

IS403

BDA

NGUYEN MINH NHUT


INFORMATION SYSTEM ENGINEERING 1

statistics

plural noun

US  /stə'tɪs-tɪks/

(infml stats, US /stæts/)

Add to word list 

a collection of numerical facts or measurements, as about people, business conditions, or weather:



Statistics in daily life



Utilizing Statistics to Measure Results



2

Trong chương này chúng ta sẽ học những nội dung:

- Tổng quan về thống kê mô tả và những công việc
- Các độ đo về vị trí (Measure of Location)
- Các độ đo về phân tán (Measure of Dispersion)
- Các độ đo về hình dạng (Shape)
- Trực quan Box-Plot, Histogram

LAB01

THỐNG KÊ MÔ TẢ



PHẦN 1

BỐI CẢNH CỦA THỐNG KÊ MÔ TẢ

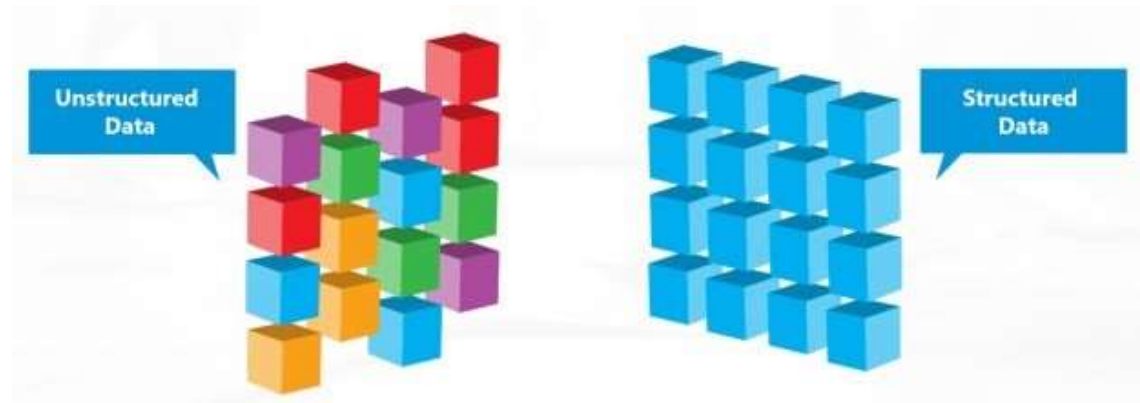
Thống kê mô tả là phương pháp tóm tắt dữ liệu thông qua các đặc điểm trung bình, phương sai và phân phối. Giúp hiểu sâu về biến số và xu hướng của chúng.



Phần 1. Bối Cảnh Thống kê Mô tả

• Dữ liệu:

- Dữ liệu thu thập từ nhiều nguồn khác nhau: **đo đạc, sự kiện, văn bản, hình ảnh và videos.**
- **Internet of Things (IoT)** đang tạo ra lượng dữ liệu rất lớn thường ở dạng phi cấu trúc (unstructured). Để ứng dụng được thống kê, cần phải đưa những kiểu dữ nguyên thủy này về dạng có cấu trúc → Một trong những dạng phổ biến nhất là **dạng bảng** với hàng và cột.



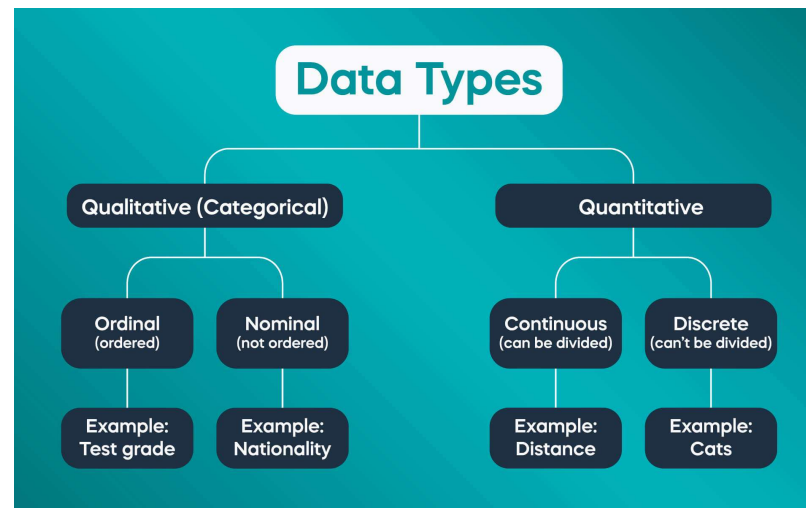
Phần 1. Bối Cảnh Thống kê Mô tả

- **Các loại dữ liệu:**

- **Dữ liệu dạng số (Numeric):** Được biểu diễn trên thang đo số

- **Dữ liệu liên tục (Continuous):** Bất kỳ giá trị nào trong một khoảng cụ thể

- **Dữ liệu rời rạc (Discrete):** Dữ liệu chỉ có thể nhận giá trị là số nguyên, như số lần xuất hiện. (Từ đồng nghĩa: số nguyên, đếm)



Phần 1. Bối Cảnh Thống kê Mô tả

• Các loại dữ liệu:

- **Dữ liệu dạng phân loại (Categorical):** Dữ liệu chỉ có thể nhận một tập hợp cụ thể các giá trị đại diện cho một nhóm các danh mục có thể xảy ra. (Từ đồng nghĩa: enums, enumerated, factors, nominal)
 - **Dữ liệu hạng mục (Nominal):** Đây là loại dữ liệu phân loại không có thứ bậc hay sắp xếp cụ thể giữa các danh mục. Các giá trị trong dữ liệu này chỉ là các nhóm hoặc loại, và không thể xác định được mức độ tương đối giữa chúng. Ví dụ: màu sắc, giới tính.
 - **Dữ liệu nhị phân (Binary):** Trường hợp đặc biệt của dữ liệu phân loại chỉ có hai danh mục giá trị, ví dụ như 0/1, đúng/sai. (Từ đồng nghĩa: dichotomous, logical, indicator, boolean)
 - **Dữ liệu thứ bậc (Ordinal):** Dữ liệu phân loại có thứ bậc là dạng dữ liệu phân loại mà có thứ tự rõ ràng. (Từ đồng nghĩa: yếu tố có thứ bậc)

Phần 1. Bối Cảnh Thống kê Mô tả

- Các loại dữ liệu:
 - Tại sao phải phân loại dữ liệu?



Phần 1. Bối Cảnh Thống kê Mô tả

• Dữ liệu dạng bảng:

- Khung frame điển hình nhất trong phân tích dữ liệu là **dữ liệu dạng bảng** (Rectangular Data) giống như bảng tính hoặc là bảng (quan hệ) trong CSDL
- Các dòng trong dữ liệu dạng bảng gọi là **records** (case)
- Các cột trong bảng còn lại là **features** (variables)
- Dataframe là **tên gọi chung** của Rectangular Data trong R và Python.
- **Thuộc tính quyết định** (Outcome): Nhiều dự án khoa học dữ liệu liên quan đến việc dự đoán một kết quả—thường là một kết quả có hoặc không
 - Biến phụ thuộc (dependent variable)
 - Phản ứng (response)
 - Mục tiêu (target)
 - Kết quả (output)

Phần 1. Bối Cảnh Thống kê Mô tả

- **Định nghĩa khái niệm Thống kê Mô tả:**

- Thống kê mô tả là một nhánh của thống kê nhằm mô tả, tổng hợp và hiểu biết các đặc tính cơ bản của dữ liệu.
- Ba công việc chính của thống kê mô tả là:
 - Ước lượng vị trí (Estimates of Location)
 - Ước lượng phân tán (Estimates of Variability)
 - Trực quan hóa dữ liệu thống kê (Visualization Statistics)



PHẦN 2

ĐO LƯỜNG GIÁ TRỊ VỊ TRÍ

Đo lường vị trí là quá trình xác định vị trí trung tâm của dữ liệu, như trung bình, trung vị và yếu vị, giúp tổng quan về xu hướng trung ương của biến số.

10



Phần 2. Đo lường giá trị vị trí

- **Định nghĩa khái niệm đo lường giá trị vị trí:**

- Các biến số có dữ liệu đo lường hoặc đếm có thể có hàng nghìn giá trị khác nhau. Một bước cơ bản trong việc khám phá dữ liệu của bạn là lấy một **"giá trị điển hình"** cho mỗi đặc trưng (biến số): ước lượng vị trí chủ yếu của dữ liệu (tức là trung bình trọng tâm của nó).
- Một số giá trị điển hình phổ biến:
 - Trung bình (Mean) (Tên khác: average)
 - Trung bình trọng số (Weight Mean) (Tên khác: Weight Average)
 - Trung vị (Median) 50th percentile
 - Percentile phần trăm dữ liệu
 - Robust
 - Outliers (Ngoại lệ)
 - Mode

Phần 2. Đo lường giá trị vị trí



- **Trung Bình (Mean)**

- Ước lượng cơ bản nhất về vị trí là trung bình, hay giá trị trung bình. Trung bình là tổng của tất cả các giá trị chia cho số lượng giá trị. Xem xét tập hợp số sau: **{3, 5, 1, 2}**. Trung bình được tính bằng cách thực hiện phép chia tổng của tất cả các giá trị cho số lượng giá trị. Trong trường hợp này, trung bình là **$(3 + 5 + 1 + 2) / 4 = 11 / 4 = 2.75$**

- **Ký hiệu của trung bình:**

\bar{x} (trung bình mẫu) hay μ (trung bình tổng thể)

- **Công thức:**

Trung bình mẫu: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

Trung bình tổng thể: $\mu = \frac{\sum_{i=1}^N x_i}{N}$

n (số lượng phần tử mẫu) hay N (số lượng phần tử tổng thể)

Phần 2. Đo lường giá trị vị trí

- **Trung Vị (Median)**

- Một tập dữ liệu số được **sắp xếp theo thứ tự** tăng dần/giảm dần. Giá trị nằm ở giữa chia tập dữ liệu ra phần trên và phần dưới được gọi là **Trung vị** (Median)
- **Cách gọi khác của Trung Vị:** 50th percentile
- Ví dụ về cách tính Trung vị
 - 1, 2, 2, 2, 2, 3, 4, **5**, 6, 7, 8, 9, 10, 12, 13
 - » Trung vị là 5
 - 2, 2, 4, 6, 6, **6**, **8**, 9, 9, 9, 10, 11
 - » Trung vị là $(6+8)/2 = 7$

Phần 2. Đo lường giá trị vị trí

• Mode

- Giá trị **phổ biến nhất** trong tập dữ liệu
- Nếu có nhiều giá trị phổ biến nhất **bằng nhau** thì mode chính là các giá trị phổ biến đó.
- Số lần xuất hiện của một giá trị trong tập dữ liệu gọi là tần số: f
- Giá trị Mode = $X[f = f_{max}]$

Ví dụ 1:

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

- Mode = $X[f = 8]$
- **Mode = 5**



PHẦN 3

ĐO LƯỜNG GIÁ TRỊ PHÂN TÁN

15

Đo lường phân tán là phương pháp đánh giá sự biến động của dữ liệu, như phạm vi, phương sai, và độ lệch chuẩn. Nó mô tả mức độ dao động, đồng thời cung cấp cái nhìn sâu sắc về sự phân bố của giá trị trong tập dữ liệu, hỗ trợ phân tích toàn diện hơn.



Phần 3. Đo lường giá trị phân tán

- **Định nghĩa khái niệm đo lường giá trị phân tán**

- **Đo lường phân tán** là một khía cạnh quan trọng trong thống kê, mô tả mức độ sự lan truyền hay phân phối của dữ liệu. Đo lường này giúp **đánh giá độ biến động** của các giá trị trong tập dữ liệu.
- Một số giá trị điển hình phổ biến:
 - Phương sai (Variance)
 - Độ lệch chuẩn (Standard Deviation)
 - Phạm Vi (Range)
 - Khoảng Tứ (IQR)
 -

Phần 3. Đo lường giá trị phân tán

- **Phạm Vi (Range)**

- Khoảng $\Delta = \text{Range} = \text{Max} - \text{Min}$

Được gọi là khoảng phạm vi

- **Phương sai (Variance)**

- Phương sai bằng tổng của các sai lệch bình phương so với giá trị trung bình chia cho $n - 1$, trong đó n là số lượng giá trị dữ liệu.

- **Ký hiệu của phương sai:**

s^2 (phương sai mẫu) hay σ^2 (phương sai tổng thể)

- **Công thức**

Phương sai mẫu: $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

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

Phần 3. Đo lường giá trị phân tán

• Độ lệch chuẩn (Deviation)

- Phương sai đo lường **mức độ biến đổi** của dữ liệu. Giá trị phương sai càng cao, tức là dữ liệu **càng phân tán** và **không tập trung quanh** giá trị trung bình. Khi dữ liệu phân tán rộng, việc dự đoán giá trị đặc trưng hoặc trung bình trở nên khó khăn. Một mô hình hoặc ước tính có thể không chính xác nếu **dữ liệu phân tán quá nhiều**.

– Ký hiệu của độ lệch chuẩn:

s (Độ lệch chuẩn mẫu) hay σ (độ lệch chuẩn tổng thể)

– Công thức

Độ lệch chuẩn mẫu: $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$

Độ lệch chuẩn tổng thể: $\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}$

Phần 3. Đo lường giá trị phân tán



- **Degrees and Freedom (N và n-1)**

- Trong các sách thống kê, luôn có một số thảo luận về tại sao chúng ta có $n - 1$ trong mẫu số của công thức phương sai, thay vì n , liên quan đến khái niệm về **bậc tự do**
- Sự phân biệt này không quan trọng vì thường n đủ lớn để sự khác biệt giữa chia cho n hoặc $n - 1$ không quá lớn.
- **Ví dụ 2:** Giả sử có một lọ với 100 viên bi. Các viên bi có trọng lượng khác nhau. Lọ này là "quần thể."
 - **Phương sai của Quần thể:** Cần tất cả 100 viên bi và sử dụng trọng lượng để tính phương sai. Ta chia cho 100 (tổng số viên bi).
 - **Phương sai của Mẫu:** Giờ, hãy tưởng chọn ngẫu nhiên chọn 5 viên bi và cân chúng. Tính phương sai cho 5 viên bi này. Nhưng lần này, chia cho 4 ($n - 1$), không phải 5.
 - **Tại sao chia cho 4 mà không chia cho 5?**

Phần 3. Đo lường giá trị phân tán

• Tứ phân vị Quantiles

- Trong thống kê, **quantile** là một khái niệm sử dụng để chia tập dữ liệu thành các phân khúc bằng nhau. Mỗi **quantile** đại diện cho một phần trăm cụ thể của tập dữ liệu. Cụ thể, **quantile thứ p** là giá trị mà $p\%$ dữ liệu trong tập hợp nhỏ hơn hoặc bằng nó.
- Quantiles thường được sử dụng để xác định các giá trị cụ thể trong tập dữ liệu và kiểm tra sự phân phối của dữ liệu.

Phần 3. Đo lường giá trị phân tán

• Tứ phân vị Quantiles

- Các ví dụ phổ biến về quantiles bao gồm:
 - **Median (hoặc 50th percentile)**: Chia tập dữ liệu thành hai phần bằng nhau. Nó là một dạng đặc biệt của quantile.
 - **Quartiles**: Là các quantile chia tập dữ liệu thành bốn phần bằng nhau. Q1 là 25th percentile, Q2 là median (50th percentile), và Q3 là 75th percentile.
 - **Percentiles**: Là các quantile chia tập dữ liệu thành một trăm phần bằng nhau.
- **Quantiles** giúp hiểu rõ hơn về sự **phân bố của dữ liệu** và cũng hữu ích trong việc phát hiện các giá trị ngoại lệ (**outliers**) trong tập dữ liệu.

Phần 3. Đo lường giá trị phân tán

- **Tứ phân vị Quantiles**

- **Cho dữ liệu sau đây:** 1, 12, 0, 7, 6, 8, 9, 10, 11, 13, 15

- Sắp xếp dãy số theo thứ tự tăng dần: 0, 1, 6, 7, 8, 9, 10, 11, 12, 13, 15.
- **Quartile 1 (Q1)** là giá trị ở vị trí thứ 25%. Vị trí này là $(25/100) * (11 + 1) = 3\text{rd position}$. Vậy $Q1 = 6$.
- **Quartile 2 (Q2)**, cũng chính là median, là giá trị ở vị trí thứ 50%. Vị trí này là $(50/100) * (11 + 1) = 6\text{th position}$. Vậy $Q2 = 9$.
- **Quartile 3 (Q3)** là giá trị ở vị trí thứ 75%. Vị trí này là $(75/100) * (11 + 1) = 9\text{th position}$. Vậy $Q3 = 12$.

Phần 3. Đo lường giá trị phân tán

- **Khoảng Tứ (IQR)**

- IQR được gọi là khoảng tứ của giữa liệu $IQR = Q3 - Q1$
- Giá trị Sàn của tứ phân vị $= Q1 - 1,5 * IQR$
- Giá trị Trần của tứ phân vị $= Q3 + 1,5 * IQR$
- **Bài toán tìm ngoại lai (Outliers)**

$x_i \in X$ và là *Outliers*: $x_i < \text{Sàn}$ hoặc $x_i > \text{Trần}$



PHẦN 4

ĐO LƯỜNG GIÁ TRỊ HÌNH DẠNG 24

Skewness và kurtosis là độ đo quan trọng trong thống kê, mô tả giá trị hình dạng của phân phối xác suất. Skewness đo độ chệch, kurtosis đo độ nhọn, cung cấp thông tin quan trọng về tính chất phân phối dữ liệu.



Phần 4. Đo lường giá trị hình dạng

- **Skewness (Độ nghiêng)**

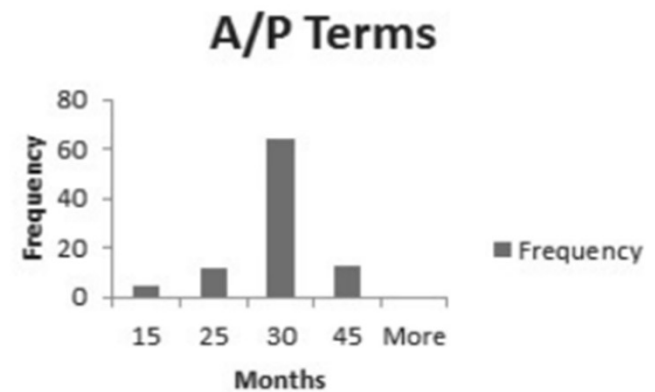
– Skewness là một đặc tính thống kê được sử dụng để **đo độ lệch** của phân phối dữ liệu. Nó chỉ ra mức độ mà dữ liệu có **xu hướng lệch về một phía** so với giá trị trung bình của nó. Skewness còn có tên khác là CS (coefficient of skewness). Khi CS dương nghiêng về bên phải, CS âm nghiêng về bên trái.

– **Công thức tính Skewness:** **Skewness:** $CS = \frac{\frac{1}{N} \sum_{i=1}^n (x_i - \mu)^3}{\sigma^3}$

Phần 4. Đo lường giá trị hình dạng

- **Skewness (Độ nghiêng)**

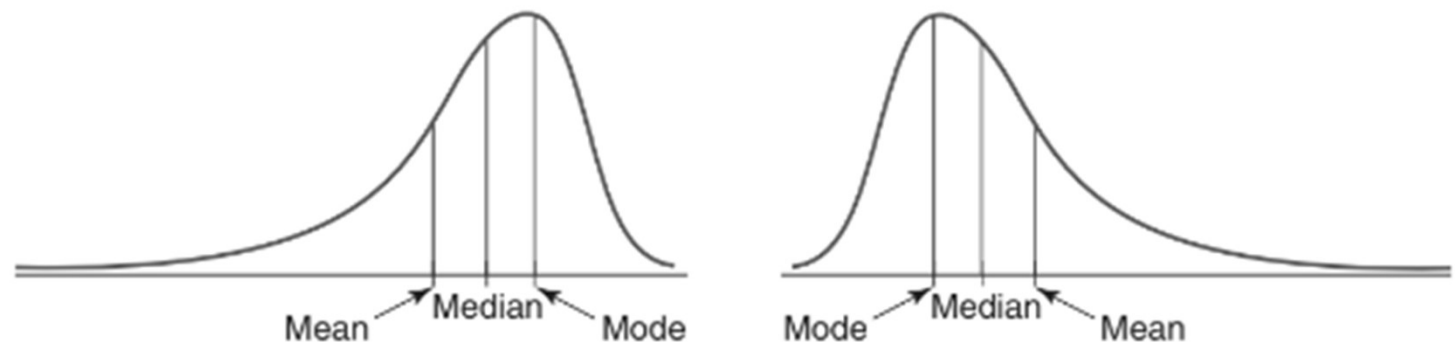
- Công thức tính Skewness: **Skewness:** $CS = \frac{\frac{1}{N} \sum_{i=1}^n (x_i - \mu)^3}{\sigma^3}$
- Cost Per Order có $CS = 1.66$
- A/P Terms có $CS = 0.66$



Phần 4. Đo lường giá trị hình dạng

- **Skewness (Độ nghiêng)**

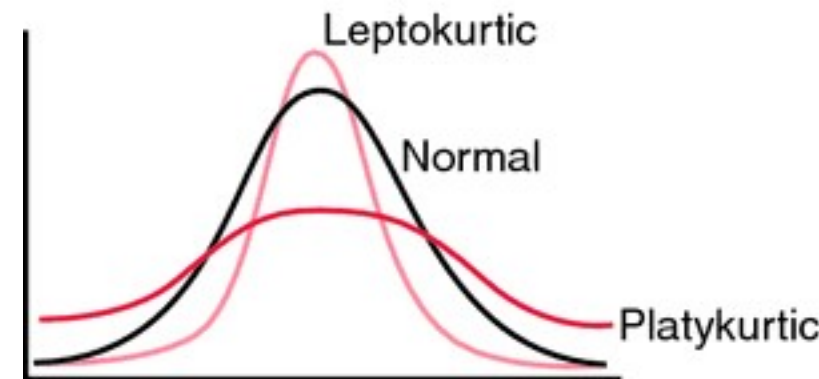
- Công thức tính Skewness: **Skewness**: $CS = \frac{\frac{1}{N} \sum_{i=1}^n (x_i - \mu)^3}{\sigma^3}$
- $CS > 1$ hoặc $CS < -1$ dữ liệu nghiêng về **chiều ngược lại**
- CS thuộc đoạn $(0,5;1]$ hoặc $[-1;-0,5)$ dữ liệu **ngiên nhẹ**
- CS càng gần về số 0 dữ liệu càng **bằng phẳng** không nghiêng.



Phần 4. Đo lường giá trị hình dạng

• Kurtosis (Độ Bè)

- Kurtosis đề cập đến **độ đỉnh** (cao và hẹp) hoặc **độ phẳng** (thấp và phẳng) của một biểu đồ tần số. Hệ số kurtosis (CK) đo lường mức độ kurtosis của một tổng thể và được tính bằng công thức như sau:
- **Công thức tính Kurtosis:** Kurtosis: $CK = \frac{\frac{1}{N} \sum_{i=1}^n (x_i - \mu)^4}{\sigma^4}$
- LeptoKurtic (Đỉnh nhọn) khi $CK > 3$
- Normal khi CK gần bằng 3
- PlatyKurtic (Đỉnh bè) khi $CK < 3$





PHẦN 5

HISTOGRAM VÀ BOXPLOT

29

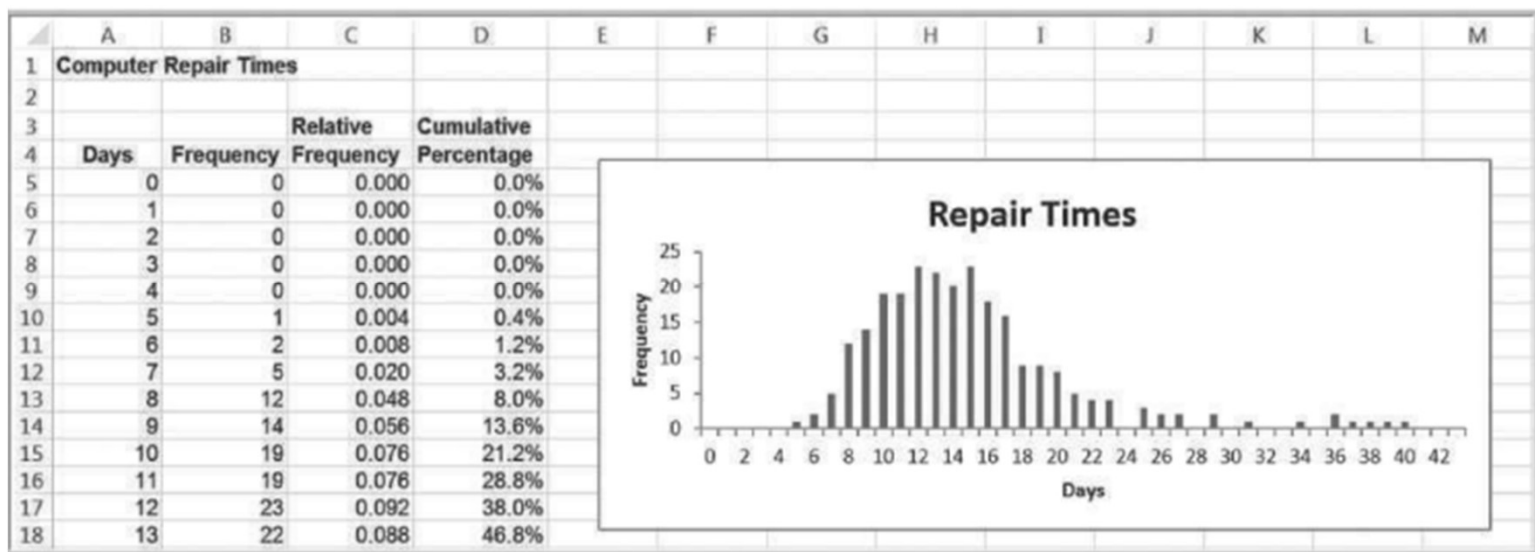
Trực quan hóa dữ liệu thống kê mô tả trên biểu đồ Histogram và BoxPlot nhằm giải thích sự phân tán dữ liệu trên hai biểu đồ này.



Phần 5. Histogram và BoxPlot

- **Histogram**

- Là biểu đồ thể giá trị **tần số** của các biến giá trị
- Ngoài ra còn đo lường giá trị phân tán, CK, CS của tập dữ liệu



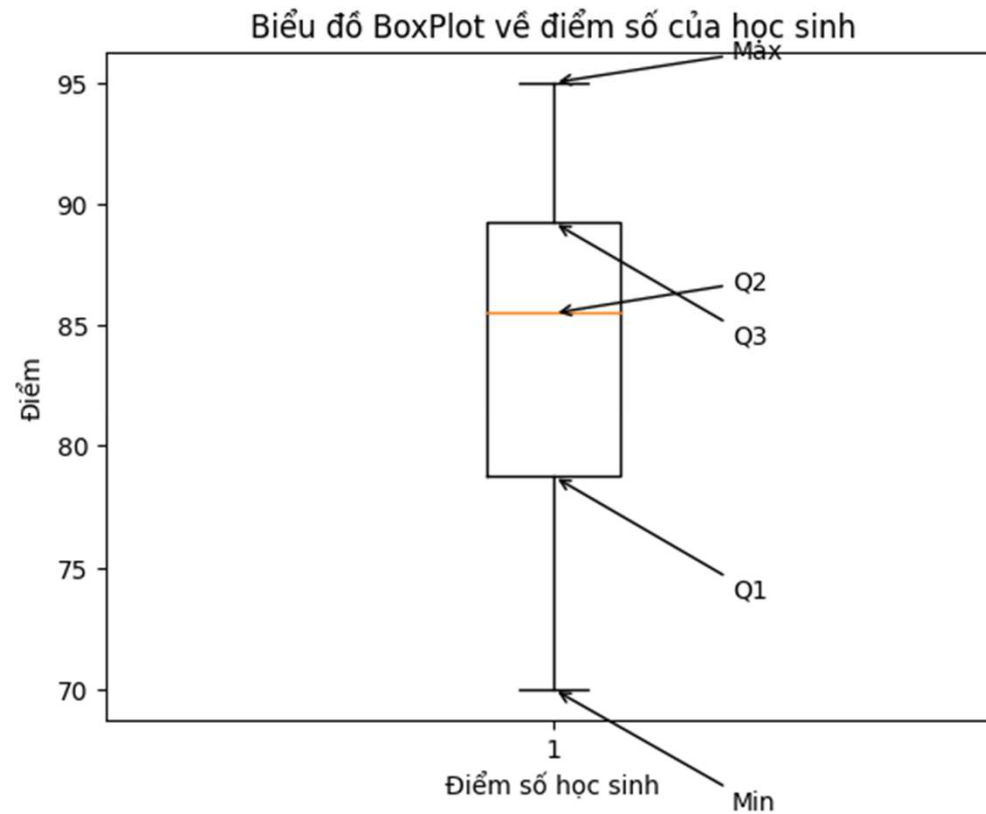
Phần 5. Histogram và BoxPlot

• BoxPlot

- Là một dạng biểu đồ được giới thiệu bởi Tukey để trực quan hóa dữ liệu phân tán.
- Với tập dữ liệu **điểm học sinh như sau**:
 - 75, 82, 90, 88, 78, 92, 85, 88, 78, 70, 95, 86, 89, 80, 92, 88, 75, 82, 91, 79
- Sắp xếp **tập dữ liệu điểm học sinh tìm tứ phân vị**
 - 70, 75, 75, 78, 78, 79, 80, 82, 82, 85, 86, 88, 88, 88, 89, 90, 91, 92, 92, 95
- Tìm Q1, Q2, Q3 của tập dữ liệu
 - **Q1** = 78.5
 - **Q2** = 85.5
 - **Q3** = 90.5
- Sàn = $Q1 - 1,5 \cdot IQR$, Trần = $Q3 + 1,5 \cdot IQR$

Phần 5. Histogram và BoxPlot

- BoxPlot

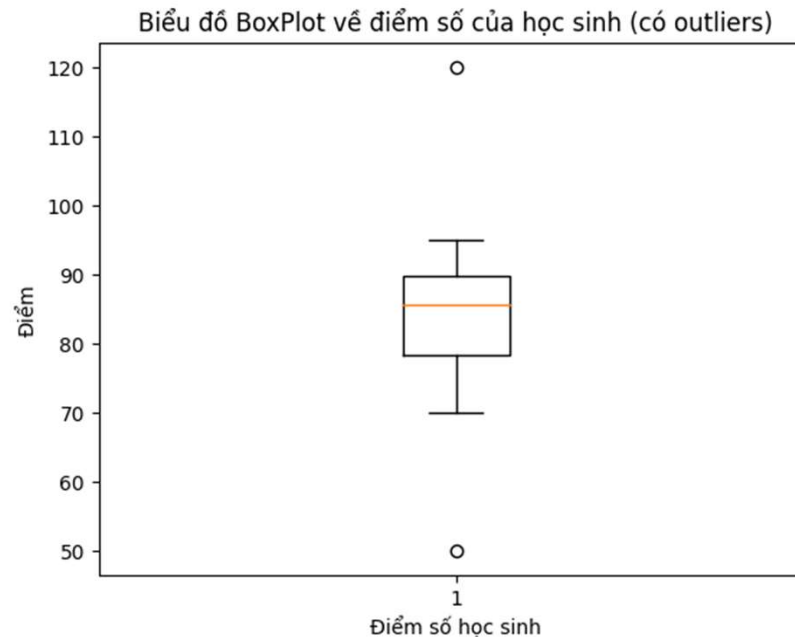


Phần 5. Histogram và BoxPlot

- **BoxPlot**

- Với tập dữ liệu **điểm học sinh như sau:**

- 75, 82, 90, 88, 78, 92, 85, 88, 78, 70, 95, 86, 89, 80, 92, 88, 75, 82, 91, 79, 120, 50



Reference

**Python
Statistics**

Nguyen Minh Nhut
UIT – Information System

- [1] Peter Bruce (Author), Andrew Bruce (Author), Peter Gedeck (Author), Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python 2nd Edition
- [2] James R Evans, Business Analytics-Pearson (2017)



CẢM ƠN ĐÃ THEO DÕI



ftisu.vn

