**1.1. THỐNG KÊ MÔ TẢ**

**1.1.1. Ôn tập lý thuyết:**

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

*\*Thống kê mô tả là:*

Về bản chất, **thống kê mô tả** được sử dụng để báo cáo hoặc mô tả các tính năng hoặc đặc điểm của dữ liệu. Chúng tóm tắt một tập dữ liệu số cụ thể, hoặc nhiều tập dữ liệu, và cung cấp thông tin chi tiết định lượng về dữ liệu đó thông qua biểu diễn số hoặc đồ họa.

Thống kê mô tả chỉ phản ánh dữ liệu mà chúng được áp dụng. Một thống kê mô tả có thể là:

* **Một thước đo xu hướng trung tâm, như trung bình, trung vị hoặc mốt:** Chúng được sử dụng để xác định điểm trung bình hoặc điểm trung tâm trong một tập dữ liệu
* **Một thước đo về độ phân tán hoặc độ biến thiên, như phương sai, độ lệch chuẩn, độ lệch hoặc phạm vi:** Những thước đo này phản ánh sự phân tán của các điểm dữ liệu
* **Một thước đo phân phối, như số lượng hoặc phần trăm của một kết quả cụ thể:** Chúng thể hiện tần suất của kết quả đó trong một tập dữ liệu

Dữ liệu thô có thể được biểu diễn dưới dạng số liệu thống kê và đồ thị, sử dụng các hình ảnh trực quan như biểu đồ hình tròn, biểu đồ đường, bảng và các cách biểu diễn khác tóm tắt dữ liệu thu thập được về một nhóm dân số nhất định.

*\*Thống kê suy luận là:*

**Thống kê suy luận** được dùng để đưa ra kết luận hoặc suy diễn về quần thể dựa trên dữ liệu mẫu. Bằng cách phân tích một mẫu ngẫu nhiên đại diện, người nghiên cứu có thể khái quát kết quả cho toàn bộ quần thể. Chất lượng và tính đại diện của mẫu là yếu tố then chốt để đảm bảo độ tin cậy của suy luận.

Các kỹ thuật phổ biến trong thống kê suy luận gồm:

* **Kiểm định giả thuyết:** xác định một kết quả có ý nghĩa thống kê hay chỉ do ngẫu nhiên.
* **Phân tích tương quan:** đánh giá mối quan hệ giữa các biến.
* **Hồi quy tuyến tính hoặc logistic:** dự đoán và phân tích mối quan hệ nhân quả.
* **Khoảng tin cậy:** ước lượng tham số quần thể trong một khoảng xác suất nhất định.

*\*Sự khác nhau giữa Thống kê mô tả và Thống kê suy luận:*

|  |  |  |
| --- | --- | --- |
| **Tiêu chí** | **Thống kê mô tả** | **Thống kê suy luận** |
| **Mục đích** | Tóm tắt, mô tả đặc điểm của dữ liệu hiện có. | Suy luận, dự đoán, kiểm định cho toàn bộ quần thể dựa trên dữ liệu mẫu. |
| **Phạm vi áp dụng** | |  | | --- | | Chỉ phản ánh tập dữ liệu đang xét. |  |  | | --- | |  | | |  | | --- | | Khái quát từ mẫu sang quần thể rộng hơn. |  |  | | --- | |  | |
| **Độ chắc chắn** | |  | | --- | | Cao, vì dựa trực tiếp vào dữ liệu thực tế. |  |  | | --- | |  | | |  | | --- | | Thấp hơn, luôn có sai số và khoảng tin cậy. |  |  | | --- | |  | |
| **Độ phức tạp** | Tính toán đơn giản (mean, median, variance, tần suất, biểu đồ…). | Tính toán phức tạp hơn (ước lượng tham số, kiểm định giả thuyết, hồi quy, khoảng tin cậy…). |
| **Ví dụ (Thử nghiệm thuốc)** | Tỷ lệ bệnh nhân gặp tác dụng phụ, độ tuổi trung bình trong mẫu. | Ước lượng khả năng xảy ra tác dụng phụ trên toàn bộ quần thể bệnh nhân. |
| **Bản chất** | |  | | --- | | Báo cáo các sự kiện, kết quả đã có. |  |  | | --- | |  | | |  | | --- | |  |  |  | | --- | | Kiểm định giả thuyết, phân tích mối quan hệ, đưa ra dự đoán. | |

***+ 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 chính thường được sử dụng gồm:*

* **Trung bình (Mean):** cho biết giá trị trung tâm của dữ liệu, phù hợp khi phân phối dữ liệu đối xứng và ít có ngoại lệ.
* **Trung vị (Median):** thể hiện giá trị nằm giữa tập dữ liệu đã sắp xếp, thường được ưu tiên khi dữ liệu bị **lệch (skewed)** hoặc có nhiều **giá trị ngoại lai (outliers)**, vì trung vị ít bị ảnh hưởng bởi các giá trị cực đoan.
* **Phương sai (Variance) và Độ lệch chuẩn (Standard Deviation):** đo mức độ phân tán của dữ liệu quanh giá trị trung bình, phản ánh mức biến thiên trong tập dữ liệu.

*\*Khi nào nên dùng trung vị thay vì trung bình?*

Nên dùng **trung vị** khi dữ liệu có **sự lệch (skewness)** hoặc tồn tại **ngoại lệ (outliers)**. Lý do là trung bình bị kéo lệch mạnh bởi các giá trị cực đoan, trong khi trung vị phản ánh trung tâm một cách “ổn định” hơn.

***+ 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)?***

*\*Làm thế nào để xác định phân bố của một tập dữ liệu?*

* **Trực quan hóa dữ liệu:** sử dụng biểu đồ tần suất (histogram), biểu đồ mật độ (density plot) hoặc Q-Q plot để so sánh với phân bố chuẩn**.**
* **Thống kê mô tả:** tính các thước đo như độ lệch (skewness) và độ nhọn (kurtosis) để đánh giá mức độ lệch và độ “dày” đuôi phân bố.
* **Kiểm định thống kê:** áp dụng các kiểm định như Shapiro-Wilk, Kolmogorov-Smirnov hoặc Anderson-Darling để kiểm tra dữ liệu có tuân theo phân bố chuẩn hay không.

*\*Các loại phân bố phổ biến:*

* **Phân bố chuẩn (Normal distribution):** dữ liệu đối xứng, hình chuông, trung bình ≈ trung vị ≈ mốt.
* **Phân bố lệch trái (Left-skewed / Negative skew):** đuôi kéo dài về bên trái, trung bình < trung vị.
* **Phân bố lệch phải (Right-skewed / Positive skew):** đuôi kéo dài về bên phải, trung bình > trung vị.
* **Một số phân bố khác thường gặp:** phân bố nhị thức, phân bố Poisson, phân bố đồng đều… (thường dùng trong xác suất và mô hình thống kê).

***+ Độ lệch chuẩn và phạm vi (range) có ý nghĩa gì trong việc đánh giá sự phân tán của dữ liệu?***

*\*Độ lệch chuẩn (Standard Deviation – SD):*

* Là thước đo mức độ phân tán trung bình của các giá trị so với trung bình cộng.
* **Ý nghĩa:**

+ Độ lệch chuẩn nhỏ → dữ liệu tập trung gần giá trị trung bình, ít biến động.

+ Độ lệch chuẩn lớn → dữ liệu phân tán rộng, biến động cao.

* **Ví dụ:** Trong điểm thi của một lớp, nếu SD thấp thì hầu hết học sinh có điểm gần giống nhau; nếu SD cao thì điểm dao động mạnh, có nhiều học sinh giỏi lẫn yếu.

*\*Phạm vi (Range):*

* Là chênh lệch giữa giá trị lớn nhất và nhỏ nhất trong tập dữ liệu.
* **Ý nghĩa:** cho ta biết mức độ trải rộng của dữ liệu, nhưng **nhạy cảm với ngoại lệ (outliers)**.
* **Ví dụ:** Nếu điểm thi dao động từ 2 đến 9 → Range = 7. Chỉ số này cho ta biết độ rộng phân bố, nhưng không cho biết dữ liệu phân bố trong khoảng đó như thế nào.

*\*So sánh & Vai trò:*

* **Phạm vi**: đơn giản, dễ tính, phù hợp để nhanh chóng mô tả độ rộng dữ liệu.
* **Độ lệch chuẩn**: chính xác hơn, phản ánh sự phân tán của toàn bộ dữ liệu quanh trung bình nên thường được ưu tiên trong phân tích thống kê.

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

*\*Các thước đo Q1, Q2, Q3 là:*

* **Q1 (Quartile 1 – tứ phân vị thứ nhất):** giá trị tại vị trí 25% dữ liệu nhỏ nhất → cho biết 25% dữ liệu nằm dưới giá trị này.
* **Q2 (Quartile 2 – trung vị/median):** giá trị tại vị trí 50% dữ liệu → chia tập dữ liệu thành hai nửa bằng nhau.
* **Q3 (Quartile 3 – tứ phân vị thứ ba):** giá trị tại vị trí 75% dữ liệu nhỏ nhất → cho biết 75% dữ liệu nằm dưới giá trị này.

*\*Sự khác biệt:*

* Q1, Q2, Q3 chia dữ liệu thành 4 phần bằng nhau (mỗi phần 25%).
* Q2 chính là **trung vị**, trong khi Q1 và Q3 cho thấy “ngưỡng thấp” và “ngưỡng cao” của dữ liệu.
* Khoảng giữa Q1 và Q3 gọi là **IQR (Interquartile Range)**, phản ánh độ phân tán trung tâm và được dùng để phát hiện ngoại lệ (outliers).

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

Giá trị thiếu (missing values) là vấn đề phổ biến trong dữ liệu thực tế. Trước khi tính toán các chỉ số thống kê mô tả (mean, median, variance...), cần xử lý chúng để tránh kết quả sai lệch.

*\*Cách xử lý phổ biến:*

* **Loại bỏ (Deletion):**

*+ Xóa hàng (row deletion):* loại bỏ các quan sát chứa giá trị thiếu.

*+ Xóa cột (column deletion):* loại bỏ biến có quá nhiều giá trị thiếu.

+ Thích hợp khi dữ liệu thiếu ít và không ảnh hưởng nhiều đến tập dữ liệu.

* **Thay thế (Imputation):**

*+ Sử dụng trung bình/median/mode:* ví dụ thay thế điểm thiếu bằng **trung vị** để giảm ảnh hưởng của ngoại lệ.

*+ Dự đoán (predictive imputation):* dùng mô hình (như KNN imputation, regression imputation) để dự đoán giá trị thiếu.

* **Giữ nguyên:**

+ Một số mô hình và thư viện thống kê/ML có thể xử lý missing values trực tiếp (ví dụ: cây quyết định có thể bỏ qua một phần dữ liệu thiếu).

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

*\*Histogram (biểu đồ tần suất):*

* Histogram cho thấy **phân bố dữ liệu** bằng cách chia dữ liệu thành các khoảng (bins) rồi đếm số lượng phần tử rơi vào từng khoảng.
* Khi đọc histogram, ta quan sát:

**+ Hình dạng phân bố**: chuẩn (chuông), lệch trái, lệch phải.

**+ Độ phân tán**: dữ liệu trải rộng hay tập trung.

**+ Đỉnh (mode)**: khoảng giá trị xuất hiện nhiều nhất.

*\*Boxplot (biểu đồ hộp):*

* Boxplot tóm tắt dữ liệu dựa trên **5 số**: min, Q1, median (Q2), Q3, max.
* Khi đọc boxplot, ta chú ý:

**+ Chiều dài hộp**: khoảng giữa Q1 và Q3 (IQR), cho thấy mức độ biến thiên của dữ liệu.

**+ Đường ở giữa hộp**: median (Q2), cho biết giá trị trung vị.

**+ Râu (whiskers)**: thể hiện phạm vi dữ liệu (không tính ngoại lệ).

**+ Dấu chấm ngoài râu**: các điểm ngoại lệ (outliers).

***+ 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 kiểm tra outlier**

* Vẽ đồ thị (boxplot, scatter plot, histogram) để phát hiện điểm bất thường.
* Dùng các tiêu chí như **Z-score** hoặc **IQR (Interquartile Range)** để đánh dấu outlier:

+ Giá trị < Q1 – 1.5 × IQR hoặc > Q3 + 1.5 × IQR → được xem là ngoại lai.

**2. Điều tra nguyên nhân**

* Có thể do **lỗi nhập liệu, đo đạc sai** → nên loại bỏ.
* Nếu là **giá trị hợp lệ nhưng hiếm gặp**, cần cân nhắc giữ lại để không mất thông tin.

**3. Phương pháp xử lý phổ biến trước khi thống kê mô tả**

* **Loại bỏ (trimming):** Xóa dữ liệu ngoại lai nếu chắc chắn là lỗi.
* **Winsorizing / Clipping:** Giới hạn giá trị ngoại lai tại một ngưỡng.
* **Imputation:** Thay thế bằng trung vị, trung bình hoặc giá trị hợp lý khác.
* **Biến đổi dữ liệu:** Dùng log, sqrt, Box-Cox để giảm ảnh hưởng outlier.
* **Dùng chỉ số “robust”:** Thay vì mean/SD, dùng median, IQR, MAD.

**4. Báo cáo kết quả rõ ràng**

Trong nghiên cứu, nên nêu rõ tiêu chí loại bỏ outlier, bao nhiêu quan sát bị loại, và ảnh hưởng đến thống kê.

**1.2. XỬ LÝ VÀ TRỰC QUAN HÓA DỮ LIỆU  
1.2.1. Ôn tập lý thuyết  
-------------------------------------------------------------------------------------------------------------------------------------  
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)?**

- Vai trò của trực quan hóa dữ liệu trong phân tích dữ liệu, bao gồm:  
 + Dễ dàng nắm bắt thông tin: hình ảnh trực quan dễ hiểu và ghi nhớ hơn so với bảng số liệu khô khan.

+ Phát hiện xu hướng và mẫu hình: Biểu đồ và đồ thị giúp dễ nhận diện các xu hướng, mẫu hình ẩn sâu trong dữ liệu

+ Xác định điểm bất thường: Các điểm dữ liệu ngoại lệ (outliers) dễ dàng được phát hiện thông qua hình ảnh trực quan.

+ Truyền đạt thông tin hiệu quả: Hình ảnh trực quan giúp truyền đạt thông tin phức tạp một cách rõ ràng và thuyết phục đến nhiều đối tượng khác nhau.

+ Hỗ trợ đưa ra quyết định: Dựa trên những thông tin được trực quan hóa, người ra quyết định có thể đưa ra những lựa chọn chính xác và kịp thời

+ Tiết kiệm thời gian: Thay vì phân tích hàng loạt con số, người dùng có thể nhanh chóng hiểu được thông tin thông qua hình ảnh trực quan.

+ Tăng cường khả năng tương tác: Nhiều công cụ trực quan hóa dữ liệu cho phép người dùng tương tác với dữ liệu, khám phá thông tin theo nhiều góc độ khác nhau.

- Khám phá dữ liệu (EDA) là giai đoạn đầu tiên trong phân tích dữ liệu, mục tiêu là hiểu dữ liệu, phát hiện vấn đề, đưa ra giả thuyết**,** nơi nhà phân tích "trò chuyện" với dữ liệu để hiểu được đặc tính, cấu trúc và những manh mối ban đầu của nó. Vì vậy, trực quan hóa dữ liệu không chỉ là công cụ hỗ trợ mà còn là nền tảng cốt lõi trong EDA, giúp hiểu dữ liệu một cách toàn diện, phát hiện các đặc điểm quan trọng và khai thác giá trị thực sự từ dữ liệu.

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

\*Histogram:

- Là biểu đồ tần suất, biểu đồ cột dọc liên tục. Mục đích phân bố của một biến số liên tục.

- Biểu đồ sử dụng trong các trường hợp:

+ Xem dữ liệu có phân bố chuẩn (Normal distribution) hay không.

+ Xác định độ lệch (Skewed) của dữ liệu.

+ Phát hiện giá trị ngoại lai (outliers).

+ So sánh sự phân bố của biến giữa các nhóm (dùng histogram chồng).

+ Nhận biết khoảng giá trị nào xuất hiện phổ biến nhất.

\*Scatter plot:

- Là biểu đồ phân tán dùng để khám phá mối **quan hệ** giữa **hai biến số liên tục.**

- Trường hợp sử dụng:

+ Nhận biết mối tương quan (correlation): tích cực (lên dốc), tiêu cực (xuống dốc), hoặc không có.

+ Xác định xu hướng và cường độ của mối quan hệ.

+ Phát hiện các cụm (clusters) dữ liệu.

+ Nhận diện các điểm ngoại lai (outliers) rõ ràng.

\*Boxplot:

- Là biểu đồ hộp dùng để tóm tắt phân phối của một biến số và so sánh phân phối đó giữa các nhóm danh mục khác nhau.

- Trường hợp sử dụng:

+ So sánh trung vị (median) và độ phân tán (spread) giữa nhiều nhóm.

+ Nhận diện nhanh các giá trị ngoại lai (outliers) - được vẽ thành các chấm riêng lẻ.

+ Hiểu được sự đối xứng của dữ liệu.

\*Bar chart:

- Là biểu đồ cột để  so sánh các giá trị của một biến danh mục (categorical variable). - Trường hợp sử dụng:  
 + So sánh số lượng hoặc tần suất giữa các danh mục khác nhau.

+ Hiển thị số liệu thống kê tóm tắt (như tổng, trung bình) cho từng nhóm.

**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)?**

Để chọn loại biểu đồ phù hợp với đặc điểm của dữ liệu cần dựa vào mục tiêu phân tích và đặc điểm dữ liệu:

- Dữ liệu phân loại: dùng để biểu diễn các nhóm hoặc danh mục (ví dụ: giới tính, nghề nghiệp, loại sản phẩm). Các biểu đồ phù hợp gồm **bar chart** để so sánh tần suất hoặc tỷ lệ giữa các nhóm, và **pie chart** để thể hiện tỷ lệ phần trăm của từng danh mục.

- Dữ liệu số: là các giá trị liên tục hoặc rời rạc (ví dụ: chiều cao, cân nặng, điểm số). Có thể dùng **histogram** để quan sát phân bố, **boxplot** để so sánh trung vị, độ phân tán và phát hiện ngoại lệ, hoặc **scatter plot** để phân tích mối quan hệ giữa hai biến số.

- Dữ liệu theo thời gian: là các dữ liệu có tính tuần tự theo mốc thời gian (ví dụ: doanh thu theo tháng, giá cổ phiếu theo ngày). Biểu đồ phù hợp nhất là **line chart** để thể hiện xu hướng, ngoài ra có thể dùng **area chart** để biểu diễn giá trị tích lũy theo thời gian.

- Dữ liệu đa biến: Khi muốn xem xét đồng thời nhiều biến (ví dụ: mối quan hệ giữa chiều cao, cân nặng và tuổi). Có thể sử dụng **heatmap** để biểu diễn ma trận tương quan, **pairplot** để quan sát mối quan hệ giữa nhiều biến số, hoặc **bubble chart** để thêm thông tin về kích thước/khối lượng.

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Thư viện** | Matplotlib | Seaborn | Plotly |
| **Mục tiêu chính** | Tạo biểu đồ từ cơ bản → nâng cao, kiểm soát chi tiết từng thành phần. | Tập trung trực quan hóa thống kê với giao diện đẹp, dễ dùng. | Trực quan hóa tương tác, hiện đại, hỗ trợ web/dashboard. |
| **Mức độ tùy chỉnh** | Rất cao (kiểm soát trục, màu, font, lưới...). | Trung bình (dựa vào Matplotlib, ít tự do hơn). | Cao nhưng hướng tới trực quan hóa tương tác, không chi tiết bằng Matplotlib. |
| **Độ phức tạp khi code** | Dài dòng, cần nhiều dòng lệnh cho biểu đồ đẹp. | Ngắn gọn, chỉ vài dòng lệnh đã có biểu đồ đẹp. | Ngắn gọn, dễ tạo biểu đồ tương tác phức tạp. |
| **Biểu đồ mặc định** | Khá đơn giản, “thô”, phải chỉnh tay nhiều. | Đẹp, trực quan, phù hợp ngay cho báo cáo thống kê. | Hiện đại, bắt mắt, có tương tác (zoom, hover, xoay 3D). |
| **Loại biểu đồ nổi bật** | Line, bar, scatter, histogram (cơ bản). | Boxplot, violin, heatmap, pairplot (thống kê nâng cao). | Biểu đồ 3D, map, gauge, sankey, biểu đồ tương tác. |
| **Tích hợp với Pandas** | Hỗ trợ nhưng phải gọi .plot() hoặc chuyển tay. | Tích hợp trực tiếp, dễ vẽ từ DataFrame. | Tích hợp tốt, đặc biệt khi làm dashboard với Dash. |
| **Tương tác** | Không có (chỉ ảnh tĩnh). | Không có (chỉ ảnh tĩnh). | Có (zoom, hover, lọc dữ liệu). |
| **Ứng dụng phù hợp** | Khi cần kiểm soát chi tiết và xuất file chất lượng cao. | Khi làm EDA và phân tích dữ liệu thống kê. | Khi cần báo cáo tương tác, dashboard hoặc web app. |
| **Hiệu năng** | Xử lý nhanh với dữ liệu nhỏ/trung bình, nhưng khi vẽ nhiều điểm (hàng triệu) thì hơi chậm. | Phụ thuộc vào Matplotlib, tốc độ tương tự nhưng có thể chậm hơn khi vẽ nhiều biểu đồ thống kê phức tạp. | Khi dữ liệu lớn có thể nặng, vì phải render HTML/JS để hỗ trợ tương tác. |
| **Khả năng xuất biểu đồ** | Xuất sang PNG, JPG, PDF, SVG chất lượng cao (rất hay dùng cho in ấn/bài báo). | Xuất tương tự Matplotlib vì nó dựa trên Matplotlib. | Xuất sang HTML (giữ được tương tác), PNG, PDF nhưng đôi khi cần thêm thư viện hỗ trợ (kaleido, orca). |

**Trong Python có nhiều thư viện trực quan hóa dữ liệu, trong đó Matplotlib, Seaborn và Plotly là ba thư viện phổ biến nhất. Mỗi thư viện có đặc điểm và ưu điểm khác nhau:**

**\***Matplotlib:

- Là thư viện nền tảng, mạnh mẽ và linh hoạt nhất, cho phép tạo hầu hết các loại biểu đồ từ cơ bản đến nâng cao.

- Ưu điểm: Kiểm soát chi tiết từng thành phần của biểu đồ, hỗ trợ nhiều định dạng xuất (PNG, PDF, SVG…)

- Nhược điểm: Biểu đồ mặc định đơn giản, muốn đẹp phải tùy chỉnh nhiều.

- Thích hợp khi cần tùy chỉnh sâu hoặc xuất bản biểu đồ.

\*Seaborn:

- Được xây dựng trên Matplotlib, cung cấp API cấp cao, dễ sử dụng và hỗ trợ trực tiếp Pandas DataFrame.

- Ưu điểm: Biểu đồ mặc định đẹp hơn, có sẵn nhiều biểu đồ thống kê (boxplot, violin plot, heatmap, pairplot…).

- Nhược điểm: Linh hoạt kém hơn Matplotlib khi cần can thiệp chi tiết.

- Thích hợp cho phân tích dữ liệu khám phá (EDA) và trực quan hóa thống kê nhanh chóng.

\*Plotly:

- Tập trung vào trực quan hóa tương tác. Biểu đồ có thể zoom, xoay 3D, hover để hiển thị chi tiết.

- Ưu điểm: Biểu đồ hiện đại, sinh động, dễ tích hợp vào báo cáo, dashboard hoặc web app.

- Nhược điểm: Ít tùy chỉnh chi tiết hơn Matplotlib và có thể nặng khi dữ liệu lớn.

- Thích hợp cho dashboard và báo cáo tương tác.

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

- **Xác định đúng mục tiêu biểu đồ:** Trước khi vẽ, cần làm rõ câu hỏi “muốn truyền tải thông tin gì? cho ai xem?” từ đó chọn loại biểu đồ phù hợp.

- **Chọn loại biểu đồ đúng với đặc điểm dữ liệu:**

+ Dữ liệu phân loại (categorical): dùng bar chart, pie chart.

+ Dữ liệu số (numerical): dùng histogram, boxplot, scatter plot.

+ Dữ liệu thời gian (time series): dùng line chart.

- **Đơn giản và rõ ràng:** Tránh nhồi nhét quá nhiều thông tin trong một biểu đồ. Loại bỏ yếu tố trang trí thừa (chartjunk), chỉ giữ lại những gì cần thiết.

- **Sử dụng màu sắc hợp lý:** Dùng màu sắc để phân biệt nhóm hoặc nhấn mạnh thông tin quan trọng. Không dùng quá nhiều màu hoặc màu tương phản mạnh gây rối mắt.

- **Đảm bảo tính nhất quán:** Giữ cùng một tông màu, thang đo, ký hiệu cho các biểu đồ trong cùng một báo cáo để dễ so sánh.

- **Hiển thị đầy đủ nhãn và đơn vị:** Có tiêu đề, nhãn trục (axis labels), chú thích (legend) rõ ràng. Ghi rõ đơn vị đo lường để tránh hiểu nhầm.

- **Nhấn mạnh điểm quan trọng:** Dùng màu nổi bật, annotation hoặc đánh dấu (highlight) để người xem nhanh chóng nhận ra insight chính.

- **Đảm bảo tính chính xác:** Không cắt trục hoặc bóp méo tỷ lệ gây sai lệch nhận thức. Biểu đồ phải phản ánh đúng bản chất dữ liệu.

**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?**

- Để tạo ra một biểu đồ đơn giản như trước hết **import** thư viện histogram hoặc Matplotlib và sau đó chuẩn bị tập dữ liệu cho trực quan hóa dữ liệu

Ví dụ: Dữ liệu lấy từ Pima Indians Diabetes



- Tiếp theo dùng hàm **figure()** để tạo khung vẽ và hàm **bar()** để tạo biểu cột (bar chart). Sau đó dùng hàm **show()** để hiển thị biểu đồ với tập dữ liệu số người mắc bệnh tiểu đường ở nữA screenshot of a computer screen

AI-generated content may be incorrect.

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

**\*** Lưu biểu đồ với **savefig()** của thư viện **Matplotlib:** savefig() (hoặc Figure.savefig()) **lưu ảnh của figure hiện tại ra file**. Tên file quyết định định dạng (ví dụ "hinh.png" → PNG, "hinh.pdf" → PDF).

\* **Xuất ra HTML:**

- Nếu chỉ cần chèn ảnh tĩnh vào báo cáo HTML, ta dùng savefig("hinh.png") rồi gắn vào HTML bằng thẻ <img src="hinh.png">.

- Nếu muốn biểu đồ **tương tác** trong HTML, ta dùng các thư viện hỗ trợ như:

+ **Plotly:** fig.write\_html("bieudo.html") → tạo file HTML có thể mở trực tiếp bằng trình duyệt, trong đó biểu đồ có thể zoom, xoay, hover để xem chi tiết.

+ **Bokeh:** output\_file("bieudo.html") và save(fig) → cũng xuất biểu đồ tương tác ra file HTML.

1. 3. Phân tích đơn biến và hai biến

1.3.1. Ôn 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?
  + Đơn biến là gì? Đơn biến là một thuật ngữ thường được sử dụng trong thống kê để mô tả một loại dữ liệu bao gồm các quan sát chỉ về một đặc điểm hoặc một thuộc tính duy nhất. Giống như tất cả các dữ liệu khác, dữ liệu đơn biến có thể được trực quan hóa bằng biểu đồ, hình ảnh, hoặc các công cụ phân tích khác sau khi dữ liệu được đo lường.
  + Phân tích đơn biến là hình thức phân tích dữ liệu đơn giản nhất. Uni nghĩa là “một”, vì vậy dữ liệu chỉ có một biến (univariate). Dữ liệu đơn biến yêu cầu phân tích từng biến riêng biệt.
  + Thông thường phân tích đơn biến có 2 mục đích. Đầu tiên là trả lời một câu hỏi nghiên cứu bằng nghiên cứu mô tả và mục đích thứ hai là tìm hiểu về cách thuộc tính thay đổi theo tác động riêng lẻ của một biến trong phân tích hồi quy.
  + Phân tích hai biến là một trong những dạng phân tích định lượng (thống kê) đơn giản nhất. Nó bao gồm việc phân tích hai biến (thường được ký hiệu là X, Y), nhằm mục đích xác định mối quan hệ thực nghiệm giữa chúng.
  + Sự khác nhau giữa phân tích đơn biến và phân tích hai biến:

|  |  |  |
| --- | --- | --- |
| **Tiêu chí** | **Univariate** | **Bivariate** |
| Số lượng biến phân tích | Chỉ tập trung vào 1 biến duy nhất | Xem xét mối quan hệ giữa hai biến |
| Mục tiêu | Hiểu phân bố, xu hướng trung tâm, độ phân tán của biến | Tìm ra sự liên hệ, tương quan hoặc ảnh hưởng giữa hai biến |
| Thước đo thống kê | Mean, Median, Mode, Variance, Std, Range, IQR | Correlation, Covariance, kiểm định giả thuyết (t-test, ANOVA, Chi-square |
| Biểu đồ trực quan | Histogram, bar chart, boxplot, density plot | Scatter plot, heatmap, boxplot (khi một biến phân loại), violin plot |
| Kết quả phân tích | Cho biết đặc điểm riêng lẻ của biến | Cho biết mối quan hệ giữa hai biến (tuyến tính, độc lập, phân biệt nhóm,...) |
| Ví dụ | Phân tích cột Age để biết tuổi trung bình của khách hàng | Phân tích Age và Income để xem tuổi có liên quan đến thu nhập không |

⇒ Phân tích đơn biến đóng vai trò như bước khởi đầu trong quá trình khám phá dữ liệu, giúp hiểu rõ đặc điểm cơ bản của từng biến trước khi đi sâu vào mối quan hệ phức tạp hơn. Còn phân tích 2 biến đi xa hơn một bước khi xem xét mối quan hệ giữa hai biến.

* 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 (univariate analysis) tập trung mô tả đặc điểm của một biến. Các thước đo thường dùng gồm:

* + Xu hướng trung tâm (Central Tendency):
    - **Trung bình (Mean):** giá trị trung bình cộng.
    - **Trung vị (Median):** giá trị ở giữa khi sắp xếp dữ liệu.
    - **Mode:** giá trị xuất hiện nhiều nhất.
  + Độ phân tán (Dispersion):
    - **Phương sai (Variance):** mức độ phân tán của dữ liệu so với trung bình.
    - **Độ lệch chuẩn (Standard Deviation):** căn bậc hai của phương sai, cho biết mức dao động dữ liệu.
    - **Khoảng (Range):** max – min.
    - **IQR (Interquartile Range):** khoảng giữa Q1 và Q3.
* Trong phân tích hai biến, làm thế nào để xác định mối quan hệ giữa hai biến (ví dụ: tương quan, nhân quả)?

Khi tiến hành phân tích hai biến, mục tiêu chính là tìm ra xem dữ liệu hai biến có mối quan hệ với nhau hay không, và nếu có thì mối quan hệ đó mạnh hay yếu, tuyến tính hay phi tuyến.

1. Tương quan (Correlation)
   * Đây là thước đo phổ biến nhất để xác định mức độ liên hệ tuyến tính giữa hai biến.
   * Hệ số tương quan Pearson (Pearson’s correlation coefficient, ký hiệu là r) có giá trị từ -1 đến 1:
     + r gần 1 ⟶ mối quan hệ tuyến tính dương mạnh (một biến tăng thì biến kia cũng tăng).
     + r gần -1 ⟶ mối quan hệ tuyến tính âm mạnh (một biến tăng thì biến kia giảm).
     + r gần 0 ⟶ ít hoặc không có mối quan hệ tuyến tính.
   * Ngoài Pearson, với dữ liệu không tuyến tính hoặc dạng phân hạng, ta có thể dùng Spearman hoặc Kendall.
2. Hiệp biến (Covariance)
   * Đo mức độ hai biến thay đổi cùng nhau, nhưng không chuẩn hóa về thang đo.
   * Nếu giá trị dương ⟶ hai biến có xu hướng cùng tăng/giảm.
   * Nếu giá trị âm ⟶ khi một biến tăng, biến kia có xu hướng giảm.
   * Tuy nhiên, khó so sánh vì giá trị phụ thuộc vào đơn vị đo ⟶ do đó hệ số tương quan thường được ưu tiên hơn.
3. Nhân quả (Causality)
   * Khác với tương quan chỉ mô tả “ hai biến có liên hệ”, nhân quả trả lời câu hỏi “biến này có gây ra thay đổi ở biến kia không”.
   * Để xác định nhân quả, chỉ phân tích thống kê là chưa đủ. Cần:
     + Thiết kế thí nghiệm (controlled experiments).
     + Phân tích chuỗi thời gian (time series analysis, Granger causality).
     + Áp dụng các phương pháp học máy/causal inference (ví dụ: DAG – Directed Acylic Graphs).
   * Một câu nổi tiếng trong thống kê là: “Correlation does not causation” (tương quan không có nghĩa là nhân quả).
4. Kết hợp trực quan hóa
   * Ngoài các thước đo, biểu đồ cũng là công cụ hữu ích:
     + **Scatter plot (biểu đồ phân tán):** quan sát xu hướng tuyến tính hay phi tuyến giữa hai biến số.
     + **Heatmap (bản đồ nhiệt):** hiển thị ma trận tương quan giữa nhiều biến, giúp so sánh nhanh.

Kết luận: Để xác định mối quan hệ giữa hai biến, ta thường bắt đầu bằng việc tính toán các thước đo như tương quan hoặc hiệp biến, kết hợp với biểu đồ trực quan hóa. Tuy nhiên, nếu muốn khẳng định mối quan hệ nhân quả, cần nhiều bằng chứng và phương pháp phân tích sâu hơn, chứ không thể chỉ dựa vào tương quan đơn thuần.

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

Khi phân tích mối quan hệ giữa hai biến, hai khái niệm thường được nhắc đến là hiệp biến (covariance) và tương quan (correlation). Cả hai đều cho ta biết liệu hai biến có thay đổi cùng nhau không, nhưng cách diễn giải và phạm vi sử dụng lại khác nhau.

1. Hiệp biến (Covariance):
   * Đo lường mức độ và chiều hướng hai biến thay đổi cùng nhau.
   * Nếu giá trị hiệp biến dương ⟶ hai biến có xu hướng tăng hoặc giảm cùng nhau.
   * Nếu giá trị hiệp biến âm ⟶khi một biến tăng thì biến kia giảm.
   * Điểm hạn chế: giá trị hiệp biến phụ thuộc vào đơn vị đo lường của biến, nên khó so sánh giữa các cặp biến khác nhau.
2. Tương quan (Correlation):
   * Là phiên bản chuẩn hóa của hiệp biến, giúp loại bỏ ảnh hưởng của đơn vị đo.
   * Có giá trị nằm trong khoảng [-1, 1], dễ dàng so sánh mức độ mạnh/yếu của mối quan hệ.
   * Hệ số tương quan càng gần ±1 ⟶ mối quan hệ tuyến tính càng mạnh.
   * Do có chuẩn hóa, tương quan được sử dụng phổ biến hơn trong phân tích dữ liệu và học máy.

Bảng so sánh:

|  |  |  |
| --- | --- | --- |
| **Đặc điểm** | **Hiệp biến (Covariance)** | **Tương quan (Correlation)** |
| Ý nghĩa | Đo lường mức độ và chiều hướng biến đổi cùng nhau | Đo mức độ tuyến tính giữa hai biến |
| Giá trị | (-∞, +∞) | [-1, 1] |
| Phụ thuộc đơn vị đo | Có (ví dụ: cm, kg, s...) | Không (chuẩn hóa) |
| Dễ so sánh giữa nhiều biến | Khó | Dễ dàng |
| Mức phổ biến trong thực tế | Ít được dùng trực tiếp | Được dùng nhiều (Pearson, Spearman, Kendall) |

Kết luận: Nói cách khác, tương quan chính là hiệp biến đã được chuẩn hóa. Nếu muốn biết hai biến có cùng xu hướng biến động hay không thì dùng hiệp biến, nhưng nếu muốn đánh giá mức độ mạnh yếu và dễ so sánh, ta dùng tương quan.

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

Trong phân tích dữ liệu, biểu đồ trực quan hóa không chỉ giúp ta mô tả dữ liệu một cách trực quan mà còn làm nổi bật những đặc điểm quan trọng mà con số thống kê thuần túy có thể che lấp. Tùy theo mục tiêu phân tích, ta sẽ chọn loại biểu đồ phù hợp.

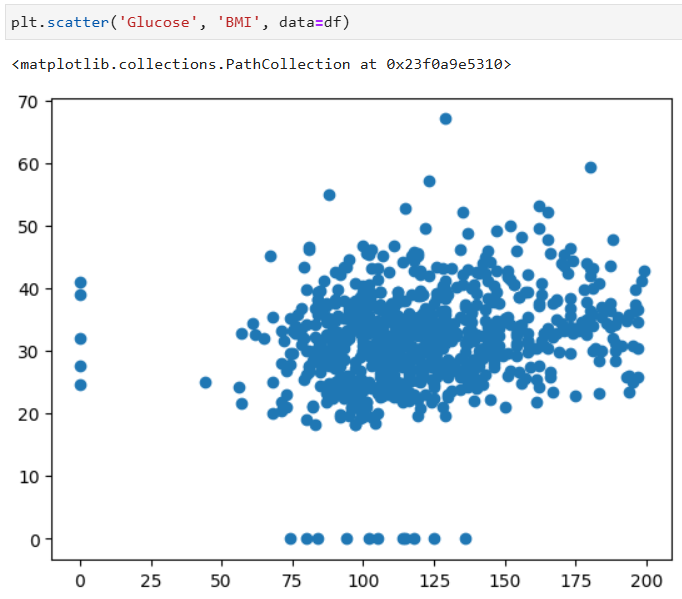
1. Trong phân tích đơn biến (Univaruate analysis)
   * Mục tiêu là khám phá đặc điểm phân bố của một biến duy nhất.
   * Các biểu đồ thường dùng:
     + **Histogram:** cho thấy tần suất xuất hiện của các giá trị, đặc biệt hữu ích với biến số liên tục.
     + **Bar chart:** phù hợp cho biến phân loại (categorical).
     + **Boxplot:** giúp phát hiện phân vị, trung vị, ngoại lệ (outlier).
   * Sử dụng khi ta muốn trả lời câu hỏi: “Biến này có phân bố ra sao? Có lệch về một phía hay không? Có giá trị bất thường không?”.
2. Trong phân tích hai biến (Bivariate analysis)
   * Mục tiêu là tìm hiểu mối quan hệ giữa hai biến.
   * Các biểu đồ thường dùng:
     + **Scatter plot:** quan sát mối liên hệ tuyến tính/phi tuyến giữa hai biến số.
     + **Heatmap:** biểu diễn ma trận tương quan, đặc biệt hữu ích khi có nhiều biến số cần so sánh.
     + **Boxplot/Violin plot** (nhóm theo biến phân loại): quan sát phân bố của một biến số trong từng nhóm của biến phân loại.
   * Sử dụng khi ta muốn trả lời câu hỏi: “Hai biến này có mối liên hệ với nhau không? Nếu có, dạng quan hệ đó là gì?”.

Kết luận: Biểu đồ trong phân tích đơn biến giúp hiểu sâu về một biến riêng lẻ, còn biểu đồ trong phân tích hai biến giúp làm rõ mối quan hệ giữa các biến. Việc lựa chọn loại biểu đồ nào phụ thuộc vào bản chất dữ liệu (số hay phân loại) và mục tiêu phân tích mà ta đang theo đuổi.

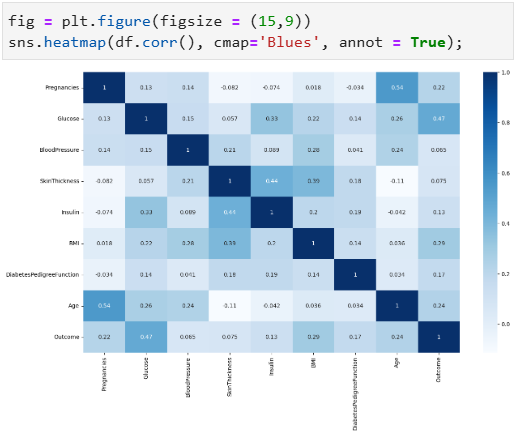
* Đoạn code mẫu để tạo biểu đồ scatter plot hoặc heatmap để phân tích mối quan hệ giữa hai biến?

Dữ liệu mẫu là dữ liệu của lab4 (pima-indians-diabetes). Cần import các thư viện là seaborn và matplotlib.

1. Code tạo biểu đồ Scatter plot:



1. Code tạo heatmap:



* 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 dữ liệu bao gồm một biến số (numerical variable) và một biến phân loại (categorical variable), ta thường muốn biết phân bố của biến số thay đổi như thế nào giữa các nhóm được định nghĩa bởi biến phân loại. Ví dụ mức lương (biến số) giữa các ngành nghề (biến phân loại) hay điểm số (biến số) giữa các lớp học (biến phân loại).

1. Boxplot (biểu đồ hộp)
   * Mô tả dữ liệu bằng các giá trị thống kê quan trọng: trung vị (median), tứ phân vị (quartiles) và ngoại lệ (outliers).
   * Giúp nhanh chóng so sánh sự khác biệt về phân bố giữa các nhóm.
2. Violin plot
   * Kết hợp ưu điểm của boxplot và phân bố mật độ (kernel density).
   * Cho thấy rõ ràng hơn hình dạng phân bố dữ liệu của từng nhóm.
   * Phù hợp khi ta muốn không chỉ thấy trung vị, mà còn quan sát sự “trải dài” hay “tập trung” của dữ liệu trong mỗi nhóm.
3. Ví dụ minh họa trong Python:
   * Code biểu đồ Boxplot:

A screenshot of a graph

AI-generated content may be incorrect.

* + Code biểu đồ Violin plot:

A screenshot of a computer screen

AI-generated content may be incorrect.

1. Kết luận:
   * Khi muốn thấy sự so sánh gọn gàng, tập trung vào trung vị và độ phân tán, hãy dùng boxplot.

Khi muốn thấy rõ hơn hình dạng phân bố dữ liệu trong từng nhóm, hãy dùng violin plot.