



Bài kiểm tra 4: Kiểm định hai mẫu và hồi quy tuyến tính đơn

Ngày 30 tháng 05 năm 2023

Thời gian làm bài: 45 phút.

Note:

- Sinh viên làm bài trên **R script**, lưu lại với tên có dạng: `"LTTK_HọTên_MSSV_Test4.R"`. Sau khi hoàn thành bài làm, copy phần code bài làm trong  sang **file text .txt** để backup. Nộp bài cả file **R script** và **file text .txt** (lưu với tên có dạng `"LTTK_HọTên_MSSV_Test4.txt"`).
- trong quá trình làm bài kiểm tra, sinh viên có thể tham khảo tài liệu "Giới thiệu về R" (đã được giới thiệu là tài liệu tham khảo môn học).
- Dùng lệnh `help(ten_ham)` để biết cú pháp và cách sử dụng một command trong .
- Bài làm cần trình bày như sau:

```
##
## Bai kiem tra 4 - Thuc hanh Ly thuyet Thong ke
## Nhom2 - Thu ... - tiet ....
##
## Ho ten: ..... - MSSV: .....
##
##*****

## Bai 1:

##-----
## Bai 2:


##-----
```

Bài 1. (4đ)


Trong sản xuất chất bán dẫn, khắc hoá chất ướt thường được sử dụng để loại bỏ silic từ mặt sau của tấm wafer trước khi kim loại hoá. Tỷ lệ ăn mòn là một đặc tính quan trọng trong quá trình này và được biết là tuân theo phân phối chuẩn. Hai phương pháp khắc khác nhau đã được so sánh bằng cách sử dụng hai mẫu ngẫu nhiên gồm 10 tấm mỏng cho mỗi dung dịch. Tỷ lệ ăn mòn quan sát được như sau (đv: mils/phút)

Mẫu 1: 9,9 10,6 9,4 10,3 9,3 10 9,6 10,3 10,2 10,1

Mẫu 2: 10,2 10 10,6 10,2 10,2 10,7 10,7 10,4 10,5 10,3

1.1. Yêu cầu: viết câu mô tả trong  để mô tả giả thuyết cần kiểm định, ví dụ:

```
## H0: mu1 - mu2 ..... D0 = ..... ;
## vs H1: mu1 - mu2 ..... D0 ;
```

1.2. Sử dụng hàm kiểm định giả thuyết thống kê trong , để đưa ra kết luận rằng dữ liệu này có hỗ trợ tuyên bố rằng tỷ lệ ăn mòn trung bình là giống nhau cho cả hai phương pháp hay không?

Sử dụng mức ý nghĩa 5% và giả sử rằng cả hai phương sai bằng nhau.

Bài 2. (4đ)

Một mẫu ngẫu nhiên 500 cư dân trưởng thành của khu vực A chỉ ra rằng 385 người ủng hộ việc tăng giới hạn tốc độ đường cao tốc lên 75 dặm một giờ, và một mẫu khác gồm 400 cư dân của khu vực B đã chỉ ra rằng 267 người ủng hộ giới hạn tăng lên. Những dữ liệu này có cho thấy có sự khác biệt về tỷ lệ ủng hộ tăng giới hạn tốc độ cho cư dân của hai khu vực này hay không? Sử dụng $\alpha = 0,05$.

2.1. Yêu cầu: viết câu mô tả trong  để mô tả giả thuyết cần kiểm định, ví dụ:

```
##      H0: p1 - p2 ..... D0 = ..... ;
## vs H1: p1 - p2 ..... D0 ;
```

Thực hiện phép kiểm định giả thuyết thống kê bằng hai cách

- Cách 1: Viết hàm tên `HypoTest.prop(p1.hat, n, p2.hat, m, D.0, alpha)` để thực hiện phép kiểm định so sánh tỷ lệ p_1 vs p_2 của hai mẫu độc lập, trong đó:


$D.0$ là giá trị sai khác giữa p_1 vs p_2 dùng để so sánh,

$p1.hat$ là tỷ lệ mẫu 1 với cỡ mẫu n ,

$p2.hat$ là tỷ lệ mẫu 2 với cỡ mẫu m

và mức ý nghĩa $\alpha = \alpha$.


Hàm `HypoTest.prop(p1.hat, n, p2.hat, m, D.0, alpha)` sẽ in ra kết luận rằng có bác bỏ H_0 hay không với mức ý nghĩa α .


- Cách 2: sử dụng hàm kiểm định cho tỷ lệ trong .

Hint: cách tạo `table` trong  với dữ liệu đã cho

```
##-----
mydata <- matrix( c( 385, 500-385, 267, 400-267 ), ncol=2, nrow=2, byrow=TRUE)
colnames(mydata) <- c('Agree', 'NotAgree')
rownames(mydata) <- c('A', 'B')
##
mytable <- as.table(mydata)
mytable
##-----
```

2.2 Tính p-giá trị bằng 2 cách:

- Cách 1: sử dụng công thức theo lý thuyết.
- Cách 2: in ra p-giá trị thông qua hàm kiểm định trong .

2.3. Yêu cầu: viết câu mô tả trong  để mô tả kết luận cuối cùng của phép kiểm định, có dạng:

```
##
##      KL: vay du lieu cho thay ..... ve ty le
##      ung ho tang toc do giua hai khu vuc nay .
##
```

Bài 3. (2đ)

Dữ liệu bên dưới mô tả về trọng lượng của các điếu thuốc lá x (g) được sản xuất từ các nhà máy khác nhau và hàm lượng nicotine y (mg) trong mỗi điếu thuốc:

x	15,8	14,9	9,0	4,5	15,0	17,0	8,6	12,0	4,1	16,0
y	0,957	0,886	0,852	0,911	0,889	0,919	0,969	1,118	0,946	1,094

3.1. Vẽ đồ thị phân tán (scatter plot) tương ứng cho dữ liệu này. (sử dụng `col = "dodgerblue"`)

3.2. Xét mô hình hồi quy tuyến tính đơn (Linear regression model) có dạng:

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon,$$


với $\beta_0, \beta_1 \in \mathbb{R}$ là các hệ số của đường thẳng hồi quy, còn ε là đại lượng sai số.

Theo lý thuyết, ước lượng bình phương nhỏ nhất (Least-Squares-Estimators) cho β_0 và β_1 , từ dữ liệu $(x_1, y_1), \dots, (x_n, y_n)$, được tính theo công thức sau:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i \cdot y_i - \frac{(\sum_{i=1}^n x_i) \cdot (\sum_{j=1}^n y_j)}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n \cdot (\bar{x})^2}, \quad \text{và} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}.$$

3.2.a Viết hàm tên `LinearRegress(x,y)` để tính và trả ra kết quả là vector chứa $(\hat{\beta}_0; \hat{\beta}_1)$.

Sử dụng hàm `LinearRegress(x,y)` vừa viết để tính $\hat{\beta}_0$ và $\hat{\beta}_1$ cho bài toán đang xét.

3.2.b Sử dụng hàm `lm` trong  để kiểm tra lại ước lượng $\hat{\beta}_0$ và $\hat{\beta}_1$ tìm được trong câu 3.2.a., dựa theo hướng dẫn như sau:

```
## Su dung ham lm trong R
model <- lm(y~x)
model$coefficients
```

3.2.c Dựa vào đường thẳng hồi quy (ước lượng bằng phương pháp Least-Squares) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$:

- dự đoán giá trị \hat{y}_1 tại $x = 5$, và ký hiệu điểm này là $(5; \hat{y}_1) \in \mathbb{R}^2$;
- dự đoán giá trị \hat{y}_2 tại $x = 12$, và ký hiệu điểm này là $(12; \hat{y}_2) \in \mathbb{R}^2$;
- dự đoán giá trị \hat{y}_3 tại $x = 18$, và ký hiệu điểm này là $(18; \hat{y}_3) \in \mathbb{R}^2$;

3.2.d Trên cùng đồ thị phân tán ở câu 3.1.:

- vẽ đường thẳng hồi quy $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$, (sử dụng `col = "red"`);
- đồng thời, vẽ thêm các điểm $(5; \hat{y}_1)$, $(12; \hat{y}_2)$ và $(18; \hat{y}_3)$.
(đối với các điểm này, sử dụng `col = "green"` và lựa chọn `pch` phù hợp (chẳng hạn `pch=17`) để phân biệt với các điểm đã được vẽ trong biểu đồ phân tán của dữ liệu)

- - - Good luck! - - -