MAS291x_Assignment_01

Bài 1:

Với mục đích so sánh trọng lượng của hai loại nước Coca-Cola thường và Coca-Cola ăn kiêng, người ta lấy ngẫu nhiên 8 lon nước mỗi loại, sau đó đo trọng lượng của từng lon. Khối lượng của 16 lon nước này được liệt kê trong bảng sau:

Thông thường	371	370	370	373	374	372	375	371
Ăn kiêng	353	355	352	354	355	356	355	357

1. Tìm trọng lượng trung bình của Coca-Cola thông thường và Coca-Cola ăn kiêng, sau đó so sánh kết quả. Trọng lượng trung bình của hai loại nước này có bằng nhau không?
2. Tìm các tứ phân của hai mẫu dữ liệu, sau đó so sánh kết quả thu được, đồng thời chỉ ra các outlier (nếu có) và xây dựng boxplot mô tả dữ liệu.
3. Tìm phương sai và độ lệch chuẩn của hai mẫu dữ liệu, sau đó so sánh kết quả.
4. Dựa vào các kết quả trên, bạn có thể kết luận rằng trọng lượng của Coca-cola thông thường và Coca-cola ăn kiêng có giống nhau hay không?

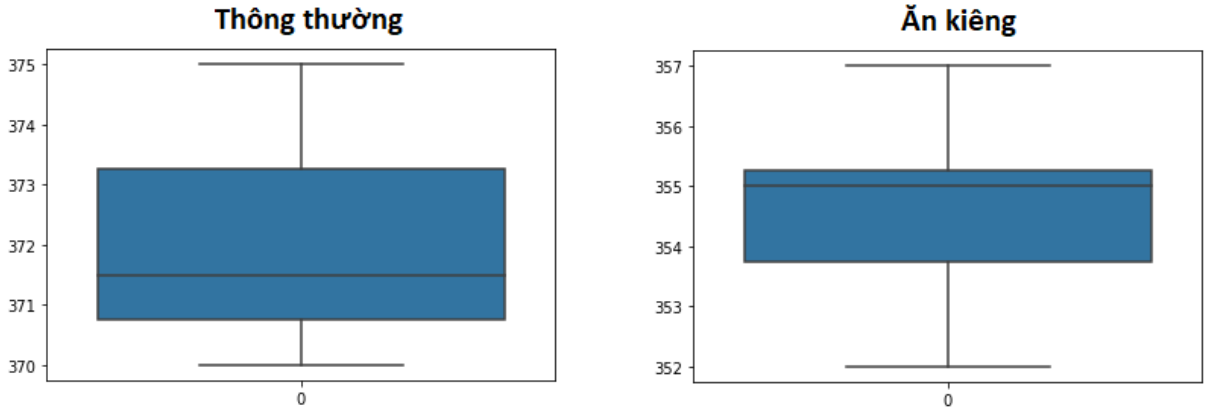
Bài làm

- Đầu tiên chúng ta sắp xếp các giá trị theo thứ tự tang dần để tìm tứ phân vị.
  - Thông thường: [370, 370, 371, 371, 372, 373, 374, 375]
  - Ăn kiêng: [352, 353, 354, 355, 355, 355, 356, 357]

	Thông thường	**Ăn kiêng**
Trung bình	372.0	354.625
Tứ phân vị đầu tiên (Q1)	370.5	353.5
Tứ phân vị thứ 3 (Q3)	373.5	355.5
Phương sai	3.0	2.234
Độ lệch chuẩn	1.732	1.495

➢ Trọng lượng trung bình của Coca thông thường lớn hơn ăn kiêng, tương tự là cả phương sai và độ lệch chuẩn cũng như thế.

- Boxpot và Outlier (nếu có).



- ➢ Từ biểu đồ boxplot, ta có thể thấy cả hai đều không có giá trị ngoại lai nào (outlier), hoặc ta có thể kiểm chứng bằng độ trải giữa (Q1 - 1.5IQR hoặc Q3 + 1.5IQR).
- ➢ Độ biến thiên giữa hai phân phối khác nhau rất rõ ràng, cụ thể là trọng lượng Coca thông thường có phạm vi dao động cao hơn so với ăn kiêng.
- ➢ Chúng ta cũng có thể kiểm định độ biến thiên (sự phân tán dữ liệu) dựa vào phương sai và độ lệch chuẩn, nếu phương sai/độ lệch chuẩn càng lớn thì độ biến thiên càng cao.
- ➢ Từ các kết quả trên, ta kết luận rằng trọng lượng của Coca thông thường và ăn kiêng là không giống nhau.

Bài 2:

Bạn muốn rút tiền mặt từ cây ATM, nhưng do trời tối nên bạn không thể nhìn thấy thẻ khi chèn thẻ vào cây. Thẻ được đưa vào với mặt trước (có in tên bạn) ở phía trên và được chèn vào cây ATM sao cho từ đầu tiên trong tên của bạn được chèn vào trước là đúng quy định.

1. Giả sử rằng bạn chèn thẻ một cách ngẫu nhiên, tính xác suất bạn chèn thẻ đúng theo quy định?
2. Giả sử rằng bạn chèn thẻ một cách ngẫu nhiên, tính xác suất thẻ được chèn không đúng quy định vào lần thử đầu tiên, nhưng được chèn đúng quy định vào lần thử thứ hai?
3. Bạn sẽ cần bao nhiêu lần thử ngẫu nhiên để chắc chắn rằng thẻ hoạt động (thẻ được chèn đúng quy định)?

Bài làm:

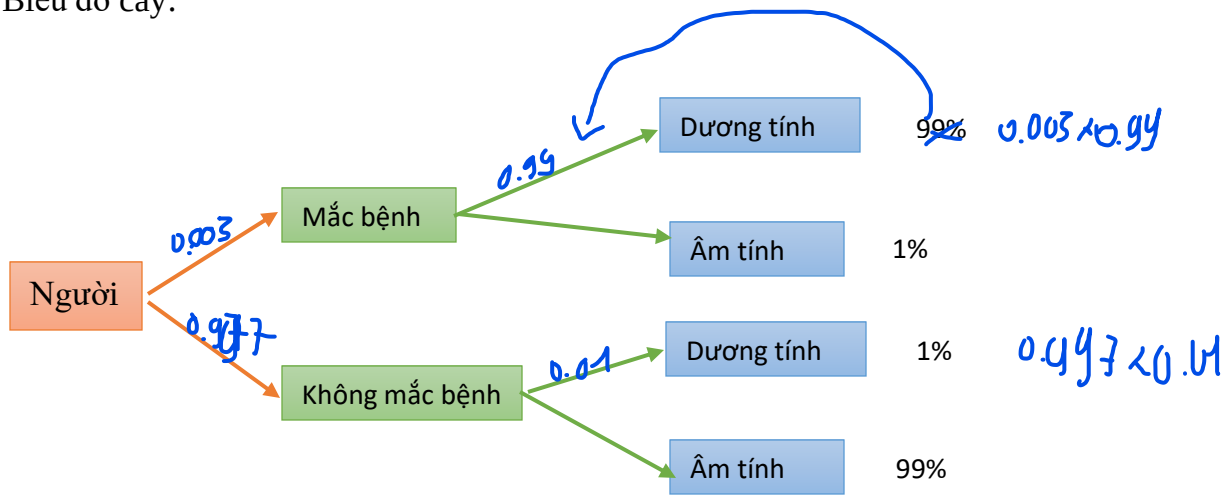
1. Thẻ ATM có 2 mặt, mỗi mặt lại có 2 hướng thẻ có thể chèn thẻ vào cây (vì thẻ ATM hình chữ nhật), vì vậy không gian mẫu sẽ là 4. Xác suất để có thể chèn thẻ đúng theo quy định (mặt có in tên nằm phía trên, và hướng chèn thẻ là chữ cái đầu tiên của tên) là: 1/4
2. Gọi A là biến cố chèn không đúng quy định ở lần đầu tiên thì: P(A) = 3/4
   Gọi B là biến cố chèn đúng quy định lần 2.
   Gọi C là biến cố chèn lần 1 sai, lần 2 đúng.
   Nếu A đã xảy ra thì lúc này không gian mẫu chỉ còn 3 cách. Khi đó P(B|A) = 1/3
   Mà C = A∩B. Do đó theo công thức nhân xác suất ta có:
   $$P(C) = P(A \cap B) = P(B|A)*P(A) = 1/3 * 3/4 = 1/4$$
3. Vì ta xem như mỗi lần thử là ngẫu nhiên và các phép thử là độc lập, vì vậy dù có theo luật số lớn (với số lượng phép thử đủ lớn) thì xác suất để chèn đúng thẻ cũng sẽ là 1/4. Vì vậy, ta sẽ không thể nào biết là sẽ thử bao nhiêu lần để để chắc chắn rằng thẻ hoạt động (vì là xác suất thì không thể nào chắc chắn được), ta chỉ có thể sử dụng định lý Bernoulli để tìm số lần có khả năng nhất mà thôi.

Bài 3:

Xét nghiệm một căn bệnh hiếm gặp luôn mang lại kết quả chính xác đến 99%. Nếu một người bị mắc bệnh, kết quả xét nghiệm sẽ dương tính với xác suất 0.99 và nếu không mắc bệnh, kết quả xét nghiệm sẽ âm tính với xác suất 0.99. Chọn một người ngẫu nhiên trong một nhóm người nhất định có xác suất bị bệnh là 0.003. Giả sử rằng người đó vừa được xét nghiệm dương tính, vậy xác suất mắc bệnh của người đó là bao nhiêu?

Bài làm:

- Biểu đồ cây:



- Gọi B lả biến có người được chọn mắc bệnh này và D là biến cố người được chọn có xét nghiệm dương tính. Áp dụng công thức Bayes, ta có:

$$P(B|D) = \frac{P(B \cap D)}{P(D)} = \frac{P(B).P(D|B)}{P(B).P(D|B)+P(\bar{B})\cdot P(D|\bar{B})}$$

$$= \frac{0.003\times 0.99}{0.003\times 0.99+0.997\times 0.01} \approx 0.23 \approx 23\%$$

- Với P(B): là xác suất người được chọn mắc bệnh.
  - P(B|D): là xác suất mắc bệnh với điều kiện lả có kết quả xét nghiệm dương tính.
  - P(D): là xác suất có kết quả xét nghiệm dương tính, ở đây ta tính xác suất biên P(D) bằng tổng các xác suất có điều kiện của biến cố D, khi ta biết kết quả của biến cố B.
  - P(D|B): là xác suất người mắc bệnh có kết quả xét nhiệm dương tính.

➢ Từ kết quả trên, ta có thể nói rằng kết quả xét nghiệm (+) không giúp ta kết luận được gì về việc người đó có mắc bệnh hay không.

Bài 4:

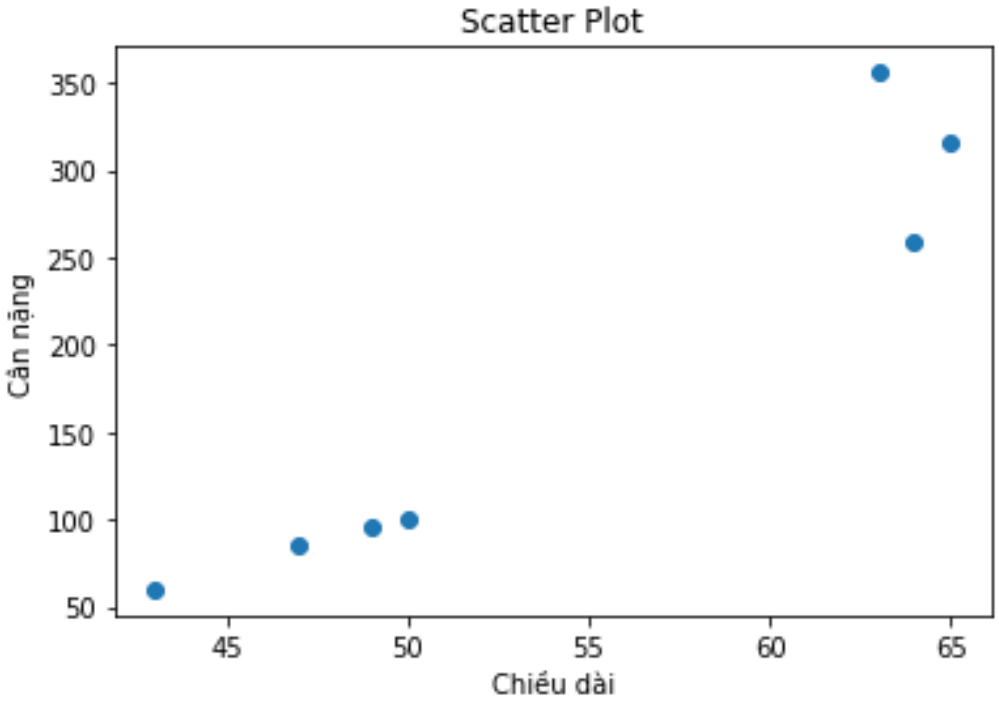
Dưới đây là chiều dài và cân nặng cơ thể của một số con gấu được chọn ngẫu nhiên:

Chiều dài (in)	43	65	63	50	49	64	47
Cân nặng (lb)	60	316	356	100	96	259	86

1. Xây dựng scatter plot mô tả dữ liệu mẫu. Scatter plot đưa ra gợi ý gì về tương quan tuyến tính giữa chiều dài và cân nặng của các con gấu?
2. Tìm hệ số tương quan giữa x và y. Tương quan này có phải là tương quan dương hay không?
3. Bạn hãy xây dựng phương trình đường hồi quy tuyến tính mô tả mối quan hệ phụ thuộc của y vào x.
4. Sử dung phương trình đường hồi quy, bạn hãy dự đoán cân nặng cho một con gấu có chiều dài 72 in.

Bài làm:

1. Biểu đồ scatter mô tả dữ liệu mẫu.



- Từ biểu đồ Scatter, ta có thể dự đoán rằng tương quan tuyến tính giữa chiều dài và cân nặng của con gấu là tương quan dương, ta sẽ kiểm chứng ở về sau.

2. Ta sẽ tìm hệ số tương quan r giữa chiều dài và cân nặng của con gấu.

Chiều dài (in) - x	Cận nặng (lb) - y	Z_x	Z_y	Z_xZ_y
43	60	-1.34	-1.06	1.42
65	316	1.24	1.17	1.45
63	356	1.00	1.51	1.51
50	100	-0.52	-0.71	0.37
49	96	-0.63	-0.75	0.47
64	259	1.12	0.67	0.75
47	86	-0.87	-0.83	0.72
$\bar{x} = 54.429$ $S_x = 8.55$	$\bar{y} = 181.857$ Sy = 114.879	$\Sigma = 0$	$\Sigma = 0$	$\Sigma = 6.69$

- $r = \frac{\Sigma z_x z_y}{n} = \frac{6.69}{7} = 0.957$ => Tương quan (+) dương chặt.

tuyến tính

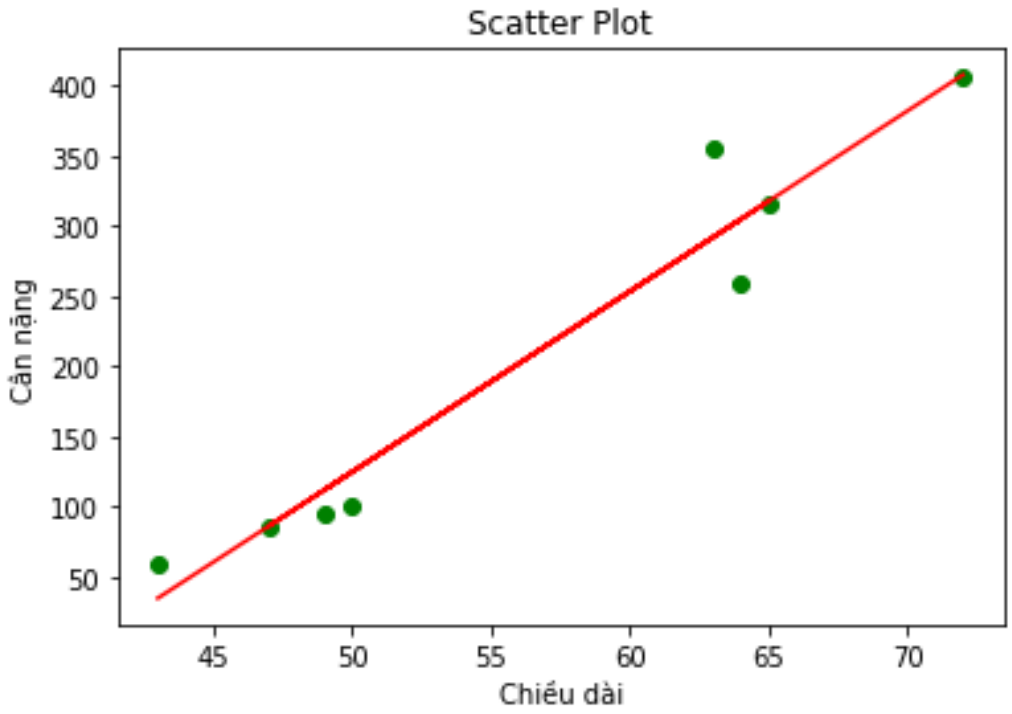
3. Phương trinh đường hồi quy tuyến tính: $\hat{y} = a + bx$

   b = r.$\frac{s_x}{s_y}$ = 0.957× $\frac{114.879}{8.55}$ = 12.86

   a = $\bar{y} - b.\bar{x}$ = 181.857 − 12.86 × 54.429 = -518.1

   ⇨ $\hat{y} = -518.1 + 12.86x$

4. Với chiều dài 72 in, con gấu sẽ có cân nặng: $\hat{y} = -518.1 + 12.86 \times 72$ = 407.82 (lb)



___End___