TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TPHCM

Khoa Công nghệ thông tin

THỐNG KÊ MÁY TÍNH VÀ ỨNG DỤNG

BÀI TẬP 4

Xây dựng khoảng tin cậy - Kiểm định giả thuyết

(Confidence interval - Hypotheses testing)

TP.HCM – 05/2019

Bài tập 3

Nội dung

1) Xây dựng khoảng tin cậy: 3
- ❖ DecisionFemale 3
- ❖ RaceF 4
- ❖ LikeF 5
- ❖ FunF 6
- ❖ AmbitiousF 7
- ❖ DecisionFemale và LikeF 8
- ❖ LikeF và AmbitiousF 9

2) Kiểm định giả thuyết 10
- a) DecisionFemale: 10
- b) RaceF: 12
- c) LikeF: 13
- d) FunF: 14
- e) AmbitiousF: 15
- f) DecisionFemale và RaceF: 15
- g) DecisionFemale và LikeF: 15
- h) LikeF và AmbitiousF: 15

Bài tập 3

1) Xây dựng khoảng tin cậy:

- ❖ DecisionFemale

```
> #DecisionFemale
> # Lọc từ dataset SpeedDating lấy DecisionFemale
> sampleDF <- subset(SpeedDating, select=c(DecisionFemale)); #sampleDF
>
> # Lấy kích thước mẫu
> n <- nrow(sampleDF); n
[1] 276
>
> #Tính tỷ lệ trên mẫu
> p_hat_No <- sum(sampleDF == 'No')/n; p_hat_No
[1] 0.5398551
>
> #.....................................Bootstrap.............................................
> #Tạo phân phối bootsrap rồi tính độ lệch chuuẩn vẽ biểu đồ hist
> #replace=TRUE: Lấy có hoàn lại
> boot_distDF <- replicate(10000, sum((sample(sampleDF, n, replace = TRUE)) == 'No')/n)
> sd(boot_distDF); hist(boot_distDF, breaks = 40)
[1] 0.02996381
>
> #Xây dựng khoảng tin cậy 95%
> anpha <- 1 - 0.95;
> quantile(boot_distDF, c(anpha/2, 1-anpha/2))
     2.5%     97.5%
0.4818841 0.5978261
```

➜ Ta có khoảng tin cậy tỷ lệ không muốn có cuộc hẹn tiếp theo là [0.48, 0.59] cho tỷ lệ

Bài tập 3

❖ RaceF

```
> #RaceF
> # Lọc từ dataset SpeedDating lấy RaceF
> sampleRF <- subset(SpeedDating, select=c(RaceF)); #sampleRF
>
> # Lấy kích thước mẫu
> n <- nrow(sampleRF); n
[1] 276
>
> #Tính tỷ lệ trên mẫu
> p_hat_Cauc <- sum(sampleRF == 'Caucasian')/n; p_hat_Cauc
[1] 0.5362319
>
> ########################################Bootstrap#################################################
> #Tạo phân phối bootsrap rồi tính độ lệch chuuẩn vẽ biểu đồ hist
> #replace=TRUE: Lấy có hoàn lại
> boot_distRF <- replicate(10000, sum((sample(sampleRF, n, replace = TRUE)) == 'Caucasian')/n)
> sd(boot_distRF); hist(boot_distRF, breaks = 40)
[1] 0.02970211
>
> #Xây dựng khoảng tin cậy 95%
> anpha <- 1 - 0.95;
> quantile(boot_distRF, c(anpha/2, 1-anpha/2))
     2.5%     97.5%
0.4782609 0.5942029
>
> ########################################CongThuc#################################################
> #Tính Sai số chuẩn
> seRF <- sqrt(p_hat_Cauc*(1 - p_hat_Cauc)/n); seRF
[1] 0.03001734
>
> #Tính phân phối chuẩn
> z <- qnorm(1 - anpha/2)
>
> #Xác định khoảng tin cậy 95%
> p_hat_Cauc + c(-z*seRF, z*seRF)
[1] 0.4773990 0.5950648
```

➜ Ta có khoảng tin cậy về tỷ lệ chủng tộc da trắng là [0.48, 0.59]

Bài tập 3

- ❖ LikeF

```
> #LikeF
> # Lọc từ dataset SpeedDating lấy LikeF
> sampleLF <- subset(SpeedDating, LikeF!="NA"  , select=c(LikeF))[[1]]; sampleLF
  [1]  7.0  7.0  6.0  7.0  5.0  6.0  6.0  6.0  7.0  8.0  7.0  8.0  8.0  3.0  6.0  7.0  8.0  7.0  2.0  7.0  7.0  6.0  6.0  6.0  6.0  8.0
 [27]  5.0  8.0  3.0  8.0  8.0  8.0  3.0  6.0  8.0  6.0  5.0  6.0  5.0  6.0  8.0  7.0  4.0  9.0  5.0  2.0  7.0  9.0  4.0  7.0  9.0  7.0
 [53]  7.0  8.0  5.0  2.0  5.0  7.0  9.0  7.0  9.0  7.0  7.0  5.0  8.0  7.0  4.0  7.0  8.0  8.0  8.0  8.0 10.0  9.0  7.0  5.0  5.0  8.0
 [79]  4.0  7.0  7.0  3.0  7.0  6.0  3.0  7.0  7.0  6.0  8.0  8.0  4.0  8.0  6.0  4.0  8.0  8.0  6.0  6.0  2.0  6.0  5.0  8.0  7.0  9.0
[105]  8.0  5.0  7.0  7.0  7.0  6.0  8.0  7.0  6.0  7.0  7.0  8.0  8.0  4.0  5.0  5.0  8.5  6.0  5.0  7.0  8.0  5.0  8.0  5.0  4.0  8.0
[131]  7.0  7.0  7.0  8.0  5.0  3.0  5.0  7.0  7.0  5.0  7.0  7.0  6.0  6.0  7.0  7.0  6.0  3.0  7.0  9.0  7.0  6.0  7.0  4.0  6.0  3.0
[157]  7.0  6.0  6.0  5.0  7.0  8.0 10.0  8.0  5.0  6.0  6.0  8.0  7.0  9.0  8.0  7.0  8.0  8.0  7.0  7.0  5.0  2.0  8.0 10.0  7.0  7.0
[183]  6.0  9.0  6.0  6.0  7.0  4.0  5.0  6.0  6.0  8.0 10.0  8.0  7.0  7.0  7.0  7.0  7.0  7.0  2.0  7.0  6.0  6.0  3.0  2.0  6.0  8.0
[209]  6.0  6.0 10.0  8.0  9.0 10.0  8.0  7.0  5.0 10.0  5.0  5.0  3.0  5.0  5.0  9.0  5.0  1.0  5.0  7.0  5.0  6.0  6.0  3.0  5.0  5.0
[235]  6.0  7.0  7.0  7.0  6.0  5.0  7.0 10.0  6.0  8.0  6.0  6.0  5.0  3.0  6.0  7.0  9.0  6.0  4.0  7.0  8.0  4.0  3.0  8.0  8.0  6.0
[261]  5.0  5.0  5.0  7.0  3.0  5.0  6.0  6.0  5.0  6.0  4.0  7.0
>
> # Lấy kích thước mẫu
> n <- length(sampleLF); n
[1] 272
>
> #Tính trung bình mẫu
> x_barLF <- mean(sampleLF); x_barLF
[1] 6.365809
>
> #Khoảng tin cậy 95%
> anpha <- 1 - 0.95; anpha
[1] 0.05

> #.....................................Bootstrap.............................................
> #Tạo phân phối bootsrap rồi tính độ lệch chuuẩn vẽ biểu đồ hist
> #replace=TRUE: Lấy có hoàn lại
> #?replicate
> boot_distLF <- replicate(10000, mean(sample(sampleLF, n, replace = TRUE)))
> sd(boot_distLF); hist(boot_distLF, breaks = 40)
[1] 0.1061237
>
> #Xây dựng khoảng tin cậy 95%
> #?quantile
> quantile(boot_distLF, c(anpha/2, 1-anpha/2), na.rm = FALSE)
    2.5%    97.5%
6.154412 6.573529
>
> #.....................................CongThuc.............................................
> #Tính Sai số chuẩn
> seLF <-sd(sampleLF)/sqrt(n); seLF
[1] 0.106423
>
> #Tính phân phối chuẩn z
> z <- qnorm(1-anpha/2);
>
> #Tính phân phối student t
> t <- qt(1-anpha/2, df = n-1)
>
> #Xác định khoảng tin cậy 95%
> x_barLF + c(-z*seLF, z*seLF)  #Dựa trên z
[1] 6.157224 6.574394
>
> x_barLF + c(-t*seLF, t*seLF)  #Dựa trên t
[1] 6.156288 6.575330
```

➜ Ta có khoảng tin cậy điểm trung bình về mức độ thích là [6.16, 6.57]

Bài tập 3

- FunF

```
> #FunF
> # Lọc từ dataset SpeedDating lấy FunF
> sampleFF <- subset(SpeedDating, FunF!="NA"  , select=c(FunF))[[1]]; sampleFF
  [1]  2  4  4  6  8  4  4  5  9  7  7  5  7  2  7  6  7  2  7  7  4  5  6  6  8  8  9  2  9  8  8  3  8  9  5  6  5  4  8  6  6  5  8
 [44]  6  6 10  9  7  5  6  8  7 10  8  5  6  7  8  8  8  8  5  9  8  4  8  8  7  8  8 10  8  5  6  5  7  3  6  7  6  7  6  2  5  7  5
 [87]  9  8  6  8  9  8  7  7  4  7  4  6  4  7  9 10  9  4  9  7 10  6  6  6  8  9  9  8  6  5  7  8  5  5  7 10  7  8  4  6  8  6  6
[130]  6  9  5  6  9  6  9  7  3  7  3  5  7  8  7  6  6 10  7  5  8  6  6  6  7  6  8  4  6  8 10  8  5  7  7  8  7  9  7  8 10  8  7
[173]  7  6  5  7 10  7  8  5  9  6  7  7  8  7  6  8  8 10  9  6  8  8  5  7  7  4  8  4  5  4  3  5  7  5  5 10 10  9 10  9  3  3 10
[216]  8  5  2  8  3  9  6  1  7  8  4  3  6  5  7  4  6  5  8  7  4  7  5 10 10  6  8  7  7  2  3  8  6  4  4  4  7  9  6  6 10 10  7
[259]  6  7  7  5  4  7  5  5  6  6  3  7
>
> # Lấy kích thước mẫu
> n <- length(sampleFF); n
[1] 270
>
> #Tính trung bình mẫu
> x_barFF <- mean(sampleFF); x_barFF
[1] 6.562963
>
> #Khoảng tin cậy 95%
> anpha <- 1 - 0.95; anpha
[1] 0.05

> #.....................................Bootstrap.............................................
> #Tạo phân phối bootsrap rồi tính độ lệch chuẩn vẽ biểu đồ hist
> #replace=TRUE: Lấy có hoàn lại
> #?replicate
> boot_distFF <- replicate(10000, mean(sample(sampleFF, n, replace = TRUE)))
> sd(boot_distFF); hist(boot_distFF, breaks = 40)
[1] 0.1190829
>
> #Xây dựng khoảng tin cậy 95%
> #?quantile
> quantile(boot_distFF, c(anpha/2, 1-anpha/2), na.rm = FALSE)
    2.5%    97.5%
6.325926 6.792593
>
> #.....................................CongThuc..............................................
> #Tính Sai số chuẩn
> seFF <-sd(sampleFF)/sqrt(n); seFF
[1] 0.1195723
>
> #Tính phân phối chuẩn z
> z <- qnorm(1-anpha/2);
>
> #Tính phân phối student t
> t <- qt(1-anpha/2, df = n-1)
>
> #Xác định khoảng tin cậy 95%
> x_barFF + c(-z*seFF, z*seFF)  #Dựa trên z
[1] 6.328606 6.797320
>
> x_barFF + c(-t*seFF, t*seFF)  #Dựa trên t
[1] 6.327546 6.798379
```

→ Ta có khoảng tin cậy điểm trung về mức độ hài hước là [6.3, 6.8]

Bài tập 3

- ❖ AmbitiousF

```
> #AmbitiousF
> # Lọc từ dataset SpeedDating lấy AmbitiousF
> sampleAF <- subset(SpeedDating, AmbitiousF!="NA"  , select=c(AmbitiousF))[[1]]; sampleAF
  [1]  2  9  3  5  5  6  6  9  8  9  7  9  7  6  7  7  8  7  2  8 10  6  5  7  9  9  8  8  9  7  8  8  5  8 10  8  9  5  7  7  5  9  8
 [44]  5 10  8 10 10  5 10 10  8  7  5  9  9  5  5  7 10 10  6  8 10  7  6  8  9  9  9  6  8  8  5  4  7  8  7  7  8  5  6  7  4  5  8
 [87]  9  8  5  9  8  8  8 10  9  8  5 10  8  8  8  9  7  8  6  8  8 10  7  5 10 10  6  9 10  8  8  6  8  6  7  5  7  4  6  6  9  7  7
[130] 10  5  8  9  6  9  7  6  7  5  5  7  8  5  9  6 10  7 10  9  8  6  5  5  6  8  7  7  8  8  8  9  6  8  9  6  7  6  6  8  8  6  9
[173]  6  7 10  6  7  9  8  9  8  8  9  8  9  8 10  8  8  6  9 10  7  9 10  9  8  5  8  8  5  9  7  5 10  8 10 10  9  3  9  9  8  9 10
[216]  8 10  9  7  5  9  8  1  3  8  4 10  7  7  5  7  8  7  7  9 10  7 10  7  9  8  8  7  7  9  9  6  6  5  5  8 10  7  8  7 10  6  5
[259]  5  7  7  5  8  8  5  7
>
> # Lấy kích thước mẫu
> n <- length(sampleAF); n
[1] 266
>
> #Tính trung bình mẫu
> x_barAF <- mean(sampleAF); x_barAF
[1] 7.428571
>
> #Khoảng tin cậy 95%
> anpha <- 1 - 0.95; anpha
[1] 0.05

> #.....................................Bootstrap.............................................
> #Tạo phân phối bootsrap rồi tính độ lệch chuuẩn vẽ biểu đồ hist
> #replace=TRUE: Lấy có hoàn lại
> #?replicate
> boot_distAF <- replicate(10000, mean(sample(sampleAF, n, replace = TRUE)))
> sd(boot_distAF); hist(boot_distAF, breaks = 40)
[1] 0.1080211
>
> #Xây dựng khoảng tin cậy 95%
> #?quantile
> quantile(boot_distAF, c(anpha/2, 1-anpha/2), na.rm = FALSE)
    2.5%    97.5%
7.214286 7.631673
>
> #.....................................CongThuc..............................................
> #Tính Sai số chuẩn
> seAF <-sd(sampleAF)/sqrt(n); seAF
[1] 0.1087166
>
> #Tính phân phối chuẩn z
> z <- qnorm(1-anpha/2);
>
> #Tính phân phối student t
> t <- qt(1-anpha/2, df = n-1)
>
> #Xác định khoảng tin cậy 95%
> x_barAF + c(-z*seAF, z*seAF)  #Dựa trên z
[1] 7.215491 7.641652
>
> x_barAF + c(-t*seAF, t*seAF)  #Dựa trên t
[1] 7.214513 7.642630
```

➜ Ta có khoảng tin cậy điểm trung bình về mức độ tham vọng là [7.2, 7.6]

Bài tập 3

- ❖ DecisionFemale và LikeF

```
> #DecisionFemale - LikeF
> #Giữ giá trị cố định khi giả lập dữ liệu
> set.seed(400)
>
> #Mức ý nghĩa anpha = 5% (Khoảng tin cậy 95%)
> anpha <- 1 - 0.95
>
> #Vẽ biểu đồ phân bố
> boxplot(LikeF ~ DecisionFemale, data=SpeedDating)
> bwplot(DecisionFemale ~ LikeF, data = SpeedDating)
>
> #Lọc từ dataset SpeedDating lấy LikeF của DecisionFemale = "Yes"
> sampleDFLF_Y <- subset(SpeedDating, DecisionFemale=='Yes'& LikeF!="NA", select=c(LikeF))[[1]]; sampleDFLF_Y
  [1]  7.0  5.0  6.0  7.0  7.0  6.0  7.0  7.0  7.0  6.0  8.0  8.0  8.0  8.0  8.0  6.0  8.0  7.0  7.0  4.0  9.0  7.0  7.0  8.0  5.0  7.0
 [27]  9.0  9.0  7.0  8.0  8.0  8.0  8.0 10.0  9.0  8.0  7.0  7.0  7.0  6.0  8.0  8.0  6.0  8.0  8.0  6.0  6.0  9.0  8.0  7.0  7.0  6.0
 [53]  7.0  8.0  5.0  8.5  7.0  8.0  8.0  7.0  7.0  7.0  7.0  7.0  6.0  7.0  9.0  7.0  6.0  7.0  3.0  7.0  6.0  6.0 10.0  8.0  6.0  8.0
 [79]  7.0  9.0  8.0  7.0  8.0  7.0  7.0  8.0 10.0  7.0  7.0  6.0  9.0  6.0  6.0  7.0  6.0  6.0  8.0 10.0  7.0  7.0  7.0  7.0  6.0 10.0
[105]  9.0 10.0  8.0 10.0  9.0  7.0  7.0  7.0  7.0 10.0  6.0  9.0  7.0  8.0  8.0  8.0  6.0  7.0  5.0  6.0  7.0
>
> #Lọc từ dataset SpeedDating lấy LikeF của DecisionFemale = "No"
> sampleDFLF_N <- subset(SpeedDating, DecisionFemale=='No'& LikeF!="NA", select=c(LikeF))[[1]]; sampleDFLF_N
  [1] 7 7 6 6 6 8 8 8 3 8 7 2 6 6 6 8 5 3 3 6 5 6 5 6 4 9 5 2 9 7 2 5 7 7 5 7 4 7 8 7 5 5 4 7 3 6 3 7 8 4 4 2 6 5 8 7 5 7 7 6 8 7 8 4 5
 [66] 6 5 5 5 4 8 7 8 5 3 5 5 7 6 7 6 3 7 4 6 5 7 8 5 6 8 5 2 4 5 8 7 7 7 2 6 6 3 2 8 6 6 8 7 5 5 5 3 5 5 5 1 5 5 6 6 3 5 5 6 7 6 5 6 8
[131] 6 6 5 3 7 6 4 4 3 5 5 5 3 6 5 6 4
>
> #Lấy kích thước mẫu sampleDFLF_Y và sampleDFLF_N
> n_Y <- length(sampleDFLF_Y); n_N <- length(sampleDFLF_N);  n_Y; n_N
[1] 125
[1] 147
>
> #Tính hiệu trung bình mẫu sampleDFLF_Y và sampleDFLF_N
> x_Y <- mean(sampleDFLF_Y); x_N <- mean(sampleDFLF_N); x_Y - x_N
[1] 1.758177
>

> ####################################Bootstrap#################################################
> #Tạo phân phối bootsrap rồi tính độ lệch chuuẩn vẽ biểu đồ hist
> #replace=TRUE: Lấy có hoàn lại
> boot_distDFLF <- replicate(10000, mean(sample(sampleDFLF_Y, n_Y, replace=TRUE)) - mean(sample(sampleDFLF_N, n_N, replace=TRUE)))
> sd(boot_distDFLF); hist(boot_distDFLF, breaks = 40)
[1] 0.1797466
>
> #Xây dựng khoảng tin cậy 95%
> quantile(boot_distDFLF, c(anpha/2, 1-anpha/2))
    2.5%    97.5%
1.408453 2.118304
>
> ####################################CongThuc#################################################
> #Tính Sai số chuẩn
> seDFLF <- sqrt(var(sampleDFLF_Y)/n_Y + var(sampleDFLF_N)/n_N); seDFLF
[1] 0.1809678
>
> #Tính phân phối chuẩn z
> z <- qnorm(1-anpha/2);
>
> #Tính phân phối student t
> t <- qt(1-anpha/2, df=min(c(n_Y-1, n_N-1)))
>
> #Xác định khoảng tin cậy 95% theo z
> (x_Y - x_N) + c(-z*seDFLF, z*seDFLF);
[1] 1.403487 2.112867
>
> #Xác định khoảng tin cậy 95% theo t
> (x_Y - x_N) + c(-t*seDFLF, t*seDFLF)
[1] 1.399991 2.116363
```

➜ Ta có khoảng tin cậy là [1.4, 2.1]

Bài tập 3

❖ LikeF và AmbitiousF

```
> #LikeF - AmbitiousF
> #Hàm tính tương quan giữa hai biến LikeF và AmbitiousF trên dataset data truyền vào
> cal_r <- function(data) cor(data$LikeF, data$AmbitiousF)
>
> #Giữ giá trị cố định khi giả lập dữ liệu
> set.seed(600)
>
> #Mức ý nghĩa anpha = 5% (Khoảng tin cậy 95%)
> anpha <- 1 - 0.95
>
> #Lọc từ dataset SpeedDating lấy LikeF và AmbitiousF
> sampleAFLF <- subset(SpeedDating, LikeF!="NA" & AmbitiousF!="NA", select=c(LikeF, AmbitiousF));
>
> #Vẽ biểu đồ phân bố
> plot(sampleAFLF$LikeF, sampleAFLF$AmbitiousF)
>
> #Xem số dòng (kích thước) của dataset
> n <- nrow(sampleAFLF); n
[1] 263
>
> #Tính tương quan giữa hai biến
> r <- cal_r(sampleAFLF); r
[1] 0.2789275
>

> ########################################Bootstrap#################################################
> #Tạo phân phối bootsrap rồi tính độ lệch chuuẩn vẽ biểu đồ hist
> #replace=TRUE: Lấy có hoàn lại
> boot_distAFLF <- replicate(10000, cal_r(sampleAFLF[sample(1:n, n, replace=TRUE), ]))
> sd(boot_distAFLF); hist(boot_distAFLF, breaks = 40)
[1] 0.06297957
>
> #Xây dựng khoảng tin cậy 95%
> quantile(boot_distAFLF, c(anpha/2, 1-anpha/2))
     2.5%     97.5%
0.1534314 0.3989089
>
> ########################################CongThuc#################################################
> #Tính Sai số chuẩn
> seAFLF <- sqrt((1-r*r)/(n-2)); seAFLF
[1] 0.05944183
>
> #Tính phân phối chuẩn z
> z <- qnorm(1-anpha/2)
>
> #Tính phân phối student t
> t <- qt(1-anpha/2, df=n-2)
>
> #Xác định khoảng tin cậy 95% theo z
> r + c(-z*seAFLF, z*seAFLF);
[1] 0.1624236 0.3954313
>
> #Xác định khoảng tin cậy 95% theo t
> r + c(-t*seAFLF, t*seAFLF)
[1] 0.1618809 0.3959741
```

➜ Ta có khoảng tin cậy tương quan giữa hai biến định lượng là [0.16, 0.4]

2) Kiểm định giả thuyết

a) DecisionFemale: Tỷ lệ người phụ nữ tham gia khảo sát không muốn một cuộc hẹn khác với đối phương nam trong cuộc hẹn đầu tiên nhiều hơn so với những người muốn

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

```
> #DecisionFemale
> # Lọc từ dataset SpeedDating lấy DecisionFemale
> sampleDF <- subset(SpeedDating, select=c(DecisionFemale)); #sampleDF
>
> n <- nrow(sampleDF)
> p0 <- 0.5
> p_hat_No <- sum(sampleDF == 'No')/n; p_hat_No
[1] 0.5398551
> anpha <- 0.05
> #.....................................Bằng Khoảng Tin Cậy.....................................
> boot_distDF <- replicate(10000, sum((sample(sampleDF, n, replace = TRUE)) == 'No')/n)
> confint <- quantile(boot_distDF, c(anpha/2, 1-anpha/2)); confint
     2.5%     97.5%
0.4818841 0.5978261
> !(confint[1] <= p0 && p0 <= confint[2])
[1] FALSE
>
```

```
> #.....................................Bằng công thức .....................................
> # Tính sai số chuẩn theo công thức p
> (seDF <- sqrt(p0*(1 - p0)/n))
[1] 0.03009646
>
> #Tính phân phối mẫu
> (z <- (p_hat_No - p0)/seDF)
[1] 1.324244
>
> #Tính p_value
> (p_value <- 1 - pnorm(z))
[1] 0.09271095
>
> #Tình giá trị tới hạn z(1-anpha)
> (crit_val <- qnorm(1 - anpha))
[1] 1.644854
>
> # Kiểm tra p_value
> # p_value < anpha => Bác bỏ H0 chấp nhận H1
> p_value < anpha;  #(1)
[1] FALSE
>
> # Kiểm tra giá trị tới hạn z(1-anpha)
> # z(1-anpha) < z => Bác bỏ H0 chấp nhận H1
> crit_val < z      #(2)
[1] FALSE
>
> #.....................................Bằng Hàm .....................................
> ?prop.test
> n <- nrow(sampleDF)
> sum_No <- sum(sampleDF == 'No')
> prop.test(sum_No, n, p = 0.5, conf.level = 1 - anpha, alternative = "greater")

        1-sample proportions test with continuity correction

data:  sum_No out of n
X-squared = 1.5978, df = 1, p-value = 0.1031
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.4885522 1.0000000
sample estimates:
        p
0.5398551
```

Bài tập 3

- Kết luận:

Vì $0.093 > 0.05 \Rightarrow \text{p} - \text{value} > \alpha$ nên ta không bác bỏ H_0. Như vậy, với mức ý nghĩa 5%, Tỷ lệ người phụ nữ tham gia khảo sát **không muốn** một cuộc hẹn khác với đối phương nam trong cuộc hẹn đầu tiên chưa chắc **nhiều hơn** so với những người **muốn**. Ngoài ra $p_0 = 0.5$ cũng nằm trong khoảng tin cậy tính theo boostrap [0.48, 0.597] nên ta không bác bỏ H_0. Kiểm định bằng hàm `prop.test` cũng cho thấy không nên bác bỏ.

Bài tập 3

b) RaceF: Tỷ lệ người phụ nữ tham gia khảo sát chủ yếu thuộc nhóm người da trắng chiếm hơn nửa

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

```
> #RaceF
> # Lọc từ dataset SpeedDating lấy DecisionFemale
> sampleRF <- subset(SpeedDating, select=c(RaceF));  #sampleRF
>
> n <- nrow(sampleRF)
> p0 <- 0.5
> p_hat_Caucasian <- sum(sampleRF == 'Caucasian')/n; p_hat_Caucasian
[1] 0.5362319
> anpha <- 0.05
> #.....................................Bằng Khoảng Tin Cậy.....................................
> boot_distRF <- replicate(10000, sum((sample(sampleRF, n, replace = TRUE)) == 'Caucasian')/n)
> confint <- quantile(boot_distRF, c(anpha/2, 1-anpha/2)); confint
     2.5%     97.5%
0.4782609 0.5942029
> !(confint[1] <= p0 && p0 <= confint[2])
[1] FALSE
```

```
> #.....................................Bằng công thức .....................................
> # Tính sai số chuẩn theo công thức p
> (seRF <- sqrt(p0*(1 - p0)/n))
[1] 0.03009646
>
> #Tính phân phối mẫu
> (z <- (p_hat_Caucasian - p0)/seRF)
[1] 1.203859
>
> #Tính p_value
> (p_value <- 1 - pnorm(z))
[1] 0.1143221
>
> #Tình giá trị tới hạn z(1-anpha)
> (crit_val <- qnorm(1 - anpha))
[1] 1.644854
>
> # Kiểm tra p_value
> # p_value < anpha => Bác bỏ H0 chấp nhận H1
> p_value < anpha;  #(1)
[1] FALSE
>
> # Kiểm tra giá trị tới hạn z(1-anpha)
> # z(1-anpha) < z => Bác bỏ H0 chấp nhận H1
> crit_val < z      #(2)
[1] FALSE
>
> #.....................................Bằng Hàm .....................................
> ?prop.test
> n <- nrow(sampleRF)
> sum_Caucasian <- sum(sampleRF == 'Caucasian')
> prop.test(sum_Caucasian, n, p = 0.5, conf.level = 1 - anpha, alternative = "greater")

        1-sample proportions test with continuity correction

data:  sum_Caucasian out of n
X-squared = 1.308, df = 1, p-value = 0.1264
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.4849385 1.0000000
sample estimates:
        p
0.5362319
```

Bài tập 3

- Kết luận:

Vì $0.11 > 0.05 \Rightarrow \text{p} - \text{value} > \alpha$ nên ta không bác bỏ H_0. Như vậy, với mức ý nghĩa 5%, Tỷ lệ người phụ nữ tham gia khảo sát chủ yếu thuộc nhóm **người da trắng** chiếm hơn nửa là chưa hợp lý.

Ngoài ra $p_0 = 0.5$ cũng nằm trong khoảng tin cậy tính theo boostrap [0.478, 0.594] nên ta không bác bỏ H_0. Kiểm định bằng hàm `prop.test` cũng cho thấy không nên bác bỏ.

c) LikeF: Người phụ nữ tham gia khảo sát chủ yếu cho đối phương nam trung bình 6,366 điểm về mức độ thích

$$\begin{cases} H_0: \mu = \mu_0 = 6.366 \\ H_1: \mu \neq 6.366 \end{cases}$$

```
> #LikeF
> # Lọc từ dataset SpeedDating lấy LikeF
> sampleLF <- subset(SpeedDating, LikeF!="NA"  , select=c(LikeF))[[1]]; sampleLF
  [1]  7.0  7.0  6.0  7.0  5.0  6.0  6.0  6.0  7.0  8.0  7.0  8.0  8.0  3.0  6.0  7.0  8.0  7.0  2.0  7.0  7.0  6.0  6.0  6.0  6.0  8.0
 [27]  5.0  8.0  3.0  8.0  8.0  8.0  3.0  6.0  8.0  6.0  5.0  6.0  5.0  6.0  8.0  7.0  4.0  9.0  5.0  2.0  7.0  9.0  4.0  7.0  9.0  7.0
 [53]  7.0  8.0  5.0  2.0  5.0  7.0  9.0  7.0  9.0  7.0  7.0  5.0  8.0  7.0  4.0  7.0  8.0  8.0  8.0  8.0 10.0  9.0  7.0  5.0  5.0  8.0
 [79]  4.0  7.0  7.0  3.0  7.0  6.0  3.0  7.0  7.0  6.0  8.0  8.0  4.0  8.0  6.0  4.0  8.0  8.0  6.0  6.0  2.0  6.0  5.0  8.0  7.0  9.0
[105]  8.0  5.0  7.0  7.0  7.0  6.0  8.0  7.0  6.0  7.0  7.0  8.0  8.0  4.0  5.0  5.0  8.5  6.0  5.0  7.0  8.0  5.0  8.0  5.0  4.0  8.0
[131]  7.0  7.0  7.0  8.0  5.0  3.0  5.0  7.0  7.0  5.0  7.0  7.0  6.0  6.0  7.0  7.0  6.0  3.0  7.0  9.0  7.0  6.0  7.0  4.0  6.0  3.0
[157]  7.0  6.0  6.0  5.0  7.0  8.0 10.0  8.0  5.0  6.0  6.0  8.0  7.0  9.0  8.0  7.0  8.0  8.0  7.0  7.0  5.0  2.0  8.0 10.0  7.0  7.0
[183]  6.0  9.0  6.0  6.0  7.0  4.0  5.0  6.0  6.0  8.0 10.0  8.0  7.0  7.0  7.0  7.0  7.0  7.0  2.0  7.0  6.0  6.0  3.0  2.0  6.0  8.0
[209]  6.0  6.0 10.0  8.0  9.0 10.0  8.0  7.0  5.0 10.0  5.0  5.0  3.0  5.0  5.0  9.0  5.0  1.0  5.0  7.0  5.0  6.0  6.0  3.0  5.0  5.0
[235]  6.0  7.0  7.0  7.0  6.0  5.0  7.0 10.0  6.0  8.0  6.0  6.0  5.0  3.0  6.0  7.0  9.0  6.0  4.0  7.0  8.0  4.0  3.0  8.0  8.0  6.0
[261]  5.0  5.0  5.0  7.0  3.0  5.0  6.0  6.0  5.0  6.0  4.0  7.0
>
> n <- length(sampleLF)
> mu0 <- 6.366
> anpha <- 0.05
> #.....................................Bằng Khoảng Tin Cậy.....................................
> boot_distLF <- replicate(10000, mean(sample(sampleLF, n, replace = TRUE)))
> confint <- quantile(boot_distLF, c(anpha/2, 1-anpha/2), na.rm = FALSE); confint
    2.5%    97.5%
6.154412 6.569853
>
> !(confint[1] <= mu0 && mu0 <= confint[2])
[1] FALSE
```

```
> #.....................................Bằng công thức .....................................
> #Tính trung bình mẫu
> (x_barLF <- mean(sampleLF))
[1] 6.365809
>
> #Tính sai số chuẩn
> (seLF <- sd(sampleLF)/sqrt(n))
[1] 0.106423
>
> #Tính phân phối student t
> (t <- abs(x_barLF - mu0)/seLF)
[1] 0.001796382
>
> #Tính p_value
> (p_value <- 2*(1 - pt(t, df = n - 1)))
[1] 0.998568
>
> #Tính giá trị tới hạn z(1-anpha)
> (crit_val <- qt(1 - anpha/2, df = n - 1))
[1] 1.968756
>
> # Kiểm tra p_value
> # p_value < anpha => Bác bỏ H0 chấp nhận H1
> p_value < anpha;
[1] FALSE
>
> # Kiểm tra giá trị tới hạn z(1-anpha)
> # z(1-anpha) < z => Bác bỏ H0 chấp nhận H1
> crit_val < t
[1] FALSE
```

- Kết luận:

Vì $0.998 > 0.05 \Rightarrow \text{p} - \text{value} > \alpha$ nên ta không bác bỏ H_0. Như vậy, với mức ý nghĩa 5%, điểm trung bình về mức độ thích người phụ nữ cho đối phương là 6,366. Ngoài ra 6,366 cũng nằm trong khoảng tin cậy [6.16, 6.57].

Bài tập 3

d) FunF: Người phụ nữ tham gia khảo sát chủ yếu cho đối phương nam trung bình 6,56 điểm về mức độ hài hước, vui tính

$$\begin{cases} H_0: \mu = \mu_0 = 6.56 \\ H_1: \mu \neq 6.56 \end{cases}$$

```
> #FunF
> # Lọc từ dataset SpeedDating lấy FunF
> sampleFF <- subset(SpeedDating, FunF!="NA"  , select=c(FunF))[[1]]; sampleFF
  [1]  2  4  4  6  8  4  4  5  9  7  7  5  7  2  7  6  7  2  7  7  4  5  6  6  8  8  9  2  9  8  8  3  8  9  5  6  5  4  8  6  6  5  8
 [44]  6  6 10  9  7  5  6  8  7 10  8  5  6  7  8  8  8  8  5  9  8  4  8  8  7  8  8 10  8  5  6  5  7  3  6  7  6  7  6  2  5  7  5
 [87]  9  8  6  8  9  8  7  7  4  7  4  6  4  7  9 10  9  4  9  7 10  6  6  6  8  9  9  8  6  5  7  8  5  5  7 10  7  8  4  6  8  6  6
[130]  6  9  5  6  9  6  9  7  3  7  3  5  7  8  7  6  6 10  7  5  8  6  6  6  7  6  8  4  6  8 10  8  5  7  7  8  7  9  7  8 10  8  7
[173]  7  6  5  7 10  7  8  5  9  6  7  7  8  7  6  8  8 10  9  6  8  8  5  7  7  4  8  4  5  4  3  5  7  5  5 10 10  9 10  9  3  3 10
[216]  8  5  2  8  3  9  6  1  7  8  4  3  6  5  7  4  6  5  8  7  4  7  5 10 10  6  8  7  7  2  3  8  6  4  4  4  7  9  6  6 10 10  7
[259]  6  7  7  5  4  7  5  5  6  6  3  7
> n <- length(sampleFF)
> mu0 <- 6.56
> anpha <- 0.05
> #.....................................Bằng Khoảng Tin Cậy.....................................
> boot_distFF <- replicate(10000, mean(sample(sampleFF, n, replace = TRUE)))
> confint <- quantile(boot_distFF, c(anpha/2, 1-anpha/2), na.rm = FALSE); confint
    2.5%    97.5%
6.329630 6.792593
>
> !(confint[1] <= mu0 && mu0 <= confint[2])
[1] FALSE
```

```
> #.....................................Bằng công thức .....................................
> #Tính trung bình mẫu
> (x_barFF <- mean(sampleFF))
[1] 6.562963
>
> #Tính sai số chuẩn
> (seFF <- sd(sampleFF)/sqrt(n))
[1] 0.1195723
>
> #Tính phân phối student t
> (t <- abs(x_barFF - mu0)/seFF)
[1] 0.02477968
>
> #Tính p_value
> (p_value <- 2*(1 - pt(t, df = n - 1)))
[1] 0.9802491
>
> #Tính giá trị tới hạn z(1-anpha)
> (crit_val <- qt(1 - anpha/2, df = n - 1))
[1] 1.968822
>
> # Kiểm tra p_value
> # p_value < anpha => Bác bỏ H0 chấp nhận H1
> p_value < anpha;
[1] FALSE
>
> # Kiểm tra giá trị tới hạn z(1-anpha)
> # z(1-anpha) < z => Bác bỏ H0 chấp nhận H1
> crit_val < t
[1] FALSE
```

- Kết luận:

Vì $0.98 > 0.05 \Rightarrow \text{p} - \text{value} > \alpha$ nên ta không bác bỏ H_0. Như vậy, với mức ý nghĩa 5%, điểm trung bình về mức độ hài hước, vui tính người phụ nữ cho đối phương là 6,56. Ngoài ra 6,56 cũng nằm trong khoảng tin cậy [6.3, 6.8].

Bài tập 3

e) AmbitiousF: Người phụ nữ tham gia khảo sát chủ yếu cho đối phương nam trung bình 7,429 điểm mức độ tham vọng

$$\begin{cases} H_0: \mu = \mu_0 = 7{,}429 \\ H_1: \mu \neq 7{,}429 \end{cases}$$

```
> #AmbitiousF
> # Lọc từ dataset SpeedDating lấy AmbitiousF
> sampleAF <- subset(SpeedDating, AmbitiousF!="NA"  , select=c(AmbitiousF))[[1]]; sampleAF
  [1]  2  9  3  5  5  6  6  9  8  9  7  9  7  6  7  7  8  7  2  8 10  6  5  7  9  9  8  8  9  7  8  8  5  8 10  8  9  5  7  7  5  9  8
 [44]  5 10  8 10 10  5 10 10  8  7  5  9  9  5  5  7 10 10  6  8 10  7  6  8  9  9  9  6  8  8  5  4  7  8  7  7  8  5  6  7  4  5  8
 [87]  9  8  5  9  8  8  8 10  9  8  5 10  8  8  8  9  7  8  6  8  8 10  7  5 10 10  6  9 10  8  8  6  8  6  7  5  7  4  6  6  9  7  7
[130] 10  5  8  9  6  9  7  6  7  5  5  7  8  5  9  6 10  7 10  9  8  6  5  5  6  8  7  7  8  8  8  9  6  8  9  6  7  6  6  8  8  6  9
[173]  6  7 10  6  7  9  8  9  8  8  9  8  9  8 10  8  8  6  9 10  7  9 10  9  8  5  8  8  5  9  7  5 10  8 10 10  9  3  9  9  8  9 10
[216]  8 10  9  7  5  9  8  1  3  8  4 10  7  7  5  7  8  7  7  9 10  7 10  7  9  8  8  7  7  9  9  6  6  5  5  8 10  7  8  7 10  6  5
[259]  5  7  7  5  8  8  5  7
>
> n <- length(sampleAF)
> mu0 <- 7.429
> anpha <- 0.05
> #.....................................Bằng Khoảng Tin Cậy.....................................
> boot_distAF <- replicate(10000, mean(sample(sampleAF, n, replace = TRUE)))
> confint <- quantile(boot_distAF, c(anpha/2, 1-anpha/2), na.rm = FALSE); confint
    2.5%    97.5%
7.217951 7.646617
>
> !(confint[1] <= mu0 && mu0 <= confint[2])
[1] FALSE

> #.....................................Bằng công thức .....................................
> #Tính trung bình mẫu
> (x_barAF <- mean(sampleAF))
[1] 7.428571
>
> #Tính sai số chuẩn
> (seAF <- sd(sampleAF)/sqrt(n))
[1] 0.1087166
>
> #Tính phân phối student t
> (t <- abs(x_barAF - mu0)/seAF)
[1] 0.003942096
>
> #Tính p_value
> (p_value <- 2*(1 - pt(t, df = n - 1)))
[1] 0.9968576
>
> #Tính giá trị tới hạn z(1-anpha)
> (crit_val <- qt(1 - anpha/2, df = n - 1))
[1] 1.968956
>
> # Kiểm tra p_value
> # p_value < anpha => Bác bỏ H0 chấp nhận H1
> p_value < anpha;
[1] FALSE
>
> # Kiểm tra giá trị tới hạn z(1-anpha)
> # z(1-anpha) < z => Bác bỏ H0 chấp nhận H1
> crit_val < t
[1] FALSE
```

- Kết luận:

Vì $0.996 > 0.05 \Rightarrow \text{p} - \text{value} > \alpha$ nên ta không bác bỏ H_0. Như vậy, với mức ý nghĩa 5%, điểm trung bình về mức độ hài hước, vui tính người phụ nữ cho đối phương là 7,429. Ngoài ra 7,429 cũng nằm trong khoảng tin cậy [7.2, 7.6].

f) DecisionFemale và RaceF: Người phụ nữ tham gia khảo sát ta thấy sự chênh lệch giữa muốn và không muốn có một cuộc hẹn khác ở các nhóm người là không cao

g) DecisionFemale và LikeF: Người phụ nữ tham gia khảo sát khi cho điểm cao về mức độ thích sẽ muốn có một cuộc hẹn khác

h) LikeF và AmbitiousF: Người phụ nữ tham gia khảo sát có xu hướng thích người đàn ông có tính cầu tiến, có tham vọng