* 1. Thống kê mô tả
     1. Ôn tập lý thuyết
* **Thống kê mô tả và thống kê suy luận**
* Thống kê mô tả (Descriptive statistics) là tập hợp các phương pháp dùng để **tóm tắt, mô tả và trình bày** dữ liệu thông qua các chỉ số (trung bình, trung vị, độ lệch chuẩn, phạm vi...) và biểu đồ (histogram, boxplot...). Nó chỉ phản ánh đặc điểm của tập dữ liệu hiện có mà không đưa ra suy luận cho quần thể lớn hơn.
* Ngược lại, thống kê suy luận (Inferential statistics) sử dụng **mẫu dữ liệu** để đưa ra **ước lượng, dự đoán hoặc kiểm định giả thuyết** cho toàn bộ quần thể. Ví dụ: dùng mẫu khảo sát để ước lượng thu nhập trung bình của toàn bộ dân số.
* **Các thước đo thống kê mô tả chính: một số chỉ số quan trọng gồm:**
* Trung bình (Mean): giá trị trung tâm, đại diện cho xu hướng chung của dữ liệu.
* Trung vị (Median): giá trị chia dữ liệu thành hai nửa bằng nhau; thích hợp dùng khi dữ liệu có nhiều ngoại lai hoặc phân bố lệch.
* Phương sai (Variance) và Độ lệch chuẩn (Standard Deviation): đo mức độ phân tán của dữ liệu xung quanh trung bình.

🡺Trong trường hợp dữ liệu có phân bố lệch hoặc tồn tại nhiều giá trị ngoại lai, nên dùng trung vị thay vì trung bình để phản ánh xu hướng trung tâm chính xác hơn.

* **Xác định phân bố dữ liệu và các dạng phổ biến**

Để xác định phân bố, có thể:

* **Trực quan**: vẽ histogram, boxplot hoặc Q-Q plot để quan sát dạng phân bố.
* **Thống kê**: tính skewness (độ lệch) và kurtosis (độ nhọn), hoặc dùng các kiểm định như Shapiro-Wilk, Kolmogorov-Smirnov.

Các dạng phân bố thường gặp:

* **Phân bố chuẩn (Normal distribution)**: đối xứng, dạng chuông, Mean ≈ Median ≈ Mode.
* **Phân bố lệch trái (Left-skewed)**: đuôi dài về bên trái, Mean < Median.
* **Phân bố lệch phải (Right-skewed)**: đuôi dài về bên phải, Mean > Median.  
  Ngoài ra còn có phân bố đều (Uniform), log-normal, Poisson, exponential... thường xuất hiện trong dữ liệu thực tế.
* **Độ lệch chuẩn và phạm vi:** Phạm vi (range) được tính bằng hiệu giữa giá trị lớn nhất và nhỏ nhất, cho thấy dữ liệu trải dài trong khoảng bao nhiêu. Tuy nhiên, phạm vi thường bị ảnh hưởng mạnh bởi các giá trị ngoại lai. Trong khi đó, độ lệch chuẩn (standard deviation) phản ánh mức độ dao động trung bình của các điểm dữ liệu quanh giá trị trung bình. Độ lệch chuẩn nhỏ nghĩa là dữ liệu tập trung, còn lớn nghĩa là dữ liệu phân tán rộng.
* **Các tứ phân vị trong boxplot (Q1, Q2, Q3)** Boxplot thể hiện dữ liệu thông qua các tứ phân vị:
* Q1 (25%): cận dưới, cho biết 1/4 dữ liệu nhỏ nhất.
* Q2 (50%): trung vị, chia tập dữ liệu thành hai nửa bằng nhau.
* Q3 (75%): cận trên, cho biết 3/4 dữ liệu nhỏ nhất.  
  Khoảng giữa Q1 và Q3 gọi là IQR (Interquartile Range), dùng để đo độ phân tán và phát hiện giá trị ngoại lai.
* **Xử lý giá trị thiếu (missing values)** trước khi tính toán thống kê mô tả, cần xử lý dữ liệu bị thiếu. Một số cách phổ biến là:
* Loại bỏ quan sát nếu số lượng thiếu ít.
* Thay thế (impute) bằng trung bình, trung vị hoặc mode.
* Dùng phương pháp nâng cao như nội suy hoặc mô hình dự đoán để ước lượng giá trị còn thiếu.
* **Cách đọc histogram và boxplot**
* Histogram cho thấy tần suất xuất hiện của dữ liệu trong từng khoảng, giúp nhận biết dạng phân bố (chuẩn, lệch trái, lệch phải, nhiều đỉnh…).
* Boxplot cho thấy nhanh sự phân tán, trung vị và các giá trị ngoại lai. Nếu hộp nghiêng về một phía, dữ liệu có xu hướng lệch về phía đó. Các điểm nằm ngoài “râu” thường được coi là ngoại lai.
* **Xử lý giá trị ngoại lai (outliers)** ngoại lai có thể do lỗi nhập liệu hoặc do bản chất dữ liệu. Nếu là lỗi, có thể loại bỏ. Nếu là giá trị thực, nên cân nhắc giữ lại và phân tích riêng. Ngoài ra, có thể dùng các kỹ thuật như log-transform, winsorizing hoặc chuẩn hóa để giảm ảnh hưởng. Trong một số trường hợp, việc so sánh giữa dữ liệu “có ngoại lai” và “không ngoại lai” mang lại cái nhìn toàn diện hơn.