**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?

1. **Phân tích đơn biến (univariate analysis)**

Định nghĩa:  phân tích một biến (variable) duy nhất tại một thời điểm.

Mục đích chính:

* Hiểu rõ đặc điểm, tính chất và phân phối của từng biến riêng lẻ.
* Mô tả dữ liệu một cách tổng quan.
* Tìm ra các giá trị bất thường (outliers) hoặc lỗi trong dữ liệu.

Các kỹ thuật và biểu đồ thường dùng:

* Với biến Định lượng (Numerical):
  + Thống kê mô tả: Trung bình (mean), trung vị (median), độ lệch chuẩn (standard deviation), min, max...
  + Biểu đồ: Histogram, Box plot, KDE plot.
* Với biến Định tính (Categorical):
  + Thống kê mô tả: Tần số (frequency), tỷ lệ phần trăm (percentage), mode (giá trị xuất hiện nhiều nhất).
  + Biểu đồ: Bar chart, Pie chart.

1. **Phân tích đa biến (bivariate analysis)**

Định nghĩa: là kỹ thuật phân tích mối quan hệ giữa hai biến khác nhau.

Mục đích chính:

* Tìm kiếm, đo lường và diễn giải mối quan hệ, liên kết hoặc sự khác biệt giữa hai biến.
* Trả lời các câu hỏi như "Biến A có liên quan đến Biến B không?", "Khi A thay đổi thì B thay đổi như thế nào?".

Các kỹ thuật và biểu đồ thường dùng (tùy thuộc vào loại biến):

* Hai biến Định lượng (Numerical - Numerical):
  + Biểu đồ: Scatter plot (biểu đồ phân tán).
  + Thước đo: Hệ số tương quan (Correlation coefficient).
* Một biến Định tính & Một biến Định lượng (Categorical - Numerical):
  + Biểu đồ: Box plot (phân nhóm), Bar chart (với giá trị trung bình), Violin plot.
  + Thước đo: So sánh trung bình (mean) giữa các nhóm.
* Hai biến Định tính (Categorical - Categorical):
  + Biểu đồ: Stacked bar chart, Heatmap (của bảng chéo).
  + Thước đo: Bảng chéo (Crosstab), Kiểm định Chi-Square.

\* Sự khác biệt chính giữa phân tích đơn biến và hai biến

|  |  |  |
| --- | --- | --- |
| Đặc điểm | Phân tích Đơn biến | Phân tích Hai biến |
| **Số lượng biến** | Một biến | Hai biến |
| **Mục đích** | Mô tả, tóm tắt đặc điểm của một biến | Tìm hiểu mối quan hệ giữa hai biến |
| **Câu hỏi trả lời** | "Dữ liệu của biến này trông như thế nào?" | "Có mối liên hệ gì giữa biến A và biến B không?" |
| **Kỹ thuật phổ biến** | Thống kê mô tả, Histogram, Bar chart | Scatter plot, Box plot phân nhóm, Correlation, Crosstab |
| **Ví dụ** | Phân phối thu nhập | Mối quan hệ giữa thu nhập và chi tiêu |

+ 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)?

1. **Cho biến định lượng (Numerical)**

Thước đo xu hướng trung tâm

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Thước đo | Công thức/Khái niệm | Khi nào Nên dùng? | Ưu điểm | Nhược điểm | Ví dụ (Dataset: [2, 3, 3, 5, 8, 10, 15, 22]) |
| Trung bình (Mean) | Tổng tất cả giá trị / Số lượng giá trị. | Khi dữ liệu **phân phối chuẩn** và **không có outliers**. | Sử dụng tất cả thông tin trong dữ liệu. | Rất nhạy cảm với **giá trị ngoại lai (outliers)**. | (2+3+3+5+8+10+15+22)/8 = **8.5** |
| Trung vị (Median) | Giá trị nằm chính giữa khi sắp xếp dữ liệu theo thứ tự. | Khi dữ liệu có **outliers** hoặc **phân phối lệch**. | **KHÔNG** bị ảnh hưởng bởi outliers. | Không sử dụng tất cả thông tin trong tập dữ liệu. | Data đã sắp xếp: [2, 3, 3, **5, 8**, 10, 15, 22] -> Trung vị = (5+8)/2 = **6.5** |
| Mode (Yếu vị) | Giá trị xuất hiện **nhiều lần nhất**. | Thường dùng cho biến định tính, hoặc khi cần biết "giá trị phổ biến nhất". | Hữu ích cho dữ liệu định tính & định lượng. | Có thể có nhiều hơn một mode, hoặc không có mode nào. | Số 3 xuất hiện 2 lần -> **Mode = 3** |

Thước đo độ phân tán

|  |  |  |  |
| --- | --- | --- | --- |
| Thước đo | Khái niệm | Mục đích | Ví dụ (Dataset: [2, 3, 3, 5, 8, 10, 15, 22]) |
| **Khoảng biến thiên (Range)** | Max - Min | Cho biết phạm vi tổng thể của dữ liệu. | 22 - 2 = **20** (Dữ liệu trải rộng 20 đơn vị) |
| **Phương sai (Variance)** | Trung bình của bình phương các độ lệch so với giá trị trung bình. | Đo lường mức độ **phân tán trung bình**. Số lớn = Phân tán nhiều. | Tính toán phức tạp hơn, ví dụ: **~40.6** |
| **Độ lệch chuẩn (Standard Deviation)** | Căn bậc hai của Phương sai. | **QUAN TRỌNG NHẤT.** Đo lường độ phân tán theo đơn vị gốc, dễ giải thích hơn. | √40.6 ≈ **6.37** (Các điểm dữ liệu thường lệch khỏi trung bình khoảng 6.37 đơn vị) |
| **Khoảng tứ phân vị (IQR - Interquartile Range)** | Q3 (75%) - Q1 (25%) | Đo lường độ phân tán của **50% dữ liệu trung tâm**. Rất hữu ích để xác định **outliers**. | Q1=3, Q3=12.5 -> IQR = 12.5 - 3 = **9.5** |

1. **Cho biến định tính**

|  |  |  |
| --- | --- | --- |
| Thước đo | Khái niệm | Ví dụ (Biến "Màu Xe ưa thích") |
| **Tần số (Frequency)** | Đếm số lần xuất hiện của mỗi hạng mục. | Đỏ: 15, Trắng: 30, Đen: 25, Xanh: 10 |
| **Tần suất (Percentage)** | Tỷ lệ phần trăm của mỗi hạng mục. | Đỏ: 18.75%, Trắng: **37.5%**, Đen: 31.25%, Xanh: 12.5% |
| **Mode (Yếu vị)** | Hạng mục xuất hiện **nhiều nhất**. | **Mode = "Trắng"** (vì có tần số cao nhất là 37.5%) |

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

a. **Mối quan hệ Tương quan (Correlation)**

Định nghĩa: Tương quan mô tả mối quan hệ thống kê giữa hai biến. Nó cho biết khi biến A thay đổi, biến B có xu hướng thay đổi theo một cách nào đó. Mối quan hệ này có thể đo lường được bằng số.

Các loại tương quan chính:

|  |  |  |  |
| --- | --- | --- | --- |
| Loại Tương quan | Biểu đồ Phân tán (Scatter Plot) | Hệ số Tương quan (r) | Ý nghĩa |
| **Tương quan Dương** | Điểm dữ liệu tạo thành một đường chéo hướng lên. | 0 < r ≤ 1 | Khi biến A **TĂNG**, biến B có xu hướng **TĂNG**. *Ví dụ: Số giờ học và Điểm thi.* |
| **Tương quan Âm** | Điểm dữ liệu tạo thành một đường chéo hướng xuống. | -1 ≤ r < 0 | Khi biến A **TĂNG**, biến B có xu hướng **GIẢM**. *Ví dụ: Số lần đi trễ và Lương thưởng.* |
| **Không có tương quan** | Điểm dữ liệu phân tán ngẫu nhiên, không có hình dạng rõ ràng. | r ≈ 0 | Không có mối quan hệ tuyến tính nào giữa hai biến. *Ví dụ: Số điện thoại và Cân nặng.* |

Cách đo lường Tương quan:

* Hệ số tương quan Pearson (r): Đo lường mức độ và hướng của mối quan hệ tuyến tính giữa hai biến định lượng.
  + r = +1: Tương quan dương hoàn hảo.
  + r = 0: Không có tương quan tuyến tính.
  + r = -1: Tương quan âm hoàn hảo.

1. **Mối quan hệ nhân quả**

Định nghĩa: Nhân quả ngụ ý rằng sự thay đổi của một biến trực tiếp gây ra sự thay đổi của biến kia. Đây là một mối quan hệ mạnh hơn rất nhiều so với tương quan.

Công thức đơn giản: Biến A → Biến B  
*Ví dụ: Hút thuốc lá (A) → Gia tăng nguy cơ ung thư phổi (B).*

Làm thế nào để xác định mối quan hệ nhân quả?  
Rất khó để chứng minh chỉ bằng số liệu thống kê. Cần:

1. Thí nghiệm có đối chứng (Gold Standard): Chia ngẫu nhiên đối tượng thành nhóm thí nghiệm và nhóm đối chứng. Nhóm thí nghiệm được can thiệp (A), nhóm đối chứng thì không. Nếu kết quả (B) khác biệt, ta có thể kết luận A gây ra B.
2. Các nghiên cứu quan sát chất lượng cao: Khi không thể làm thí nghiệm (ví dụ: nghiên cứu tác hại của thuốc lá), cần phải loại trừ tất cả các yếu tố gây nhiễu tiềm ẩn.
3. Sử dụng các mô hình kinh tế lượng: Như Hồi quy với Biến công cụ (Instrumental Variables) để cô lập ảnh hưởng nhân quả.

Cách để xác định mối quan hệ giữa 2 biến:

Bước 1: Xác định loại biến

* 2 biến Định lượng: Dùng Scatter plot và hệ số tương quan Pearson.
* 1 biến Định tính & 1 biến Định lượng: Dùng Box plot (so sánh trung vị/trung bình giữa các nhóm).
* 2 biến Định tính: Dùng Bảng chéo (Crosstab) và biểu đồ Stacked Bar.

Bước 2: Trực quan hóa

* Luôn luôn vẽ biểu đồ! Một biểu đồ tốt có thể cho bạn thấy ngay hình dạng của mối quan hệ (tuyến tính, phi tuyến, không có) và phát hiện outliers.

Bước 3: Đo lường mức độ (nếu có)

* Tính hệ số tương quan, hoặc so sánh sự khác biệt giữa các nhóm.

Bước 4: Diễn giải và Đặt câu hỏi về Tính nhân quả

* Hỏi: "Mối quan hệ này có ý nghĩa gì trong thực tế?"
* Cảnh giác với các biến gây nhiễu: Có yếu tố thứ ba nào có thể giải thích cho mối quan hệ này không?
* Xem xét hướng của mối quan hệ: Có chắc A gây ra B, hay thực ra là B gây ra A, hay chúng ảnh hưởng lẫn nhau?

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

**a. Hiệp phương sai (Covariance)**

Định nghĩa: Hiệp phương sai đo lường hướng của mối quan hệ tuyến tính giữa hai biến. Nó cho biết khi một biến tăng, biến kia có xu hướng tăng hay giảm.

Công thức (ý tưởng): Cov(X, Y) = Σ [ (Xᵢ - Mean(X)) \* (Yᵢ - Mean(Y)) ] / (n-1)

Cách hiểu đơn giản:

* Nó nhìn vào xem các cặp điểm (X, Y) cùng nằm về một phía so với trung bình của chúng hay không.
* Giá trị dương (+): Khi X > trung bình, Y cũng có xu hướng > trung bình (và ngược lại). → Hướng đi lên.
* Giá trị âm (-): Khi X > trung bình, Y lại có xu hướng < trung bình (và ngược lại). → Hướng đi xuống.
* Giá trị gần 0: Không có mối quan hệ tuyến tính rõ ràng.

Vấn đề lớn của Covariance:

* Đơn vị đo lường kỳ quặc: Đơn vị của nó là tích của đơn vị của X và Y (ví dụ: cm\*kg). Điều này làm cho chúng ta không thể so sánh mối quan hệ giữa các cặp biến khác nhau.
* Không biết được "mạnh hay yếu": Một giá trị +100 có mạnh hơn +10 không? Không chắc, vì nó phụ thuộc vào quy mô của dữ liệu.

Tóm lại, Covariance trả lời câu hỏi: "Hướng của mối quan hệ là gì?"

**b.** **Tương quan (Correlation - thường là Pearson)**

Định nghĩa: Tương quan là một phiên bản chuẩn hóa của hiệp phương sai. Nó không chỉ cho biết hướng mà còn cho biết độ mạnh của mối quan hệ tuyến tính.

Công thức (ý tưởng): Corr(X, Y) = Cov(X, Y) / (σₓ \* σᵧ)  
(Chia hiệp phương sai cho tích của độ lệch chuẩn của X và Y)

Cách hiểu đơn giản:

* Bằng cách chia cho độ lệch chuẩn, chúng ta "loại bỏ" ảnh hưởng của đơn vị và quy mô dữ liệu.
* Giá trị luôn nằm trong khoảng từ -1 đến +1.
  + +1: Tương quan dương tuyệt đối.
  + -1: Tương quan âm tuyệt đối.
  + 0: Không có tương quan tuyến tính.

Ưu điểm vượt trội của Correlation:

* Không có đơn vị: Cho phép so sánh trực tiếp mối quan hệ giữa các cặp biến khác nhau. Ví dụ, bạn có thể so sánh độ mạnh của mối quan hệ (Chiều cao - Cân nặng) với (Số giờ học - Điểm thi).
* Diễn giải dễ dàng: Giá trị càng gần ±1, mối quan hệ tuyến tính càng mạnh.

Tóm lại, Correlation trả lời câu hỏi: "Hướng và sức mạnh của mối quan hệ là gì?"

Sự khác biệt then chốt (Tóm tắt bằng bảng)

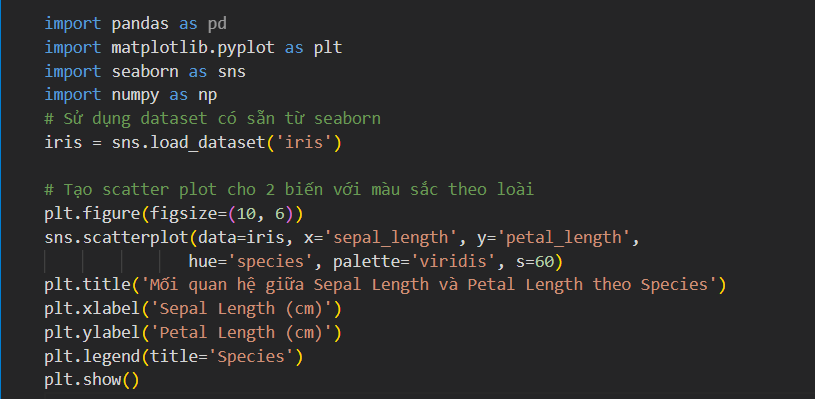
|  |  |  |
| --- | --- | --- |
| Đặc điểm | Hiệp phương sai (Covariance) | Tương quan (Correlation) |
| **Định nghĩa** | Đo lường **hướng** của mối quan hệ tuyến tính. | Đo lường **hướng và sức mạnh** của mối quan hệ tuyến tính. |
| **Giá trị** | Từ -∞ đến +∞ | Từ **-1 đến +1** |
| **Đơn vị** | Có đơn vị (phụ thuộc vào đơn vị của X và Y). | **Không có đơn vị** (chuẩn hóa). |
| **Khả năng so sánh** | **KHÔNG** thể so sánh giữa các tập dữ liệu khác nhau. | **CÓ** thể so sánh giữa các tập dữ liệu khác nhau. |
| **Ý nghĩa** | Cho biết hai biến thay đổi cùng chiều hay ngược chiều. | Cho biết hai biến thay đổi cùng/ngược chiều và mức độ chặt chẽ của xu hướng đó. |

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

a. Đối với phân tích đơn biến

1. Khi bắt đầu dự án phân tích
   * Mục tiêu: Khám phá và làm quen với dữ liệu
   * Ví dụ: Xem phân phối tuổi của toàn bộ nhân viên công ty
2. Khi cần làm sạch dữ liệu
   * Mục tiêu: Phát hiện giá trị bất thường (outliers), lỗi dữ liệu
   * Ví dụ: Dùng box plot để tìm người có thu nhập = 0 hoặc quá cao bất thường
3. Khi cần mô tả đặc điểm một nhóm
   * Mục tiêu: Tóm tắt và trình bày đặc điểm của một tập hợp
   * Ví dụ: Báo cáo phân bố trình độ học vấn của nhân viên
4. Khi kiểm tra giả thuyết về phân phối
   * Mục tiêu: Xem dữ liệu có phân phối chuẩn không, có bị lệch không
   * Ví dụ: Kiểm tra xem điểm thi có phân phối chuẩn không trước khi áp dụng các phương pháp thống kê

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



A graph with different colored dots

AI-generated content may be incorrect.

+ 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?

Trường hợp lý tưởng:

* + Biến phân loại: Có từ 2-6 nhóm (quá nhiều nhóm sẽ khó nhìn)
  + Biến số: Cần so sánh hình dạng phân phối chi tiết giữa các nhóm
  + Mục tiêu: Hiểu không chỉ vị trí mà cả "hình dạng" sự khác biệt

Các bước thực hiện:

Bước 1: Chuẩn bị dữ liệu

* Biến phân loại: Dạng category (ví dụ: "Miền Bắc", "Miền Trung", "Miền Nam")
* Biến số: Dạng numerical (ví dụ: thu nhập, điểm số, chiều cao)

Bước 2: Lựa chọn thư viện

* Seaborn (khuyến nghị): sns.violinplot() - dễ dùng, đẹp mắt
* Matplotlib: plt.violinplot() - nhiều tuỳ chỉnh hơn nhưng phức tạp hơn

Bước 3: Vẽ violin plot cơ bản

sns.violinplot(x='biến\_phân\_loại', y='biến\_số', data=dataset)

Bước 4: Tuỳ chỉnh nâng cao (tuỳ chọn)

* Màu sắc: Dùng palette để phân biệt các nhóm
* Hiển thị bên trong:
  + inner='box': Hiển thị box plot bên trong
  + inner='quartile': Hiển thị các đường tứ phân vị
  + inner='stick': Hiển thị các điểm dữ liệu
* Điều chỉnh độ mượt: Tham số bw để kiểm soát bandwidth