**1.1.1. Ôn tập lý thuyết**

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

**Trả lời:**

* Thống kê mô tả: Là mô tả dữ liệu đang có những gì bằng các số đo (min, max, median, mean, …) hoặc histogram.
* Thống kê suy luận: Là suy luận ra giữa các quần thể, đặt ra giả thuyết, kiểm thử , ước lượng, …

**+ 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?**

**Trả lời:**

* **Trung bình (Mean)**: Là giá trị trung bình của một đối tượng, tuy nhiên nếu dữ liệu bị “lệch” thì giá trị sẽ không chính xác. Dùng khi phân bố tương đối và ít “lệch”.
* **Trung vị (Median)**: Giá trị ở vị trí 50%, thể hiện rõ giá trị bị “lệch. Dùng khi dữ liệu lệch hoặc mất cân đối
* **Phương sai (Variance)**: mức độ phân tán quanh mean (đơn vị bình phương).
* **Độ lệch chuẩn (SD)**: căn bậc hai của phương sai; cùng đơn vị với dữ liệu, dễ diễn giải hơn variance.

**+ 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)?**

**Trả lời:**

**Cách xác định phân bố:**

* Histogram (thấy được mối tương quan giữa các biến)
* Boxplot (thấy được IQR, và outlier của dữ liệu)

**Dạng phổ biến:**

* Gaussian (phân phối chuẩn): dữ liệu đối xứng, 1 đỉnh duy nhất và mean ~ median.
* Lệch phải (skwed): phân bố bị trải dài bên phải
* Lệch trái (skwed): phân bố bị trải dài bên trái
* Đa đỉnh: dữ liệu có nhiều biến bất thường

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

* SD: cho thấy mức độ dao động quanh mean, mạnh khi phân bố gần chuẩn
* Range (max – min): dễ trực quan, nhìn nhanh outlier.

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

* Q1 (25%): 25% dữ liệu nằm ≤ Q1.
* Q2 (Median, 50%): “giữa” dữ liệu.
* Q3 (75%): 75% dữ liệu nằm ≤ Q3.

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

**Trả lời:**

* **Hiểu nguyên nhân:** MCAR (missing completely at random), MAR (missing at random), MNAR (missing not at random).
* **Chiến lược:**
* Ít và rải rác: Drop dòng có NA (cần kiểm tra lại).
* Impute:
  + Số: median (bền vững) hoặc mean (khi chuẩn), KNN, mô hình hồi quy.
  + Phân loại: mode (giá trị phổ biến).
* **Không nên:** tự động thay 0 cho NA (trừ khi 0 có nghĩa thật).

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

**Trả lời:**

**Histogram**

* Nhìn tâm (đỉnh), độ rộng (phân tán), đuôi (lệch), số đỉnh (đa nhóm).

**Boxplot**

* Median trong hộp: vị trí gần Q1/Q3 cho thấy lệch.
* Chiều cao hộp (IQR): phân tán của “khu vực trung tâm”.
* Whiskers dài ngắn: bất đối xứng → lệch.
* Chấm ngoài whiskers: outliers; nhiều chấm phía phải → lệch phải.

**+ 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?**

**Trả lời:**

* Kiểm tra giá trị outliers đó do đâu (nhập, sai đơn vị). Nếu có thì cần sửa lại hoặc loại bỏ.
* Xác định biến đó có phạm vi bao nhiêu, xem xét các giá trị đó có hợp lý không?
* Xác định đo bằng median, IQR, SD để nhìn độ phân tán
* Báo cáo về cách xử lý đối với outliers và lý do