1.1. THỐNG KÊ MÔ TẢ

## 1.1.1. 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 phương pháp tóm tắt và trình bày các đặc điểm của một tập dữ liệu cụ thể( trung bình, trung vị, mốt, phương sai, độ lệch chuẩn, …) hay các biểu đồ trực quan như biểu đồ histogram, biểu đồ cột, để giúp người đọc dễ dàng hình dung và hiểu rõ về dữ liệu thu thập được

- Nó khác thống kê suy luận cốt lõi là thống kê mô tả những gì có trong dữ liệu, còn thống kê suy luận mở rộng những gì quan sát được từ mẫu để rút ra kết luận về toàn bộ nhóm lớn hơn

### **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?**

-Các thước đo thống kê mô tả như trung bình, trung vị, mốt, phương sai, độ lệch chuẩn, khoảng và khoảng tứ phân vị được dùng để tóm tắt, mô tả sự tập trung(xu hướng trung tâm) và phân tán của một tập dữ liệu(độ phân tán)

-Nên dùng trung vị thay vì trung bình khi dữ liệu có giá trị ngoại lệ (dữ liệu bất thường) hoặc không phân phối đều, vì trung vị ít bị ảnh hưởng bởi các giá trị cực đoan, giúp phản ánh trung tâm dữ liệu một cách chính xác hơn

### **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)?**

-Để xác định phân bố của một tập dữ liệu, bạn có thể sử dụng các biểu đồ như biểu đồ histogram để xem hình dạng trực quan và phân tích các tham số thống kê như độ lệch (skewness) và độ nhọn (kurtosis) để đo lường tính bất đối xứng và độ dày của đuôi phân bố

-Các loại phân bố phổ biến bao gồm phân bố chuẩn (đường cong hình chuông, đối xứng), phân bố lệch trái (đuôi dài về bên trái, trung bình nhỏ hơn trung vị) và phân bố lệch phải (đuôi dài về bên phải, trung bình lớn hơn trung vị)

### **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 đo lường mức độ phân tán trung bình của dữ liệu so với giá trị trung bình, trong khi phạm vi (range) cho biết sự chênh lệch giữa giá trị lớn nhất và nhỏ nhất trong một tập dữ liệu. độ lệch chuẩn thấp cho thấy dữ liệu tập trung quanh trung bình, còn độ lệch chuẩn cao cho thấy dữ liệu phân tán rộng. phạm vi cho biết tổng quát về độ trải dài của dữ liệu, nhưng nhạy cảm với các giá trị ngoại lai hơn độ lệch chuẩn, vì nó chỉ dựa trên hai giá trị cực đoan nhất.

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

-Trong biểu đồ hộp (boxplot), Q1 (tứ phân vị thứ nhất) là giá trị ở phân vị thứ 25, tức là 25% dữ liệu nhỏ hơn giá trị này. Q2 (trung vị hay median) là giá trị ở phân vị thứ 50, chia đôi tập dữ liệu, với 50% dữ liệu nhỏ hơn và 50% lớn hơn. Q3 (tứ phân vị thứ ba) là giá trị ở phân vị thứ 75, tức là 75% dữ liệu nhỏ hơn giá trị này. hộp trong biểu đồ hộp đánh dấu khoảng giữa Q1 và Q3(khoảng tứ phân vị), với đường thẳng bên trong hộp biểu thị Q2.

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

-Trong quá trình phân tích dữ liệu khám phá (EDA), tôi tiến hành kiểm tra giá trị thiếu bằng các hàm isnull() trong pandas. Kết quả cho thấy một số biến có chứa giá trị bị thiếu, cụ thể như cột Age và Gender. Sau khi đánh giá mức độ ảnh hưởng, tôi nhận thấy tỷ lệ thiếu ở mức chấp nhận được và có thể xử lý thay vì loại bỏ toàn bộ dữ liệu.

-Đối với biến số liên tục (Age), tôi thay thế giá trị thiếu bằng trung vị (median) để tránh ảnh hưởng bởi các giá trị ngoại lai. Với biến phân loại (Gender), tôi thay thế giá trị thiếu bằng giá trị xuất hiện nhiều nhất (mode). Riêng những cột có tỷ lệ thiếu quá cao (>50%), tôi quyết định loại bỏ để không làm sai lệch kết quả phân tích.

-Sau khi xử lý, bộ dữ liệu không còn giá trị thiếu, đảm bảo tính đầy đủ và sẵn sàng cho các bước thống kê mô tả cũng như trực quan hóa tiếp theo.

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

-Histogram hiển thị tần suất xuất hiện của dữ liệu trong các khoảng (bin) khác nhau.Để đọc và diễn giải biểu đồ histogram, ta quan sát hình dạng các cột để xác định phân phối, độ tập trung và độ phân tán của dữ liệu:

Đọc:

* Trục x: biểu thị các khoảng giá trị của dữ liệu.
* Trục y: biểu thị tần suất (số lượng) các điểm dữ liệu trong mỗi khoảng.

Diễn giải:

* Hình dạng phân phối: xem các cột có đối xứng không, hay bị lệch về bên trái (lệch trái) hoặc bên phải (lệch phải).
* Độ tập trung: cột nào có tần suất cao nhất cho biết giá trị dữ liệu thường tập trung ở đó.
* Độ phân tán: độ rộng của các cột và khoảng cách giữa các cột thể hiện mức độ phân tán của dữ liệu.
* Các mode: nếu có nhiều cột có tần suất cao, đó là các "mode", cho thấy nhiều cụm dữ liệu.

-Boxplot (biểu đồ hộp) tập trung vào các phân vị (quartile) của dữ liệu.Đối với boxplot, ta xem xét vị trí đường trung vị, độ dài của hộp và "râu" để hiểu phân vị, độ lệch và các giá trị ngoại lai:

Đọc:

* Đường ngang trong hộp (median): đại diện cho phân vị thứ 2 (P50) hoặc trung vị.
* Biên trên của hộp (Q3): đại diện cho phân vị thứ 3 (P75).
* Biên dưới của hộp (Q1): đại diện cho phân vị thứ 1 (P25).
* Đường "râu" (whiskers): kéo dài từ hộp đến các giá trị tối thiểu và tối đa trong một phạm vi nhất định.
* Các điểm riêng biệt: đại diện cho các giá trị ngoại lai (outliers), là những điểm nằm ngoài phạm vi "râu".

Diễn giải:

* Độ lệch: nếu đường trung vị không nằm chính giữa hộp, dữ liệu bị lệch. nếu đường trung vị lệch về phía râu ngắn hơn, dữ liệu bị lệch về phía râu dài hơn.
* Phân tán: độ dài của hộp cho biết sự phân tán của 50% dữ liệu ở giữa (khoảng cách iqr - interquartile range).
* Giá trị ngoại lai: các điểm riêng biệt cho thấy những giá trị bất thường trong tập dữ liệu.

### **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?**

-Khi gặp dữ liệu có giá trị ngoại lai, tôi sẽ kiểm tra bằng boxplot và quy tắc IQR để xác định mức độ ảnh hưởng. Nếu outlier do lỗi nhập liệu, tôi loại bỏ; nếu là giá trị hợp lý nhưng lệch phân phối, tôi sử dụng trung vị thay vì trung bình hoặc áp dụng biến đổi log/winsorization để giảm tác động. Nhờ vậy, thống kê mô tả phản ánh dữ liệu chính xác và ổn định hơn.

1.2. XỬ LÝ VÀ TRỰC QUAN HÓA DỮ LIỆU

## 1.2.1. Ôn tập lý thuyết

## **1. Trực quan hóa dữ liệu có vai trò gì trong phân tích dữ liệu? Tại sao nó quan trọng trong khám phá dữ liệu (EDA)?**

* Giúp hiểu nhanh cấu trúc và phân phối dữ liệu, phát hiện dữ liệu lệch, nhiều mode hoặc bất thường.
* Hỗ trợ phát hiện dữ liệu thiếu, giá trị ngoại lai, lỗi đo lường.
* Làm rõ mối quan hệ giữa các biến, ví dụ tương quan hay xu hướng.
* Hình thành giả thuyết và định hướng bước xử lý, xây dựng mô hình.
* Giúp truyền đạt kết quả dễ hiểu, kể cả cho người không chuyên về dữ liệu.

### **2. Các loại biểu đồ phổ biến (như histogram, scatter plot, boxplot, bar chart) được sử dụng trong các trường hợp nào?**

* Histogram: dùng cho biến số liên tục, để xem phân phối giá trị.
* Scatter plot: dùng để thể hiện quan hệ giữa hai biến số, kiểm tra tương quan.
* Boxplot: dùng để so sánh phân phối, tìm trung vị, độ phân tán và outlier giữa nhiều nhóm.
* Bar chart: dùng để so sánh độ lớn giữa các nhóm phân loại hoặc dữ liệu đếm.
* Line chart: dùng cho dữ liệu chuỗi thời gian để thể hiện xu hướng theo thời gian.

### **3. Làm thế nào để chọn loại biểu đồ phù hợp với đặc điểm của dữ liệu (ví dụ: dữ liệu phân loại, dữ liệu số, dữ liệu thời gian)?**

* Dữ liệu số liên tục: histogram, KDE plot, boxplot, scatter plot.
* Dữ liệu phân loại: bar chart, countplot, mosaic plot.
* So sánh giữa phân loại và số: bar chart cho tổng/mean, boxplot để so sánh phân phối.
* Dữ liệu chuỗi thời gian: line chart, area chart, scatter có trục thời gian.
* Nhiều biến số: pair plot, heatmap tương quan, ma trận scatter.

### **4. Sự khác biệt giữa các thư viện trực quan hóa trong Python như Matplotlib, Seaborn và Plotly là gì?**

* Matplotlib: thư viện nền tảng, nhiều tùy chỉnh chi tiết, nhưng cú pháp dài; phù hợp báo cáo nghiên cứu.
* Seaborn: xây dựng trên Matplotlib, cú pháp ngắn gọn, đẹp mặc định, mạnh cho biểu đồ thống kê.
* Plotly: tạo biểu đồ tương tác (zoom, hover, filter), dễ xuất HTML, thích hợp dashboard và trình bày web.

### **5. Những nguyên tắc thiết kế nào cần tuân thủ để tạo ra một biểu đồ trực quan hóa dễ hiểu và hiệu quả?**

* Mỗi biểu đồ phải trả lời một câu hỏi cụ thể, tránh dàn trải.
* Có tiêu đề, nhãn trục, đơn vị, chú giải rõ ràng.
* Dùng tỷ lệ trục đúng, tránh gây hiểu nhầm; cân nhắc log-scale nếu dữ liệu lệch mạnh.
* Hạn chế màu sắc và hiệu ứng thừa, tránh 3D không cần thiết.
* Giữ sự nhất quán về font, màu sắc, định dạng.
* Nhấn mạnh insight chính bằng annotation, màu sắc nổi bật.

### **6. Làm thế nào để tạo một biểu đồ đơn giản như histogram hoặc bar chart bằng Matplotlib? Bạn có thể chia sẻ đoạn code mẫu không?**

import numpy as np

import matplotlib.pyplot as plt

# Dữ liệu giả định

np.random.seed(42)

scores = np.random.normal(loc=7.0, scale=1.5, size=300) # điểm số

classes = ["A", "B", "C", "D"]

counts = [35, 80, 60, 25]

# Histogram

plt.figure()

plt.hist(scores, bins=15, edgecolor="black")

plt.title("Phân phối điểm số")

plt.xlabel("Điểm")

plt.ylabel("Tần suất")

plt.tight\_layout()

plt.show()

# Bar chart

plt.figure()

plt.bar(classes, counts)

plt.title("Số lượng học sinh theo lớp")

plt.xlabel("Lớp")

plt.ylabel("Số lượng")

plt.tight\_layout()

plt.show()

### **7. Làm thế nào để xuất biểu đồ từ Python ra các định dạng như PNG, PDF hoặc HTML để sử dụng trong báo cáo?**

* Xuất bằng Matplotlib:

plt.savefig("chart.png", dpi=300, bbox\_inches="tight") # PNG

plt.savefig("chart.pdf", bbox\_inches="tight") # PDF

plt.savefig("chart.svg", bbox\_inches="tight") # SVG

* Xuất HTML tương tác với Plotly:

import plotly.express as px

fig = px.histogram(x=scores, nbins=15, title="Phân phối điểm số")

fig.write\_html("chart.html", include\_plotlyjs="cdn") # mở trong trình duyệt

1.3. Phân tích đơn biến và hai biến

## 1.3.1. Ôn lý thuyết

### **1. Phân tích đơn biến (Univariate analysis) là gì? Khác gì với phân tích hai biến (Bivariate analysis)?**

* Univariate analysis: chỉ tập trung phân tích một biến duy nhất. Mục đích là hiểu:  
  + Dữ liệu phân bố thế nào (phân phối chuẩn, lệch trái, lệch phải).
  + Độ tập trung quanh giá trị nào (mean, median, mode).
  + Độ phân tán, outlier.
  + Biểu đồ thường dùng: histogram, bar chart, boxplot.
* Bivariate analysis: nghiên cứu mối quan hệ giữa hai biến (numeric–numeric, numeric–categorical, categorical–categorical).  
  + Ví dụ: chiều cao và cân nặng (numeric–numeric), điểm toán theo giới tính (numeric–categorical).
  + Biểu đồ thường dùng: scatter plot, heatmap (numeric–numeric), boxplot/violin (numeric–categorical), bảng chéo/stacked bar (categorical–categorical).

Khác biệt cốt lõi:

* Univariate: “biến này trông như thế nào?”
* Bivariate: “biến A có liên quan thế nào đến biến B?”

### **2. Các thước đo thống kê trong phân tích đơn biến**

* Xu hướng trung tâm (central tendency):  
  + Mean (trung bình): giá trị trung bình cộng.
  + Median (trung vị): giá trị ở giữa khi sắp xếp dữ liệu.
  + Mode (mode): giá trị xuất hiện nhiều nhất.
* Độ phân tán (dispersion):  
  + Range = max – min.
  + Variance (phương sai): mức độ phân tán trung bình quanh mean.
  + Standard deviation (độ lệch chuẩn) = căn bậc hai variance.
  + IQR (Interquartile Range) = Q3 – Q1.
* Hình dạng phân phối (distribution shape):  
  + Skewness: độ lệch trái/phải.
  + Kurtosis: độ nhọn của phân phối.
* Phát hiện ngoại lệ (outliers):  
  + Quy tắc Tukey: điểm ≤ Q1 – 1.5×IQR hoặc ≥ Q3 + 1.5×IQR

### **3. Trong phân tích hai biến, làm thế nào để xác định mối quan hệ giữa hai biến?**

Tuỳ loại dữ liệu, ta có cách khác nhau:

* Numeric – Numeric:  
  + Dùng correlation coefficient (Pearson cho tuyến tính, Spearman cho quan hệ đơn điệu).
  + Biểu đồ: scatter plot, regression line.
  + Nếu cần suy luận nguyên nhân: phải dùng thêm thí nghiệm hoặc mô hình nhân quả.
* Numeric – Categorical:  
  + So sánh trung bình giữa các nhóm (t-test, ANOVA).
  + Biểu đồ: boxplot, violin plot.
* Categorical – Categorical:  
  + Dùng Chi-square test để kiểm tra độc lập.
  + Biểu đồ: bảng chéo (cross-tab), stacked bar chart.

### 4. Sự khác biệt giữa Correlation và Covariance

* Covariance:  
  + Đo lường hai biến thay đổi cùng nhau như thế nào.
  + Nếu >0: biến cùng tăng/giảm. Nếu <0: một tăng, một giảm.
  + Có đơn vị (phụ thuộc vào scale của biến).
  + Giá trị không giới hạn.
* Correlation:  
  + Chuẩn hoá từ covariance → giá trị nằm trong [-1, 1].
  + Không có đơn vị.
  + Dễ so sánh giữa các cặp biến khác nhau.

### **5. Khi nào nên sử dụng biểu đồ trực quan hóa trong phân tích đơn biến so với phân tích hai biến?**

* Phân tích Đơn biến:
  + Sử dụng để hiểu phân phối của dữ liệu số( numerical), ví dụ: histogram, boxplot/violin plot)
  + Hiểu tần suất của dữ liệu phân loại( categorical), ví dụ: bar chart, pie-chart
* Phân tích Hai biến:
  + Sử dụng để khám phá mối quan hệ giữa hai biến.
  + Ví dụ:  
     Scatter plot cho hai biến số; Boxplot hoặc Violin plot để so sánh một biến số giữa các nhóm phân loại.

### **6. Đoạn code mẫu để tạo biểu đồ scatter plot hoặc heatmap để phân tích mối quan hệ giữa hai biến?**

import seaborn as sns

import matplotlib.pyplot as plt

iris = sns.load\_dataset("iris")

# Scatter plot

sns.scatterplot(data=iris, x="sepal\_length", y="petal\_length", hue="species")

plt.title("Scatter: Sepal vs Petal Length")

plt.show()

# Heatmap

corr = iris.drop(columns=["species"]).corr()

sns.heatmap(corr, annot=True, cmap="coolwarm", square=True)

plt.title("Correlation Heatmap")

plt.show()

### **7. Làm thế nào để trực quan hóa mối quan hệ giữa một biến số và một biến phân loại bằng biểu đồ boxplot hoặc violin plot trong Python?**

* Khi có một biến số (numeric) và một biến phân loại (categorical), ta thường dùng:  
  + Boxplot: hiển thị trung vị, tứ phân vị (Q1, Q3), phạm vi giá trị, và các outlier của biến số theo từng nhóm phân loại.
  + Violin plot: tương tự boxplot nhưng còn biểu diễn thêm đường cong mật độ (density) của biến số trong từng nhóm.