Báo cáo bài tập 7

1612174 - Phùng Tiến Hào - tienhaophung@gmail.com

12/05/2019

Contents

Dữ liệu khảo sát: SpeedDating trong package Lock5withR

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

```
require(Lock5withR) # Load package
library(Lock5withR)
library(mosaic)

# head(SpeedDating)
attach(SpeedDating) # Avoid dollar sign before each varibles name

```

1 Chi-square Goodness of Fit Test for single catogorical variable

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 và phản hồi No bằng nhau" với mức ý nghĩa (significance level) 5%.

Tỉ lệ kỳ vọng (Expected proportion) là 0.5 cho các phản hồi "Yes" và "No".

Ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : p_{Yes} = p_{No} = 0.5 \\ H_1 : p_{Yes} \neq p_{No} \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

Lưu ý: Để dùng Chi-square để kiểm định thì các ô quan sát phải ít nhất 5.

```
> # Load data
> data <- DecisionMale
> # Significant level
> alpha <- 0.05
> # Frequency table
> t <- table(data); t
data
No Yes
130 146
> # Chisq test
> res <- chisq.test(t); res

Chi-squared test for given probabilities

data:  t
X-squared = 0.92754, df = 1, p-value = 0.3355

> # If p.value < alpha, we ignore H0
> (res$p.value < alpha)
[1] FALSE
> # Expected values
> res$expected
```

```
No Yes
138 138

```

Vì $p - value > \alpha$ nên ta không bác bỏ H_0. Như vậy, với mức ý nghĩa 5%, ta không có đủ căn cứ để bác bỏ "tỉ lệ phản hồi "Yes" bằng "No"".

Nhận xét:

- Ta thấy rằng expected values và observed values chênh lệch không nhiều, do đó không đóng góp nhiều vào $\mathcal{X}^2$.

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 6 dân tộc: 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 các dân tộc phân bố không đều nhau" với mức ý nghĩa (significance level) 5%.

Gọi p_i là tỉ lệ dân tộc nữ (Caucasian, Asian,..., Other) trong trường $\hat{p}_i$ là tỉ lệ dân tộc nữ trong mẫu dữ liệu. Với $i = 1, 2, \ldots, 6$

Ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0 : p_i = 0.1666667 \\ H_1 : p_i \neq 0.1666667 \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

Lưu ý: Để dùng Chi-square để kiểm định thì các ô quan sát phải ít nhất 5.

```
> # Load data
> data <- RaceF
> # Significant level
> alpha <- 0.05
>
> # Frequency table
> t <- table(data); t
data
Asian    Black Caucasian   Latino    Other
4       70       15      148       23       16
> t.prob <- prop.table(t); t.prob
data
Asian     Black  Caucasian     Latino     Other
0.01449275 0.25362319 0.05434783 0.53623188 0.08333333 0.05797101
>
> # Chisq test
> res <- chisq.test(t, p = rep(1/6, 6)); res

Chi-squared test for given probabilities

data:  t
X-squared = 329, df = 5, p-value < 2.2e-16

```

```
> # If p.value < alpha, we ignore H0
> (res$p.value < alpha)
[1] TRUE
> # Expected values
> res$expected
Asian    Black Caucasian   Latino    Other
46       46       46       46       46       46
```

Vì $p - value < \alpha$ nên ta bác bỏ H_0 và chấp nhận H_1. Như vậy, với mức ý nghĩa 5%, ta chấp nhận "tỉ lệ các dân tộc nữ phân bố không đều".

Nhận xét:

- Ta thấy rằng giữa observed values và expected values chênh lệch khá nhiều. Đặc biệt là Caucasian, Asian và Null đóng góp nhiều vào $\mathcal{X}^2$.
- Vì thế nên khả năng để giả thuyết xảy ra là rất thấp, cụ thể ta có $p - value < 2.2e - 16$ rất bé.

2 Chi-Square Test of Independence (significant association) for two catogorical variables

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

Khảo sát 2 biến định tính DecisionMale và RaceF

```
# 2 bien dinh tinh
tab1 = table(DecisionMale, RaceF)
# Them margin
addmargins(tab1)
>
 RaceF
DecisionMale     Asian Black Caucasian Latino Other Sum
No    2    32     7       72      7    10 130
Yes   2    38     8       76     16     6 146
Sum   4    70    15      148     23    16 276

# 2-way table
# Ti le chung toc nu (Asian, Black, ...) nhan phan hoi
prop.table(tab1, margin = 1)
>
           RaceF
DecisionMale              Asian      Black  Caucasian     Latino      Other
No  0.01538462 0.24615385 0.05384615 0.55384615 0.05384615 0.07692308
Yes 0.01369863 0.26027397 0.05479452 0.52054795 0.10958904 0.04109589

barplot(tab1, legend = TRUE)
```

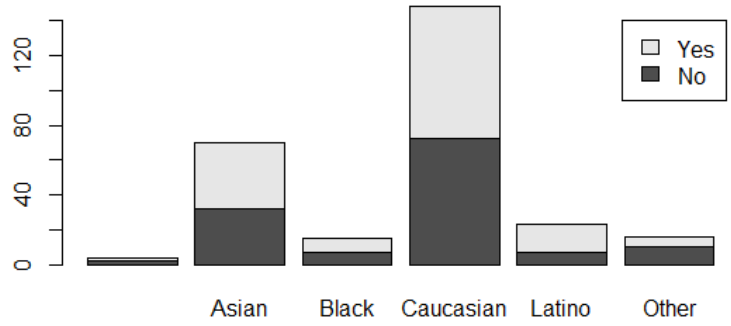

Figure 1: Segmented barchart của DecisionMale và RaceF

Giả sử, ta cần khảo sát sự liên kết giữa dân tộc nữ và sự phản hồi (Yes/No) của nam cho quần thể (population). 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 "Giữa RaceF và DecisionMale có mỗi liên kết với nhau" với mức ý nghĩa (significance level) 5%.

Ta thực hiện kiểm định giả thuyết:

$$\begin{cases} H_0: & \text{RaceF và DecisionMale độc lập nhau} \\ H_1: & \text{RaceF và DecisionMale có sự liên kết với nhau} \end{cases}$$

với mức ý nghĩa $\alpha = 0.5\%$

Trực quan hóa bảng tần suất (2-way frequency table)

```
> # Load data
> data <- data.frame(DecisionMale, RaceF)
> # Significant level
> alpha <- 0.05
>
> # Frequency table
> t <- table(data); t
RaceF
DecisionMale   Asian Black Caucasian Latino Other
No   2    32    7       72     7    10
Yes  2    38    8       76    16     6
> # P(RaceF|DecisionMale)
> t.prob <- prop.table(t, margin = 1); t.prob
RaceF
DecisionMale             Asian     Black  Caucasian     Latino      Other
No  0.01538462 0.24615385 0.05384615 0.55384615 0.05384615 0.07692308
Yes 0.01369863 0.26027397 0.05479452 0.52054795 0.10958904 0.04109589

> library("graphics")
> # shade: color graph
> # las = 1: horizontal labels
> mosaicplot(t(t), shade = TRUE, las = 1, main = "data")
```

```
23
```

1

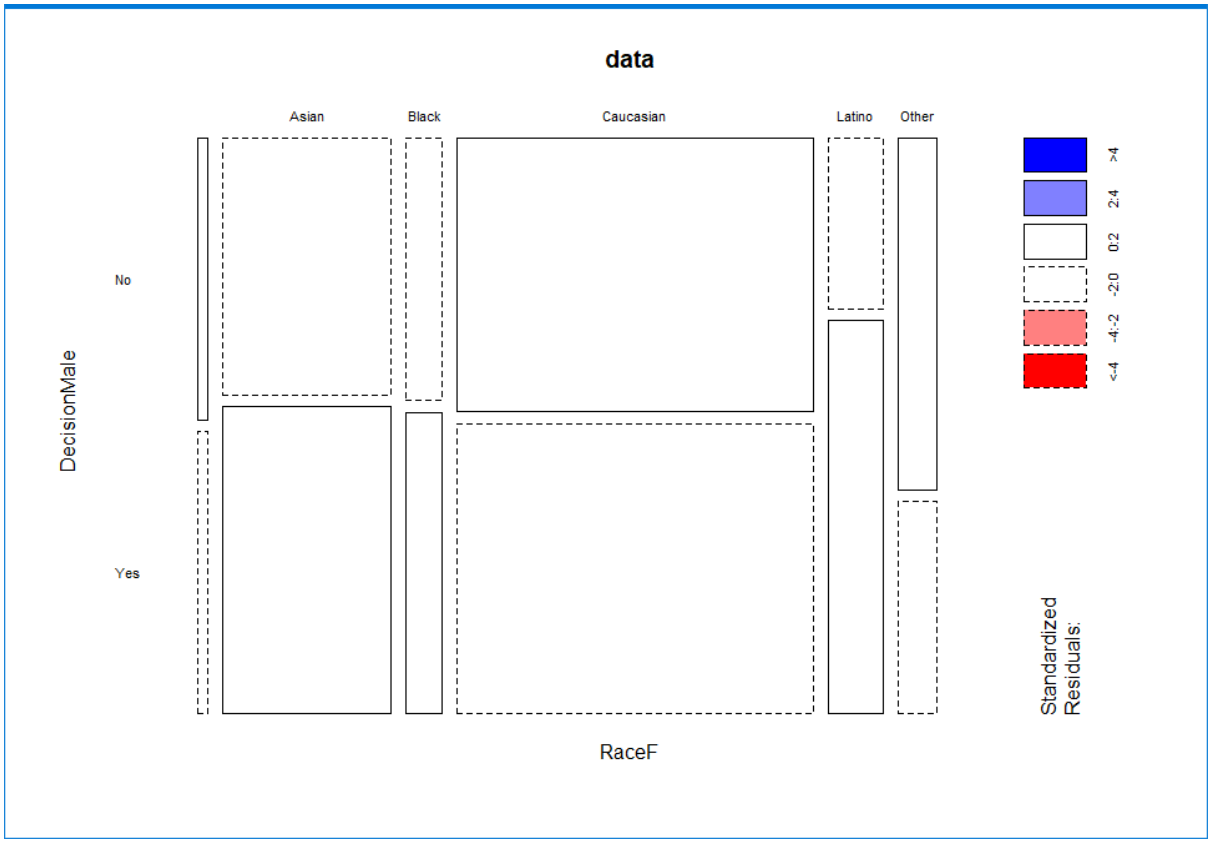

Figure 2: Màu đỏ biểu thị observed values bé hơn expected values
Màu xanh biểu thị observed values lớn hơn expected values
(với điều kiện dữ liệu phải là ngẫu nhiên)

Nhận xét

- Nhìn vào mosaicplot thì ta thấy rằng bảng dữ liệu của chúng ta thấy rằng sự chênh lệch giữa observed values và expected values rất bé. Một phần nào cho ta thấy được giữa RaceF và DecisionMale hầu như không có liên kết.
- Các ô (cell) màu trắng nét liền biểu thị độ lệch dương và màu trắng nét đứt biểu thị độ lệch âm nhưng ta thấy rằng các độ lệch này rất bé.

Với mỗi ô thì ta có thể tính được expected value tương ứng:

$$e = \frac{row.sum * col.sum}{grand.total}$$

Chi-square statistic tính như sau:

$$\mathcal{X}^2 = \sum \frac{(o-e)^2}{e}$$

[1]mosaicplot() là hàm built-in của R package graphics

với $\begin{cases} \text{o: observed value} \\ \text{e: expected value} \end{cases}$

Tính Chi-square statistic trong R

```
> # Chisq test
> res <- chisq.test(t); res
Warning message:
In chisq.test(t) : Chi-squared approximation may be incorrect

Pearson Chi-squared test

data:  t
X-squared = 4.2977, df = 5, p-value = 0.5074

> # If p.value < alpha, we ignore H0
> (res$p.value < alpha)
[1] FALSE

> # Observed values
> res$observed
RaceF
DecisionMale    Asian Black Caucasian Latino Other
No   2    32    7       72     7    10
Yes  2    38    8       76    16     6
> # Expected values
> round(res$expected, 2)
RaceF
DecisionMale     Asian Black Caucasian Latino Other
No  1.88 32.97  7.07    69.71  10.83  7.54
Yes 2.12 37.03  7.93    78.29  12.17  8.46

```

Vì $p - value > \alpha$ nên ta không bác bỏ H_0. Như vậy, với mức ý nghĩa 5%, ta không có đủ căn cứ để bác bỏ "RaceF và DecisionMale độc lập hay nói cách khác cả hai không có sự liên kết".

Nhìn vào observed values table và expected values table, ta thấy được rằng chênh lệch ở đây rất ít.

Để biết rõ về bản chất của sự phụ thuộc giữa 2 biến RaceF và DecisionMale, ta sẽ tiếp tục tính dư lượng chuẩn hóa (Standardized residuals hoặc Pearson residuals) cho từng ô để biết được ô nào đóng góp nhiều vào Chi-square $\mathcal{X}^2$:

$$r = \frac{o - e}{\sqrt{e}}$$

Pearson residuals được lấy từ kết quả của chisq.test():

```
> # Pearson Residuals: Do lech giua observed values and expected values
> round(res$residuals, 3)
RaceF
DecisionMale     Asian  Black Caucasian Latino  Other
No   0.084 -0.169 -0.025     0.274 -1.165  0.897
Yes -0.080  0.160  0.023    -0.259  1.099 -0.847

> # Visualize Pearson residuals
> library(corrplot)
corrplot 0.84 loaded
```

```
Warning message:
package ‘’corrplot was built under R version 3.5.3
> corrplot(res$residuals, is.cor = FALSE)

```

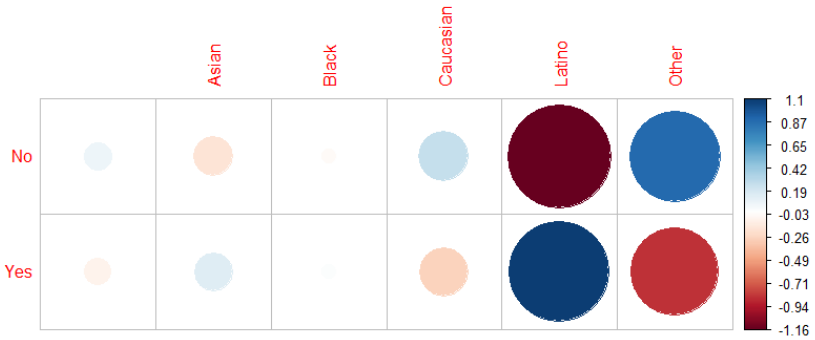

Figure 3: Correlation plot

Chú thích:

- Kích thước của hình tròn là tỉ lệ thuận với mức độ đóng góp của ô đó.
- Hệ số tương quan ở đây khác với của 2 biến định lượng là không có dao động trong miền [-1, 1]

Nhận xét:

- Màu đỏ biểu thị mối liên kết âm (negative association). Như ta thấy, hàng No và cột Lattino; hàng Yes và các cột Other, Caucasian có mối liên kết âm (negative association) - nghĩa là khi cái này tăng thì cái kia giảm. Ngụ ý là sự không thích (repulsion) , ta có thể thấy dân tộc Latino là nhận nhiều phản hồi No nhất (tức là bị từ chối) và ngược lại, dân tộc Other và Caucasian lại nhận liên kết âm với phản hồi Yes.
- Ngược lại, màu xanh biểu thị mối liên kết dương (positive association) - nghĩa là cả hai đều cùng tăng. Như ta thấy, hàng Yes và cột Latino; hàng No và cột Other và Caucasian có liên kết dương. Ngụ ý là sự thu hút (attraction). Điều đó cho thấy người Latino có xu hướng nhận được phản hồi Yes cao. Tương tự, hàng No, cột Other và Caucasian có liên kết dương mạnh, có thể hiểu đơn giản là khi số lượng dân tộc Other và Caucasian tăng thì khả năng họ nhận được phản No (bị từ chối) cũng tăng theo.

Bây giờ, để biết được được mức độ phần trăm đóng góp của các ô cho Chi-square $\mathcal{X}^2$, ta tính theo công thức sau:

$$contrib = \frac{r^2}{\mathcal{X}^2}$$

```
> # Contribution (Percentage %) of given cell to total chi−squre
> contrib <− 100∗res$residuals^2 / res$statistic
> round(contrib, 3)
RaceF
DecisionMale      Asian  Black Caucasian Latino  Other
No   0.166  0.665  0.014    1.750 31.561 18.742
Yes  0.148  0.592  0.012    1.558 28.102 16.688
```

```
> # Visualiza contribution
> corrplot(contrib, is.cor = FALSE)
```

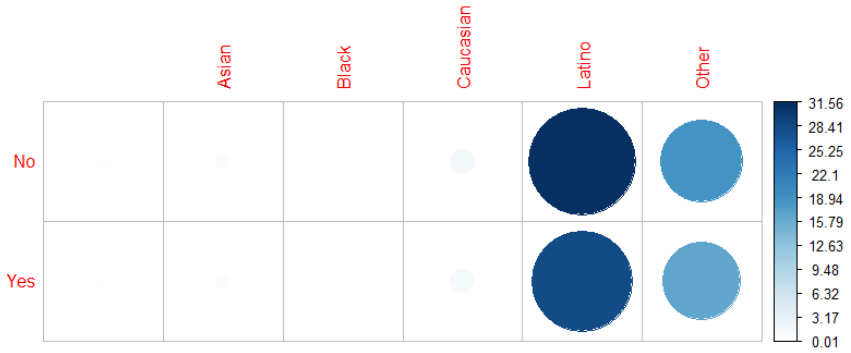

Figure 4: Contribution in percentage (%) plot

Sự đóng góp tương đối của mỗi ô vào tổng Chi bình phương cho thấy một số dấu hiệu về bản chất của sự phụ thuộc giữa các hàng và cột của bảng tần suất.

Từ ảnh trên, ta kết luận được:

- Hàng ”No” có liên kết mạnh với cột Latino và Other
- Hàng ”Yes” cũng có liên kết mạnh với Latino và Other
- Từ bảng trên tính bằng R, ta thấy các ô đóng góp nhiều cho Chi-squre là No/Latino (31.561%), No/Other (18.742%), Yes/Latino (28.102%) và Yes/Other (16.688%).
- Tổng cộng 4 ô trên đóng góp tới tận 95.048% vào tổng Chi bình phương và vì vậy chúng chiếm phần lớn sự khác biệt giữa các giá trị kì vọng và giá trị quan sát.

References

[1] Randall Pruim and Lana Park. Lock5WithR. Chapter 7: Chi-Squared Tests for Categorical Variables. PDF.

[2] R Users Guide. Chapter 7: Chi-Squared Tests for Categorical Variables. PDF.

[3] Chi-Square Test of Independence in R. (n.d.). Retrieved from http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r

[4] Chi-square Goodness of Fit Test in R. (n.d.). Retrieved from http://www.sthda.com/english/wiki/chi-square-goodness-of-fit-test-in-r