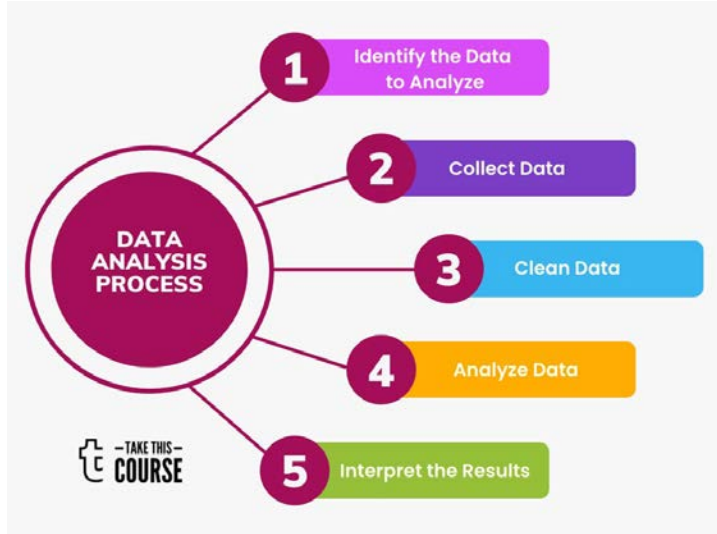


PHÂN TÍCH DỮ LIỆU

(Data Analysis)



THE DATA ANALYSIS PROCESS



Lê Văn Hạnh
levanhhanhvn@gmail.com

NỘI DUNG MÔN HỌC

PHẦN 1 TỔNG QUAN & THU THẬP DỮ LIỆU CHO VIỆC PHÂN TÍCH

1. Khoa học dữ liệu
2. Thu thập dữ liệu
3. Tìm hiểu dữ liệu

PHẦN 2: TIỀN XỬ LÝ DỮ LIỆU (*Data Preprocessing*)

4. Nhiệm vụ chính trong tiền xử lý dữ liệu
5. PANDAS
6. Thao tác với các định dạng khác nhau của tập tin dữ liệu
7. Làm sạch và Chuẩn bị dữ liệu
8. Sắp xếp dữ liệu: nối, kết hợp và định hình lại
9. Tổng hợp dữ liệu và các tác vụ trên nhóm

PHẦN 3 TRỰC QUAN HÓA DỮ LIỆU (*Data Visualization*)

10. Đồ thị và Biểu đồ
11. Vẽ đồ thị và Trực quan hóa



PHẦN 1 TỔNG QUAN & THU THẬP DỮ LIỆU CHO VIỆC PHÂN TÍCH

Chương 3

TÌM HIỂU DỮ LIỆU (Getting to Know Your Data)



Lê Văn Hạnh

levanhhanhvn@gmail.com




NỘI DUNG CHƯƠNG 3

1. Đối tượng dữ liệu và kiểu thuộc tính
2. Mô tả các thống kê cơ bản
3. Trực quan hóa dữ liệu
4. Đo lường sự tương đồng và khác biệt của dữ liệu
5. Bài tập

1. ĐỐI TƯỢNG DỮ LIỆU VÀ KIỂU THUỘC TÍNH

1.1.- Đối tượng dữ liệu (data object)

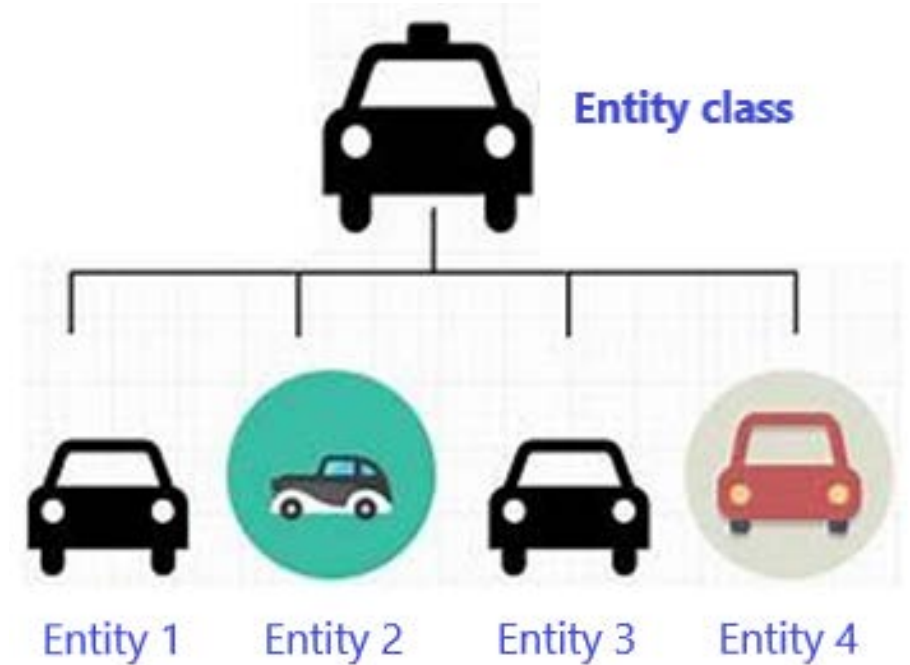
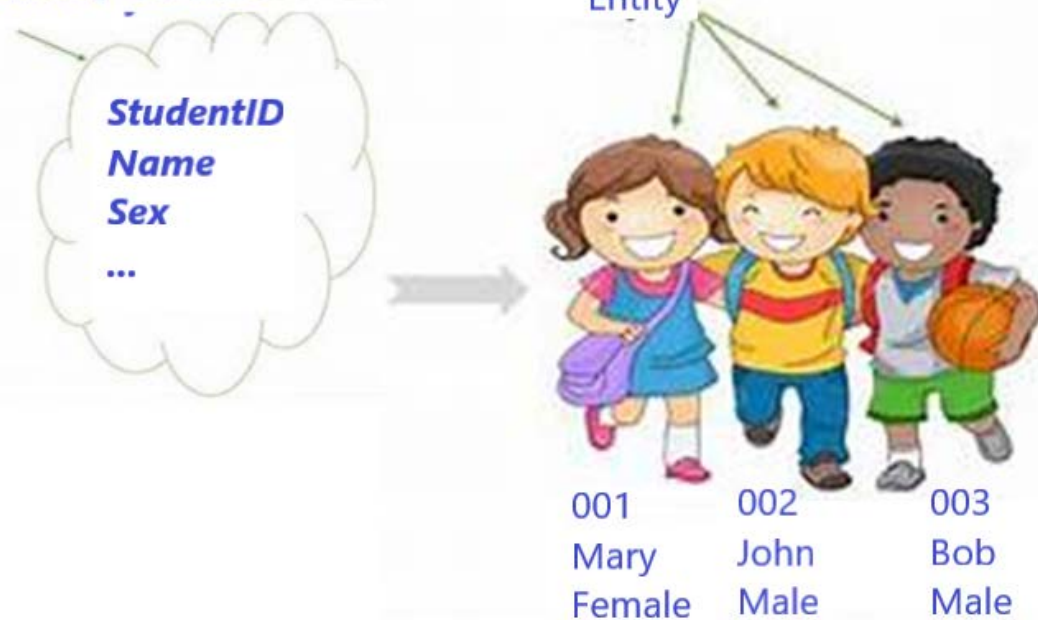
- Đối tượng dữ liệu cũng có thể được gọi là:
 - mẫu (samples)
 - đồ mẫu (examples)
 - thể hiện (instances)
 - điểm dữ liệu (data points)
 - đối tượng (objects).
- Các đối tượng dữ liệu thường được mô tả bằng các thuộc tính (attributes).

Object Attributes				
Color	Red Green Blue	Black White Purple	Yellow Other?	
Shape	Round 	Square 	Rectangular  Other?	
Size	LARGE MEDIUM SMALL Other?			
Number	1 One	2 Some	3 A Few	4 Many Other?
Use	▪ To eat ▪ To wear ▪ To play ▪ To write ▪ To cook Other?			

1.1.- Đối tượng dữ liệu (data object)

- Đối tượng dữ liệu đại diện cho một thực thể (entity) như trong cơ sở dữ liệu (CSDL) bán hàng, các đối tượng có thể là khách hàng, mặt hàng trong cửa hàng và doanh số bán hàng; trong cơ sở dữ liệu y tế, đối tượng có thể là bệnh nhân; trong cơ sở dữ liệu đại học, các đối tượng có thể là sinh viên, giáo sư và khóa học.

Student Entity Class



1.1.- Đối tượng dữ liệu (data object)

- Tập dữ liệu được tạo thành từ các đối tượng dữ liệu.
- Nếu các đối tượng dữ liệu được lưu trữ trong cơ sở dữ liệu thì chúng là các bộ dữ liệu. Nghĩa là, các hàng của cơ sở dữ liệu tương ứng với các đối tượng dữ liệu và các cột tương ứng với các thuộc tính.

<i>MSSV</i>	<i>Ten</i>	<i>Phai Nu</i>	<i>DiaChi</i>	<i>DienThoai</i>	<i>MaKhoa</i>
C0001F	Bùi Thúy An	1	223 Trần Hưng Đạo, HCM	38132202	CNTT
C0002M	Nguyễn Thanh Tùng	0	140 Cống Quỳnh, Sóc Trăng	38125678	CNTT
T0003M	Nguyễn Thành Long	0	112/4 Cống Quỳnh, HCM	0918345623	TOAN
C0004F	Hoàng Thị Hoa	1	90 Nguyễn Văn Cừ, HCM	38320123	CNTT
T0005M	Trần Hồng Sơn	0	54 Cao Thắng, Hà Nội	38345987	TOAN

1.2.- Thuộc tính (Attributes)

- Thuộc tính là một trường dữ liệu (data field), biểu thị một đặc tính hoặc tính năng của đối tượng dữ liệu.
- Các thuộc tính được dùng để mô tả một đối tượng, ví dụ: khách hàng có thể bao gồm ID khách hàng, tên và địa chỉ.
- Tập hợp các thuộc tính dùng để mô tả một đối tượng nhất định được gọi là vectơ thuộc tính (*attribute vector* hoặc vectơ đặc trưng - feature vector).

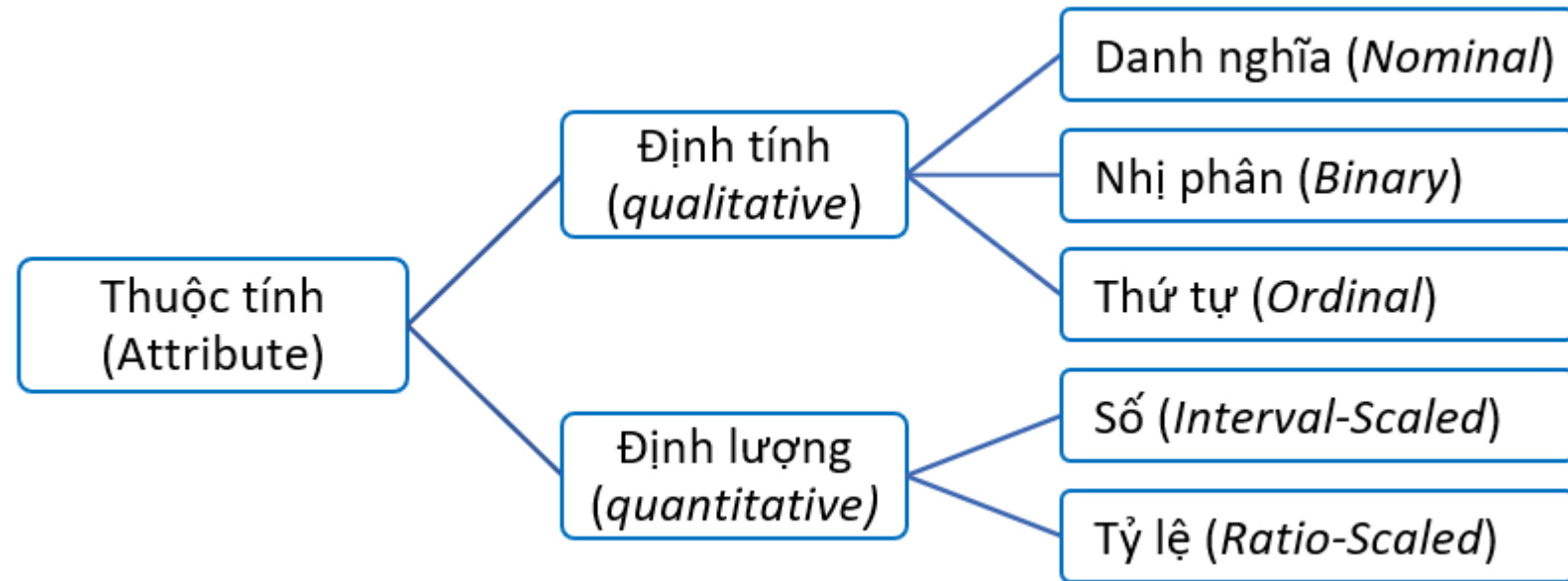
Attributes

<i>MSSV</i>	<i>Ten</i>	<i>Phai Nu</i>	<i>DiaChi</i>	<i>DienThoai</i>	<i>MaKhoa</i>
C0001F	Bùi Thúy An	1	223 Trần Hưng Đạo, HCM	38132202	CNTT
C0002M	Nguyễn Thanh Tùng	0	140 Cống Quỳnh, Sóc Trăng	38125678	CNTT
T0003M	Nguyễn Thành Long	0	112/4 Cống Quỳnh, HCM	0918345623	TOAN
C0004F	Hoàng Thị Hoa	1	90 Nguyễn Văn Cừ, HCM	38320123	CNTT
T0005M	Trần Hồng Sơn	0	54 Cao Thắng, Hà Nội	38345987	TOAN

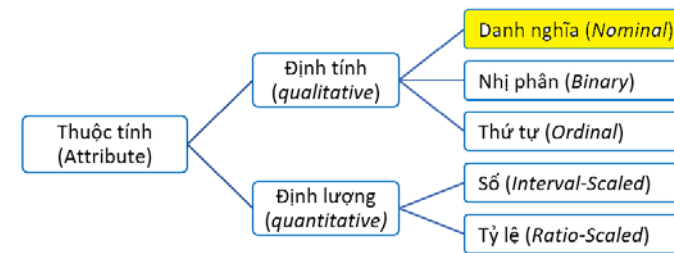
Attribute vector

1.2.- Thuộc tính (Attributes)

- Kiểu dữ liệu của thuộc tính được xác định bởi tập hợp các giá trị mà thuộc tính có thể có.
- Phân loại kiểu dữ liệu của thuộc tính:



1.3.- Thuộc tính định danh (Qualitative attributes)



1.3.1.- Thuộc tính danh nghĩa (Nominal attributes)

- Thuộc tính *Danh nghĩa* chứa các giá trị là ký hiệu hoặc tên của sự vật.
- Mỗi giá trị đại diện cho một số loại danh mục, mã hoặc trạng thái và do đó các thuộc tính danh nghĩa cũng được gọi là thuộc tính phân loại.
- Các giá trị này thuộc dạng này thường không quan tâm đến thứ tự giữa các giá trị.
- Trong khoa học máy tính, các giá trị còn được gọi là kiểu liệt kê (enumerations).
- Ví dụ: giá trị có thể có của 1 số thuộc tính như:
 - Màu tóc: đen, nâu, vàng, đỏ, nâu vàng, xám và trắng.
 - Tình trạng hôn nhân có thể mang các giá trị độc thân, đã kết hôn, đã ly hôn và góa bụa.
 - Nghề nghiệp: giáo viên, nha sĩ, lập trình viên, nông dân, v.v.

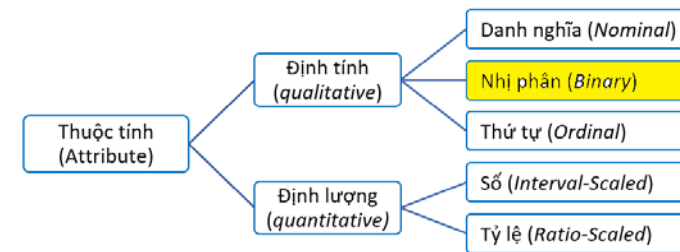
1.3.- Thuộc tính định danh (*Qualitative attributes*)

1.3.1- Thuộc tính danh nghĩa (*Nominal Attributes*)

- Giá trị của một thuộc tính danh nghĩa là các ký hiệu hoặc “tên của sự vật/hiện tượng”, nhưng vẫn có thể biểu diễn các ký hiệu hoặc “tên” đó bằng các con số. Ví dụ: với màu tóc, có thể gán mã 0 cho màu đen, 1 cho màu nâu, v.v.
- Mặc dù một thuộc tính danh nghĩa có thể có các số nguyên làm giá trị, nhưng nó không được coi là thuộc tính số vì các số nguyên không được sử dụng để định lượng. Tức là các phép toán trên giá trị của thuộc tính danh nghĩa là không có ý nghĩa. VD: Sẽ vô nghĩa khi trừ xếp loại học tập “*giỏi*” cho xếp loại “*trung bình*”.
- Vì các giá trị thuộc tính danh nghĩa không có bất kỳ thứ tự có ý nghĩa nào về chúng và không mang tính định lượng, nên việc tìm giá trị trung bình (*mean*) hoặc giá trị trung vị (*median*) cho một thuộc tính như vậy là vô nghĩa. Tuy nhiên, một điều đáng quan tâm là giá trị xuất hiện phổ biến nhất của thuộc tính. Giá trị này, được gọi là mode, là một trong những thước đo của xu hướng trung tâm (*measures of central tendency*).

1.3.- Thuộc tính định danh (Qualitative attributes)

1.3.2- Thuộc tính nhị phân (Nominal Attributes)



- Thuộc tính nhị phân là một thuộc tính danh nghĩa chỉ nhận 1 trong 2 giá trị (hoặc trạng thái) là không có (0, false, no) hoặc có (1, true, yes).
- Các thuộc tính nhị phân được gọi là Boolean nếu hai trạng thái tương ứng là đúng (true hay yes) và sai (false hay no).
- Ví dụ Với bệnh nhân,
 - Thuộc tính “có hút thuốc” = 1 cho biết bệnh nhân có hút thuốc, trong khi 0 chỉ ra rằng bệnh nhân không hút thuốc.
 - Thuộc tính “Loãng xương”= true cho biết bệnh nhân bị loãng xương và false cho biết là không bị.

1.3.- Thuộc tính định danh (*Qualitative attributes*)

1.3.2. - Thuộc tính nhị phân (*Nominal Attributes*)

- Phân loại thuộc tính nhị phân

- *Thuộc tính nhị phân có tính đối xứng:*

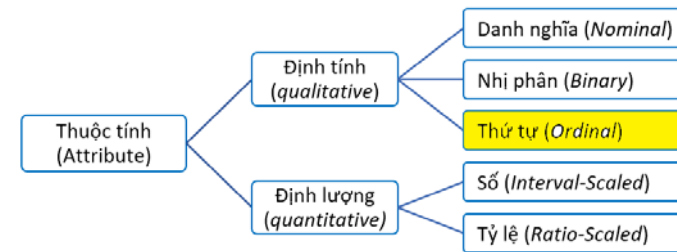
- Nếu cả hai trạng thái của nó có giá trị như nhau (có cùng trọng số); nghĩa là, không có ưu tiên nào về việc kết quả nào sẽ được mã hóa là 0 hoặc 1.
- Ví dụ thuộc tính giới tính có trạng thái nam và nữ.

- *Thuộc tính nhị phân là không đối xứng:*

- Nếu kết quả của các trạng thái không quan trọng như nhau, chẳng hạn như kết quả dương tính và âm tính của xét nghiệm y tế về HIV.
- Theo quy ước, ta mã hóa kết quả quan trọng nhất (thường là kết quả hiếm nhất) bằng 1
- Ví dụ: thuộc tính HIV=1 là dương tính và =0 là âm tính.

1.3.- Thuộc tính định danh (Qualitative attributes)

1.3.3.- Thuộc tính thứ tự (Ordinal attributes)



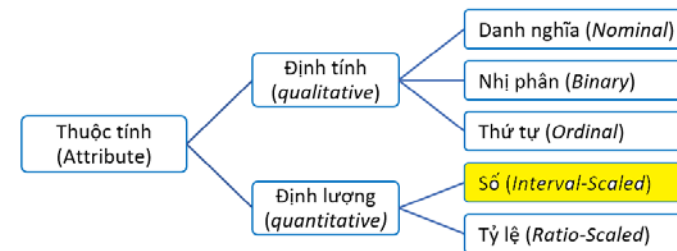
- Thuộc tính thứ tự là một thuộc tính có các giá trị có thể có thứ tự hoặc thứ hạng có ý nghĩa giữa chúng với nhau, nhưng độ lớn giữa các giá trị kế tiếp không được biết.
- Ví dụ:
 - Giả sử “size” đồ uống tại một cửa hàng có thể là: small, medium, large, big. Qua đó, ta không thể biết được độ chênh lệch giữa các “size” là bao nhiêu?
 - Giả sử có quy định về thuộc tính thứ tự cấp bậc gồm: binh nhì, binh nhất, hạ sĩ, trung sĩ, ... cho các cấp bậc trong quân đội. Cũng không cho biết được độ chênh lệch giữa các “cấp bậc” là bao nhiêu?
 - Sự hài lòng của khách hàng được chia thành các mức độ theo thứ tự sau: 0: rất không hài lòng, 1: hơi không hài lòng, 2: bình thường, 3: hài lòng và 4: rất hài lòng. => Không cho biết được độ chênh lệch giữa các “mức hài lòng”?

1.3.- Thuộc tính định danh (Qualitative attributes)

1.3.3.- Thuộc tính thứ tự (Ordinal attributes)

- Các thuộc tính thứ tự cũng có thể thu được từ việc rời rạc hóa các đại lượng số bằng cách chia phạm vi giá trị thành một số hữu hạn các danh mục được sắp xếp.
- Xu hướng trung tâm của thuộc tính thứ tự có thể được biểu thị bằng giá trị phổ biến nhất (mode) và trung vị (median) của nó (giá trị ở giữa trong một chuỗi có thứ tự), nhưng không thể xác định được giá trị trung bình.

1.4.- Thuộc tính định lượng (Quantitative attributes)



1.4.1.- Thuộc tính tỷ lệ theo khoảng (Interval-Scale attributes)

- Thuộc tính được chia tỷ lệ theo khoảng có kích thước bằng nhau.
- Các giá trị của các thuộc tính chia tỷ lệ theo khoảng có thứ tự và có thể dương, bằng 0 hoặc âm.
- Do các thuộc tính được chia tỷ lệ theo khoảng là số, nên ta có thể so sánh và định lượng sự khác biệt giữa các giá trị. Cụ thể là có thể tính các giá trị:
 - Trung bình (mean)
 - Trung vị (median)
 - Giá trị phổ biến (mode hay còn được gọi là giá trị có xu hướng trung tâm - measures of central tendency).

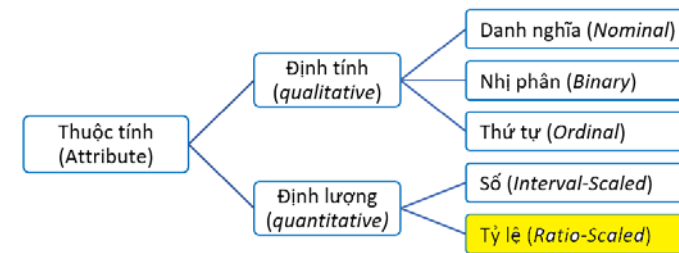
1.4.- Thuộc tính định lượng (Quantitative attributes)

1.4.1.- Thuộc tính tỷ lệ (Interval-Scale attributes)

- Ví dụ 1: về ngày dương lịch, năm 2002 và 2010 cách nhau 8 năm.
⇒ Lưu ý: Không có điểm 0 thực sự cho ngày dương lịch nên Năm 0 không tương ứng với thời điểm bắt đầu.
- Ví dụ 2: Thuộc tính nhiệt độ được chia tỷ lệ theo khoảng. Giả sử ta có giá trị nhiệt độ ngoài trời cho một số ngày khác nhau, trong đó mỗi ngày là một đối tượng.
 - Bằng cách sắp xếp các giá trị, có được thứ hạng của các đối tượng theo nhiệt độ.
 - Ngoài ra, có thể định lượng sự khác biệt giữa các giá trị. VD: nhiệt độ 20°C cao hơn 5 độ so với nhiệt độ 15°C .⇒ Lưu ý: Mặc dù có thể tính toán sự khác biệt giữa các giá trị nhiệt độ, nhưng ta không thể nói về một giá trị nhiệt độ là bội số của giá trị khác. VD không thể nói rằng 10°C ấm gấp đôi 5°C . Nghĩa là, *không thể nói về các giá trị dưới dạng tỷ lệ*.

1.4.- Thuộc tính định lượng (Quantitative attributes)

1.4.2.- Thuộc tính tỷ lệ (Ratio-Scale attributes)



- Thuộc tính tỷ lệ là thuộc tính số có điểm 0 vốn có. Nghĩa là:
 - Nếu một phép đo được chia tỷ lệ, ta có thể coi một giá trị là bội số (hoặc tỷ lệ) của một giá trị khác.
 - Do các giá trị được sắp xếp theo thứ tự nên có thể tính toán sự khác biệt giữa các giá trị cũng như giá trị trung bình (mean), trung vị (median) và giá trị phổ biến (mode).
- Ví dụ:
 - Các thuộc tính để đo tọa độ trọng lượng, chiều cao, tiền tệ, vĩ độ và kinh độ
 - Người A có 25\$, người B có 10\$, có thể nói A giàu hơn B 2.5 lần.

1.5.- Thuộc tính rời rạc và liên tục (Discrete versus Continuous Attributes)

- Các giải thuật phân loại (Classification algorithms) được phát triển từ lĩnh vực machine learning thường coi các thuộc tính là rời rạc hoặc liên tục (discrete or continuous). Mỗi loại có thể được xử lý khác nhau.
- Thuộc tính rời rạc có một tập hợp các giá trị hữu hạn hoặc vô hạn đếm được, các giá trị này có thể được biểu diễn thay thế dưới dạng số nguyên. Mỗi thuộc tính màu tóc, người hút thuốc, xét nghiệm y tế và kích cỡ đồ uống đều có một số giá trị hữu hạn và do đó chúng rời rạc. Lưu ý rằng các thuộc tính rời rạc có thể có các giá trị số, chẳng hạn như 0 và 1 cho thuộc tính nhị phân hoặc các giá trị từ 0 đến 110 cho thuộc tính tuổi. Một thuộc tính được coi là vô hạn nếu tập hợp các giá trị có thể là vô hạn nhưng các giá trị đó có thể được đặt ở dạng tương ứng một-một với các số tự nhiên. Ví dụ: thuộc tính ID khách hàng là vô hạn. Số lượng khách hàng có thể tăng lên vô tận, nhưng trên thực tế, tập hợp các giá trị thực tế có thể đếm được (trong đó các giá trị có thể được đặt ở dạng tương ứng 1-1 với tập hợp các số nguyên). Mã Zip là một ví dụ khác.

1.5.- Thuộc tính rời rạc và liên tục (Discrete versus Continuous Attributes)

- Các giải thuật phân loại (Classification algorithms) được phát triển từ lĩnh vực machine learning thường coi các thuộc tính là rời rạc hoặc liên tục. Mỗi loại có thể được xử lý khác nhau.
- *Thuộc tính liên tục*: theo nghĩa cổ điển, các giá trị liên tục là số thực, trong khi giá trị số có thể là số nguyên hoặc số thực. Trong thực tế, các giá trị thực được biểu diễn bằng một số hữu hạn các chữ số và thường được biểu diễn dưới dạng các biến có dấu chấm động.

1.5.- Thuộc tính rời rạc và liên tục (Discrete versus Continuous Attributes)

- Thuộc tính rời rạc:

- Có một tập hợp các giá trị hữu hạn hoặc vô hạn đếm được, các giá trị này có thể được biểu diễn thay thế dưới dạng số nguyên.
 - Ví dụ: Mỗi thuộc tính màu tóc, người hút thuốc, xét nghiệm y tế và kích cỡ đồ uống đều có một số giá trị hữu hạn và do đó chúng rời rạc.
 - Lưu ý rằng các thuộc tính rời rạc có thể có các giá trị số, chẳng hạn như 0 và 1 cho thuộc tính nhị phân hoặc các giá trị từ 0 đến 110 cho thuộc tính tuổi.
- Một thuộc tính được coi là vô hạn nếu tập hợp các giá trị có thể là vô hạn nhưng các giá trị đó có thể được đặt ở dạng tương ứng một-một với các số tự nhiên.

Ví dụ: thuộc tính ID khách hàng là vô hạn. Số lượng khách hàng có thể tăng lên vô tận, nhưng trên thực tế, tập hợp các giá trị thực tế có thể đếm được (trong đó các giá trị có thể được đặt ở dạng tương ứng 1-1 với tập hợp các số nguyên).

NỘI DUNG CHƯƠNG 3

1. Đối tượng dữ liệu và kiểu thuộc tính
2. Mô tả các thống kê cơ bản
3. Trực quan hóa dữ liệu
4. Đo lường sự tương đồng và khác biệt của dữ liệu
5. Tóm tắt



2. THỐNG KÊ MÔ TẢ (*Descriptive Statistics*)

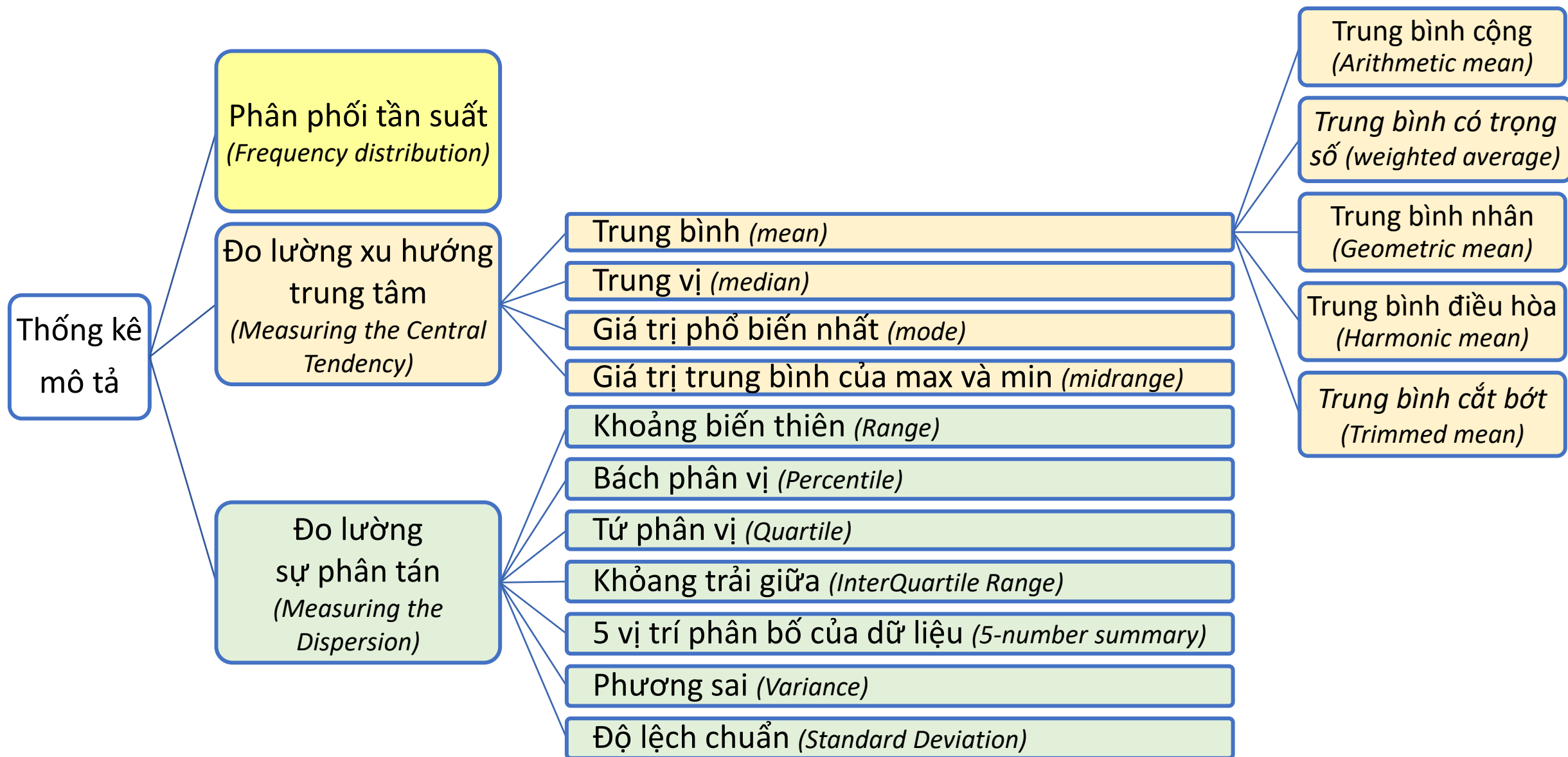
2.1. Giới thiệu

Thống kê mô tả là các hệ số mô tả ngắn gọn hay tóm tắt một tập dữ liệu nhất định, có thể là đại diện cho toàn bộ hoặc một mẫu của một tổng thể.

2.2. Phân loại

- Các thông số ***Đo lường xu hướng trung tâm*** mô tả vị trí trung tâm của phân phối tập dữ liệu.
- Các thông số ***đo lường biến động*** (hay các biện pháp ***đo lường sự phân tán***) hỗ trợ việc phân tích mức độ lan truyền trong phân phối của một tập dữ liệu.

2.2. Phân loại



2.3. Đo lường xu hướng trung tâm (*Measuring the Central Tendency*)

2.3.1. Trung bình (mean)

- Đại lượng trung bình thể hiện trung tâm về mặt giá trị của tập dữ liệu.
- Đại lượng trung bình gồm:
 - **Trung bình cộng** (*Arithmetic mean*)
 - Là thước đo phổ biến nhất và dễ hiểu nhất về xu hướng trung tâm trong tập dữ liệu.
 - Trung bình cộng bao gồm trung bình cộng đơn giản và trung bình cộng có trọng số.
 - **Trung bình có trọng số** (*weighted arithmetic mean or the weighted average*): khi mỗi giá trị x_i trong tập hợp có thể gắn với trọng số w_i (với $i = 1, \dots, N$). Các trọng số phản ánh tầm quan trọng hoặc tần suất xuất hiện gắn liền với các giá trị tương ứng của chúng.
 - **Trung bình nhân** (*Geometric mean*): (hay trung bình hình học), cho biết xu hướng trung tâm hoặc giá trị điển hình của một tập hợp số bằng cách sử dụng tích các giá trị của chúng.
 - **Trung bình điều hòa** (*Harmonic mean*): Thường được sử dụng để tìm giá trị trung bình của các quan sát được biểu diễn bởi tỉ số của hai giá trị có hai đơn vị đo khác nhau chẳng hạn như tốc độ di chuyển trung bình trong một khoảng thời gian.

2.3. Đo lường xu hướng trung tâm

2.3.1. Trung bình (mean)

- Cách tính:

- **Trung bình cộng** (*Arithmetic mean*)

- Trung bình cộng đơn giản:

- Công thức $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$

- Ví dụ $\bar{x} = \frac{7+5+2+6+9+7+8}{7} = \frac{44}{7} = 6.286$

- Trung bình cộng có trọng số (*Weighted arithmetic mean*)

- Công thức $\bar{x} = \frac{\sum_{i=1}^n x_i W_i}{n} = \frac{x_1 W_1 + x_2 W_2 + \dots + x_n W_n}{n}$

- Ví dụ $\bar{x} = \frac{(7 \times 4) + (5 \times 2) + (2 \times 3) + (6 \times 4) + (9 \times 4) + (7 \times 2) + (8 \times 3)}{7} = \frac{142}{7} = 20.286$

- Lưu ý:

- Giá trị của trung bình cộng dễ bị ảnh hưởng bởi các giá trị ngoại lệ và các phân phối bất đối xứng.
 - Không sử dụng đại lượng trung bình cộng đối với dữ liệu định danh.
 - Trung bình cộng hạn chế sử dụng với dữ liệu định lượng theo thang đo khoảng.

Trọng số (w)	Giá trị (x)
4	7
2	5
3	2
4	6
4	9
2	7
3	8

2.3. Đo lường xu hướng trung tâm

2.3.1. Trung bình (mean)

- Cách tính:

- **Trung bình nhân** (*Geometric mean*)

- Công thức $\bar{x} = (\prod_{i=1}^n x_i)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \dots x_n}$

- Ví dụ $\bar{x} = \sqrt[7]{7 * 5 * 2 * 6 * 9 * 7 * 8} = \sqrt[7]{211680} = 5.765161$

Trọng số (w)	Giá trị (x)
4	7
2	5
3	2
4	6
4	9
2	7
3	8

2.3. Đo lường xu hướng trung tâm

2.3.1. Trung bình (mean)

- Cách tính:

- **Trung bình điều hòa (Harmonic mean)**

○ Công thức
$$\bar{x} = n \times \left(\sum_{i=1}^n \frac{1}{x_i} \right)^{-1}$$

hoặc: là nghịch đảo của trung bình cộng:
$$\bar{x} = \left\{ \frac{1}{x_i} \right\}_{i=1}^n$$

○ Ví dụ
$$\bar{x} = 7 \times \left\{ \frac{1}{7} + \frac{1}{5} + \frac{1}{2} + \frac{1}{6} + \frac{1}{9} + \frac{1}{7} + \frac{1}{8} \right\}^{-1}$$

$$= 7 \times \left\{ \frac{360 + 504 + 1260 + 420 + 280 + 360 + 315}{2520} \right\}^{-1} = 7 \times \left(\frac{3499}{2520} \right)^{-1} = 7 \times \frac{2520}{3499} = 5.041$$

Trọng số (w)	Giá trị (x)
4	7
2	5
3	2
4	6
4	9
2	7
3	8

2.3. Đo lường xu hướng trung tâm

2.3.1. Trung bình (mean)

- Cách tính:

- **Trung bình cắt bớt (*Trimmed mean*)**

- ▣ Một vấn đề lớn với giá trị trung bình là ảnh hưởng với các giá trị ngoại lệ (outlier). Ví dụ:
 - Mức lương trung bình tại một công ty có thể bị đẩy lên đáng kể do mức lương trung bình của một số nhà quản lý được trả lương cao.
 - Hoặc điểm trung bình của một lớp có thể bị kéo xuống khá nhiều bởi một vài điểm rất thấp.
- ▣ Để bù lại hiệu ứng do một số lượng nhỏ các giá trị cực trị gây ra, thay vào đó, ta có thể sử dụng giá trị trung bình đã được cắt bớt (*trimmed mean*), là giá trị trung bình thu được sau khi loại bỏ các giá trị ở mức cực cao và cực thấp.

2.3. Đo lường xu hướng trung tâm

2.3.1. Trung bình (mean)

- Cách tính:

- **Trung bình cắt bớt** (*Trimmed mean*)

▣ Ví dụ: Giả sử có các giá trị sau về tiền lương (đơn vị tính là 1.000\$), được hiển thị theo thứ tự tăng dần: 30, 36, 48, 50, 52, 52, 56, 60, 63, 70, 70, 110.

Tính *trimmed mean* như sau: loại bỏ 2 giá trị lớn và nhỏ nhất (là 110 và 30), sau đó mới tính bình quân:

$$\bar{x} = \frac{36 + 48 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70}{10} = \frac{557}{10} = 55.7$$

Trong khi nếu tính trung bình mean theo cách thông thường thì

$$\bar{x} = \frac{30 + 36 + 48 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12} = \frac{697}{12} = 58.083$$

2.3. Đo lường xu hướng trung tâm

2.3.2. Trung vị (median)

- Trung vị là một số tách giữa nửa lớn hơn và nửa bé hơn của một mẫu, một quần thể, hay một phân bố xác suất. Trung vị là giá trị giữa, có nghĩa $\frac{1}{2}$ quan sát sẽ có các giá trị nhỏ hơn hay bằng số trung vị, và $\frac{1}{2}$ quan sát sẽ có giá trị bằng hoặc lớn hơn số trung vị.
- Với tập dữ liệu có số lượng 1 quan sát n (đã được sắp xếp), khi:
 - n là số lẻ: quan sát ở vị trí thứ $[(n+1)/2]$ là số trung vị.

VD: $\bar{x} = x_4 = 7$

1	2	3	4	5	6	7
2	5	6	7	7	8	9

- n là số chẵn: số trung vị là giá trị trung bình cộng của 2 quan sát nằm ở vị trí $n/2$ và $[(n+2)/2]$

VD: $\bar{x} = x_3 + x_4 = 6 + 7 = 6.5$

1	2	3	4	5	6
2	5	6	7	7	8

Giá trị (x)
7
5
2
6
9
7
8

2.3. Đo lường xu hướng trung tâm

2.3.2. Trung vị (median)

- Một số lưu ý:
 - Giá trị của trung vị không chịu ảnh hưởng của các giá trị ngoại lệ và dễ tính toán.
 - Không thể dùng trung vị để dự đoán vì không chính xác bằng trung bình (mean).
 - Trung vị thường được dùng để thay thế hoặc bổ sung nhằm điều chỉnh 1 số hạn chế khi sử dụng giá trị trung bình như trường hợp tập dữ liệu bị lệch (không đối xứng), trung vị là thước đo tốt hơn về tâm của dữ liệu.

2.3. Đo lường xu hướng trung tâm

2.3.3. Giá trị phổ biến nhất (mode)

- Giới thiệu

- *Mode* là trung tâm về mức độ tập trung dữ liệu.
- *Mode* là giá trị xuất hiện nhiều lần nhất trong tập dữ liệu.
- Một tập dữ liệu có thể có 1, 2 hoặc 3 mode và cũng có thể không có mode nào.

- Cách tính: Đếm số lần xuất hiện của các giá trị, giá trị xuất hiện nhiều nhất chính là số mode.

- Ví dụ 1: trong dữ liệu ở bảng bên có giá trị là 7 xuất hiện nhiều lần nhất (2 lần) nên chỉ có 1 mode là 7 (*unimodal*).

- Ví dụ 2: dữ liệu trong hình sau là *bimodal* với hai *mode* là 52.000\$ và 70.000\$.

1	2	3	4	5	6	7	8	9	10	11	12
30	36	48	50	52	52	56	60	63	70	70	110

Giá trị (x)
7
5
2
6
9
7
8

2.3. Đo lường xu hướng trung tâm

2.3.3. Giá trị phổ biến nhất (mode)

- *Một số lưu ý:*

- *Mode* là đại lượng thống kê mô tả duy nhất có thể vận dụng cho dữ liệu định tính.
- *Mode* không bị ảnh hưởng bởi các giá trị ngoại lệ.
- *Mode* chỉ ổn định khi lượng giá trị nhiều và sẽ khó xác định rõ nếu dữ liệu chỉ có một số ít giá trị.

2.3. Đo lường xu hướng trung tâm

2.3.4. Midrange

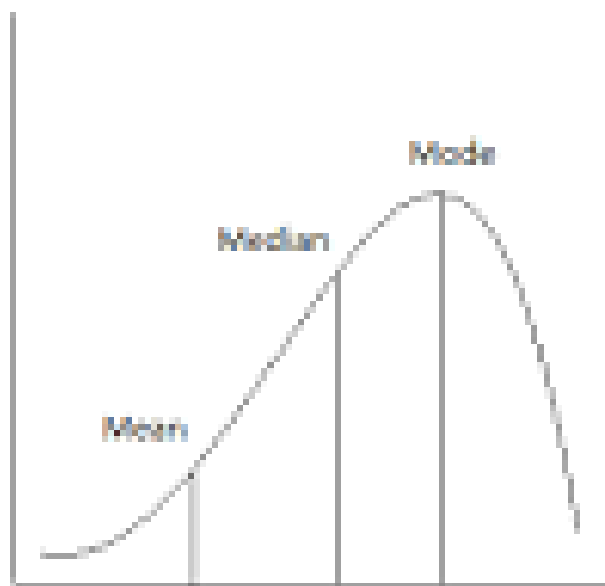
- Là giá trị trung bình của 2 giá trị lớn nhất và nhỏ nhất trong tập dữ liệu.
- Midrange cũng có thể được sử dụng để đánh giá xu hướng trung tâm của tập dữ liệu số.
- Phép đo này dễ dàng tính toán bằng cách sử dụng các hàm tổng hợp SQL là tính trung bình giá trị của 2 hàm max() và min().
- Ví dụ: với dãy số liệu sau:

1	2	3	4	5	6	7	8	9	10	11	12
30	36	48	50	52	52	56	60	63	70	70	110

$$\text{midrange} = \frac{30.000 + 110.000}{2} = 70.000$$

2.3. Đo lường xu hướng trung tâm

2.3.5. Sự liên hệ giữa mean – median - mode bằng đồ thị



Negatively skewed data
(Left skew - dữ liệu lệch âm)
 $\text{mean} < \text{median} < \text{mode}$



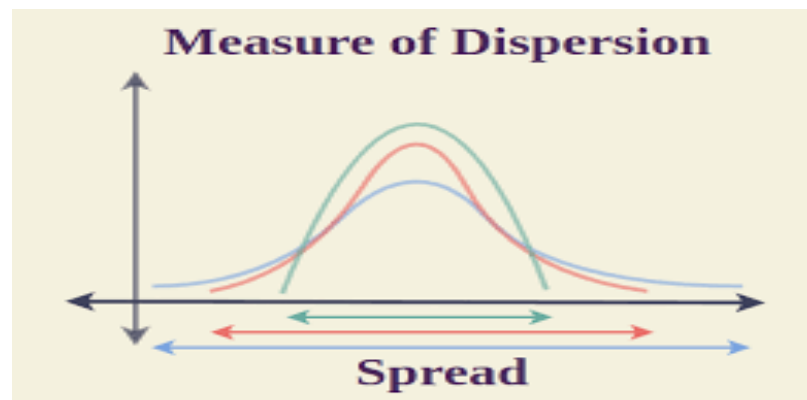
Normal Distribution
(dữ liệu cân đối)
 $\text{median} = \text{mean} = \text{mode}$



Positively skewed data
(Right skew - dữ liệu lệch dương)
 $\text{mode} < \text{median} < \text{mean}$

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

- Các thông số đo lường biến động (hay các biện pháp đo lường sự phân tán - *dispersion*): hỗ trợ việc phân tích mức độ lan truyền trong phân phối của một tập dữ liệu. Ví dụ, trong khi các thông số Đo lường xu hướng trung tâm có thể cung cấp mức trung bình của tập dữ liệu, nó lại không mô tả cách dữ liệu được phân phối như thế nào trong tập hợp đó.
- Vì vậy, mặc dù bình quân của dữ liệu có thể là 65 trong 100, vẫn có thể có các điểm dữ liệu ở điểm 1 và 100 trong tập dữ liệu. Các thông số đo lường biến động giúp xác định điều này bằng cách mô tả hình dạng và mức độ phân tán của tập dữ liệu.



2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.1. Khoảng biến thiên (Range)

- Là đại lượng đo mức độ trải dài của một tập dữ liệu nhất định từ giá trị nhỏ nhất đến giá trị lớn nhất.
- Công thức tính khoảng biến thiên: **range** = x_{\max} - x_{\min}
Trong đó:
 - x_{\max} là giá trị lớn nhất
 - x_{\min} là giá trị nhỏ nhất
- Ví dụ: với tập dữ liệu $X=\{7, 5, 2, 9, 7, 8\}$. Ta có $x_{\max}=9$ và $x_{\min}=2$
 \Rightarrow Khoảng biến thiên R là: $9 - 2 = 7$
- Range được sử dụng trong rất nhiều tình huống, như:
 - Tìm ra sự phân tán điểm kiểm tra trong một lớp học.
 - Xác định phạm vi giá cả của một dịch vụ,
 - Xác định nhiệt độ khoảng chênh lệch nhiệt độ trong 1 ngày tại 1 vùng, ...

Giá trị (x)
7
5
2
6
9
7
8

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.1. Khoảng biến thiên (Range)

- Một số đặc điểm của range:
 - Là đại lượng dễ hiểu và dễ tính toán.
 - Do chỉ sử dụng 2 giá trị MAX và MIN của tập dữ liệu nên không thể dùng để đánh giá sự phân tán của tập dữ liệu.
 - Là thang đo tương đối tốt đối với các bộ dữ liệu nhỏ, nhưng độ tin cậy sẽ ít đi khi áp dụng với các bộ dữ liệu lớn có độ dàn trải của dữ liệu cũng lớn.
 - Dữ liệu ngoại lệ (*Outliers*):
 - Là một điểm dữ liệu có sự khác biệt đáng kể so với các quan sát khác. Dữ liệu ngoại lệ có thể xuất hiện do sự thay đổi thang đo hoặc do lỗi từ dữ liệu thu thập.
 - Một giá trị ngoại lệ có thể gây ra vấn đề nghiêm trọng trong quá trình phân tích dữ liệu. Thông thường dữ liệu ngoại lệ sẽ bị loại khỏi tập dữ liệu.
 - Do đó, không nên sử dụng đại lượng range đối với các bộ dữ liệu có giá trị ngoại lệ.

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.2. Bách phân vị (*Percentile*)

- Là đại lượng dùng để ước tính tỷ lệ dữ liệu trong một tập số liệu rơi vào vùng cao hơn hoặc thấp hơn so với một giá trị cho trước. Bách phân vị chia dữ liệu có thứ tự theo hàng trăm.
- Ví dụ: cho phân vị thứ $p=85 \in [0;100]$ và giá trị v_p (20) tại vị trí p thì:
 - Có ít nhất $p\%$ các quan sát có giá trị $\leq v_p$, tức là có 85% các quan sát có giá trị ≤ 20
 - Có ít nhất $(100-p)\%$ các quan sát có giá trị $\geq v_p$, và có 15% ($=100-85$) các quan sát có giá trị > 20

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.2. Bách phân vị (Percentile)

- Xác định giá trị của phân vị thứ p :

- B1. Sắp xếp dữ liệu theo thứ tự từ nhỏ nhất đến lớn nhất.

- B2. Tính chỉ số i :

$$i = \frac{p \times (n+1)}{100}$$

Trong đó:

- i là vị trí của giá trị dữ liệu tại phân vị thứ p

- p là phân vị thứ p

- n là tổng số quan sát

- B3. Xác định giá trị vp

- Nếu i LÀ số nguyên thì phân vị thứ p là giá trị dữ liệu ở vị trí thứ i trong tập dữ liệu.

- Nếu i KHÔNG phải là số nguyên thì làm tròn i lên và làm tròn i xuống số nguyên gần nhất, sau đó tính trung bình hai giá trị dữ liệu ở hai vị trí này trong tập dữ liệu

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.2. Bách phân vị (Percentile)

- Ví dụ: Một tập dữ liệu A gồm điểm số của 29 học viên trong 1 lớp học theo thang điểm từ 0-100. Dữ liệu đã được sắp thứ tự tăng dần).

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
18	21	22	25	26	27	29	30	31	33	36	37	41	42	47	52	55	57	58	62	64	67	69	71	72	73	74	76	77

• Tìm phân vị thứ 70?

□ **B2.** Tính chỉ số i :
$$i = \frac{p \times (n+1)}{100} = \frac{70 \times (29+1)}{100} = \frac{2100}{100} = 21$$

- **B3.** Xác định giá trị v_p : Vì $i_{70}=21$ là số nguyên nên phân vị thứ 70 có giá trị là $A[21]=64$.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
18	21	22	25	26	27	29	30	31	33	36	37	41	42	47	52	55	57	58	62	64	67	69	71	72	73	74	76	77

- Có thể kết luận 70% học viên có điểm thấp hơn 64 và 30% có điểm trên 64.

• Tìm phân vị thứ 81?

□ **B2.** Tính chỉ số i :
$$i = \frac{p \times (n+1)}{100} = \frac{81 \times (29+1)}{100} = \frac{2430}{100} = 24.3$$

- **B3.** Xác định giá trị v_p : Vì $i_{81}=24.3$ là số lẻ nên phân vị thứ 81 có giá trị là $(A[24] + A[25])/2 = (71 + 72)/2 = 71.5$.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
18	21	22	25	26	27	29	30	31	33	36	37	41	42	47	52	55	57	58	62	64	67	69	71	72	73	74	76	77

- Có thể kết luận 81% học viên có điểm thấp hơn 71.5 và 19% có điểm trên 71.5

2.4. Đo lường sự phân tán của dữ liệu (**Measuring the Dispersion**)

2.4.3. **Tứ phân vị** (*Quartile*)

- Tứ phân vị (Quartile) là một trường hợp đặc biệt của bách phân vị. Tứ phân vị có 3 giá trị, đó là tứ phân vị thứ nhất (Q1), thứ hai (Q2), và thứ ba (Q3). Ba giá trị này chia một tập hợp dữ liệu đã sắp xếp theo thứ tự thành 4 phần có số lượng quan sát đều nhau.
- Cách xác định giá trị các tứ phân vị:
 - Tứ phân vị thứ nhất Q1 bằng trung vị phần dưới, tương đương với bách phân vị thứ 25.
 - Tứ phân vị thứ hai Q2 chính bằng giá trị trung vị, tương đương với bách phân vị thứ 50.
 - Tứ phân vị thứ ba Q3 bằng trung vị phần trên, tương đương với bách phân vị thứ 75.

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.3. Tứ phân vị (Quartile)

- Ví dụ xác định giá trị của $Q1$, $Q2$, $Q3$ với tập dữ liệu được cho như sau:

$X = \{1; 11,5; 6; 7,2; 4; 8; 9; 10; 6,8; 8,3; 2; 2; 10; 1\}$.

- Sắp xếp lại tập X theo thứ tự tăng dần: $X = \{1; 1; 2; 2; 4; 6; 6,8; 7,2; 8; 8,3; 9; 10; 10; 11,5\}$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
X	1	1	2	2	4	6	6,8	7,2	8	8,3	9	10	10	11,5
							Q2↑							
	X1							X2						
	1	1	2	2	4	6	6,8	7,2	8	8,3	9	10	10	11,5
							Q1↑							
									Q3↑					

- **Q2**: Tập dữ liệu có 14 quan sát, giá trị trung vị (median) nằm giữa giá trị thứ 7 (6,8) và giá trị thứ 8 (7,2). Giá trị trung vị là trung bình cộng của 2 giá trị này $\Rightarrow Q2 = (6,8 + 7,2) / 2 = 7$
- **Q1**: là giá trị giữa của nửa dưới dữ liệu tương ứng với tập dữ liệu $X_1 = \{1; 1; 2; 2; 4; 6; 6,8\}$. Tập X_1 có 7 giá trị, do đó giá trị trung vị của tập dữ liệu X_1 là 2. $\Rightarrow Q1 = 2$
- **Q3**: là giá trị nửa trên của dữ liệu tương ứng với tập dữ liệu $X_2 = \{7,2; 8; 8,3; 9; 10; 10; 11,5\}$. Tập X_2 có 7 giá trị, do đó giá trị trung vị của tập dữ liệu X_2 là 9. $\Rightarrow Q3 = 9$

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.3. Tứ phân vị (Quartile)

- Ví dụ

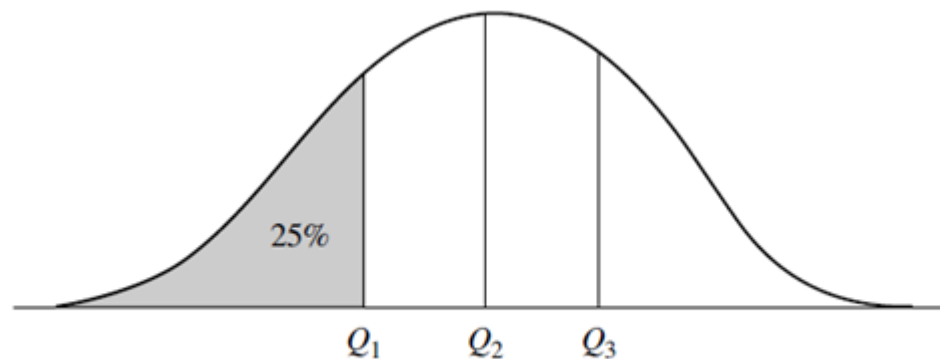
- Kết luận: $\frac{1}{4}$ tập dữ liệu có giá trị ≤ 2 , $\frac{3}{4}$ tập dữ liệu có giá trị ≥ 2 . Tương tự kết luận với Q2 và Q3.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
X	1	1	2	2	4	6	6,8	7,2	8	8,3	9	10	10	11,5
							Q2↑							
	X1						X2							
	1	1	2	2	4	6	6,8	7,2	8	8,3	9	10	10	11,5
	Q1↑						Q3↑							

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.4. Quantile

- Là các điểm cắt/điểm mốc chia một khoảng phân phối xác suất ra thành các khoảng với xác suất giống nhau. Như vậy, quartile và percentile là các trường hợp đặc biệt của quantile.
- Giả sử dữ liệu của thuộc tính X được sắp xếp theo thứ tự số tăng dần. Có thể chọn một số điểm dữ liệu nhất định để phân chia phân phối dữ liệu thành các tập hợp liên tiếp có kích thước bằng nhau (tương đối vì sẽ có những dữ liệu không thể chia đều nhau). Những điểm dữ liệu này được gọi là phân vị (quantiles).
- Phân vị thứ k (k^{th} q-quantiles) cho một phân bố dữ liệu nhất định là giá trị x sao cho có nhiều nhất k/q của các giá trị dữ liệu nhỏ hơn x và có nhiều nhất $(q - k)/q$ của các giá trị dữ liệu lớn hơn x , trong đó k là số nguyên sao cho $0 < k < q$. Có $q - 1$ q-phân vị ($q-1$ q-quantiles).



3-quantiles: Q_1 , Q_2 , Q_3

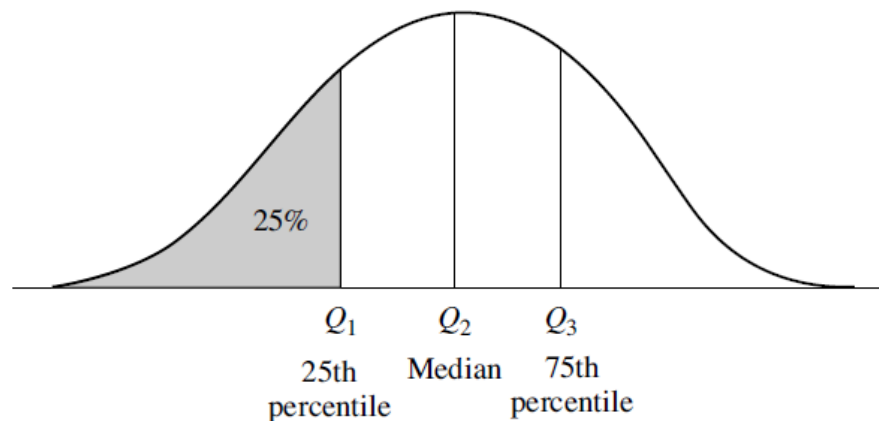
2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.5. 2-quantiles

- Phân vị 2 (*2-quantile*) là điểm dữ liệu phân chia nửa dưới và nửa trên của phân phối dữ liệu. *2-quantile* tương ứng với trung vị.

2.4.6. 4-quantiles (hay *Quartiles*)

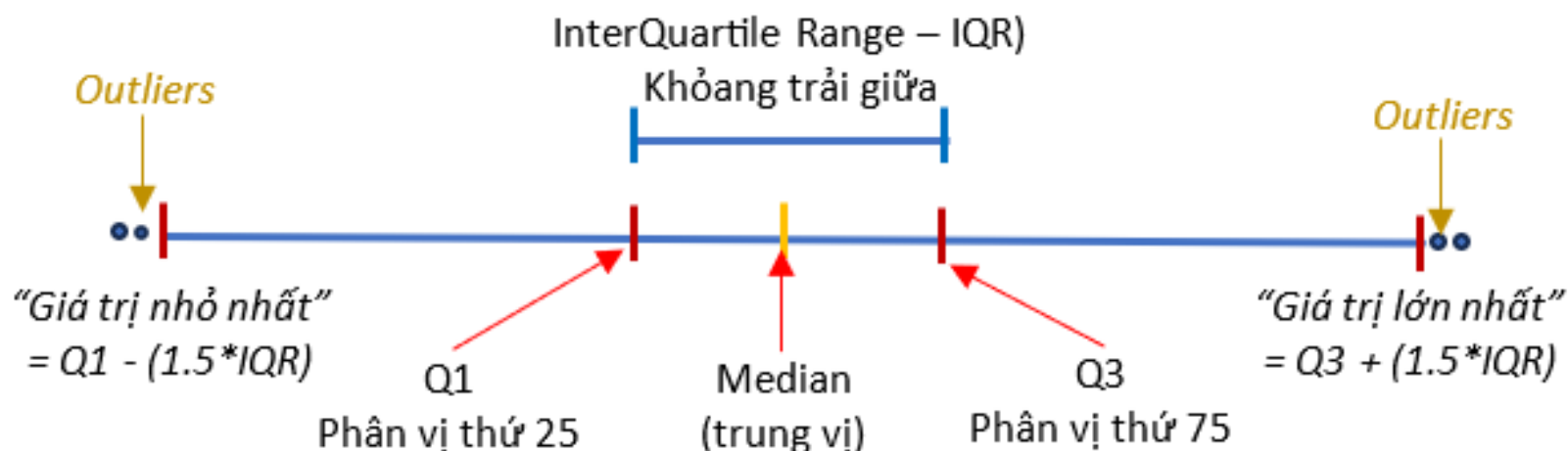
- Phân vị 4 (*4-quantiles*) là ba điểm dữ liệu chia phân phối dữ liệu thành bốn phần bằng nhau; mỗi phần đại diện cho một phần tư phân phối dữ liệu. Chúng thường được gọi là tứ phân vị (quartiles).



2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.7. Khoảng trải giữa (InterQuartile Range)

- Khoảng trải giữa (*InterQuartile Range* – *IQR*) hay còn gọi là khoảng tứ phân vị của tập dữ liệu. Khoảng trải giữa là một con số cho biết mức độ lan truyền của nửa giữa hoặc 50% phần giữa của tập dữ liệu. IQR thường được sử dụng thay cho khoảng biến thiên (*Range*) vì nó loại trừ hầu hết giá trị bất thường hay giá trị ngoại lệ (*Outliers*) của dữ liệu.
- Công thức tính IQR: $IQR = Q3 - Q1$

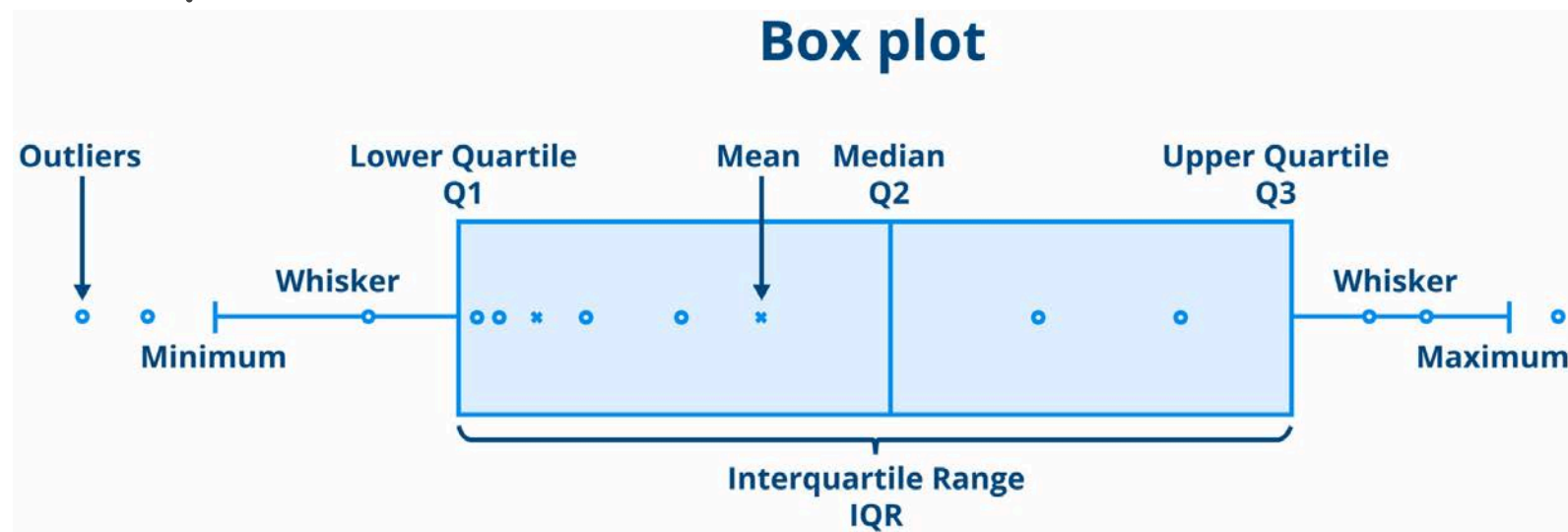


2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.7. Khoảng trải giữa (InterQuartile Range)

- Giá trị ngoại lệ:

- IQR có thể giúp xác định các giá trị ngoại lệ.
- Một giá trị bị nghi ngờ là một giá trị ngoại lệ nếu nó nhỏ hơn $1,5 \times \text{IQR}$ dưới phần tư đầu tiên ($Q1 - 1,5 \times \text{IQR}$) hoặc lớn hơn $(1,5 \times \text{IQR})$ trên phần tư thứ ba ($Q3 + 1,5 \times \text{IQR}$).
- Các giá trị ngoại lệ luôn yêu cầu việc rà soát, kiểm tra lại dữ liệu. Những điểm dữ liệu đặc biệt này có thể do lỗi hoặc do sự bất thường trong dữ liệu nhưng cũng có thể là chìa khóa để hiểu dữ liệu.

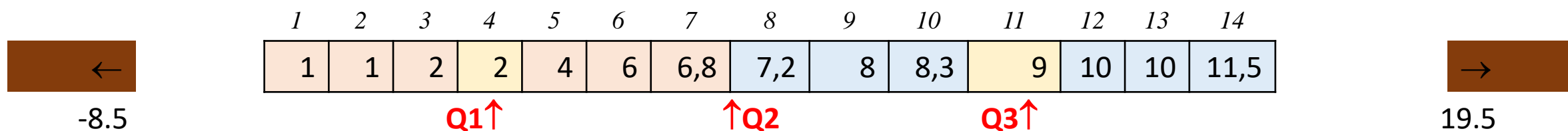


2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.7. Khoảng trải giữa (InterQuartile Range)

- Giá trị ngoại lệ:

- Ví dụ 1: cho 14 giá sau (đã được sắp xếp tăng dần)



- $Q_2 = (6,8 + 7,2) / 2 = 7.0$
- $Q_1 = 2$
- $Q_3 = 9$
- $IQR = Q_3 - Q_1 = 9 - 2 = 7$
- Lower Outliers = $Q_1 - 1.5 * IQR = 2 - 10.5 = -8.5$
- Upper Outliers = $Q_3 + 1.5 * IQR = 9 + 10.5 = 19.5$

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.7. InterQuartile Range (*IQR*)

- Ví dụ 1-12.- cho 12 giá trị quan sát, đã được sắp xếp theo thứ tự tăng dần:
30, 36, 48, 50, 52, 52, 56, 60, 63, 70, 70, 110.

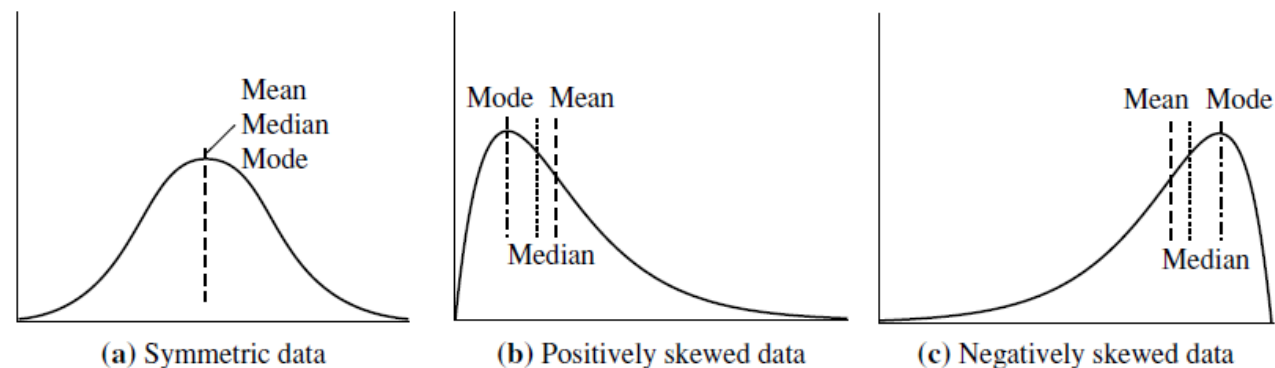
1	2	3	4	5	6	7	8	9	10	11	12
30	36	48	50	52	52	56	60	63	70	70	110
		Q ₁			Q ₂			Q ₃			

Các phần tư cho dữ liệu này lần lượt là giá trị thứ 3, 6 và 9 trong danh sách được sắp xếp. Do đó, $Q_1 = 48$ và Q_3 là 63. Do đó, **$IQR = 63 - 48 = 15$** .

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.8. Five-Number Summary (5 vị trí của phân bố của dữ liệu)

- Không có thước đo mức độ chênh lệch bằng số nào (ví dụ: IQR) để mô tả sự phân bố dữ liệu là đối xứng hay bị lệch. Hãy xem lại sự phân bố dữ liệu đối xứng và lệch của hình sau.

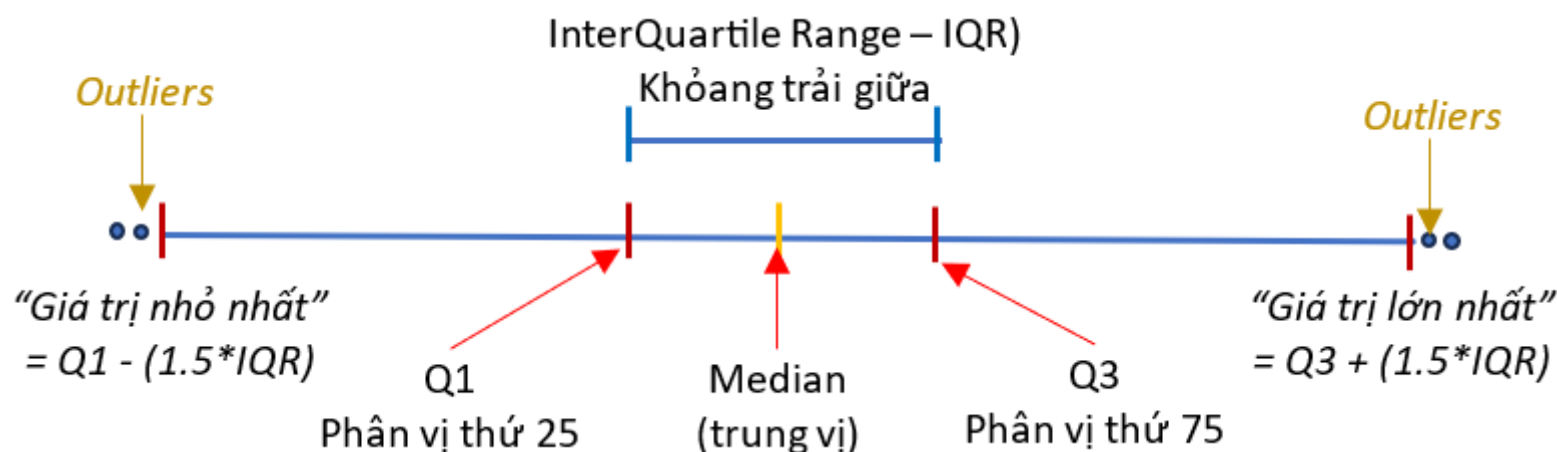


- Trong phân bố đối xứng, trung vị (và các thước đo khác của xu hướng trung tâm) chia dữ liệu thành hai nửa có kích thước bằng nhau. Điều này không xảy ra đối với các bản phân phối bị lệch. Do đó, sẽ có nhiều thông tin hơn nếu cung cấp cả hai tứ phân vị Q1 và Q3, cùng với giá trị trung vị. Nguyên tắc chung để xác định các ngoại lệ bị nghi ngờ là chọn ra các giá trị giảm ít nhất $1,5 \times \text{IQR}$ trên tứ phân vị thứ ba hoặc dưới tứ phân vị thứ nhất.

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.8. Five-Number Summary (5 vị trí của phân bố của dữ liệu)

- Vì Q_1 , trung vị và Q_3 cùng nhau không chứa thông tin về điểm cuối (ví dụ: đuôi) của dữ liệu nên có thể thu được bản tóm tắt đầy đủ hơn về hình dạng của phân phối bằng cách cung cấp các giá trị dữ liệu thấp nhất và cao nhất.



- Điều này được gọi là 5 vị trí phân bố của dữ liệu (five-number summary). Tóm tắt năm số của một phân bố bao gồm trung vị (Q_2), tứ phân vị Q_1 và Q_3 , và các quan sát riêng lẻ nhỏ nhất và lớn nhất, được viết theo thứ tự nhỏ nhất (*Minimum*), Q_1 , Trung vị (*Median*), Q_3 , lớn nhất (*Maximum*).

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.9. Phương sai (Variance)

- Phương sai là thước đo độ biến thiên của các giá trị xung quanh giá trị trung bình số học của chúng.
- Khi giá trị của phương sai càng lớn cho biết dữ liệu có sự dàn trải rộng hơn, ngược lại cho biết dữ liệu được gom về gần mức trung bình hơn.
- Ký hiệu của phương sai là σ^2
- Phương sai của N quan sát, x_1, x_2, \dots, x_N , đối với thuộc tính số X là:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \bar{x}^2$$

Trong đó:

- \bar{x} là giá trị trung bình (mean) của các quan sát.
- σ là độ lệch chuẩn của các quan sát và là căn bậc hai của phương sai σ^2 .

2.4. Đo lường sự phân tán của dữ liệu (*Measuring the Dispersion*)

2.4.9. Phương sai (*Variance*)

- **Ví dụ:** Cho dữ liệu gồm 12 giá trị và đã được sắp theo thứ tự tăng dần: 30, 36, 48, 50, 52, 52, 56, 60, 63, 70, 70, 110. Ta có $\bar{x} = 58.083$.

Xác định phương sai:

$$\sigma^2 = \frac{1}{12} \left(\sum_{i=1}^N x_i^2 \right) - \bar{x}^2 = \frac{1}{12} (30^2 + 36^2 + 48^2 + \dots + 110^2) - 58.083^2 \approx 377.4484$$

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.9. Phương sai (Variance)

- **Ví dụ 1:** Cho mẫu dữ liệu về thời gian (giây) chạy cự ly 500m và 1500m của một nhóm gồm 5 người như sau, so sánh phương sai chạy của 2 cự ly: $T_{500} = \{55.2, 58.8, 62.4, 54, 59.4\}$
 $T_{1500} = \{271.2, 261, 276, 282, 270\}$

- Tính giá trị trung bình: $\bar{x}_{500} = \frac{55.2+58.8+62.4+54+59.4}{5} = 57.96$

$$\bar{x}_{1500} = \frac{271.2+261+276+282+270}{5} = 272.04$$

- Tính phương sai: $S_{500}^2 = \frac{1}{5-1} ((55.2 - 57.96)^2 + (58.8 - 57.96)^2 + (62.4 - 57.96)^2 + (54 - 57.96)^2 + (59.4 - 57.96)^2) = 11.45$

$$S_{1500}^2 = \frac{1}{5-1} ((271.2 - 272.04)^2 + (261 - 272.04)^2 + (276 - 272.04)^2 + (282 - 272.04)^2 + (270 - 272.04)^2) = 60.41$$

- So sánh: $S_{1500}^2 > S_{500}^2$ cho thấy T1500 có sự biến động mạnh hơn, tức dữ liệu có sự dàn trải rộng hơn giữa thời gian chạy của 5 người này.

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.10. Độ lệch chuẩn (*Standard deviation*)

- Độ lệch chuẩn (Standard deviation) là thước đo độ phân tán của các giá trị trong một tập dữ liệu đã cho từ giá trị trung bình của chúng. Nó cho biết trung bình mỗi giá trị nằm bao xa so với giá trị trung bình.
- Ký hiệu: σ
- Độ lệch chuẩn là căn bậc hai của phương sai.

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.10. Độ lệch chuẩn (*Standard deviation*)

- **Ví dụ:** Cho dữ liệu gồm 12 giá trị và đã được sắp theo thứ tự tăng dần: 30, 36, 48, 50, 52, 52, 56, 60, 63, 70, 70, 110. Ta có $\bar{x} = 58.083$.

Xác định phương sai:

$$\sigma^2 = \frac{1}{12} \left(\sum_{i=1}^N x_i^2 \right) - \bar{x}^2 = \frac{1}{12} (30^2 + 36^2 + 48^2 + \dots + 110^2) - 58.083^2 \approx 377.4484$$

Xác định độ lệch chuẩn

$$\sigma = \sqrt{377.4484} = 19.428$$

2.4. Đo lường sự phân tán của dữ liệu (Measuring the Dispersion)

2.4.6. Độ lệch chuẩn (Standard deviation)

- **Ví dụ:** Cho mẫu dữ liệu về thời gian (giây) chạy cự ly 500m và 1500m của một nhóm gồm 5 người như sau, so sánh phương sai chạy của 2 cự ly:

$$T_{500} = \{55.2, 58.8, 62.4, 54, 59.4\}$$

$$T_{1500} = \{271.2, 261, 276, 282, 270\}$$

- Tính giá trị trung bình: $\bar{x}_{500} = \frac{55.2+58.8+62.4+54+59.4}{5} = 57.96$

$$\bar{x}_{1500} = \frac{271.2+261+276+282+270}{5} = 272.04$$

- Tính phương sai: $S_{500} =$

$$\sqrt{\frac{1}{5-1} ((55.2 - 57.96)^2 + (58.8 - 57.96)^2 + (62.4 - 57.96)^2 + (54 - 57.96)^2 + (59.4 - 57.96)^2)} = 3.38$$

$$S_{1500}^2 =$$

$$\sqrt{\frac{1}{5-1} ((271.2 - 272.04)^2 + (261 - 272.04)^2 + (276 - 272.04)^2 + (282 - 272.04)^2 + (270 - 272.04)^2)} = 7.77$$

- So sánh: Kết luận: Độ lệch chuẩn của cự ly 500m cho biết thời gian chạy 500m của 5 người này chỉ lệch trung bình 3.38s so với thời gian chạy trung bình 500m là 57.96s. Nhưng độ lệch chuẩn của cự ly 1500m đến 7.77s cho thấy với cự ly dài hơn thì thành tích trung bình của 5 vận động viên sẽ có sự khác biệt đáng kể hơn so với cự ly 500m.

2.5. Phân phối tần suất (*Frequency distribution*)

- Một tập dữ liệu được tạo thành từ sự phân phối các giá trị hoặc điểm số.
Trong bảng hoặc biểu đồ, bạn có thể tóm tắt tần suất của mọi giá trị có thể có của một biến theo số hoặc tỷ lệ phần trăm.

- Phân loại:

- Bảng phân phối tần số đơn giản

Ví dụ \Rightarrow

<i>Gender</i>	<i>Number</i>
Male	182
Female	235
Other	27

KL:

- Phụ nữ nhiều hơn nam giới
- Những người có giới tính khác đã tham gia vào nghiên cứu.

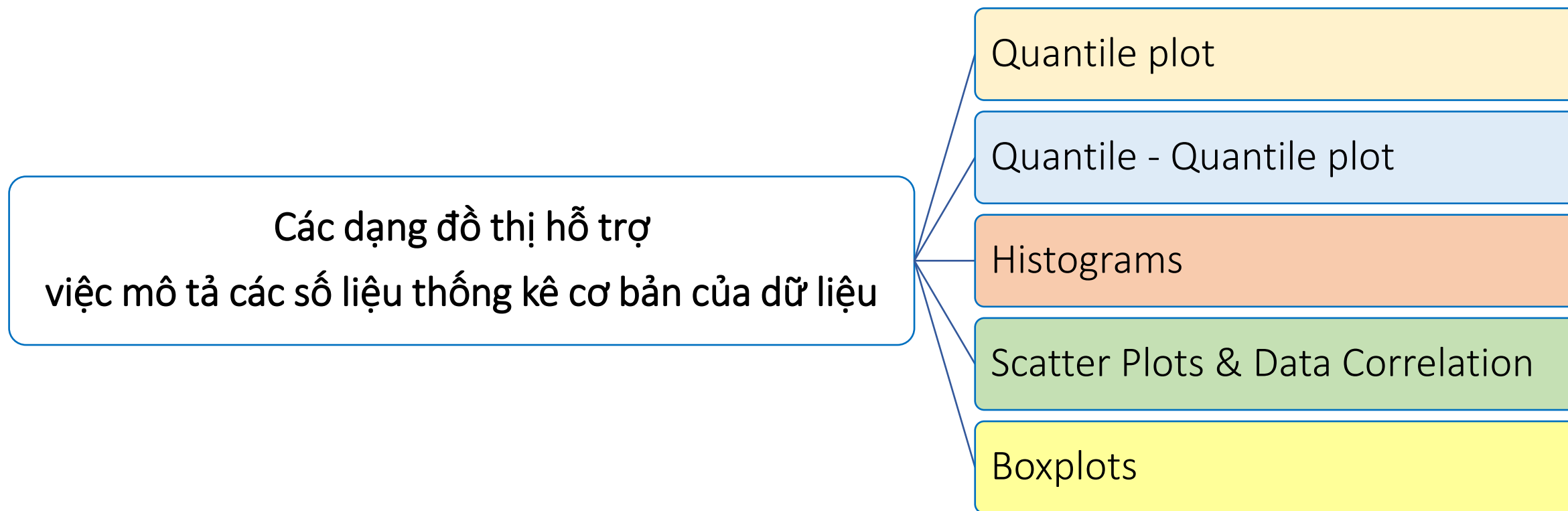
- Bảng phân phối tần số theo nhóm

<i>Library visits in the past year</i>	<i>Percent</i>
0–4	6%
5–8	20%
9–12	42%
13–16	24%
17+	8%

KL:

- Hầu hết mọi người đã ghé thăm thư viện từ 5 đến 16 lần trong năm qua.

2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu



Các dạng đồ thị giúp xác định nhiễu và các ngoại lệ, sẽ đặc biệt hữu ích cho việc làm sạch dữ liệu sau này

2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

2.5.1. Quantile Plot (Biểu đồ phân vị)

- Quantile plot là một cách đơn giản và hiệu quả để có cái nhìn đầu tiên về phân bố dữ liệu **đơn biến** (univariate).
- Hiển thị tất cả dữ liệu cho thuộc tính nhất định (cho phép người dùng đánh giá cả hành vi tổng thể và các lần xuất hiện bất thường).
- Vẽ đồ thị thông tin về các phân vị (Q1, Q2, Q3).
- Cho phép so sánh các phân phối khác nhau dựa trên phân vị

Giả sử x_i , với $i = 1$ đến N , là dữ liệu được sắp xếp theo thứ tự tăng dần sao cho x_1 là quan sát nhỏ nhất và x_N là lớn nhất đối với một số thuộc tính thứ tự hoặc số X . Mỗi quan sát, x_i , được ghép với tỷ lệ phần trăm, f_i , chỉ ra rằng khoảng $f_i \times 100\%$ dữ liệu nằm dưới giá trị, x_i . Ta nói “xấp xỉ” vì có thể không có giá trị nào có chính xác một phần, f_i , của dữ liệu bên dưới x_i . Lưu ý rằng phân vị 0,25 tương ứng với tứ phân vị Q1, phân vị 0,50 là trung vị (median) và phân vị 0,75 là Q3. Với

$$f_i = \frac{i-0.5}{N}$$

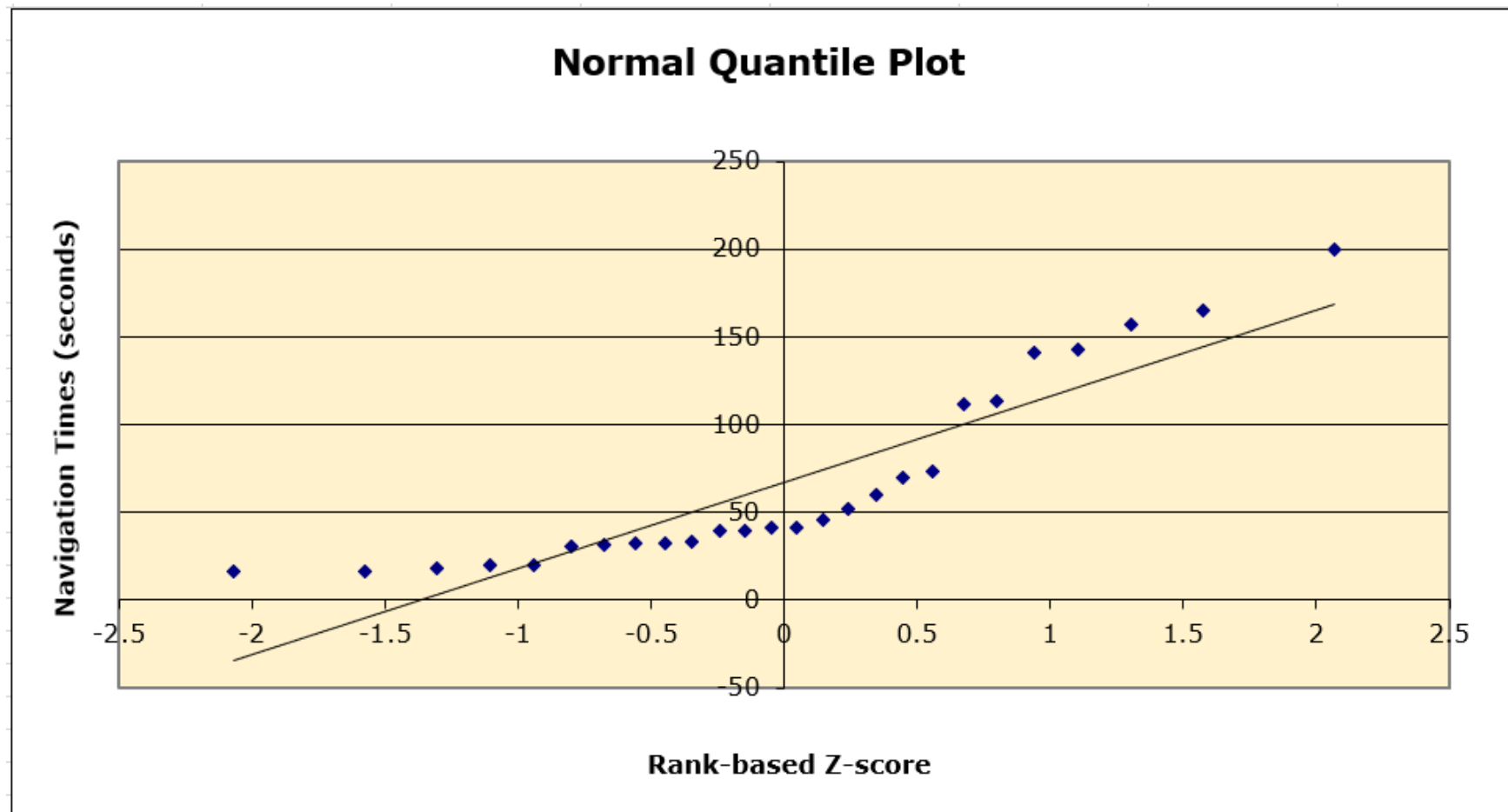
2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

2.5.1. Quantile Plot (Biểu đồ phân vị)

- Ví dụ: cho bảng dữ liệu sau:

rank-based z-score	time
-2.069901831	16.042
-1.574444965	16.606
-1.303782672	18.367
-1.104835744	20.03
-0.942075775	20.042
-0.801094529	30.726
-0.67448975	31.538
-0.557884763	32.428
-0.448425483	32.589
-0.344102463	33.522
-0.243404178	39.5
-0.145120941	39.619
-0.048223074	41.362
0.048223074	41.673
0.145120941	45.874
0.243404178	52.135
0.344102463	59.999
0.448425483	69.86
0.557884763	72.879
0.67448975	111.137
0.801094529	113.141
0.942075775	140.862
1.104835744	143.063
1.303782672	157.113
1.574444965	165.308
2.069901831	199.531

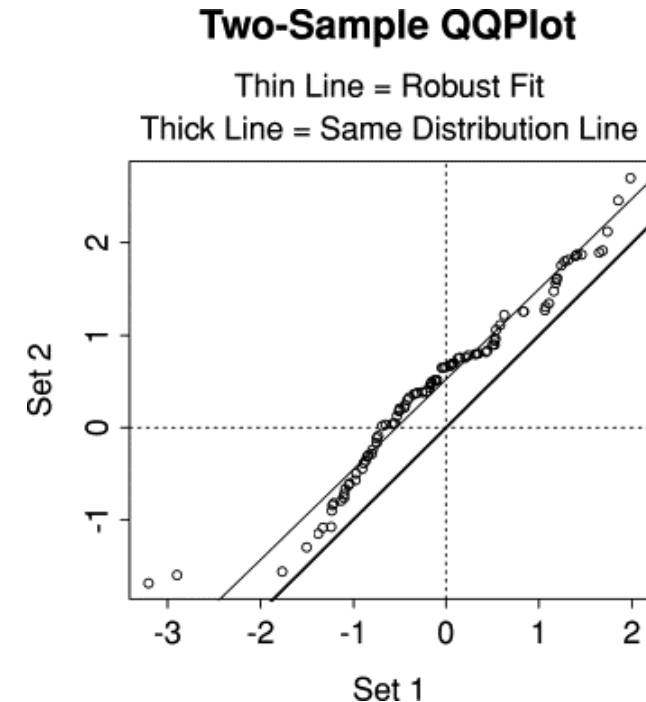
Đồ thị phân vị tương ứng:



2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

2.5.2. Quantile–Quantile Plot (*biểu đồ lượng tử-phân vị*)

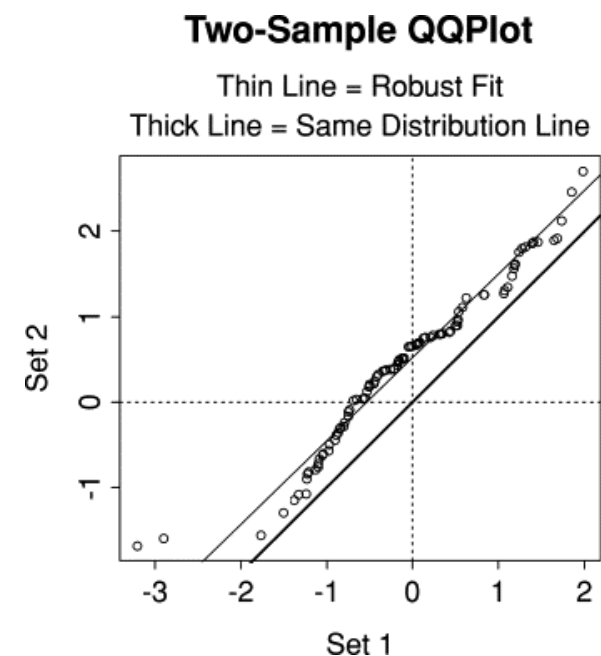
- Biểu đồ quantile–quantile plot hay Q–Q plot là biểu đồ xác suất, một phương pháp đồ họa để so sánh hai phân phối xác suất bằng cách vẽ các phân vị của chúng với nhau. Một điểm (x, y) trên biểu đồ tương ứng với một trong các phân vị của phân phối thứ hai (tọa độ y) được vẽ trên cùng một phân vị của phân phối thứ nhất (tọa độ x). Điều này xác định một đường cong tham số trong đó tham số là chỉ số của khoảng lượng tử.
- Nếu hai phân phối được so sánh là tương tự nhau, thì các điểm trong đồ thị Q–Q sẽ xấp xỉ nằm trên đường đồng nhất $y = x$. Nếu các phân phối có quan hệ tuyến tính, các điểm trong đồ thị Q–Q sẽ xấp xỉ nằm trên một đường thẳng, nhưng không nhất thiết nằm trên đường thẳng $y = x$.



2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

2.5.2. Quantile–Quantile Plot (biểu đồ lượng tử-phân vị)

- Giả sử có hai tập hợp quan sát về biến đổi của thuộc tính đơn giá, được lấy từ hai chi nhánh khác nhau. Đặt x_1, \dots, x_N là dữ liệu từ nhánh đầu tiên và y_1, \dots, y_M là dữ liệu từ nhánh thứ hai, trong đó mỗi tập dữ liệu được sắp xếp theo thứ tự tăng dần.
- Nếu $M = N$ (tức là số điểm trong mỗi tập hợp là như nhau), thì ta chỉ cần vẽ đồ thị y_i theo x_i , trong đó y_i và x_i đều là phân vị $(i - 0,5)/N$ của tập dữ liệu tương ứng của chúng.
- Nếu $M < N$ (tức là nhánh thứ hai có ít quan sát hơn nhánh thứ nhất), chỉ có thể có M điểm trên biểu đồ Q-q. Ở đây, y_i là phân vị $(i - 0,5)/M$ của y



2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

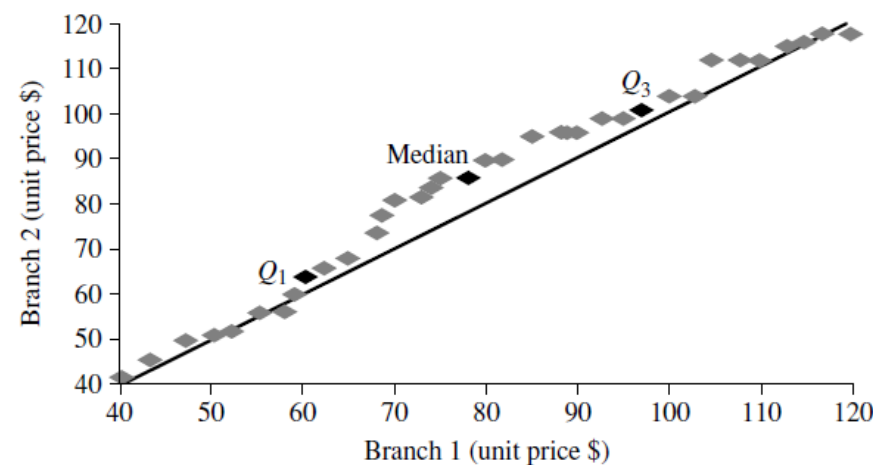
2.5.2. Quantile–Quantile Plot (biểu đồ lượng tử-phân vị)

- Ví dụ: cho dữ liệu đơn giá của các mặt hàng được bán tại hai chi nhánh của 1 công ty trong một khoảng thời gian nhất định. Mỗi điểm tương ứng với cùng một lượng tử cho mỗi tập dữ liệu và hiển thị đơn giá của các mặt hàng được bán ở chi nhánh 1 so với chi nhánh 2 cho lượng tử đó. Các điểm tối hơn tương ứng với dữ liệu cho Q_1 , trung vị (*median*) và Q_3 .)

Qua ví dụ, ta thấy

- Tại Q_1 : đơn giá của các mặt hàng bán ở chi nhánh 1 thấp hơn một chút so với chi nhánh 2. Nói cách khác, 25% mặt hàng bán ở chi nhánh 1 nhỏ hơn hoặc bằng 60\$, trong khi 25% % mặt hàng được bán tại chi nhánh 2 nhỏ hơn hoặc bằng 64\$.
- Tại Q_2 : (phân vị thứ 50): 50% mặt hàng được bán ở nhánh 1 có giá dưới 78\$, trong khi 50% mặt hàng ở nhánh 2 có giá dưới 85\$.

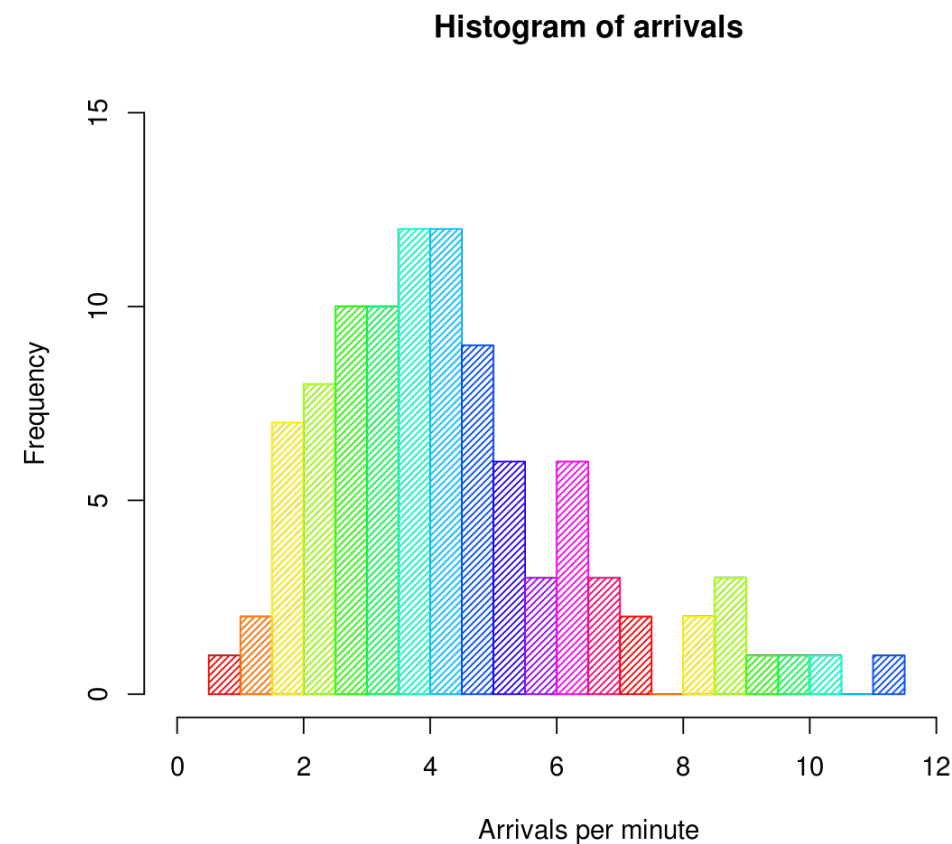
⇒ Nhìn chung, có sự dịch chuyển trong phân phối của chi nhánh 1 so với chi nhánh 2 ở chỗ đơn giá các mặt hàng bán ở chi nhánh 1 có xu hướng thấp hơn so với chi nhánh 2.



2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

2.5.3. Histograms (biểu đồ hoặc biểu đồ tần suất - frequency histograms)

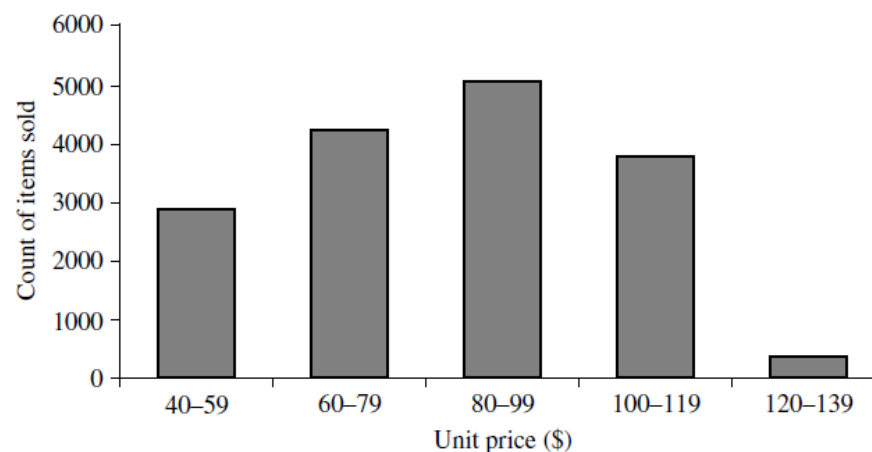
- “Histos” có nghĩa là cột, và “gram” có nghĩa là biểu đồ, vì vậy biểu đồ là biểu đồ các cực.
- Histogram là một phương pháp đồ họa để tóm tắt sự phân bố của một thuộc tính nhất định, X . Nếu X là danh nghĩa, chẳng hạn như số lượng chuyến bay đến trong mỗi ngày hay mẫu ô tô hoặc loại mặt hàng bán được trong mỗi tháng, ... thì thanh dọc được vẽ cho mỗi giá trị đã biết của X . Chiều cao của thanh cho biết tần suất (frequency - hay số lượng) của giá trị X đó.
- Biểu đồ kết quả thường được gọi là biểu đồ thanh (bar chart).



2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

2.5.3. Histograms (*biểu đồ hoặc biểu đồ tần suất - frequency histograms*)

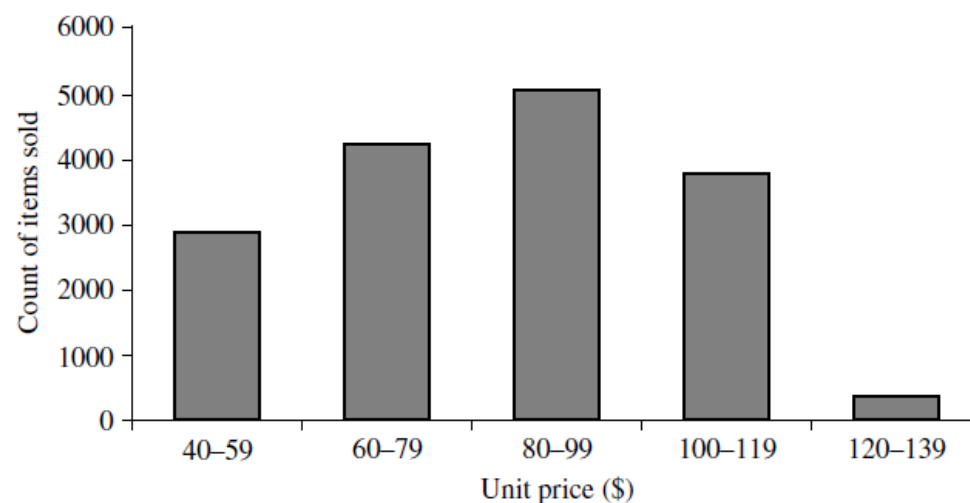
- Nếu kiểu dữ liệu của X là số thì thuật ngữ biểu đồ (histogram) được ưa thích hơn. Phạm vi giá trị của X có thể được phân chia thành các phạm vi con liên tiếp rời rạc.
- Các phạm vi con, được gọi là nhóm (buckets) hoặc thùng (bins), là các tập hợp con rời rạc của phân phối dữ liệu cho X .
- Phạm vi của mỗi nhóm được gọi là chiều rộng (width). Thông thường, các buckets có chiều rộng bằng nhau. Ví dụ: thuộc tính giá có phạm vi giá trị từ \$1 đến \$200 (làm tròn đến đô la gần nhất) có thể được phân chia thành các phạm vi con từ 1 đến 20, 21 đến 40, 41 đến 60, v.v.



2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

2.5.3. Histograms (*biểu đồ hoặc biểu đồ tần suất - frequency histograms*)

- Đối với mỗi phạm vi con, một thanh được vẽ có chiều cao biểu thị tổng số mục được quan sát trong phạm vi con.



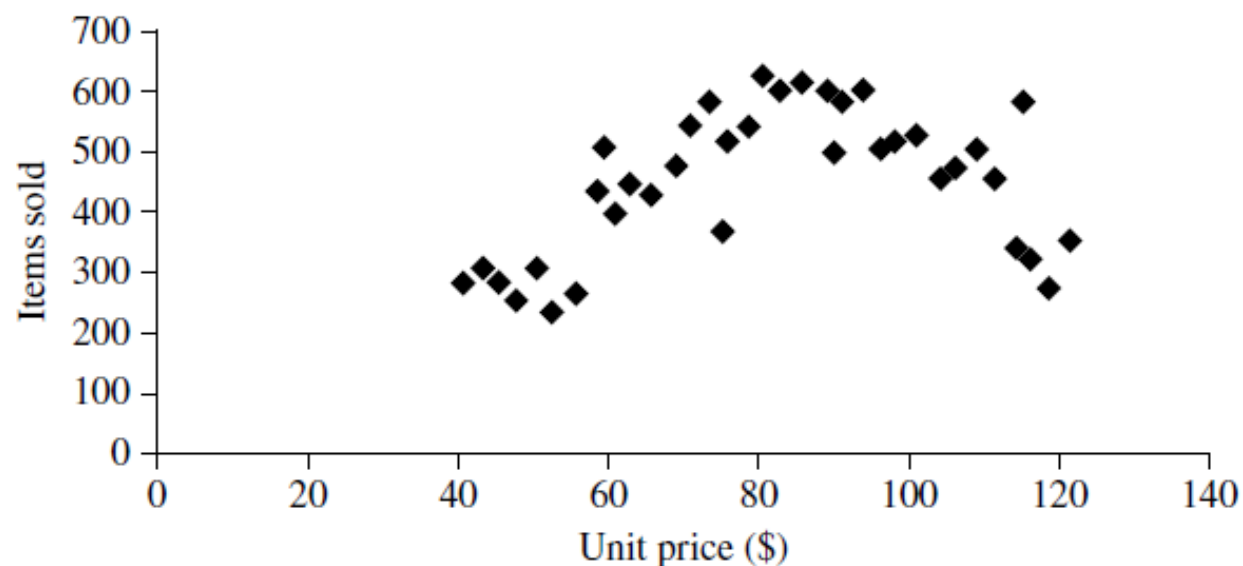
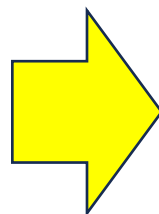
- Mặc dù biểu đồ được sử dụng rộng rãi nhưng chúng có thể không hiệu quả bằng các phương pháp biểu đồ phân vị (*quantile plot*), biểu đồ q-q (*q-q plot*) và biểu đồ hộp (*boxplot*) trong việc so sánh các nhóm quan sát đơn biến.

2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

2.5.4. Scatter Plots and Data Correlation (*biểu đồ phân tán và tương quan dữ liệu*)

- Biểu đồ phân tán (scatter plot) là một trong những phương pháp đồ họa hiệu quả nhất để xác định xem có tồn tại mối quan hệ, mẫu hoặc xu hướng giữa hai thuộc tính số hay không.
- Để xây dựng một biểu đồ phân tán, mỗi cặp giá trị được coi là một cặp tọa độ theo nghĩa đại số và được vẽ dưới dạng các điểm trong mặt phẳng.
- Minh họa biểu đồ phân tán cho tập hợp dữ liệu trong bảng:

Unit price (\$)	Count of items sold
40	275
43	300
47	250
-	-
74	360
75	515
78	540
-	-
115	320
117	270
120	350



2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

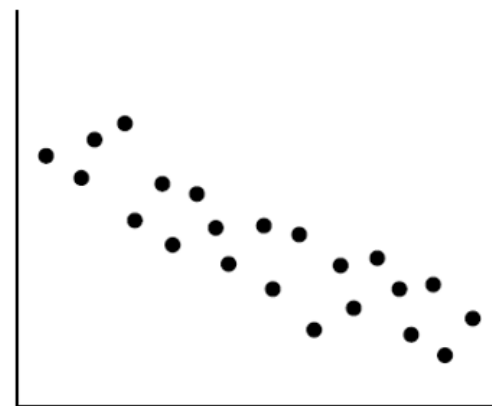
2.5.4. Scatter Plots and Data Correlation (biểu đồ phân tán và tương quan dữ liệu)

- Minh họa 1:

- Cho thấy các ví dụ về mối tương quan tích cực và tiêu cực giữa hai thuộc tính.
- (Hình a): mô hình được vẽ dốc từ phía dưới bên trái sang phía trên bên phải \Rightarrow giá trị của X tăng khi giá trị của Y tăng, cho thấy mối tương quan dương
- (Hình b): mô hình được vẽ dốc từ phía trên bên trái xuống phía dưới bên phải \Rightarrow giá trị của X tăng khi giá trị của Y giảm, cho thấy mối tương quan âm (tương quan nghịch).



(a) Mối tương quan dương
(positive correlation)

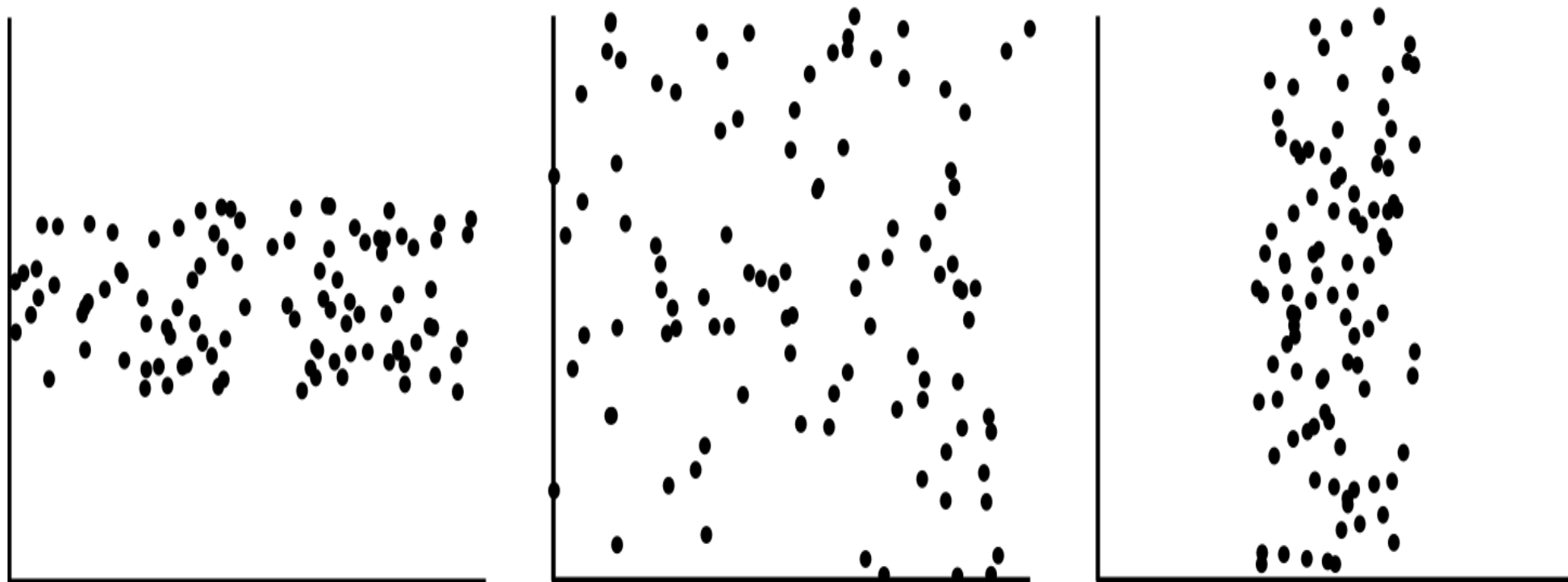


(b) Mối tương quan âm
(negative correlation)

2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

2.5.4. Scatter Plots and Data Correlation (biểu đồ phân tán và tương quan dữ liệu)

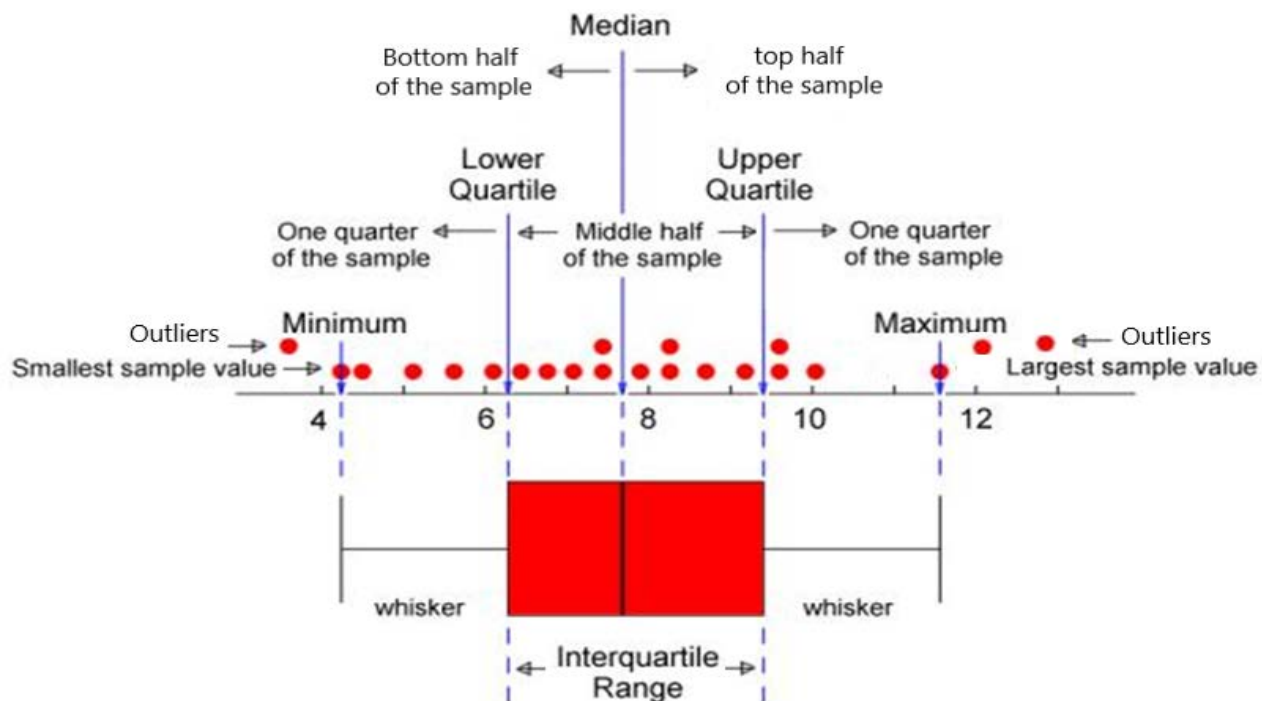
- *Minh họa 2*: cho thấy ba trường hợp không có mối quan hệ tương quan giữa hai thuộc tính



2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

2.5.5. Boxplots (hay Box and Whisker plot)

- Do John Tukey sáng tạo ra năm 1977.
- Là dạng biểu đồ mô tả 5 vị trí phân bố của dữ liệu (*5-number summary*) như sau:
 - Độ dài của hộp biểu thị phạm vi của 50% *Dữ liệu trung tâm (IQR)*.
 - Đường kẻ giữa hộp là giá trị *Trung vị* của tập dữ liệu (*median*).
 - Các đường kẻ dưới và trên (khi biểu đồ ở dạng đứng) hoặc đường kẻ bên trái và bên phải (khi biểu đồ ở dạng nằm ngang) của hộp tương ứng với *Q1* và *Q3*.
 - Các đường kẻ bên ngoài (còn được gọi là râu của biểu đồ) mô tả phạm vi của những dữ liệu ngoài khoảng 25% và 75% dữ liệu trung tâm.
 - Các dấu chấm bên ngoài (nếu có) là các giá trị ngoại lai (*outliers*).



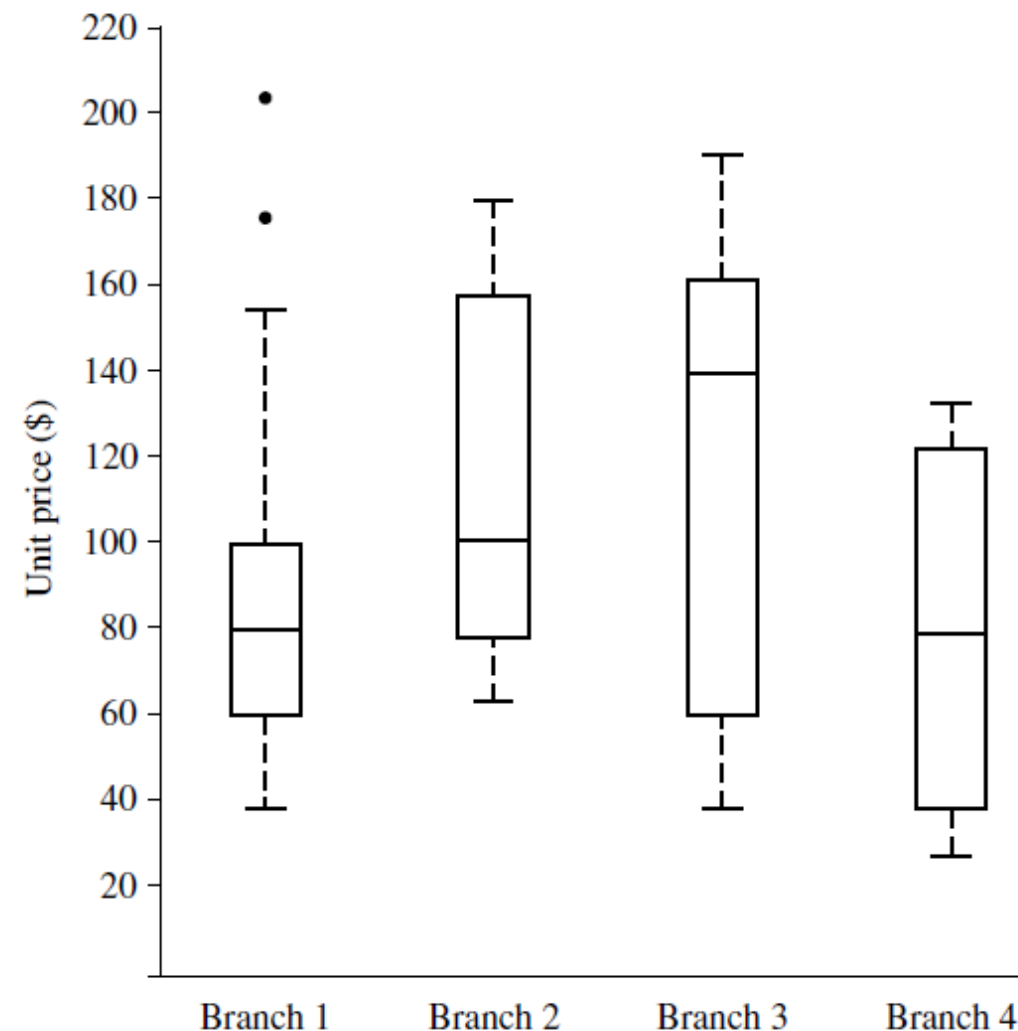
2.6. Sử dụng đồ họa để mô tả các số liệu thống kê cơ bản của dữ liệu

2.5.5. Boxplots (hay Box and Whisker plot)

- Minh họa: về dữ liệu đơn giá của các mặt hàng được bán tại bốn chi nhánh của 1 công ty trong một khoảng thời gian nhất định.

Đối với chi nhánh 1:

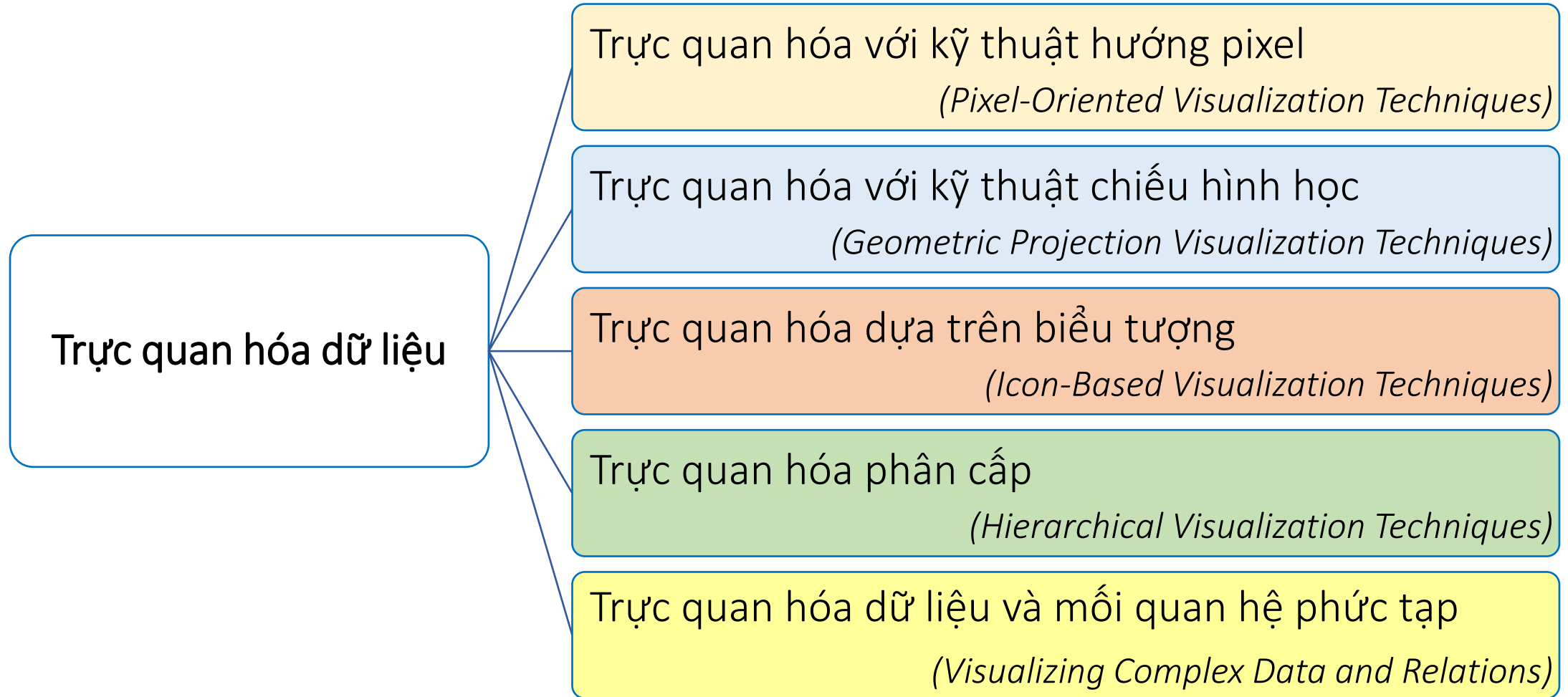
- Mean (Giá) của các mặt hàng là 80\$
- Q1 là 60\$
- Q3 là 100\$.
- IQR là 40 ($=100 - 60$)
- Hai quan sát ngoại lệ được vẽ riêng lẻ, vì giá trị 175 và 202 của chúng $>1,5 \cdot \text{IQR}$.



NỘI DUNG CHƯƠNG 3

1. Đối tượng dữ liệu và kiểu thuộc tính
2. Mô tả các thống kê cơ bản
3. Trực quan hóa dữ liệu
4. Đo lường sự tương đồng và khác biệt của dữ liệu
5. Bài tập

3. TRỰC QUAN HÓA DỮ LIỆU (*Data Visualization*)



3.1.- *Trực quan hóa với kỹ thuật hướng pixel (Pixel-Oriented Visualization Techniques)*

- Một cách đơn giản để trực quan hóa giá trị của các chiều (value of a dimension) là sử dụng pixel trong đó màu của pixel phản ánh giá trị của các chiều.
- Đối với tập dữ liệu có m chiều, kỹ thuật hướng pixel (*pixel-oriented techniques*) tạo ra m cửa sổ trên màn hình, mỗi cửa sổ cho mỗi chiều. Giá trị m chiều của bản ghi được ánh xạ tới m pixel tại các vị trí tương ứng trong cửa sổ. Màu sắc của các pixel phản ánh các giá trị tương ứng.
- Bên trong một cửa sổ, các giá trị dữ liệu được sắp xếp theo thứ tự chung nào đó được chia sẻ bởi tất cả các cửa sổ. Việc sắp xếp trật tự toàn cục có thể đạt được bằng cách sắp xếp tất cả các bản ghi dữ liệu theo cách có ý nghĩa đối với nhiệm vụ hiện tại.

3. Trực quan hóa dữ liệu (Data Visualization)

3.1.- Trực quan hóa với kỹ thuật hướng pixel (Pixel-Oriented Visualization Techniques)

- Ví dụ: Giả sử một công ty có dữ liệu về khách hàng, bao gồm: thu nhập (income), hạn mức tín dụng (credit_limit), khối lượng giao dịch (transaction volume) và độ tuổi (age). Có phân tích mối tương quan giữa thu nhập và các thuộc tính khác thông qua trực quan hóa dữ liệu hay không?
 - Có thể sắp xếp tất cả khách hàng theo thứ tự thu nhập tăng dần và sử dụng thứ tự này để sắp xếp dữ liệu khách hàng trong bốn cửa sổ trực quan
 - Màu pixel được chọn sao cho giá trị càng nhỏ thì độ bóng càng nhạt.

Nhận xét dựa trên hình ảnh:

- Hạn mức tín dụng tăng khi thu nhập tăng;
- Khách hàng có thu nhập ở mức trung bình có nhiều khả năng mua hàng của công ty hơn;
- Không có mối tương quan rõ ràng giữa thu nhập và độ tuổi.



(a) income



(b) credit_limit



(c) transaction_volume



(d) age

3. Trực quan hóa dữ liệu (Data Visualization)

3.1.- *Trực quan hóa với kỹ thuật hướng pixel (Pixel-Oriented Visualization Techniques)*

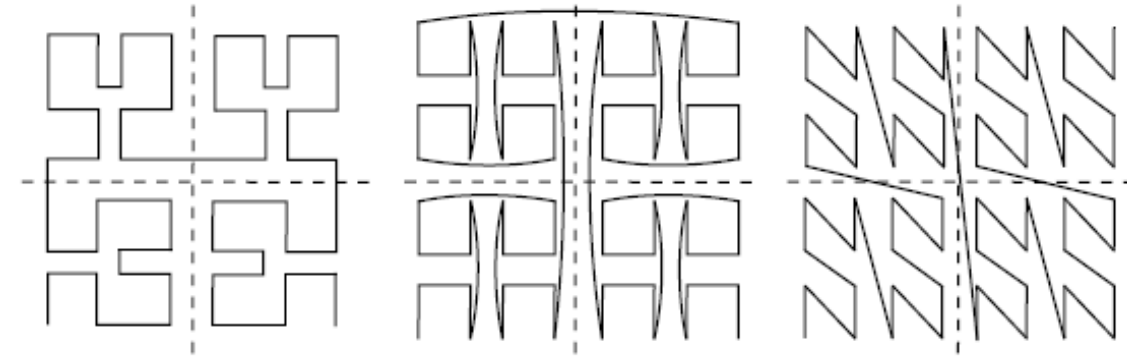
- Việc lấp đầy một cửa sổ bằng cách sắp xếp các bản ghi dữ liệu theo cách tuyến tính có thể không hoạt động tốt đối với một cửa sổ rộng do:
 - Pixel đầu tiên trong hàng cách xa pixel cuối cùng ở hàng trước, mặc dù chúng nằm cạnh nhau theo thứ tự chung.
 - Hơn nữa, một pixel nằm cạnh pixel phía trên nó trong cửa sổ, mặc dù cả hai pixel này không nằm cạnh nhau theo thứ tự chung.
- Để giải quyết vấn đề này, ta có thể bố trí các bản ghi dữ liệu theo *đường cong lấp đầy khoảng trống (space-filling curve to fill)* để lấp đầy các cửa sổ.
- Đường cong lấp đầy khoảng trống là một đường cong có phạm vi bao phủ toàn bộ siêu khối đơn vị n chiều (n -dimensional unit hypercube). Vì cửa sổ hiển thị là 2-D nên ta có thể sử dụng bất kỳ đường cong lấp đầy không gian 2-D (*2-D space-filling curve*) nào.

3. Trực quan hóa dữ liệu (Data Visualization)

3.1.- Trực quan hóa với kỹ thuật hướng pixel (Pixel-Oriented Visualization Techniques)

Các cửa sổ không nhất thiết phải là hình chữ nhật.

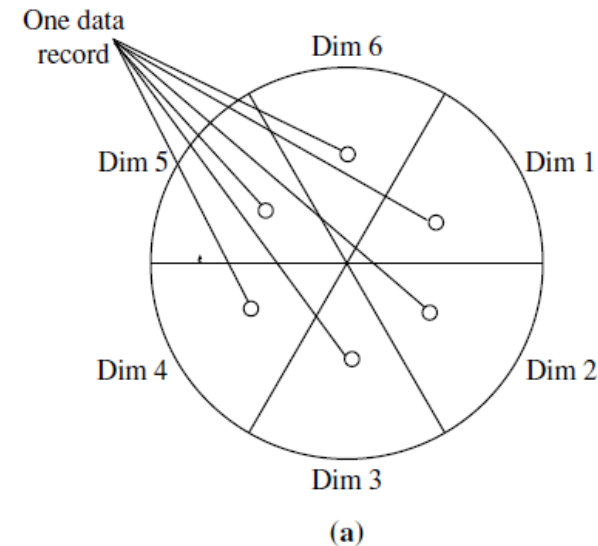
- Minh họa: Một số đường cong 2-D space-filling curve thường được sử dụng
- Minh họa: Kỹ thuật phân đoạn đường tròn (circle segment technique). Kỹ thuật này có thể dễ dàng so sánh kích thước vì các cửa sổ kích thước được đặt cạnh nhau và tạo thành một vòng tròn
 - (a) Biểu diễn một bản ghi dữ liệu theo các đoạn hình tròn.
 - (b) Bố trí các pixel theo các đoạn hình tròn



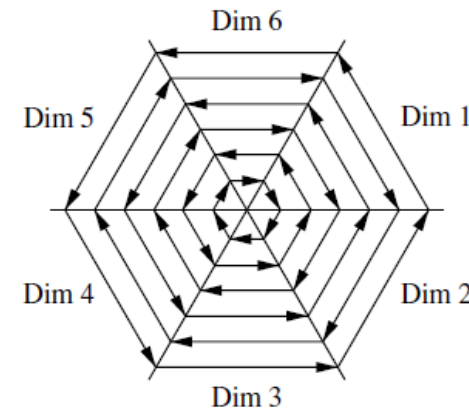
(a) Hilbert curve

(b) Gray code

(c) Z-curve



(a)



(b)

3.2.- *Trực quan hóa với kỹ thuật chiếu hình học*

(Geometric Projection Visualization Techniques)

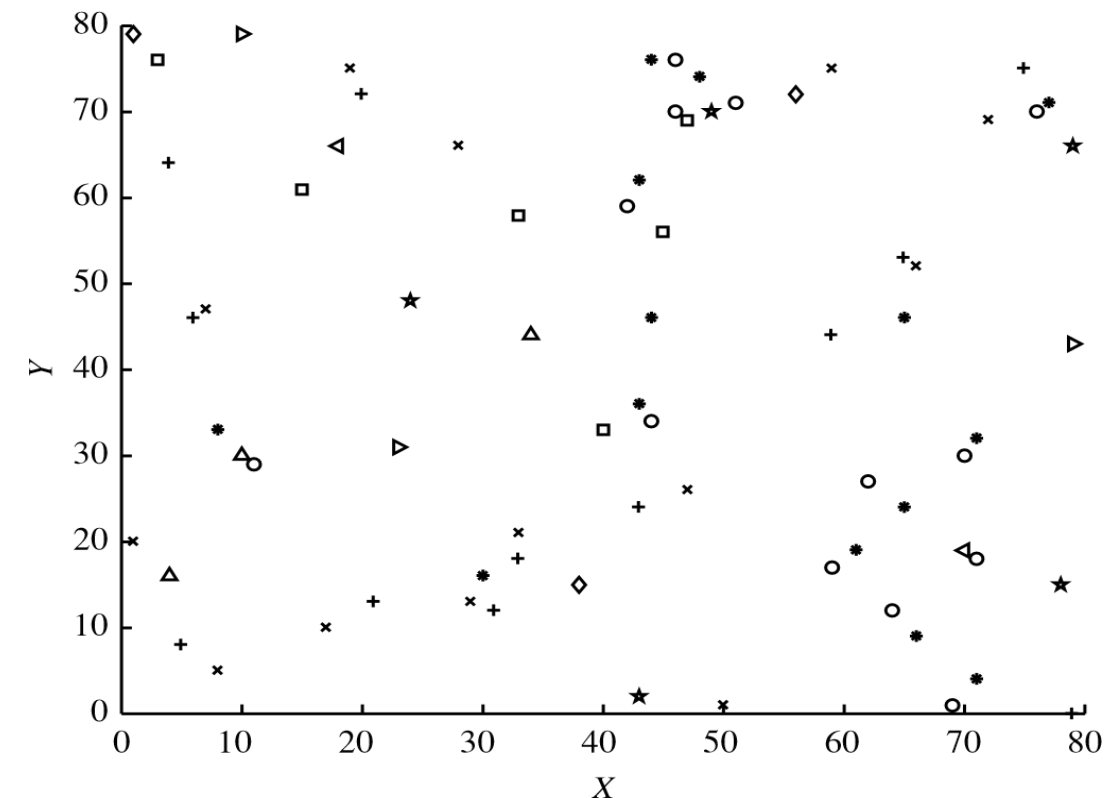
- Một hạn chế của kỹ thuật hiển thị theo định hướng pixel là không thể giúp hiểu sự phân bố dữ liệu trong không gian đa chiều. Ví dụ, chúng không chỉ ra liệu có một vùng dày đặc trong không gian con đa chiều hay không.
- Kỹ thuật chiếu hình học (*Geometric projection techniques*) giúp người dùng tìm ra những phép chiếu thú vị của tập dữ liệu đa chiều.
- Thách thức chính mà các kỹ thuật chiếu hình học cố gắng giải quyết là làm thế nào để hiển thị không gian nhiều chiều trên màn hình 2-D.

3. Trực quan hóa dữ liệu (Data Visualization)

3.2.- Trực quan hóa với kỹ thuật chiếu hình học (Geometric Projection Visualization Techniques)

3.2.1.- Biểu đồ phân tán (scatter plot)

- Biểu đồ phân tán (scatter plot) hiển thị các điểm dữ liệu 2-D bằng tọa độ Descartes. Chiều thứ ba có thể được thêm vào bằng cách sử dụng các màu sắc hoặc hình dạng khác nhau để thể hiện các điểm dữ liệu khác nhau.
- **Minh họa 1:** X và Y là hai thuộc tính không gian và chiều thứ ba được thể hiện bằng các hình dạng khác nhau. Thông qua hình ảnh trực quan này, ta có thể thấy rằng các điểm thuộc loại “+” và “x” có xu hướng được đặt cùng một vị trí.

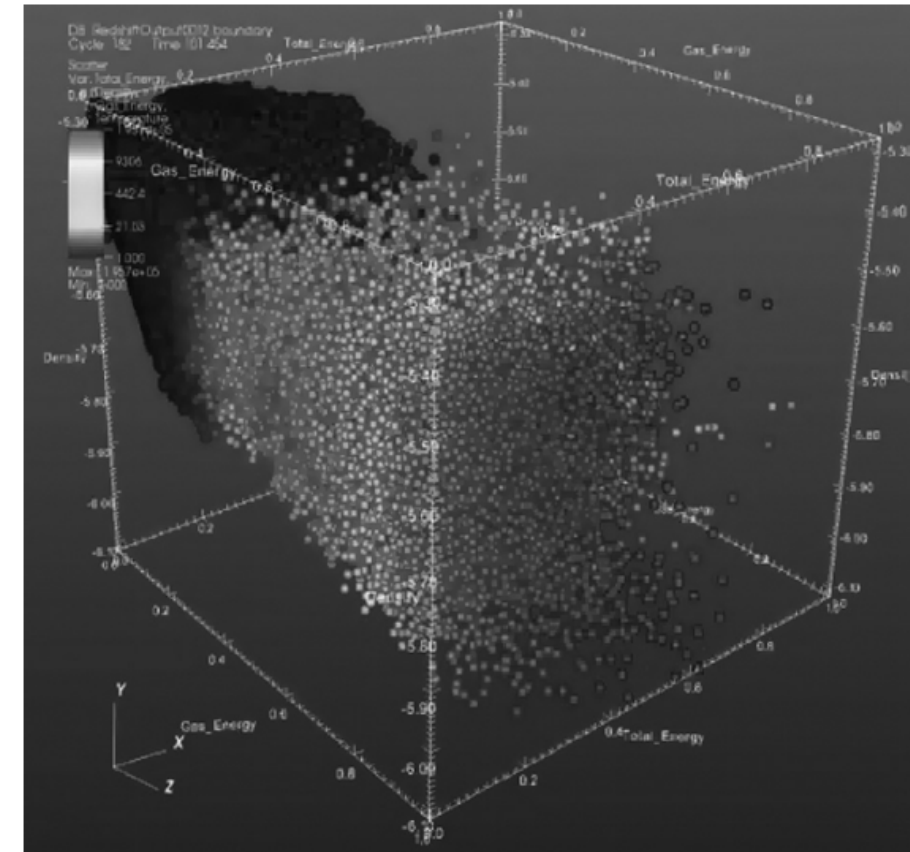


3. Trực quan hóa dữ liệu (Data Visualization)

3.2.- Trực quan hóa với kỹ thuật chiếu hình học (Geometric Projection Visualization Techniques)

3.2.1.- Biểu đồ phân tán (scatter plot)

- **Minh họa 2:** Biểu đồ phân tán 3-D sử dụng ba trục trong hệ tọa độ Descartes. Nếu sử dụng màu sắc, biểu đồ có thể hiển thị tối đa các điểm dữ liệu 4-D
- Đối với các tập dữ liệu có nhiều hơn bốn chiều, biểu đồ phân tán thường không hiệu quả.



3. Trực quan hóa dữ liệu (Data Visualization)

3.2.- Trực quan hóa với kỹ thuật chiếu hình học (Geometric Projection Visualization Techniques)

3.2.2.- Ma trận biểu đồ phân tán (scatter plot matrix technique)

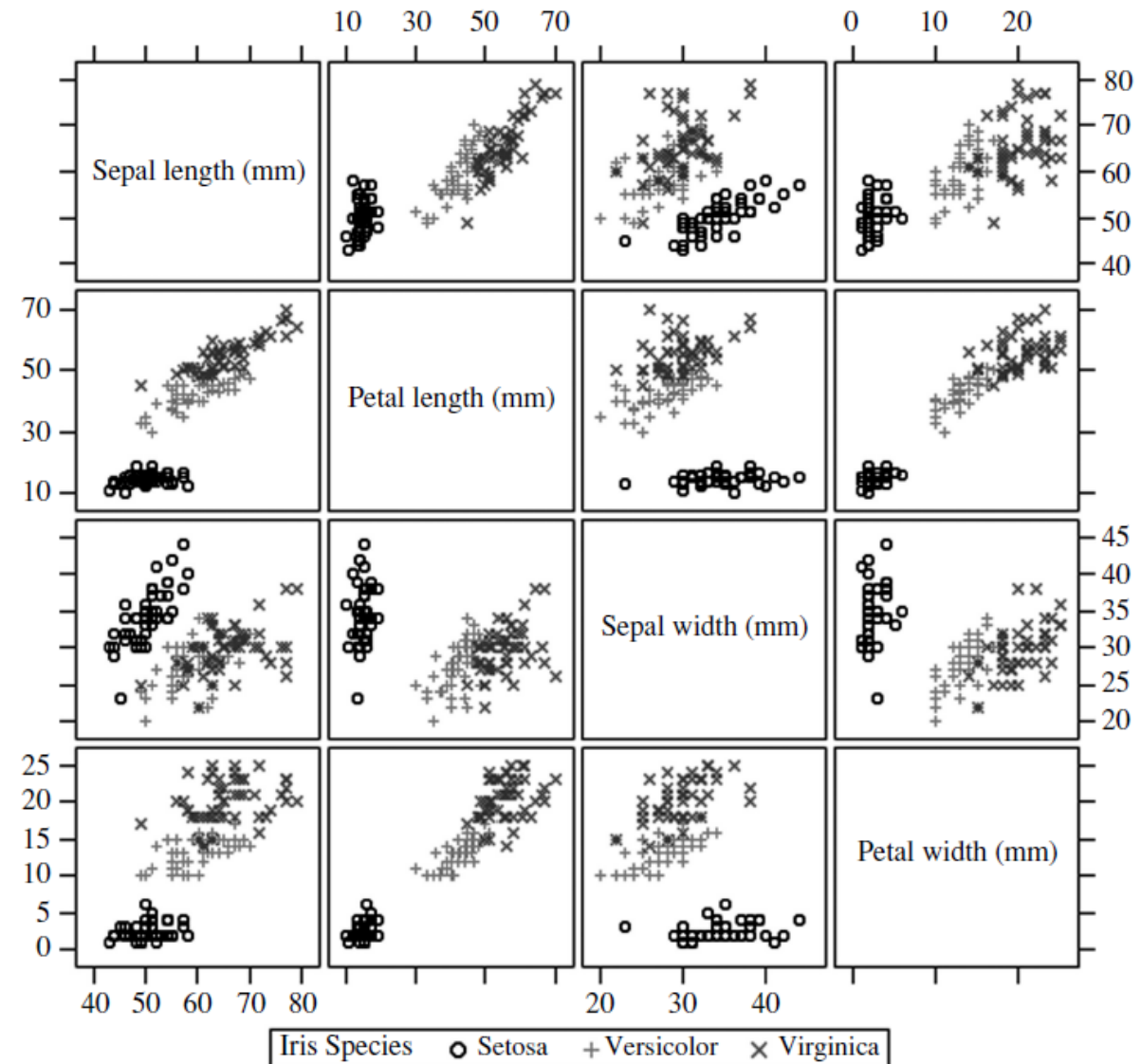
- Kỹ thuật ma trận biểu đồ phân tán là một phần mở rộng hữu ích cho biểu đồ phân tán.
- Đối với tập dữ liệu nhiều chiều, ma trận biểu đồ phân tán là một lưới $n \times n$ gồm các biểu đồ phân tán 2-D cung cấp hình ảnh trực quan của từng chiều với mọi chiều khác.

3. Trực quan hóa dữ liệu (Data Visualization)

3.2.- Trực quan hóa với kỹ thuật chiếu hình học (Geometric Projection Visualization Techniques)

3.2.2.- Ma trận biểu đồ phân tán (scatter plot matrix technique)

- *Minh họa:* về bộ dữ liệu bao gồm 450 mẫu từ mỗi loài trong số ba loài hoa Iris. Có năm chiều trong tập dữ liệu: chiều dài và chiều rộng của đài hoa và cánh hoa cũng như loài.

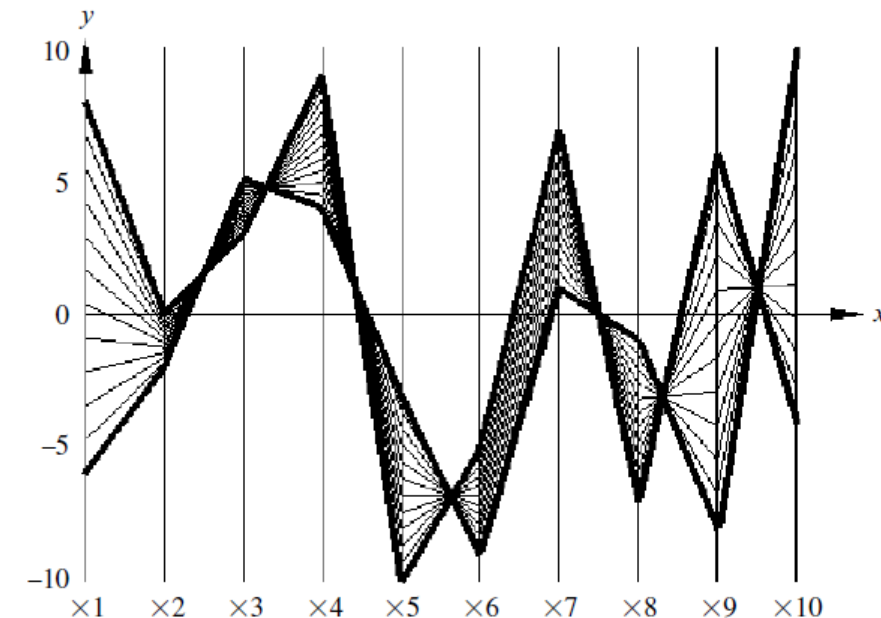


3. Trực quan hóa dữ liệu (Data Visualization)

3.2.- Trực quan hóa với kỹ thuật chiếu hình học (Geometric Projection Visualization Techniques)

3.2.3.- Kỹ thuật tọa độ song song (parallel coordinates technique)

- Ma trận biểu đồ phân tán trở nên kém hiệu quả hơn khi số chiều tăng lên.
- Một kỹ thuật phổ biến khác, được gọi là tọa độ song song (*parallel coordinates technique*) có thể xử lý được nhiều chiều hơn.
- Để trực quan hóa các điểm dữ liệu n chiều, kỹ thuật tọa độ song song vẽ n trục cách đều nhau, một trục cho mỗi chiều, song song với một trong các trục hiển thị.
- Hạn chế chính của kỹ thuật tọa độ song song là không thể hiển thị một cách hiệu quả tập dữ liệu gồm nhiều bản ghi do sự lộn xộn và chồng chéo về mặt hình ảnh thường làm giảm khả năng đọc của hình ảnh trực quan và khiến cho các mẫu khó tìm thấy.

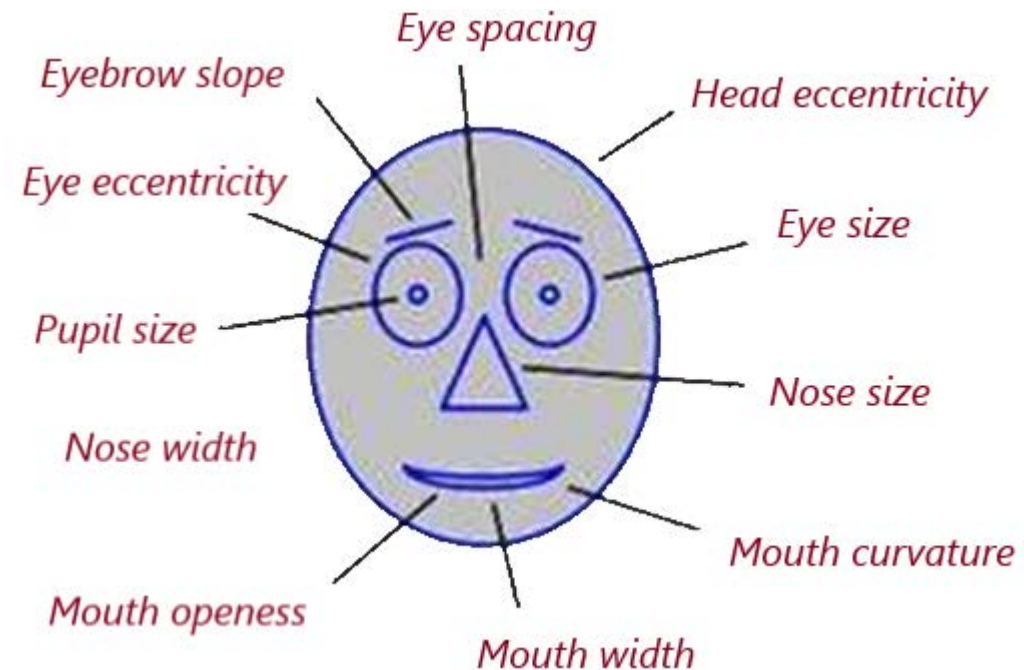


3.3.- Trực quan hóa dựa trên biểu tượng (Icon-Based Visualization Techniques)

Kỹ thuật này sử dụng các biểu tượng nhỏ để thể hiện các giá trị dữ liệu đa chiều

3.3.1. Khuôn mặt Chernoff (Chernoff faces)

- Được nhà thống kê Herman Chernoff giới thiệu vào năm 1973.
- Khuôn mặt Chernoff giúp tiết lộ xu hướng trong dữ liệu. Mỗi mặt (hoạt hình) đại diện cho một điểm dữ liệu n chiều ($n \leq 18$).
- Ví dụ: kích thước có thể được ánh xạ tới các đặc điểm khuôn mặt như: kích thước mắt, khoảng cách giữa hai mắt, chiều dài mũi, chiều rộng mũi, độ cong miệng, chiều rộng miệng, độ mở miệng, kích thước đồng tử, độ nghiêng của lông mày, độ lệch tâm của mắt và độ lệch tâm của đầu.



3. Trực quan hóa dữ liệu (Data Visualization)


3.3.- Trực quan hóa dựa trên biểu tượng (Icon-Based Visualization Techniques)

3.3.1.Khuôn mặt Chernoff (Chernoff faces)

1

PoissonDistribution[3.67853]


#	FacePart	Statistic	Value
1	NoseLength	Mean	3.6
2	EyeSize	StandardDeviation	1.8
3	EyeSlant	Kurtosis	2.8
4	EyesVerticalPosition	Median	3.
5	FaceLength	QuartileDeviation	1.5
6	MouthSmile	PearsonChiSquareTest	0
7	FaceColor	PearsonChiSquareTest	0



2

GammaDistribution[4.2399, 2]


#	FacePart	Statistic	Value
1	NoseLength	Mean	8.
2	EyeSize	StandardDeviation	3.8
3	EyeSlant	Kurtosis	4.8
4	EyesVerticalPosition	Median	7.4
5	FaceLength	QuartileDeviation	2.3
6	MouthSmile	PearsonChiSquareTest	0.00041
7	FaceColor	PearsonChiSquareTest	0.00041



3

GammaDistribution[4.51739, 2]


#	FacePart	Statistic	Value
1	NoseLength	Mean	8.9
2	EyeSize	StandardDeviation	4.4
3	EyeSlant	Kurtosis	4.
4	EyesVerticalPosition	Median	8.1
5	FaceLength	QuartileDeviation	3.
6	MouthSmile	PearsonChiSquareTest	0.00046
7	FaceColor	PearsonChiSquareTest	0.00046



4

NormalDistribution[3.39784, 3.05667]


#	FacePart	Statistic	Value
1	NoseLength	Mean	3.2
2	EyeSize	StandardDeviation	3.
3	EyeSlant	Kurtosis	2.9
4	EyesVerticalPosition	Median	3.3
5	FaceLength	QuartileDeviation	2.2
6	MouthSmile	PearsonChiSquareTest	0.54
7	FaceColor	PearsonChiSquareTest	0.54



5

GammaDistribution[3.78043, 2]


#	FacePart	Statistic	Value
1	NoseLength	Mean	7.5
2	EyeSize	StandardDeviation	3.8
3	EyeSlant	Kurtosis	4.
4	EyesVerticalPosition	Median	6.9
5	FaceLength	QuartileDeviation	2.2
6	MouthSmile	PearsonChiSquareTest	0.021
7	FaceColor	PearsonChiSquareTest	0.021



6

GammaDistribution[2.57564, 2]


#	FacePart	Statistic	Value
1	NoseLength	Mean	5.
2	EyeSize	StandardDeviation	3.2
3	EyeSlant	Kurtosis	4.2
4	EyesVerticalPosition	Median	4.1
5	FaceLength	QuartileDeviation	1.8
6	MouthSmile	PearsonChiSquareTest	0
7	FaceColor	PearsonChiSquareTest	0



7

PoissonDistribution[3.33805]


#	FacePart	Statistic	Value
1	NoseLength	Mean	3.2
2	EyeSize	StandardDeviation	1.8
3	EyeSlant	Kurtosis	3.
4	EyesVerticalPosition	Median	3.
5	FaceLength	QuartileDeviation	1.
6	MouthSmile	PearsonChiSquareTest	0
7	FaceColor	PearsonChiSquareTest	0



8

GammaDistribution[4.0564, 2]


#	FacePart	Statistic	Value
1	NoseLength	Mean	7.9
2	EyeSize	StandardDeviation	3.8
3	EyeSlant	Kurtosis	5.3
4	EyesVerticalPosition	Median	7.5
5	FaceLength	QuartileDeviation	2.1
6	MouthSmile	PearsonChiSquareTest	0.0027
7	FaceColor	PearsonChiSquareTest	0.0027



9

NormalDistribution[2.86128, 0.228569]


#	FacePart	Statistic	Value
1	NoseLength	Mean	2.9
2	EyeSize	StandardDeviation	0.24
3	EyeSlant	Kurtosis	3.1
4	EyesVerticalPosition	Median	2.9
5	FaceLength	QuartileDeviation	0.15
6	MouthSmile	PearsonChiSquareTest	0.99
7	FaceColor	PearsonChiSquareTest	0.99



10

PoissonDistribution[3.51499]


#	FacePart	Statistic	Value
1	NoseLength	Mean	3.5
2	EyeSize	StandardDeviation	1.8
3	EyeSlant	Kurtosis	3.
4	EyesVerticalPosition	Median	4.
5	FaceLength	QuartileDeviation	1.5
6	MouthSmile	PearsonChiSquareTest	0
7	FaceColor	PearsonChiSquareTest	0



11

GammaDistribution[3.69953, 2]


#	FacePart	Statistic	Value
1	NoseLength	Mean	7.4
2	EyeSize	StandardDeviation	4.2
3	EyeSlant	Kurtosis	4.1
4	EyesVerticalPosition	Median	6.6
5	FaceLength	QuartileDeviation	2.6
6	MouthSmile	PearsonChiSquareTest	0.0024
7	FaceColor	PearsonChiSquareTest	0.0024



12

PoissonDistribution[1.8384]

#	FacePart	Statistic	Value
1	NoseLength	Mean	1.9
2	EyeSize	StandardDeviation	1.4
3	EyeSlant	Kurtosis	4.3
4	EyesVerticalPosition	Median	2.
5	FaceLength	QuartileDeviation	1.
6	MouthSmile	PearsonChiSquareTest	0
7	FaceColor	PearsonChiSquareTest	0



3. Trực quan hóa dữ liệu (Data Visualization)

3.3.- Trực quan hóa dựa trên biểu tượng (Icon-Based Visualization Techniques)

3.3.1. Khuôn mặt Chernoff (Chernoff faces)

- Đặc điểm của các khuôn mặt Chernoff:
 - Kích thước mắt và độ nghiêng của lông mày được coi là quan trọng.
 - Giúp người dùng dễ dàng xử lý dữ liệu hơn.
 - Tạo điều kiện thuận lợi cho việc trực quan hóa các quy luật và sự bất thường có trong dữ liệu, mặc dù sức mạnh của chúng trong việc liên kết nhiều mối quan hệ còn hạn chế.
 - Một số hạn chế:
 - Các giá trị dữ liệu cụ thể không được hiển thị.
 - Các đặc điểm trên khuôn mặt có tầm quan trọng khác nhau. Điều này có nghĩa là độ giống nhau của hai khuôn mặt (đại diện cho hai điểm dữ liệu đa chiều) có thể khác nhau tùy thuộc vào thứ tự các kích thước được gán cho các đặc điểm khuôn mặt.
- ⇒ Vì vậy, bản đồ này nên được lựa chọn cẩn thận.

3. Trực quan hóa dữ liệu (Data Visualization)

3.3.- *Trực quan hóa dựa trên biểu tượng (Icon-Based Visualization Techniques)*

3.3.1. Khuôn mặt Chernoff (Chernoff faces)

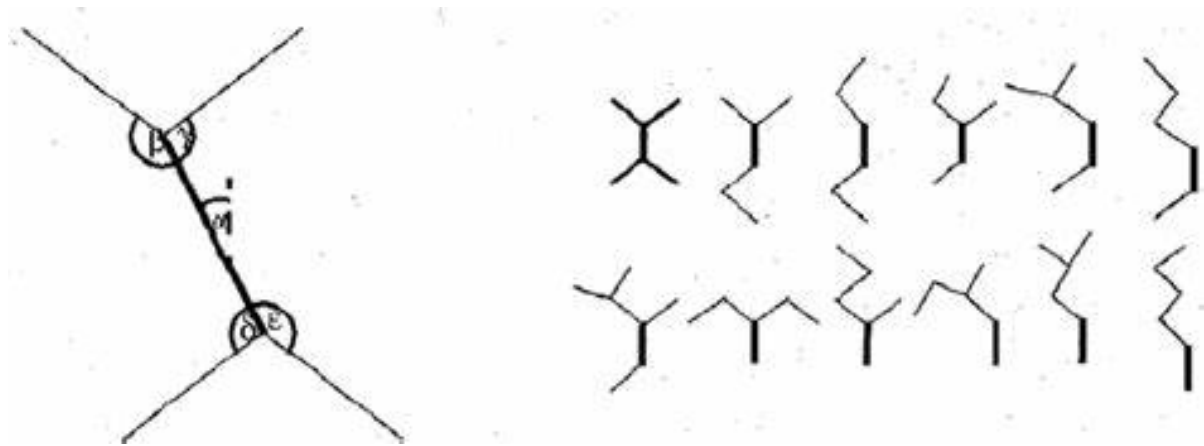
- Vì một khuôn mặt có tính đối xứng dọc (dọc theo trục y) nên bên trái và bên phải của khuôn mặt giống hệt nhau, điều này gây lãng phí không gian.
- Các khuôn mặt Chernoff không đối xứng được đề xuất như một sự mở rộng cho kỹ thuật ban đầu. Khuôn mặt Chernoff không đối xứng tăng gấp đôi số lượng đặc điểm khuôn mặt, do đó cho phép hiển thị tới 36 chiều.

3. Trực quan hóa dữ liệu (Data Visualization)

3.3.- Trực quan hóa dựa trên biểu tượng (Icon-Based Visualization Techniques)

3.3.2. Kỹ thuật trực quan hóa hình que (stick figure visualization technique)

- Kỹ thuật trực quan hóa hình que ánh xạ dữ liệu đa chiều thành hình que năm mảnh, trong đó mỗi hình có bốn chi và một cơ thể. Hai chiều được ánh xạ tới trục hiển thị (x và y) và các chiều còn lại được ánh xạ tới góc và/hoặc chiều dài của các chi.



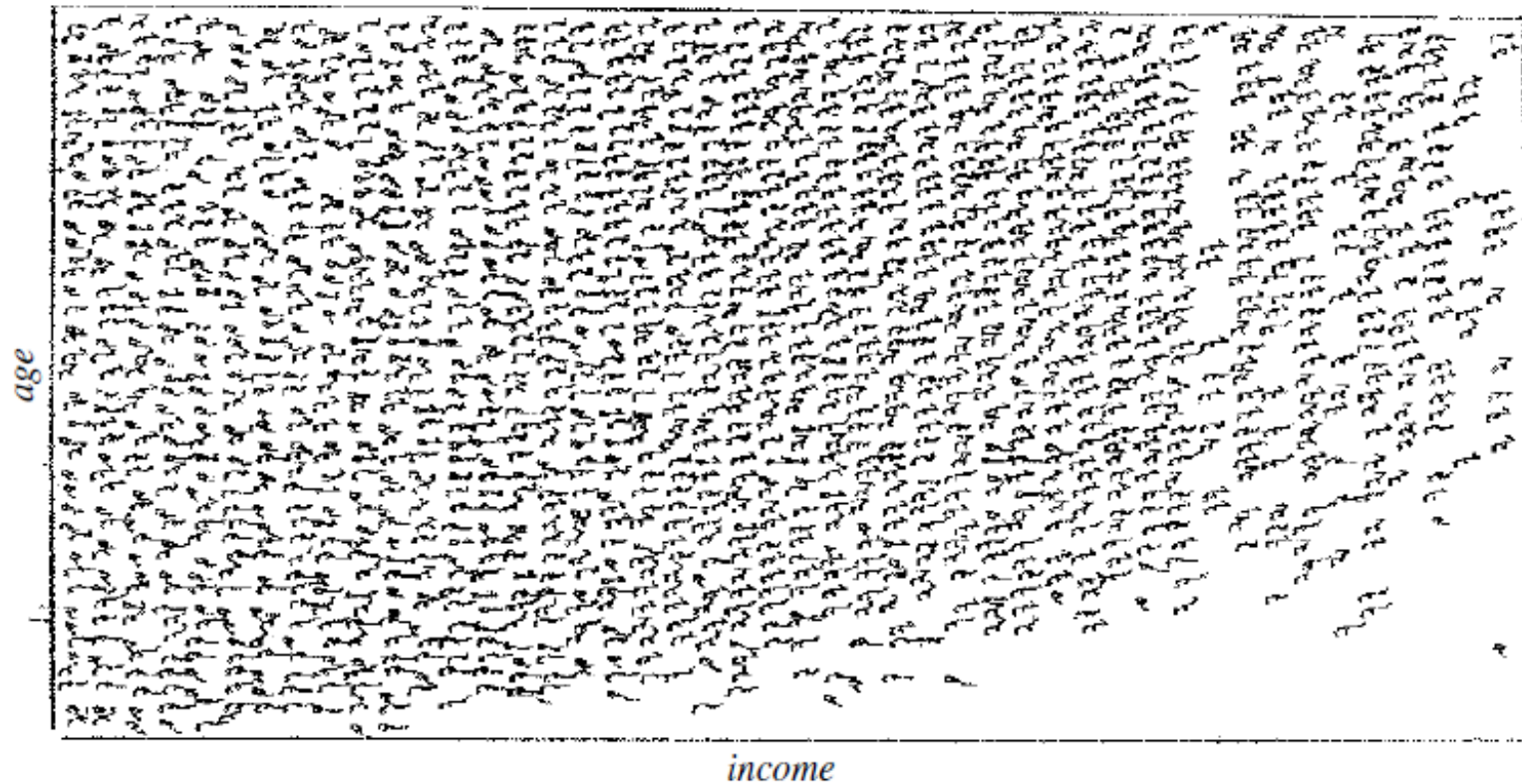
	Low job married	High job married	Low job single	High job single
Male low education				
Male high education				
Female low education				
Female high education				

3. Trực quan hóa dữ liệu (Data Visualization)

3.3.- Trực quan hóa dựa trên biểu tượng (Icon-Based Visualization Techniques)

3.3.2. Kỹ thuật trực quan hóa hình que (stick figure visualization technique)

- **Minh họa:** Hiện thị dữ liệu điều tra dân số, trong đó độ tuổi và thu nhập được ánh xạ tới các trục hiển thị và các khía cạnh còn lại (giới tính, trình độ học vấn, v.v.) được ánh xạ thành các số liệu cố định. Nếu các mục dữ liệu tương đối dày đặc so với hai kích thước hiển thị, thì hình ảnh trực quan thu được sẽ hiển thị các mẫu kết cấu, phản ánh xu hướng dữ liệu.



3.4.- Trực quan hóa phân cấp (Hierarchical Visualization Techniques)

Các kỹ thuật trực quan được thảo luận cho đến nay tập trung vào việc trực quan hóa nhiều chiều cùng một lúc. Tuy nhiên, đối với một tập dữ liệu lớn có nhiều chiều, sẽ khó hình dung được tất cả các chiều cùng một lúc. Kỹ thuật trực quan hóa phân cấp phân chia tất cả các chiều thành các tập hợp con (tức là các không gian con). Các không gian con được hiển thị theo cách phân cấp.

3.4.- *Trực quan hóa phân cấp (Hierarchical Visualization Techniques)*

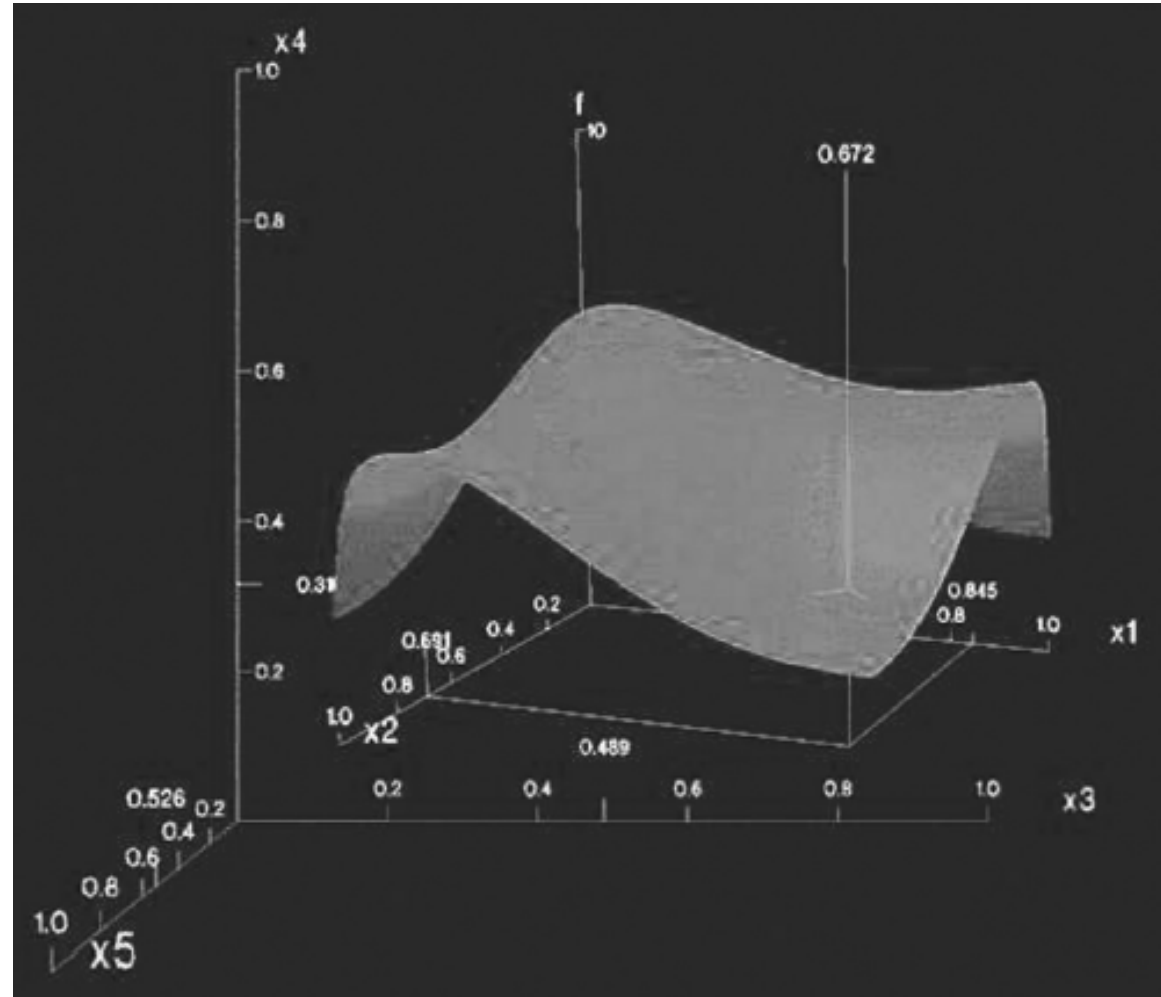
3.3.1. *Worlds-within-Worlds*

- *Worlds-within-Worlds* còn được gọi là *n-Vision*, là một phương pháp trực quan hóa phân cấp tiêu biểu.
- Giả sử muốn trực quan hóa tập dữ liệu *6-D*, trong đó các kích thước là F, X_1, \dots, X_5 . Ta muốn quan sát chiều F thay đổi như thế nào so với các chiều khác. Trước tiên, có thể sửa các giá trị của kích thước X_3, X_4, X_5 thành một số giá trị đã chọn, chẳng hạn như c_3, c_4, c_5 . Sau đó, ta có thể hình dung F, X_1, X_2 bằng cách sử dụng biểu đồ 3-D, được gọi là *world*. Vị trí gốc của *the inner world* nằm ở điểm (c_3, c_4, c_5) của *the outer world*, là một biểu đồ 3D khác sử dụng các kích thước X_3, X_4, X_5 . Người dùng có thể tương tác để thay đổi:
 - *Vị trí gốc* của thế giới bên trong để quan sát những kết quả thay đổi của *the inner world*.
 - *Kích thước* được sử dụng trong *the inner world* và *the outer world*.
- Với nhiều chiều hơn, có thể sử dụng nhiều cấp độ “world” hơn, đó là lý do tại sao phương pháp này được gọi là “*Worlds-within-Worlds*”.

3. Trực quan hóa dữ liệu (Data Visualization)

3.4.- *Trực quan hóa phân cấp (Hierarchical Visualization Techniques)*

3.3.1. *Worlds-within-Worlds*



3. Trực quan hóa dữ liệu (Data Visualization)

3.4.- Trực quan hóa phân cấp (Hierarchical Visualization Techniques)

3.3.2. *Tree-maps*

- Là một ví dụ khác về phương pháp trực quan hóa phân cấp, bản đồ cây hiển thị dữ liệu phân cấp dưới dạng tập hợp các hình chữ nhật lồng nhau.
- Minh họa: Hình bên hiển thị một bản đồ dạng cây trực quan hóa các tin tức (news) trên Google. Tất cả các tin tức được sắp xếp thành bảy loại, mỗi loại được hiển thị trong một hình chữ nhật lớn có màu sắc riêng. Trong mỗi danh mục (tức là mỗi hình chữ nhật ở cấp cao nhất), các câu chuyện tin tức lại được phân chia thành các danh mục con nhỏ hơn.



3.5.- Trực quan hóa dữ liệu và mối quan hệ phức tạp (Visualizing Complex Data and Relations)

- Trong những ngày đầu, kỹ thuật trực quan hóa chủ yếu dành cho dữ liệu số. Gần đây, ngày càng có nhiều dữ liệu phi số, chẳng hạn như văn bản và mạng xã hội. Trực quan hóa và phân tích dữ liệu như vậy thu hút rất nhiều sự quan tâm.
- Có nhiều kỹ thuật trực quan mới dành riêng cho các loại dữ liệu này.

3. Trực quan hóa dữ liệu (Data Visualization)

3.5.- *Trực quan hóa dữ liệu và mối quan hệ phức tạp (Visualizing Complex Data and Relations)*

- Có nhiều kỹ thuật trực quan mới dành riêng cho các loại dữ liệu này.

- **Tag cloud** (đám mây thẻ):

- Nhiều người trên Web gắn thẻ các đối tượng khác nhau như hình ảnh, bài viết trên blog và đánh giá sản phẩm.
- *tag cloud* là hình ảnh trực quan về số liệu thống kê của các thẻ do người dùng tạo. Trong tag cloud, các thẻ được liệt kê theo thứ tự bảng chữ cái hoặc theo thứ tự ưu tiên của người dùng. Tầm quan trọng của thẻ được biểu thị bằng kích thước phông chữ hoặc màu sắc.
- Hình minh họa cho thấy một đám mây thẻ để hiển thị các thẻ phổ biến được sử dụng trong một trang Web.

animals architecture art asia australia autumn baby band barcelona beach berlin bike bird
birds birthday black blackandwhite blue bw california canada canon car cat
chicago china christmas church city clouds color concert cute dance day de dog
england europe fall family fashion festival film florida flower flowers food
football france friends fun garden geotagged germany girl girls graffiti green
halloween hawaii holiday home house india iphone ireland island italia italy japan july kids la
lake landscape light live london love macro me mexico model mountain mountains museum
music nature new newyork newyorkcity night nikon nyc ocean old paris
park party people photo photography photos portrait red river rock san
sanfrancisco scotland sea seattle show sky snow spain spring street summer
sun sunset taiwan texas thailand tokyo toronto tour travel tree trees trip uk urban
usa vacation washington water wedding white winter yellow york zoo

3. Trực quan hóa dữ liệu (Data Visualization)

3.5.- *Trực quan hóa dữ liệu và mối quan hệ phức tạp (Visualizing Complex Data and Relations)*

- **Tag cloud** (đám mây thẻ):
 - ▣ Các đám mây thẻ thường được sử dụng theo hai cách:
 - Sử dụng kích thước của thẻ để biểu thị số lần thẻ được áp dụng cho mục này bởi những người dùng khác nhau.
 - Sử dụng kích thước của thẻ để biểu thị số lượng mục mà thẻ đã được áp dụng, tức là mức độ phổ biến của thẻ.

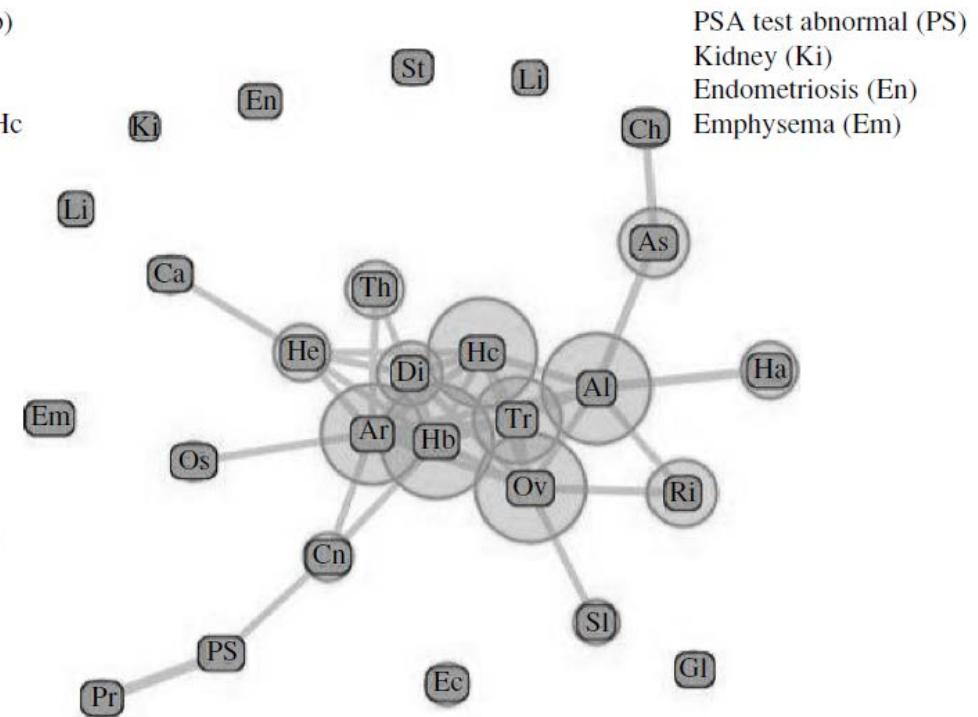
animals architecture art asia australia autumn baby band barcelona beach berlin bike bird
birds birthday black blackandwhite blue bw california canada canon car cat
chicago china christmas church city clouds color concert cute dance day de dog
england europe fall family fashion festival film florida flower flowers food
football france friends fun garden geotagged germany girl girls graffiti green
halloween hawaii holiday home house india iphone ireland island italia italy japan july kids la
lake landscape light live london love macro me mexico model mountain mountains museum
music nature new newyork newyorkcity night nikon nyc ocean old paris
park party people photo photography photos portrait red river rock san
sanfrancisco scotland sea seattle show sky snow spain spring street summer
sun sunset taiwan texas thailand tokyo toronto tour travel tree trees trip uk urban
usa vacation washington water wedding white winter yellow york zoo

3.5.- Trực quan hóa dữ liệu và mối quan hệ phức tạp (Visualizing Complex Data and Relations)

- **Biểu đồ ảnh hưởng của bệnh tật**

- Để hình dung mối tương quan giữa các bệnh tật.
- Các nút trong biểu đồ là các bệnh.
- Kích thước của mỗi nút tỷ lệ thuận với mức độ phổ biến của bệnh tương ứng.
- Hai nút được liên kết bằng một cạnh nếu các bệnh tương ứng có mối tương quan chặt chẽ. Độ rộng của một cạnh tỷ lệ thuận với độ mạnh của mô hình tương quan của hai bệnh tương ứng.

High blood pressure (Hb)
Allergies (Al)
Overweight (Ov)
High cholesterol level (Hc)
Arthritis (Ar)
Trouble seeing (Tr)
Risk of diabetes (Ri)
Asthma (As)
Diabetes (Di)
Hayfever (Ha)
Thyroid problem (Th)
Heart disease (He)
Cancer (Cn)
Sleep disorder (Sl)
Eczema (Ec)
Chronic bronchitis (Ch)
Osteoporosis (Os)
Prostate (Pr)
Cardiovascular (Ca)
Glaucoma (Gl)
Stroke (St)
Liver condition (Li)



NỘI DUNG CHƯƠNG 3

1. Đối tượng dữ liệu và kiểu thuộc tính
2. Mô tả các thống kê cơ bản
3. Trực quan hóa dữ liệu
4. Đo lường sự tương đồng và khác biệt của dữ liệu
5. Bài tập

4. ĐO LƯỜNG SỰ TƯƠNG ĐỒNG VÀ KHÁC BIỆT CỦA DỮ LIỆU

(Measuring Data Similarity and Dissimilarity)

- Ma trận dữ liệu (*the data matrix*) và ma trận sai phân (*dissimilarity matrix*)
- Đo lường thuộc tính danh nghĩa (*nominal attributes*)
- Đo lường thuộc tính nhị phân (*binary attributes*)
- Đo lường thuộc tính số (*numeric attributes*)
- Đo lường thuộc tính thứ tự (*ordinal attributes*)
- Đo lường thuộc tính hỗn hợp (*mixed attributes*)
- Đo lường với Độ tương tự cosin (*Cosine Similarity*)

4.1. Ma trận dữ liệu (*the data matrix*) và ma trận sai phân (*dissimilarity matrix*)

- Trong thực tế, các đối tượng đều được mô tả bằng nhiều thuộc tính. Vì vậy, ta cần thay đổi cách ký hiệu. Giả sử có n đối tượng (ví dụ: người, vật phẩm hoặc khóa học) được mô tả bởi p thuộc tính (còn gọi là số đo hoặc đặc điểm, chẳng hạn như tuổi, chiều cao, cân nặng hoặc giới tính).
- Các đối tượng là $x_1 = (x_{11}, x_{12}, \dots, x_{1p})$, $x_2 = (x_{21}, x_{22}, \dots, x_{2p})$, v.v..., trong đó x_{ij} là giá trị cho đối tượng x_i của thuộc tính thứ j . Để ngắn gọn, sau đây ta gọi đối tượng x_i là đối tượng i . Các đối tượng có thể là các bộ dữ liệu trong cơ sở dữ liệu quan hệ và còn được gọi là mẫu dữ liệu (*data samples*) hoặc vector đặc trưng (*feature vectors*).

4.1. Ma trận dữ liệu (*the data matrix*) và ma trận sai phân (*dissimilarity matrix*)

- **Cấu trúc dữ liệu:** Các thuật toán phân cụm dựa trên bộ nhớ chính và các thuật toán lân cận gần nhất thường hoạt động trên một trong hai cấu trúc dữ liệu sau:
 - **Datamatrix** (ma trận dữ liệu hoặc cấu trúc theo từng thuộc tính của đối tượng - *object-by-attribute structure*):
 - Cấu trúc này lưu trữ n đối tượng dữ liệu dưới dạng bảng quan hệ hoặc ma trận n -by- p (n đối tượng $\times p$ thuộc tính):

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Mỗi hàng tương ứng với một đối tượng. Là một phần trong ký hiệu của ta, có thể sử dụng f để lập chỉ mục thông qua các thuộc tính p .

4.1. Ma trận dữ liệu (the data matrix) và ma trận sai phân (dissimilarity matrix)

- **Cấu trúc dữ liệu:** thường dùng trong các thuật toán lân cận gần nhất:
- **Dissimilarity matrix** (Ma trận sai phân hoặc cấu trúc từng đối tượng - *object-by-object structure*): Cấu trúc này lưu trữ một tập hợp các giá trị gần đúng có sẵn cho tất cả các cặp n đối tượng. Nó thường được biểu diễn bằng bảng *n-by-n*:

0					
d(2,1)	0				
d(3,1)	d(3,2)	0			
...	0		
...	0	
d(n,1)	d(n,2)	d(n,3)	...	d(n-1)	0

trong đó:

- $d(i, j)$ là độ khác nhau được đo hoặc “sự khác biệt” giữa các đối tượng i và j . Nói chung, $d(i, j)$ là một số không âm, gần bằng 0 khi các đối tượng i và j rất giống nhau hoặc tỷ lệ “gần” nhau và càng lớn thì chúng càng khác nhau.
- Lưu ý rằng $d(i, i) = 0$; nghĩa là, sự khác biệt giữa một đối tượng và chính nó là 0.
- Do ma trận đang xét là ma trận vuông ($n \times n$) nên $d(i, j) = d(j, i)$. Nên để dễ đọc, ở đây không hiển thị các phần tử $d(i, j)$.

4.1. Ma trận dữ liệu (*the data matrix*) và ma trận sai phân (*dissimilarity matrix*)

- Độ đo độ tương tự thường có thể được biểu diễn dưới dạng hàm của độ đo độ khác biệt. Ví dụ, đối với dữ liệu danh nghĩa,

$$sim(i, j) = 1 - d(i, j) \quad \text{Công thức 1-8}$$

trong đó: $sim(i, j)$ là độ tương tự giữa đối tượng i và j .

- **Ma trận dữ liệu** (*data matrix*) được tạo thành từ hai thực thể hoặc “thứ”, cụ thể là hàng (đối với đối tượng) và cột (đối với thuộc tính). Vì vậy, ma trận dữ liệu thường được gọi là ma trận hai chế độ (*two-mode matrix*).
- **Ma trận sai phân** (*dissimilarity matrix*) so sánh sự khác biệt giữa 2 đối tượng thuộc cùng một loại thực thể do đó được gọi là ma trận một chế độ (*one-mode matrix*). Nhiều thuật toán phân cụm và láng giềng gần nhất hoạt động trên ma trận sai phân.

Dữ liệu ở dạng **ma trận dữ liệu** có thể được chuyển đổi thành **ma trận sai phân** trước khi áp dụng các thuật toán đó.

4.2. Đo lường thuộc tính Danh nghĩa (*nominal attributes*)

- Một thuộc tính danh nghĩa có thể có hai hoặc nhiều trạng thái (*states*). Ví dụ: màu bản đồ là một thuộc tính danh nghĩa có thể có năm trạng thái: đỏ, vàng, lục, hồng và xanh lam.
- Đặt số trạng thái của một thuộc tính danh nghĩa là m . Các trạng thái có thể được biểu thị bằng các chữ cái, ký hiệu hoặc một tập hợp số nguyên, chẳng hạn như $1, 2, \dots, m$. Lưu ý rằng các số nguyên đó chỉ được sử dụng để xử lý dữ liệu và không đại diện cho bất kỳ thứ tự cụ thể nào.
- Tính toán sự khác biệt giữa 2 đối tượng i và j được mô tả bằng các thuộc tính danh nghĩa được tính dựa trên tỷ lệ không khớp (*ratio of mismatches*):

$$d(i,j) = \frac{p-m}{p} \quad \text{Công thức 1-9}$$

trong đó

- m là số lượng kết quả trùng khớp (tức là số lượng thuộc tính mà i và j ở cùng trạng thái)
- p là tổng số thuộc tính mô tả các đối tượng.
- Các trọng số có thể được gán để tăng tác dụng của m hoặc để gán trọng số lớn hơn cho các kết quả trùng khớp trong các thuộc tính có số lượng trạng thái lớn hơn.

4.2. Đo lường thuộc tính Danh nghĩa (nominal attributes)

- Ví dụ Cho dữ liệu như bảng sau. Với thuộc tính test-1 là thuộc tính dạng danh nghĩa (nominal)

Object Identifier	test-1 (nominal)
1	code A
2	code B
3	code C
4	code A

Cần xây dựng ma trận sau:

$$\begin{bmatrix} 0 & & & \\ d(2, 1) & 0 & & \\ d(3, 1) & d(3, 2) & 0 & \\ d(4, 1) & d(4, 2) & d(4, 3) & 0 \end{bmatrix}$$

Dựa trên công thức: $d(i,j)=\frac{p-m}{p}$

Vì chỉ có một thuộc tính danh nghĩa, nên $p = 1$. Và xác lập giá trị cho $d(i,j)$ như sau:

- $d(i, j) = 0$ nếu đối tượng i và j khớp nhau (giống nhau)
 - $d(i, j) = 1$ nếu đối tượng khác nhau.
- $$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Với cách tính này, ta nhận được kết quả như sau:

Nhận xét: tất cả các đối tượng đều khác nhau ngoại trừ đối tượng 1 và 4 (cùng có giá trị là “code A”) nên $d(4,1) = 0$.

- Ngoài ra, độ tương tự có thể được tính là

$$\text{sim}(i, j) = 1 - d(i, j) = \frac{m}{p}$$

Công thức 1-10

4.2. Đo lường thuộc tính Danh nghĩa (*nominal attributes*)

- Khoảng cách gần giữa các đối tượng được mô tả bởi các thuộc tính danh nghĩa có thể được tính toán bằng cách sử dụng sơ đồ mã hóa thay thế (*alternative encoding scheme*). Các thuộc tính danh nghĩa có thể được mã hóa bằng các **thuộc tính nhị phân bất đối xứng** (*asymmetric binary attributes*) bằng cách tạo thuộc tính nhị phân mới cho mỗi trạng thái m . Đối với một đối tượng có giá trị trạng thái nhất định, thuộc tính nhị phân biểu thị trạng thái đó được đặt thành 1, trong khi các thuộc tính nhị phân còn lại được đặt thành 0.

Ví dụ: để mã hóa màu bản đồ thuộc tính danh nghĩa, một thuộc tính nhị phân có thể được tạo cho mỗi thuộc tính năm màu được liệt kê trước đó. Đối với một đối tượng có màu vàng, thuộc tính màu vàng được đặt thành 1, trong khi bốn thuộc tính còn lại được đặt thành 0. Các thước đo độ gần cho dạng mã hóa này có thể được tính bằng các phương pháp được thảo luận trong tiêu mục tiếp theo.

4.3. Đo lường thuộc tính nhị phân (*binary attributes*)

Thuộc tính nhị phân chỉ có một trong hai trạng thái: 0 và 1, trong đó 0 có nghĩa là thuộc tính không có và 1 có nghĩa là thuộc tính đó có mặt.

Ví dụ, với thuộc tính người hút thuốc mô tả một bệnh nhân, 1 chỉ ra rằng bệnh nhân hút thuốc, trong khi 0 chỉ ra rằng bệnh nhân không hút thuốc.

Việc xử lý các thuộc tính nhị phân dưới dạng số có thể gây hiểu nhầm. Do đó, các phương pháp dành riêng cho dữ liệu nhị phân là cần thiết để tính toán độ khác nhau.

4.3. Đo lường thuộc tính nhị phân (binary attributes)

- Tính toán sự khác biệt giữa hai thuộc tính nhị phân bằng ma trận sai phân từ dữ liệu nhị phân đã cho. Nếu tất cả các thuộc tính nhị phân được coi là có cùng trọng số, ta có bảng thống kê trong ma trận 2×2 :

trong đó:

- q là số thuộc tính bằng 1 cho cả hai đối tượng i và j,
- r là số thuộc tính bằng nhau 1 đối với đối tượng i nhưng bằng 0 đối với đối tượng j,
- s là số thuộc tính bằng 0 đối với đối tượng i nhưng bằng 1 đối với đối tượng j
- t là số thuộc tính bằng 0 đối với cả đối tượng i và j.

Bảng thống kê cho các thuộc tính nhị phân

		Object j		sum
		1	0	
Object i	1	q	r	q + r
	0	s	t	s + t
	sum	q + s	r + t	p

4. Đo lường sự tương đồng và khác biệt của dữ liệu (Measuring Data Similarity and Dissimilarity)

4.3. Đo lường thuộc tính nhị phân (binary attributes)

		Object j		sum
		1	0	
Object i	1	q	r	q + r
	0	s	t	s + t
sum		q + s	r + t	p

- Tính độ khác biệt giữa hai thuộc tính nhị phân i và j

- Đối với thuộc tính nhị phân đối xứng (symmetric binary attributes) - mỗi trạng thái đều có giá trị như nhau):

$$d(i, j) = \frac{r+s}{q+r+s+t}$$

Công thức 11

- Đối với thuộc tính nhị phân KHÔNG đối xứng (asymmetric binary attributes) - sự phù hợp của hai số 1 (kết quả khớp dương) khi đó được coi là có ý nghĩa hơn so với sự phù hợp của hai số 0 (kết quả khớp âm) \Rightarrow do đó kết quả khớp âm t là không quan trọng. Như xét nghiệm bệnh gồm kết quả dương tính (1) và âm tính (0):

$$d(i, j) = \frac{r+s}{q+r+s}$$

Công thức 12

- Tính độ tương đồng giữa hai thuộc tính nhị phân i và j

- Đối với thuộc tính nhị phân KHÔNG đối xứng (được gọi là hệ số Jaccard):

$$sim(i, j) = \frac{q}{q+r+s} = 1 - d(i, j)$$

Công thức 13

Khi cả hai thuộc tính nhị phân **đối xứng** và **bất đối xứng** xuất hiện trong cùng một tập dữ liệu, có thể áp dụng cách tiếp cận thuộc tính **hỗn hợp**

4.3. Đo lường thuộc tính nhị phân (binary attributes)

- Ví dụ: Giả sử bảng hồ sơ bệnh nhân chứa các thuộc tính tên (*name*), giới tính (*gender*), sốt (*fever*), ho (*cough*), *test-1*, *test-2*, *test-3* và *test-4*, trong đó *name* là định danh đối tượng, *gender* là thuộc tính đối xứng và các thuộc tính còn lại là nhị phân bất đối xứng.

<i>name</i>	<i>gender</i>	<i>fever</i>	<i>cough</i>	<i>test-1</i>	<i>test-2</i>	<i>test-3</i>	<i>test-4</i>
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N

Đối với các giá trị thuộc tính không đối xứng, đặt:

- 1 cho Y (Yes) và P (Positive)
- 0 cho N (No hoặc Negative)
- Đối với thuộc tính nhị phân đối xứng (*symmetric binary attributes*) - mỗi trạng thái đều có giá trị như nhau):

<i>name</i>	<i>fever</i>	<i>cough</i>	<i>test-1</i>	<i>test-2</i>	<i>test-3</i>	<i>test-4</i>
Jack	1	0	1	0	0	0
Jim	1	1	0	0	0	0
Mary	1	0	1	0	1	0
...

4. Đo lường sự tương đồng và khác biệt của dữ liệu (Measuring Data Similarity and Dissimilarity)

4.3. Đo lường thuộc tính nhị phân (binary attributes)

- Ví dụ: Giả sử khoảng cách giữa các đối tượng (bệnh nhân) chỉ được tính dựa trên các thuộc tính bất đối xứng. Theo công thức 12, khoảng cách giữa mỗi cặp của ba bệnh nhân Jack, Mary và Jim là:

name	gender	fever	cough	test-1	test-2	test-3	test-4
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N

name	fever	cough	test-1	test-2	test-3	test-4
Jack	1	0	1	0	0	0
Jim	1	1	0	0	0	0
Mary	1	0	1	0	1	0
...

$d_{(Obj1, Obj2)}$		Obj1	
		1	0
Obj2	1	q	r
	0	s	t

$d_{(Jack, Jim)}$		Jim	
		1	0
Jack	1	1	1
	0	1	3

$$d_{(Jack, Jim)} = \frac{r + s}{q + r + s} = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$d_{(Jack, Mary)}$		Mary	
		1	0
Jack	1	2	0
	0	1	3

$$d_{(Jack, Mary)} = \frac{r + s}{q + r + s} = \frac{0 + 1}{1 + 1 + 1} = 0.33$$

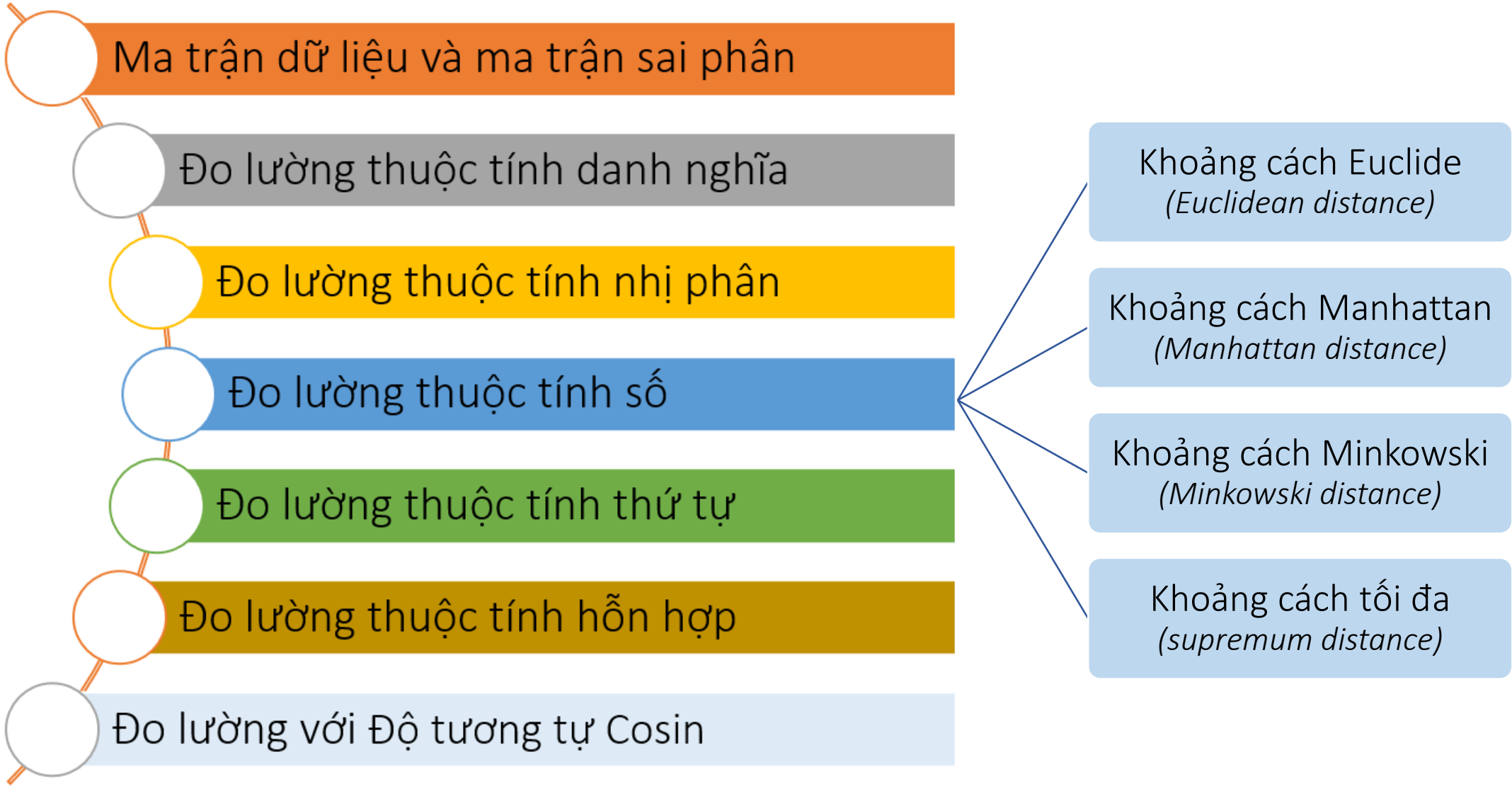
$d_{(Jim, Mary)}$		Mary	
		1	0
Jim	1	1	1
	0	2	2

$$d_{(Jim, Mary)} = \frac{r + s}{q + r + s} = \frac{1 + 2}{1 + 2 + 1} = 0.75$$

KL: dựa trên độ khác biệt

- Jim và Mary **khó** có thể mắc bệnh tương tự vì có giá trị khác biệt cao (=0.75).
- Jack và Mary: **có** nhiều khả năng mắc bệnh tương tự nhất (=0.33)

4.4. Đo lường thuộc tính số (numeric attributes)



4.4. Đo lường thuộc tính số (*numeric attributes*)

- Việc chuẩn hóa dữ liệu cố gắng cung cấp cho tất cả các thuộc tính một trọng số bằng nhau. Nó có thể hữu ích hoặc không hữu ích trong một ứng dụng cụ thể.
- Trong một số trường hợp, dữ liệu được chuẩn hóa trước khi áp dụng tính toán khoảng cách. Điều này liên quan đến việc chuyển đổi dữ liệu để nằm trong phạm vi nhỏ hơn (*smaller range*) hoặc phổ biến hơn (*common range*), chẳng hạn như $[-1; 1]$ hoặc $[0,0; 1,0]$.
- Nói chung, việc thể hiện một thuộc tính theo đơn vị nhỏ hơn sẽ dẫn đến phạm vi lớn hơn cho thuộc tính đó và do đó có xu hướng mang lại cho các thuộc tính đó hiệu ứng (*effect*) hoặc “trọng lượng” (*weight*) lớn hơn.

4.4. Đo lường thuộc tính số (*numeric attributes*)

4.4.1. Khoảng cách Euclide (*Euclidean distance*)

- Thước đo khoảng cách phổ biến nhất là khoảng cách *Euclide*, tức là đường thẳng (*straight line*) hoặc “theo đường chim bay” (*as the crow flies*).
- Cho $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ và $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ là hai đối tượng được mô tả bởi p thuộc tính số.
- Khoảng cách *Euclide* giữa các đối tượng i và j được định nghĩa:

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$

Công thức 14

4.4. Đo lường thuộc tính số (*numeric attributes*)

4.4.2. Khoảng cách Manhattan (*Manhattan distance*)

- Một thước đo phổ biến khác là khoảng cách *Manhattan* (hoặc *city block* - khối thành phố), được đặt tên như vậy vì nó là khoảng cách tính theo khối giữa hai điểm bất kỳ trong thành phố (chẳng hạn như 2 khối phía dưới và 3 khối phía trên để có tổng cộng 5 khối).
- Khoảng cách *Manhattan* giữa các đối tượng i và j được định nghĩa:

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

Công thức 15

4. Đo lường sự tương đồng và khác biệt của dữ liệu (*Measuring Data Similarity and Dissimilarity*)

4.4. Đo lường thuộc tính số (*numeric attributes*)

4.4.2. Khoảng cách Manhattan (*Manhattan distance*)

- Cả khoảng cách *Euclidean* và *Manhattan* đều thỏa mãn các tính chất toán học sau:
 - ***Non-negativity*** (không âm): $d(i, j) \geq 0$: khoảng cách là số không âm.
 - ***Identity of indiscernibles***: $d(i, i) = 0$: khoảng cách của một vật tới chính nó là 0.
 - ***Symmetry*** (tính đối xứng): $d(i, j) = d(j, i)$: khoảng cách là một hàm đối xứng.
 - ***Triangle inequality*** (bất đẳng thức tam giác): $d(i, j) \leq d(i, k) + d(k, j)$: đi thẳng từ vật i đến vật j trong không gian không khác gì việc đi đường vòng qua bất kỳ vật k nào khác.

Một thước đo thỏa mãn các điều kiện này được gọi là metric.

Lưu ý: đặc tính không âm được bao hàm trong ba đặc tính còn lại.

4. Đo lường sự tương đồng và khác biệt của dữ liệu (*Measuring Data Similarity and Dissimilarity*)

4.4. Đo lường thuộc tính số (*numeric attributes*)

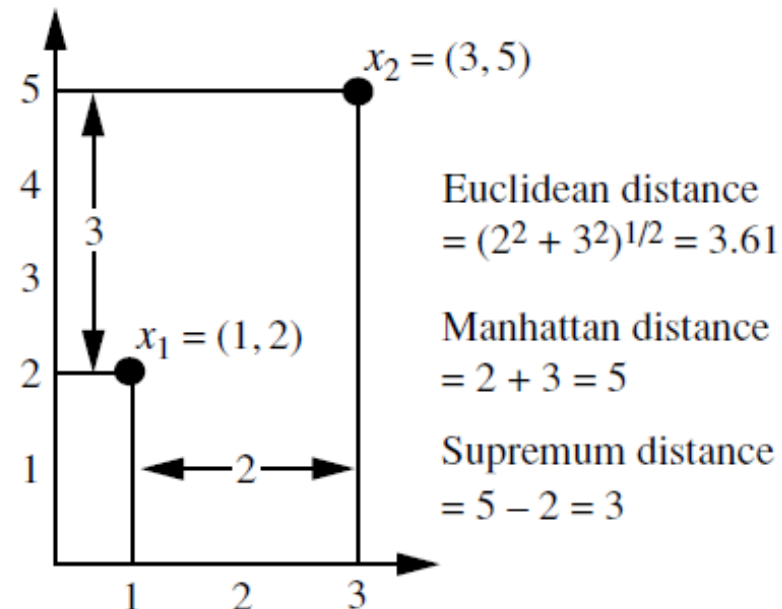
4.4.2. Khoảng cách Manhattan (*Manhattan distance*)

- Ví dụ: Cho $x_1 = (1, 2)$ và $x_2 = (3, 5)$.
- Khoảng cách Euclide giữa hai điểm x_1 và x_2 :

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2} = \sqrt{(1 - 3)^2 + (2 - 5)^2} = \sqrt{2^2 + 3^2} = 3.61$$

- Khoảng cách Manhattan giữa hai điểm x_1 và x_2 :

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| = |1 - 3| + |2 - 5| = 5$$



4. Đo lường sự tương đồng và khác biệt của dữ liệu (*Measuring Data Similarity and Dissimilarity*)

4.4. Đo lường thuộc tính số (*numeric attributes*)

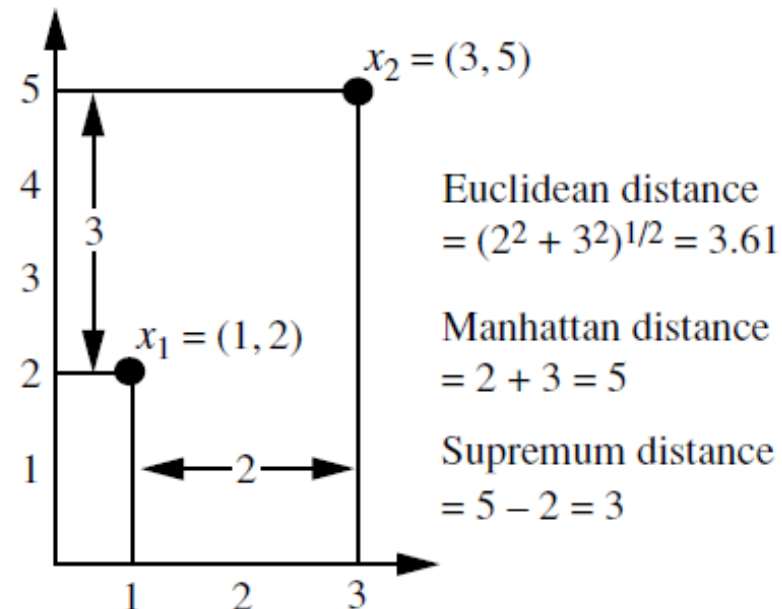
4.4.2. Khoảng cách Manhattan (*Manhattan distance*)

- Ví dụ: Cho $x_1 = (1, 2)$ và $x_2 = (3, 5)$.
- Khoảng cách Euclide giữa hai điểm x_1 và x_2 :

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2} = \sqrt{(1 - 3)^2 + (2 - 5)^2} = \sqrt{2^2 + 3^2} = 3.61$$

- Khoảng cách Manhattan giữa hai điểm x_1 và x_2 :

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| = |1 - 3| + |2 - 5| = 5$$



4.4. Đo lường thuộc tính số (*numeric attributes*)

4.4.3. Khoảng cách Minkowski (*Minkowski distance*)

- Khoảng cách *Minkowski* là sự khái quát hóa của khoảng cách *Euclidean* và *Manhattan*. Khoảng cách này được định nghĩa như sau:

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ip} - x_{jp}|^h} \quad \text{Công thức 1-16}$$

trong đó h là một số thực sao cho $h \geq 1$.

- Khoảng cách như vậy còn được gọi là chuẩn L_p (L_p norm). Với ký hiệu p là số lượng thuộc tính ám chỉ ký hiệu về h .
 - Khi $h=1$: khoảng cách *Minkowski* biểu thị khoảng cách *Manhattan* khi $h = 1$ (tức là L_1)
 - Khi $h=2$: khoảng cách *Minkowski* biểu thị khoảng cách *Euclidean* khi $h = 2$ (tức là L_2).

4.4. Đo lường thuộc tính thứ tự (*ordinal attributes*)

- Các giá trị của thuộc tính thứ tự tuy có ý nghĩa về có thứ tự hoặc thứ hạng của chúng, tuy nhiên độ lớn giữa các giá trị kế tiếp vẫn chưa được xác định. Ví dụ thuộc tính *size* bao gồm các giá trị *small*, *medium*, *large*.
- Các thuộc tính thứ tự cũng có thể thu được từ việc rời rạc hóa các thuộc tính số bằng cách chia phạm vi giá trị thành một số hữu hạn các loại. Các loại này được tổ chức thành cấp bậc. Nghĩa là, phạm vi của thuộc tính số có thể được ánh xạ tới thuộc tính thứ tự f có trạng thái M_f .

4.4. Đo lường thuộc tính thứ tự (ordinal attributes)

- Ví dụ:

- Thuộc tính về Phạm vi nhiệt độ (tính bằng độ C - Celsius) : được chia tỷ lệ theo khoảng có thể được tổ chức thành các trạng thái sau:

Temperature	Celsius
very cold	<-30
cold	-30 → -10
moderate	-10 → +10
warm	+10 → +30
hot	>30

- Hoặc người ta chia mức độ sốt ở trẻ thành 4 mức như sau:

Mức độ sốt	Vị trí lấy nhiệt độ	
	ở nách	ở hậu môn
Sốt nhẹ	37.5 – 38 ^o C	38-38.5 ^o C
Sốt vừa	38-38.5 ^o C	38.5- 39 ^o C
Sốt cao	38.5- 39 ^o C	39- 40 ^o C
Sốt rất cao	> 39 ^o C	> 40 ^o C

4. Đo lường sự tương đồng và khác biệt của dữ liệu (*Measuring Data Similarity and Dissimilarity*)

4.4. Đo lường thuộc tính thứ tự (*ordinal attributes*)

- Gọi M là số trạng thái có thể có của một thuộc tính *Thứ tự*. Các trạng thái có thứ tự này xác định thứ hạng $1, \dots, M_f$.
- Xử lý các thuộc tính *Thứ tự*:
 - Được tính toán khá giống với các thuộc tính số khi cần xác định độ khác nhau giữa các đối tượng.
 - Giả sử f là một thuộc tính từ một tập các thuộc tính *Thứ tự* mô tả n đối tượng. Việc tính toán độ sai phân (hay sự khác biệt - *dissimilarity*) đối với f bao gồm các bước sau:
 - i. Giá trị của f cho đối tượng thứ i là x_{if} và f có các trạng thái có thứ tự M_f , biểu thị hạng $1, \dots, M_f$. Thay thế mỗi x_{if} bằng thứ hạng tương ứng của nó, $r_{if} \in \{1, \dots, M_f\}$.
 - ii. Vì mỗi thuộc tính thứ tự có thể có số trạng thái khác nhau nên thường cần phải ánh xạ phạm vi của từng thuộc tính vào $[0,0; 1,0]$ để mỗi thuộc tính có trọng số bằng nhau. Thực hiện chuẩn hóa dữ liệu đó bằng cách thay thế thứ hạng của đối tượng thứ i trong thuộc tính thứ f bằng

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

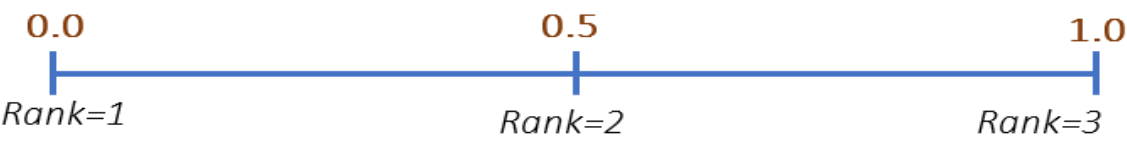
Công thức 19

- iii. Độ khác biệt có thể được tính toán bằng cách sử dụng bất kỳ thước đo khoảng cách nào đã được mô tả cho các thuộc tính số, sử dụng z_{if} để biểu thị giá trị f cho đối tượng thứ i .

4. Đo lường sự tương đồng và khác biệt của dữ liệu (Measuring Data Similarity and Dissimilarity)

4.4. Đo lường thuộc tính thứ tự (ordinal attributes)

- Ví dụ Giả sử rằng ta có dữ liệu mẫu được hiển thị trước đó trong bảng
 - i. Thay thế từng giá trị cho thuộc tính test-2 bằng thứ hạng của nó, với thứ hạng này do ta tự quy ước. Giả sử có quy ước như sau:
⇒ 4 đối tượng sẽ được xếp hạng tương ứng lần lượt là 1, 3 , 2 và 1.
 - ii. Chia các thứ hạng vào các mốc giá trị trong khoảng [0,0; 1,0] sao cho phạm vi của chúng bằng nhau.
⇒ Gán hạng 1 đến giá trị 0,0, hạng 2 đến 0,5 và hạng 3 đến 1,0.



iii. Có thể sử dụng khoảng cách Euclide, dẫn đến ma trận sai phân:

KL: Do $d(2,1)=1.0$ và $d(4,2)=1.0 \Rightarrow$ các cặp đối tượng 1 và 2; 2 và 4 là khác nhau nhất. Điều này có ý nghĩa trực quan vì đối tượng 1 và 4 đều là xuất sắc nên nằm ở 1 đầu, còn đối tượng 2 nằm ở đầu đối diện.

Object Identifier	test-2 (ordinal)
1	excellent
2	fair
3	good
4	excellent

(Dữ liệu ban đầu)

Ranking	Value
1	excellent
2	good
3	fair

(Quy ước về thứ hạng)

Object Identifier	test-2 (ordinal)	Ranking
1	excellent	1
2	fair	3
3	good	2
4	excellent	1

(Xếp hạng cho dữ liệu gốc)

Object Identifier	1	2	3	4
1	0			
2	1.0	0		
3	0.5	0.5	0	
4	0	1.0	0.5	0

4.5. Đo lường thuộc tính hỗn hợp (*mixed attribute types*)

- Trong nhiều cơ sở dữ liệu thực, các đối tượng được mô tả bằng sự kết hợp của nhiều kiểu thuộc tính.
- Tính toán sự khác biệt giữa các đối tượng thuộc loại thuộc tính hỗn hợp:
 - **Cách 1:** nhóm từng loại thuộc tính lại với nhau, thực hiện phân tích khai thác dữ liệu riêng biệt (ví dụ: phân cụm) cho từng loại. Điều này là khả thi nếu những phân tích này thu được kết quả tương thích. Tuy nhiên, trong các ứng dụng thực tế, việc phân tích riêng biệt từng loại thuộc tính sẽ không tạo ra kết quả tương thích.
 - **Cách 2:** thực hiện một phân tích duy nhất dựa trên việc xử lý tất cả các loại thuộc tính cùng nhau. Cách này thường là cách tiếp cận thích hợp hơn. Như vậy, cần kết hợp các thuộc tính khác nhau thành một ma trận sai phân (dissimilarity matrix) duy nhất, đưa tất cả các thuộc tính có ý nghĩa vào một thang đo chung của khoảng $[0,0, 1,0]$.

4. Đo lường sự tương đồng và khác biệt của dữ liệu (*Measuring Data Similarity and Dissimilarity*)

4.5. Đo lường thuộc tính hỗn hợp (*mixed attribute types*)

- Giả sử tập dữ liệu chứa p thuộc tính thuộc loại hỗn hợp. Sự khác biệt $d(i, j)$ giữa các đối tượng i và j được định nghĩa là:

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}} \quad \text{Công thức 20}$$

trong đó giá trị của chỉ báo $\delta_{ij}^{(f)}$

- **= 0** nếu 1 trong 2 trường hợp sau:
 - (1) Thiếu x_{if} hoặc x_{jf} (nghĩa là không có phép đo thuộc tính f cho đối tượng i hoặc đối tượng j),
 - (2) Hoặc $x_{if} = x_{jf} = 0$ và thuộc tính (f) là nhị phân bất đối xứng.
- **= 1** trong các trường hợp còn lại.

Sự đóng góp của thuộc tính f vào sự khác biệt (d) giữa i và j (tức là $d_{ij}^{(f)}$) được tính toán phụ thuộc vào loại của nó:

- Nếu f là số (*numeric*): $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$, trong đó h chạy trên tất cả các đối tượng không thiếu (*nonmissing objects*) cho thuộc tính f .
- Nếu f là danh nghĩa (*nominal*) hoặc nhị phân (*binary*): $d_{ij}^{(f)} = 0$ nếu $x_{if} = x_{jf}$; ngược lại $d_{ij}^{(f)} = 1$.
- Nếu f là thứ tự (*ordinal*): tính thứ hạng r_{if} và $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, và coi z_{if} là số.

4.5. Đo lường thuộc tính hỗn hợp (mixed attribute types)

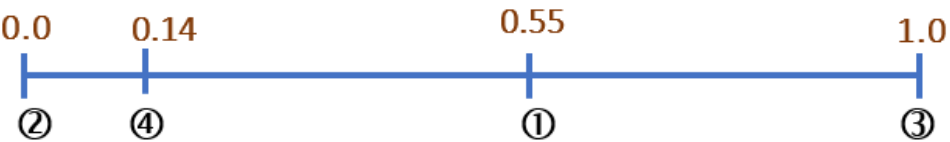
- Ví dụ: cho bảng dữ liệu như hình bên (trong đó các thuộc tính test-1 và test-2 đã được tính ma trận sai phân trong các ví dụ trước)
- Tính ma trận sai phân cho thuộc tính test-3 (kiểu dữ liệu số - *numeric*). Tức là tính $d_{ij}^{(3)}$. Theo trường hợp đối với các thuộc tính số, ta đặt $max_h x_h = 64$ và $min_h x_h = 22$. Sử dụng công thức 20 để chuẩn hóa các giá trị của ma trận sai phân.

(phân bố các giá trị lên trục \Rightarrow)

Ma trận sai phân thu được cho *test-3* \Rightarrow

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Object Identifier	test-3 (numeric)		Vị trí
1	45	$= (45-22)/(64-22) = 23/42$	0.55
2	22	$= (22-22)/(64-22) = 0/42$	0
3	64	$= (64-22)/(64-22) = 42/42$	1
4	28	$= (28-22)/(64-22) = 6/42$	0.14



Object Identifier	1	2	3	4
1	0			
2	0.55	0		
3	0.45	1.0	0	
4	0.41	0.14	0.86	0

4.5. Đo lường thuộc tính hỗn hợp (mixed attribute types)

- Ví dụ:
 - Sử dụng các ma trận sai phân cho ba thuộc tính vào công thức 20. Chỉ báo $d_{ij}^{(f)} = 1$ cho mỗi thuộc tính trong số ba thuộc tính, f.

□ Lần lượt tính giá trị cho từng phần tử trong ma trận sai phân. Ví dụ tính d(1,3) như sau:

Ma trận sai phân của các thuộc tính											
Thuộc tính test-1				Thuộc tính test-2				Thuộc tính test-3			
0				0				0			
1	0			1.0	0			0.55	0		
1	1	0		0.5	0.5	0		0.45	1.0	0	
0	1	1	0	0	1.0	0.5	0	0.41	0.14	0.86	0

$$d(1,3) = \frac{1(1) + 1(0.50) + 1(0.45)}{1 + 1 + 1} = 0.65$$

□ Thực hiện tương tự cho các d(ij) khác, ta thu được ma trận sai phân thu được cho dữ liệu được mô tả bởi ba thuộc tính của các loại hỗn hợp là

0			
0.85	0		
0.65	0.83	0	
0.13	0.71	0.79	0

4.6. Đo lường với Độ đo tương tự Cosin (Cosine Similarity)

- Một tài liệu có thể được biểu thị bằng hàng nghìn thuộc tính, mỗi thuộc tính ghi lại tần suất của một từ cụ thể (chẳng hạn như từ khóa - *keyword*) hoặc cụm từ (*phrase*) trong tài liệu. Vì vậy, mỗi tài liệu là một đối tượng được biểu diễn bằng cái gọi là vectơ tần suất (*term-frequency vector* hay còn được gọi là *vectơ tài liệu - Document Vector*).
- Ví dụ: trong bảng sau, ta thấy *Document1* chứa năm trường hợp của từ “*team*”, trong khi khúc côn cầu (*hockey*) xuất hiện ba lần. Từ huấn luyện viên (*coach*) không có trong toàn bộ *Document1*, được biểu thị bằng giá trị đếm bằng 0. Dữ liệu như vậy có thể rất bất đối xứng.

<i>Document</i>	<i>team</i>	<i>coach</i>	<i>hockey</i>	<i>baseball</i>	<i>soccer</i>	<i>penalty</i>	<i>score</i>	<i>win</i>	<i>loss</i>	<i>season</i>
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

4.6. Đo lường với Độ đo tương tự Cosin (*Cosine Similarity*)

- Các vector tần suất thuật ngữ thường rất dài và thưa thớt (*sparse* - nghĩa là chúng có nhiều giá trị 0). Các ứng dụng sử dụng các cấu trúc như vậy bao gồm:
 - Truy xuất thông tin (*information retrieval*)
 - Phân cụm tài liệu văn bản (*text document clustering*)
 - Phân loại sinh học (*biological taxonomy*)
 - Lập bản đồ đặc điểm gen (*gene feature mapping*).
 - Các thước đo khoảng cách truyền thống đã nghiên cứu không hoạt động tốt đối với dữ liệu số thưa thớt như vậy. Ví dụ: hai vector tần suất có thể có nhiều giá trị 0 chung, nghĩa là các tài liệu tương ứng không chia sẻ nhiều từ, nhưng điều này không làm cho chúng giống nhau.
- ⇒ Cần một biện pháp tập trung vào những từ chung của hai tài liệu và tần suất xuất hiện của những từ đó. Nói cách khác, cần một thước đo cho dữ liệu số bỏ qua các kết quả không trùng khớp.

4.6. Đo lường với Độ đo tương tự Cosin (*Cosine Similarity*)

- Độ tương tự cosine (*Cosine similarity*) là thước đo độ tương tự có thể được sử dụng để so sánh các tài liệu hoặc đưa ra thứ hạng các tài liệu đối với một vector từ (words) truy vấn nhất định. Cho x và y là hai vector để so sánh. Sử dụng thước đo cosin làm hàm tương tự, ta có:

$$\text{sim}(x, y) = \frac{x \cdot y}{\|x\| \times \|y\|}$$

Trong đó:

- $x \cdot y$ (từ đây trở đi sẽ được ký hiệu là $x^t \cdot y$) tương đương như việc nhân 2 ma trận $1 \times n * n \times 1$.
- $\|x\|$ là chuẩn *Euclidean* (*Euclidean norm*) của vector $x = (x_1, x_2, \dots, x_p)$,
được định nghĩa là $\sqrt{x_1^2 + x_2^2 + \dots + x_p^2}$. Về mặt khái niệm, đó là độ dài của vector.
- Tương tự, $\|y\|$ là chuẩn *Euclidean* của vector y .

4.6. Đo lường với Độ đo tương tự Cosin (*Cosine Similarity*)

- Số đo tính *Cosin* của góc giữa các vector x và y . Giá trị *cosine*
 - **=0** có nghĩa là hai vector vuông góc với nhau 90 độ (trực giao - *orthogonal*) và **không khớp** nhau (*have no match*).
 - **Càng gần 1** thì góc càng nhỏ và **độ khớp** (*match*) giữa các vector **càng lớn**.
- Lưu ý rằng vì độ đo tương tự *cosine* không tuân theo tất cả các tính chất xác định độ đo số liệu nên được gọi là độ đo phi hệ số (*nonmetric measure*).

4.6. Đo lường với Độ đo tương tự Cosin (Cosine Similarity)

- Ví dụ: Giả sử x và y là hai vector đầu tiên trong Bảng 4.

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Nghĩa là cần xác định xem $x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$ và $y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$ giống nhau thế nào? Sử dụng công thức 21 để tính độ tương tự cosine giữa x và y, ta có:

$$sim(x, y) = \frac{x \times y}{\|x\| \times \|y\|} = x^t * y = (5 \times 3) + (0 \times 0) + (3 \times 2) + (0 \times 0) + (2 \times 1) + (0 \times 1) + (0 \times 0) + (2 \times 1) + (0 \times 0) + (0 \times 1)$$

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = \sqrt{42} = 6.48$$

$$\|y\| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = \sqrt{17} = 4.12$$

$$sim(x, y) = \frac{x \times y}{\|x\| \times \|y\|} = \frac{25}{6.48 \times 4.12} = 0.94$$

Nếu ta sử dụng thước đo độ tương tự Cosine để so sánh 2 document này thì chúng sẽ được coi là khá giống nhau.

4.6. Đo lường với Độ đo tương tự Cosin (*Cosine Similarity*)

- Khi các thuộc tính có giá trị nhị phân, độ tương tự *Cosine* có thể được hiểu theo các đặc điểm hoặc thuộc tính được chia sẻ. Giả sử một đối tượng x sở hữu thuộc tính thứ i nếu $x_i = 1$. Khi đó $x^t \cdot y$ là số thuộc tính được sở hữu (tức là được chia sẻ) bởi cả x và y và $|x|.|y|$ là giá trị trung bình hình học (*geometric mean*) của số thuộc tính mà x và y sở hữu (*possessed*). Vì vậy, $sim(x, y)$ là thước đo mức độ sở hữu tương đối của các thuộc tính chung.
- Một biến thể đơn giản của độ tương tự *Cosine* cho kịch bản trước đó là

$$sim(x, y) = \frac{x \times y}{(x \times x) + (y \times y) - (x \times y)}$$

Đây là tỷ lệ giữa số thuộc tính được chia sẻ bởi x và y với số thuộc tính mà x hoặc y sở hữu. Hàm này, được gọi là *hệ số Tanimoto* (*Tanimoto coefficient*) hoặc *khoảng cách Tanimoto* (*Tanimoto distance*), thường được sử dụng trong tìm kiếm thông tin và phân loại sinh học.

NỘI DUNG CHƯƠNG 3

1. Đối tượng dữ liệu và kiểu thuộc tính
2. Mô tả các thống kê cơ bản
3. Trực quan hóa dữ liệu
4. Đo lường sự tương đồng và khác biệt của dữ liệu
5. Bài tập

5. BÀI TẬP

- i. Yêu cầu chung cho nhóm bài tập này: tính toán các giá trị và vẽ đồ thị với những yêu cầu sau đây cho từng trường hợp:
- a. Tính các giá trị Mean, Median, Mode? Cho biết mode của dữ liệu là gì? (ví dụ: bimodal, trimodal, v.v.).
 - b. Midgange của dữ liệu là gì?
 - c. Tính các phân vị thứ 32 và 87?
 - d. Tính giá trị của các tứ phân vị thứ nhất (Q1) và tứ phân vị thứ ba (Q3) của dữ liệu.
 - e. Đưa ra bản tóm tắt năm số (five-number summary) của dữ liệu.
 - f. Tính phương sai
 - g. Tính độ lệch chuẩn.
 - h. Lần lượt vẽ từng biểu đồ sau của dữ liệu: boxplot, biểu đồ phân tán (*scatter plot*), biểu đồ q-q (*q-q plot*), Histogram.
- 1) Trường hợp 1: Cho phân bố về cân nặng như sau:



5. BÀI TẬP

i. Yêu cầu chung cho nhóm bài tập này: tính toán các giá trị và vẽ đồ thị với những yêu cầu sau đây cho từng trường hợp:

2) Trường hợp 2: Giả sử dữ liệu phân tích bao gồm thuộc tính tuổi. Giá trị tuổi của các bộ dữ liệu là (theo thứ tự tăng dần) 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.

3) Trường hợp 2: Đây là số liệu thống kê về thời gian chờ đợi của 10 bệnh nhân để được bác sĩ khám bệnh:

5 mins	17 mins	8 mins	2 mins	55 mins
9 mins	22 mins	11mins	16 mins	5 mins

4) Trường hợp 3: Giả sử rằng một bệnh viện đã kiểm tra dữ liệu về độ tuổi và lượng mỡ trong cơ thể của 18 người lớn được chọn ngẫu nhiên và cho kết quả như sau:

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

5. BÀI TẬP

ii. Cho dữ liệu về 4 vận động viên nước Mỹ có thành tích tốt nhất khi tham gia thi đấu môn bơi lội tại Olympic Rio 2016 như sau.

Giả sử, người ta muốn đo lường sự tương đồng giữa 3 VDV còn lại với VDV Michael Phelps dựa trên các thuộc tính: *sex*, *date_of_birth* (chỉ quan tâm năm sinh), *height* và *weight*. Hãy xác định xem VDV nào có độ tương đồng gần với VDV Michael Phelps nhất? Với:

- a. Cách 1: sử dụng độ đo thuộc tính tổng hợp
- b. Cách 2: sử dụng độ đo Cosin
- c. So sánh kết quả của 2 cách trên

<i>name</i>	<i>sex</i>	<i>date_of_birth</i>	<i>height</i>	<i>weight</i>	<i>gold</i>	<i>silver</i>	<i>bronze</i>	<i>total</i>
Michael Phelps	male	6/30/1985	1.94	90	5	1	0	6
Katie Ledecky	female	3/17/1997	1.83	72	4	1	0	5
Simone Manuel	female	8/2/1996	1.78	72	2	2	0	4
Nathan Adrian	male	12/7/1988	1.99	102	2	0	2	4

5. BÀI TẬP

- iii. Cho dữ liệu thử trong bảng sau của 3 nước, mỗi nước gồm 7 VĐV Yêu cầu:
- a. Tính các giá trị Mean, Median, Mode, Midgange, five-number summary của các thuộc tính age, height và weight?
 - b. Vẽ đồ thị scatter dựa trên 2 thuộc tính height và age của cả 5 nước trên cùng 1 đồ thị.
 - c. Vẽ đồ thị boxplot dựa trên thuộc tính weight của cả 3 nước trên cùng 1 đồ thị.
 - d. Giả sử, người ta muốn đo lường sự tương đồng giữa 3 VĐV tiêu biểu của mỗi nước, cụ thể là các VĐV có id lần lượt là 3, 8, 15. Hãy xác định độ tương đồng của 3 VĐV này?

id	name	nationality	sex	height	weight	medal	age
1	Shelly Francis	USA	female	1.58	65	Silver	58
2	Phillip Dutton	USA	male	1.68	68	Bronze	53
3	Lauren Hernandez	USA	female	1.53	48	Gold	16
4	Kanak Jha	USA	male	1.66	51	No	16
5	Laura Zeng	USA	female	1.61	43	Silver	17
6	Sydney McLaughlin	USA	female	1.76	59	No	17
7	Aria Fischer	USA	female	1.83	78	Gold	17
8	Angelina Melnikova	RUS	female	1.51	44	Gold	16
9	Iaroslav Potapov	RUS	male	1.88	72	No	17
10	Seda Tutkhalian	RUS	female	1.42	35	Gold	17
11	Daria Chikunova	RUS	female	1.77	59	Bronze	17
12	Arina Openysheva	RUS	female	1.68	59	Silver	17
13	Inessa Merkulova	RUS	female	1.7	65	No	52
14	Andrey Mitin	RUS	male	1.74	80	Bronze	46
15	Philippe Rozier	FRA	male	1.73	63	Gold	53
16	Karen Tebar	FRA	female	1.59	57	No	52
17	Marine Boyer	FRA	female	1.6	50	Bronze	16
18	Oreane Lechenault	FRA	female	1.34	37	Silver	16
19	Loan His	FRA	female	1.6	51	No	17
20	Louise Vanhille	FRA	female	1.67	55	No	18
21	Marie Wattel	FRA	female	1.81	71	Gold	19

