Bài giảng môn học:

Khai phá dữ liệu (7080508)

CHƯƠNG 2:
THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU
(Phần 2)

08/2024

Nội dung chương 2 – Phần 2

1.1 Tóm tắt mô tả dữ liệu

1.2 Đo lường xu hướng trung tâm (Central tendency)

1.3 Đo lường xu hướng phân tán (Dispersion)

1.4 Đo lường vị trí (Position)

1.5 Đo lường mức độ tương quan của dữ liệu (Correlation)

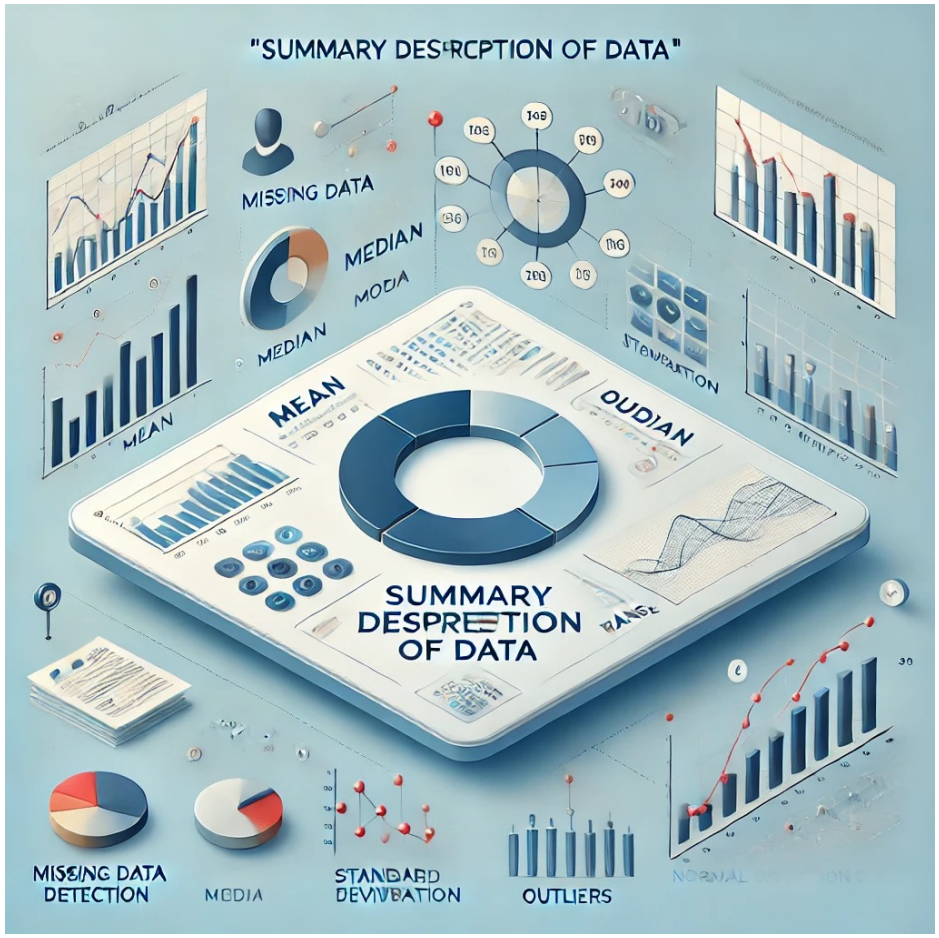
1.6 Bài tập

1.1 Tóm tắt mô tả dữ liệu

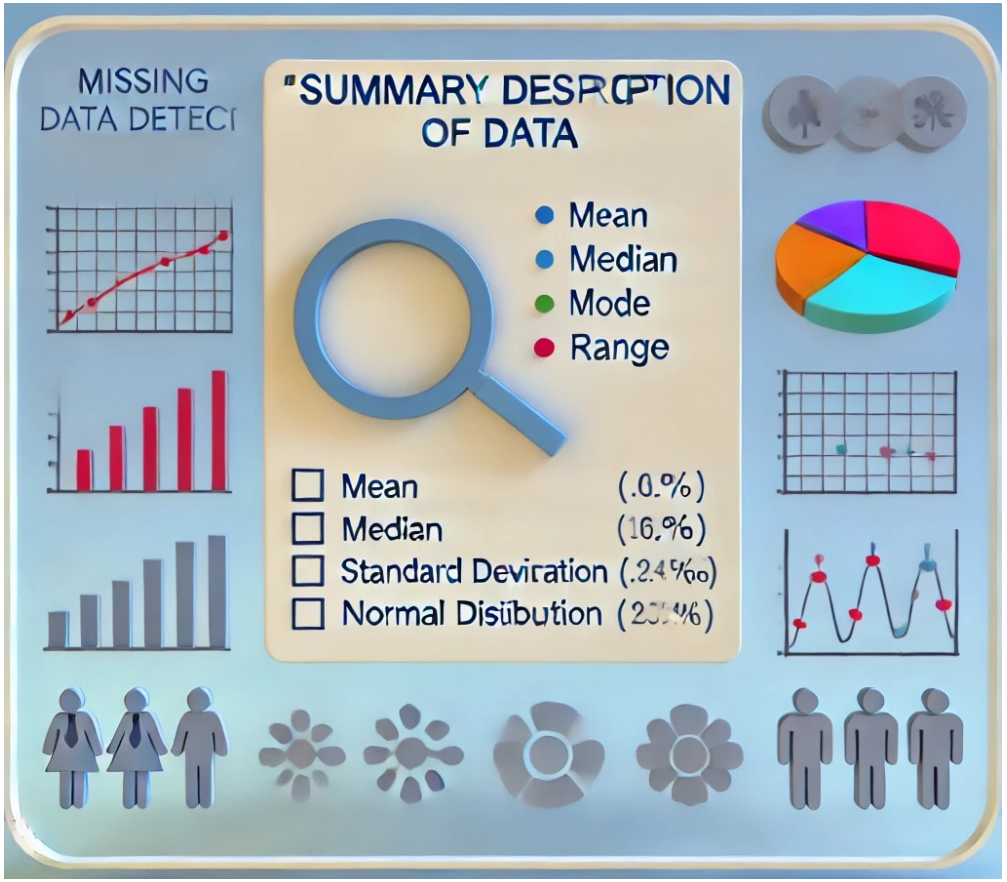
Tóm tắt mô tả dữ liệu

Tóm tắt mô tả dữ liệu (Summary Description) là một bước quan trọng trong việc hiểu tập dữ liệu vì nó cung cấp một cái nhìn tổng quan về các đặc trưng chính của dữ liệu. Việc tóm tắt dữ liệu giúp nắm bắt thông tin quan trọng một cách nhanh chóng, giúp đưa ra các quyết định chính xác hơn trong quá trình phân tích dữ liệu.

Tóm tắt mô tả dữ liệu thường bao gồm các chỉ số thống kê như giá trị trung bình, trung vị, độ lệch chuẩn, min, max, phân vị (quartiles), v.v. Các chỉ số này cung cấp một cái nhìn sơ bộ về phân phối dữ liệu và giúp xác định xu hướng chung, giá trị bất thường, hoặc sự sai lệch trong dữ liệu.



Tóm tắt mô tả dữ liệu



Tóm tắt mô tả dữ liệu là một công cụ quan trọng giúp hiểu sâu hơn về cấu trúc và hành vi của dữ liệu.

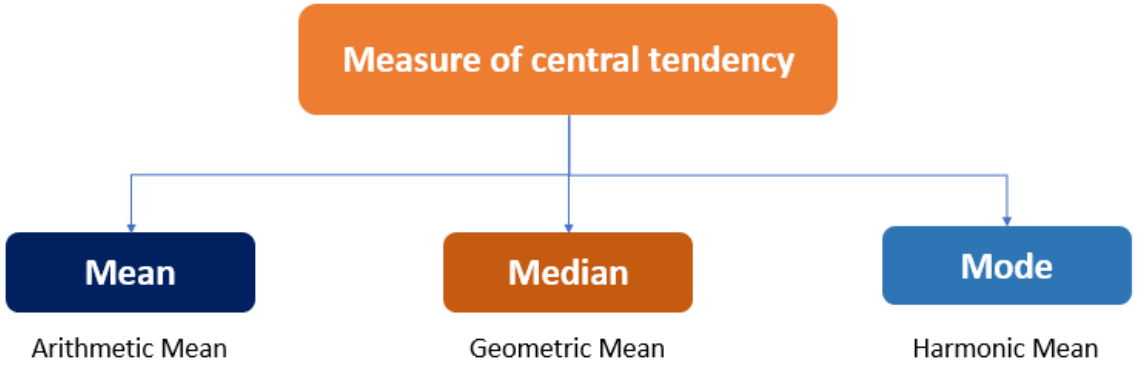
Nó không chỉ giúp tiết kiệm thời gian trong việc xác định các đặc trưng quan trọng mà còn hỗ trợ trong quá trình tiền xử lý và ra quyết định, đảm bảo rằng các phân tích tiếp theo được thực hiện trên nền tảng dữ liệu chất lượng.

1.2 Đo lường xu hướng trung tâm

Đo lường xu hướng trung tâm của dữ liệu

Các số đo xu hướng trung tâm: là những thống kê cho chúng ta biết đặc điểm cơ bản của một tập dữ liệu. Các số đo này dùng để mô tả một giá trị tiêu biểu hoặc điểm trung tâm của một tập dữ liệu. Nó cung cấp thông tin về đặc điểm chung của tập dữ liệu, thay vì các giá trị cá nhân cụ thể, giúp người ta hiểu rõ hơn về sự phân bố của dữ liệu và xác định được các xu hướng chung mà dữ liệu thể hiện

Ba số đo xu hướng trung tâm phổ biến nhất là giá trị Mean, Median và Mode



Đo lường xu hướng trung tâm của dữ liệu

Giá trị trung bình (Mean): Là tổng giá trị của từng quan sát trong một tập dữ liệu chia cho số lượng quan sát.

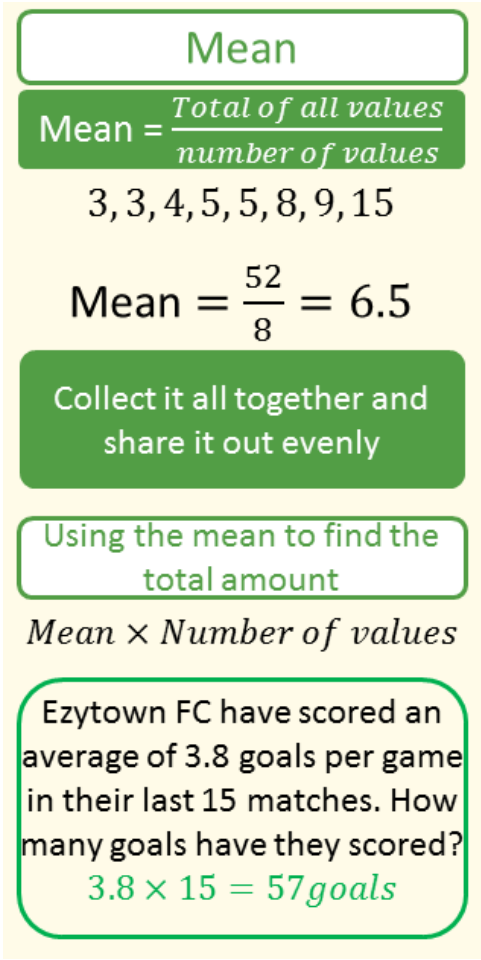
Ví dụ: Cho một phân bố gồm 11 quan sát sau:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60 → Mean = 56,5

+ Mean có thể được sử dụng cho cả dữ liệu số liên tục và rời rạc.

+ Mean không thể được tính cho dữ liệu phân loại (categorical data), vì các giá trị không thể được cộng lại với nhau.

+ Vì mean bao gồm mọi giá trị trong phân bố, nên nó bị ảnh hưởng bởi các giá trị ngoại lệ và phân bố lệch.



Đo lường xu hướng trung tâm của dữ liệu

Giá trị trung vị (Median): là giá trị nằm ở giữa trong phân bố khi các giá trị được sắp xếp theo thứ tự tăng dần hoặc giảm dần.Trung vị chia phân bố thành hai nửa. Trong một phân bố với số lượng quan sát lẻ, giá trị trung vị là giá trị ở giữa. Khi phân bố có số lượng quan sát chẵn, giá trị trung vị là trung bình cộng của hai giá trị ở giữa.

Ví dụ:

Cho 1 phân bố gồm 11 quan sát: 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

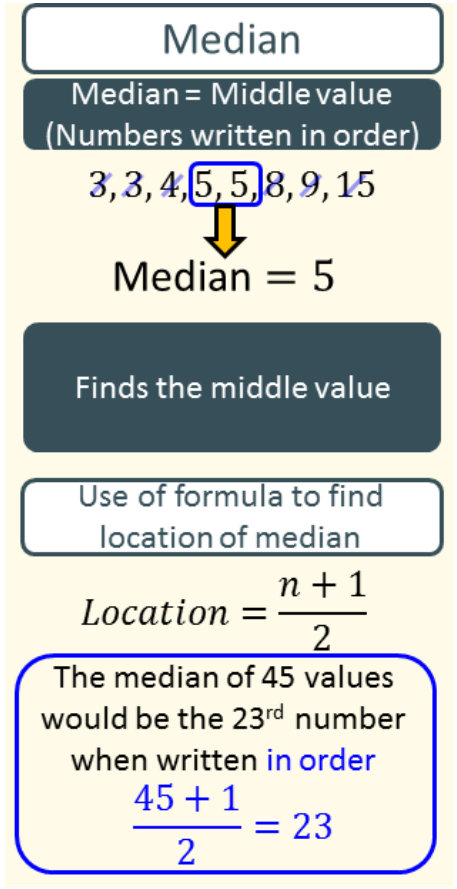
→ Median = 57

Cho 1 phân bố với 12 quan sát: 52, 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

→ Median =56,5

+ Trung vị ít bị ảnh hưởng bởi các giá trị ngoại lệ và dữ liệu lệch hơn so với trung bình cộng, và thường là số đo xu hướng trung tâm được ưa chuộng khi phân bố không đối xứng.

+ Trung vị không thể được xác định cho dữ liệu danh mục định danh vì nó không thể được sắp xếp theo thứ tự logic.



Đo lường xu hướng trung tâm của dữ liệu

Giá trị Mode: Là giá trị xuất hiện thường xuyên nhất trong một phân bố.

Ví dụ:

Cho phân bố gồm 11 quan sát: 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

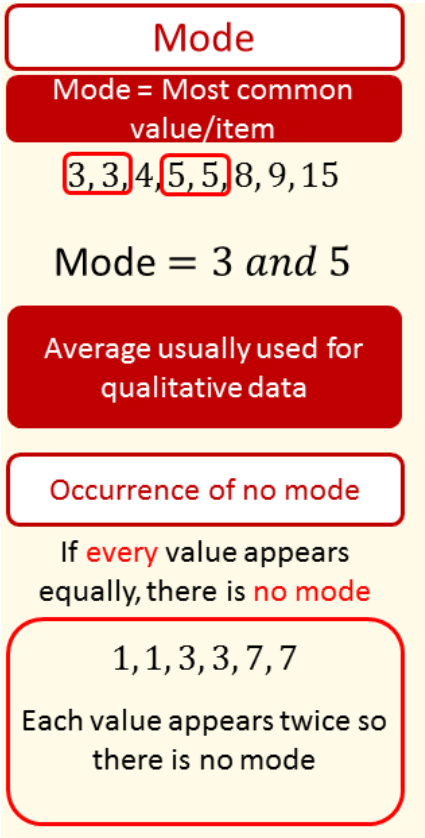
→ Mode = 54

Cho phân bố gồm 12 quan sát: 52, 53, 54, 55, 56, 57, 58, 58, 60, 60, 61,62

→ Mode =58 và 60

+ Số Mode được xác định cho cả dữ liệu số và dữ liệu phân loại

+ Sự tồn tại của nhiều hơn một số mốt có thể hạn chế khả năng của số mốt trong việc mô tả trung tâm hoặc giá trị điển hình của phân bố vì không thể xác định một giá trị duy nhất để mô tả trung tâm.



Đo lường xu hướng trung tâm của dữ liệu

Ví dụ:

Giả sử rằng bạn chạy 100 m trong sáu lần, mỗi lần chạy bạn dùng đồng hồ đo lại thời gian chạy (tính bằng giây) và kết quả 6 lần chạy của bạn gồm sáu giá trị như sau:

$$x = \{25.1, 21.2, 17.9, 23.0, 24.6, 19.5\}$$

Giả sử sau khi chạy hết 6 lần, bạn chạy tiếp lần thứ 7. Lần này đột nhiên chân bạn bị đau và bạn đi bộ thay vì chạy và kết quả thời gian của lần này là 79.9 giây. Bạn cố gắng thử thêm lần nữa và kết quả vẫn 79.9 giây. Bây giờ ta có Sample về 8 lần chạy như sau:

$$x=\{25.1, 21.2, 17.9, 23.0, 24.6, 19.5, 79.9, 79.9\}$$

Dữ liệu này cho bạn biết những thông tin gì?

Độ đo	6	8
Mean	21.9 giây	36.4 giây
Median	22.1 giây	23.8 giây
Mode	Not available	79.9 giây

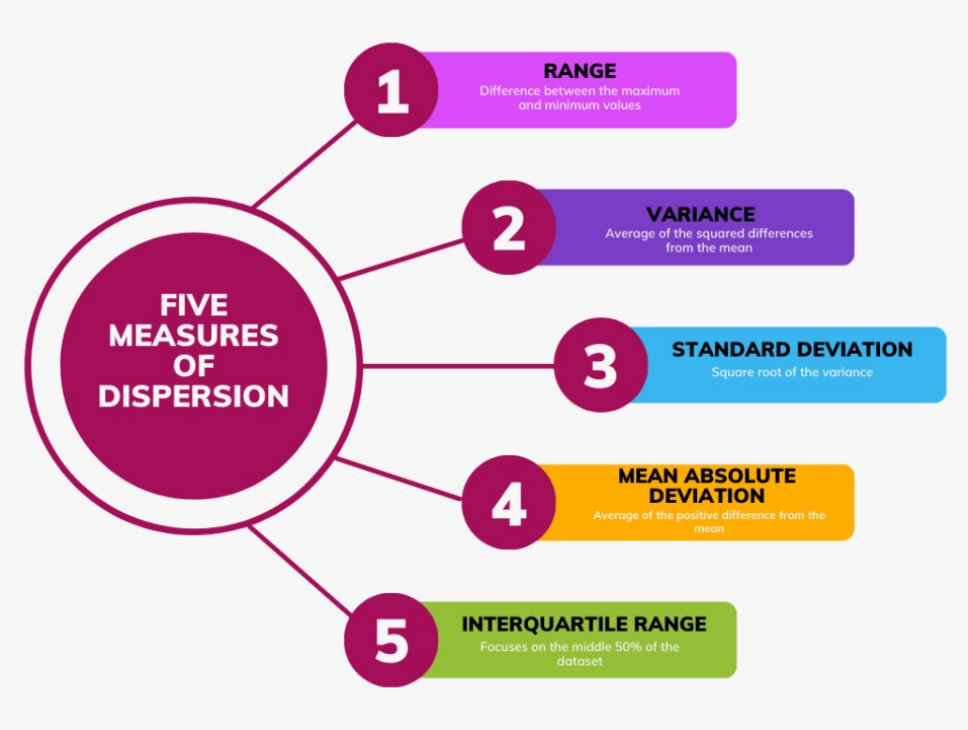
1.3 Đo lường xu hướng phân tán

Đo lường xu hướng phân tán của dữ liệu

Các số đo độ phân tán là các số thực không âm giúp đánh giá mức độ phân tán của dữ liệu xung quanh một giá trị trung tâm. Ứng dụng quan trọng nhất của các số đo độ phân tán là giúp hiểu được phân bố của dữ liệu. Khi dữ liệu trở nên đa dạng hơn, giá trị của số đo độ phân tán sẽ tăng lên.

+ Một số đo độ phân tán được sử dụng phổ biến nhất:

- ✓ Phạm vi (range)
- ✓ Phương sai (variance)
- ✓ Độ lệch chuẩn (standard deviation)
- ✓ Độ lệch trung bình (mean deviation)
- ✓ Đo lường vị trí (Interquartile Range (IQR))



Đo lường xu hướng phân tán của dữ liệu

Phạm vi (range): Là sự chênh lệch giữa giá trị nhỏ nhất và giá trị lớn nhất trong một tập dữ liệu.

Ví dụ: Tập dữ liệu B: 1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11

Phạm vi là 10, là sự chênh lệch giữa giá trị cao nhất (11) và giá trị thấp nhất (1).

+ Range càng cao thì mức độ biến thiên càng lớn

+ Range bị ảnh hưởng rất lớn nếu có dữ liệu ngoại lai

Đo lường xu hướng phân tán của dữ liệu

Phương sai (variance): Là giá trị đặc trưng cho độ phân tán (biến thiên) của các số liệu trong bộ số liệu so với giá trị trung bình của bộ số liệu.

$$s^2 = \frac{\sum_i^n (X_i - \bar{X})^2}{n-1}$$

Trong đó: $\bar{X}$: là giá trị trung bình của bộ số liệu

X_i :Là các giá trị của bộ số liệu

n: số phần tử của bộ số liệu

+ Variance cho biết mức độ đồng nhất hoặc không đồng nhất của các giá trị trong một tập dữ liệu.

+ Variance không bao giờ âm, vì nó là tổng của bình phương độ lệch trung bình từ giá trị trung bình

+ Variance của một hằng số luôn = 0.

+ Dễ bị ảnh hưởng bởi các giá trị ngoại lai

Đo lường xu hướng phân tán của dữ liệu

Ví dụ: Tính Phương sai của data set bao gồm các điểm: 46, 69, 32, 60, 52, 41

Bước 1: Tính giá trị trung bình: $\bar{X}$ = (46+69+32+60+52+41)/6= 50

Bước 2: Tìm độ lệch của từng điểm so với giá trị trung bình ($X_{i-} - \overline{X}$)

46	69	32	60	52	41
-4	19	-18	10	2	-9

Bước 3: Bình phương mỗi độ lệnh so với giá trị trung bình $(X_{i-} - \overline{X})^2$

16	361	324	100	4	81

16 + 361 + 324 + 100 + 4 + 81 = 886

Áp dụng công thức: $\mathbf{S^2 = 886/(6-1) = 177,2}$

Đo lường xu hướng phân tán của dữ liệu

Độ lệch chuẩn (standard deviation): Là đại lượng để đo lường mức độ biến thiên của các giá trị trong tập dữ liệu so với giá trị trung bình của tập dữ liệu đó. Nó được xác định bằng căn bậc hai

$$S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Trong đó: $\bar{X}$: là giá trị trung bình của bộ số liệu

x_i :Là các giá trị của bộ số liệu

n: số phần tử của bộ số liệu

+ Độ lệch chuẩn cho ta biết về sự biến thiên, từng giá trị quan sát có mối liên hệ tập trung như thế nào xung quanh giá trị trung bình.

+ Nếu độ lệch chuẩn bằng 0 => phương sai bằng 0 => các giá trị quan sát cũng chính là giá trị trung bình hay nói cách khác không có sự biến thiên nào cả.

+ Nếu độ lệch chuẩn càng lớn => sự biến thiên xung quang giá trị trung bình càng lớn

1.4 Đo lường vị trí

Đo lường vị trí của dữ liệu

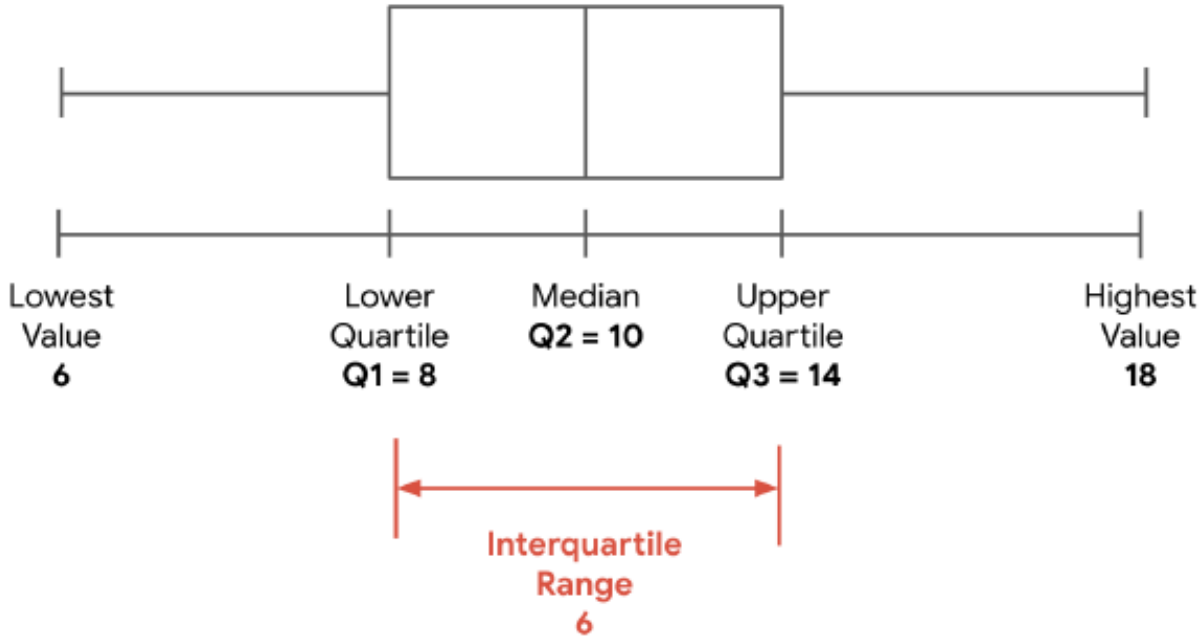
Interquartile Range (IQR) Là một thước đo thống kê dùng để đánh giá độ phân tán của dữ liệu, đặc biệt là để đo độ phân tán ở giữa dữ liệu. IQR đại diện cho khoảng giữa của 50% dữ liệu trong tập dữ liệu đã sắp xếp theo thứ tự tăng dần. IQR được tính bằng cách lấy hiệu giữa giá trị của quartile thứ ba (Q3) và quartile thứ nhất (Q1): **IQR=Q3−Q1**

Trong đó:

Q1: Là giá trị mà 25% các điểm dữ liệu nằm dưới nó

Q2: 50% dữ liệu nằm dưới giá trị này còn được gọi là trung vị.

Q3: Là giá trị mà 75% các điểm dữ liệu nằm dưới nó.



Đo lường vị trí của dữ liệu

❖ **Cách xác định vị trí Q1, Q3**

Q1=(n+1)/4

Q3= 3(n+1)/4

Trong đó: n là số lượng phần tử trong tập dữ liệu

Nếu vị trí là một số nguyên, Q1, Q3 là giá trị tại vị trí đó trong tập dữ liệu.

Nếu vị trí không phải là số nguyên thì Q1, Q3 được lấy trung bình của hai giá trị gần nhất.

Đo lường vị trí của dữ liệu

Ví dụ: Giả sử bạn có tập dữ liệu sau, đại diện cho điểm kiểm tra của 10 học sinh
55, 60, 62, 65, 68, 70, 72, 75, 78, 80

Bước 1: Sắp xếp theo thứ tự tăng dần: 55, 60, 62, 65, 68, 70, 72, 75, 78, 80

Bước 2: Xác định giá trị Q1 và Q3

- ✓ **Q1**: Tập dữ liệu có 10 giá trị, vị trí của Q1 sẽ nằm giữa giá trị thứ 2 và thứ 3

  → Q1=(60+62)/2=61

- ✓ **Q3**: Vị trí của Q3 nằm giữa giá trị thứ 8 và thứ 9 → Q3=(75+78)/22=76.5

Bước 3: Tính IQR: Q3−Q1=76.5−61=15.5

Kết luận: 50% giá trị ở giữa (tức là một nửa dữ liệu) sẽ nằm trong phạm vi từ 61 đến 76.5.

Điều này cho thấy rằng phần giữa của dữ liệu có mức độ biến động không quá lớn, tập trung trong khoảng 15.5 đơn vị giữa hai giá trị này. IQR càng nhỏ thì dữ liệu càng tập trung, IQR càng lớn thì dữ liệu càng phân tán.

1.5 Đo lường mức độ tương quan

Đo lường mức độ tương quan của dữ liệu

- Dựa trên các dữ liệu đã có, phân tích tương quan có thể cho thấy mức độ mà một thuộc tính có thể được suy diễn hoặc được quyết định bởi một thuộc tính khác.
- Hệ số tương quan: dùng để đánh giá độ tương quan giữa 02 thuộc tính. Cụ thể, hệ số tương quan giữa 02 thuộc tính x và y được xác định:

$$r_{x,y} = \frac{\overline{xy} - \bar{x}\bar{y}}{S_x S_y}$$

Trong đó:

- n: số bộ dữ liệu.
- xi,yi là các giá trị tương ứng với 02 thuộc tính x và y trong bộ i.
- $\bar{x}$, $\bar{y}$, tương ứng là các giá trị trung bình trên x và y.
- Sx, Sy tương ứng là độ lệch chuẩn của x và y.

Đo lường mức độ tương quan của dữ liệu

- **Ta có: $-1 \leq r_{x,y} \leq 1$**
  + Nếu $r_{x,y} > 0$: x, y có mối tương quan dương (giá trị ứng với x tăng thì giá trị ứng với y cũng tăng). Giá trị $r_{x,y}$ càng lớn thể hiện tính tương quan giữa 02 thuộc tính càng mạnh $\Longrightarrow$ Có thể loại bỏ một trong 02 thuộc tính (x hoặc y) vì nó là dư thừa.

  + Nếu $r_{x,y} = 0$: Không tồn tại mối liên hệ tương quan. x và y là 02 thuộc tính hoàn toàn độc lập.

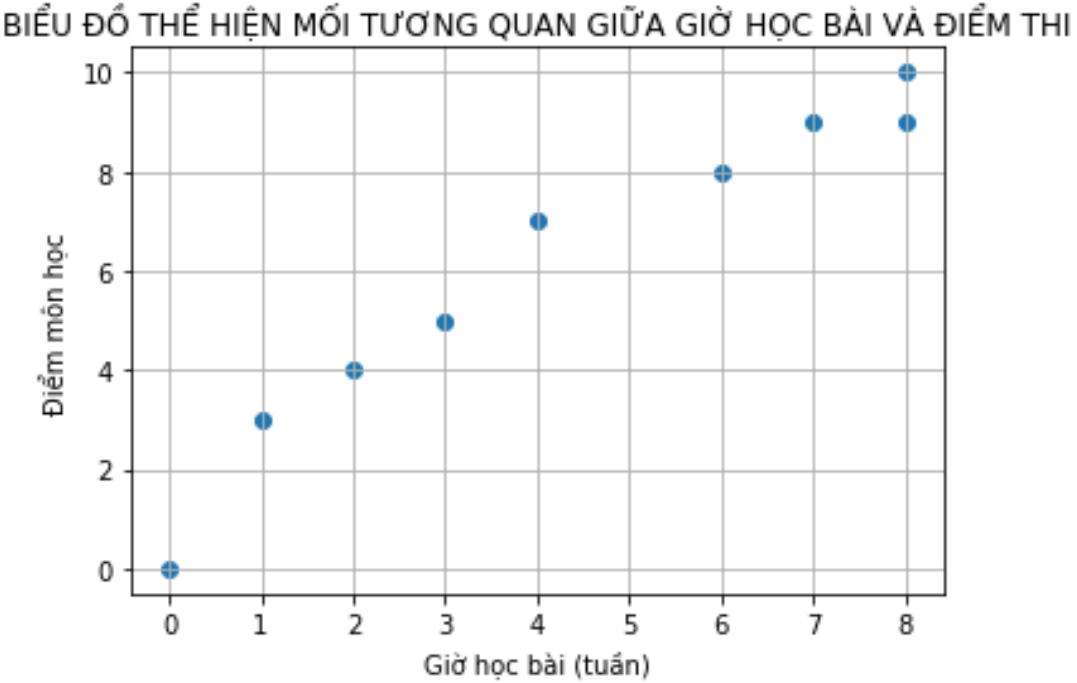
  + Nếu $r_{x,y} < 0$: x, y có mối tương quan âm (giá trị ứng với x tăng thì giá trị ứng với y giảm và ngược lại) $\Longrightarrow$ x và y là 02 thuộc tính trái ngược nhau
- **Chú ý:**
  + Hệ số tương quan đo lường mức độ quan hệ tuyến tính giữa hai biến
  + Hệ số tương quan đo lường mức độ quan hệ tuyến tính giữa hai biến

Đo lường mức độ tương quan của dữ liệu

Ví dụ về mối tương quan giữa **thời gian dành cho việc học bài** với **điểm thi nhận được**!

```
#corrcoef: Hệ số tương quan
#Thời gian dành cho học bài
a_giohoc = np.array([4,7,1,2,8,0,3,8,6])
#Điểm thi nhận được:
b_diem   = np.array([7,9,3,4,9,0,5,10,8])
co = np.corrcoef(a_giohoc,b_diem)
print(type(co))
print('Hệ số tương quan: \n', co)
```

```
<class 'numpy.ndarray'>
Hệ số tương quan:
 [[1.         0.96995403]
 [0.96995403 1.        ]]
```



Đo lường mức độ tương quan của dữ liệu

Ví dụ:

Theo dõi mức lãi suất (Y) và tỷ lệ lạm phát (X) ở một số nước ta có số liệu sau:

Y	17,5	15,6	9,8	5,3	7,9	10,0	19,2	13,1
X	14,2	11,7	6,4	2,1	4,8	8,1	15,4	9,8

Tính hệ số tương quan mẫu X,Y

Ta có $\overline{X} = 9{,}0625$; $\overline{Y}$=12,3; $\overline{XY} = 130{,}9813$

S_X2=21,24; S_X =4,6

S_Y2=23,72; S_Y =4,87

Vậy hệ số tương quan mẫu sẽ là: $r=\frac{130{,}9813-9{,}0625*12{,}3}{4{,}6*4{,}87}$ =0,87

1.5 Bài tập

Thực hành

Bài số 1: Giả sử giá trị của thuộc tính tuổi như sau:

Tuổi	13	15	16	16	19	20	20	21	22	22	25	25	25	25	30	33	33	35	35	35	35	36	40	45	46	52	70

a) Tính giá trị trung bình và trung vị của tập dữ liệu trên

b) Tính giá trị mode và kết luận tập dữ này có đặc điểm gì

c) Tính giá trị Q1, và Q3

Bài số 2: Giả sử dữ liệu kiểm tra liên quan giữ tuổi và sự béo phì của bệnh viện trên 18 người chọn ngẫu nhiên

Age	23	23	27	27	39	41	47	49	50	52	54	54	56	57	58	58	60	61
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

a) Tính giá trị trung bình, trung vị và độ lệch chuẩn của 2 thuộc tính age và tỷ lệ béo %fat cho tập dữ liệu trên

b) Tính hệ số tương quan giữa 2 thuộc tính trên. Kết luận 2 thuộc tính có quan hệ gì với nhau?

Thực hành

Yêu cầu:

1. xxx


Q
A