Phân tích dữ liệu sức khỏe

Tràn Anh Quân

Ngày 20 tháng 3 năm 2025

1 Giới thiệu

Quản lý cân nặng đóng vai trò quan trọng trong việc duy trì sức khỏe và phòng ngừa các bệnh lý liên quan đến lối sống như béo phì, tiểu đường và tim mạch. Việc phân tích dữ liệu về thói quen tập luyện và lối sống có thể cung cấp những hiểu biết quan trọng trong việc tối ưu hóa chiến lược quản lý cân nặng.

Nghiên cứu này sử dụng bộ dữ liệu FitLife360, bao gồm thông tin theo dõi sức khỏe và hoạt động thể chất của 3.000 người tham gia trong vòng một năm. Dữ liệu bao gồm các thông tin nhân khẩu học, thói quen tập luyện, chỉ số sinh lý và các yếu tố lối sống như thời gian ngủ, mức độ căng thẳng và mức độ hydrat hóa. Việc khai thác và phân tích dữ liệu này có thể giúp xác định các yếu tố có ảnh hưởng đáng kể đến cân nặng, từ đó hỗ trợ các biện pháp điều chỉnh phù hợp.

Mục tiêu nghiên cứu:

Phân tích tác động của các loại hình tập luyện và cường độ tập luyện đến lượng calo tiêu thụ.

Đánh giá ảnh hưởng của giấc ngủ, mức độ căng thẳng và lượng nước tiêu thụ đến quản lý cân nặng.

Xác định mối quan hệ giữa các chỉ số sinh lý (nhịp tim, huyết áp) với quá trình kiểm soát cân nặng.

Ứng dụng hồi quy tuyến tính để xây dựng mô hình dự đoán mức độ tiêu hao calo dựa trên các biến số liên quan.

Ứng dụng thực tiễn

Hỗ trợ cá nhân xây dựng chương trình tập luyện phù hợp với thể trạng và mục tiêu kiểm soát cân nặng.

Đưa ra khuyến nghị về chế độ sinh hoạt hợp lý nhằm tối ưu hóa hiệu quả tập luyện.

Cung cấp cơ sở dữ liệu cho chuyên gia sức khỏe trong việc thiết kế các chương trình quản lý cân nặng dựa trên bằng chứng thực nghiệm.

Việc phân tích dữ liệu này sẽ góp phần nâng cao hiểu biết về mối quan hệ giữa thói quen tập luyện, lối sống và cân nặng, đồng thời hỗ trợ các giải pháp khoa học trong lĩnh vực quản lý sức khỏe cá nhân.

2 Giới thiệu chung về các phương pháp phân tích dữ liệu

Thống kê mô tả (Descriptive Statistics)

Thống kê mô tả là tập hợp các phương pháp được sử dụng để tổ chức, tóm tắt và trình bày dữ liệu một cách có ý nghĩa, nhằm cung cấp cái nhìn tổng quan về tập dữ liệu mà không cần đưa ra kết luận suy luận hay dự đoán. Nó tập trung vào ba khía cạnh chính: **xu hướng trung tâm**, **độ phân tán**, và **hình dạng phân phối**. Dưới đây là lý thuyết chi tiết cho từng phần:

1. Đo lường xu hướng trung tâm (Measures of Central Tendency)

Xu hướng trung tâm thể hiện giá trị "đại diện" hoặc "trung tâm" của tập dữ liệu, giúp hiểu được vị trí chung của các quan sát.

I. **Mean (Trung bình cộng):**
Đây là giá trị trung bình của tất cả các quan sát trong tập dữ liệu, được tính bằng cách cộng tất cả các giá trị và chia cho số lượng phần tử.
Công thức:

$$\text{Mean}(\mu) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Trong đó:

- x_i: Giá trị của phần tử thứ i.
- n: Tổng số phần tử trong tập dữ liệu.

Ý nghĩa: Mean phản ánh giá trị trung bình tổng quát, nhưng nhạy cảm với các giá trị ngoại lai (outliers).

Ứng dụng: Được dùng trong kinh tế để tính trung bình thu nhập, điểm số, v.v.

II. **Median (Trung vị):**
Median là giá trị nằm ở vị trí trung tâm khi tập dữ liệu được sắp xếp theo thứ tự tăng dần hoặc giảm dần.

- Nếu n lẻ: Median là giá trị ở vị trí $\frac{n+1}{2}$.
- Nếu n chẵn: Median là trung bình cộng của hai giá trị ở vị trí $\frac{n}{2}$ và $\frac{n}{2}+1$.

Ý nghĩa: Median không bị ảnh hưởng bởi ngoại lai, phù hợp với dữ liệu lệch.
Ứng dụng: Phân tích thu nhập để tránh tác động của giá trị cực đại.

III. **Mode (Mốt):**
Mode là giá trị xuất hiện với tần suất cao nhất trong tập dữ liệu. Có thể có một mode, nhiều mode, hoặc không có mode.
Ý nghĩa: Mode xác định giá trị phổ biến nhất, hữu ích cho dữ liệu phân loại.
Ứng dụng: Xác định sản phẩm bán chạy nhất trong thị trường.

IV. **Midrange (Trung tuyến):**
Midrange là trung bình cộng của giá trị lớn nhất và nhỏ nhất trong tập dữ liệu.
Công thức:

$$\text{Midrange} = \frac{x_{\max} + x_{\min}}{2}$$

Ý nghĩa: Midrange cung cấp một ước lượng đơn giản về trung tâm dữ liệu, nhưng rất nhạy cảm với ngoại lai do chỉ dựa vào hai giá trị cực trị.
Ứng dụng: Được dùng trong các bài toán đơn giản như đo nhiệt độ trung bình trong ngày.

2. Đo lường độ phân tán (Measures of Dispersion)

Độ phân tán mô tả mức độ biến thiên hoặc lan rộng của dữ liệu xung quanh xu hướng trung tâm.

I. **Range (Khoảng biến thiên):**
Range là chênh lệch giữa giá trị lớn nhất và nhỏ nhất.
Công thức:

$$\text{Range} = x_{\max} - x_{\min}$$

Ý nghĩa: Range đơn giản nhưng không phản ánh phân bố chi tiết của dữ liệu.
Ứng dụng: Kiểm soát chất lượng (ví dụ: chênh lệch nhiệt độ).

II. **Variance (Phương sai):**
Variance đo lường mức độ phân tán trung bình của dữ liệu so với mean.
Công thức:

- Cho tổng thể:
$$\text{Variance}(\sigma^2) = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$
- Cho mẫu:
$$\text{Variance}(s^2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Ý nghĩa: Variance lớn cho thấy dữ liệu phân tán mạnh.
Ứng dụng: Đo lường rủi ro tài chính.

III. **Standard Deviation (Độ lệch chuẩn):**
Standard deviation là căn bậc hai của variance.
Công thức:

- Cho tổng thể:
$$\text{Standard Deviation}(\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$
- Cho mẫu:
$$\text{Standard Deviation}(s) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Ý nghĩa: Dễ diễn giải hơn variance, phổ biến trong phân phối chuẩn.
Ứng dụng: Đánh giá độ tin cậy của phép đo.

IV. **Interquartile Range (IQR - Khoảng tứ phân vị):**
IQR là khoảng cách giữa phân vị thứ ba (Q3) và phân vị thứ nhất (Q1), tức là phạm vi chứa 50% dữ liệu ở giữa.
Công thức:
$$\text{IQR} = Q_3 - Q_1$$

Ý nghĩa: IQR không bị ảnh hưởng bởi ngoại lai, đo lường độ phân tán của phần trung tâm dữ liệu.
Ứng dụng: Dùng trong boxplot để phát hiện giá trị ngoại lai.

3. Hình dạng phân phối (Shape of Distribution)

Hình dạng phân phối cung cấp thông tin về cách dữ liệu được phân bố.

I. **Skewness (Độ lệch):**
Skewness đo lường mức độ bất đối xứng của phân phối.
Công thức:

$$\text{Skewness} = \frac{\sum_{i=1}^{n}(x_i - \mu)^3/n}{\sigma^3}$$

- > 0: Lệch phải.
- < 0: Lệch trái.
- $= 0$: Đối xứng.

Ý nghĩa: Hỗ trợ chọn phương pháp phân tích thống kê.
Ứng dụng: Phân tích lợi nhuận tài chính.

II. **Kurtosis (Độ nhọn):**
Kurtosis đo lường độ nhọn hoặc phẳng của phân phối.
Công thức:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n}(x_i - \mu)^4/n}{\sigma^4}$$

- $= 3$: Phân phối chuẩn.
- > 3: Nhọn, đuôi dày.
- < 3: Phẳng, đuôi mỏng.

Ý nghĩa: Đo lường khả năng xảy ra giá trị ngoại lai.
Ứng dụng: Phân tích rủi ro.

III. **Percentiles (Phân vị):**
Percentiles chia tập dữ liệu thành 100 phần bằng nhau, trong đó phân vị thứ p (Pp) là giá trị mà $p\%$ dữ liệu nằm dưới nó.
Cách tính: Sắp xếp dữ liệu, vị trí của Pp là $\frac{p}{100} \times (n+1)$.
Ý nghĩa: Mô tả vị trí tương đối của dữ liệu, phổ biến là Q1 (25%), Q2 (50%, tức median), Q3 (75%).
Ứng dụng: Đánh giá phân phối thu nhập hoặc điểm số.

Ứng dụng của Thống kê mô tả

Thống kê mô tả giúp:

1. **Tóm tắt dữ liệu:** Chuyển đổi dữ liệu thô thành số liệu dễ hiểu.
2. **Trực quan hóa:** Dùng histogram, boxplot, scatter plot để minh họa.
3. **Cơ sở cho suy luận:** Hỗ trợ ước lượng tham số, kiểm định giả thuyết.
4. **So sánh dữ liệu:** Đánh giá sự khác biệt giữa các nhóm.

Kết luận

Thống kê mô tả là công cụ mạnh mẽ để khám phá dữ liệu, kết hợp các chỉ số xu hướng trung tâm, độ phân tán và hình dạng phân phối. Việc lựa chọn chỉ số phù hợp và diễn giải trong bối cảnh cụ thể là yếu tố then chốt để áp dụng hiệu quả.

Hồi quy tuyến tính

Mục tiêu: Tìm ra sự ảnh hưởng của từng biến lên chỉ số cân nặng, đồng thời xây dựng mô hình dự đoán giá trị của biến phụ thuộc dựa trên các biến độc lập và đánh giá mức độ phù hợp của mô hình với dữ liệu thực tế.

Hồi quy đơn biến

Khái niệm Hồi quy tuyến tính đơn biến là phương pháp thống kê mô hình hóa mối quan hệ tuyến tính giữa một biến độc lập X và một biến phụ thuộc Y. Mô hình được biểu diễn dưới dạng:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

Trong đó:

- Y: Biến phụ thuộc (ví dụ: chỉ số cân nặng).
- X: Biến độc lập (ví dụ: chiều cao).
- β_0: Hệ số chặn (intercept), giá trị của Y khi $X = 0$.
- β_1: Hệ số góc (slope), thể hiện mức độ thay đổi của Y khi X tăng thêm một đơn vị.

- ϵ: Sai số ngẫu nhiên, đại diện cho các yếu tố không giải thích được bởi mô hình, thường được giả định tuân theo phân phối chuẩn với kỳ vọng bằng 0 và phương sai không đổi ($\epsilon \sim N(0, \sigma^2)$).

Các giả định của hồi quy đơn biến Để mô hình hồi quy tuyến tính đơn biến có ý nghĩa thống kê, cần thỏa mãn các giả định sau:

- Quan hệ giữa X và Y là tuyến tính.
- Sai số ϵ có phân phối chuẩn với kỳ vọng bằng 0.
- Phương sai của sai số ($\text{Var}(\epsilon)$) là không đổi (đồng nhất phương sai - homoscedasticity).
- Các quan sát là độc lập với nhau (không có tự tương quan).

Các bước thực hiện

1. **Thu thập và xử lý dữ liệu:** - Thu thập cặp dữ liệu (X, Y) từ mẫu. - Kiểm tra và xử lý giá trị ngoại lai, dữ liệu bị thiếu (missing data).
2. **Ước lượng tham số bằng phương pháp bình phương tối thiểu (OLS):** - Tìm β_0 và β_1 sao cho tổng bình phương sai số $\sum(Y_i - \hat{Y}_i)^2$ nhỏ nhất, với $\hat{Y}_i = \beta_0 + \beta_1 X_i$. - Công thức ước lượng:

$$\beta_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}, \quad \beta_0 = \bar{Y} - \beta_1 \bar{X}$$

3. **Đánh giá mô hình:**
   - **Hệ số xác định R^2:** Đo lường mức độ giải thích của mô hình, dao động từ 0 đến 1 (1 là hoàn hảo).

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

   - **Kiểm định ý nghĩa hệ số hồi quy:** Sử dụng kiểm định t để đánh giá β_1 có khác 0 hay không (p-value < 0.05).
   - **Kiểm tra giả định hồi quy:** - Dùng biểu đồ phân tán (residual plot) kiểm tra tuyến tính và đồng nhất phương sai. - Kiểm tra phân phối chuẩn của sai số bằng biểu đồ Q-Q hoặc kiểm định Shapiro-Wilk.

4. **Dự báo giá trị mới:** Sử dụng $\hat{Y} = \beta_0 + \beta_1 X$ để dự đoán Y cho giá trị X mới, kèm theo khoảng tin cậy (confidence interval).

Ý nghĩa

- Xác định mức độ ảnh hưởng của một yếu tố đến kết quả (ví dụ: chiều cao ảnh hưởng bao nhiêu đến cân nặng).
- Dự đoán xu hướng dựa trên một biến duy nhất (ví dụ: dự đoán cân nặng từ chiều cao).
- Hỗ trợ ra quyết định trong các vấn đề đơn giản.

Ví dụ thực tế Dự đoán cân nặng (Y) dựa trên chiều cao (X): Nếu $\beta_1 = 0.5$, mỗi cm chiều cao tăng thêm làm tăng cân nặng trung bình 0.5 kg (giả sử $\beta_0 = -50$).

Hồi quy đa biến

Khái niệm Hồi quy tuyến tính đa biến mở rộng mô hình đơn biến bằng cách đưa vào nhiều biến độc lập $X_1, X_2, \ldots, X_n$ để giải thích biến phụ thuộc Y:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \tag{2}$$

Trong đó:

- $X_1, X_2, \ldots, X_n$: Các biến độc lập (ví dụ: chiều cao, tuổi, giới tính).
- $\beta_1, \beta_2, \ldots, \beta_n$: Hệ số hồi quy, thể hiện mức độ ảnh hưởng của từng biến.
- Các ký hiệu khác tương tự hồi quy đơn biến.

Các giả định của hồi quy đa biến Ngoài các giả định của hồi quy đơn biến, hồi quy đa biến yêu cầu thêm:

- Không có đa cộng tuyến (multicollinearity) giữa các biến độc lập (có thể kiểm tra bằng VIF - Variance Inflation Factor).
- Mối quan hệ giữa các biến độc lập và biến phụ thuộc là tuyến tính.

Các bước thực hiện

1. **Chuẩn bị dữ liệu, kiểm tra đa cộng tuyến:** - Thu thập dữ liệu đầy đủ cho tất cả các biến. - Kiểm tra tương quan giữa các biến độc lập (correlation matrix) và tính VIF (VIF $>$ 10 thường chỉ ra đa cộng tuyến). - Chuẩn hóa dữ liệu nếu cần (ví dụ: đưa về thang đo chung).

2. **Ước lượng hệ số bằng phương pháp OLS:** - Tối ưu hóa $\beta_0, \beta_1, \ldots, \beta_n$ để giảm thiểu tổng bình phương sai số. - Sử dụng ma trận (matrix algebra) trong các phần mềm như R hoặc Python để giải.

3. **Đánh giá mô hình:**

   - **Hệ số R^2 và Adjusted R^2:** R^2 đo mức độ giải thích tổng quát, Adjusted R^2 điều chỉnh theo số biến để tránh overfitting.
   - **Kiểm định F:** Đánh giá ý nghĩa tổng thể của mô hình (p-value < 0.05).
   - **Kiểm tra ý nghĩa từng hệ số hồi quy:** Dùng kiểm định t cho từng β_i.
   - **Kiểm tra các giả định hồi quy:** - Biểu đồ residual vs fitted để kiểm tra tuyến tính và đồng nhất phương sai. - Kiểm tra đa cộng tuyến bằng VIF. - Kiểm tra tự tương quan bằng Durbin-Watson test.

4. **Dự báo giá trị mới:** Dùng mô hình đã xây dựng để dự đoán Y từ tập hợp giá trị $X_1, X_2, \ldots, X_n$, kèm khoảng dự báo (prediction interval).

Ý nghĩa

- Phân tích đồng thời nhiều yếu tố ảnh hưởng đến kết quả (ví dụ: cân nặng chịu ảnh hưởng từ chiều cao, tuổi, chế độ ăn).
- Cải thiện khả năng dự báo bằng cách xem xét toàn diện các yếu tố.
- Hỗ trợ ra quyết định trong các vấn đề phức tạp, đa chiều.

Ví dụ thực tế Dự đoán cân nặng (Y) dựa trên chiều cao (X_1), tuổi (X_2), và số giờ tập thể dục mỗi tuần (X_3): Nếu $\beta_1 = 0.4$, $\beta_2 = 0.1$, $\beta_3 = -0.2$, thì cân nặng tăng 0.4 kg mỗi cm chiều cao, tăng 0.1 kg mỗi năm tuổi, và giảm 0.2 kg mỗi giờ tập thể dục (giả sử $\beta_0 = -30$).

2.1 Cây quyết định, Random Forest