TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

**BÁO CÁO**

**THỰC HÀNH: PHÂN TÍCH KHÁM PHÁ DỮ LIỆU**

Môn học: Nhập môn máy học

Giảng viên hướng dẫn: **Đỗ Như Tài**

Nhóm lớp học: **Nhóm 01 (Thứ 7, từ tiết 6 đến tiết 10 )**

Sinh viên thực hiện: **Lê Khánh Hoàng** **MSSV:** **3122410125**

**TP.HCM, ngày 18 tháng 5 năm 2025**

**NHẬN XÉT CỦA GIẢNG VIÊN**

………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

**Mục lục**

**Phần 1: THỐNG KÊ MÔ TẢ**

* + 1. Ôn tập lý thuyết
    2. Lý do chọn đề tài
    3. Bài tập thực hành 1
    4. Bài tập thực hành 2

**Phần 2: XỬ LÍ VÀ TRỰC QUAN HOÁ DỮ LIỆU**

* + 1. Ôn tập lý thuyết
    2. Bài làm mẫu
    3. Bài tập thực hành 1
    4. Bài tập thực hành 2

**Phần 3: PHÂN TÍCH ĐƠN BIẾN VÀ HAI BIẾN**

* + 1. Ôn tập lý thuyết
    2. Bài làm mẫu
    3. Bài tập thực hành 1
    4. Bài tập thực hành 2

**PHẦN 1: THỐNG KÊ MÔ TẢ**

**1.1.1 Ôn tập lý thuyết**

**+ Thống kê mô tả là gì? Nó khác gì với thống kê suy luận (inferential statistics)?**

**Thống kê mô tả (Descriptive Statistics) là gì?**

Thống kê mô tả là tập hợp các phương pháp dùng để tóm tắt, tổ chức và trình bày dữ liệu một cách dễ hiểu, giúp mô tả các đặc điểm chính của tập dữ liệu mà không đưa ra kết luận vượt ra ngoài dữ liệu đã thu thập. Nó tập trung vào việc cung cấp một cái nhìn tổng quan về dữ liệu thông qua các số liệu, biểu đồ, hoặc bảng, thường được sử dụng trong giai đoạn khám phá dữ liệu (EDA).

Các thành phần chính của thống kê mô tả:

1. Đo lường xu hướng trung tâm (Central Tendency):
   * Mean (Trung bình): Tổng các giá trị chia cho số lượng giá trị, ví dụ: trung bình điểm thi của một lớp.
   * Median (Trung vị): Giá trị ở giữa khi dữ liệu được sắp xếp theo thứ tự, ít nhạy cảm với giá trị ngoại lai hơn mean.
   * Mode (Mode): Giá trị xuất hiện nhiều nhất trong tập dữ liệu.
2. Đo lường độ phân tán (Dispersion):
   * Variance (Phương sai): Đo lường mức độ phân tán của dữ liệu so với trung bình.
   * Standard Deviation (Độ lệch chuẩn): Căn bậc hai của phương sai, dễ diễn giải hơn vì cùng đơn vị với dữ liệu.
   * Range (Khoảng giá trị): Sự chênh lệch giữa giá trị lớn nhất và nhỏ nhất.
   * Interquartile Range (IQR): Khoảng cách giữa tứ phân vị thứ ba (Q3) và thứ nhất (Q1), dùng để phát hiện ngoại lai.
3. Phân bố (Distribution):
   * Mô tả hình dạng của dữ liệu, như phân bố chuẩn (normal), lệch trái/phải (skewed), hoặc đa đỉnh (multimodal).
   * Thường được trực quan hóa qua histogram, boxplot, hoặc density plot.
4. Tóm tắt định tính:
   * Với dữ liệu phân loại (categorical), thống kê mô tả bao gồm tần suất (frequency) hoặc tỷ lệ phần trăm của từng danh mục, ví dụ: số lượng khách hàng theo khu vực.

Ví dụ:

Giả sử bạn có dữ liệu về doanh thu hàng tháng của một cửa hàng (triệu VND): [500, 600, 550, 700, 300]. Thống kê mô tả có thể bao gồm:

* Mean: (500 + 600 + 550 + 700 + 300) / 5 = 530 triệu.
* Median: 550 triệu (sắp xếp: 300, 500, 550, 600, 700).
* Range: 700 - 300 = 400 triệu.
* Trực quan hóa: Vẽ histogram hoặc bar chart để thấy xu hướng doanh thu.

Ứng dụng:

* Cung cấp cái nhìn tổng quan về dữ liệu trong EDA.
* Hỗ trợ báo cáo và trình bày dữ liệu cho các đối tượng không chuyên.
* Dùng để kiểm tra chất lượng dữ liệu (phát hiện giá trị thiếu, ngoại lai).

**Thống kê suy luận (Inferential Statistics) là gì?**

Thống kê suy luận là tập hợp các phương pháp dùng để đưa ra kết luận hoặc dự đoán về một tổng thể (population) dựa trên dữ liệu từ một mẫu (sample). Nó vượt ra ngoài việc mô tả dữ liệu hiện có để suy ra các đặc tính hoặc kiểm tra giả thuyết về tổng thể, thường kèm theo mức độ không chắc chắn (uncertainty) được biểu thị qua xác suất.

Các thành phần chính của thống kê suy luận:

1. Ước lượng (Estimation):
   * Ước lượng điểm (Point Estimation): Sử dụng một giá trị duy nhất để ước lượng tham số tổng thể, ví dụ: dùng trung bình mẫu để ước lượng trung bình tổng thể.
   * Ước lượng khoảng (Interval Estimation): Cung cấp một khoảng giá trị (khoảng tin cậy - confidence interval) mà tham số tổng thể có khả năng nằm trong đó, ví dụ: “Trung bình doanh thu tổng thể nằm trong khoảng 500 ± 50 triệu với độ tin cậy 95%.”
2. Kiểm định giả thuyết (Hypothesis Testing):
   * Kiểm tra xem một giả thuyết về tổng thể có đúng không, ví dụ: “Liệu trung bình doanh thu của cửa hàng có lớn hơn 600 triệu không?”
   * Sử dụng p-value hoặc các bài kiểm định như t-test, chi-square test để đánh giá.
3. Dự đoán (Prediction):
   * Sử dụng mô hình (như hồi quy tuyến tính, logistic regression) để dự đoán giá trị dựa trên mẫu, ví dụ: dự đoán doanh thu tương lai dựa trên xu hướng hiện tại.
4. Mô hình hóa (Modeling):
   * Xây dựng các mô hình thống kê hoặc máy học để suy ra mối quan hệ giữa các biến, như mối quan hệ giữa chi phí quảng cáo và doanh thu.

Ví dụ:

Tiếp tục với dữ liệu doanh thu [500, 600, 550, 700, 300]:

* Thống kê suy luận: Dựa trên mẫu này, bạn muốn suy ra liệu trung bình doanh thu của tất cả các tháng (tổng thể) có lớn hơn 600 triệu không.
  + Thực hiện t-test để kiểm tra giả thuyết.
  + Tính khoảng tin cậy 95% cho trung bình tổng thể.
  + Xây dựng mô hình hồi quy để dự đoán doanh thu dựa trên các yếu tố như mùa vụ.

Ứng dụng:

* Đưa ra kết luận về tổng thể từ dữ liệu mẫu, ví dụ: đánh giá hiệu quả của một chiến dịch quảng cáo.
* Kiểm tra giả thuyết, như so sánh doanh thu giữa hai khu vực.
* Dự đoán xu hướng hoặc xây dựng mô hình trong phân tích kinh doanh, khoa học, hoặc y học.

**Sự khác biệt giữa Thống kê mô tả và Thống kê suy luận:**

|  |  |  |
| --- | --- | --- |
| Tiêu chí | Thống kê mô tả | Thống kê suy luận |
| Mục tiêu | Tóm tắt và mô tả dữ liệu hiện có. | Suy ra đặc tính của tổng thể từ mẫu. |
| Phạm vi | Chỉ áp dụng cho dữ liệu đã thu thập. | Mở rộng kết luận ra tổng thể chưa biết. |
| Ví dụ công cụ | Mean, median, variance, histogram, boxplot. | t-test, ANOVA, confidence interval, regression. |
| Độ không chắc chắn | Không liên quan đến xác suất. | Dựa trên xác suất để đánh giá độ tin cậy. |
| Trực quan hoá | Sử dụng biểu đồ để tóm tắt (bar chart, histogram). | Ít tập trung vào trực quan, thường dùng để kiểm tra mô hình. |
| Ví dụ câu hỏi | “Doanh thu trung bình tháng này là bao nhiêu?” | “Liệu doanh thu trung bình có vượt 600 triệu không?” |
| Ứng dụng trong EDA | Cung cấp cái nhìn ban đầu về dữ liệu. | Kiểm tra giả thuyết hoặc dự đoán sau EDA. |

**Minh họa qua ví dụ thực tế:**

Giả sử bạn có dữ liệu doanh thu hàng tháng của một cửa hàng trong 6 tháng: [500, 600, 550, 700, 300, 650].

* **Thống kê mô tả**:
  + Tính: Mean = 550, Median = 575, Standard Deviation ≈ 139.3.
  + Vẽ histogram để thấy phân bố doanh thu.
  + Báo cáo: “Doanh thu trung bình là 550 triệu, với phần lớn dao động từ 300 đến 700 triệu.”
  + Trực quan hoá:

import matplotlib.pyplot as plt

data = [500, 600, 550, 700, 300, 650]

plt.hist(data, bins=5, color='#1f77b4', edgecolor='black')

plt.title('Phân Bố Doanh Thu Hàng Tháng')

plt.xlabel('Doanh Thu (triệu VND)')

plt.ylabel('Tần Suất')

plt.savefig('doanh\_thu\_histogram.png', dpi=300, bbox\_inches='tight')

plt.show()

* **Thống kê suy luận:**
  + Kiểm tra giả thuyết: “Liệu trung bình doanh thu tổng thể có lớn hơn 600 triệu không?” (sử dụng t-test).
  + Tính khoảng tin cậy 95% cho trung bình tổng thể.
  + Dự đoán doanh thu tháng tiếp theo dựa trên mô hình hồi quy.
  + Ví dụ code (t-test):

from scipy.stats import ttest\_1samp

data = [500, 600, 550, 700, 300, 650]

t\_stat, p\_value = ttest\_1samp(data, popmean=600)

print(f"p-value: {p\_value:.4f}")

# Nếu p-value < 0.05, bác bỏ giả thuyết rằng trung bình = 600 triệu.

**Kết luận:**

* **Thống kê mô tả** giúp bạn hiểu dữ liệu hiện có thông qua các số liệu tóm tắt và trực quan hóa, là bước đầu tiên trong EDA.
* **Thống kê suy luận** cho phép đưa ra kết luận về tổng thể hoặc dự đoán, thường được thực hiện sau khi đã hiểu dữ liệu qua thống kê mô tả.
* Trong thực tế, hai phương pháp này thường được sử dụng kết hợp: thống kê mô tả để khám phá và làm sạch dữ liệu, sau đó thống kê suy luận để kiểm tra giả thuyết hoặc xây dựng mô hình.

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

**1. Các thước đo thống kê mô tả chính và ý nghĩa**

**Trung bình (Mean)**

* Là giá trị “đại diện” cho dữ liệu, được tính bằng tổng các giá trị chia cho số lượng quan sát.
* Dùng để mô tả xu hướng trung tâm khi dữ liệu phân bố **tương đối đều** và không có giá trị ngoại lai (outlier).  
  Ví dụ: Điểm trung bình một lớp = tổng điểm / số sinh viên.

**Trung vị (Median)**

* Là giá trị nằm giữa tập dữ liệu khi đã sắp xếp theo thứ tự.
* Chia dữ liệu thành hai nửa bằng nhau (50% nhỏ hơn, 50% lớn hơn).
* Không bị ảnh hưởng bởi các giá trị ngoại lai.  
  Ví dụ: Nếu lương của 7 nhân viên là [5, 6, 7, 8, 100, 110, 120], thì **trung bình = 51,7** nhưng **trung vị = 8** → trung vị phản ánh thực tế “mức lương điển hình” hơn.

**Phương sai (Variance)**

* Đo mức độ **phân tán** của dữ liệu quanh giá trị trung bình.
* Công thức: s2=∑(xi−xˉ)2n−1s^2 = \frac{\sum (x\_i - \bar{x})^2}{n-1}s2=n−1∑(xi​−xˉ)2​ (với mẫu).
* Giá trị càng lớn → dữ liệu càng “loãng” so với trung bình.

**Độ lệch chuẩn (Standard Deviation – SD)**

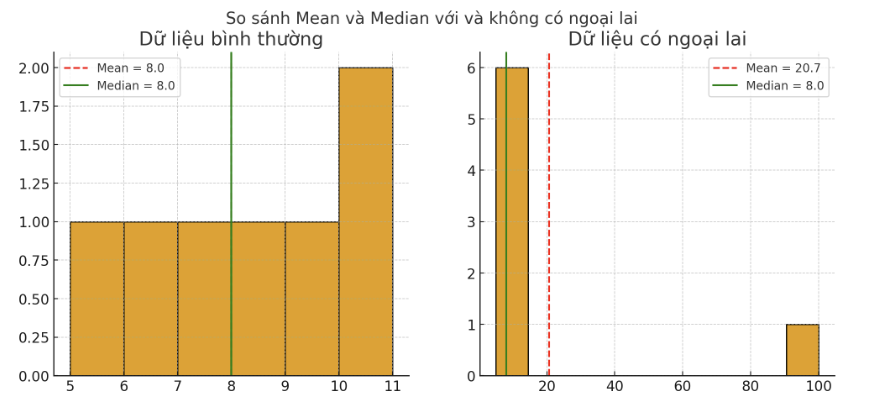
* Là căn bậc hai của phương sai, có cùng đơn vị với dữ liệu.
* Dễ diễn giải hơn phương sai.
* Cho biết mức độ dao động trung bình của các giá trị quanh trung bình.  
  Ví dụ: Chiều cao trung bình = 168 cm, SD = 5 cm → đa số sinh viên cao từ **163–173 cm**.

**2. Khi nào nên dùng trung vị thay vì trung bình?**

* Khi dữ liệu **bị lệch (skewed distribution)**.  
  Ví dụ: Thu nhập thường có vài người rất giàu → trung bình bị kéo cao, trung vị phản ánh “người điển hình” hơn.
* Khi có **giá trị ngoại lai (outliers)**.  
  Ví dụ: Điểm thi của cả lớp 40 sinh viên là 7–8 điểm, chỉ có 1 bạn được 0 → trung bình bị kéo xuống, trung vị vẫn giữ nguyên.
* Khi muốn mô tả **vị trí trung tâm “ổn định”** của dữ liệu thay vì “giá trị đại diện nhạy cảm” như trung bình.

Tóm lại:

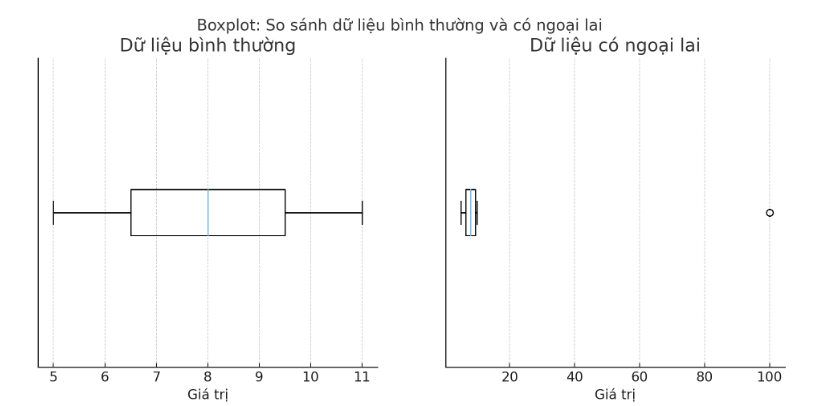
* **Trung bình** tốt khi dữ liệu “đẹp”, phân bố chuẩn hoặc gần chuẩn.
* **Trung vị** tốt khi dữ liệu lệch, có ngoại lai hoặc khi muốn mô tả “điển hình”.



Kết quả: Khi có ngoại lai (giá trị 100), **trung bình tăng vọt từ 8 → 20.7**, trong khi **trung vị vẫn giữ nguyên = 8** → minh chứng rằng trung vị ổn định hơn khi dữ liệu bị lệch.

* Ở **dữ liệu bình thường**, Mean (đường đỏ) và Median (đường xanh) gần như trùng nhau.
* Ở **dữ liệu có ngoại lai**, Mean bị kéo lệch sang phải (giá trị cao), trong khi Median vẫn giữ nguyên vị trí trung tâm.

Đây chính là lý do tại sao khi dữ liệu có ngoại lai hoặc bị lệch, ta nên dùng **Median** thay vì **Mean**.



* **Dữ liệu bình thường**: hộp (IQR) nằm gọn, không có điểm ngoại lai.
* **Dữ liệu có ngoại lai**: xuất hiện điểm nằm ngoài hộp (giá trị 100) → ngoại lai, làm kéo dài đường whisker.

Boxplot cho thấy Median (đường ngang trong hộp) vẫn ổn định, còn Mean thì dễ bị lệch bởi ngoại lai.

**+ Làm thế nào để xác định phân bố của một tập dữ liệu? Các loại phân bố phổ biến là gì (ví dụ: phân bố chuẩn, lệch trái, lệch phải)?**

**1. Cách xác định phân bố của tập dữ liệu**

Để xác định phân bố của một tập dữ liệu, bạn có thể sử dụng các phương pháp sau:

**a. Trực quan hóa dữ liệu**

Trực quan hóa là cách trực tiếp và hiệu quả để nhận diện hình dạng phân bố của dữ liệu. Các công cụ phổ biến bao gồm:

* **Biểu đồ Histogram**:
  + Vẽ histogram để thấy tần suất xuất hiện của các giá trị trong các khoảng (bins).
  + Hình dạng của histogram giúp nhận biết phân bố (chuẩn, lệch trái, lệch phải, đa đỉnh, v.v.).
* **Biểu đồ hộp (Boxplot)**:
  + Hiển thị trung vị (Q2), tứ phân vị (Q1, Q3), và giá trị ngoại lai.
  + Boxplot giúp xác định độ lệch (skewness) và sự phân tán của dữ liệu.
  + Nếu trung vị gần Q1 hơn Q3, dữ liệu có thể lệch phải; nếu gần Q3 hơn Q1, dữ liệu có thể lệch trái.
* **Biểu đồ mật độ (Density Plot)**:
  + Tương tự histogram nhưng mượt mà hơn, giúp dễ dàng nhận ra hình dạng phân bố.
* **Biểu đồ Q-Q (Quantile-Quantile Plot)**:
  + So sánh các phân vị của dữ liệu với phân vị của phân bố chuẩn lý thuyết.
  + Nếu các điểm nằm gần đường thẳng, dữ liệu có khả năng theo phân bố chuẩn.

**b. Tính các chỉ số thống kê mô tả**

Các chỉ số thống kê giúp định lượng đặc điểm phân bố:

* **Trung bình (Mean)**, **Trung vị (Median)**, **Mode**:
  + Nếu **trung bình ≈ trung vị ≈ mode**, dữ liệu có thể theo phân bố chuẩn hoặc đối xứng.
  + Nếu **trung bình > trung vị**, dữ liệu có khả năng **lệch phải** (positive skew).
  + Nếu **trung bình < trung vị**, dữ liệu có khả năng **lệch trái** (negative skew).
* Độ lệch (Skewness):



* + Công thức: Skewness=
  + Skewness ≈ 0: Phân bố đối xứng (gần chuẩn).
  + Skewness > 0: Lệch phải.
  + Skewness < 0: Lệch trái.
* **Độ nhọn (Kurtosis)**:
  + Đo lường độ “nhọn” hoặc “phẳng” của phân bố so với phân bố chuẩn.
  + Kurtosis = 3 (hoặc 0 khi chuẩn hóa): Phân bố chuẩn.
  + Kurtosis > 3: Phân bố nhọn hơn (leptokurtic).
  + Kurtosis < 3: Phân bố phẳng hơn (platykurtic).
* **Phạm vi (Range)** và **Khoảng tứ phân vị (IQR)**:
  + Cung cấp thông tin về độ phân tán, hỗ trợ đánh giá mức độ lan tỏa của dữ liệu.

**c. Kiểm định thống kê**

Các kiểm định thống kê giúp xác định xem dữ liệu có phù hợp với một phân bố cụ thể (thường là phân bố chuẩn) hay không:

* **Kiểm định Shapiro-Wilk**: Kiểm tra xem dữ liệu có theo phân bố chuẩn hay không (p-value < 0.05 bác bỏ giả thuyết phân bố chuẩn).
* **Kiểm định Kolmogorov-Smirnov**: So sánh phân bố dữ liệu với phân bố lý thuyết.
* **Kiểm định Anderson-Darling**: Nhạy hơn với phân bố chuẩn, đặc biệt ở đuôi phân bố.
* **Kiểm định D’Agostino’s K²**: Kết hợp skewness và kurtosis để kiểm tra phân bố chuẩn.

**d. Phân tích đặc điểm của dữ liệu**

* **Tính chất của biến**: Xác định biến là liên tục (continuous) hay rời rạc (discrete), vì điều này ảnh hưởng đến loại phân bố có thể xảy ra.
* **Bối cảnh dữ liệu**: Hiểu ngữ cảnh thu thập dữ liệu (ví dụ, thu nhập thường lệch phải, thời gian chờ thường lệch phải, điểm số có thể gần chuẩn).

**2. Các loại phân bố phổ biến**

Dưới đây là các loại phân bố phổ biến, đặc biệt là phân bố chuẩn, lệch trái, và lệch phải, cùng với đặc điểm và ví dụ:

**a. Phân bố chuẩn (Normal Distribution)**

* **Đặc điểm**:
  + Hình chuông, đối xứng quanh trung bình.
  + Trung bình = trung vị = mode.
  + Skewness ≈ 0, Kurtosis ≈ 3 (hoặc 0 khi chuẩn hóa).
  + Khoảng 68% dữ liệu nằm trong khoảng ±1 độ lệch chuẩn, 95% trong ±2 độ lệch chuẩn, 99.7% trong ±3 độ lệch chuẩn.
* **Ví dụ**:
  + Chiều cao, cân nặng của một quần thể đồng nhất.
  + Điểm số trong các bài kiểm tra chuẩn hóa (như IQ).
* **Ứng dụng**:
  + Phân bố chuẩn là nền tảng cho nhiều kiểm định thống kê (như t-test, ANOVA).
  + Dữ liệu gần chuẩn thường dễ phân tích và mô hình hóa.

**b. Phân bố lệch phải (Positively Skewed Distribution)**

* **Đặc điểm**:
  + Đuôi phân bố kéo dài về phía bên phải (giá trị lớn).
  + Trung bình > trung vị > mode.
  + Skewness > 0.
  + Thường xuất hiện trong dữ liệu có giới hạn dưới (như 0) nhưng không có giới hạn trên.
* **Ví dụ**:
  + Thu nhập cá nhân (nhiều người có thu nhập thấp, ít người có thu nhập rất cao).
  + Thời gian chờ (như thời gian chờ tại bệnh viện).
  + Thời gian sự cố (thời gian đến khi một thiết bị hỏng).
* **Ứng dụng**:
  + Có thể cần biến đổi dữ liệu (như log hoặc căn bậc hai) để đưa về gần phân bố chuẩn trước khi phân tích.

**c. Phân bố lệch trái (Negatively Skewed Distribution)**

* **Đặc điểm**:
  + Đuôi phân bố kéo dài về phía bên trái (giá trị nhỏ).
  + Trung bình < trung vị < mode.
  + Skewness < 0.
  + Thường xuất hiện trong dữ liệu có giới hạn trên nhưng không có giới hạn dưới.
* **Ví dụ**:
  + Điểm số trong các bài kiểm tra dễ (nhiều người đạt điểm cao, ít người điểm thấp).
  + Tuổi nghỉ hưu (nhiều người nghỉ hưu ở tuổi cao, ít người nghỉ sớm).
* **Ứng dụng**:
  + Tương tự lệch phải, có thể cần biến đổi để phân tích.

**d. Phân bố đều (Uniform Distribution)**

* **Đặc điểm**:
  + Tất cả giá trị trong một khoảng có xác suất xảy ra như nhau.
  + Không có đỉnh rõ ràng, histogram gần giống hình chữ nhật.
  + Skewness ≈ 0 (đối xứng).
* **Ví dụ**:
  + Kết quả của một lần tung xúc xắc (mỗi mặt từ 1 đến 6 có xác suất bằng nhau).
  + Số ngẫu nhiên được tạo bởi máy tính trong một khoảng cố định.
* **Ứng dụng**:
  + Thường dùng trong mô phỏng hoặc thử nghiệm ngẫu nhiên.

**e. Phân bố đa đỉnh (Bimodal/Multimodal Distribution)**

* **Đặc điểm**:
  + Có hai hoặc nhiều đỉnh (mode) trong histogram.
  + Có thể chỉ ra sự tồn tại của nhiều nhóm hoặc cụm trong dữ liệu.
* **Ví dụ**:
  + Chiều cao của một quần thể gồm cả nam và nữ (hai đỉnh tương ứng với hai giới).
  + Điểm số của một bài kiểm tra có hai nhóm học sinh (giỏi và yếu).
* **Ứng dụng**:
  + Gợi ý cần phân tích dữ liệu theo các nhóm riêng biệt.

**f. Phân bố lũy thừa (Power Law Distribution)**

* **Đặc điểm**:
  + Đuôi dài, thường lệch phải mạnh.
  + Một số ít giá trị rất lớn, phần lớn giá trị nhỏ.
* **Ví dụ**:
  + Quy mô dân số của các thành phố (vài thành phố lớn, nhiều thành phố nhỏ).
  + Lượt truy cập các trang web (vài trang rất phổ biến, nhiều trang ít truy cập).
* **Ứng dụng**:
  + Thường gặp trong các lĩnh vực như kinh tế, mạng xã hội, hoặc khoa học mạng.

**3. Ví dụ minh họa**

Giả sử tập dữ liệu: [1, 2, 2, 3, 3, 4, 4, 4, 5, 10]

**Bước 1: Trực quan hóa**

**Dữ liệu**

Tập dữ liệu: [1, 2, 2, 3, 3, 4, 4, 4, 5, 10]

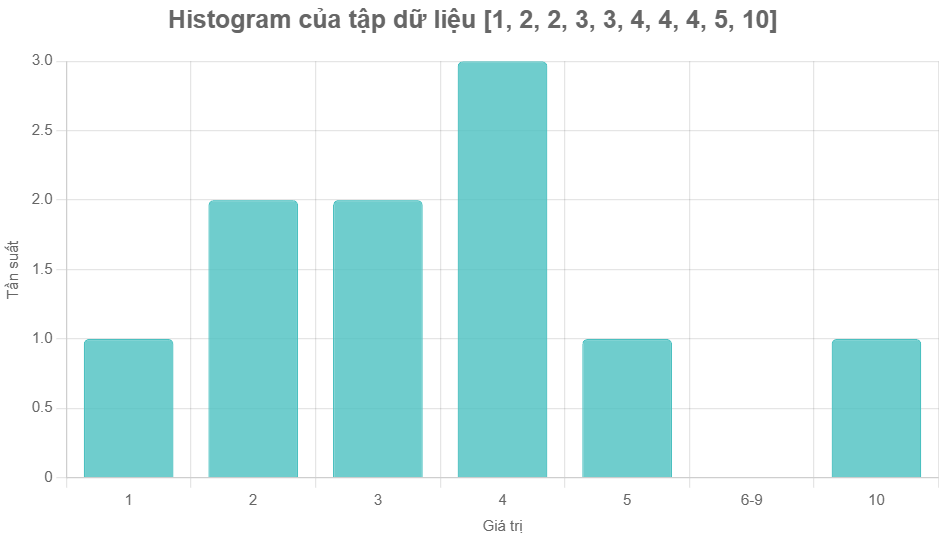
* **Histogram**: Hiển thị tần suất xuất hiện của các giá trị, giúp nhận diện hình dạng phân bố (lệch phải, lệch trái, hay đối xứng).
* **Boxplot**: Hiển thị trung vị, tứ phân vị, và giá trị ngoại lai, giúp xác định độ lệch và phân tán.

**a. Histogram**

Histogram sẽ đếm tần suất các giá trị trong các khoảng (bins). Vì dữ liệu có các giá trị từ 1 đến 10, tôi sẽ chia thành các khoảng hợp lý (ví dụ, các khoảng 1-2, 2-3, ..., 9-10) để thể hiện tần suất.

Dựa trên tập dữ liệu:

* Giá trị 1: 1 lần
* Giá trị 2: 2 lần
* Giá trị 3: 2 lần
* Giá trị 4: 3 lần
* Giá trị 5: 1 lần
* Giá trị 10: 1 lần



**Giải thích Histogram**:

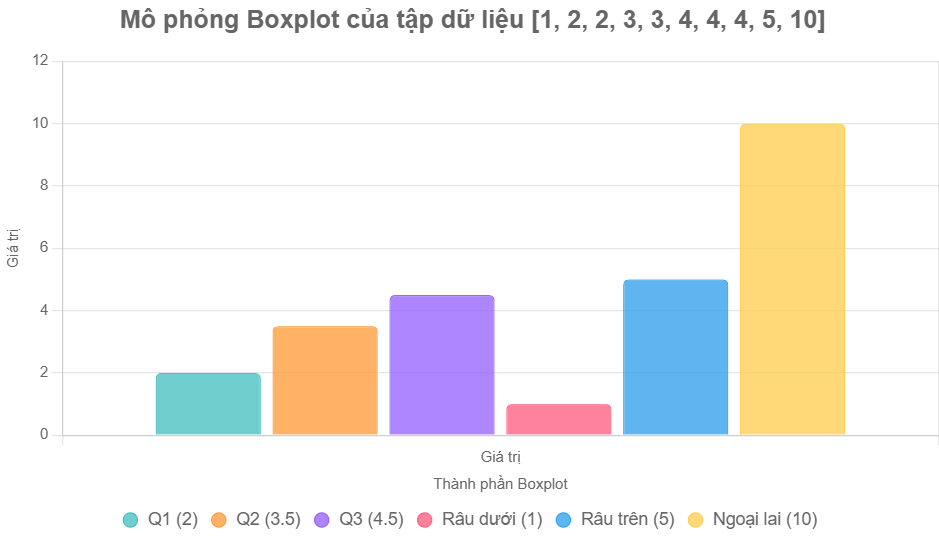
* Biểu đồ cho thấy tần suất cao nhất ở giá trị 4 (3 lần), và có một đuôi dài bên phải với giá trị 10 (1 lần).
* Đuôi dài bên phải (giá trị 10) và sự tập trung ở các giá trị nhỏ hơn (1-5) cho thấy phân bố **lệch phải** (positive skew).

**b. Boxplot**

Để vẽ boxplot, chúng ta cần tính các thành phần:

* **Tập dữ liệu đã sắp xếp**: [1, 2, 2, 3, 3, 4, 4, 4, 5, 10]
* **Q1 (tứ phân vị thứ nhất)**: Vị trí 25% (khoảng vị trí 2.5) → Trung bình của 2 và 2 = **2**.
* **Q2 (trung vị)**: Vị trí 50% (vị trí 5.5) → Trung bình của 3 và 4 = **3.5**.
* **Q3 (tứ phân vị thứ ba)**: Vị trí 75% (khoảng vị trí 7.5) → Trung bình của 4 và 5 = **4.5**.
* **IQR (khoảng tứ phân vị)**: Q3−Q1=4.5−2=2.5 Q3 - Q1 = 4.5 - 2 = 2.5 Q3−Q1=4.5−2=2.5.
* **Râu (Whiskers)**:
  + Râu dưới: Q1−1.5×IQR=2−1.5×2.5=2−3.75=−1.75 Q1 - 1.5 \times IQR = 2 - 1.5 \times 2.5 = 2 - 3.75 = -1.75 Q1−1.5×IQR=2−1.5×2.5=2−3.75=−1.75 (nhưng giới hạn bởi giá trị nhỏ nhất = 1).
  + Râu trên: Q3+1.5×IQR=4.5+1.5×2.5=4.5+3.75=8.25 Q3 + 1.5 \times IQR = 4.5 + 1.5 \times 2.5 = 4.5 + 3.75 = 8.25 Q3+1.5×IQR=4.5+1.5×2.5=4.5+3.75=8.25 (giá trị tối đa trong phạm vi = 5, vì 10 là ngoại lai).
* **Giá trị ngoại lai**: Giá trị 10 > 8.25, nên 10 là ngoại lai.

Boxplot không được hỗ trợ trực tiếp trong Chart.js, nhưng tôi sẽ mô phỏng bằng cách hiển thị các thành phần chính (Q1, Q2, Q3, râu, và ngoại lai) dưới dạng một biểu đồ cột đơn giản với chú thích.



**Giải thích Boxplot**:

* **Hộp**: Từ Q1 (2) đến Q3 (4.5), chứa 50% dữ liệu ở giữa, với trung vị Q2 (3.5).
* **Râu**: Từ giá trị nhỏ nhất (1) đến giá trị lớn nhất trong phạm vi không ngoại lai (5).
* **Ngoại lai**: Giá trị 10 nằm ngoài Q3+1.5×IQR=8.25 Q3 + 1.5 \times IQR = 8.25 Q3+1.5×IQR=8.25, xác nhận đuôi dài bên phải.
* **Nhận xét**: Trung vị (3.5) gần Q1 (2) hơn Q3 (4.5), và sự hiện diện của ngoại lai ở giá trị 10 củng cố rằng phân bố **lệch phải**.

**Bước 2: Phân tích phân bố**

* **Histogram**: Đuôi dài bên phải (giá trị 10) và sự tập trung ở các giá trị nhỏ (1-5) cho thấy phân bố lệch phải.
* **Boxplot**: Trung vị gần Q1 hơn Q3, và có ngoại lai ở giá trị cao (10), xác nhận phân bố lệch phải.
* **Chỉ số thống kê** (từ ví dụ trước):
  + Trung bình (3.8) > Trung vị (3.5) > Mode (4), skewness > 0, phù hợp với phân bố lệch phải.
* **Kết luận**: Tập dữ liệu có phân bố **lệch phải** (positively skewed), với đuôi dài ở các giá trị lớn và sự tập trung ở các giá trị nhỏ.

**4. Lưu ý khi xác định phân bố**

* **Kích thước mẫu**: Mẫu nhỏ có thể không phản ánh đúng phân bố thực tế.
* **Giá trị ngoại lai**: Có thể làm sai lệch hình dạng phân bố, cần xử lý trước (xem phương pháp xử lý giá trị thiếu hoặc ngoại lai).
* **Biến đổi dữ liệu**: Nếu dữ liệu lệch mạnh, có thể áp dụng biến đổi (log, căn bậc hai) để đưa về gần phân bố chuẩn.
* **Ngữ cảnh**: Phân bố phụ thuộc vào đặc điểm của biến (ví dụ, thu nhập thường lệch phải, điểm số có thể gần chuẩn).

**5. Kết luận**

* **Xác định phân bố** cần kết hợp trực quan hóa (histogram, boxplot, Q-Q plot) và các chỉ số thống kê (trung bình, trung vị, skewness, kurtosis, kiểm định).
* **Phân bố phổ biến**:
  + **Phân bố chuẩn**: Đối xứng, hình chuông, trung bình = trung vị = mode.
  + **Lệch phải**: Đuôi dài bên phải, trung bình > trung vị.
  + **Lệch trái**: Đuôi dài bên trái, trung bình < trung vị.
  + **Phân bố đều**: Xác suất đồng đều.
  + **Đa đỉnh**: Nhiều đỉnh, gợi ý nhiều nhóm dữ liệu.
  + **Lũy thừa**: Đuôi dài mạnh, ít giá trị lớn, nhiều giá trị nhỏ.
* Hiểu rõ phân bố giúp lựa chọn phương pháp phân tích phù hợp (ví dụ, kiểm định tham số cho phân bố chuẩn, kiểm định phi tham số cho phân bố không chuẩn).

**+ 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 (range) là hai thước đo quan trọng để đánh giá sự phân tán của dữ liệu trong thống kê, mỗi cái cung cấp một góc nhìn khác nhau về mức độ biến động của dữ liệu:

1. **Độ lệch chuẩn (Standard Deviation)**:
   * **Ý nghĩa**: Độ lệch chuẩn đo lường mức độ mà các giá trị trong tập dữ liệu lệch khỏi giá trị trung bình (mean). Nó cho biết dữ liệu phân tán chặt chẽ hay rộng rãi quanh giá trị trung bình.
   * **Cách hiểu**:
     + Nếu độ lệch chuẩn nhỏ, dữ liệu tập trung gần giá trị trung bình, nghĩa là sự phân tán thấp.
     + Nếu độ lệch chuẩn lớn, dữ liệu phân tán xa giá trị trung bình, nghĩa là sự phân tán cao.
   * **Ưu điểm**:
     + Phản ánh chính xác hơn sự phân tán vì nó xem xét tất cả các giá trị trong tập dữ liệu, không chỉ giá trị cực đại hay cực tiểu.
     + Nhạy với các biến động nhỏ trong dữ liệu.
   * **Ứng dụng**: Được sử dụng rộng rãi trong phân tích thống kê, tài chính, khoa học, v.v., đặc biệt khi cần so sánh mức độ biến động giữa các tập dữ liệu.
2. **Phạm vi (Range)**:
   * **Ý nghĩa**: Phạm vi là hiệu số giữa giá trị lớn nhất (max) và giá trị nhỏ nhất (min) trong tập dữ liệu. Nó cho biết khoảng cách từ giá trị thấp nhất đến cao nhất.
   * **Cách hiểu**:
     + Phạm vi lớn cho thấy dữ liệu có sự khác biệt lớn giữa các giá trị cực đại và cực tiểu.
     + Phạm vi nhỏ cho thấy các giá trị trong tập dữ liệu khá gần nhau.
   * **Ưu điểm**:
     + Dễ tính toán và dễ hiểu.
     + Cung cấp cái nhìn nhanh về mức độ lan tỏa của dữ liệu.
   * **Hạn chế**:
     + Chỉ dựa vào hai giá trị (max và min), nên không phản ánh được sự phân bố của các giá trị khác trong tập dữ liệu.
     + Nhạy cảm với các giá trị ngoại lai (outliers), có thể làm sai lệch bức tranh về sự phân tán.

**So sánh và ý nghĩa trong đánh giá sự phân tán**:

* **Độ lệch chuẩn** cung cấp thông tin chi tiết hơn về sự phân tán vì nó xem xét toàn bộ tập dữ liệu, trong khi **phạm vi** chỉ cho một cái nhìn tổng quát và đơn giản.
* Ví dụ: Nếu hai tập dữ liệu có cùng phạm vi nhưng độ lệch chuẩn khác nhau, tập dữ liệu có độ lệch chuẩn nhỏ hơn sẽ có các giá trị tập trung hơn quanh trung bình.
* Trong thực tế, độ lệch chuẩn thường được ưu tiên hơn phạm vi khi cần phân tích kỹ lưỡng, nhưng phạm vi hữu ích khi cần đánh giá nhanh hoặc khi dữ liệu có ít giá trị ngoại lai.

Tóm lại, cả hai thước đo đều quan trọng, nhưng độ lệch chuẩn thường được sử dụng trong các phân tích phức tạp, trong khi phạm vi phù hợp cho các đánh giá sơ bộ hoặc trực quan.

**Minh họa bằng 2 bộ dữ liệu có cùng Range nhưng SD khác nhau để thấy rõ sự khác biệt:**

Tạo hai bộ dữ liệu có cùng phạm vi nhưng độ lệch chuẩn khác nhau, sau đó giải thích và trình bày trực quan để làm rõ sự phân tán.

**Hai bộ dữ liệu**

* **Bộ dữ liệu 1**: [2, 3, 4, 5, 6]
* **Bộ dữ liệu 2**: [2, 2, 4, 6, 6]

Cả hai bộ dữ liệu này có:

* **Phạm vi (Range)**: Giá trị lớn nhất (6) - Giá trị nhỏ nhất (2) = **4**.
* Tuy nhiên, sự phân tán của các giá trị quanh giá trị trung bình sẽ khác nhau, dẫn đến độ lệch chuẩn khác nhau.

**Bước 1: Tính toán các thông số**

**Bộ dữ liệu 1: [2, 3, 4, 5, 6]**

* **Trung bình (Mean)**: (2 + 3 + 4 + 5 + 6) / 5 = 20 / 5 = 4



* Độ lệch chuẩn (SD): Công thức:
  + Các độ lệch bình phương: (2-4)^2 = 4, (3-4)^2 = 1, (4-4)^2 = 0, (5-4)^2 = 1, (6-4)^2 = 4
  + Tổng: 4 + 1 + 0 + 1 + 4 = 10
  + Phương sai: 10/5 = 2
  + Độ lệch chuẩn: sqrt{2} ​≈1.414

**Bộ dữ liệu 2: [2, 2, 4, 6, 6]**

* **Trung bình (Mean)**: (2 + 2 + 4 + 6 + 6) / 5 = 20 / 5 = 4
* **Độ lệch chuẩn (SD)**:
  + Các độ lệch bình phương: (2-4)^2 = 4, (2-4)^2 = 4, (4-4)^2 = 0, (6-4)^2 = 4, (6-4)^2 = 4
  + Tổng: 4 + 4 + 0 + 4 + 4 = 16
  + Phương sai: 16 / 5 = 3.2
  + Độ lệch chuẩn: sqrt{3.2} ​≈1.789

**Kết quả**

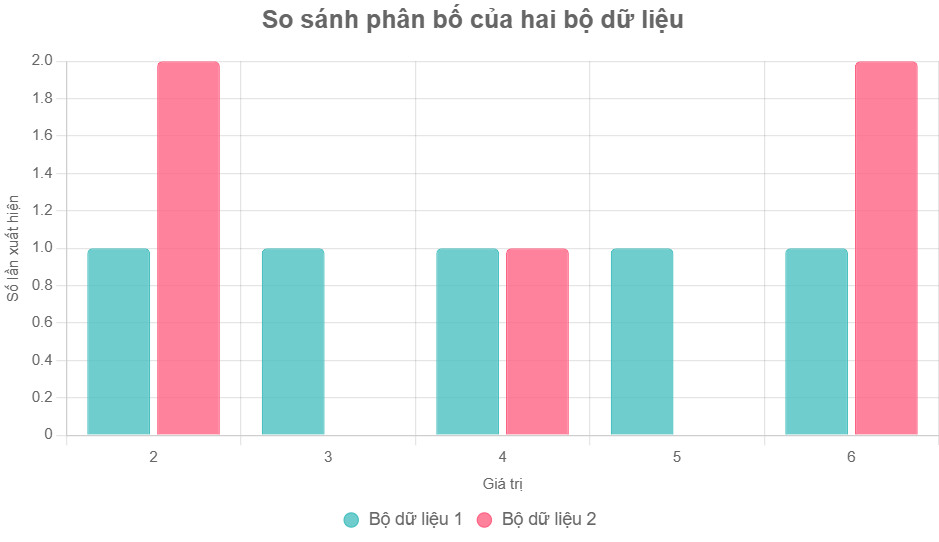
* **Phạm vi (Range)**: Cả hai bộ dữ liệu đều có phạm vi = 4.
* **Độ lệch chuẩn (SD)**:
  + Bộ dữ liệu 1: ≈1.414\approx 1.414≈1.414
  + Bộ dữ liệu 2: ≈1.789\approx 1.789≈1.789

**Phân tích**

* **Bộ dữ liệu 1** ([2, 3, 4, 5, 6]): Các giá trị phân bố đều quanh trung bình (4), khoảng cách giữa các giá trị khá đồng đều (tăng dần từng đơn vị). Điều này dẫn đến độ lệch chuẩn nhỏ hơn, cho thấy dữ liệu ít phân tán hơn.
* **Bộ dữ liệu 2** ([2, 2, 4, 6, 6]): Các giá trị tập trung nhiều hơn ở hai đầu (2 và 6), với ít giá trị gần trung bình (4). Điều này làm tăng độ lệch chuẩn, cho thấy dữ liệu phân tán nhiều hơn, mặc dù phạm vi vẫn bằng 4.

**Minh hoạ trực quan:**

Dưới đây là biểu đồ cột (bar chart) minh họa hai bộ dữ liệu để thấy rõ sự khác biệt trong phân bố giá trị:



**Giải thích biểu đồ**

* **Bộ dữ liệu 1**: Các cột có chiều cao đều (mỗi giá trị xuất hiện 1 lần), thể hiện sự phân bố đồng đều, dẫn đến độ lệch chuẩn nhỏ hơn.
* **Bộ dữ liệu 2**: Các cột cao hơn ở hai đầu (2 và 6 xuất hiện 2 lần), cho thấy dữ liệu tập trung ở các giá trị cực đại và cực tiểu, dẫn đến độ lệch chuẩn lớn hơn.

**Kết luận**

Mặc dù cả hai bộ dữ liệu có cùng phạm vi (4), độ lệch chuẩn của **Bộ dữ liệu 2** lớn hơn, phản ánh sự phân tán cao hơn do các giá trị tập trung ở hai đầu. Điều này cho thấy **độ lệch chuẩn** cung cấp thông tin chi tiết hơn về sự phân tán so với **phạm vi**, vốn chỉ dựa vào giá trị lớn nhất và nhỏ nhất.

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

Trong thống kê, các thước đo **Q1**, **Q2**, **Q3** là các **tứ phân vị** (quartiles) được sử dụng để mô tả sự phân bố của dữ liệu, đặc biệt trong **biểu đồ hộp (boxplot)**. Chúng giúp chia tập dữ liệu thành các phần bằng nhau và cung cấp thông tin về sự phân tán, trung tâm, và hình dạng của dữ liệu. Dưới đây là giải thích chi tiết về sự khác biệt và ý nghĩa của Q1, Q2, Q3 trong biểu đồ hộp:

**1. Định nghĩa các tứ phân vị**

* **Q1 (Tứ phân vị thứ nhất)**:
  + Là giá trị mà 25% dữ liệu nằm dưới nó (hay còn gọi là **tứ phân vị dưới**).
  + Q1 đại diện cho giá trị ở vị trí 25% của tập dữ liệu đã được sắp xếp theo thứ tự tăng dần.
* **Q2 (Tứ phân vị thứ hai)**:
  + Là **trung vị** (median) của tập dữ liệu, tức là giá trị mà 50% dữ liệu nằm dưới và 50% nằm trên.
  + Q2 chia tập dữ liệu thành hai nửa bằng nhau và đại diện cho giá trị trung tâm.
* **Q3 (Tứ phân vị thứ ba)**:
  + Là giá trị mà 75% dữ liệu nằm dưới nó (hay còn gọi là **tứ phân vị trên**).
  + Q3 đại diện cho giá trị ở vị trí 75% của tập dữ liệu đã được sắp xếp.

**2. Vai trò trong biểu đồ hộp (Boxplot)**

Biểu đồ hộp sử dụng Q1, Q2, Q3 cùng với các giá trị khác (như giá trị cực đại, cực tiểu, và giá trị ngoại lai) để biểu diễn sự phân bố của dữ liệu. Các thành phần chính của boxplot liên quan đến Q1, Q2, Q3 bao gồm:

* **Hộp (Box)**: Được tạo bởi Q1 (cạnh dưới) và Q3 (cạnh trên), thể hiện **khoảng tứ phân vị (IQR)**, nơi chứa 50% dữ liệu ở giữa (từ 25% đến 75%).
* **Đường giữa hộp**: Là Q2 (trung vị), biểu thị giá trị trung tâm của tập dữ liệu.
* **Râu (Whiskers)**: Thường kéo dài từ Q1 đến giá trị nhỏ nhất và từ Q3 đến giá trị lớn nhất trong phạm vi không bị coi là ngoại lai (thường tính bằng Q1−1.5×IQR Q1 - 1.5 \times IQR Q1−1.5×IQR và Q3+1.5×IQR Q3 + 1.5 \times IQR Q3+1.5×IQR).
* **Giá trị ngoại lai (Outliers)**: Các điểm dữ liệu nằm ngoài phạm vi của râu, tức là nhỏ hơn Q1−1.5×IQR Q1 - 1.5 \times IQR Q1−1.5×IQR hoặc lớn hơn Q3+1.5×IQR Q3 + 1.5 \times IQR Q3+1.5×IQR.

**3. Sự khác biệt giữa Q1, Q2, Q3**

* **Vị trí trong phân bố**:
  + **Q1**: Đánh dấu ranh giới giữa 25% dữ liệu thấp nhất và 75% còn lại.
  + **Q2**: Là trung tâm của dữ liệu, chia đôi tập dữ liệu.
  + **Q3**: Đánh dấu ranh giới giữa 75% dữ liệu thấp hơn và 25% dữ liệu cao nhất.
* **Ý nghĩa về phân tán**:
  + **Khoảng tứ phân vị (IQR = Q3 - Q1)**: Đo lường mức độ phân tán của 50% dữ liệu ở giữa. IQR lớn hơn cho thấy dữ liệu phân tán rộng hơn.
  + **Q2**: Không chỉ là trung tâm mà còn giúp đánh giá tính đối xứng của phân bố (nếu Q2 gần giữa Q1 và Q3, dữ liệu có thể đối xứng; nếu lệch, dữ liệu có thể bị lệch).
* **Ứng dụng**:
  + Q1 và Q3 giúp xác định phạm vi dữ liệu chính và phát hiện giá trị ngoại lai.
  + Q2 cung cấp thông tin về giá trị trung tâm, ít bị ảnh hưởng bởi các giá trị cực đại hơn trung bình cộng.
  + Cả ba tứ phân vị giúp so sánh sự phân bố giữa các tập dữ liệu khác nhau.

**4. Ví dụ minh họa**

Giả sử một tập dữ liệu: [2, 3, 4, 5, 6, 7, 8, 9, 10]

* Sắp xếp theo thứ tự tăng dần: [2, 3, 4, 5, 6, 7, 8, 9, 10]
* **Q1**: Giá trị tại vị trí 25% (khoảng vị trí 2.5) → Trung bình của 3 và 4 = **3.5**
* **Q2 (trung vị)**: Giá trị tại vị trí 50% (vị trí 5) = **6**
* **Q3**: Giá trị tại vị trí 75% (khoảng vị trí 7.5) → Trung bình của 8 và 9 = **8.5**
* **IQR**: Q3−Q1=8.5−3.5=5 Q3 - Q1 = 8.5 - 3.5 = 5 Q3−Q1=8.5−3.5=5

Trong biểu đồ hộp:

* Hộp kéo dài từ 3.5 (Q1) đến 8.5 (Q3).
* Đường giữa hộp ở 6 (Q2).
* Râu kéo dài từ giá trị nhỏ nhất (2) đến lớn nhất (10), trừ khi có ngoại lai.

**5. Ý nghĩa trong phân tích dữ liệu**

* **Q1, Q2, Q3** cung cấp cái nhìn tổng quan về sự phân bố mà không bị ảnh hưởng nhiều bởi giá trị ngoại lai, không giống như trung bình cộng hay độ lệch chuẩn.
* **Boxplot** sử dụng các tứ phân vị để trực quan hóa sự phân tán, đối xứng, và phát hiện các giá trị bất thường.
* **So sánh tập dữ liệu**: Khi so sánh nhiều tập dữ liệu, Q1, Q2, Q3 trong boxplot giúp dễ dàng nhận ra sự khác biệt về trung tâm, độ phân tán, và hình dạng phân bố.

**Kết luận**

* **Q1**: Đại diện cho 25% dữ liệu thấp nhất, là cạnh dưới của hộp.
* **Q2**: Là trung vị, biểu thị giá trị trung tâm và chia đôi dữ liệu.
* **Q3**: Đại diện cho 75% dữ liệu thấp hơn, là cạnh trên của hộp.
* Cùng nhau, chúng tạo nên biểu đồ hộp, cung cấp một cách trực quan và hiệu quả để đánh giá sự phân bố, phân tán, và các đặc điểm khác của dữ liệu.

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

Xử lý giá trị thiếu (missing values) là một bước quan trọng trước khi tính toán các chỉ số thống kê mô tả (như trung bình, trung vị, độ lệch chuẩn, tứ phân vị, phạm vi, v.v.) để đảm bảo kết quả chính xác và không bị sai lệch. Dưới đây là các phương pháp phổ biến để xử lý giá trị thiếu, cùng với cách áp dụng và ưu/nhược điểm của từng phương pháp:

**1. Hiểu bản chất của giá trị thiếu**

Trước khi xử lý, cần xác định:

* **Loại giá trị thiếu**:
  + **MCAR (Missing Completely at Random)**: Giá trị thiếu không phụ thuộc vào bất kỳ yếu tố nào (ngẫu nhiên hoàn toàn).
  + **MAR (Missing at Random)**: Giá trị thiếu phụ thuộc vào các biến khác trong dữ liệu nhưng không phụ thuộc vào chính giá trị bị thiếu.
  + **MNAR (Missing Not at Random)**: Giá trị thiếu liên quan trực tiếp đến giá trị của chính nó.
* **Tỷ lệ giá trị thiếu**: Nếu tỷ lệ thiếu nhỏ (ví dụ <5%), xử lý đơn giản có thể đủ. Nếu tỷ lệ lớn, cần phương pháp phức tạp hơn.
* **Mục đích phân tích**: Tùy thuộc vào mục tiêu (ví dụ, tính trung bình hay xây dựng mô hình), phương pháp xử lý sẽ khác nhau.

**2. Các phương pháp xử lý giá trị thiếu**

Dưới đây là các phương pháp phổ biến:

**a. Loại bỏ giá trị thiếu (Deletion)**

* **Cách thực hiện**:
  + **Listwise deletion (Loại bỏ toàn bộ hàng)**: Xóa toàn bộ hàng (bản ghi) chứa giá trị thiếu.
  + **Pairwise deletion**: Chỉ loại bỏ giá trị thiếu trong các tính toán cụ thể, giữ lại dữ liệu cho các phép tính khác.
* **Khi sử dụng**:
  + Phù hợp khi tỷ lệ giá trị thiếu nhỏ và dữ liệu bị thiếu là MCAR.
  + Ví dụ: Nếu một tập dữ liệu có 100 hàng và chỉ 2 hàng thiếu giá trị, loại bỏ 2 hàng này thường không ảnh hưởng lớn.
* **Ưu điểm**:
  + Đơn giản, dễ thực hiện.
  + Không cần giả định về dữ liệu.
* **Nhược điểm**:
  + Làm giảm kích thước tập dữ liệu, có thể dẫn đến mất thông tin quan trọng.
  + Không phù hợp nếu dữ liệu thiếu không ngẫu nhiên (MAR hoặc MNAR) hoặc tỷ lệ thiếu lớn.

**b. Thay thế bằng giá trị thống kê (Imputation with Statistical Values)**

* **Cách thực hiện**:
  + **Thay bằng trung bình (Mean Imputation)**: Thay giá trị thiếu bằng trung bình của cột.
  + **Thay bằng trung vị (Median Imputation)**: Thay giá trị thiếu bằng trung vị, phù hợp với dữ liệu bị lệch (skewed).
  + **Thay bằng mode (Mode Imputation)**: Thay giá trị thiếu bằng giá trị xuất hiện nhiều nhất, thường dùng cho dữ liệu phân loại (categorical).
* **Khi sử dụng**:
  + Phù hợp khi dữ liệu thiếu là MCAR hoặc MAR và tỷ lệ thiếu không quá lớn.
  + Ví dụ: Trong một cột điểm số, thay các giá trị thiếu bằng trung bình hoặc trung vị của cột đó.
* **Ưu điểm**:
  + Giữ nguyên kích thước tập dữ liệu.
  + Đơn giản và nhanh chóng.
* **Nhược điểm**:
  + Làm giảm sự biến thiên của dữ liệu (ví dụ, thay bằng trung bình làm giảm độ lệch chuẩn).
  + Không phản ánh tốt mối quan hệ giữa các biến.
  + Có thể gây sai lệch nếu dữ liệu bị lệch hoặc có giá trị ngoại lai.

**c. Thay thế bằng giá trị dự đoán (Predictive Imputation)**

* **Cách thực hiện**:
  + Sử dụng các thuật toán học máy (như hồi quy tuyến tính, KNN, hoặc Random Forest) để dự đoán giá trị thiếu dựa trên các biến khác trong tập dữ liệu.
  + Ví dụ: Dùng mô hình KNN để tìm các hàng tương tự và thay giá trị thiếu bằng giá trị trung bình của các hàng lân cận.
* **Khi sử dụng**:
  + Phù hợp với dữ liệu MAR, khi có mối quan hệ rõ ràng giữa các biến.
  + Thường dùng khi dữ liệu phức tạp và tỷ lệ giá trị thiếu đáng kể.
* **Ưu điểm**:
  + Chính xác hơn vì xem xét mối quan hệ giữa các biến.
  + Phù hợp với dữ liệu phức tạp.
* **Nhược điểm**:
  + Phức tạp, đòi hỏi thời gian và tài nguyên tính toán.
  + Có thể dẫn đến sai lệch nếu mô hình dự đoán không phù hợp.

**d. Thay thế bằng giá trị cố định (Constant Imputation)**

* **Cách thực hiện**:
  + Thay giá trị thiếu bằng một giá trị cố định, ví dụ 0, một giá trị cụ thể, hoặc một nhãn như “Unknown” cho dữ liệu phân loại.
* **Khi sử dụng**:
  + Phù hợp khi giá trị thiếu mang ý nghĩa cụ thể (ví dụ, “0” có thể biểu thị không có giá trị trong một số ngữ cảnh).
* **Ưu điểm**:
  + Đơn giản và giữ nguyên kích thước dữ liệu.
* **Nhược điểm**:
  + Có thể làm sai lệch phân bố dữ liệu.
  + Chỉ phù hợp trong một số trường hợp cụ thể.

**e. Sử dụng giá trị nội suy (Interpolation)**

* **Cách thực hiện**:
  + Ước lượng giá trị thiếu dựa trên các giá trị lân cận, thường áp dụng cho dữ liệu chuỗi thời gian.
  + Các phương pháp: Nội suy tuyến tính, nội suy đa thức, hoặc nội suy spline.
* **Khi sử dụng**:
  + Phù hợp với dữ liệu có tính liên tục, như chuỗi thời gian (ví dụ, nhiệt độ hàng ngày).
* **Ưu điểm**:
  + Phù hợp với dữ liệu có xu hướng hoặc mẫu rõ ràng.
  + Giữ được tính liên tục của dữ liệu.
* **Nhược điểm**:
  + Không phù hợp với dữ liệu không liên tục hoặc dữ liệu phân loại.
  + Kết quả có thể không chính xác nếu xu hướng không rõ ràng.

**f. Giữ nguyên giá trị thiếu**

* **Cách thực hiện**:
  + Không xử lý giá trị thiếu mà sử dụng các thuật toán thống kê hoặc mô hình học máy có khả năng xử lý giá trị thiếu (như một số cây quyết định hoặc mô hình XGBoost).
* **Khi sử dụng**:
  + Phù hợp khi thuật toán phân tích có thể tự xử lý giá trị thiếu.
  + Hoặc khi việc thay thế giá trị thiếu có thể làm sai lệch dữ liệu.
* **Ưu điểm**:
  + Tránh làm sai lệch dữ liệu do thay thế không chính xác.
* **Nhược điểm**:
  + Không phải tất cả các chỉ số thống kê (như trung bình, độ lệch chuẩn) có thể tính toán trực tiếp với giá trị thiếu.
  + Yêu cầu công cụ/phương pháp hỗ trợ.

**3. Quy trình xử lý giá trị thiếu trước khi tính toán thống kê**

1. **Xác định giá trị thiếu**:
   * Kiểm tra tập dữ liệu để xác định tỷ lệ và mô hình của giá trị thiếu (dùng các công cụ như Python với pandas.isna() hoặc R với is.na()).
2. **Đánh giá tác động**:
   * Xác định xem giá trị thiếu có ảnh hưởng lớn đến chỉ số thống kê hay không.
   * Ví dụ: Trung bình và độ lệch chuẩn nhạy với giá trị thiếu, trong khi trung vị và tứ phân vị ít bị ảnh hưởng hơn.
3. **Chọn phương pháp phù hợp**:
   * Nếu tỷ lệ thiếu nhỏ và ngẫu nhiên (MCAR), có thể dùng loại bỏ hoặc thay bằng trung bình/trung vị.
   * Nếu tỷ lệ thiếu lớn hoặc dữ liệu MAR/MNAR, cân nhắc dự đoán hoặc nội suy.
4. **Kiểm tra sau xử lý**:
   * So sánh phân bố dữ liệu trước và sau khi xử lý (dùng biểu đồ như histogram, boxplot).
   * Đảm bảo các chỉ số thống kê (như Q1, Q2, Q3, độ lệch chuẩn) không bị sai lệch nghiêm trọng.
5. **Ghi nhận phương pháp xử lý**:
   * Ghi lại cách xử lý giá trị thiếu để đảm bảo tính minh bạch và tái lập trong phân tích.

**4. Ví dụ minh họa**

Giả sử tập dữ liệu: [2, 4, NaN, 6, 8]

* **Loại bỏ hàng chứa NaN**:
  + Dữ liệu còn lại: [2, 4, 6, 8]
  + Trung bình: (2+4+6+8)/4=5(2 + 4 + 6 + 8) / 4 = 5(2+4+6+8)/4=5
  + Trung vị: 5
  + Độ lệch chuẩn: ≈2.582\approx 2.582≈2.582
* **Thay bằng trung bình**:
  + Trung bình của [2, 4, 6, 8] = 5, thay NaN bằng 5.
  + Dữ liệu mới: [2, 4, 5, 6, 8]
  + Trung bình: 5
  + Trung vị: 5
  + Độ lệch chuẩn: ≈2.236\approx 2.236≈2.236 (nhỏ hơn do dữ liệu ít phân tán hơn).
* **Thay bằng nội suy tuyến tính** (nếu là chuỗi thời gian):
  + NaN nằm giữa 4 và 6, nội suy tuyến tính: (4+6)/2=5(4 + 6) / 2 = 5(4+6)/2=5
  + Kết quả tương tự như thay bằng trung bình trong trường hợp này.

**5. Lưu ý khi tính toán các chỉ số thống kê**

* **Trung bình và độ lệch chuẩn**: Nhạy cảm với cách xử lý giá trị thiếu, đặc biệt nếu thay bằng trung bình hoặc trung vị, vì có thể làm giảm sự biến thiên.
* **Trung vị, Q1, Q3**: Ít nhạy hơn, nhưng nếu loại bỏ quá nhiều dữ liệu, các giá trị này có thể không đại diện cho tập dữ liệu gốc.
* **Phạm vi (Range)**: Có thể bị ảnh hưởng nếu giá trị thiếu nằm ở các cực trị (max/min).
* **Boxplot**: Xử lý giá trị thiếu trước khi vẽ để đảm bảo Q1, Q2, Q3 và râu được tính toán chính xác.

**6. Kết luận**

* Không có phương pháp xử lý giá trị thiếu nào là tối ưu cho mọi trường hợp. Lựa chọn phương pháp phụ thuộc vào bản chất dữ liệu, tỷ lệ giá trị thiếu, và mục tiêu phân tích.
* **Đề xuất**:
  + Nếu tỷ lệ thiếu nhỏ và ngẫu nhiên, **loại bỏ** hoặc **thay bằng trung bình/trung vị** là đủ.
  + Nếu dữ liệu phức tạp hoặc tỷ lệ thiếu lớn, cân nhắc **dự đoán bằng mô hình** hoặc **nội suy**.
  + Luôn kiểm tra tác động của phương pháp xử lý lên các chỉ số thống kê và phân bố dữ liệu để đảm bảo kết quả đáng tin cậy.

**+ Bạn có thể giải thích cách đọc và diễn giải một biểu đồ histogram hoặc boxplot từ dữ liệu thực tế không?**

**Tập dữ liệu thực tế**

Giả sử chúng ta có điểm số của 50 học sinh trong một bài kiểm tra, với thang điểm từ 0 đến 100:

Dữ liệu: [45, 50, 55, 60, 62, 65, 65, 68, 70, 70, 72, 75, 75, 78, 80, 80, 82, 85, 85, 85, 88, 90, 90, 92, 95, 95, 98, 100, 60, 62, 65, 68, 70, 72, 75, 78, 80, 82, 85, 88, 90, 92, 95, 50, 55, 60, 65, 70, 75, 80, 85]

Dữ liệu này có thể đại diện cho điểm số thực tế trong một lớp học, với một số học sinh đạt điểm cao, một số đạt điểm trung bình, và một số ít điểm thấp.

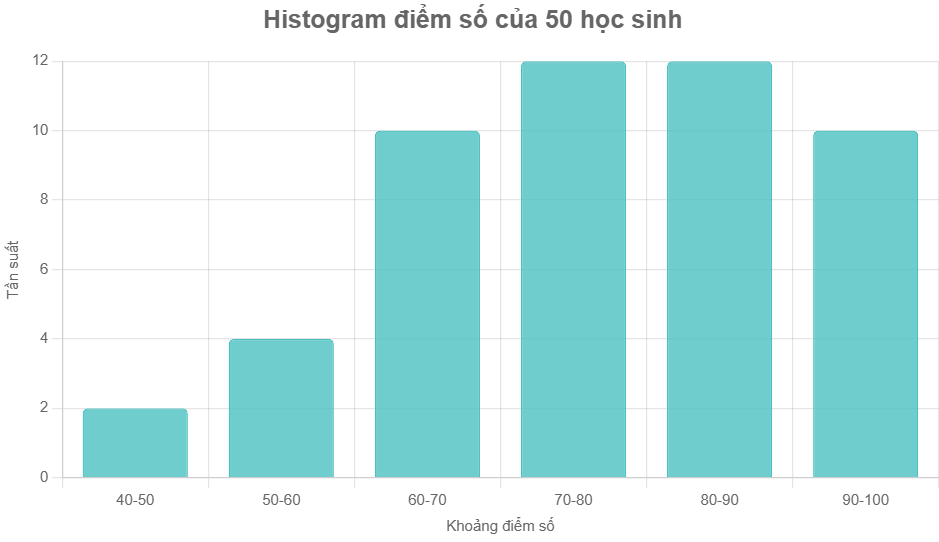
**1. Biểu đồ Histogram**

**a. Vẽ Histogram**

Histogram chia dữ liệu thành các khoảng (bins) và hiển thị tần suất (số lần xuất hiện) của các giá trị trong mỗi khoảng. Tôi sẽ chia dữ liệu thành các khoảng điểm (ví dụ: 40-50, 50-60, ..., 90-100) để vẽ histogram.

Dựa trên dữ liệu, tần suất trong các khoảng như sau:

* 40-50: 2 (45, 50)
* 50-60: 4 (50, 55, 55, 60)
* 60-70: 10 (60, 60, 62, 62, 65, 65, 65, 68, 68, 70)
* 70-80: 12 (70, 70, 72, 72, 75, 75, 75, 78, 78, 80, 80, 80)
* 80-90: 12 (80, 82, 82, 85, 85, 85, 85, 88, 88)
* 90-100: 10 (90, 90, 92, 92, 95, 95, 95, 98, 100)



**b. Cách đọc Histogram**

* **Trục X (ngang)**: Đại diện cho các khoảng giá trị của dữ liệu (điểm số từ 40 đến 100, chia thành các khoảng 10 điểm).
* **Trục Y (dọc)**: Đại diện cho tần suất, tức là số lần xuất hiện của điểm số trong mỗi khoảng.
* **Hình dạng phân bố**:
  + Histogram cho thấy tần suất cao nhất ở khoảng 70-80 và 80-90 (mỗi khoảng có 12 học sinh), tiếp theo là 60-70 và 90-100 (10 học sinh mỗi khoảng).
  + Có ít học sinh ở khoảng thấp (40-50: 2 học sinh) và khoảng 50-60 (4 học sinh).
  + Phân bố hơi **lệch trái** vì có nhiều giá trị tập trung ở phía điểm cao (70-100) và đuôi dài hơn ở phía điểm thấp (40-60).

**c. Diễn giải Histogram**

* **Hình dạng phân bố**: Phân bố không hoàn toàn đối xứng, có xu hướng **lệch trái** nhẹ (negative skew) vì đuôi bên trái (điểm thấp) dài hơn một chút so với bên phải.
* **Tập trung dữ liệu**: Phần lớn học sinh đạt điểm từ 70 đến 90, cho thấy đa số học sinh có kết quả tốt hoặc trung bình khá.
* **Độ phân tán**: Dữ liệu trải dài từ 40 đến 100, nhưng tập trung chủ yếu ở khoảng 60-100, cho thấy bài kiểm tra có thể không quá khó (ít học sinh điểm rất thấp).
* **Bối cảnh thực tế**: Nếu đây là bài kiểm tra dễ, phân bố lệch trái là hợp lý vì nhiều học sinh đạt điểm cao.

**2. Biểu đồ Boxplot**

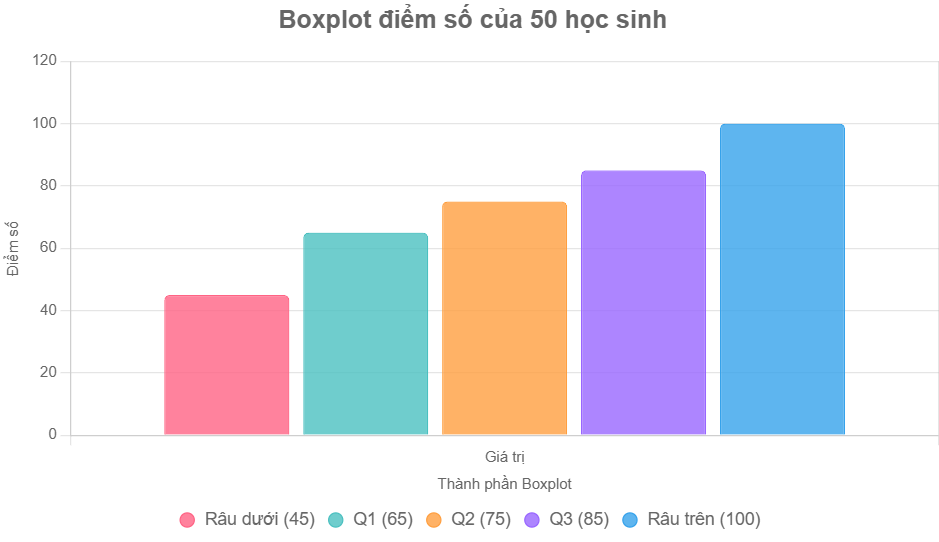
**a. Tính toán các thành phần Boxplot**

Để vẽ boxplot, cần tính các giá trị: Q1, Q2, Q3, IQR, râu, và giá trị ngoại lai (nếu có).

* **Sắp xếp dữ liệu**: [45, 50, 50, 55, 55, 60, 60, 60, 62, 62, 65, 65, 65, 65, 68, 68, 70, 70, 70, 70, 72, 72, 75, 75, 75, 75, 78, 78, 80, 80, 80, 80, 82, 82, 85, 85, 85, 85, 88, 88, 90, 90, 92, 92, 95, 95, 95, 98, 100]
* **Q1 (tứ phân vị thứ nhất)**: Vị trí 25% (khoảng vị trí 12.5) → Trung bình của giá trị thứ 12 và 13 = (65 + 65) / 2 = **65**.
* **Q2 (trung vị)**: Vị trí 50% (khoảng vị trí 25.5) → Trung bình của giá trị thứ 25 và 26 = (75 + 75) / 2 = **75**.
* **Q3 (tứ phân vị thứ ba)**: Vị trí 75% (khoảng vị trí 37.5) → Trung bình của giá trị thứ 37 và 38 = (85 + 85) / 2 = **85**.
* **IQR (khoảng tứ phân vị)**: Q3−Q1=85−65=20 Q3 - Q1 = 85 - 65 = 20 Q3−Q1=85−65=20.
* **Râu**:
  + Râu dưới: Q1−1.5×IQR=65−1.5×20=65−30=35 → Giá trị nhỏ nhất là 45 (> 35), nên râu dưới là **45**.
  + Râu trên: Q3+1.5×IQR=85+1.5×20=85+30=115 → Giá trị lớn nhất là 100 (< 115), nên râu trên là **100**.
* **Giá trị ngoại lai**: Không có giá trị nào nằm ngoài khoảng [35, 115], nên không có ngoại lai.

**b. Vẽ Boxplot**

Vì Chart.js không hỗ trợ boxplot trực tiếp, tôi sẽ mô phỏng bằng cách hiển thị các thành phần chính (Q1, Q2, Q3, râu) dưới dạng biểu đồ cột với chú thích.



**c. Cách đọc Boxplot**

* **Hộp (Box)**: Kéo dài từ Q1 (65) đến Q3 (85), chứa 50% dữ liệu ở giữa (từ 25% đến 75% của điểm số).
* **Đường giữa hộp (Q2)**: Trung vị ở 75, nghĩa là 50% học sinh đạt điểm dưới 75 và 50% đạt điểm trên 75.
* **Râu (Whiskers)**:
  + Râu dưới: Từ 45 (điểm thấp nhất) đến Q1 (65).
  + Râu trên: Từ Q3 (85) đến 100 (điểm cao nhất).
* **Ngoại lai**: Không có giá trị nào nằm ngoài khoảng [35, 115], nên không có ngoại lai.
* **Độ lệch**:
  + Trung vị (75) gần Q3 (85) hơn Q1 (65), cho thấy dữ liệu hơi **lệch trái** (khoảng cách từ Q2 đến Q1 lớn hơn từ Q2 đến Q3).

**d. Diễn giải Boxplot**

* **Trung tâm dữ liệu**: Trung vị là 75, cho thấy điểm số trung bình của lớp nằm ở mức khá tốt.
* **Độ phân tán**: IQR = 20 (từ 65 đến 85) cho thấy 50% học sinh có điểm số nằm trong khoảng này, tương đối tập trung.
* **Độ lệch**: Trung vị gần Q3 hơn Q1, xác nhận phân bố lệch trái nhẹ, với nhiều học sinh đạt điểm cao hơn trung vị.
* **Phạm vi**: Điểm số trải dài từ 45 đến 100, cho thấy sự khác biệt đáng kể giữa học sinh thấp nhất và cao nhất.
* **Ngoại lai**: Không có ngoại lai, nghĩa là không có điểm số nào bất thường so với phân bố chung.
* **Bối cảnh thực tế**: Boxplot cho thấy bài kiểm tra có thể không quá khó, vì phần lớn học sinh đạt điểm từ 65 trở lên, và không có học sinh nào đạt điểm cực thấp (dưới 35).

**3. So sánh và diễn giải tổng hợp**

* **Histogram**:
  + Cho thấy phân bố tổng quát, với tần suất cao ở khoảng 70-90, xác nhận đa số học sinh đạt điểm khá.
  + Đuôi dài hơn ở phía điểm thấp (40-60) cho thấy phân bố lệch trái nhẹ.
  + Giúp dễ dàng nhận ra các khoảng điểm phổ biến và mức độ tập trung của dữ liệu.
* **Boxplot**:
  + Cung cấp thông tin chi tiết về trung tâm (trung vị = 75), độ phân tán (IQR = 20), và phạm vi (45-100).
  + Xác nhận phân bố lệch trái qua vị trí của trung vị (gần Q3 hơn Q1).
  + Hữu ích để phát hiện ngoại lai (trong trường hợp này không có).

**Tính toán bổ sung để xác nhận phân bố:**

* **Trung bình**: (45+50+...+100)/50≈74.6(45 + 50 + ... + 100) / 50 \approx 74.6(45+50+...+100)/50≈74.6
* **Trung vị**: 75
* **Mode**: 85 (xuất hiện 4 lần)
* **Skewness**: Trung bình (74.6) < Trung vị (75) < Mode (85), xác nhận phân bố **lệch trái** (negative skew).

**4. Diễn giải trong bối cảnh thực tế**

* **Kết quả học tập**: Phân bố lệch trái cho thấy bài kiểm tra có thể dễ hoặc học sinh được chuẩn bị tốt, vì nhiều học sinh đạt điểm cao (70-100) và ít học sinh đạt điểm thấp (40-60).
* **Hành động cải thiện**: Giáo viên có thể tập trung hỗ trợ một số ít học sinh có điểm thấp (dưới 60) để cải thiện kết quả tổng thể.
* **So sánh với các lớp khác**: Histogram và boxplot có thể được dùng để so sánh phân bố điểm số giữa các lớp hoặc các kỳ thi khác, giúp đánh giá mức độ khó của bài kiểm tra hoặc chất lượng giảng dạy.

**5. Lưu ý khi đọc và diễn giải**

* **Histogram**:
  + Kích thước khoảng (bin size) ảnh hưởng đến hình dạng histogram. Nếu khoảng quá lớn hoặc quá nhỏ, phân bố có thể bị méo mó.
  + Cần xem xét bối cảnh dữ liệu để diễn giải (ví dụ, điểm số lệch trái có thể do bài kiểm tra dễ).
* **Boxplot**:
  + Nhạy với giá trị ngoại lai, nên cần kiểm tra xem ngoại lai có ý nghĩa thực tế hay là lỗi dữ liệu.
  + Vị trí trung vị so với Q1 và Q3 giúp xác định độ lệch nhanh chóng.
* **Kết hợp cả hai**: Histogram cung cấp cái nhìn tổng quát về hình dạng phân bố, trong khi boxplot chi tiết hơn về trung tâm, phân tán, và ngoại lai.

**+ 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?**

Giá trị ngoại lai (outliers) là những điểm dữ liệu khác biệt đáng kể so với phần còn lại của tập dữ liệu, có thể làm sai lệch các chỉ số thống kê mô tả như trung bình, độ lệch chuẩn, hoặc phạm vi. Xử lý giá trị ngoại lai trước khi tính toán thống kê mô tả là một bước quan trọng để đảm bảo kết quả phản ánh chính xác đặc điểm của tập dữ liệu. Dưới đây là các bước và phương pháp xử lý giá trị ngoại lai, cùng với hướng dẫn cách áp dụng trong bối cảnh thực tế.

**1. Xác định giá trị ngoại lai**

Trước khi xử lý, cần xác định giá trị ngoại lai bằng các phương pháp sau:

**a. Sử dụng biểu đồ trực quan**

* **Boxplot**:
  + Giá trị ngoại lai thường được xác định là các điểm nằm ngoài khoảng [Q1−1.5×IQR, Q3+1.5×IQR], trong đó IQR = Q3 - Q1 là khoảng tứ phân vị.
  + Ví dụ: Nếu Q1=20, Q3 = 50, thì IQR = 30, và giá trị ngoại lai là các điểm < 20 − 1.5 × 30 = −25 hoặc > 50 + 1.5 × 30 = 95.
* **Histogram hoặc Scatter Plot**:
  + Giúp nhận diện các giá trị bất thường bằng cách quan sát các điểm xa rời khỏi phần lớn dữ liệu hoặc các đỉnh bất thường trong phân bố.

**b. Phương pháp thống kê**

* **Z-score (Điểm chuẩn hóa)**:
  + Tính Z = (x−μ) / σ​, với μ là trung bình và σ là độ lệch chuẩn.
  + Giá trị có |Z| > 3 (hoặc ngưỡng khác, ví dụ 2.5) thường được coi là ngoại lai, vì chúng nằm ngoài 99.7% dữ liệu trong phân bố chuẩn.
* **Phương pháp IQR**:
  + Như đã đề cập trong boxplot, các giá trị ngoài [Q1−1.5×IQR,Q3+1.5×IQR] là ngoại lai.
* **Kiểm định thống kê**: Các kiểm định như Grubbs’ Test hoặc Dixon’s Q Test có thể được dùng để xác định ngoại lai trong các tập dữ liệu nhỏ.

**c. Hiểu bối cảnh dữ liệu**

* Xác định xem giá trị ngoại lai có ý nghĩa thực tế hay không (ví dụ, điểm số 100 trong bài kiểm tra là hợp lý, nhưng 1000 là lỗi nhập liệu).
* Kiểm tra xem ngoại lai có phải do lỗi đo lường, nhập liệu sai, hay phản ánh một hiện tượng thực tế (như thu nhập cực cao của một cá nhân trong tập dữ liệu).

**2. Các phương pháp xử lý giá trị ngoại lai**

Sau khi xác định, có nhiều cách để xử lý giá trị ngoại lai, tùy thuộc vào mục đích phân tích và bản chất của dữ liệu. Dưới đây là các phương pháp phổ biến:

**a. Loại bỏ giá trị ngoại lai (Deletion)**

* **Cách thực hiện**:
  + Xóa các điểm dữ liệu được xác định là ngoại lai khỏi tập dữ liệu.
  + Ví dụ: Nếu giá trị 1000 trong tập điểm số [45, 50, ..., 100, 1000] là lỗi, loại bỏ 1000.
* **Khi sử dụng**:
  + Phù hợp khi ngoại lai rõ ràng là lỗi nhập liệu hoặc không có ý nghĩa thực tế.
  + Tỷ lệ ngoại lai nhỏ (ví dụ, <5% dữ liệu).
* **Ưu điểm**:
  + Đơn giản, đảm bảo các chỉ số thống kê không bị sai lệch bởi các giá trị bất thường.
* **Nhược điểm**:
  + Làm giảm kích thước mẫu, có thể mất thông tin quan trọng nếu ngoại lai mang ý nghĩa thực tế.
  + Không phù hợp nếu ngoại lai là một phần tự nhiên của dữ liệu (ví dụ, thu nhập cao trong phân tích kinh tế).

**b. Thay thế giá trị ngoại lai (Imputation)**

* **Cách thực hiện**:
  + Thay giá trị ngoại lai bằng một giá trị hợp lý, như:
    - **Trung bình hoặc trung vị**: Thay ngoại lai bằng trung bình hoặc trung vị của tập dữ liệu (không tính ngoại lai).
    - **Ranh giới IQR**: Thay ngoại lai bằng giá trị gần nhất trong khoảng [Q1−1.5×IQR,Q3+1.5×IQR] (ví dụ, thay giá trị > Q3+1.5×IQR bằng Q3+1.5×IQR.
    - **Dự đoán**: Sử dụng mô hình học máy (như KNN, hồi quy) để dự đoán giá trị thay thế dựa trên các biến khác.
* **Khi sử dụng**:
  + Khi muốn giữ nguyên kích thước tập dữ liệu.
  + Phù hợp với dữ liệu có ngoại lai do lỗi đo lường hoặc khi ngoại lai không quá quan trọng.
* **Ưu điểm**:
  + Giữ nguyên số lượng quan sát.
  + Giảm tác động của ngoại lai lên các chỉ số như trung bình và độ lệch chuẩn.
* **Nhược điểm**:
  + Có thể làm sai lệch phân bố nếu thay thế không chính xác.
  + Làm giảm sự biến thiên tự nhiên của dữ liệu.

**c. Biến đổi dữ liệu (Transformation)**

* **Cách thực hiện**:
  + Áp dụng biến đổi toán học để giảm tác động của ngoại lai, ví dụ:
    - **Log transformation**: log(x) hoặc log(x+1) để giảm độ lớn của các giá trị cực đại.
    - **Căn bậc hai**: sqrt{x} để nén các giá trị lớn.
    - **Chuẩn hóa (Standardization)**: Chuyển dữ liệu về dạng Z-score để giảm ảnh hưởng của các giá trị xa.
  + Ví dụ: Nếu tập dữ liệu có giá trị ngoại lai 1000, áp dụng log(1000) = 3 sẽ làm giá trị này gần hơn với các giá trị khác.
* **Khi sử dụng**:
  + Phù hợp với dữ liệu lệch mạnh (như thu nhập, thời gian chờ) hoặc khi ngoại lai là một phần tự nhiên của dữ liệu.
* **Ưu điểm**:
  + Giữ nguyên tất cả dữ liệu và giảm tác động của ngoại lai mà không cần xóa.
  + Giúp dữ liệu gần hơn với phân bố chuẩn, thuận lợi cho phân tích thống kê.
* **Nhược điểm**:
  + Có thể làm phức tạp việc diễn giải kết quả (ví dụ, giá trị log không trực quan như giá trị gốc).
  + Không phù hợp với dữ liệu có giá trị âm hoặc bằng 0 (trừ khi điều chỉnh).

**d. Giữ nguyên giá trị ngoại lai**

* **Cách thực hiện**:
  + Không xử lý ngoại lai mà phân tích dữ liệu với sự hiện diện của chúng, sử dụng các chỉ số thống kê ít nhạy cảm với ngoại lai (như trung vị, IQR thay vì trung bình, độ lệch chuẩn).
* **Khi sử dụng**:
  + Khi ngoại lai có ý nghĩa thực tế quan trọng (ví dụ, thu nhập cực cao của một CEO trong phân tích tài chính).
  + Khi sử dụng các phương pháp thống kê phi tham số hoặc mô hình học máy có khả năng xử lý ngoại lai (như cây quyết định, Random Forest).
* **Ưu điểm**:
  + Bảo toàn thông tin gốc của dữ liệu.
  + Phù hợp với các phân tích cần giữ nguyên đặc điểm tự nhiên của dữ liệu.
* **Nhược điểm**:
  + Các chỉ số như trung bình và độ lệch chuẩn có thể bị sai lệch nghiêm trọng.
  + Cần chọn phương pháp phân tích phù hợp để giảm thiểu tác động của ngoại lai.

**e. Phân tích riêng biệt (Segmentation)**

* **Cách thực hiện**:
  + Tách tập dữ liệu thành hai nhóm: nhóm không có ngoại lai và nhóm chứa ngoại lai, sau đó phân tích riêng.
  + Ví dụ: Phân tích điểm số học sinh thông thường và học sinh có điểm cực cao/thấp riêng biệt.
* **Khi sử dụng**:
  + Khi ngoại lai đại diện cho một nhóm đặc biệt (ví dụ, học sinh xuất sắc hoặc yếu kém).
* **Ưu điểm**:
  + Cho phép phân tích chi tiết từng nhóm, giữ được ý nghĩa của ngoại lai.
* **Nhược điểm**:
  + Làm phức tạp phân tích, đặc biệt với tập dữ liệu nhỏ.
  + Có thể không phù hợp nếu ngoại lai không thuộc một nhóm đặc biệt.

**3. Quy trình xử lý giá trị ngoại lai trước khi tính toán thống kê mô tả**

Dưới đây là quy trình tổng quát:

1. **Xác định ngoại lai**:
   * Sử dụng boxplot, Z-score, hoặc IQR để tìm các giá trị ngoại lai.
   * Kiểm tra bối cảnh: Ngoại lai là lỗi hay có ý nghĩa thực tế?
2. **Đánh giá tác động**:
   * Tính thử các chỉ số thống kê (trung bình, độ lệch chuẩn, trung vị, v.v.) với và không có ngoại lai để xem mức độ ảnh hưởng.
   * Ví dụ: Trung bình và độ lệch chuẩn nhạy cảm với ngoại lai, trong khi trung vị và IQR ít bị ảnh hưởng.
3. **Chọn phương pháp xử lý**:
   * Nếu ngoại lai là lỗi (như nhập liệu sai): Loại bỏ hoặc thay thế.
   * Nếu ngoại lai có ý nghĩa thực tế: Giữ nguyên hoặc biến đổi dữ liệu.
   * Nếu cần so sánh: Phân tích riêng biệt hoặc sử dụng chỉ số ít nhạy cảm (trung vị, IQR).
4. **Kiểm tra sau xử lý**:
   * Vẽ lại histogram hoặc boxplot để kiểm tra phân bố sau khi xử lý.
   * So sánh các chỉ số thống kê trước và sau xử lý để đảm bảo không làm sai lệch dữ liệu.
5. **Ghi nhận phương pháp**:
   * Ghi lại cách xử lý ngoại lai để đảm bảo tính minh bạch và tái lập trong phân tích.

**4. Ví dụ minh họa**

Giả sử tập dữ liệu điểm số: [45, 50, 55, 60, 65, 70, 75, 80, 85, 1000] (1000 là ngoại lai có thể do lỗi nhập liệu).

**Bước 1: Xác định ngoại lai**

* **Boxplot**:
  + Q1 = 55, Q2 = 67.5, Q3 = 80, IQR = 80 - 55 = 25.
  + Ranh giới: [Q1−1.5×IQR,Q3+1.5×IQR]=[55−1.5×25,80+1.5×25]=[17.5,117.5]
  + Giá trị 1000 > 117.5, là ngoại lai.
* **Z-score**:
  + Trung bình (có 1000) ≈ 124.5, độ lệch chuẩn ≈ 297.5.
  + Z-score của 1000: Z=(1000−124.5) / 297.5 ≈ 2.94, gần ngưỡng 3, xác nhận là ngoại lai.

**Bước 2: Xử lý ngoại lai**

* **Phương án 1: Loại bỏ**:
  + Dữ liệu mới: [45, 50, 55, 60, 65, 70, 75, 80, 85]
  + Trung bình: 65, Trung vị: 65, Độ lệch chuẩn: ≈12.9, Phạm vi: 85 - 45 = 40.
* **Phương án 2: Thay thế bằng Q3 + 1.5 \times IQR**:
  + Thay 1000 bằng 117.5.
  + Dữ liệu mới: [45, 50, 55, 60, 65, 70, 75, 80, 85, 117.5]
  + Trung bình: 70.25, Trung vị: 67.5, Độ lệch chuẩn: ≈21.7.
* **Phương án 3: Biến đổi log**:
  + Áp dụng log10​(x): [1.65, 1.70, 1.74, 1.78, 1.81, 1.85, 1.88, 1.90, 1.93, 3]
  + Trung bình log: ≈1.82, Độ lệch chuẩn log: ≈0.39.
* **Phương án 4: Giữ nguyên**:
  + Trung bình: 124.5, Trung vị: 67.5, Độ lệch chuẩn: ≈297.5.
  + Trung bình bị sai lệch nghiêm trọng, nhưng trung vị vẫn đáng tin cậy.

**Bước 3: Diễn giải**

* Nếu 1000 là lỗi nhập liệu, **loại bỏ** hoặc **thay thế** là hợp lý để đảm bảo trung bình và độ lệch chuẩn chính xác.
* Nếu 1000 có ý nghĩa (ví dụ, điểm bonus), có thể **giữ nguyên** và dùng trung vị/IQR hoặc **biến đổi log** để phân tích.

**5. Lưu ý khi xử lý ngoại lai**

* **Bối cảnh dữ liệu**: Luôn xem xét ý nghĩa thực tế của ngoại lai trước khi xử lý.
* **Tác động đến chỉ số thống kê**:
  + Trung bình và độ lệch chuẩn nhạy cảm với ngoại lai.
  + Trung vị, Q1, Q3, và IQR ít bị ảnh hưởng, nên ưu tiên khi ngoại lai chưa được xử lý.
* **Kiểm tra sau xử lý**: So sánh phân bố (histogram, boxplot) trước và sau xử lý để đảm bảo không làm mất đặc điểm tự nhiên của dữ liệu.
* **Phương pháp phân tích**: Nếu giữ nguyên ngoại lai, hãy chọn các phương pháp thống kê phi tham số hoặc mô hình ít nhạy cảm với ngoại lai.

**6. Kết luận**

* **Xác định ngoại lai**: Dùng boxplot, Z-score, hoặc kiểm định thống kê, kết hợp với bối cảnh dữ liệu.
* **Phương pháp xử lý**:
  + **Loại bỏ**: Phù hợp với lỗi nhập liệu, tỷ lệ ngoại lai nhỏ.
  + **Thay thế**: Giữ kích thước mẫu, dùng trung vị hoặc ranh giới IQR.
  + **Biến đổi**: Giảm tác động của ngoại lai, đặc biệt với dữ liệu lệch mạnh.
  + **Giữ nguyên**: Khi ngoại lai có ý nghĩa thực tế, dùng trung vị/IQR.
  + **Phân tích riêng**: Khi ngoại lai đại diện cho một nhóm đặc biệt.
* **Trước khi tính thống kê mô tả**: Ưu tiên xử lý ngoại lai nếu chúng làm sai lệch trung bình hoặc độ lệch chuẩn; nếu không, sử dụng các chỉ số ít nhạy cảm như trung vị và IQR.