# 1.Thống kê mô tả là gì?

-Thống kê mô tả là phương pháp thu thập, tổ chức, phân tích và diễn giải dữ liệu nhằm tóm tắt, mô tả các đặc điểm chính của một tập dữ liệu hoặc mẫu nghiên cứu, có thể là các chỉ số như trung bình, trung vị, độ lệch chuẩn, hoặc các biểu đồ trực quan như histogram. Mục tiêu chính của nó là làm cho dữ liệu phức tạp trở nên dễ hiểu hơn, cung cấp cái nhìn tổng quan mà không đưa ra kết luận hoặc suy luận về một quần thể lớn hơn.

**+Nó khác gì với thống kê suy luận (inferential statistics)?**

-Thống kê mô tả (descriptive statistics) là việc tóm tắt và mô tả các đặc điểm chính của một tập dữ liệu, giúp hiểu rõ hơn về dữ liệu đó thông qua các chỉ số như trung bình, trung vị, phương sai và biểu đồ. Ngược lại, thống kê suy luận (inferential statistics) sử dụng dữ liệu từ một mẫu để rút ra kết luận, đưa ra dự đoán hoặc khái quát hóa cho một quần thể lớn hơn.

|  |  |  |
| --- | --- | --- |
|  | Thống kê mô tả | Thống kê suy luận |
| Mục tiêu | Tóm tắt, sắp xếp và trình bày dữ liệu một cách có ý nghĩa và ngắn gọn để phác họa một bức tranh rõ ràng về dữ liệu đã thu thập. | Rút ra kết luận, kiểm định giả thuyết hoặc đưa ra dự đoán về một quần thể lớn hơn dựa trên dữ liệu thu thập được từ một mẫu đại diện. |
| Phạm vi | Chỉ tập trung vào mô tả các đặc điểm của tập dữ liệu đang được phân tích, không cố gắng khái quát hóa sang các tập dữ liệu khác. | Sử dụng dữ liệu mẫu để suy diễn và đưa ra những kết luận có tính khái quát hóa cho toàn bộ dân số. |
| Phương pháp | Sử dụng các biện pháp đo lường xu hướng tập trung (như trung bình, trung vị, yếu vị), các biện pháp đo lường độ phân tán (như phương sai, độ lệch chuẩn), và các hình dạng phân phối (như độ lệch, độ nhọn). | Sử dụng các công cụ phân tích thống kê và xác suất để ước lượng các tham số của quần thể hoặc kiểm định các giả thuyết. |
| Ví dụ | Tính điểm trung bình của học sinh trong một lớp học, hoặc biểu diễn số lượng bán hàng hàng ngày bằng biểu đồ. | Sử dụng dữ liệu khảo sát ý kiến của một nhóm nhỏ khách hàng để dự đoán ý kiến chung của tất cả khách hàng, hoặc ước tính doanh số của sản phẩm tương ớt mới dựa trên dữ liệu bán hàng trong quá khứ. |

**Sự khác biệt chính:**

* **Mục đích:**

Mô tả dữ liệu có sẵn so với suy luận/dự đoán về một quần thể lớn hơn.

* **Phạm vi:**

Chỉ mô tả mẫu so với khái quát hóa từ mẫu sang quần thể.

# 2. Các thước đo thống kê mô tả chính (ví dụ: trung bình, trung vị, phương sai, độ lệch chuẩn) được sử dụng để làm gì?

-Các thước đo thống kê mô tả chính như trung bình, trung vị (xu hướng trung tâm) và phương sai, độ lệch chuẩn (phân tán) được dùng để tóm tắt và mô tả các đặc điểm cơ bản của một tập dữ liệu, giúp người dùng hiểu rõ cấu trúc, tính biến động và sự phân bố của dữ liệu đó một cách cô đọng và dễ hiểu.

|  |  |
| --- | --- |
| Trung bình (Mean) | +Dùng để biểu diễn **giá trị đại diện chung** cho dữ liệu.  +Thích hợp khi dữ liệu phân phối tương đối cân đối (không có quá nhiều ngoại lệ). |
| Trung vị (Median) | +Dùng để tìm **giá trị ở giữa** khi sắp xếp dữ liệu.  +Cho biết 50% dữ liệu nhỏ hơn và 50% lớn hơn.  +Ít bị ảnh hưởng bởi **giá trị ngoại lai** (outliers). |
| Phương sai (Variance) | Đo mức độ **phân tán tổng thể** của dữ liệu quanh giá trị trung bình. |
| Độ lệch chuẩn (Standard deviation) | +Dễ hiểu hơn phương sai (vì cùng đơn vị với dữ liệu gốc).  +Cho biết dữ liệu **thường lệch khỏi trung bình bao nhiêu** |

**+Trong trường hợp nào thì nên dùng trung vị thay vì trung bình?**

Nên dùng trung vị (median) thay vì trung bình (mean) khi dữ liệu có các giá trị ngoại lai (cực đoan) hoặc dữ liệu bị lệch phân phối, vì trung vị ít bị ảnh hưởng bởi các giá trị này hơn và đại diện chính xác hơn cho giá trị trung tâm hoặc điển hình của tập dữ liệu đó.

Ví dụ:

* **Thu nhập:**

Nếu tính thu nhập trung bình của một nhóm người, một vài cá nhân có thu nhập cực cao có thể làm tăng giá trị trung bình lên đáng kể. Tuy nhiên, trung vị sẽ cho biết mức thu nhập của gia đình "điển hình" tốt hơn, vì nó không bị ảnh hưởng bởi những thu nhập "khủng" đó.

* **Hiệu suất bán hàng:**

Khi đánh giá hiệu suất bán hàng của một nhóm nhân viên, trung vị sẽ giúp loại bỏ ảnh hưởng của một vài nhân viên có hiệu suất cực kỳ cao hoặc thấp, cho nhà quản lý một cái nhìn chính xác hơn về hiệu suất chung của nhóm.

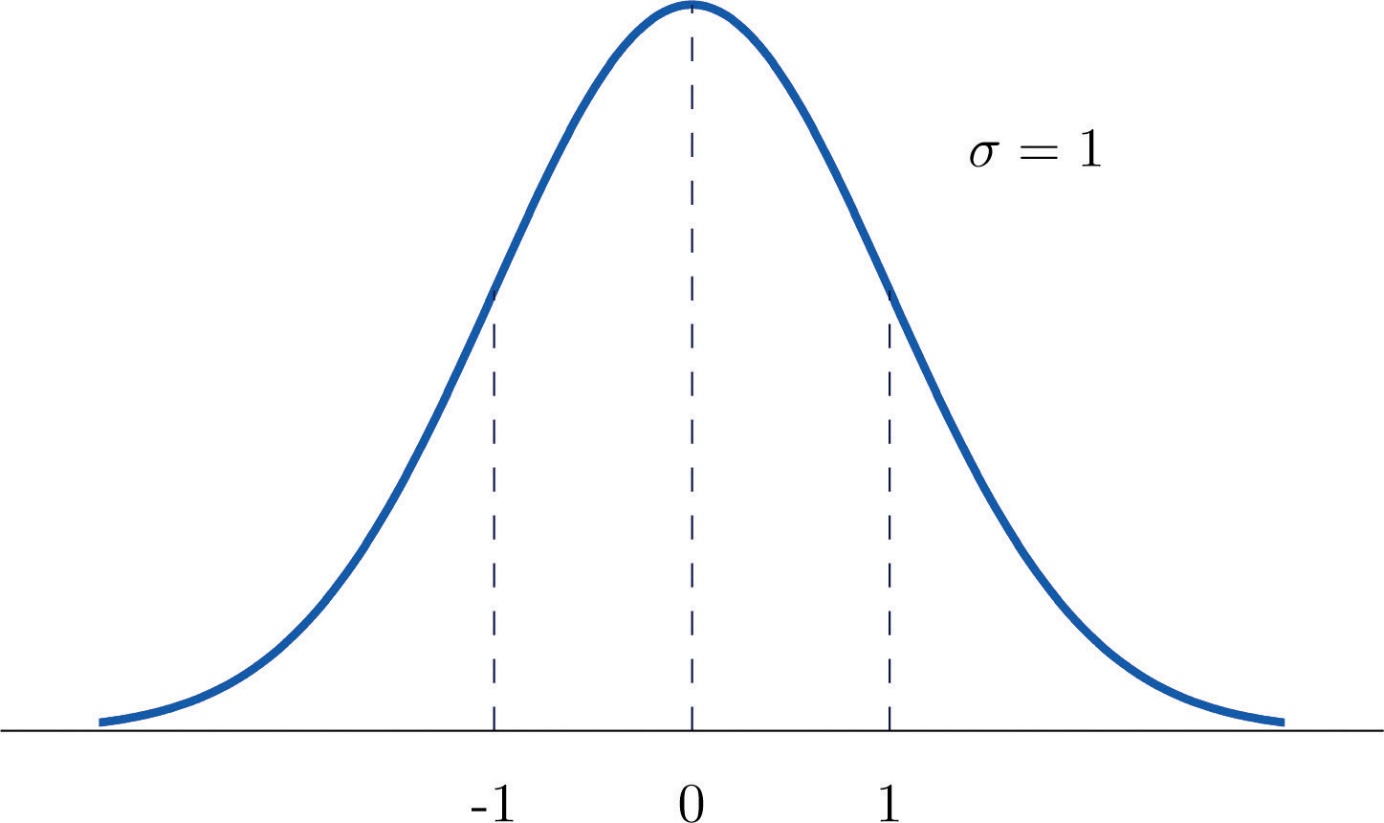
# 3. Làm thế nào để xác định phân bố của một tập dữ liệu?

-Để xác định phân bố của một tập dữ liệu, bạn có thể sử dụng biểu đồ Histogram để trực quan hóa hình dạng phân phối, biểu đồ xác suất (Q-Q plot) để kiểm tra tính chuẩn, và áp dụng kiểm định độ phù hợp phân phối như kiểm định Kolmogorov-Smirnov hoặc Shapiro-Wilk để đánh giá xem dữ liệu có tuân theo một phân phối giả định (ví dụ: phân phối chuẩn) hay không.

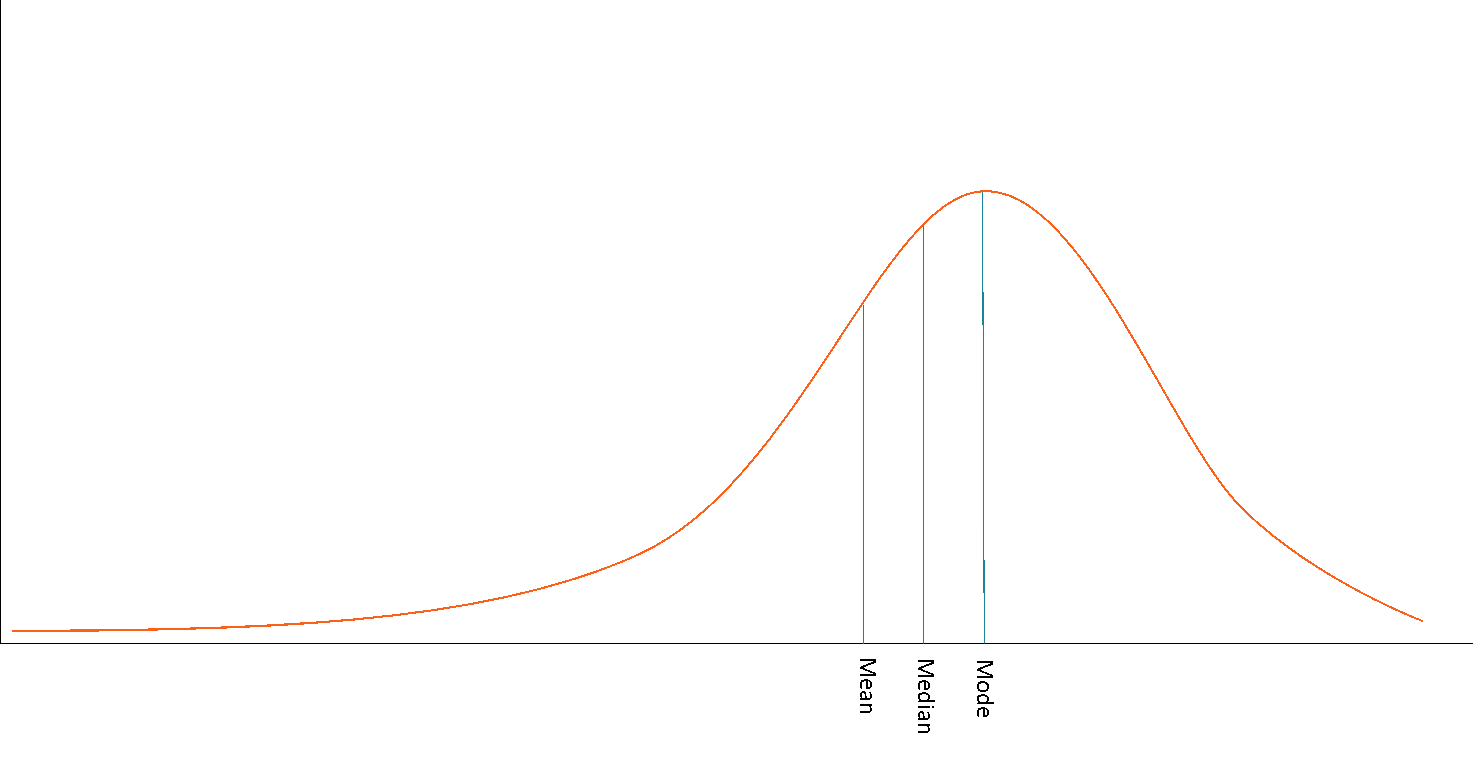
**+Các loại phân bố phổ biến là gì (ví dụ: phân bố chuẩn, lệch trái, lệch phải)?**

**-**Các loại phân bố phổ biến bao gồm Phân bố chuẩn, Phân phối lệch trái (lệch âm) và Phân phối lệch phải (lệch dương). Phân bố chuẩn là dạng đối xứng hình chuông, còn phân phối lệch trái có đuôi dài về bên trái và trung vị lớn hơn trung bình, trong khi phân phối lệch phải có đuôi dài về bên phải và trung bình lớn hơn trung vị.

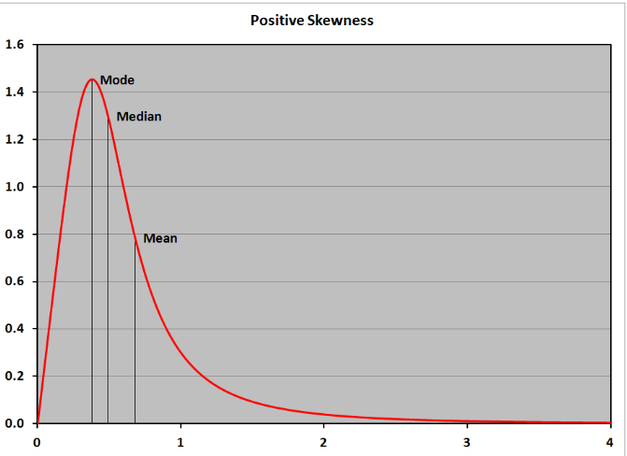
* **Phân phối chuẩn (chuẩn mực)**
  + **Đặc điểm**: Là một phân bố đối xứng, có hình dạng giống như quả chuông.
  + **Độ lệch**: Độ lệch (skewness) bằng 0.
  + **Trung tâm**: Các thước đo xu hướng trung tâm như trung bình, trung vị và mốt có cùng giá trị.
  + **Ví dụ**: Tỷ lệ chiều cao của con người trong một quần thể lớn.



* **Phân phối lệch trái (Lệch âm)**
  + **Đặc điểm**: Phần đuôi của đồ thị kéo dài về phía bên trái, tức là về phía các giá trị âm.
  + **Độ lệch**: Trung vị (median) lớn hơn trung bình (mean), vì các giá trị nhỏ làm giảm giá trị trung bình.
  + **Ví dụ**: Điểm thi của một lớp học có nhiều học sinh giỏi, nên điểm số cao tập trung nhiều ở bên phải, và một vài điểm thấp làm kéo đuôi về bên trái.



* **Phân phối lệch phải  (Lệch dương)**
  + **Đặc điểm**: Phần đuôi kéo dài về phía bên phải, về phía các giá trị dương.
  + **Độ lệch**: Trung bình (mean) lớn hơn trung vị (median), vì các giá trị cao làm tăng giá trị trung bình.
  + **Ví dụ**: Phân phối thu nhập của người dân trong một quốc gia, nơi đa số có thu nhập trung bình, nhưng có một số ít người có thu nhập rất cao làm kéo đuôi về bên phải.



# 4. Độ lệch chuẩn và phạm vi (range) có ý nghĩa gì trong việc đánh giá sự phân tán của dữ liệu?

-Độ lệch chuẩn và phạm vi đều đánh giá sự phân tán dữ liệu, nhưng độ lệch chuẩn đo lường sự biến động trung bình của các điểm dữ liệu so với giá trị trung bình, cho thấy mức độ tập trung hoặc phân tán xung quanh trung tâm, trong khi phạm vi đơn giản là hiệu số giữa giá trị lớn nhất và nhỏ nhất, chỉ cung cấp một cái nhìn đơn giản về sự lan truyền mà không phản ánh sự phân bố của các giá trị còn lại.   
Độ lệch chuẩn (Standard Deviation)

* **Ý nghĩa:**

Đo lường mức độ các giá trị trong tập dữ liệu phân tán như thế nào so với giá trị trung bình.

* **Giải thích:**
  + **Độ lệch chuẩn cao:** Các giá trị dữ liệu có xu hướng phân tán rộng, cách xa giá trị trung bình, cho thấy tính biến động cao.
  + **Độ lệch chuẩn thấp:** Các giá trị dữ liệu tập trung gần giá trị trung bình, cho thấy dữ liệu ít biến động và ổn định hơn.
* **Ứng dụng:**

Giúp hiểu được sự biến động và rủi ro trong dữ liệu, ví dụ như rủi ro của các khoản đầu tư.

Phạm vi (Range)

* **Ý nghĩa:**

Là hiệu số giữa giá trị lớn nhất và giá trị nhỏ nhất trong tập dữ liệu, cho biết toàn bộ độ trải dài của dữ liệu.

* **Giải thích:**
  + Phạm vi chỉ cho biết hai thái cực của dữ liệu.
  + Một phạm vi lớn có thể chỉ ra một tập dữ liệu rất phân tán hoặc chỉ là có hai giá trị cực đoan trong một tập dữ liệu ổn định.
* **Hạn chế:**

Phạm vi không cung cấp thông tin về sự phân bố của các giá trị ở giữa, và có thể bị ảnh hưởng mạnh bởi các giá trị ngoại lai (outliers).

# 5. Sự khác biệt giữa các thước đo như Q1, Q2, Q3 trong biểu đồ hộp (boxplot) là gì?

Trong biểu đồ hộp, Q1 là tứ phân vị thứ nhất, đại diện cho giá trị mà 25% dữ liệu nhỏ hơn nó, còn Q2 (hay trung vị) là giá trị mà 50% dữ liệu nhỏ hơn nó, và Q3 là tứ phân vị thứ ba, nơi 75% dữ liệu nhỏ hơn nó. Q1, Q2, và Q3 là các điểm quan trọng trong hộp của biểu đồ, giúp chia dữ liệu thành các phần tư bằng nhau và thể hiện sự phân bố của dữ liệu.

|  |  |  |
| --- | --- | --- |
|  | **Giải thích chi tiết** | **Vai trò của Q1, Q2, Q3 trong biểu đồ hộp** |
| **Q1 (Tứ phân vị thứ nhất)** | Là trung vị (median) của nửa dưới của tập dữ liệu (các giá trị nhỏ hơn Q2). Nó đánh dấu điểm 25% của dữ liệu, nghĩa là 25% các điểm dữ liệu có giá trị thấp hơn Q1. | Các đường kẻ bên trong biểu đồ hộp cho bạn biết vị trí của Q1, Q2 (trung vị), và Q3. |
| **Q2 (Trung vị)** | Là trung vị của toàn bộ tập dữ liệu. Nó chia dữ liệu thành hai nửa bằng nhau, với 50% dữ liệu nhỏ hơn Q2 và 50% lớn hơn Q2. | Khoảng cách giữa Q1 và Q3 được gọi là khoảng tứ phân vị ([IQR](https://www.google.com/search?rlz=1C1GCEA_enVN1175VN1175&cs=0&sca_esv=36e49ebf526829de&sxsrf=AE3TifM2vZEWUvfJ0wq8To2FrCp080ti2Q%3A1759215609632&q=IQR&sa=X&ved=2ahUKEwiP8rvr9P-PAxVCZfUHHeZ3G30QxccNegQILBAB&mstk=AUtExfD14as2jEddtS988iF6KnO3JHYNesva8EAMcKPa99QBTvQjwdmqbJ1jPwGvPWmZqq1nZHiiAJFnYm2VOAqclkfvQa0MRnMkAtzr8fdUjWv-UocIh_eexGiUw8sPV_Q6gkqyr53IJ2aVGlrnzthvuv7kgoyTPSYjYkU7KuJCopJLRpw&csui=3)). IQR là một thước đo quan trọng cho sự biến động của dữ liệu ở phần giữa. |
| **Q3 (Tứ phân vị thứ ba)** | Là trung vị của nửa trên của tập dữ liệu (các giá trị lớn hơn Q2). Nó đánh dấu điểm 75% của dữ liệu, nghĩa là 75% các điểm dữ liệu có giá trị thấp hơn Q3 | Vị trí của Q2 trong hộp và độ dài của các "râu" có thể chỉ ra liệu dữ liệu có bị lệch trái (nhiều giá trị lớn) hay lệch phải (nhiều giá trị nhỏ). |

# 6. Làm thế nào để xử lý giá trị thiếu (missing values) trước khi tính toán các chỉ số thống kê mô tả?

Thiếu dữ liệu là một vấn đề đau đầu thường gặp trong bất kỳ lĩnh vực nào liên quan đến bộ dữ liệu. Nó có thể phát sinh do nhiều lý do, từ lỗi của con người trong quá trình thu thập dữ liệu đến những hạn chế của phương pháp thu thập dữ liệu. May mắn thay, có những chiến lược để giải quyết vấn đề thiếu dữ liệu và giảm thiểu tác động của nó đến phân tích của bạn. Dưới đây là hai phương pháp chính:

* **Xóa:** Phương pháp này bao gồm việc xóa các hàng hoặc cột có giá trị bị thiếu. Đây là một phương pháp đơn giản, nhưng có thể gây ra vấn đề nếu một phần đáng kể dữ liệu của bạn bị thiếu. Việc loại bỏ quá nhiều dữ liệu có thể ảnh hưởng đến độ tin cậy của kết luận.
* **Quy nạp:** Phương pháp này thay thế các giá trị bị thiếu bằng các ước tính. Có nhiều kỹ thuật quy nạp khác nhau, mỗi kỹ thuật đều có ưu và nhược điểm riêng. Dưới đây là một số kỹ thuật phổ biến:
  + **Quy nạp Trung bình/Trung vị/Mốt:** Thay thế các mục bị thiếu bằng giá trị trung bình (mean), giá trị ở giữa (median) hoặc giá trị thường xuyên nhất (mốt) của cột tương ứng. Đây là một phương pháp nhanh chóng và dễ dàng, nhưng có thể gây ra sai lệch nếu dữ liệu bị thiếu không được phân phối ngẫu nhiên.
  + **K-Láng giềng gần nhất (KNN Imputation):** Phương pháp này tìm các điểm dữ liệu gần nhất (láng giềng) dựa trên các đặc trưng có sẵn và sử dụng giá trị của chúng để ước tính giá trị bị thiếu. KNN hữu ích khi bạn có nhiều dữ liệu và các giá trị bị thiếu nằm rải rác.
  + **Quy nạp dựa trên mô hình:** Phương pháp này bao gồm việc tạo ra một mô hình thống kê để dự đoán các giá trị bị thiếu dựa trên các đặc điểm khác trong dữ liệu. Đây có thể là một kỹ thuật mạnh mẽ, nhưng đòi hỏi nhiều chuyên môn hơn và có thể tốn kém về mặt tính toán.pen\_spark.

# 7. Bạn có thể giải thích cách đọc và diễn giải một biểu đồ histogram hoặc boxplot từ dữ liệu thực tế

không?

1. Histogram (biểu đồ tần suất)

🔹 Cách đọc

Trục X (Ox): các khoảng giá trị (bins).

Trục Y (Oy): số lượng hoặc tần suất dữ liệu rơi vào khoảng đó.

Hình dạng: cho biết phân phối dữ liệu (chuẩn, lệch trái, lệch phải, đa đỉnh…).

🔹 Diễn giải

Nếu chuẩn → histogram đối xứng, hình chuông.

Nếu lệch phải → cột cao tập trung bên trái, đuôi kéo dài về bên phải.

Nếu lệch trái → cột cao tập trung bên phải, đuôi kéo dài về bên trái.

Nhiều đỉnh → dữ liệu có nhiều nhóm khác nhau.

Ví dụ: Histogram điểm thi của 100 SV

Cột cao nhất ở khoảng 7–8 → nhiều SV đạt điểm 7–8.

Đuôi kéo dài về bên phải đến 10 → có ít SV đạt điểm tuyệt đối.  
→ Kết luận: phân bố lệch phải nhẹ, điểm tập trung quanh 7–8.

2. Boxplot (biểu đồ hộp)

\*Cách đọc

Đường giữa hộp = Median (Q2).

Đáy hộp = Q1 (25%).

Đỉnh hộp = Q3 (75%).

Chiều dài hộp = IQR = Q3 – Q1 → độ phân tán của 50% dữ liệu trung tâm.

Râu (whiskers): min, max (không tính ngoại lệ).

Điểm lẻ ngoài râu: outliers.

\*Diễn giải

Median nằm giữa hộp → dữ liệu cân đối.

Median lệch về gần đáy hộp → phân bố lệch phải.

Median lệch về gần đỉnh hộp → phân bố lệch trái.

Hộp càng dài → dữ liệu phân tán mạnh.

Nhiều outliers → dữ liệu có giá trị bất thường.

Ví dụ: Boxplot lương nhân viên

Q1 = 8 triệu, Median = 12 triệu, Q3 = 15 triệu.

Có vài outliers ở 50–100 triệu.  
→ Kết luận: đa số nhân viên có lương 8–15 triệu, một số ít có thu nhập rất cao kéo phân bố lệch phải.

# 8. Khi gặp một tập dữ liệu có giá trị ngoại lai (outliers), bạn sẽ xử lý chúng như thế nào trước khi thực hiện thống kê mô tả?

Các cách xử lý outliers trước khi tính thống kê mô tả

1) Xác định xem outlier có hợp lý hay không

Đầu tiên phải phân biệt:

Outlier do lỗi nhập liệu/thu thập → cần sửa hoặc loại bỏ.

Outlier thực sự của hiện tượng → có thể giữ lại (vì nó mang ý nghĩa).

Dùng biểu đồ boxplot, z-score, IQR để phát hiện.

2)Các cách xử lý

-Loại bỏ outlier (Remove)

Nếu dữ liệu lớn và outliers ít → có thể bỏ để tránh ảnh hưởng đến thống kê.

Ví dụ: lương của 1 CEO quá cao trong khảo sát nhân viên.

Dùng thước đo ít nhạy với outlier

Thay mean bằng median.

Thay variance/SD bằng IQR (Q3 – Q1).

Cách này giữ nguyên dữ liệu nhưng vẫn phản ánh trung tâm và độ phân tán chính xác hơn.

Biến đổi dữ liệu (Transformation)

Dùng log, sqrt, winsorization (giới hạn giá trị lớn/nhỏ về một ngưỡng).

Giúp giảm ảnh hưởng của các giá trị quá cực đoan.

Thay thế (Imputation)

Nếu outlier rõ ràng là lỗi, có thể thay bằng median hoặc giá trị gần đó.

Phân tích riêng biệt

Với dữ liệu quan trọng, có thể tách thành 2 nhóm: dữ liệu chính + outliers → so sánh riêng.