BÀI TẬP 5
(Lấy mẫu lại)
THỐNG KÊ MÁY TÍNH VÀ ỨNG DỤNG

Câu 1. (2 điểm) 13 hình chữ nhật được chọn ngẫu nhiên từ một tập các hình chữ nhật có chiều dài 2 cạnh kề W, H được cho trong bảng sau

No.	1	2	3	4	5	6	7	8	9	10	11	12	13
W	8.63	4.37	4.92	7.59	7.84	5.13	2.82	6.89	6.77	6.06	3.31	2.82	3.01
H	11.89	6.97	6.53	6.14	11.22	8.87	7.10	6.85	7.94	10.09	6.21	4.25	4.73

a) Tính hệ số tương quan mẫu giữa W và H.

b) Kiểm định giả thuyết "W và H có tương quan" bằng kiểm định hệ số tương quan trong `scipy` (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html).

c) Dùng kĩ thuật lấy mẫu lại hoán vị, kiểm định giả thuyết "W và H có tương quan" và so sánh kết quả với Câu (b).

Câu 2. (3 điểm) Nếu $X \sim \mathcal{U}(0, \theta)$ với $\theta \geq 1$, tức là X có phân phối đều trong khoảng $[0, \theta]$ thì X có kì vọng là $\frac{\theta}{2}$, trung vị là $\frac{\theta}{2}$, độ lệch chuẩn là $\frac{\theta}{2\sqrt{3}}$. Hơn nữa, $P(X \leq 1) = \frac{1}{\theta}$. Từ đó, cho mẫu ngẫu nhiên $X_1, X_2, \dots, X_n \sim \mathcal{U}(0, \theta)$, ta có thể dùng các ước lượng sau cho θ

$$T_1 = 2\bar{X}, \quad T_2 = 2\hat{m}, \quad T_3 = 2\sqrt{3}S, \quad T_4 = \frac{\sum_{i=1}^{n} \mathbb{I}(X_i \leq 1)}{n}, \quad T_5 = \min(X_1, \dots, X_n) + \max(X_1, \dots, X_n),$$

với $\bar{X}, \hat{m}, S, \min, \max$ lần lượt là trung bình, trung vị, độ lệch chuẩn, giá trị lớn nhất, nhỏ nhất của mẫu; kí hiệu $\mathbb{I}(Q)$ cho giá trị 1 nếu Q đúng và 0 nếu Q sai.

Bảng sau đây là một mẫu ngẫu nhiên cỡ $n = 24$ sinh từ phân phối $\mathcal{U}(0, \theta)$

```
3.1209    3.6235    4.4852    0.2718    1.6783    4.1182
0.7732    1.0495    0.2336    3.4155    0.8832    4.2080
0.2738    2.7319    2.9840    1.1990    4.3821    4.3974
3.0841    1.6558    2.8287    2.8890    1.4964    3.3925
```

a) Tính các giá trị ước lượng T_1, T_2, T_3, T_4, T_5 cho θ từ mẫu dữ liệu đã cho.

b) Dùng kĩ thuật bootstrapping, so sánh sai số chuẩn của các ước lượng trên.

c) Giả sử ta có thêm thông tin là θ được sinh từ phân phối chuẩn với kì vọng 5, độ lệch chuẩn 1. Dùng kĩ thuật suy diễn Bayes để ước lượng θ. So sánh sai số của ước lượng này với các ước lượng trên.

Câu 3. (5 điểm) Từ bộ dữ liệu `California Housing` trên trang `scikit-learn` (https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset) , dùng kĩ thuật kiểm tra chéo, chọn ra mô hình "tốt nhất" giải thích giá nhà (target) theo các đặc trưng (feature).

--- HẾT ---