# 1.3.1 Câu hỏi lý thuyết

## + 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** là quá trình khám phá và tìm hiểu **từng thuộc tính (biến) trong tập dữ liệu một cách độc lập**. Mục tiêu chính là để hiểu rõ hình dạng và đặc điểm phân phối của mỗi biến mà không xét đến mối quan hệ của nó với các biến khác.

Sự khác biệt với Phân tích Hai biến (Bivariate Analysis): **Phân tích hai biến** tập trung vào **mối quan hệ giữa hai biến và cách chúng thay đổi cùng nhau**. Mục tiêu là để xem liệu có sự tương tác hay liên hệ nào giữa một cặp thuộc tính hay không.

Bảng tóm tắt sự khác biệt chính:

|  |  |  |
| --- | --- | --- |
| Tiêu chí | Phân tích đơn biến | Phân tích hai biến |
| Số lượng biến | Phân tích **một biến** tại một thời điểm | Phân tích **hai biến** tại một thời điểm |
| Mục tiêu chính | **Mô tả** đặc điểm của một biến duy nhất: phân phối, xu hướng trung tâm, độ phân tán, độ lệch, giá trị ngoại lai… | **Khám phá mối liên hệ** giữa hai biến: tương quan thuận, tương quan nghịch, hoặc không có tương quan |
| Ví dụ | Thuộc tính *tuổi* được phân bổ như thế nào? hoặc Giá trị trung bình của thuộc tính *khối lượng* là bao nhiêu? | Thuộc tính bệnh tiểu đường và *số lần mang thai* có liên quan đến nhau không? |

## + 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)?

Phân tích đơn biến tập trung vào việc tìm hiểu từng thuộc tính một cách độc lập. Các thước đo thống kê mô tả là công cụ chính để có được cái nhìn sâu sắc về hình dạng của mỗi thuộc tính. Chúng ta sẽ có 2 nhóm thước đo chính:

1. **thống kê mô tả:**

* **Count (Số lượng):** Tổng số lượng quan sát (hàng) không bị thiếu dữ liệu.
* **Mean (Giá trị trung bình):** Giá trị trung bình cộng của tất cả các quan sát. Đây là một trong những thước đo bạn đã đề cập.
* **Standard Deviation (Độ lệch chuẩn):** Một thước đo cho thấy mức độ phân tán hoặc lan rộng của dữ liệu so với giá trị trung bình. Đây cũng là một thước đo bạn đã đề cập.
* **Minimum Value (Giá trị nhỏ nhất):** Giá trị nhỏ nhất trong tập dữ liệu của thuộc tính đó.
* **25th Percentile (Phân vị thứ 25):** Giá trị mà tại đó 25% dữ liệu có giá trị nhỏ hơn hoặc bằng nó.
* **50th Percentile (Median - Trung vị):** Đây chính là **trung vị**, là giá trị nằm ở giữa của tập dữ liệu đã được sắp xếp. 50% dữ liệu có giá trị nhỏ hơn hoặc bằng giá trị này. Đây là thước đo thứ ba bạn đã đề cập.
* **75th Percentile (Phân vị thứ 75):** Giá trị mà tại đó 75% dữ liệu có giá trị nhỏ hơn hoặc bằng nó.
* **Maximum Value (Giá trị lớn nhất):** Giá trị lớn nhất trong tập dữ liệu của thuộc tính đó.

Bảng kết quả thống kê mô tả trong tài liệu là một ví dụ rõ ràng về cách áp dụng các thước đo này cho từng thuộc tính trong bộ dữ liệu về bệnh tiểu đường Pima.

1. **Độ lệch của Phân phối (Skewness)**

Đây là một thước đo thống kê quan trọng khác được sử dụng trong phân tích đơn biến để hiểu về hình dạng phân phối của một thuộc tính.

* **Skew (Độ lệch):** Thước đo này cho biết một phân phối có bị lệch hay bị "bóp méo" về một phía so với phân phối chuẩn (hình chuông) hay không.
* **Ý nghĩa:** Nhiều thuật toán học máy hoạt động tốt hơn khi các biến đầu vào có phân phối chuẩn (Gaussian). Việc biết một thuộc tính có bị lệch hay không cho phép chúng ta thực hiện các bước tiền xử lý dữ liệu để cải thiện độ chính xác của mô hình.
* **Cách đọc giá trị:** Các giá trị gần bằng 0 cho thấy độ lệch ít. Giá trị dương cho thấy phân phối lệch sang phải, và giá trị âm cho thấy phân phối lệch sang trái.
* Về các ví dụ bạn đã đưa ra:
* **Trung bình (Mean):** Là một thước đo thống kê cơ bản được liệt kê trong phần thống kê mô tả.
* **Mode:** Là giá trị xuất hiện nhiều nhất trong dữ liệu. Nó thuộc nhóm thước đo xu hướng trung tâm (measures of central tendency).
* **Trung vị (Median):** Là Phân vị thứ 50 (50th Percentile) và cũng là một thước đo quan trọng. Các biểu đồ hộp (Box plots) cũng trực quan hóa giá trị trung vị này một cách rõ ràng.
* **Độ lệch chuẩn (Standard Deviation):** Có, đây là một thước đo độ phân tán chính được sử dụng.

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

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

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

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

## + 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?