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

**1.1. Phân tích đơn biến (Univariate Analysis)**

Định nghĩa: Quá trình phân tích một biến duy nhất để hiểu đặc điểm, phân bố, và xu hướng của nó mà không liên quan đến các biến khác.

Mục đích: Xác định phân bố (chuẩn, lệch), phát hiện ngoại lai, tính các thống kê cơ bản.

Ví dụ: Phân tích cột "Income" trong marketing\_campaign.csv để xem phân bố thu nhập.

**1.2. Phân tích hai biến (Bivariate Analysis)**

Định nghĩa: Quá trình phân tích mối quan hệ giữa hai biến để khám phá xu hướng hoặc tương quan.

Mục đích: Đánh giá mức độ liên kết (dương, âm, hoặc không có) giữa hai biến.

Ví dụ: Phân tích mối quan hệ giữa "Income" và "MntWines" để xem thu nhập ảnh hưởng đến chi tiêu cho rượu.

**1.3. Sự khác biệt**

Số lượng biến: Đơn biến (1 biến), hai biến (2 biến).

Mục tiêu: Đơn biến mô tả, hai biến khám phá mối quan hệ.

Phương pháp: Đơn biến dùng histogram/boxplot; hai biến dùng scatter plot/heatmap.

## 2. Các thước đo thống kê nào thường được sử dụng trong phân tích đơn biến?

* **Trung bình (Mean)**: .Nhạy với ngoại lai (ví dụ: trung bình "Income").
* **Trung vị (Median)**: Giá trị giữa khi sắp xếp dữ liệu, không bị ảnh hưởng bởi ngoại lai (ví dụ: trung vị "Income").
* **Mode (Mốt)**: Giá trị xuất hiện nhiều nhất (ví dụ: mode của "Education").
* **Độ lệch chuẩn (Standard Deviation)**: . Đo phân tán (ví dụ: độ lệch chuẩn "MntWines").
* **Phạm vi (Range)**: Range= max(x) - min(x) (ví dụ: phạm vi "culmen\_length\_mm").
* **Tứ phân vị (Quartiles) và IQR**: Q1, Q2, Q3; IQR = Q3 - Q1 (ví dụ: IQR "new\_cases").
* **Phân vị (Percentiles)**: Giá trị dưới một phần trăm nhất định (ví dụ: phân vị 75 của "Income").

**Ứng dụng**: Sử dụng pandas (.mean(), .median(), .std()) hoặc numpy/scipy.stats như trong tài liệu (trang 2).

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

### 3.1. Tương quan (Correlation)

* **Định nghĩa**: Đo lường mức độ tuyến tính (hệ số Pearson: -1 đến +1).
* **Công thức**: wps
* **Ví dụ**: Tương quan giữa "Income" và "MntWines".

### 3.2. Hiệp biến (Covariance)

* **Định nghĩa**: Đo hướng và cường độ mối quan hệ tuyến tính, phụ thuộc đơn vị.
* **Công thức**: wps​.
* **Ví dụ**: Hiệp biến giữa "culmen\_length\_mm" và "body\_mass\_g".

### 3.3. Phân tích nhân quả (Causality)

* **Cách xác định**: Không thể từ dữ liệu quan sát (chỉ tương quan); cần thử nghiệm hoặc mô hình nhân quả (DAG).
* **Ví dụ**: "Income" cao có dẫn đến "MntWines" cao không? Cần kiểm tra yếu tố khác.

**Phương pháp**: Biểu đồ (scatter), thống kê (tương quan Pearson), mô hình (hồi quy).

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

* **Định nghĩa**:
  + **Hiệp biến**: Đo hướng và cường độ, giá trị phụ thuộc đơn vị.
  + **Tương quan**: Chuẩn hóa hiệp biến, giá trị từ -1 đến +1.
* **Công thức**:
  + Covariance: wps
  + Correlation:wps
* **Sự khác biệt**:
  + **Quy mô**: Hiệp biến thay đổi theo đơn vị (ví dụ: 500,000); tương quan cố định [-1, 1].
  + **So sánh**: Tương quan dễ so sánh giữa cặp biến; hiệp biến không.
  + **Dễ hiểu**: Tương quan (0.7) rõ ràng hơn hiệp biến.
* **Ví dụ**: Cov(Income, MntWines) = 500,000, r = 0.6.

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

### 5.1. Phân tích đơn biến

* **Khi nào dùng**: Để hiểu phân bố, xu hướng trung tâm, hoặc ngoại lai của một biến.
* **Biểu đồ**:
  + **Histogram**: Phân bố (ví dụ: "Income").
  + **Boxplot**: Ngoại lai và tứ phân vị (ví dụ: "new\_cases").
  + **Pie chart/Bar chart**: Biến phân loại (ví dụ: "species").

### 5.2. Phân tích hai biến

* **Khi nào dùng**: Để khám phá mối quan hệ hoặc xu hướng giữa hai biến.
* **Biểu đồ**:
  + **Scatter plot**: Mối quan hệ tuyến tính (ví dụ: "Income" vs "MntWines").
  + **Heatmap**: Ma trận tương quan.
  + **Boxplot/Violin với hue**: So sánh nhóm (ví dụ: "Income" theo "Education").

### 5.3. Quy tắc

* Đơn biến mô tả, hai biến liên kết. Chọn dựa trên loại biến và câu hỏi.

## 6. Đoạn code mẫu để tạo biểu đồ scatter plot hoặc heatmap?

### 6.1. Scatter Plot

**Python:**

**import pandas as pd**

**import seaborn as sns**

**import matplotlib.pyplot as plt**

**marketing\_data = pd.read\_csv("marketing\_campaign.csv", skipinitialspace=True)**

**marketing\_data["Income"] = pd.to\_numeric(marketing\_data["Income"], errors='coerce')**

**marketing\_data["Income"] = marketing\_data["Income"].fillna(marketing\_data["Income"].median())**

**plt.figure(figsize=(10, 6))**

**sns.scatterplot(data=marketing\_data, x="Income", y="MntWines", hue="Education")**

**plt.title("Mối quan hệ giữa Income và MntWines theo Education")**

**plt.xlabel("Thu nhập (Income)")**

**plt.ylabel("Chi tiêu cho rượu (MntWines)")**

**plt.legend(title="Trình độ học vấn")**

**plt.show()**

**6.2. Heatmap**

**Python:**

**plt.figure(figsize=(10, 8))**

**numeric\_data = marketing\_data.select\_dtypes(include=['int64', 'float64'])**

**correlation\_matrix = numeric\_data.corr()**

**sns.heatmap(correlation\_matrix, annot=True, cmap="coolwarm", vmin=-1, vmax=1)**

**plt.title("Ma trận tương quan giữa các biến số")**

**plt.show()**

## 7. Làm thế nào để trực quan hóa mối quan hệ giữa một biến số và một biến phân loại?

### 7.1. Boxplot

**Python:**

**plt.figure(figsize=(12, 6))**

**sns.boxplot(data=marketing\_data, x="Education", y="Income")**

**plt.title("Phân bố thu nhập theo trình độ học vấn")**

**plt.xlabel("Trình độ học vấn")**

**plt.ylabel("Thu nhập (Income)")**

**plt.xticks(rotation=45)**

**plt.show()**

**7.2. Violin Plot**

**Python:**

**plt.figure(figsize=(12, 6))**

**sns.violinplot(data=marketing\_data, x="Education", y="Income")**

**plt.title("Phân bố thu nhập theo trình độ học vấn (Violin Plot)")**

**plt.xlabel("Trình độ học vấn")**

**plt.ylabel("Thu nhập (Income)")**

**plt.xticks(rotation=45)**

**plt.show()**