<u>Lab 01</u>: Ôn tập lý thuyết

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

- Thống kê mô tả là nhánh của thống kê dùng để mô tả, tóm tắt và trình bày dữ liệu đã thu thập được.
- Nó không rút ra kết luận hay dự đoán nào vượt ra ngoài dữ liệu mẫu.

	Thống kê mô tả	Thống kê suy luận
Mục tiêu	Mô tả và tóm tắt dữ liệu có sẵn	Suy luận/Dự đoán về quần thể
Phạm vi	Chỉ áp dụng cho mẫu hiện tại	Mở rộng kết luận cho toàn bộ quần thể
Dựa vào xác suất	Không	Có
Công cụ	Mean, media, mode, charts	t-test, chi-square, regression

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

Thước đo	Mục đích chính	Vai trò trong phân tích dữ liệu
Trung bình	Xác định giá trị trung tâm hoặc giá trị điển hình của dữ liệu.	Dùng để tính tổng thể, là nền tảng cho nhiều phân tích thống kê nâng cao (hồi quy, kiểm định t,...).
Trung vị	Xác định giá trị ở giữa của tập dữ liệu đã sắp xếp.	Cung cấp thước đo trung tâm ổn định khi dữ liệu có ngoại lai hoặc bị lệch.
Phương sai	Đo lường mức độ phân tán của dữ liệu xung quanh giá trị	Cung cấp thông tin về sự biến động. Phương sai lớn → dữ

	trung bình.	liệu phân tán rộng.
Độ lệch chuẩn	Đo lường độ phân tán của dữ liệu, nhưng ở đơn vị giống với dữ liệu gốc (là căn bậc hai của phương sai).	Giúp trực quan hóa mức độ rủi ro, biến động hoặc sự khác biệt giữa các điểm dữ liệu so với trung bình

- Nên dùng Trung vị (Median) thay cho Trung bình (Mean) trong các trường hợp sau:

Trường hợp 1: Dữ liệu có Giá trị Ngoại lai (Outliers) outlier (hoặc giá trị cực đoan)

- **Vấn đề:** Trung bình cộng rất nhạy cảm với các giá trị quá lớn hoặc quá nhỏ. Một hoặc hai giá trị ngoại lai có thể kéo giá trị trung bình lên hoặc xuống đáng kể, làm cho nó không còn đại diện cho phần lớn dữ liệu nữa.
- **Ưu điểm của Trung vị:** Trung vị chỉ phụ thuộc vào giá trị ở vị trí giữa, nên nó bền vững hơn và ít bị ảnh hưởng bởi ngoại lai.

Trường hợp 2: Dữ liệu có Phân phối Bị lệch (Skewed Distribution)

- **Vấn đề:** Khi phân phối dữ liệu không đối xứng (ví dụ: lệch phải như dữ liệu về thu nhập, thời gian chờ đợi), trung bình sẽ bị kéo về phía đuôi dài hơn, không còn nằm ở vị trí trung tâm nhất của dữ liệu.

3. 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)?

Có thể xác định dạng phân bố của dữ liệu bằng các cách sau:

(1) Trực quan hóa dữ liệu

- *Histogram (biểu đồ tần suất):* cho thấy dữ liệu tập trung ở đâu và có bị lệch không.
- *Boxplot (biểu đồ hộp):* giúp phát hiện độ lệch và các giá trị ngoại lệ (outliers).
- *Q-Q Plot (Quantile–Quantile plot):* so sánh dữ liệu thực tế với phân bố chuẩn.

(2) Thống kê mô tả

Kiểm tra độ lệch (skewness) và độ nhọn (kurtosis) của dữ liệu

Nếu skew ≈ 0: dữ liệu có phân bố chuẩn (normal)

Nếu skew > 0: dữ liệu lệch phải (right-skewed)
Nếu skew < 0: dữ liệu lệch trái (left-skewed)

(3) Kiểm định thống kê

Dùng kiểm định Shapiro–Wilk hoặc Kolmogorov–Smirnov để kiểm tra tính chuẩn hóa

Để xác định phân bố dữ liệu, cần kết hợp cả **biểu đồ trực quan**, **chỉ số thống kê** và **kiểm định thống kê** để đảm bảo đánh giá chính xác.

4. Độ lệch chuẩn và phạm vi (range) có ý nghĩa gì trong việc đánh giá sự phân tán của dữ liệu?
    - Ý nghĩa của Độ lệch chuẩn (Standard Deviation - SD):
        + Độ lệch chuẩn là thước đo được sử dụng phổ biến và có ý nghĩa nhất trong việc đánh giá sự phân tán.

- *Ý nghĩa:* Độ lệch chuẩn đo lường mức độ trung bình mà các điểm dữ liệu lệch khỏi giá trị trung bình (Mean) của tập dữ liệu.
- *Giá trị và Diễn giải:*
    - SD nhỏ: Các giá trị dữ liệu tập trung chặt chẽ xung quanh giá trị trung bình. Dữ liệu có tính ổn định cao và ít biến động.
    - SD lớn: Các giá trị dữ liệu phân tán rộng rãi ra xa giá trị trung bình. Dữ liệu biến động mạnh và kém ổn định.
- *Ưu điểm*: Độ lệch chuẩn sử dụng tất cả các điểm dữ liệu trong tính toán và có cùng đơn vị với dữ liệu gốc (vì nó là căn bậc hai của Phương sai), giúp diễn giải trực quan và dễ so sánh hơn.
- *Ứng dụng:* Thường được dùng để đánh giá rủi ro (trong tài chính), độ tin cậy của kết quả đo lường (trong khoa học) và là thành phần quan trọng trong các kiểm định thống kê nâng cao (như t-test, hồi quy).

- Ý nghĩa của Phạm vi (Range)

+ Phạm vi là thước đo đơn giản nhất để đánh giá sự phân tán.

- **Ý nghĩa:** Phạm vi cho biết khoảng cách tuyệt đối giữa giá trị lớn nhất và giá trị nhỏ nhất trong tập dữ liệu.
- **Giá trị và Diễn giải:** Phạm vi càng lớn, độ dàn trải tổng thể của dữ liệu càng lớn.
- **Ưu điểm:** Rất dễ tính toán và dễ hiểu.
- **Hạn chế quan trọng:** Phạm vi chỉ sử dụng hai giá trị (Min và Max) và hoàn toàn bị ảnh hưởng bởi bất kỳ giá trị ngoại lai (outlier) nào.

Nó không cung cấp thông tin về cách các điểm dữ liệu khác được phân tán ở giữa.

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

- Sự khác biệt chính giữa Q1, Q2, và Q3 nằm ở vị trí của chúng trong tập dữ liệu đã được sắp xếp và tỷ lệ phần trăm dữ liệu mà mỗi điểm đại diện.

	Tên gọi	Tỉ lệ phần trăm	Vị trí trong boxplot	Ý nghĩa
Q1	Tứ phân vị thứ nhất	25%	Cạnh dưới/trái của hộp.	25% số liệu có giá trị nhỏ hơn hoặc bằng Q1.
Q2	Tứ phân vị thứ hai	50%	Đường trung tâm (ngăn cách) bên trong hộp.	Chính là Trung vị (Median). 50% số liệu nhỏ hơn hoặc bằng Q2.
Q3	Tứ phân vị thứ ba	75%	Cạnh trên/phải của hộp.	75% số liệu có giá trị nhỏ hơn hoặc bằng Q3.

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

Bước 1: Kiểm tra giá trị thiếu

Sử dụng các hàm kiểm tra để xác định cột nào có giá trị bị thiếu

Bước 2: Xử lý giá trị thiếu

Tùy vào mức độ và loại dữ liệu, ta có thể chọn một trong các phương pháp sau:

(1) Loại bỏ giá trị thiếu

- Dùng khi số lượng giá trị thiếu ít và không ảnh hưởng nhiều đến dữ liệu.
- Có thể loại bỏ hàng hoặc cột chứa giá trị thiếu

(2) Thay thế (điền) giá trị thiếu bằng giá trị phù hợp

Với dữ liệu số (numeric):

- Dùng mean, median, hoặc mode để thay thế.
- Mean phù hợp khi dữ liệu phân bố chuẩn.
- Median phù hợp khi dữ liệu bị lệch.

Với dữ liệu phân loại (categorical):

- Điền bằng mode (giá trị xuất hiện nhiều nhất).

(3) Sử dụng kỹ thuật suy diễn nâng cao

Áp dụng khi dữ liệu phức tạp hoặc có nhiều giá trị thiếu.

Một số phương pháp phổ biến:

- KNN Imputer: dự đoán giá trị thiếu dựa trên các điểm dữ liệu gần nhất.
- Iterative Imputer (MICE): ước lượng giá trị thiếu dựa trên các biến khác.

3. Kiểm tra lại dữ liệu sau khi xử lý

4. Kết luận

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

- Cách đọc và diễn giải biểu đồ Histogram
    + Histogram cho thấy **hình dạng** và **tần suất** phân phối của một biến định lượng, tức là cách các giá trị dữ liệu được nhóm lại và phân bổ như thế nào.
    + Cách đọc các thành phần:

Thành phần	Ý nghĩa
Trục X	Biểu diễn phạm vi giá trị của biến dữ liệu, được chia thành các khoảng (bins).
Trục Y	Biểu diễn tần suất (Frequency) hoặc tỉ lệ phần trăm (Relative Frequency) của các quan sát rơi

	vào mỗi khoảng trên trục X.
Các cột	Chiều cao của mỗi cột cho biết có bao nhiêu điểm dữ liệu nằm trong khoảng giá trị mà cột đó đại diện.

+ Cách diễn giải hình dạng phân khối:
  Phân phối chuẩn (Normal/Symmetric):

- Các cột cao nhất nằm ở giữa.
- Các cột giảm dần đều về hai phía, tạo thành hình quả chuông (bell-shaped).
- **Ý nghĩa:** Dữ liệu cân bằng, Giá trị Trung bình (Mean), Trung vị (Median) và Mode nằm gần nhau.

  Phân phối lệch trái (Skewed Left / Negative Skew):

- Phần lớn dữ liệu tập trung ở bên **phải** (giá trị cao).
- "Đuôi" của biểu đồ kéo dài về bên **trái** (giá trị thấp).
- **Ý nghĩa:** Giá trị trung bình **nhỏ hơn** trung vị (Mean<Median).

  Phân phối lệch phải (Skewed Right / Positive Skew):

- Phần lớn dữ liệu tập trung ở bên **trái** (giá trị thấp).
- "Đuôi" của biểu đồ kéo dài về bên **phải** (giá trị cao).
- **Ý nghĩa:** Giá trị trung bình **lớn hơn** trung vị (Mean>Median).

  Phân phối Đa mode (Multimodal):

- Biểu đồ có hai hoặc nhiều đỉnh (cột cao) riêng biệt.
- **Ý nghĩa:** Dữ liệu có thể đến từ hai hoặc nhiều nhóm dân số khác nhau (ví dụ: chiều cao của nam và nữ được gộp chung).

- Cách đọc và diễn giải Biểu đồ Boxplot:

+ Cách đọc các thành phần:

Thành phần	Ký hiệu	Ý nghĩa
Đường trung tâm của hộp	Q2	Trung vị

Cạnh dưới/trái của hộp	Q1	Tứ phân vị thứ nhất
Cạnh trên/phải của hộp	Q3	Tứ phân vị thứ ba
Chiều dài của hộp	IQR(Q3-Q1)	Khoảng Tứ phân vị, chứa 50% dữ liệu ở giữa (mô tả độ phân tán cốt lõi).
Râu (Whiskers)	Max (không phải ngoại lai) và Min (không phải ngoại lai)	Kéo dài đến giá trị lớn nhất và nhỏ nhất không được coi là ngoại lai.
Các dấu chấm/ngôi sao ngoài râu	Outliers	Các giá trị ngoại lai

- Cách diễn giải phân phối và biến động:

Tính ổn định/Phân tán:

- **Hộp càng ngắn** (IQR nhỏ): Dữ liệu càng tập trung, biến động càng thấp.
- **Hộp càng dài** (IQR lớn): Dữ liệu phân tán rộng, biến động cao.

Tính đối xứng (Symmetry) và Độ lệch (Skewness):

- **Đối xứng:** Đường Q2 (Trung vị) nằm gần giữa hộp và hai râu có chiều dài gần bằng nhau.
- **Lệch phải (Positive Skew):** Đường Q2 nằm gần cạnh dưới (Q1), và râu trên (Max) dài hơn râu dưới.
- **Lệch trái (Negative Skew):** Đường Q2 nằm gần cạnh trên (Q3), và râu dưới (Min) dài hơn râu trên.

Giá trị Ngoại lai (Outliers):

- Sự xuất hiện của các chấm bên ngoài râu chỉ ra những quan sát có giá trị cực đoan, cần được điều tra xem chúng là lỗi nhập liệu hay là dữ liệu có ý nghĩa đặc biệt.

Ví dụ Diễn giải Dữ liệu Thực tế

- **Sử dụng Histogram:** Nếu bạn thấy một Histogram về **thu nhập** lệch phải, điều đó cho thấy phần lớn dân số có thu nhập thấp hoặc trung bình, nhưng có một số ít người có thu nhập rất cao kéo dài "đuôi" về phía phải.
- **Sử dụng Boxplot:** Khi so sánh doanh số bán hàng của hai khu vực A và B bằng Boxplot:
    - Nếu hộp của khu vực A **ngắn hơn** khu vực B, khu vực A có doanh số ổn định hơn.
    - Nếu **Trung vị (Q2)** của khu vực A cao hơn Q3 của khu vực B, điều này chỉ ra rằng ít nhất **50%** doanh số của A cao hơn **75%** doanh số của B, cho thấy A vượt trội rõ rệt.

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?

1. Xác định và Phân tích Nguồn gốc

Trước khi thực hiện bất kỳ thay đổi nào, ta cần xác định xem ngoại lai là do lỗi hay là dữ liệu có ý nghĩa:

- **Lỗi nhập liệu/Đo lường:** Nếu giá trị ngoại lai là do lỗi đánh máy, lỗi cảm biến, hoặc lỗi thu thập dữ liệu (ví dụ: chiều cao là 2000 cm), nên **xóa bỏ** quan sát đó hoặc **thay thế** bằng giá trị thiếu (NaN) để xử lý sau.
- **Giá trị tự nhiên/Có ý nghĩa:** Nếu ngoại lai là một giá trị thực tế, hợp lý nhưng hiếm gặp (ví dụ: thu nhập cực cao, một sự kiện thời tiết cực đoan), việc xóa bỏ có thể làm mất thông tin quan trọng. Trong trường hợp này, nên xem xét sử dụng các phương pháp thống kê **bền vững** (robust) hoặc **biến đổi dữ liệu**.

2. Các Phương pháp Xử lý Chính

Tùy thuộc vào nguồn gốc và mức độ ảnh hưởng của ngoại lai, ta có thể áp dụng các phương pháp sau:

A. Loại bỏ (Deletion)

- **Mô tả:** Xóa các quan sát chứa giá trị ngoại lai.

- **Khi sử dụng:** Chỉ áp dụng khi chắc chắn ngoại lai là **lỗi** hoặc khi tập dữ liệu **rất lớn** và việc loại bỏ không làm ảnh hưởng đến tính đại diện của mẫu.

B. Biến đổi Dữ liệu (Transformation)

- **Mô tả:** Áp dụng các hàm toán học để giảm ảnh hưởng của các giá trị cực lớn, làm cho phân phối dữ liệu gần với phân phối chuẩn hơn.
- **Ví dụ phổ biến:**
  - **Logarit (log):** Thường dùng cho các biến lệch dương (positive skewed) như thu nhập, giá cả.
  - **Căn bậc hai (x):** Cũng giúp giảm độ lệch.
- **Lưu ý:** Sau khi biến đổi, các chỉ số thống kê mô tả sẽ được tính trên dữ liệu đã biến đổi, nên việc diễn giải kết quả có thể phức tạp hơn.

C. Winsorizing (Capping)

- **Mô tả:** Thay thế các giá trị ngoại lai bằng giá trị nằm ở **ngưỡng nhất định** (ví dụ: 5% hoặc 95% của phân phối). Thay vì xóa, bạn "kẹp" chúng lại.
- **Ví dụ:** Thay thế mọi giá trị lớn hơn **phân vị thứ 95** bằng chính giá trị của phân vị thứ 95 đó.
- **Ưu điểm:** Giữ lại tất cả các quan sát, không làm giảm kích thước mẫu, và giảm thiểu ảnh hưởng của các giá trị cực đoan.

3. Sử dụng Thống kê Bền vững (Robust Statistics)

Nếu không muốn/không thể thay đổi dữ liệu gốc, hãy chọn các chỉ số thống kê mô tả ít nhạy cảm với ngoại lai:

Chỉ số Thống kê Mô tả	Chỉ số bền vững
Giá trị Trung bình (Mean)	Trung vị
Độ lệch chuẩn	Khoảng Tứ phân vị (IQR) $(Q3-Q1)$

Phạm vi	Khoảng Tứ phân vị (IQR)