Exercise 2 - Sep 29

Opened: Thứ Hai, 29 tháng 9 2025, 12:00 AM
Due: Thứ Hai, 29 tháng 9 2025, 1:40 PM

1. Hãy nêu 3 tiêu chí quan trọng để đánh giá **chất lượng dữ liệu** trong quá trình *Data Pre-processing*.

2. Cho tập dữ liệu S = {2, 12, 14, 5, 5, 6, 8, 9, 10, 2, 4}. Hãy tính **Q1, Q2, Q3 và IQR** của tập dữ liệu trên.

3. Cho 2 thuộc tính A = {2, 4, 6} và B = {1, 3, 5}. Hãy tính Covariance và Correlation Coefficient giữa A và B.

4. Bảng dữ liệu khảo sát mối quan hệ giữa giới tính (Nam/Nữ) và sở thích đọc sách (Tiểu thuyết/Khoa học) như

	Tiểu thuyết	**Khoa học**	**Tổng**
Nam	40	20	60
Nữ	30	50	80
Tổng	70	70	140

a. Hãy thiết lập giả thuyết H0 và H1 cho kiểm định Chi-squared.

b. Tính giá trị chi-squared cho bảng dữ liệu trên.

c. Với mức ý nghĩa alpha = 0.05, cho biết kết luận kiểm định.

1. 3 tiêu chí quan trọng để đánh giá **chất lượng dữ liệu** trong quá trình *Data Pre-processing là:* Accuracy, Completeness và Consistency

2.

- Sắp xếp dữ liệu theo thứ tự tăng dần: 2,2,4,5,5,6,8,9,10,12,14 - gồm 11 phần tử

- Do đó Q2 là phần tử thứ 6, hay Q2 = 6

- Nửa dưới: {2,2,4,5,5}, do đó Q1 = 4

- Nửa trên: {8,9,10,12,14}, do đó Q3 = 10

- IQR = Q3−Q1=10−4=6

3.

- A_mean = 4, B_mean = 3, σA = σB = sqrt(8/3)

- Covariance(A,B) = ((2-4)(1-3) + (4-4)(3-3) + (6-4)(5-3))/3 = 8/3

- Correlation_Coefficient(A,B) = (8/3) / (sqrt(8/3))^2 = 1

4.

a.

H0: Giới tính và Sở thích đọc sách là độc lập

H1: Giới tính và Sở thích đọc sách có tương quan

b. o11 = 40, o12 = 20, o21 = 30, o22 = 50

e11 = (60x70)/140 = 30; e12 = (60x70)/140 = 30; e21 = (80x70)/140 = 40; e22 = (80x70)/140 = 40

Giá trị chi-squared: χ^2 = (40-30)^2/30 + (20-30)^2/30 + (30-40)^2/40 + (50-40)^2/40 = 35/3

df (degree of freedom): (2-1)x(2-1)=1

c. Với df = 1, giá trị của χ^2 cần thiết để bác bỏ H0 tại alpha = 0.05 là 3.841, mà 35/3 > 3.841 nên ta có thể bác bỏ H0 và khẳ
định Kết luận rằng hai thuộc tính này tương quan