alpha$ nên ta không bác bỏ H_0 . Tương tự, ta có giá trị tới hạn (critical value) $\text{crit_val} > \Delta \hat{p}$ nên ta không bác bỏ H_0 .

Như vậy, với mức ý nghĩa 5%, tỉ lệ nữ châu Á nhận phản hồi "Yes" bằng với phản hồi "No".

2.2 Biến định tính và biến định lượng

Chọn 1 biến định tính và 1 biến định lượng: DecisionMale (yes/no), AttractiveM (1-10)

```
1 # Tính favorite statistics
2 > favstats(AttractiveM ~ DecisionMale)
3 DecisionMale min Q1 median Q3 max mean sd n missing
4 1 No 1 5 5 6 10 5.641732 1.694877 127 3
5 2 Yes 5 7 8 8 10 7.595890 1.357375 146 0
6
7 # Vẽ boxplot
8 boxplot(AttractiveM ~ DecisionMale, xlab = "DecisionMale", ylab = "AttractiveM")
9
```

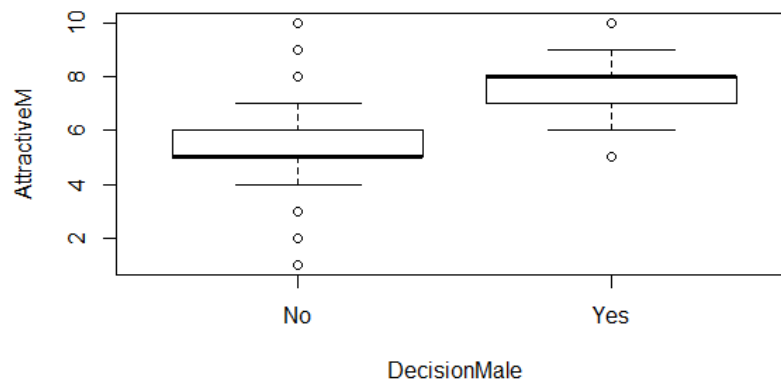


Figure 5: Side-by-side boxplots

2.2.1 Kiểm định cho độ chênh lệch kì vọng giữa mức độ hấp dẫn của nữ trong phản hồi "Yes" và mức độ hấp dẫn của nữ trong phản hồi "No"

Giả sử, ta cần khảo sát kì vọng chênh lệch giữa mức độ hấp dẫn của nữ trong phản hồi "Yes" và trong phản hồi "No" cho quần thể (population) là toàn bộ sinh viên nữ của trường Columbia.

Từ tổng thể, ta thu thập được một mẫu dữ liệu ngẫu nhiên (random sample) gồm 276 quan sát. Dựa vào mẫu dữ liệu này, ta kiểm định nghi vấn "Kì vọng mức độ hấp dẫn của nữ trong phản hồi Yes sẽ cao hơn trong phản hồi No" với mức ý nghĩa (significance level) 5%.

Gọi $\mu_{AttractiveM|Yes}$ là kì vọng mức độ hấp dẫn của nữ nhận phản hồi "Yes" và $\mu_{AttractiveM|No}$ là kì vọng mức độ hấp dẫn của nữ nhận phản hồi "No" trong trường. Từ đây, suy ra độ chênh lệch kì vọng giữa phản hồi Yes và phản hồi No trong trường là $\Delta\mu$:

$$\Delta\mu = \mu_{AttractiveM|Yes} - \mu_{AttractiveM|No}$$

Gọi $\bar{x}_{AttractiveM|Yes}$ là mức độ hấp dẫn trung bình của nữ nhận phản hồi "Yes" và $\bar{x}_{AttractiveM|No}$ là mức độ hấp dẫn trung bình của nữ nhận phản hồi "No" trong mẫu dữ liệu. Từ đây, suy ra độ chênh lệch kì vọng giữa phản hồi Yes và phản hồi No trong mẫu dữ liệu là $\Delta\bar{x}$:

$$\Delta\bar{x} = \bar{x}_{AttractiveM|Yes} - \bar{x}_{AttractiveM|No} = 1.954158$$

Để tính đến sự biến động của $\Delta\bar{x}$ theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : \Delta\mu = \Delta\mu_0 = 0 \\ H_1 : \Delta\mu > 0 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

```

1 # TK can tinh
2 diffmean <- function(data1, data2) {
3   # Lay index
4   index <- 1:(n1+n2) %in% sample(1:(n1+n2), n1)
5   # random sample cua yes
6   rand_sample1 <- c(data1, data2)[index]
7   # random sample cua no
8   rand_sample2 <- c(data1, data2)[!index]
9   return(mean(rand_sample1, na.rm = TRUE) - mean(rand_sample2, na.rm = TRUE))
10 }
11
12 randomization <- function(B){
13   return (replicate(B, diffmean(sample1, sample2)))
14 }
15
16 > # Sample
17 > sample1 <- subset(SpeedDating, DecisionMale=='Yes', select=c(AttractiveM))[[1]];
18 > sample2 <- subset(SpeedDating, DecisionMale=='No', select=c(AttractiveM))[[1]];
19 > # Kich thuoc mau
20 > n1 <- length(sample1); n2 <- length(sample2); n1; n2
21 [1] 146
22 [1] 130
23
24 > # tb mau 1, tb mau 2 va diff mean cua mau 1 va 2
25 > (x_1 <- mean(sample1, na.rm = TRUE)); (x_2 <- mean(sample2, na.rm = TRUE)); (diff_x <- x_1 - x_2)
26 [1] 7.59589
27 [1] 5.641732
28 [1] 1.954158
29 > # tham so mac dinh, muc y nghia
30 > (diff_mu0 <- 0); (alpha <- 0.05)
31 [1] 0
32 [1] 0.05
33
34 > # Lay phan phoi cua randomization
35 > rand_dist <- randomization(10000); hist(rand_dist)
36 > # Tinh p-value trong kiem dinh mot phia (one-tailed)
37 > (p_value <- mean(rand_dist >= diff_x))
38 [1] 0
39 > # Tinh gia tri toi han (critical value) voi muc phan vi 1-alpha
40 > (crit_val <- quantile(rand_dist, 1 - alpha, names = FALSE))
41 [1] 0.3517241
42 > # Kiem tra xem p_value co be hon alpha, neu co thi bac bo H0
43 > p_value < alpha; crit_val < diff_x
44 [1] TRUE
45 [1] TRUE
46

```

Vì $p\text{-value} < \alpha$ nên ta bác bỏ H_0 , chấp nhận H_1 . Tương tự, ta có giá trị tới hạn (critical value) $\text{crit_val} < \Delta\bar{x}$ nên ta bác bỏ H_0 , chấp nhận H_1 .

Như vậy, với mức ý nghĩa 5%, kì vọng mức độ quyền rũ trong phản hồi "Yes" cao hơn phản hồi "No".

2.2.2 Kiểm định cho độ chênh lệch trung vị (median) giữa mức độ hấp dẫn của nữ trong phản hồi "Yes" và mức độ hấp dẫn của nữ trong phản hồi "No"

Giả sử cùng tổng thể và mẫu dữ liệu ở câu trên nhưng ta muốn xây dựng khoảng tin cậy 95% cho trung vị (median) thay vì trung bình của mức độ hấp dẫn trong phản hồi "Yes"/"No". Mặc dù trung bình thường được sử dụng như là con số mô tả trọng tâm của phân phối nhưng nó lại rất nhạy cảm với ngoại lệ (outlier).

Ta có thể thấy trong 5 ở mỗi phản hồi "Yes" và "No" đều có các outlier xuất hiện đặc biệt nhất là ở phản hồi "No", có những điểm số bất thường như 8, 9, 10 vẫn nằm trong phản hồi "No".

Gọi $med_{AttractiveM|Yes}$ là median mức độ hấp của quyền rũ trong phản hồi "Yes" và $med_{AttractiveM|No}$ là median mức độ quyền rũ của nữ trong phản hồi "No" của trường. Từ đây, suy ra độ chênh lệch trung vị giữa phản hồi Yes và phản hồi No trong trường là Δmed :

$$\Delta med = med_{AttractiveM|Yes} - med_{AttractiveM|No}$$

Gọi $\hat{med}_{AttractiveM|Yes}$ là median mức độ hấp dẫn của nữ trong phản hồi "Yes" và $\hat{med}_{AttractiveM|No}$ là median mức độ hấp dẫn của nữ trong phản hồi "No" của mẫu dữ liệu. Từ đây, suy ra độ chênh lệch trung vị giữa phản hồi Yes và phản hồi No trong mẫu dữ liệu là $\Delta \hat{med}$:

$$\Delta \hat{med} = \hat{med}_{AttractiveM|Yes} - \hat{med}_{AttractiveM|No} = 3$$

Để tính đến sự biến động của $\Delta \hat{med}$ theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : \Delta med = \Delta med_0 = 0 \\ H_1 : \Delta med > 0 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

Ta có thể thấy rằng, median kháng nhiễu tốt hơn so với mean dựa vào số liệu thống kê trên.

```

1 # TK can tinh
2 diffmed <- function(data1, data2) {
3   # Lay index
4   index <- 1:(n1+n2) %in% sample(1:(n1+n2), n1)
5   # random sample cua yes
6   rand_sample1 <- c(data1, data2)[index];
7   # random sample cua no
8   rand_sample2 <- c(data1, data2)[!index];
9   return(median(rand_sample1, na.rm = TRUE) - median(rand_sample2, na.rm = TRUE))
10 }
11
12 randomization <- function(B){
13   return (replicate(B, diffmed(sample1, sample2)))
14 }
15
16 > # Sample

```

```

17 > sample1 <- subset(SpeedDating, DecisionMale=='Yes', select=c(AttractiveM))[[1]];
18 > sample2 <- subset(SpeedDating, DecisionMale=='No', select=c(AttractiveM))[[1]];
19 > # Kich thuoc mau
20 > n1 <- length(sample1); n2 <- length(sample2); n1; n2
21 [1] 146
22 [1] 130
23
24 > # med mau 1, med mau 2 va diff med của mau 1 va 2
25 > (med_1 <- median(sample1, na.rm = TRUE)); (med_2 <- median(sample2, na.rm = TRUE)); (diff_med <-
    med_1 - med_2)
26 [1] 8
27 [1] 5
28 [1] 3
29 > # tham so mac dinh, muc y nghia
30 > (diff_med0 <- 0); (alpha <- 0.05)
31 [1] 0
32 [1] 0.05
33
34 > # Lay phan phoi của randomization
35 > rand_dist <- randomization(10000); hist(rand_dist)
36 > # Tinh p-value trong kiem dinh mot phia (one-tailed)
37 > (p_value <- mean(rand_dist >= diff_med))
38 [1] 0
39 > # Tinh gia tri toi han (critical value) voi muc phan vi 1-alpha
40 > (crit_val <- quantile(rand_dist, 1 - alpha, names = FALSE))
41 [1] 1
42 > # Kiem tra xem p_value có bé hơn alpha, nếu có thì bác bỏ H0
43 > p_value < alpha; crit_val < diff_med
44 [1] TRUE
45 [1] TRUE
46

```

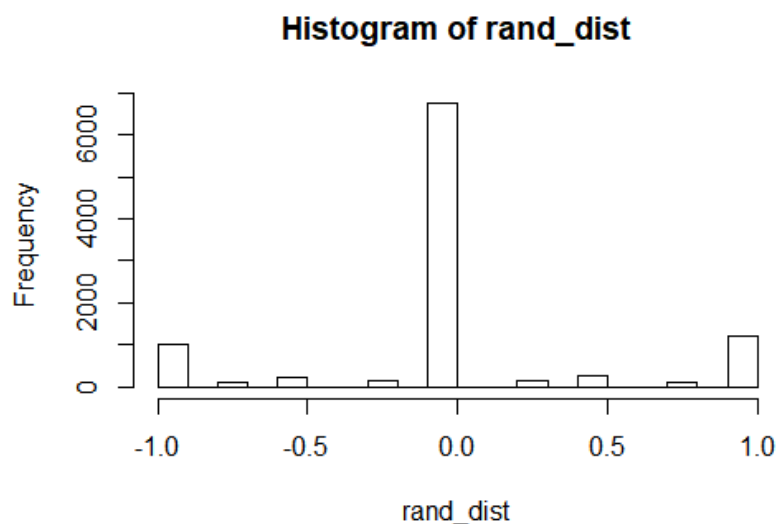


Figure 6: Histogram của randomization distribution

Vì $p\text{-value} < \alpha$ nên ta bác bỏ H_0 , chấp nhận H_1 . Tương tự, ta có giá trị tới hạn (critical value) $\text{crit_val} < \Delta\hat{\text{med}}$ nên ta bác bỏ H_0 , chấp nhận H_1 .

Như vậy, với mức ý nghĩa 5%, trung vị mức độ quyền rũ trong phản hồi "Yes" cao hơn phản hồi "No".

2.3 Biến định lượng và biến định lượng

Chọn 2 biến định lượng: AttractiveM (1-10) và LikeM (1-10)

```
1 # Correlation of 2 quantative variables: AttractiveM and LikeM
2 > cor(AttractiveM, LikeM, use = "complete.obs") # Avoid missing values
3 [1] 0.7240187
4
5 # Fit regression line
6 lmInfo <- lm(LikeM~AttractiveM)
7 > summary(lmInfo) # get more info
8 Call:
9 lm(formula = LikeM ~ AttractiveM)
10
11 Residuals:
12 Min    1Q  Median    3Q   Max
13 -4.6225 -0.6225  0.0914  0.8054  3.6611
14
15 Coefficients:
16 Estimate Std. Error t value Pr(>|t|)
17 (Intercept)  1.91100    0.28616   6.678 1.37e-10 ***
18 AttractiveM  0.71394    0.04132  17.279 < 2e-16 ***
19 ---
20 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
21
22 Residual standard error: 1.232 on 271 degrees of freedom
23 (3 observations deleted due to missingness)
24 Multiple R-squared:  0.5242, Adjusted R-squared:  0.5224
25 F-statistic: 298.6 on 1 and 271 DF, p-value: < 2.2e-16
26
27 # Graphical display: scatterplot
28 plot(AttractiveM, LikeM, main = "Scatter plot example", pch=19)
29 # Add fit lines
30 abline(lm(LikeM~AttractiveM), col="red") # regression line (y~x)
31
32 plot(lmInfo$residuals, pch = 16, col = "red") #Plot residual de xem du lieu co phan bo ngau nhieu khong?
33
```

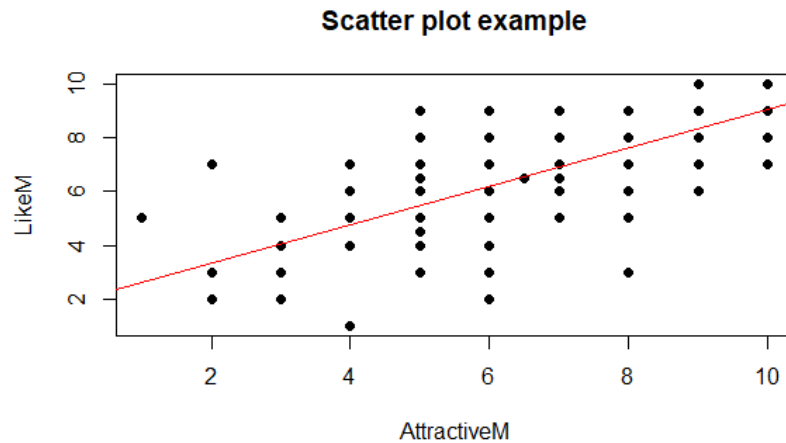


Figure 7: Scatterplot của 2 biến định lượng và có linear regression line

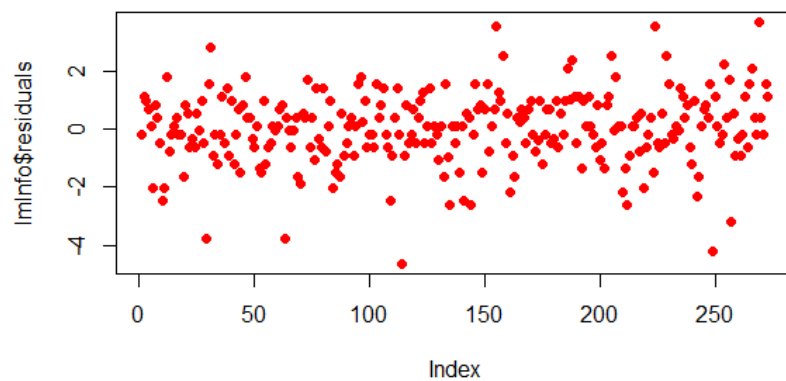


Figure 8: Residuals plot

Nhận xét:

- Nhìn vào Residuals, ta thấy rằng độ lệch giữa giá trị dự đoán và giá trị quan sát vẫn còn chênh lệch khá nhiều.
- Tiếp theo, để đánh giá model này có tốt hay không thì ta cần nhìn vào $R^2 = 0.5242$ thì ta thấy nó gần 0.5. Điều này có thể tạm chấp nhận là model này khá tốt.
- Nhưng đến đây ta chưa thể vội kết luận rằng model này tốt. Do đó, ta cần plot residuals để xem phân bố của chúng có ngẫu nhiên không. Nếu không ngẫu nhiên mà có thể là có 1 hidden pattern mà model chưa xét tới. Điều này sẽ ảnh hưởng đến khả năng dự đoán khi mà dữ liệu tăng.
- Nhìn vào hình 8, ta đã có thể yên tâm kết luận rằng model này là tốt vì các residuals phân bố ngẫu nhiên (không có hidden pattern như: curve,...)

2.3.1 Kiểm định cho hệ số tương quan giữa AttractiveM và LikeM

Giả sử, ta cần khảo sát hệ số tương quan (correlation coefficient) giữa mức độ hấp dẫn và mức độ thích của nam giới đánh giá cho nữ (AttractiveM và LikeM) cho quần thể (population) là toàn bộ sinh viên nữ của trường Columbia.

Từ tổng thể, ta thu thập được một mẫu dữ liệu ngẫu nhiên (random sample) gồm 276 quan sát. Dựa vào mẫu dữ liệu này, ta kiểm định nghi vấn "hệ số tương quan giữa AttractiveM và LikeM lớn hơn 0.5" với mức ý nghĩa (significance level) 5%.

Gọi ρ là hệ số tương quan giữa AttractiveM và LikeM của sinh viên nữ trong trường và r là hệ số tương quan giữa AttractiveM và LikeM của sinh viên nữ trong mẫu dữ liệu. Từ các thống kê tính được bằng R, ta có:

$$r = 0.7240187$$

Để tính đến sự biến động của r theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : \rho = \rho_0 = 0.5 \\ H_1 : \rho > 0.5 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

```
1 library("ecodist") # Generate data.frame with specific correlation
2
3 #TK can tinh
4 stat <- function(data){
5   #Tinh correlation
6   return (cor(data$AttractiveM, data$LikeM, use = "complete.obs")) # Avoid missing values
7 }
8
9 randomization <- function(B){
10   return (replicate(B, stat(sample(nullsample, n, replace = TRUE))))
11 }
12
13 > # Sample
14 > sample <- data.frame(AttractiveM, LikeM)
15 > # Kich thuoc mau
16 > (n <- nrow(sample))
17 [1] 276
18
19 > # tham so mac dinh, correlation tren mau, muc y nghĩa
20 > (cor0 <- 0.5); (cor_hat <- stat(sample)); (alpha <- 0.05)
21 [1] 0.5
22 [1] 0.7240187
23 [1] 0.05
24 > nullsample <- corgen(len = n, r = 0.5, epsilon = 0.01) # Mau tuong thich voi H0
25 > #rename column
26 > names(nullsample)[1] = "AttractiveM"
27 > names(nullsample)[2] = "LikeM"
28 > stat(nullsample)
29 [1] 0.5089513
30
31 > # Lay phan phoi cua randomization
32 > rand_dist <- randomization(10000)
33 > # Tinh p-value trong kiem dinh mot phia (one-tailed)
```

```

34 > (p_value <- mean(rand_dist >= cor_hat))
35 [1] 0
36 > # Tính giá trị tới hạn (critical value) với mức phân vi 1-alpha
37 > (crit_val <- quantile(rand_dist, 1 - alpha, names = FALSE))
38 [1] 0.5089513
39 > # Kiểm tra xem p_value có bé hơn alpha, nếu có thì bác bỏ H0
40 > p_value < alpha; crit_val < cor_hat
41 [1] TRUE
42 [1] TRUE
43

```

Vì $p\text{-value} < \alpha$ nên ta bác bỏ H_0 , chấp nhận H_1 . Tương tự, ta có giá trị tới hạn (critical value) $\text{crit_val} < r$ nên ta bác bỏ H_0 , chấp nhận H_1 .

Như vậy, với mức ý nghĩa 5%, Hệ số tương quan giữa AttractiveM và LikeM lớn hơn 0.5.

Nhận xét:

- Ta có thể thấy rằng đây là 1 liên kết dương mạnh. (do $\rho > 0.5$)
- Điều này có nghĩa là mức độ hấp dẫn của nữ AttractiveM tăng thì mức độ thích của nam dành cho nữ LikeM cũng tăng.

2.3.2 Kiểm định cho hệ số hồi qui β (regression slope) của regression line

Giả sử, ta cần khảo sát hệ số hồi qui của best-fit line: β (slope) giữa mức độ hấp dẫn và mức độ thích của nam giới đánh giá cho nữ (AttractiveM và LikeM) cho quần thể (population) là toàn bộ sinh viên nữ của trường Columbia.

Từ tổng thể, ta thu thập được một mẫu dữ liệu ngẫu nhiên (random sample) gồm 276 quan sát. Dựa vào mẫu dữ liệu này, ta kiểm định nghi vấn "hệ số hồi qui của regression line giữa AttractiveM và LikeM lớn hơn 0.5" với mức ý nghĩa (significance level) 5%.

Gọi β là regression slope của regression line giữa AttractiveM và LikeM của sinh viên trong trường và b là regression slope của regression line giữa AttractiveM và LikeM của sinh viên trong mẫu dữ liệu. Từ các thống kê tính được bằng R, ta có:

$$b = 0.71394$$

Để tính đến sự biến động của b theo mẫu dữ liệu $n = 276$ thu thập từ tổng thể, ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : \beta = \beta_0 = 0.5 \\ H_1 : \beta > 0.5 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

Ở đây, tôi sẽ dùng phương pháp kiểm định bằng khoảng tin cậy (confident interval) cho β với độ tin cậy là $1 - \alpha = 95\%$ trên $[a, \infty]$:

$$P(\beta < a) = \alpha$$

```

1 # Các TK cần tính
2 stat <- function(data){

```



```

3  #Tìm best-fit line
4  lmInfo <- lm(data$LikeM~data$AttractiveM)
5  return (lmInfo$coefficients[2])
6  }
7  # Bootstrap
8  bootstrap <- function(B){
9    return (replicate(B, stat(sample(sample, nrow(data), replace = TRUE))))
10 }
11
12 > # tham so mac dinh, correlation tren mau, muc y nghia
13 > (slope0 <- 0.5); (slope_hat <- stat(sample)); (alpha <- 0.05)
14 [1] 0.5
15 data$AttractiveM
16 0.7139398
17 [1] 0.05
18 >
19 > boots_dist <- bootstrap(10000) # Tìm phân phối của bootstrap
20 > (se <- sd(boots_dist, na.rm = TRUE)) # Tính standard deviation (missing value sẽ bị bỏ qua)
21 [1] 0.0471093
22 > (conf_boots <- quantile(boots_dist, c(alpha, 1), names = FALSE)) # Tìm khoảng tin cậy cho correlation
23 [1] 0.6365502 0.9118599
24 > # Nếu cor0 nằm ngoài khoảng tin cậy thì ta sẽ bác bỏ H0
25 > !(conf_boots[1] <= slope0 && slope0 <= conf_boots[2])
26 [1] TRUE
27

```

Vì β_0 nằm ngoài khoảng tin cậy (confident interval) với độ tin cậy $1 - \alpha = 95\%$ nên ta bác bỏ H_0 và chấp nhận H_1

Như vậy, với mức ý nghĩa 5%, Hệ số hồi qui β của regression line giữa AttractiveM và LikeM lớn hơn 0.5.

References

- [1] Hoang, Vu Quoc and An, Le Huong Thao. LAB 06 - KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ. PDF.