Bài tập thực hành lab05

Trả lời các câu hỏi lý thuyết

Thành viên nhóm : Nguyễn Tuấn Đạt 3123410070

Châu Hải Đăng

Trần Đại Thắng

Bài làm

Phần 1.1: Thống kê mô tả

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

* Thống kê mô tả là tập hợp các phương pháp tóm tắt, trình bày và mô tả dữ liệu hiện có. Nó không đưa ra kết luận hay suy luận cho tổng thể, mà chỉ phản ánh thực trạng của dữ liệu mẫu.

Ví dụ: Giả sử ta có dataset về chiều cao của 100 học sinh:

+ Trung bình chiều cao: 165cm

+ Độ lệch chuẩn: 7.2cm

+ Min = 148cm, Max= 180cm

Có các công cụ thường dùng:

* Số đo tập trung: Mean, Median, Mode
* Độ phân tán: Variance, Standard deviation, Range, IQR
* Biểu đồ: Histogram, Boxplot, Pie chart

Thống kê suy luận đi xa hơn trong thống kê mô tả. Nó dùng dữ liệu mẫu để suy luận, ước lượng hoặc kiểm định giả thuyết cho toàn bộ quần thể. Thay vì chỉ “ mô tả dữ liệu ta có”, thống kê suy luận giúp đưa ra kết luận, dự đoán dựa trên xác suất và mẫu ngẫu nhiên.

Ví dụ: về chiều cao của 100 học sinh:

+ Trung bình mẫu 165cm, SD = 7,2cm ta muốn suy luận là học sinh trường cao từ trong khoảng từ bao nhiêu đến bao nhiêu hoặc kiểm định giả thiết “Mam cao hơn nữ một cách có ý nghĩa thông kê hay không?”

Có các công cụ thường dùng:

* Ước lượng tham số: Confidence Interval
* Kiểm định giả thiết: t-test, ANOVA, Chi-square
* Mô hình hóa: Hồi quy, Logistic Regression.

Tóm lại thống kê mô tả là kể chuyện về dữ liệu đang có. Thống kê suy luận là dự đoán, tổng quát hóa dữ liệu mẫu ra quần thể.

1. **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?**

* - Trung bình (Mean):

+ Định nghĩa: Tổng tất cả giá trị chia cho số lượng quan sát.

+ Ý nghĩa: Đại diện cho giá trị trung tâm của dữ liệu.

+ Ưu điểm: Dễ tính, sử dụng rộng rãi.

+ Nhược điểm: Nhạy cảm với outlier (giá trị ngoại lai).

* Trung vị (Median):

+ Định nghĩa: Giá trị ở giữa khi sắp xếp dữ liệu.

+ Ý nghĩa: Đại diện cho giá trị trung tâm, ít bị ảnh hưởng bởi outlier.

+ Ưu điểm: Ổn định khi dữ liệu lệch (skewed).

+ Nhược điểm: Không phản ánh sự biến động toàn bộ dữ liệu.

* Mốt (Mode):

+ Định nghĩa: Giá trị xuất hiện nhiều nhất

+ Ý nghĩa: Cho biết giá trị phổ biến trong dữ liệu

+ Ứng dụng: Dữ liệu phân loại (ví dụ mày sắc giấy được mua nhiều nhất)

* Phương sai (Variance)

+ Định nghĩa: Đo độ phân tán bằng trung bình bình phương khoảng cách đến mean.

+ Ý nghĩa: Cho biết dữ liệu dao động nhiều hay ít quanh giá trị trung tâm.

+ Đơn vị: Bình phương của đơn vị gốc.

* Độ lệch chuẩn (Standard Deviation – SD):

+ Định nghĩa: Căn bậc 2 của phương sai.

+ Ý nghĩa: Đo mức độ phân tán thực tế của dữ liệu (trở về đơn vị gốc)

+ Ứng dụng: So sánh mức biến động giữa các biến

* Khoảng tứ phân vị (IQR – Interquartile Range)

+ Định nghĩa: Hiệu số giữa Q3 (75%) và Q1 (25%)

+ Ý nghĩa: Đo độ phân tán giữa các giá trị trung tâm (loại bỏ outlier)

+ Ứng dụng: Xác định outlier bằng quy tắc IQR

* Nên dùng trung vị thay cho trung bình ở các trường hợp:

+ Phân phối lệch: ví dụ như thu nhập cá nhân, giá nhà,...

+ Có outlier mạnh: ví dụ như dữ liệu lỗi (tuổi = 200),...

+ Khi quan tâm đến “ giá trị điển hình” ví dụ báo cáo mức lượng điển hình.

1. 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)?

* Để kiểm tra dữ liệu phân bố theo dạng nào ta có thể kết hợp nhiều cách:
* Trực quan hóa: Histogram, KDE, Boxplot
* Thống kê mô tả: Skewness (Độ lệch), Kurtosis (Độ nhọn)
* Kiểm định thống kê: Shapiro-Wilk test, Kolmogorov-Smirnov test,
* Q-Q Plot: So sánh quantile của dữ liệu với quantile lý thuyết chuẩn.

Các loại phân bố phổ biến:

* Phân bố chuẩn: Đối xứng, hình chuông, nhiều hiện tượng tự nhiên, Mean và Median và Mode gần bằng nhau
* Phân bố lệch phải: Đuôi kéo dài về phía bên phải, Mean > Median.
* Phân bố lệch trái: Đuôi kéo dài về phái bên trái, Mean < Median.
* Phân bố đồng đều: Mọi giá trị trong khoảng điều có xác suất gần nhau
* Phân bố nhọn/bẹt : nhọn thì giá trị gần Mean, đuôi dày còn bẹt thì phân tán đuôi mỏng.

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

* Phạm vi (range)
* Định nghĩa: Là khoảng giá trị mà dữ liệu trải rộng.
* Ý nghĩa: Cho biết dữ liệu trải dài từ nhỏ nhất đến lớn nhất, dễ hiểu, dễ tính.
* Hạn chế: Rất nhạy cảm với outlier

Độ lệch chuẩn:

* Định nghĩa: Đo trung bình khoảng các của từng giá trị đến mean.
* Ý nghĩa: Cho biết mức độ dao động của dữ liệu quanh giá trị trung bình
* Ưu điểm: Phản ánh sự biến động toàn bộ dữ liệu, dùng được trong nhiều mô hình thống kê và ML

Tóm lại

* Range: nhìn toàn cảnh, dễ bị outlier bóp méo
* SD : nhìn độ biến động thực tế, đáng tin cậy hơn.

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

* Trong thống kê, ta chia dữ liệu sau khi sắp xếp từ nhỏ đến lớn thành 4 phân bằng nhau gọi là tứ phân vị:
* Q1: Điểm chia tại 25% dữ liệu nhỏ nhất
* Q2: Trung vị của dữ liệu chia dữ liệu ra thành 2 nửa bằng nhau.
* Q3: Điểm chia tại 75% dữ liệu nhỏ nhất

Tóm lại:

* Q1: giá trị thấp nhưng không có cực trị
* Q2: giá trị điển hình ở giữa
* Q3: giá trị cao nhưng không có cực trị

Một boxplot gồm các thành phần chính:

* Hộp: Q1 🡪 Q3
* Đường giữa hộp: Q2
* Chiều dài hộp: Q3 – Q1
* Điểm ngoài: Giá trị ngoài khoảng trên

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

* Tùy thuộc vào ngữ cảnh và bản chất dữ liệu, ta có nhiều chiến lược:
* Xóa dữ liệu thiếu: Xóa hàng, xóa cột
* Thay thế bằng giá trị thống kê: Trung bình, trung vị, mode
* Một số kỹ thuật nâng cao hơn: KNN imputation, Regression imputation,

Tóm lại:

* Tỷ lệ giá trị thiếu (ít thì xoá, nhiều thì cần impute)
* Bản chất dữ liệu ( số liệu chuẩn dùng mean, lệch dùng median, categorical dùng mode)
* Mục đích phân tích ( EDA đơn giản hay chuyên sâu)

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

* Histogram:
* Cách đọc:

+ Trục x: Các giá trị của biến

+ Trục y: Tần suất

+ Chiều cao của cột: Cho biết có bao nhiêu dữ liệu nằm trong khoảng đó.

* Diễn giải:

+ Nếu Histogram cân đối và chuông: Dữ liệu gần với phân bố chuẩn

+ Nếu lệch trái: Phần đuôi dữ liệu hướng về phía trái.

+ Nếu lệch phải: Đuôi dài về phía phải.

+ Đỉnh (mode) : Vị trí cột cao nhất 🡪 giá trị thường gặp.

+ Độ rộng trải: Histogram dàn trải rộng 🡪 dữ liệu phân tán cao

Boxplot:

* Cách đọc:

+ Q1: Giá trị mà 25% dữ liệu nhỏ hơn

+ Q2: Giá trị trung vị

+ Q3: Giá trị mà 75% dữ liệu nhỏ hơn

+ Hộp: Khoảnh cách giữa Q1 và Q3

+ Điểm rời rạc: là outlier.

* Diễn giải:

+ Hộp cao: Dữ liệu biến động mạnh

+ Đường median lệch về 1 phía: dữ liệu không đối xứng

+ Nhiều outlier: dữ liệu có giá trị bất thường

1. 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?

* Tùy mục tiêu và dữ liệu, ta có nhiều cách:
* Giữ nguyên: Nếu outlier có ý nghĩa thực tế
* Loại bỏ: outlier là do lỗi nhập liệu hoặc cảm biến
* Winsorizing: thay outlier bằng ngưỡng gần nhất trong phân vị.
* Biến đổi dữ liệu: áp dụng log-transform, square root đề làm giảm ảnh hưởng outlier
* Sử dụng thước đo robust thay vì mean/std: median thay cho mean, IQR thay cho std.

Phần 1.2: Xử lý và trực quan hóa dữ liệu:

1. 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)?

* Trực quan hóa dữ liệu là biến dữ liệu thành hình ảnh như biểu đồ, đồ thị,... có vai trò là:
* Hiểu nhanh cấu trúc và phân phối dữ liệu.
* Phát hiện vấn đề và bất thường
* Giao tiếp và thuyết phục
* Định hướng phân tích sâu hơn

Trực quan hóa dữ liệu quan trọng trong EDA vì:

* EDA = khám phá dữ liêu nên ta cần hình dung nó nên hình ảnh sẽ giúp ta dễ dàng hơn trong việc hình dung dữ liệu.
* Xem nhanh các đặc điểm cơ bản: phân bố, trung tâm, độ phân tán
* Phát hiện patterns: xu hướng, chu kỳ, cụm dữ liệu
* Xác định dữ liệu bất thường: missing values, outliers, noise
* Ra quyết định bước kế tiếp: biến nào cần phân tích thêm, biến nào cần loại bỏ,...

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

* Histogram: dùng khi muốn xem phân bố của một biến liên tục
* Cho biết dữ liệu có tuân theo phân bố chuẩn, lệch, có đa đỉnh hay không
* Giúp xác định outliers hoặc vùng giá trị hiếm gặp

Scatter Plot: dùng khi muốn khám phá mối quan hệ giữa 2 biến số

* Nếu dữ liệu phân bố thành đường chéo 🡪 có quan hệ tuyến tính
* Nếu dữ liệu phân bô tán loạn 🡪 ít mỗi liên hệ
* Có thể phát hiện cụm dữ liệu hoặc outliers

Boxplot: dùng khi muốn tóm tắt phân bố một biến và phát hiện outliers

* Thể hiện median, Q1, Q2, Q3, min, max.
* Giúp so sánh phân bố của nhiều nhóm 1 lúc

Bar Chart: dùng khi muốn so sánh giá trị giữa các nhóm (categorical)

* Thích hợp cho dữ liệu rời rạc (giới tính, ngành học, sản phẩm)
* Không dùng cho dữ liệu liên tục.

Tóm lại:

* Histogram: phân bố của một biến liên tục
* Scatter plot: mối quan hệ giữa 2 biến
* Boxplot: Tóm tắt phân bố và phát hiện outliers
* Bar Chart: so sánh dữ liệu giữa các nhóm phân loại

1. 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)?

* Nguyên tắc chung khi chọn biểu đồ:
* Dữ liệu phân loại (categorical): so sánh số lượng/tỷ lệ giữa các nhóm
* Dữ liệu số (numerical / continuous): xem phân bố, trung tâm, độ phân tán
* Dữ liệu chuỗi thời gian: theo dõi sự thay đổi qua thời gian
* Kết hợp nhiều biến: dùng scatter, heatmap, pairplot

Lựa chọn theo từng loại dữ liệu:

* Dữ liệu phân loại : Bar chart, Pie chart, Stacked bar chart
* Dữ liệu số liên tục: Histogram, Boxplot, kDE, Scatter plot
* Dữ liệu chuỗi thời gian: Line chart, Area chart, Candlestick chart
* Dữ liệu nhiều biến: Scatter plot ma trận, Heatmap, Bubble chart

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

* Matplotlib
* Đặc điểm:

+ “Ông tổ” của trực quan hóa trong Python (ra đời từ 2003).

+ Rất linh hoạt: bạn có thể tùy chỉnh gần như mọi thứ (màu sắc, font chữ, ticks, annotations).

* Điểm mạnh:

+ Kiểm soát chi tiết, tùy chỉnh cao.

+ Phù hợp để xuất bản biểu đồ cho báo cáo khoa học.

+ Có thể tạo từ biểu đồ cơ bản đến nâng cao (line, bar, scatter, histogram, 3D plot).

* Điểm yếu:

+ Code dài dòng, dễ mất thời gian chỉnh sửa.

+ Giao diện mặc định không đẹp thường phải kết hợp Seaborn để có style đẹp.

Seaborn

* Đặc điểm:

+ Xây dựng trên Matplotlib nhưng cú pháp ngắn gọn, đẹp hơn

+ Hỗ trợ trực tiếp cho dữ liệu dạng DataFrame (Pandas).

+ Tích hợp sẵn các biểu đồ thống kê nâng cao: boxplot, violin plot, pairplot, heatmap.

* Điểm mạnh:

+ Giao diện mặc định đẹp.

+ Dễ dàng làm EDA (khám phá dữ liệu).

+ Có biểu đồ trực quan hóa mối quan hệ và phân bố rất mạnh.

* Điểm yếu:

+ Ít tùy biến hơn Matplotlib (nếu cần kiểm soát chi tiết rất sâu thì phải gọi Matplotlib).

+ Không hỗ trợ biểu đồ tương tác

Plotly

* Đặc điểm:

+ Hỗ trợ biểu đồ tương tác

+ Dễ dùng cho dashboard, web app, và khi trình bày dữ liệu cho người không chuyên.

+ Có phiên bản tích hợp với Dash để làm web dashboard.

* Điểm mạnh:

+ Biểu đồ đẹp, hiện đại, có thể zoom, hover, pan.

+ phù hợp cho data storytelling và trình chiếu báo cáo

+ Hỗ trợ rất nhiều loại biểu đồ nâng cao

* Điểm yếu:

+ Cú pháp dài hờn Seaborn một chút

+ Đôi khi hơi nặng lúc hiển thị dữ liệu lớn

+ Phụ thuộc vào Javascript

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

* Nguyên tắc thiếc kế biểu đồ trực quan hóa dữ liệu
* Biểu đồ phải phục vụ mục tiêu phân tích: xác định rõ muốn người xem so sánh, xu hướng hay là phân bố.
* Đơn giản và tránh rườm rà: Tránh biểu đồ 3D hiệu ứng, giữ biểu đồ tối giản
* Chọn thang đo hợp lý
* Màu sắc có ý nghĩa: dùng màu để nhấn mạnh thông tin, tránh lạm dụng
* Chú thích rõ ràng: Trục X, Y phải có tên, có tiêu đề ngắn gọn, rõ ràng
* Nhấn mạnh Insight, không chỉ dữ liệu: highlight phần quan trọng
* Tính nhất quán: tránh thay đổi trục hoặc scale giữa các biểu đồ

Tóm lại, một biểu đồ trực quan hóa phải:

* Đúng mục tiêu
* Dễ đọc, dễ hiểu
* Chính xác, không gây hiểu nhầm
* Thẩm mỹ vừa đủ

1. 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?

* Histogram dùng để xem phân bố của dữ liệu số.

Code mẫu: xem phân bố điểm thi của 100 sinh viên:

A computer screen shot of text

AI-generated content may be incorrect.

Kết quả:

A graph with blue bars

AI-generated content may be incorrect.

Bar chart dùng để so sánh giữa các nhóm (categorical data)

Code mẫu:

A screen shot of a computer program

AI-generated content may be incorrect.

Kết quả:

A graph with orange bars

AI-generated content may be incorrect.

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

* Xuất biểu đồ tĩnh ( PNG, JPG, PDF, SVG, EPS) với matplotlib

Sau khi vẽ biểu đồ chỉ cần thêm dòng:

A black background with white text

AI-generated content may be incorrect.

Nếu muốn chèn biểu đồ động vào báo cáo web hoặc notebook dùng plotly

Xuất ra file PDF: 

Phần 1.3: Phân tích đơn biến và đa biến

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:
* Định nghĩa: là phân tích một biến duy nhất trong tập dữ liệu
* Kỹ thuật thường dùng:

+ Với biến số: Thống kê mô tả, sử dụng histogram, Boxplot

+ Với biến phân loại: Tần suất, sử dụng Bar chart, Pie chart

Phân tích hai biến:

* Định nghĩa: là phân tích mối quan hệ giữa hai biến.
* Kỹ thuật thường dùng:

+ Số -- số: Hệ số tương quan, dùng Scatter plot, Regression line

+ số -- phân loại: So sánh trung bình bằng Boxplot

+ Phân loại – Phân loại: bảng chéo, Heatmap, Mosaic plot

Tóm lại:

* Univariate: chỉ tập trung vào một biến để hiểu rõ phân bố
* Bivariate: tập trung vào mối quan hệ giữa hai biến để phát hiện sự tương quan.

1. 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)?

* Thước đo xu hướng trung tâm: Mean, Median, Mode

Thước đo phân tán: Range, Variance, SD, IQR

Thước đo hình dạng phân bố: Skewness, Kurtosis

1. Trong phân tích hai biến, làm thế nào để xác định mối quan hệ giữa hai biến (ví dụ: tương quan, nhân quả)?
2. Sự khác biệt giữa tương quan (correlation) và hiệp biến (covariance) trong phân tích hai biến là gì?
3. Khi nào nên sử dụng biểu đồ trực quan hóa trong phân tích đơn biến so với phân tích hai biến?
4. Đoạn code mẫu để tạo biểu đồ scatter plot hoặc heatmap để phân tích mối quan hệ giữa hai biến?
5. 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?