**Assignment 1**

**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. Trọng lượng trung bình của Coca-cola thông thường:

(371+370+370+373+374+372+375+371) / 8 = 372

Trọng lượng trung bình Coca-cola ăn kiêng:

(353+355+352+354+355+356+355+357) / 8 = 354.62

**So sánh:** Trọng lượng trung bình của Coca-cola thông thường **lớn hơn** Trọng lượng trung bình Coca-cola ăn kiêng khoảng gần 20g

**2. Các tứ phân của hai mẫu dữ liệu:**

Đầu tiên ta sắp xếp lại bảng theo thứ tự tăng dần:

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

**Tứ phân vị gồm 3 giá trị Q1, Q2, Q3 chia tập dữ liệu thành 4 phần bằng nhau, đều gồm n/4=2 phần tử mỗi khoảng:**

**Coca Thông thường:**

Q1: Trung bình cộng giá trị thứ 2 và 3. Có giá trị ( 370+371) / 2 = 370.5

Q2:Trung bình cộng giá trị thứ 4 và 5. Có giá trị ( 371+372) / 2 = 371.5

Q3: Trung bình cộng giá trị thứ 6và 7. Có giá trị ( 373+374) / 2 = 373.5

**Coca Ăn kiêng:**

Q1: Trung bình cộng giá trị thứ 2 và 3. Có giá trị ( 353+354) / 2 = 353.5

Q2: Trung bình cộng giá trị thứ 4 và 5. Có giá trị ( 355+355) / 2 = 355

Q3: Trung bình cộng giá trị thứ 6 và 7. Có giá trị ( 355+356) / 2 = 355.5

**So sánh:**

Tứ phân vị của Coca thường có giá trị giữa các khoảng đồng đều (~1.5) hơn so với tứ phân vị của Coca ăn kiêng. Độ trải giữa(IQR) của Coca thường (Q3-Q1) rộng hơn IQR của Coca ăn kiêng, từ đó có thể nhận xét là dữ liệu của Coca Thường có đô phân tán rộng hơn so với dữ liệu của Coca Ăn kiêng.

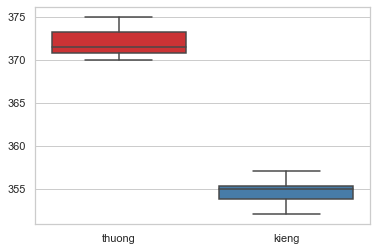
**\*Tìm các giá trị Outlier:**

Các Outliner( nếu có) sẽ nằm ngoài khoảng Q1-1.5IQR tới Q3+1.5IQR:

Với trường hợp Coca bình thường: Từ ( 370.5-1.5IQR) =366 tới (373.5+1.5IQR)=378 suy ra không có Outlier trong tập dữ liệu này.

Với trường hợp Coca ăn kiêng: Từ ( 353.5-1.5IQR) 350.5 tới (355.5+1.5IQR) 358.5 suy ra không có Outlier trong tập dữ liệu này.

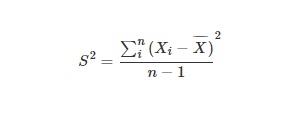
**\*Boxplot tương ứng**



**3. Tìm phương sai và độ lệch chuẩn của hai mẫu dữ liệu, sau đó so sánh kết quả.**

Phương sai là trung bình của tổng bình phương độ lệch của các giá trị dữ liệu so với giá trị trung bình mean:

Từ đó:



Áp dụng công thức ta tìm được giá trị Phương sai của Coca bình thường: 3.429, Coca Ăn kiêng: 2.554

Độ lệch chuẩn s bằng căn bậc 2 của phương sai:

Độ lệch chuẩn của dữ liệu Coca bình thường: s=1.852, Coca Ăn kiêng: s= 1.598

Phương sai và độ lệch chuẩn của dữ liệu Coca Bình thường lớn hơn dữ liệu của Coca Ăn kiêng, từ đó cho thấy độ phân tán dữ liệu trọng lương của Coca Bình thường lớn hơn, trải rộng hơn so với dữ liệu Coca còn lại.

4. Kết luận:

Từ việc tính toán Range IQR, Variance (Phương sai), SD (Độ lệch chuẩn) của dữ liệu trọng lượng 2 loại Coca trên ta có nhận xét tổng thể như sau:

+ Coca Bình thường có Range IQR, Phương sai và độ lệch chuẩn đều lớn hơn loại Coca dành cho ăn kiêng nên độ dàn trải của dữ liệu Coca bình thường sẽ rộng hơn, bên cạch đó các phần tử dữ liệu của Coca bình thường cũng có xu hướng lệch xa hơn so với giá trị trung bình.

Ngược lại dữ liệu của Coca ăn kiêng có độ dàn trải thấp, chênh lệch so với giá trị trung bình ít hơn.

**Kết luận tổng quan thì 2 mẫu trong lượng của 2 loại Coca là khác nhau: Mặc dù không có các giá trị ngoại lai Outlier trong dữ liệu nhưng dữ liệu của Coca Bình thường có độ phân tán dàn trải, ngược lại Coca ăn kiêng có xu hướng tập trung hơn.**

Bài 2:

Gọi A là biến cố để đưa thẻ vào đúng mặt. B là biến cố đưa vào đúng chiều hợp lệ (có tên)

1. Ta có P(A) là xác suất xảy ra biến cố A là 0.5 ( Có 2 trường hợp xấp và ngửa có tỉ lệ ngẫu nhiên bằng nhau bằng 0.5)

Tương tự P(B) là xác suất xảy ra biến cố đưa vào đúng chiều hợp lệ: 0.5

Vì xác suất xảy ra biến cố A và B là ngẫu nhiên và độc lập nên xác suất xảy ra đồng thời cả A và B là:

P(AB) = P(A) x P(B)= 0.5\*0.5= 0.25

2. Vì biến cố để thẻ chèn thẻ không đúng quy định là biến cố đối của chèn thẻ đúng quy định nên sẽ có xác suất bằng

1- P(AB)=0.75

Vậy xác suất để thẻ được chèn không đúng quy định vào lần thử đầu tiên, nhưng được chèn đúng quy định vào lần thử thứ hai là: 0.75 x 0.25=0.1875

3. Vì xác suất để thẻ đưa vào hợp lệ là P(AB) =0.25, nên theo lý thuyết số lần cần thử để đưa thẻ vào đúng quy định bằng 1/0.25=4 lần thử ngẫu nhiên.

**Bài 3:**

Đầu tiên ta có thể khẳng định đây là xác suất có điều kiện được suy luận theo trường phái Bayes.

Vì xác suất xảy ra A,B ảnh hưởng đến xác suất xảy ra biến cố còn lại nên 2 biến cố A và B không độc lập.

Gọi A là biến cố một người bị mắc bệnh. Ta có P(A)=0.003

Suy ra xác suất để người đó không mắc bệnh là P()=1- P(A)=0.997

P(B) là biến cố một người được chọn ngẫu nhiên xét nghiệm dương tính.

P(B|A) là xác suất một người xét nghiệm dương tính khi bị bệnh. P(B|A)=0.99

P(|) là xác suất một người xét nghiệm âm tính khi không bị bệnh. P(|)= 0.99

P(A|B) là biến cố một người mắc bệnh khi người đó có kết quả xét nghiệm dương tính

Từ đó ta có các kí hiệu: P(B**∩**A) , P(B**∩**), P(), P(**∩**A) lần lượt là xác suất xảy ra đồng thời 2 biến cố được nêu ở trên.

**Decision Tree ( Sơ đồ cây)**

P()

P(|)

0.99

P()=0.997

P(B|)

0.01

P(B**∩**)

P(**∩**A)

P(|A)0.01

P(A) 0.003

P(B|A) 0.99

P(B**∩**A)

Dựa trên phân hoạch không gian mẫu bởi 2 tập rời nhau và xác suất đối trong cùng 1 điều kiện

P(B|A) + P(|A) = 1 ta tính xác suất của biến cố một người được chọn ngẫu nhiên xét nghiệm

dương tính:

P(B)= P(B|A) \* P(A) + P(B|)\* P() = 0.99\*0.003+0.01\*0.997=0.0129

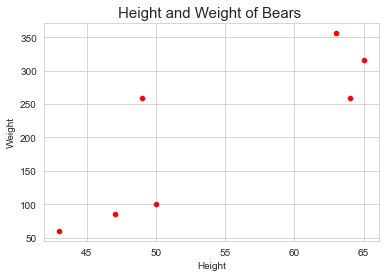
Từ đó, biến cố một người mắc bệnh khi người đó có kết quả xét nghiệm dương tính là:

P(A|B)= =0.99 × 0.003 / 0.0129= 0.23

Vậy giả sử rằng người đó vừa được xét nghiệm dương tính, vậy xác suất mắc bệnh của người đó là 23%

**Bài 4:**

**1. Scatter Plot mô tả dữ liệu**

****

Nhìn vào scatter plot ta có thể thấy được xu hướng tương quan giữa cân nặng và chiều cao của các cá thể gấu. Có thể thấy chiều dài càng lớn thì các cá thể gấu cũng có cân nặng càng lớn theo tuyến tính dương.

**2. Tìm hệ số tương quan giữa x và y. Tương quan này có phải là tương quan dương hay không?**

Đầu tiên ta tìm giá trị Mean- trung bình và Standard Deviation ( độ lệch chuẩn) của 2 biến x- Chiều dài và y- cân nặng theo công thức dưới đây :

Ta có kết quả được tổng hợp ở dạng bảng:

|  |  |  |
| --- | --- | --- |
|  | x- Chiều dài | y- cân nặng |
| Mean - Trung bình mẫu | 54.43 | 205.14 |
| S - Độ lệch chuẩn | 9.24 | 120.5 |

Để tìm hệ số tương quan trước tiên ta phải tìm các giá trị z-score từng phần tử của các biến x, y theo công thức:

z-score= Trong đó là Mean của mẫu, s là độ lệch chuẩn của dữ liệu

Sau đó ta tính hệ số tương quan **r** theo công thức:

r=

**Sau khi tính tổng các cặp Zx\*Zy, áp dụng công thức trên ta tính được hệ số tương quan r bằng 0.853**

**Ta có hệ số tương quan có giới hạn -1 r 1, nên giá trị r bằng 0.853 là tương quan tuyến tính dương.**

**3. Bạn hãy xây dựng phương trình đường hồi quy tuyến tính mô tả mối quan hệ phụ thuộc của y vào x.**

Để xây dựng phương trình hồi quy tuyến tính có dạng **y= a+bx**  ta phải dựa vào hệ số tương quan r để tìm hệ số góc slope b và hằng số- intercept a

Ta có công thức:

**Hệ số góc b = Hệ số tương quan r \* (Độ lệch chuẩn của y / Độ lệch chuẩn của x ) =** 0.853\*(120.5/9.24)= 11.134

**Hằng số a = Mean y – b \* (Mean x)** = 205.14 - 11.134\*54.43= - 400.88

Vậy đường hồi quy tuyến tính sẽ là : **y= -400.88+11.134x**

**4. Sử dung phương trình đường hồi quy, bạn hãy dự đoán cân nặng cho một con gấu có chiều dài 72 .**

Ta có thể nhận thấy giá trị chiều dài 72 vượt quá các giá trị trong dữ liệu chiều dài của biến x, nên đây có thể xem là trường hợp ngoại suy ( dự đoán cho giá trị vượt quá phạm vị của biến độc lập ). Vì không có cách nào để biết được một mối quan hệ giữa giá trị vượt quá phạm vi của biến độc lập có còn tuân theo quy luật của hồi quy tuyến tính hay không nên trường hợp ngoại suy này không đáng tin cây.

Tuy nhiên vì mục đích tham khảo ta có thể dễ dàng dự đoán được giá trị của y bằng cách thay x= 72 vào phungơ trình đường hồi quy tuyến tính: y= -400.88+11.1348\*72 = 400.8256. Vậy có thể dự đoán tương đối (không đáng tin) rằng 1 con gấu dài 72 inch có thể tới cân nặng hơn 400 kí.