

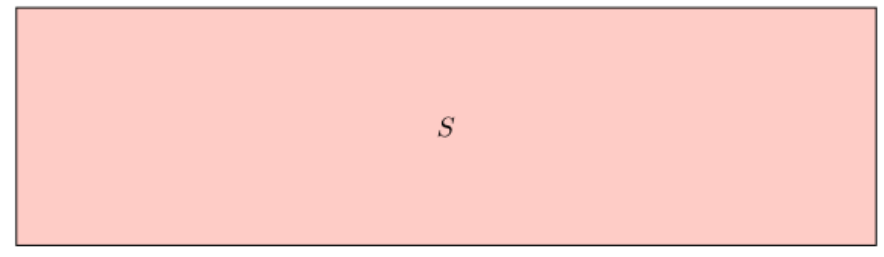
# Thống kê và tiền xử lý dữ liệu

- ▶ 2.0 Một số vấn đề về thống kê
- ▶ 2. Tiền xử lý dữ liệu
  - ▶ Tổng quan về giai đoạn tiền xử lý dữ liệu
  - ▶ Tóm tắt mô tả về dữ liệu
  - ▶ Làm sạch dữ liệu
  - ▶ Tích hợp dữ liệu
  - ▶ Biến đổi dữ liệu
  - ▶ Thu giảm dữ liệu
  - ▶ Rời rạc hóa dữ liệu

## 2.1 Nhắc lại kiến thức cơ bản về thống kê và xác suất



## 2.1 Nhắc lại kiến thức về thống kê và xác suất cơ bản



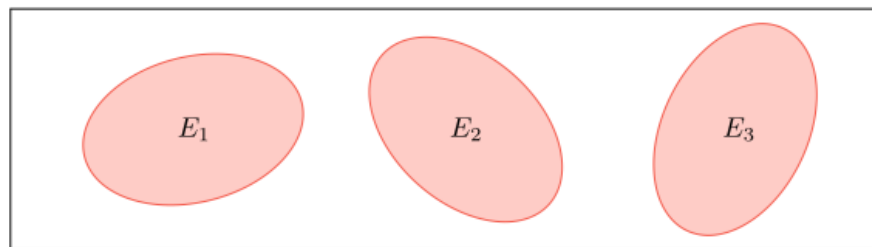
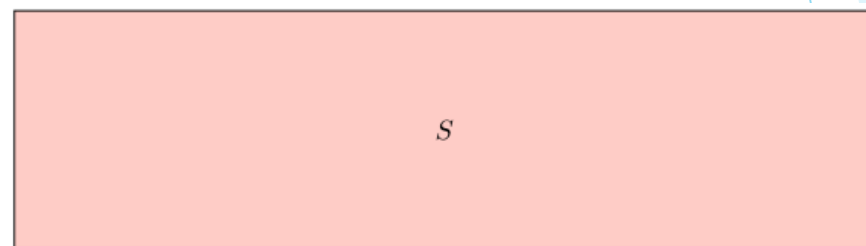
- ▶ Giới thiệu về Xác suất và Tổ hợp
- ▶ Không gian mẫu – Một tập hợp các kết cục có thể xảy ra của một phép thử được gọi là không gian mẫu của phép thử và được kí hiệu là S.
- ▶ Sự kiện (hay còn gọi là biến cố) – Bất kỳ một tập hợp con E nào của không gian mẫu đều được gọi là một sự kiện. Một sự kiện là một tập các kết cục có thể xảy ra của phép thử. Nếu kết quả của phép thử chứa trong E, chúng ta nói sự kiện E đã xảy ra.
- ▶ Tiên đề của xác suất – Với mỗi sự kiện E, chúng ta kí hiệu  $P(E)$  là xác suất sự kiện E xảy ra.

(1)  $0 \leq P(E) \leq 1$

(2)  $P(S) = 1$

(3)

$$P\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n P(E_i)$$



(1)  $0 \leq P(E) \leq 1$

(2)  $P(S) = 1$

(3)  $P\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n P(E_i)$

## 2.1 Nhắc lại kiến thức về thống kê và xác suất cơ bản

- ▶ Hoán vị – Hoán vị là một cách sắp xếp  $r$  phần tử từ một nhóm  $n$  phần tử, theo một thứ tự nhất định. Số lượng cách sắp xếp như vậy là  $P(n, r)$ , được định nghĩa như sau:

$$P(n, r) = \frac{n!}{(n - r)!}$$

- ▶ Tổ hợp – Một tổ hợp là một cách sắp xếp  $r$  phần tử từ  $n$  phần tử, không quan trọng thứ tự. Số lượng cách sắp xếp như vậy là  $C(n, r)$ , được định nghĩa như sau:

$$C(n, r) = \frac{P(n, r)}{r!} = \frac{n!}{r!(n - r)!}$$

- ▶ Ghi chú: Chúng ta lưu ý rằng với  $0 \leq r \leq n$ , ta có  $P(n, r) > C(n, r)$

## 2.1 Nhắc lại kiến thức về thống kê và xác suất cơ bản

- ▶ Xác suất có điều kiện
- ▶ □ Định lí Bayes - Với các sự kiện A và B sao cho  $P(B) > 0$ , ta có:

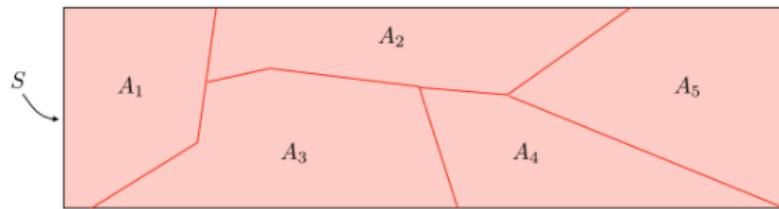
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- ▶ Ghi chú: ta có  $P(A \cap B) = P(A)P(B|A) = P(A|B)P(B)$

## 2.1 Nhắc lại kiến thức về thống kê và xác suất cơ bản

- ▶ Phân vùng - Cho  $\{A_i, i \in [1, n]\}$  sao cho với mỗi  $i$ ,  $A_i \neq \emptyset$ . Chúng ta nói rằng  $\{A_i\}$  là một phân vùng nếu có:

$$\forall i \neq j, A_i \cap A_j = \emptyset \quad \text{và} \quad \bigcup_{i=1}^n A_i = S$$



- ▶ Ghi chú: với bất  $P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$  nào, ta có :



## 2.1 Nhắc lại kiến thức về thống kê và xác suất cơ bản

- ▶ Định lý Bayes mở rộng - Cho  $\{A_i, i \in [[1, n]]\}$  là một phân vùng của không gian mẫu. Ta có:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^n P(B|A_i)P(A_i)}$$

- ▶ Sự kiện độc lập – Hai sự kiện A và B được coi là độc lập khi và chỉ khi ta có:
- ▶  $P(A \cap B) = P(A)P(B)$

## 2.1 Tổng quan về tiền xử lý dữ liệu.

- ▶ Tổng quan về giai đoạn tiền xử lý dữ liệu
- ▶ Giai đoạn tiền xử lý dữ liệu:
  - ▶ Các kỹ thuật datamining đều thực hiện trên các cơ sở dữ liệu, nguồn dữ liệu lớn. Đó là kết quả của quá trình ghi chép liên tục thông tin phản ánh hoạt động của con người, các quá trình tự nhiên...
  - ▶ Các dữ liệu lưu trữ hoàn toàn là dưới dạng thô, chưa sẵn sàng cho việc phát hiện, khám phá thông tin ẩn chứa trong đó. Do vậy chúng cần phải qua giai đoạn tiền xử lý dữ liệu trước khi tiến hành bất kỳ một phân tích nào.

## 2.1 Tổng quan về tiền xử lý dữ liệu.

### 2.1.1. Tại sao phải tiền xử lý dữ liệu?

Dữ liệu trong thế giới thực (mà chúng ta muốn phân tích bằng cách áp dụng các kỹ thuật khai phá dữ liệu) thường:

- **Không hoàn chỉnh** (incomplete): thiếu vắng các giá trị hoặc các thuộc tính đáng quan tâm, hoặc chỉ chứa các dữ liệu gộp nhóm.
- **Chứa đựng các giá trị nhiễu** (noisy): bao gồm các lỗi hoặc các giá trị lệch quá xa ra ngoài phạm vi mong đợi.
- **Không nhất quán** (inconsistent).

#### Lý do:

- ❑ Kích thước dữ liệu quá lớn.
- ❑ Được thu thập từ nhiều nguồn khác nhau.

⇒ Chất lượng dữ liệu thấp sẽ dẫn tới những kết quả khai phá tồi.

***Tiền xử lý dữ liệu là quá trình áp dụng các kỹ thuật nhằm nâng cao chất lượng dữ liệu và từ đó giúp nâng cao chất lượng kết quả khai phá.***

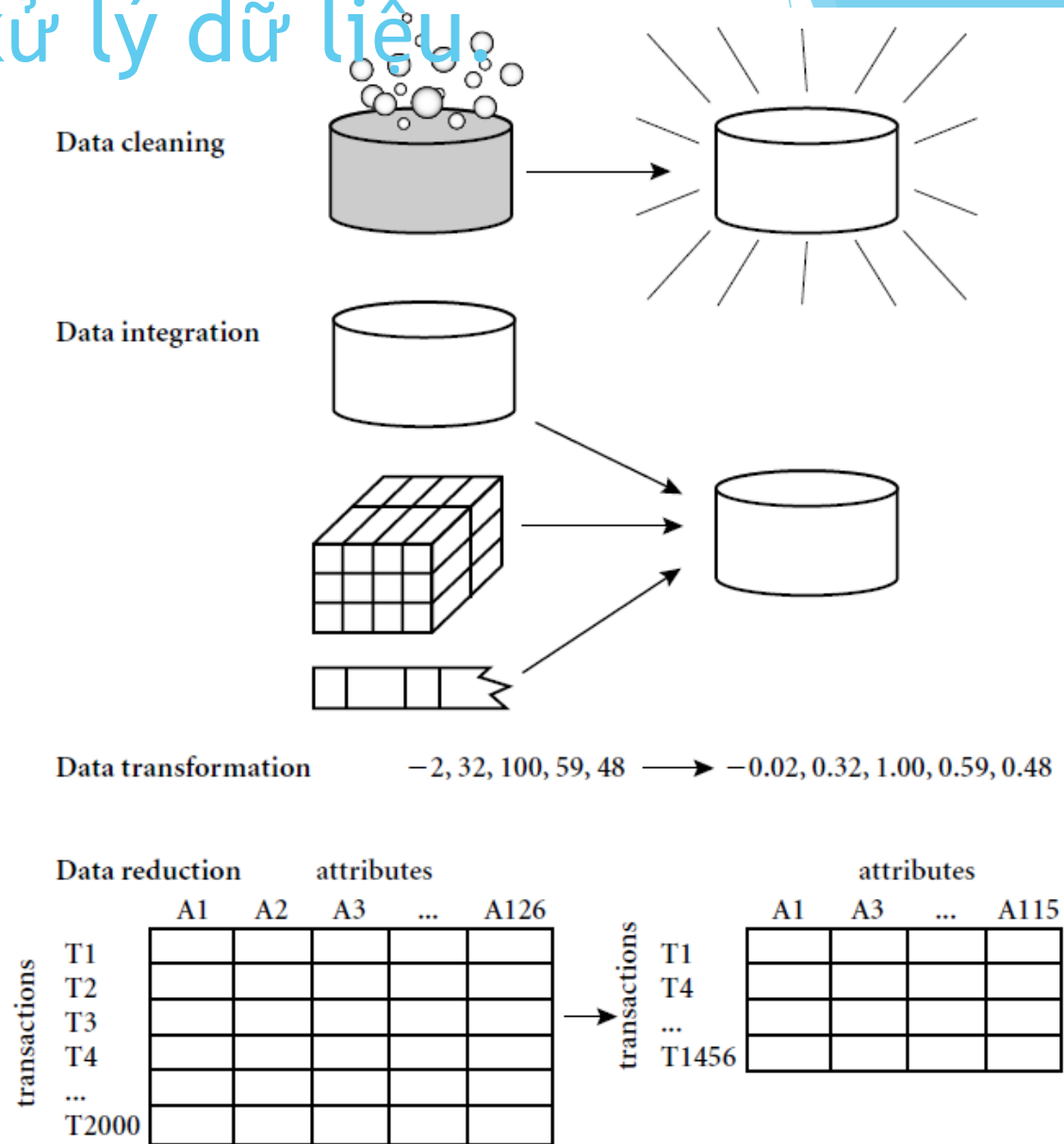
## 2.1 Tổng quan về tiền xử lý dữ liệu.

- ▶ Chất lượng dữ liệu (data quality):
  - ▶ Tính chính xác (accuracy): giá trị được ghi nhận đúng với giá trị thực.
  - ▶ Tính hiện hành (currency/timeliness): giá trị được ghi nhận không bị lỗi thời.
  - ▶ Tính toàn vẹn (completeness): tất cả các giá trị dành cho một biến/thuộc tính đều được ghi nhận.
  - ▶ Tính nhất quán (consistency): tất cả giá trị dữ liệu đều được biểu diễn như nhau trong tất cả các trường hợp.

## 2.1 Tổng quan về tiền xử lý dữ liệu

### ► Bốn bước tiền xử lý dữ liệu

- Data cleaning
- Data integration
- Data transformation
- Data reduction



## 2.1 Tổng quan về tiền xử lý dữ liệu.

- ▶ Các kỹ thuật tiền xử lý dữ liệu:
  - ▶ Làm sạch dữ liệu (data cleaning/cleansing):
    - ▶ Loại bỏ nhiễu (remove noise)
    - ▶ Tóm tắt hoá dữ liệu
    - ▶ Xử lý dữ liệu bị thiếu
    - ▶ Hiệu chỉnh những phần dữ liệu không nhất quán (correct data inconsistencies)
  - ▶ Tích hợp dữ liệu (data integration):
    - ▶ Trộn dữ liệu (merge data) từ nhiều nguồn khác nhau vào một kho dữ liệu
    - ▶ Vấn đề dư thừa
    - ▶ Phát hiện và xử lý xung đột về dữ liệu.

- ▶ Biến đổi dữ liệu (data transformation):
  - ▶ Chuẩn hoá dữ liệu (data normalization)
  - ▶ Làm trơn dữ liệu
  - ▶ Tổng quát hoá dữ liệu
  - ▶ Xây dựng thuộc tính (feature generation)
- ▶ Thu giảm dữ liệu (data reduction):
  - ▶ Thu giảm kích thước dữ liệu (nghĩa là giảm số phần tử) bằng kết hợp dữ liệu (data aggregation)
  - ▶ Loại bỏ các đặc điểm dư thừa (redundant features)
    - ▶ Giảm số chiều/thuộc tính dữ liệu
    - ▶ Gom cụm dữ liệu

## 2.2. Tóm tắt mô tả về dữ liệu

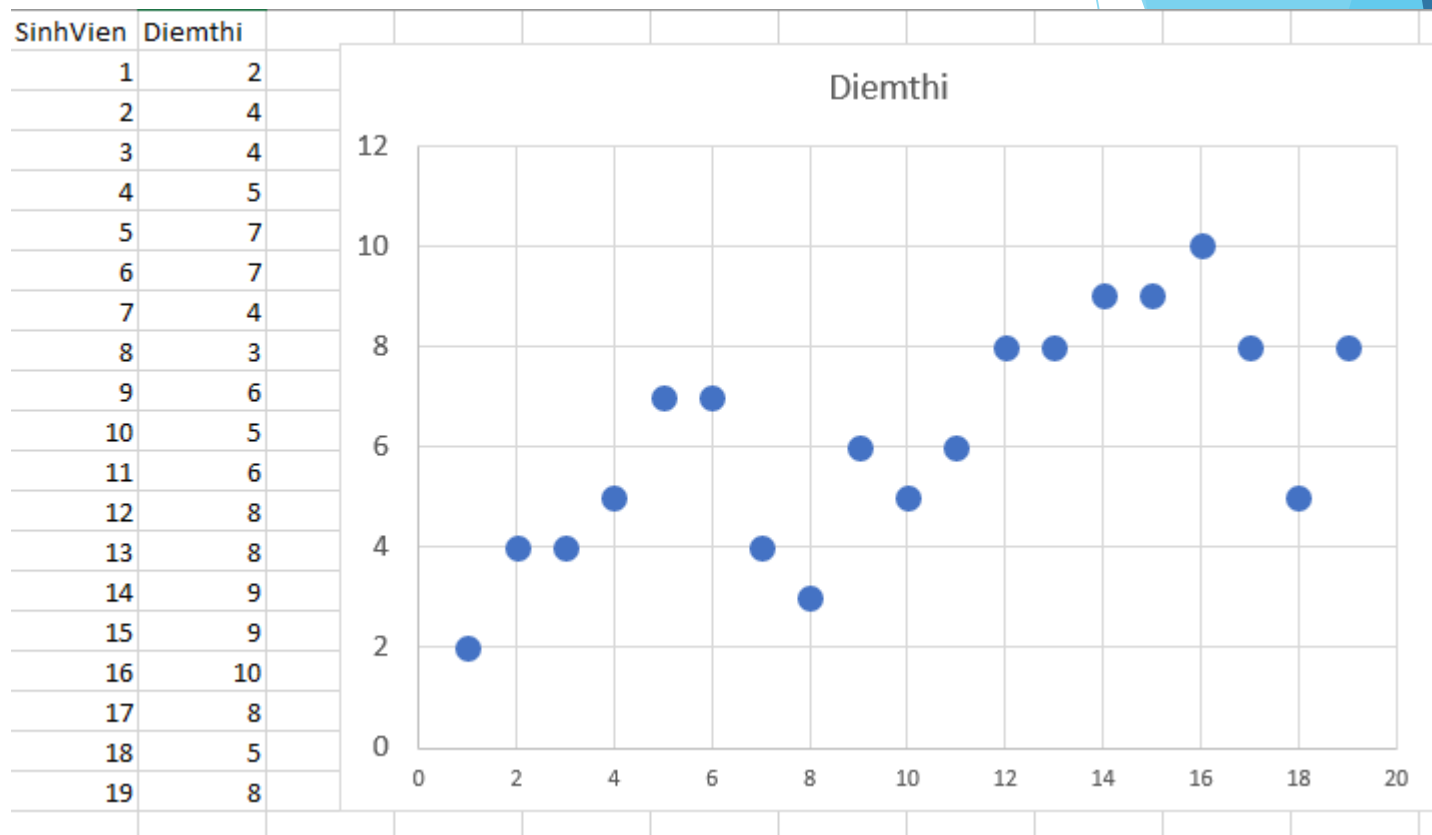
### 2.2.1. Khái niệm về tóm tắt mô tả dữ liệu

**Tóm tắt mô tả dữ liệu** (descriptive data summarization) là kỹ thuật được sử dụng nhằm xác định những đặc trưng điển hình và những đặc điểm nổi bật (highlight) của dữ liệu (những giá trị được xem là nhiễu (noise) hoặc vượt ngoài phạm vi mong đợi (outliers)).



## 2.2. Tóm tắt mô tả về dữ liệu

- Thấy được gì từ dữ liệu này



- ▶ Nghiên cứu tính chất hay đặc trưng của dữ liệu, cần quan tâm tới: những tính chất tiêu biểu của dữ liệu: Xu hướng tập trung (central tendency) và sự phân tán (dispersion) của dữ liệu.
- ▶ Xu hướng tập trung của dữ liệu được thể hiện qua:
  - ▶ Mean, median, mode, midrange
- ▶ Xu hướng phân tán:
  - ▶ Tứ phân vị (quartile), khoảng tứ phân vị (interquartile range – IRQ), phương sai (variance).
  - ▶ Từ đó, có thể xác định những giá trị biên (outliers) và các giá trị nhiễu (noise).

### ► Giá trị trung bình (Mean)

Xét dãy gồm N phần tử  $\{x_1, x_2, \dots, x_N\}$ . Giá trị trung bình (mean) được xác định bởi công thức:

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

Nếu mỗi phần tử  $x_i$  có một trọng số  $w_i$  đi kèm thì giá trị trung bình gọi là *trung bình dựa trên trọng số* (weighted average) và được xác định bởi:

$$\bar{x} = \frac{\sum_{i=1}^N x_i w_i}{\sum_{i=1}^N w_i} = \frac{x_1 w_1 + x_2 w_2 + \dots + x_N w_N}{w_1 + w_2 + \dots + w_N}$$

### ► Trung vị (Median)

Xét dãy gồm N phần tử được **sắp có thứ tự**  $\{x_1, x_2, \dots, x_N\}$ .

Nếu N là số nguyên lẻ ( $N=2K+1$ ) thì trung vị  $Med = x_{[N/2]+1}$  (phần tử chính giữa dãy).

Nếu N là số nguyên chẵn ( $N=2K$ ) thì trung vị  $Med = (x_{N/2} + x_{N/2+1})/2$  (trung bình cộng của hai phần tử chính giữa dãy).

## ► Giá trị mode

Mode là giá trị có tần suất xuất hiện lớn nhất trong tập dữ liệu đang xét. Giả sử tập dữ liệu đang xét chứa  $N$  giá trị khác nhau  $v_1, v_2, \dots, v_N$ . Gọi tần suất xuất hiện của giá trị  $v_i$  là  $f(v_i)$ . Khi đó:

$$f(\text{mode}) = \max_{1 \leq i \leq N} \{f(v_i)\}$$

Một tập dữ liệu có thể có nhiều giá trị mode.

## ► Khoảng trung bình (midrange)

Khoảng trung bình cũng có thể được sử dụng để xác định độ tập trung của dữ liệu. Khoảng trung bình được xác định là trung bình cộng của các giá trị lớn nhất và nhỏ nhất trong tập dữ liệu.

$$\text{midrange} = \frac{\text{max} + \text{min}}{2}$$

## Đánh giá sự phân tán của dữ liệu

- ▶ Tứ phân vị
  - ▶ Nhất-tứ phân vị (first quartile) là 25-thập phân vị ( $Q_1$ )
  - ▶ Nhị- tứ phân vị (second quartile) là 50-thập phân vị ( $Q_2$ )
  - ▶ Tam-tứ phân vị (third quartile) là 75-thập phân vị ( $Q_3$ )
  - ▶ Khoảng liên tứ phân vị (interquartile range - IQR):
$$IQR = Q_3 - Q_1$$
    - ▶ Outliers: thường được xác định là nằm trên  $Q_3$  hay dưới  $Q_1$  một khoảng  $= 1.5 * IQR$
- ▶ Cách tính tứ phân: ví dụ  $Q_1$  ( $25^{th}$ )  $\rightarrow \text{round-down}(0.25 * (n+1)) \rightarrow$  số thứ tự trong dãy số tăng dần.
- ▶ Giá trị của  $Q_1 = (\text{Số liền trước} + \text{số liền sau}) / 2$
- ▶

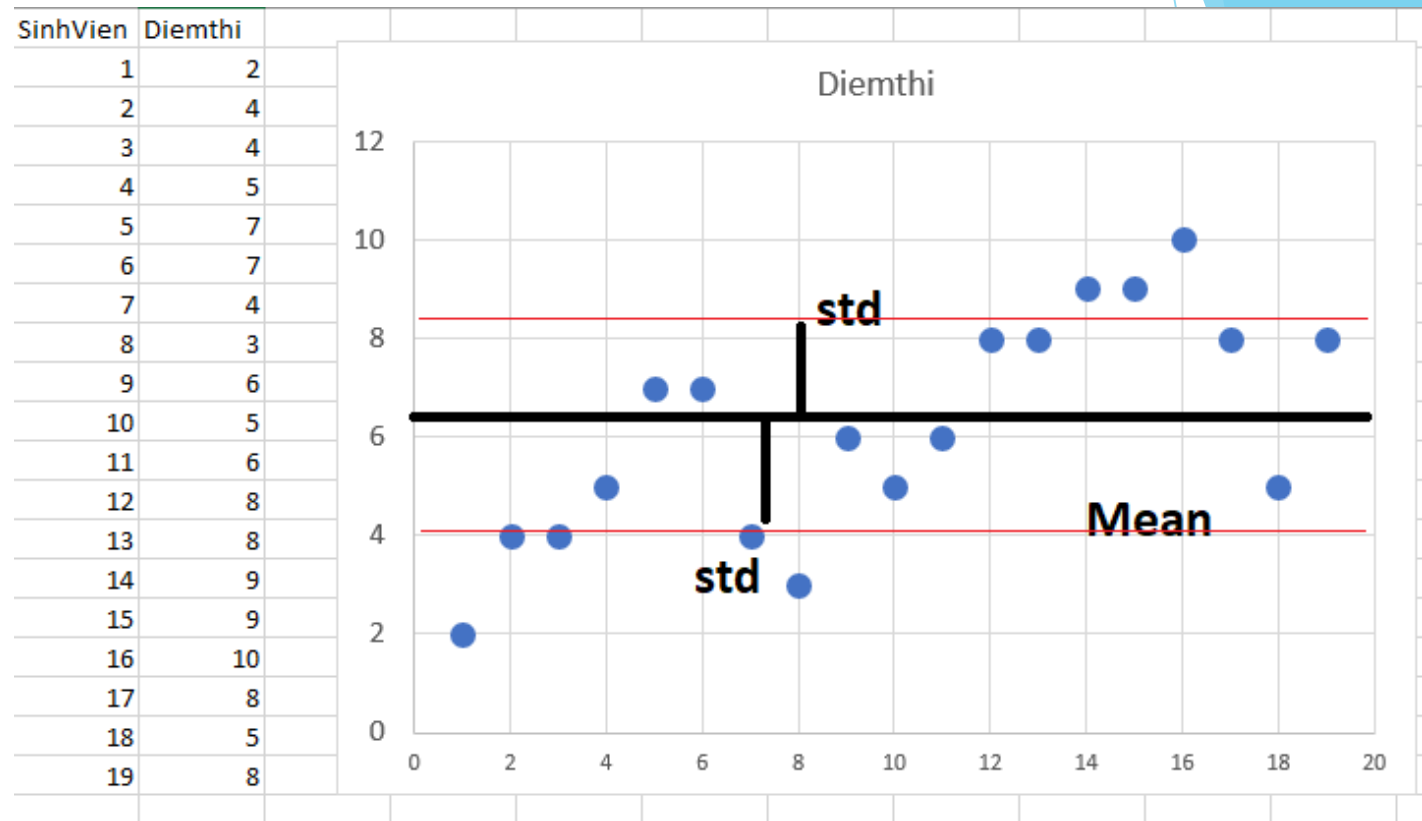
► Phương sai (variance) = 
$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

►  $\bar{x}$  : là giá trị trung bình của  $X_1, \dots, X_n$

**Độ lệch chuẩn** (standard deviation)  $\sigma$  được xác định bằng căn bậc 2 của phương sai.

► Lưu ý:

- *Độ lệch chuẩn phân bố xung quanh giá trị trung bình và chỉ được sử dụng khi giá trị trung bình được chọn làm giá trị đặc trưng cho trung tâm của dãy.*
- *$\sigma = 0$  có nghĩa là không có sự phân bố phương sai, tất cả các giá trị đều bằng nhau.*

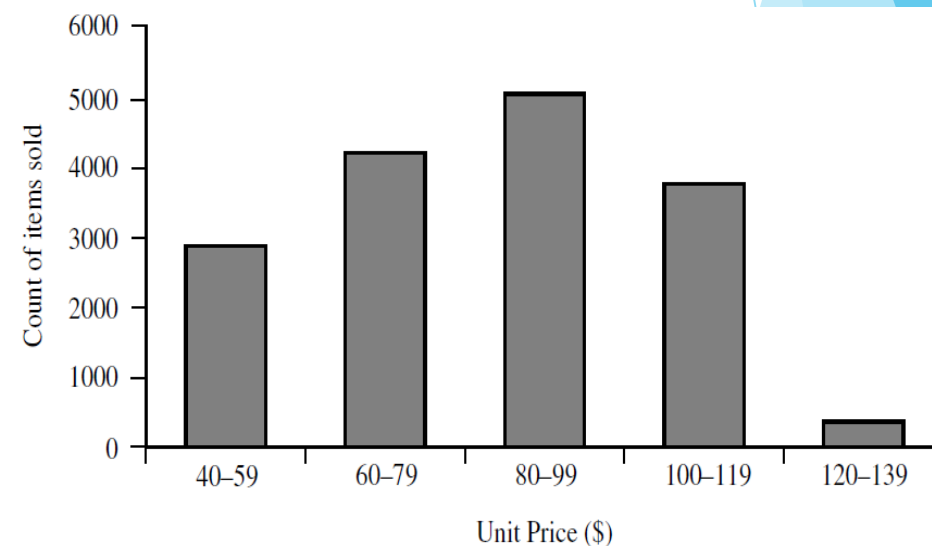


## 2.2.4. Biểu diễn tóm tắt mô tả dữ liệu dưới dạng đồ thị

### 2.2.4.1. Biểu đồ tần suất (frequency histograms)

- ▶ Là phương pháp biểu diễn tóm tắt sự phân bố của một thuộc tính cho trước nào đó dưới dạng trực quan.
- ▶ Biểu đồ tần suất ứng với một thuộc tính A nào đó sẽ chia sự phân bố dữ liệu của A thành các tập không giao nhau gọi là bucket (thường thì độ rộng của các bucket là bằng nhau).
- ▶ Mỗi bucket được biểu diễn bằng một hình chữ nhật có chiều cao tương ứng là số lượng hay tần suất của các giá trị có trong bucket.

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

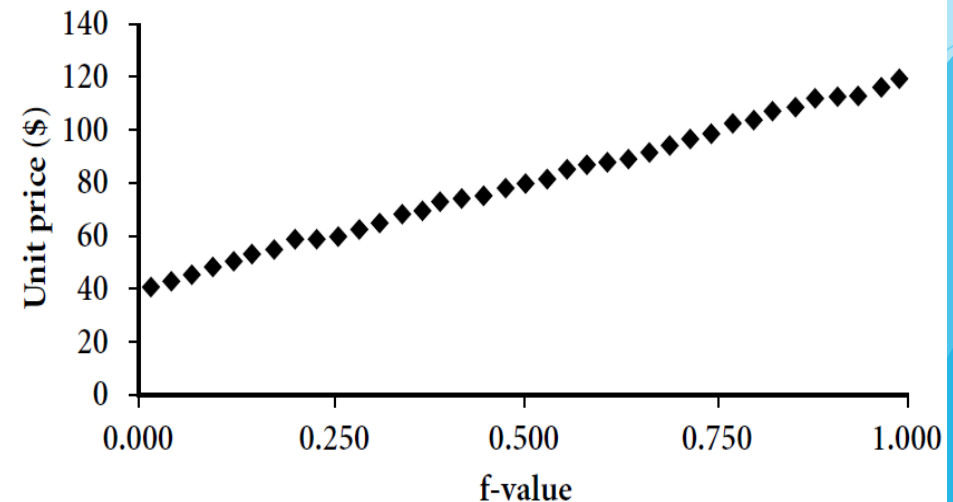




### 2.2.4.2. Đồ thị phân vị (quantile plot):

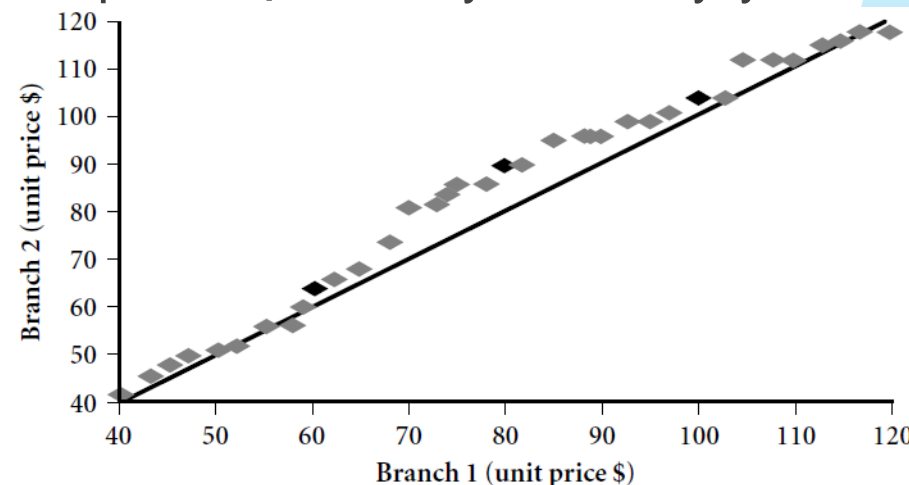
- ▶ Là cách thức đơn giản và hiệu quả để cho ta một cái nhìn về sự phân bố của dữ liệu đơn biến.
- ▶ Cho phép biểu diễn toàn bộ dữ liệu ứng với thuộc tính cho trước.
- ▶ Biểu diễn đồ thị thông tin phân vị (quantile information).
- ▶ Kỹ thuật biểu diễn:
  - ❖ Dãy giá trị  $x_i$  sẽ được sắp tăng dần từ  $x_1$  tới  $x_N$ . Mỗi giá trị  $x_i$  sẽ được đi kèm với một giá trị  $f_i$  là tỷ lệ phần trăm các giá trị dữ liệu trong dãy nhỏ hơn hoặc bằng  $x_i$ .
  - ❖ Giá trị  $f_i$  có thể tính bởi công thức:  $f_i = \frac{i - 0.5}{N}$
  - ❖ Trên đồ thị,  $x_i$  được biểu diễn theo  $f_i$ .

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



### 2.2.4.3. Đồ thị song phân vị (quantile-quantile plot):

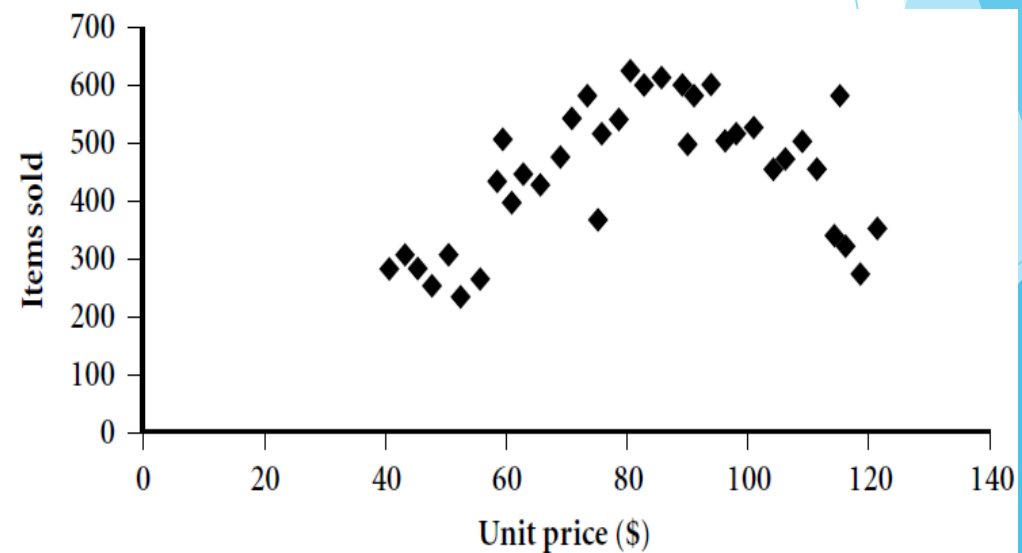
- ▶ Biểu diễn mối liên hệ giữa phân vị của một phân bố đơn biến này với phân vị của một phân bố đơn biến khác.
- ▶ Đây là công cụ trực quan mạnh mẽ cho phép quan sát sự thay đổi khi chuyển từ phân bố này sang một phân bố khác.
- ▶ Kỹ thuật biểu diễn:
  - ❖ Giả sử chúng ta có hai dãy giá trị của cùng một biến ngẫu nhiên được thu thập độc lập nhau: dãy  $x = \{x_1, x_2, \dots, x_N\}$  và dãy  $y = \{y_1, y_2, \dots, y_M\}$
  - ❖ Nếu  $N = M$ : biểu diễn  $Y_i$  theo  $X_i$  trong đó  $X_i, Y_i$  tương ứng là các phân vị của dãy  $x$  và dãy  $y$  xác định theo công thức  $(i - 0.5)/N$ .
  - ❖ Nếu  $M < N$ : biểu diễn  $Y_i$  theo  $X_i$  và chỉ có  $M$  điểm biểu diễn trên đồ thị. Trong đó  $X_i, Y_i$  tương ứng là các phân vị của dãy  $x$  và dãy  $y$  xác định theo công thức  $(i - 0.5)/M$ .



#### 2.2.4.4. Đồ thị phân tán (scatter plot):

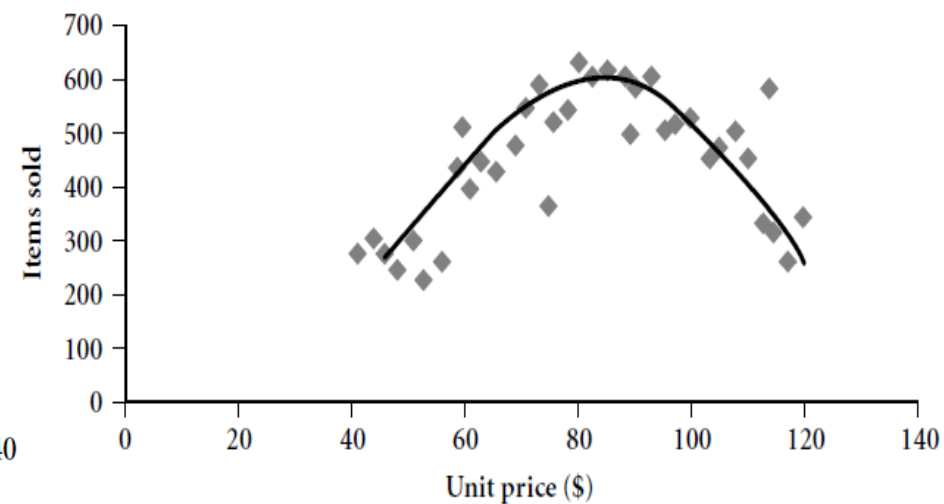
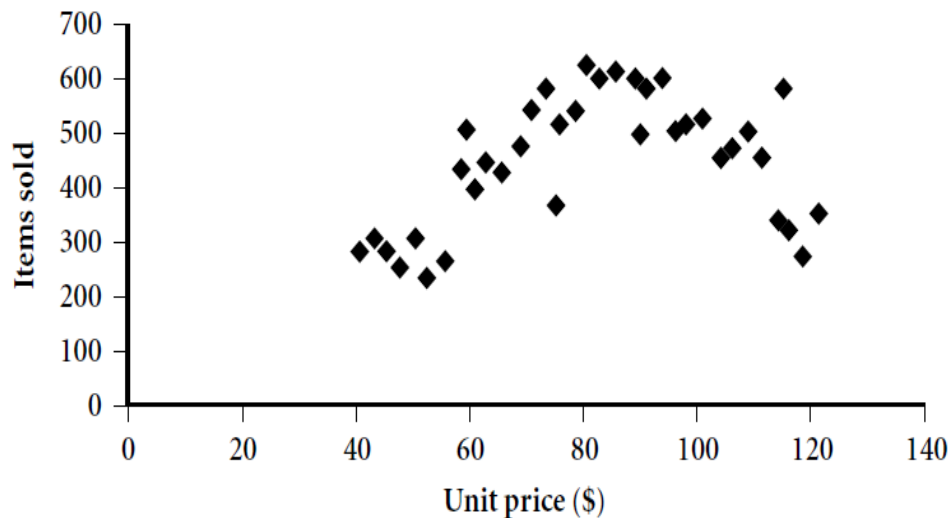
- ▶ Là phương pháp hiệu quả để xác định xem liệu có xuất hiện mối quan hệ, các mẫu hay xu hướng giữa 02 thuộc tính mang giá trị số hay không.
- ▶ Mỗi cặp giá trị được biểu diễn bằng một cặp tọa độ (tương ứng với một điểm trên mặt phẳng tọa độ).
- ▶ Cung cấp một cái nhìn sơ bộ về dữ liệu để thấy được các cụm điểm và các giá trị kỳ dị (outliers) cũng như phát hiện khả năng tồn tại của các mối liên hệ phụ thuộc.

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



### 2.2.4.5. Đường loess

- ▶ Là công cụ biểu diễn đồ thị quan trọng cho phép bổ sung một đường cong “trơn” vào đồ thị phân tán nhằm cung cấp một sự hình dung tốt hơn về mẫu độc lập (loess = local regression: hồi quy cục bộ).
- ▶ Để khớp với đường cong hồi quy, các giá trị cần được thiết lập với 02 tham số là  $\alpha$ -tham số độ trơn và  $\lambda$ -bậc của đa thức hồi quy.
- ▶ Cần chọn  $\alpha$  để tạo ra một đường cong “trơn” nhất có thể nhưng không làm biến dạng mẫu dữ liệu được phản ánh.



## 2.3. LÀM SẠCH DỮ LIỆU

- ▶ **Làm sạch dữ liệu (data cleaning)** là kỹ thuật giúp xử lý sự thiếu vắng giá trị, loại bỏ nhiễu và các giá trị không mong muốn cũng như giải quyết vấn đề không nhất quán dữ liệu.
  - ▶ **Xử lý sự thiếu vắng giá trị (missing values)**
  - ▶ **Xử lý dữ liệu nhiễu (noisy data)**
  - ▶ **Xử lý dữ liệu không nhất quán (inconsistent data)**

## 2.3. LÀM SẠCH DỮ LIỆU

### ▶ **Xử lý sự thiếu vắng giá trị (missing values)**

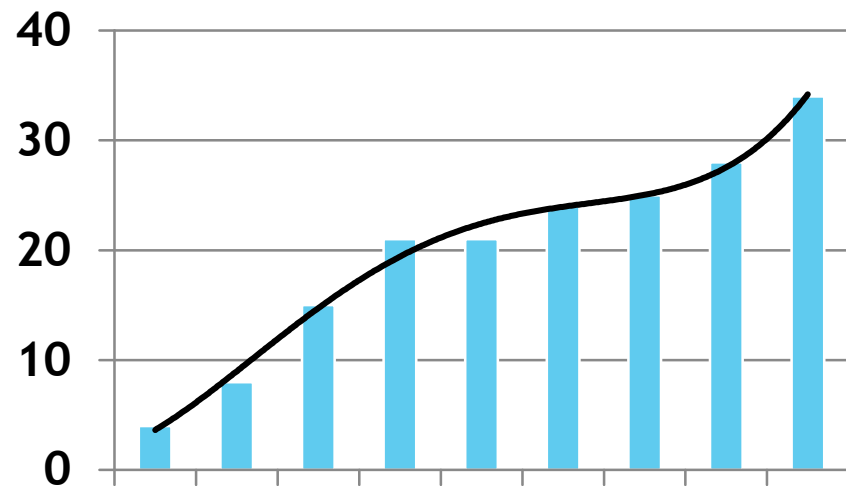
- ▶ Dữ liệu không có sẵn khi cần sử dụng, thường là bị thiếu các đặc trưng trong một số bản ghi.
- ▶ Do nguyên nhân, nhập dữ liệu, hoặc do cố ý con người muốn test mô hình.

### ▶ **Xử lý:**

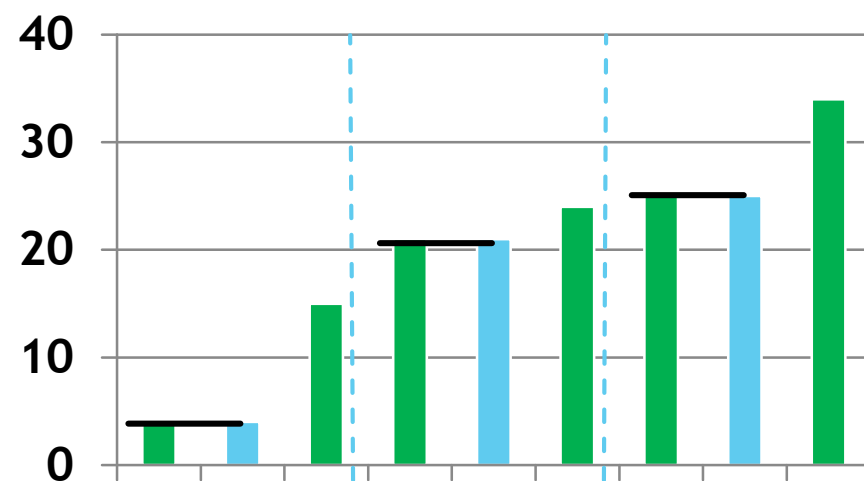
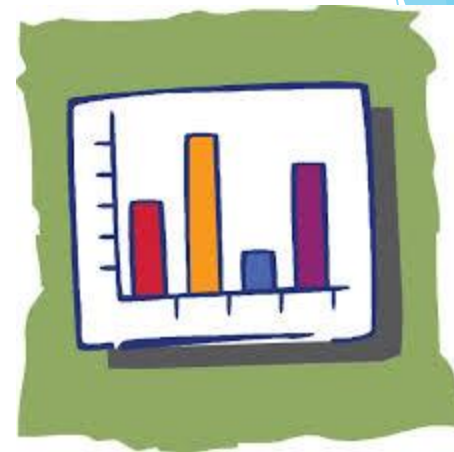
- ▶ Bỏ qua
- ▶ Sử dụng các giá trị (hằng) quy ước để thay cho các giá trị thiếu: Thay thế các giá trị thiếu bằng các giá trị (hằng) quy ước giống nhau (vd: “unknown”). Cách này có thể gây hiểu lầm cho hệ thống KPDL khi nghĩ rằng “unknown” là một giá trị đáng quan tâm.
- ▶ Sử dụng giá trị trung bình để thay cho các giá trị thiếu: Sử dụng giá trị trung bình của một thuộc tính để thay thế cho các giá trị thiếu trên thuộc tính đó.
- ▶ ....

## 2.3. LÀM SẠCH DỮ LIỆU

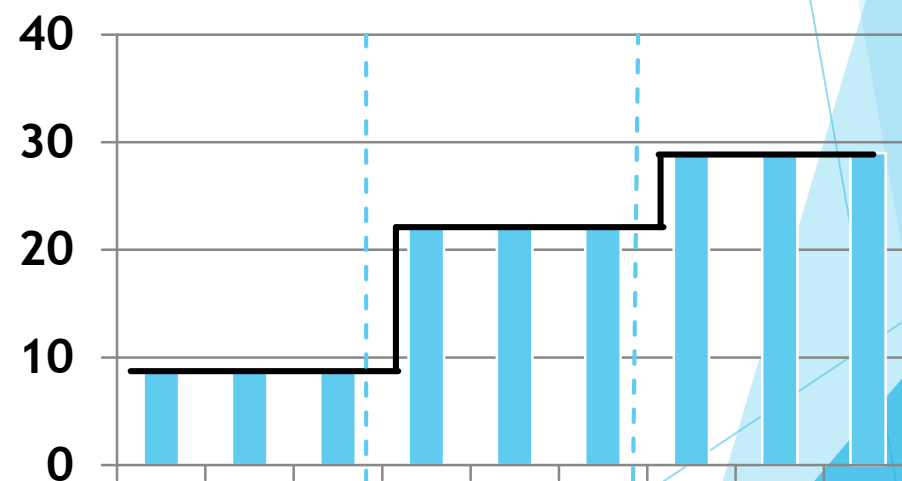
- ▶ Xử lý dữ liệu nhiễu (noisy data)
- ▶ **Nhiều (noise)** là những lỗi ngẫu nhiên hoặc những giá trị “lệch chuẩn”.
- ▶ **⇒ Làm thế nào để làm “mượt” (smooth) dữ liệu và loại bỏ nhiễu?**
- ▶ **“Đóng thùng” (binning):**
  - Là phương pháp làm “trơn” một giá trị dữ liệu đã được sắp xếp dựa trên các giá trị xung quanh (làm “trơn” cục bộ).
  - Các giá trị dữ liệu đã được sắp xếp sẽ được phân chia vào các “thùng chứa” (gọi là bin/bucket) có kích thước bằng nhau. Có 2 kiểu phân chia:
    - ❖ **Equal-frequency:** Các “thùng chứa” chứa số giá trị như nhau.
    - ❖ **Equal-width:** Các “thùng chứa” có khoảng giá trị biến động (từ giá trị min đến giá trị max của thùng) là như nhau.
  - Có 2 kỹ thuật phổ biến:
    - ❖ **Làm trơn trung bình/trung vị (smoothing by bin means/median):** mỗi giá trị trong “thùng chứa” sẽ được thay thế bằng trung bình cộng (hoặc trung vị) của toàn bộ các giá trị ban đầu có trong “thùng chứa” đó.
    - ❖ **Làm trơn dựa trên biên (smoothing by boundaries):** giá trị lớn nhất và giá trị nhỏ nhất trong “thùng chứa” sẽ được chọn làm biên. Mỗi giá trị trong thùng chứa sẽ được thay thế bằng giá trị biên gần nhất.



*Dữ liệu được sắp xếp*



*Làm trơn dựa trên biên*



*Làm trơn trung bình*



## 2.3. LÀM SẠCH DỮ LIỆU

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

**Partition into (equal-frequency) bins:**

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

**Smoothing by bin means:**

Bin 1: 9, 9, 9

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

**Smoothing by bin boundaries:**

Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

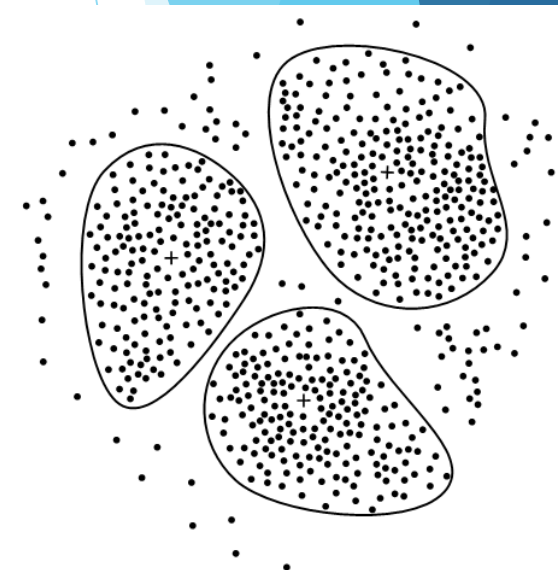
## 2.3. LÀM SẠCH DỮ LIỆU

### ► Hồi quy (regression):

- Dữ liệu có thể được làm trơn bằng cách khớp dữ liệu với một hàm hồi quy.
- Hồi quy tuyến tính đòi hỏi phải tìm ra đường thẳng tối ưu khớp với 2 biến (thuộc tính). Từ đó, một thuộc tính có thể được sử dụng để dự đoán thuộc tính còn lại.
- Hồi quy tuyến tính kép là sự mở rộng của hồi quy tuyến tính khi mà có nhiều hơn 02 biến (thuộc tính) và dữ liệu sẽ được khớp với đồ thị không gian là một mặt đa chiều.

### ► Phân cụm (clustering):

- Giá trị bất thường (outliers) có thể được phát hiện bằng kỹ thuật phân cụm khi mà các giá trị tương tự nhau được đưa vào cùng nhóm (cụm). Các giá trị không thuộc về một cụm nào cả có thể xem là bất thường.



## 2.3. LÀM SẠCH DỮ LIỆU

- ▶ Xử lý dữ liệu không nhất quán
  - ▶ Dữ liệu được ghi nhận khác nhau cho cùng một đối tượng/thực thể
    - ▶ Ví dụ: ngày/tháng/năm    8/23/2021; 23/8/2021
    - ▶ Dữ liệu được ghi nhận không phản ánh đúng ngữ nghĩa cho các đối tượng
- ▶ Nguyên nhân:
  - ▶ Sự không nhất quán trong các quy ước đặt tên hay mã dữ liệu
  - ▶ Định dạng không thống nhất tại các nơi nhập dữ liệu

## 2.3. LÀM SẠCH DỮ LIỆU

- ▶ Cách Xử lý:
  - ▶ Tận dụng sự ràng buộc của dữ liệu, để kiểm tra nhận dạng
  - ▶ Sử lý bằng tay, nếu không quá nhiều
  - ▶ Chuẩn hoá hay biến đổi tự động.

## 2.4. TÍCH HỢP

### 2.4.1. Tích hợp dữ liệu (Data Integration)

- Kết hợp dữ liệu từ nhiều nguồn khác nhau thành một kho dữ liệu thống nhất.
- Các nguồn dữ liệu khác nhau: cơ sở dữ liệu, data cube, tập tin phẳng,...
- Các vấn đề phải đối mặt:
  - ❖ **Tích hợp lược đồ (schema integration) và khớp các đối tượng (object matching):** cùng một thực thể trong thế giới thực có thể được phản ánh trong dữ liệu từ các nguồn khác nhau  $\Rightarrow$  cần phải khớp lại các đối tượng này. VD: Vấn đề về định danh thực thể
  - ❖ **Sự dư thừa (redundancy):**
    - ✓ Một thuộc tính có thể dư thừa nếu có thể được suy diễn từ một hay một tập các thuộc tính khác.
    - ✓ Sự không nhất quán trong thuộc tính hay do cách đặt tên có thể gây ra sự dư thừa trong tập dữ liệu kết quả.
    - ✓ Dư thừa dữ liệu có thể được phát hiện thông qua phân tích tương quan (correlation analysis).

## Phân tích dựa trên hệ số tương quan

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

$$r_{A,B} = \frac{\sum_{i=1}^N (a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^N (a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

### Trong đó:

- ✓ N: số bộ dữ liệu.
- ✓  $a_i, b_i$  là các giá trị tương ứng với 02 thuộc tính A và B trong bộ i.
- ✓  $\bar{A}, \bar{B}$  tương ứng là các giá trị trung bình trên A và B.
- ✓  $\sigma_A, \sigma_B$  tương ứng là độ lệch chuẩn của A và B.

Ta luôn có  $-1 \leq r_{A,B} \leq 1$  và:

- **Nếu  $r_{A,B} > 0$ :** A, B có mối tương quan dương (giá trị ứng với A tăng thì giá trị ứng với B cũng tăng). Giá trị  $r_{A,B}$  càng lớn thể hiện tính tương quan giữa 02 thuộc tính càng mạnh  $\Rightarrow$  Có thể loại bỏ một trong 02 thuộc tính (A hoặc B) vì nó là dư thừa.
- **Nếu  $r_{A,B} = 0$ :** Không tồn tại mối liên hệ tương quan. A và B là 02 thuộc tính hoàn toàn độc lập.
- **Nếu  $r_{A,B} < 0$ :** A, B có mối tương quan âm (giá trị ứng với A tăng thì giá trị ứng với B giảm và ngược lại)  $\Rightarrow$  A và B là 02 thuộc tính trái ngược nhau.



## Phân tích tương quan đối với dữ liệu rời rạc

Mỗi quan hệ tương quan giữa 02 thuộc tính A và B có thể được đặc trưng bởi phép đo Khi - Bình phương (Chi-square)  $\chi^2$

- ❖ Giả sử thuộc tính A có c giá trị khác nhau  $a_1, a_2, \dots, a_c$  và B có r giá trị khác nhau  $b_1, b_2, \dots, b_r$ .
- ❖ Các bộ dữ liệu đặc trưng bởi A, B được biểu diễn dưới dạng một bảng ngẫu nhiên (contingency table) với các cột là c giá trị khác nhau của A và các dòng là r giá trị khác nhau của B.
- ❖ Ký hiệu  $(A_i, B_j)$  là sự kiện thuộc tính A nhận giá trị  $a_i$  và thuộc tính B nhận giá trị  $b_j$ . Mỗi sự kiện  $(A_i, B_j)$  có thể có sẽ chiếm trọn một ô trong bảng.
- ❖ Giá trị Khi - Bình phương  $\chi^2$  có thể được xác định qua công thức:

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Trong đó:

- $o_{ij}$  là tần suất quan sát được hay tần suất biểu kiến (observed frequency) của sự kiện  $(A_i, B_j)$
- $e_{ij}$  là tần xuất kỳ vọng (expected frequency) của sự kiện  $(A_i, B_j)$ .



Tần xuất kỳ vọng (expected frequency) của sự kiện  $(A_i, B_j)$  có thể tính bởi công thức:

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{N}$$

Trong đó:

$N$ : số lượng các bộ dữ liệu.

$\text{count}(A=a_i)$ : số lượng các bộ có thuộc tính  $A$  nhận giá trị  $a_i$ .

$\text{count}(B=b_j)$ : số lượng các bộ có thuộc tính  $B$  nhận giá trị  $b_j$ .

**Chú ý:**

*Độ đo Khi - Bình phương dùng để kiểm tra giả thiết về tính độc lập của 02 thuộc tính  $A$  và  $B$ . Việc kiểm tra này dựa trên mức độ chú ý (significance level) với  $(r-1)(c-1)$  bậc tự do.*

	<i>male</i>	<i>female</i>	Total
<i>fiction</i>	250 (90)	200 (360)	450
<i>non_fiction</i>	50 (210)	1000 (840)	1050
Total	300	1200	1500

$$e_{11} = \frac{\text{count}(\text{male}) \times \text{count}(\text{fiction})}{N} = \frac{300 \times 450}{1500} = 90$$

$$\begin{aligned} \chi^2 &= \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} \\ &= 284.44 + 121.90 + 71.11 + 30.48 = 507.93. \end{aligned}$$

Với số bậc tự do là  $(2-1)(2-1) = 1$ , mức độ chú ý là 0.001 thì để đảm bảo 02 thuộc tính A, B là độc lập, giá trị  $\chi^2 = 10.828$  (**đề nghị SV tham khảo thêm các giáo trình về xác suất thống kê**)

⇒ Giá trị tính được là  $507.93 > 10.828$  nên A và B là 02 thuộc tính phụ thuộc chặt chẽ.

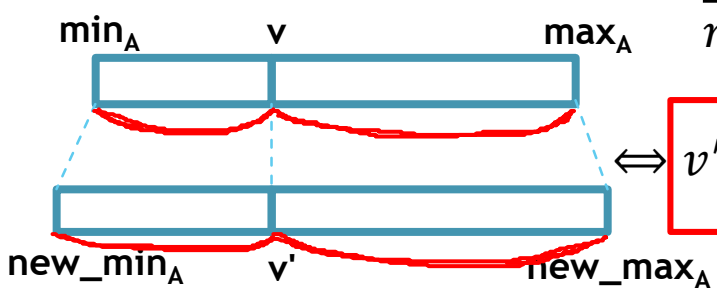
## 2.5 Biến đổi dữ liệu (Data Transformation)

Dữ liệu được chuyển đổi hoặc hợp nhất thành các dạng phù hợp cho việc khai phá. Chuyển dạng dữ liệu liên quan tới các vấn đề sau đây:

- **Làm trơn (Smoothing):** Loại bỏ các nhiễu (noisy) khỏi dữ liệu. Các kỹ thuật được sử dụng bao gồm: đóng thùng (binning), hồi quy (regression), phân cụm (clustering).
- **Gộp nhóm (Aggregation):** các thao tác tóm tắt hay gộp nhóm được áp dụng với dữ liệu. Bước này thường được sử dụng để xây dựng data cube cho phân tích dữ liệu từ nhiều nguồn.
- **Khởi tạo dữ liệu (Generalization of the data):** dữ liệu thô được thay thế bởi các khái niệm ở mức cao hơn thông qua việc sử dụng lược đồ khái niệm.
- **Xây dựng thuộc tính (Attribute construction):** các thuộc tính mới được xây dựng và thêm vào từ tập thuộc tính đã có để hỗ trợ quá trình khai phá (tăng độ chính xác và sự dễ hiểu của cấu trúc trong dữ liệu nhiều chiều (high-dimensional data)). Bằng cách kết hợp các thuộc tính  $\Rightarrow$  phát hiện ra các thông tin bị thiếu liên quan đến mối quan hệ giữa các thuộc tính (hữu ích cho quá trình khai phá).

- **Chuẩn hóa (Normalization):** Dữ liệu thuộc tính được chuyển đổi tương ứng với các phạm vi biểu diễn nhỏ hơn như  $[-1,1]$  hoặc  $[0,1]$ .

**Chuẩn hóa min-max:** thực hiện việc chuyển đổi tuyến tính dựa trên dữ liệu gốc. Gọi  $\min_A$ ,  $\max_A$  là giá trị lớn nhất và nhỏ nhất của thuộc tính  $A$ . Chuẩn hóa min-max sẽ ánh xạ một giá trị  $v$  của  $A$  tương ứng với một giá trị  $v'$  trong khoảng  $[\text{new\_min}_A, \text{new\_max}_A]$  thông qua công thức:

$$\frac{v - \min_A}{\max_A - \min_A} = \frac{v' - \text{new\_min}_A}{\text{new\_max}_A - \text{new\_min}_A}$$


$$\Leftrightarrow v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

**Ví dụ:** Giả sử giá trị lớn nhất và nhỏ nhất của thuộc tính income là \$12,000 và \$98,000. Người ta định ánh xạ miền giá trị của thuộc tính income tương ứng với khoảng  $[0.0, 1.0]$ . Hỏi giá trị  $v = \$73,600$  của income sẽ tương ứng với giá trị ánh xạ  $v'$  bằng bao nhiêu trong khoảng  $[0.0, 1.0]$ ?

$$\min_A = \$12,000$$

$$\max_A = \$98,000$$

$$\text{new\_min}_A = 0.0$$

$$\text{new\_max}_A = 0.1$$

$$v = \$73,600$$

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

$$= \frac{73,600 - 12,000}{98,000 - 12,000} (1 - 0) + 0 = 0.716$$

**Chuẩn hóa z-score:** các giá trị ứng với thuộc tính A được chuẩn hóa dựa trên giá trị trung bình và độ lệch chuẩn của A. Một giá trị  $v$  của A sẽ được chuẩn hóa tương ứng với một giá trị  $v'$  thông qua công thức:

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

Chuẩn hóa z-score rất hữu dụng khi:

- ❖ Không biết giá trị lớn nhất và nhỏ nhất thực tế của thuộc tính A.
- ❖ Các giá trị kỳ dị (outliers) chi phối chuẩn hóa min-max

**Ví dụ:** Giả sử rằng giá trị trung bình và độ lệch chuẩn của thuộc tính income tương ứng là \$54,000 và \$16,000. Một giá trị  $v = \$73,600$  của income sẽ được chuẩn hóa tương ứng với giá trị  $v'$  bằng bao nhiêu?

$$v' = \frac{v - \bar{A}}{\sigma_A} = \frac{73,600 - 54,000}{16,000} = 1.225$$

**Chuẩn hóa thập phân (decimal scaling):** dịch chuyển dấu phẩy thập phân của các giá trị ứng với thuộc tính A. Số vị trí di chuyển phụ thuộc vào giá trị tuyệt đối lớn nhất của A. Một giá trị  $v$  của A được chuẩn hóa thập phân tương ứng với một giá trị  $v'$  theo công thức:

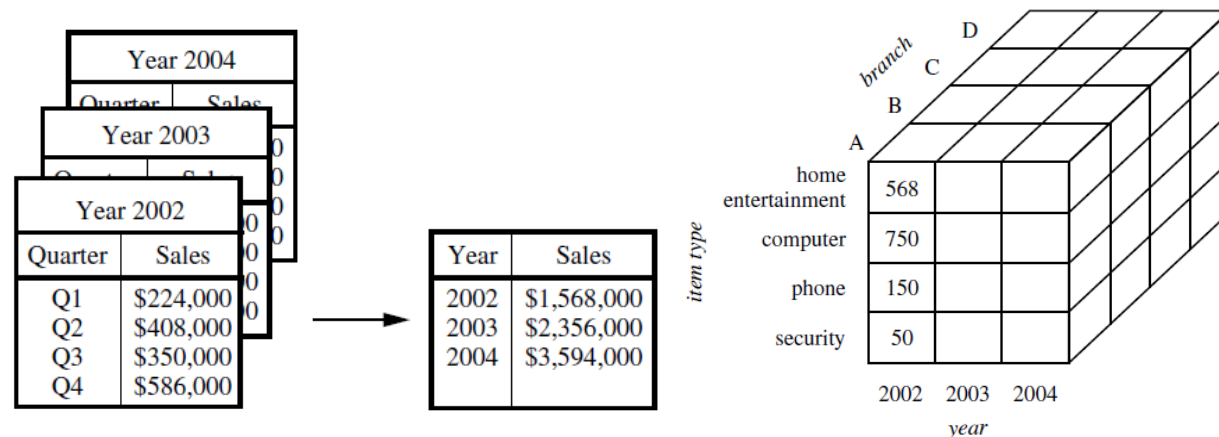
$$v' = \frac{v}{10^j}$$

( $j$  là số nguyên nhỏ nhất sao cho  $\text{Max}(|v'|) < 1$  )

**Ví dụ:** Giả sử thuộc tính A có miền giá trị là  $[-986, 917]$ . Giá trị tuyệt đối lớn nhất của A là 986. Như vậy, ta chọn  $j = 3$ . Khi đó thì một giá trị  $v = 817$  sẽ được chuẩn hóa thành  $v' = 0.817$

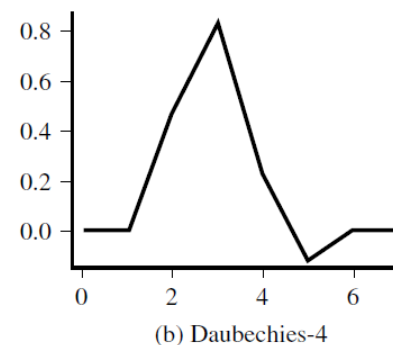
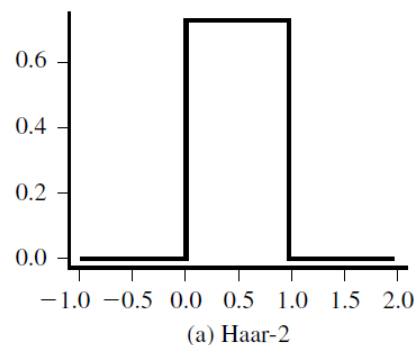
## 2.6. RÚT GỌN DỮ LIỆU

**2.6.1. Gộp nhóm dữ liệu dưới dạng data cube:** Các thao tác gộp nhóm sẽ được áp dụng trên dữ liệu để tạo ra một data cube



**2.6.2. Lựa chọn tập thuộc tính (Attribute subset selection):** Các thuộc tính thừa hoặc không thích hợp sẽ được phát hiện và loại bỏ.

**2.6.3. Giảm số chiều dữ liệu (Dimensionality reduction):** Các cơ chế mã hóa (encoding) sẽ được áp dụng để làm giảm kích thước dữ liệu.



## 2.7 Rời rạc hoá dữ liệu

- ▶ Giảm số lượng giá trị của một thuộc tính liên tục (continuous attribute) bằng cách chia khoảng bằng nhau (intervals)
- ▶ Các nhãn (labels) được gán cho các khoảng và được dùng thay cho giá trị thực của chúng
- ▶ Các thuộc tính có thể được phân hoạch theo một phân cấp (hierarchical) hay ở nhiều mức độ phân giải khác nhau (multiresolution)



- ▶ Các phương pháp rời rạc hoá dữ liệu cho các thuộc tính số
  - ▶ Binning
  - ▶ Histogram analysis
  - ▶ Interval merging bằng phân tích Chi ( $\chi^2$ ) bình phương
  - ▶ Phân tích nhóm/cụm
  - ▶ Rời rạc dựa vào entropy