**BÀI TẬP 4**
**THỐNG KÊ MÁY TÍNH VÀ ỨNG DỤNG**

**Câu 1.** (3 điểm) Dữ liệu của 3 đại lượng $X, Y, Z$ được cho trong bảng sau

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{X}$ | 0.55 | 0.72 | 0.60 | 0.54 | 0.42 | 0.65 | 0.44 | 0.89 | 0.96 | 0.38 | 0.79 | 0.53 | 0.57 |
| $\boldsymbol{Y}$ | 1.85 | 0.14 | 0.17 | 0.04 | 1.67 | 1.56 | 1.74 | 1.96 | 1.60 | 0.92 | 1.56 | 0.24 | 1.28 |
| $\boldsymbol{Z}$ | 12.80 | 15.19 | 14.68 | 14.57 | 12.46 | 13.49 | 12.47 | 14.05 | 14.69 | 13.05 | 14.05 | 14.33 | 13.45 |

*Phần I.*

a) Tính hệ số tương quan mẫu giữa $X$ và $Z$.

b) Kiểm định giả thuyết "$X$ và $Z$ 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 "$X$ và $Z$ có tương quan" và so sánh kết quả với Câu (b).

*Phần II.* Xét mô hình hồi qui tuyến tính

$$Z = aX + bY + c + \varepsilon.$$

với $a, b, c$ là các hệ số và $\varepsilon$ là lỗi.

a) Ước lượng các hệ số hồi qui $a, b, c$.

b) Dùng kĩ thuật bootstrapping, xây dựng khoảng tin cậy 95% cho $a, b, c$.

c) Giả sử ta dùng mô hình hồi qui trên để dự đoán giá trị cho $Z$ là $z_0$ tại $x_0 = 0.5, y_0 = 1$. Dùng kĩ thuật bootstrapping, ước lượng sai số dự đoán và xây dựng khoảng tin cậy 95% cho $z_0$.

**Câu 2.** (3 điểm) Nếu $X$ có phân phối Poisson với tham số $\lambda > 0$ thì $X$ có kì vọng và phương sai đều là $\lambda$, còn yếu vị (mode) xấp xỉ $\lambda - \frac{1}{2}$. Từ đó, khi có mẫu dữ liệu của $X$ ta có thể dùng các ước lượng sau để ước lượng $\lambda$

$$T_1 = \bar{X}, \quad T_2 = S^2, \quad T_3 = \hat{m} + \frac{1}{2}$$

với $\bar{X}, S^2, \hat{m}$ lần lượt là trung bình, phương sai và yếu vị mẫu.

Bảng sau đây là một mẫu dữ liệu cỡ $n = 40$ sinh từ phân phối Poissson với tham số $\lambda$.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 3 | 3 | 1 | 5 | 4 | 0 | 2 | 3 | 3 |
| 5 | 3 | 6 | 2 | 1 | 3 | 5 | 2 | 5 | 2 |
| 6 | 3 | 3 | 6 | 6 | 3 | 3 | 4 | 3 | 2 |
| 2 | 1 | 6 | 4 | 2 | 4 | 3 | 5 | 4 | 2 |

a) Tính các giá trị ước lượng $T_1, T_2, T_3$ cho $\lambda$ 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 $3 \leq \lambda \leq 4$. Dùng kĩ thuật suy diễn Bayes để ước lượng $\lambda$. So sánh sai số của ước lượng này với các ước lượng trên.

**Câu 3.** (4 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).

*Lưu ý:* Trình bày bài làm (lời giải, công thức Toán, mã Python, kết quả, …) trong tập tin notebook.

--- HẾT ---