1. Phân tích đơn biến (univariate analysis) là gì? Nó khác gì với phân tích hai biến (bivariate analysis) trong khám phá dữ liệu?

**\*Phân tích đơn biến (Univariate Analysis)**

* **Định nghĩa**: Là quá trình phân tích một biến (một cột dữ liệu) tại một thời điểm.
* **Mục tiêu**: Hiểu rõ đặc điểm phân phối, xu hướng trung tâm, độ phân tán, và hình dạng phân phối của biến đó.
* **Kỹ thuật thường dùng**:
  + Với **biến định lượng (số)**: dùng thống kê mô tả (mean, median, variance, std), histogram, boxplot.
  + Với **biến định tính (categorical)**: dùng tần suất (frequency), tỷ lệ phần trăm, bar chart, pie chart.

Ví dụ: Phân tích điểm thi Toán của học sinh trong lớp để xem phân phối điểm, điểm trung bình, mức độ lệch chuẩn, số học sinh đạt điểm cao/thấp.

**\*Phân tích hai biến (Bivariate Analysis)**

* **Định nghĩa**: Là quá trình phân tích mối quan hệ giữa **hai biến** trong dữ liệu.
* **Mục tiêu**: Xác định xem hai biến có mối liên hệ, tương quan, hoặc phụ thuộc lẫn nhau không.
* **Kỹ thuật thường dùng**:
  + **Biến định lượng – định lượng**: scatter plot, hệ số tương quan (correlation).
  + **Biến định tính – định tính**: bảng chéo (crosstab), kiểm định Chi-square, stacked bar chart.
  + **Định lượng – định tính**: boxplot theo nhóm, group mean, ANOVA.
* Ví dụ: Phân tích mối quan hệ giữa **giờ học thêm** (biến định lượng) và **điểm thi Toán** (biến định lượng) để xem có tương quan dương không. Hoặc phân tích **giới tính** (biến định tính) và **kết quả đậu/rớt** (biến định tính).

**SO SÁNH**

| **Tiêu chí** | **Phân tích đơn biến** | **Phân tích hai biến** |
| --- | --- | --- |
| **Số biến** | 1 | 2 |
| **Mục tiêu** | Hiểu đặc điểm và phân phối của biến | Tìm mối quan hệ giữa hai biến |
| **Ví dụ** | Phân phối điểm thi Toán | Quan hệ giữa giờ học thêm và điểm thi Toán |

2. Các thước đo thống kê nào thường được sử dụng trong phân tích đơn biến (ví dụ: trung bình, trung vị, mode, độ lệch chuẩn)?

**a. Thước đo xu hướng trung tâm (Measures of Central Tendency)**

Dùng để chỉ ra giá trị “điển hình” của dữ liệu.

* **Trung bình (Mean)**: Giá trị trung bình cộng.
* **Trung vị (Median)**: Giá trị nằm giữa tập dữ liệu (khi sắp xếp).
* **Mode (Giá trị thường gặp nhất)**: Giá trị xuất hiện nhiều nhất trong dữ liệu.

*Ví dụ*: Với điểm thi [5, 7, 7, 8, 10]:

* Mean = 7.4
* Median = 7
* Mode = 7

**b. Thước đo độ phân tán (Measures of Dispersion)**

Cho biết dữ liệu **phân tán nhiều hay ít** quanh trung tâm.

* **Phạm vi (Range)** = Max – Min
* **Phương sai (Variance)**: Độ chênh lệch bình phương so với trung bình.
* **Độ lệch chuẩn (Standard Deviation)**: Căn bậc hai của phương sai → dễ hiểu hơn phương sai.
* **IQR (Interquartile Range)**: Khoảng tứ phân vị (Q3 – Q1), giúp phát hiện dữ liệu lệch hoặc ngoại lệ.

**c. Thước đo hình dạng phân phối (Shape of Distribution)**

Cho biết dữ liệu có đối xứng hay lệch.

* **Skewness (Độ lệch)**:
  + Skew > 0: lệch phải (đuôi dài bên phải).
  + Skew < 0: lệch trái (đuôi dài bên trái).
* **Kurtosis (Độ nhọn)**:
  + Kurtosis cao → phân phối nhọn (nhiều dữ liệu tập trung).
  + Kurtosis thấp → phân phối bẹt.

**d. Thống kê tần suất (Frequency Statistics)**

Thường áp dụng cho biến phân loại (categorical).

* **Tần suất (Frequency)**: Số lần xuất hiện của mỗi giá trị.
* **Tỷ lệ phần trăm (%)**: Tần suất chia tổng số quan sát.

Ví dụ: Trong lớp có 40 sinh viên: 25 nữ (62.5%), 15 nam (37.5%).

3. Trong phân tích hai biến, làm thế nào để xác định mối quan hệ giữa hai biến (ví dụ: tương quan, nhân quả)?

**I. Xác định mối quan hệ qua thống kê và trực quan hóa**

Tùy loại biến, ta có cách khác nhau:

**a) Cả hai biến định lượng (numeric – numeric)**

* **Scatter plot (biểu đồ phân tán)** → xem dữ liệu có xu hướng tuyến tính hay không.
* **Hệ số tương quan Pearson (r)**: đo mức độ tuyến tính (từ –1 đến +1).
  + r gần +1 → tương quan dương mạnh.
  + r gần –1 → tương quan âm mạnh.
  + r gần 0 → ít hoặc không tuyến tính.
* **Spearman rank correlation**: dùng khi dữ liệu không tuyến tính nhưng có quan hệ theo thứ hạng.

Ví dụ: **giờ học thêm** ↔ **điểm thi Toán**.

**b) Biến định tính – định tính (categorical – categorical)**

* **Bảng chéo (crosstab/contingency table)** để xem phân bố.
* **Kiểm định Chi-square** để kiểm tra hai biến có độc lập hay không.
* **Tỷ lệ phần trăm so sánh** giữa các nhóm.

Ví dụ: **giới tính** ↔ **đậu/rớt**.

**c) Định tính – định lượng (categorical – numeric)**

* **So sánh trung bình theo nhóm** (group mean).
* **Boxplot** để thấy sự khác biệt phân phối.
* **Kiểm định t-test hoặc ANOVA** để xem sự khác biệt có ý nghĩa thống kê.

Ví dụ: **ngành học** ↔ **điểm GPA trung bình**.

**II. Phân biệt tương quan và nhân quả**

* **Tương quan (Correlation)**: hai biến có mối liên hệ (cùng tăng/giảm, hoặc thay đổi cùng nhau).
* **Nhân quả (Causation)**: biến A **tác động trực tiếp** làm thay đổi biến B.

Lưu ý:

* “Có tương quan không có nghĩa là có nhân quả”.
* Ví dụ:
  + **Tích cực giả**: Số lượng kem bán ra và số ca đuối nước đều tăng vào mùa hè → tương quan, nhưng không phải nhân quả.
  + **Nhân quả thực**: Số giờ học tăng → điểm số có thể tăng.

**III. Cách kiểm tra nhân quả**

Để xác định nhân quả, không chỉ dựa vào thống kê đơn giản mà cần:

* **Thí nghiệm có kiểm soát (Controlled experiments)** → thay đổi biến độc lập và đo biến phụ thuộc.
* **Phân tích hồi quy (Regression analysis)** → kiểm tra tác động của một biến lên biến khác khi kiểm soát các yếu tố khác.
* **Phương pháp thống kê nâng cao**: Granger causality (trong chuỗi thời gian), mô hình SEM (Structural Equation Modeling), v.v.

4. Sự khác biệt giữa tương quan (correlation) và hiệp biến (covariance) trong phân tích hai biến là gì?

**1. Hiệp biến (Covariance)**

* **Định nghĩa**: Đo mức độ **hai biến thay đổi cùng nhau**.
* **Công thức**: A black background with white letters and numbers

  AI-generated content may be incorrect.
* **Ý nghĩa**:
  + Cov(X,Y) > 0 → Khi X tăng thì Y cũng có xu hướng tăng (quan hệ đồng biến).
  + Cov(X,Y) < 0 → Khi X tăng thì Y có xu hướng giảm (quan hệ nghịch biến).
  + Cov(X,Y) ≈ 0 → Ít hoặc không có mối quan hệ tuyến tính.

Nhược điểm: Giá trị hiệp biến **phụ thuộc đơn vị đo lường** → khó so sánh.  
Ví dụ: Nếu X là “cân nặng (kg)” và Y là “chiều cao (cm)”, Cov có đơn vị **kg·cm**, nên khó diễn giải trực tiếp.

**2. Tương quan (Correlation)**

* **Định nghĩa**: Là dạng **chuẩn hóa** của hiệp biến → để loại bỏ ảnh hưởng của đơn vị đo.
* **Công thức**: A math equation with white text

  AI-generated content may be incorrect.

Trong đó: σX,σY\sigma\_X, \sigma\_YσX​,σY​ là độ lệch chuẩn của X và Y.

* **Ý nghĩa**:
  + Corr(X,Y) ∈ [–1, 1].
  + Corr gần +1 → tương quan dương mạnh.
  + Corr gần –1 → tương quan âm mạnh.
  + Corr ≈ 0 → không có quan hệ tuyến tính.

Ưu điểm: Không phụ thuộc đơn vị đo, dễ so sánh.

| **Tiêu chí** | **Hiệp biến (Covariance)** | **Tương quan (Correlation)** |
| --- | --- | --- |
| **Phạm vi giá trị** | (–∞ , +∞) | [–1 , +1] |
| **Đơn vị** | Phụ thuộc vào đơn vị của biến (vd: kg·cm) | Không có đơn vị (chuẩn hóa) |
| **Ý nghĩa** | Chỉ cho biết chiều quan hệ (cùng chiều/ ngược) | Cho biết cả chiều và mức độ quan hệ |
| **Dễ so sánh** | Khó so sánh giữa các cặp biến khác nhau | Dễ so sánh do chuẩn hóa |

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

**A. Trong phân tích đơn biến (Univariate Analysis)**

Mục tiêu: Hiểu **đặc điểm phân phối** của một biến (một cột dữ liệu).

* **Khi nên dùng biểu đồ**:
  + Muốn xem **phân phối dữ liệu** (chuẩn, lệch trái/phải, nhiều đỉnh).
  + Muốn phát hiện **ngoại lệ (outlier)**.
  + Muốn so sánh **tần suất xuất hiện** của các giá trị.
* **Biểu đồ thường dùng**:
  + **Histogram** → phân phối của biến số (numeric).
  + **Boxplot** → trung vị, tứ phân vị, ngoại lệ.
  + **Bar chart / Pie chart** → biến phân loại (categorical).

Ví dụ: Vẽ histogram điểm thi Toán để xem nhiều học sinh đạt mức điểm nào.

**B. Trong phân tích hai biến (Bivariate Analysis)**

Mục tiêu: Hiểu **mối quan hệ giữa hai biến**.

* **Khi nên dùng biểu đồ**:
  + Muốn xem **xu hướng tuyến tính hay phi tuyến** giữa hai biến số.
  + Muốn so sánh **sự khác biệt giữa các nhóm**.
  + Muốn phát hiện **mối quan hệ mạnh/yếu, cùng chiều/ngược chiều**.
* **Biểu đồ thường dùng**:
  + **Scatter plot** → hai biến số (numeric-numeric).
  + **Line chart** → biến liên tục theo thời gian.
  + **Boxplot theo nhóm** → numeric vs categorical.
  + **Stacked bar chart / Mosaic plot** → categorical vs categorical.

Ví dụ: Vẽ scatter plot giữa **giờ học thêm** và **điểm thi** để xem có xu hướng tăng theo nhau không.

| **Loại phân tích** | **Khi nào cần trực quan hóa** | **Biểu đồ phù hợp** |
| --- | --- | --- |
| **Đơn biến** | Hiểu phân phối, phát hiện outlier, tần suất | Histogram, Boxplot, Bar chart, Pie chart |
| **Hai biến** | Hiểu mối quan hệ giữa hai biến | Scatter plot, Line chart, Boxplot nhóm, Stacked bar chart |

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

|  |  |
| --- | --- |
| scatter plot | heatmap |
| import matplotlib.pyplot as plt  import seaborn as sns  import pandas as pd  # Ví dụ: dataset nhỏ  data = pd.DataFrame({  "hours\_study": [2, 3, 4, 5, 6, 7, 8],  "exam\_score": [50, 55, 60, 65, 70, 75, 80]  })  # Vẽ scatter plot  plt.figure(figsize=(6,4))  sns.scatterplot(x="hours\_study", y="exam\_score", data=data, color="blue", s=80)  plt.title("Scatter Plot: Hours of Study vs Exam Score")  plt.xlabel("Hours of Study")  plt.ylabel("Exam Score")  plt.show() | # Dataset mẫu  df = pd.DataFrame({  "math\_score": [65, 70, 75, 80, 85, 90],  "reading\_score": [60, 68, 74, 82, 86, 92],  "writing\_score": [62, 72, 78, 84, 88, 94]  })  # Tính ma trận tương quan  corr = df.corr()  # Vẽ heatmap  plt.figure(figsize=(6,4))  sns.heatmap(corr, annot=True, cmap="coolwarm", fmt=".2f")  plt.title("Heatmap: Correlation Matrix")  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?