

Excel for Data Analysis (2)

TimeSeries Team

Ngày 18 tháng 8 năm 2025

Mục lục

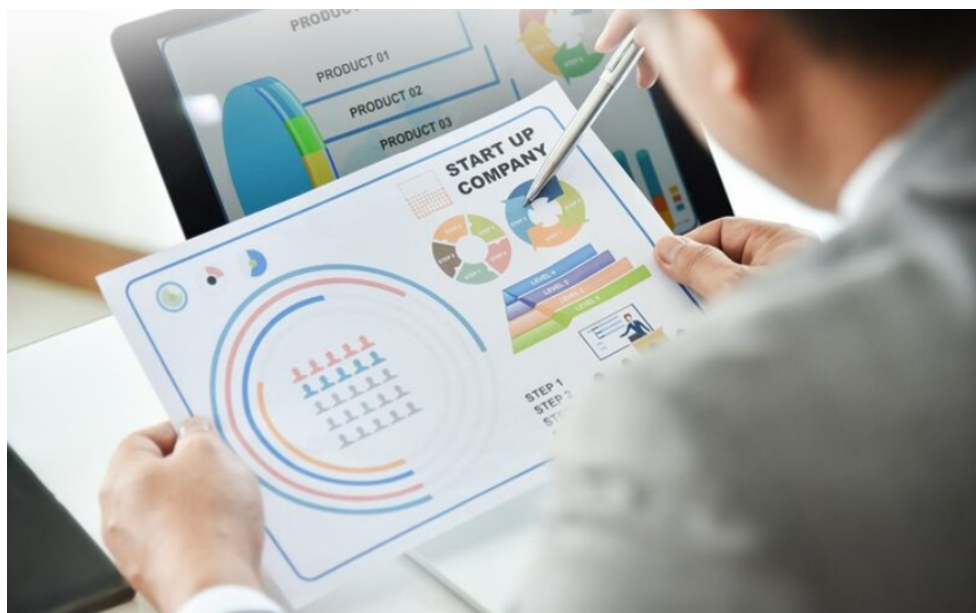
I. Trục quan hóa dữ liệu	3
1 Trục quan hóa dữ liệu là gì?	3
2 Kiến thức cần biết	4
2.1 Thành phần của trục quan hóa dữ liệu	4
2.2 Các bước thực hiện trục quan hóa dữ liệu	4
2.3 Một số phương pháp trục quan hóa dữ liệu điển hình	5
2.3.1 Biểu đồ cột- Bar plot	5
2.3.2 Biểu đồ đường - Line Chart	5
2.3.3 Biểu đồ tròn - Pie Chart	5
2.3.4 Biểu đồ phân phối - Histograms	5
2.3.5 Heatmap	5
2.3.6 Các nguyên tắc chọn phương tiện trục quan hóa dữ liệu	6
II. Kiểm Định Giả Thuyết Là Gì	7
1 Khái niệm & tầm quan trọng	7
2 Các khái niệm cốt lõi	7
2.1 Miền bác bỏ	7
2.2 Các bước làm bài toán kiểm định	8
2.2.1 Sử dụng miền tiêu chuẩn	9
2.2.2 Sử dụng xác suất ý nghĩa p-value	9
2.2.3 Sử dụng ước lượng khoảng	9
2.3 Tổng kết	10
3 Kiểm định giả thuyết thống kê bằng Python	10
3.1 Kiểm định tính chuẩn	10
3.1.1 Shapiro-Wilk Test	10
3.1.2 D'Agostino's K^2 Test	11
3.1.3 Anderson-Darling Test	11
3.2 Kiểm định tương quan - Correlation Tests	12
3.2.1 Hệ số tương quan Pearson	12
3.2.2 Hệ số tương quan thứ hạng của Spearman	12
3.2.3 Hệ số tương quan thứ hạng của Kendall	13
3.2.4 Kiểm định Chi-Squared	13
3.3 Kiểm định giả thuyết thống kê tham số	14
3.3.1 Kiểm định t-student	14
3.3.2 Kiểm định t-Paired Student	14
3.3.3 Kiểm định t-ANOVA	15
3.3.4 Kiểm định giả thuyết thống kê phi tham số	15

3.3.5	Kiểm định Wilcoxon có dấu hiệu-hạng	16
3.3.6	Kiểm định Kruskal-Wallis H	16
3.3.7	Kiểm tra Friedman	17
4	Kiểm định giả thiết thống kê bằng Excel	18
III.	Tiền Xử Lý Dữ Liệu & Phân Tích Hồi Quy	19
1	Tiền xử lý dữ liệu là gì?	19
1.1	Khái niệm	19
1.2	Vai trò	19
1.3	Các bước thực hiện trong tiền xử lý dữ liệu	19
1.4	Các phương pháp xử lý dữ liệu phổ biến	20
2	Hàm Excel Phổ Biến Trong Tiền Xử Lý Dữ Liệu	21
3	Mô Hình Hồi Quy Tuyến Tính	22
3.1	Tại sao phải Phân Tích Hồi Quy?	23
3.2	Các Bước Phân Tích Hồi Quy Trong Excel	23
3.3	Thực hành	23
3.3.1	Hồi quy đơn biến	23
3.3.2	Hồi quy đa biến	23

I. Trực quan hóa dữ liệu

1 Trực quan hóa dữ liệu là gì?

Khi làm việc với các tệp dữ liệu phức tạp, việc trực quan hoá dữ liệu là một bước quan trọng để hiểu và phân tích các thông tin quan trọng. Vậy **Data Visualization** là gì?



Trước khi tìm hiểu Data Visualization là gì, chúng ta hãy phân tích ví dụ sau:

Thông thường khi gửi bảng báo cáo cho cấp trên, bạn có thể gửi một tập hợp các con số và thông tin chi tiết, điều này gây khó khăn cho việc đọc hiểu và phân tích. Tuy nhiên, nếu tệp dữ liệu này được trực quan hóa, kết quả sẽ hoàn toàn khác biệt. Thay vì chỉ nhìn vào những con số và chữ cái khô khan, Data Visualization cho phép chúng ta biểu diễn dữ liệu theo cách mà não bộ có thể dễ dàng nhận biết và xử lý.

Data Visualization là cách biểu diễn thông tin dưới dạng hình ảnh, biểu đồ hoặc đồ thị để minh họa dữ liệu dễ hiểu nhất. Mục đích của việc trực quan hóa dữ liệu là biến các nguồn dữ liệu thành những tệp thông tin trực quan, dễ quan sát và dễ hiểu.

Tại sao phải trực quan hóa dữ liệu?

Thực tế, việc trực quan hoá dữ liệu là rất cần thiết để người dùng có thể tiếp cận và xử lý dữ liệu dễ dàng hơn. Trực quan hóa dữ liệu cũng có thể giúp người dùng hiểu được những thông tin phức tạp hơn so với các bảng tính hoặc báo cáo. Chúng ta có thể hình dung trực quan hóa dữ liệu như là một ngôn ngữ để chia sẻ, ngay cả khi không được đào tạo chính thức và bài bản thì bạn vẫn có thể nắm vững được những thông tin cơ bản qua các biểu đồ.

Bộ não con người thường bị thu hút bởi những màu sắc và các biểu đồ hơn là các thông tin số liệu khô khan. Vì thế, chúng có thể dễ tiếp thu dữ liệu dưới dạng đồ thị hoặc biểu đồ để hình dung ra lượng lớn dữ liệu phức tạp.

Một số lợi ích của việc Trực quan hóa dữ liệu đó là:

- Nắm bắt dữ liệu nhanh chóng

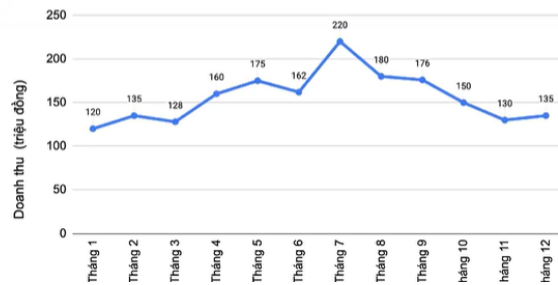
❌ Khó nắm bắt xu hướng

Tháng	Doanh thu (triệu đồng)
Tháng 1	120
Tháng 2	135
Tháng 3	128
Tháng 4	160
Tháng 5	175
Tháng 6	162
Tháng 7	220
Tháng 8	180
Tháng 9	176
Tháng 10	150
Tháng 11	130
Tháng 12	135



🕒 Dễ dàng nắm bắt xu hướng thay đổi doanh số theo mùa

Doanh thu có xu hướng tăng vào mùa hè và bắt đầu giảm khi chuyển sang thu và đông



- Ra quyết định chính xác
- Rút ngắn thời gian làm báo cáo
- Dễ dàng truyền đạt thông tin
- Cải thiện hiệu quả hoạt động sản xuất

2 Kiến thức cần biết

2.1 Thành phần của trực quan hóa dữ liệu

Trực quan hóa dữ liệu được hình thành từ 3 thành phần chính sau đây:









- **Thông điệp:** Trình bày mục đích của việc trực quan số liệu. Nhà quản lý sẽ làm việc và quyết định kết quả mong muốn đạt được sau khi phân tích dữ liệu. Ví dụ: Dự đoán doanh thu bán hàng hay đo lường hiệu suất làm việc của công - nhân viên.
- **Dữ liệu:** Sau khi xác định thông điệp, các nhà phân tích tiến hành xử lý dữ liệu: chỉnh sửa định dạng, làm sạch dữ liệu, loại bỏ thông tin không liên quan và phân tích kỹ lưỡng hơn. Sau đó, các phương thức trình bày dữ liệu trực quan sẽ được sử dụng giúp bộ phận chiến lược lên kế hoạch phù hợp.
- **Phương tiện trực quan:** Các nhà khoa học dữ liệu tạo ra các biểu đồ hoặc đồ thị để làm sinh động dữ liệu chính, đơn giản dữ liệu phức tạp nhằm chia sẻ thông tin chuyên sâu nhất. Các phương thức được cân nhắc sử dụng để thông tin được liên mạch và dễ hiểu một cách có hệ thống.

2.2 Các bước thực hiện trực quan hóa dữ liệu

Quá trình thực hiện trực quan dữ liệu hóa bao gồm 5 bước.

1. **Xác định mục tiêu:** Xác định các vấn đề cần trả lời.
2. **Thu thập dữ liệu:** Tổng hợp dữ liệu nội bộ và bên ngoài liên quan đến mục tiêu xác định.
3. **Chọn lọc dữ liệu:** Loại bỏ dữ liệu thừa, không liên quan, thực hiện các phép tính toán để phân tích và chuyển đổi loại dữ liệu để có thể sử dụng.
4. **Lựa chọn phương tiện trực quan hóa:** Có rất nhiều loại biểu đồ giúp trình bày dữ liệu hiệu quả. Người trình bày có thể dựa vào mối quan hệ giữa các điểm dữ liệu và thông tin muốn thể hiện để chọn.

2.3 Một số phương pháp trực quan hóa dữ liệu điển hình

	Biểu đồ cột (Bar Chart) So sánh giá trị giữa các nhóm		Biểu đồ đường (Line Chart) Hiện thị xu hướng theo thời gian
	Biểu đồ tròn (Pie Chart) Thể hiện tỷ lệ các phần trong tổng thể		Biểu đồ phân tán (Scatter Plot) Thể hiện mối liên hệ giữa hai biến số
	Biểu đồ hộp (Box Plot) Thể hiện phân phối dữ liệu và phát hiện giá trị bất thường		Heatmap Biểu diễn dữ liệu bằng màu sắc, thể hiện mức độ quan hệ
	Biểu đồ Phân Phối (Histogram) Thể hiện tần suất xuất hiện của dữ liệu		Biểu đồ Water Flow Thể hiện sự thay đổi giá trị qua các giai đoạn

2.3.1 Biểu đồ cột- Bar plot

Thường được sử dụng để **so sánh** các nhóm dữ liệu khác nhau trong cùng một giai đoạn hay các giai đoạn khác nhau của cùng một dữ liệu. Có 4 loại biểu đồ thường được sử dụng :

- vertical bar
- hozirontal bar
- stacked vertical bar
- stacked hozirontal bar

[Xem thêm](#)

2.3.2 Biểu đồ đường - Line Chart

Thường được sử dụng để **xem xu hướng** theo thời gian, so sánh nhiều loại dữ liệu. [Xem thêm](#)

2.3.3 Biểu đồ tròn - Pie Chart

Thường được sử dụng để **hiển thị tỷ lệ %**, tối ưu hóa 7 phần. [Xem thêm](#)

2.3.4 Biểu đồ phân phối - Histograms

Thường được sử dụng để **phân bố tần suất**, phát hiện ngoại lai, kiểm tra tính chuẩn, xác định mô hình thống kê. [Xem thêm](#)

2.3.5 Heatmap

Thường được sử dụng để **xu hướng** thông qua màu sắc. [Xem thêm](#)

2.3.6 Các nguyên tắc chọn phương tiện trực quan hóa dữ liệu

Trong trực quan hóa dữ liệu, đơn giản vẫn luôn là tốt. Dữ liệu nên được trình bày một cách đơn giản và hiệu quả để người xem có thể dễ dàng nắm bắt thông tin mà không cần phải xem lại nhiều lần hay suy nghĩ quá nhiều. Cần hạn chế việc sử dụng các yếu tố phức tạp và không rõ ràng, đồng thời tăng cường tính minh bạch của thông tin bằng cách sử dụng tiêu đề, chú thích, và các hình ảnh giải thích rõ ràng.

- Chọn đúng loại biểu đồ: chọn theo mục tiêu.
- Đơn giản hóa nội dung và hiệu quả
- Gắn nhãn rõ ràng: đảm bảo ai cũng có thể hiểu đúng
- Sử dụng màu sắc phù hợp

II. Kiểm Định Giả Thuyết Là Gì

Kiểm định giả thuyết là một công cụ cơ bản được sử dụng trong nghiên cứu khoa học để xác nhận hoặc bác bỏ các giả thuyết về các tham số quần thể dựa trên dữ liệu mẫu. Nó cung cấp một khuôn khổ có cấu trúc để đánh giá ý nghĩa thống kê của một giả thuyết và rút ra kết luận về bản chất thực sự của một quần thể. Kiểm định giả thuyết được sử dụng rộng rãi trong các lĩnh vực như sinh học, tâm lý học, kinh tế và kỹ thuật để xác định hiệu quả của các phương pháp điều trị mới, khám phá mối quan hệ giữa các biến số và đưa ra quyết định dựa trên dữ liệu. Tuy nhiên, mặc dù có tầm quan trọng, kiểm định giả thuyết có thể là một chủ đề khó hiểu và khó áp dụng đúng cách.

1 Khái niệm & tầm quan trọng

Kiểm định giả thuyết là quá trình thống kê để xác định liệu có đủ bằng chứng ủng hộ hay bác bỏ một giả thuyết hay không từ dữ liệu.

Tại Sao Phải Kiểm Định Giả Thuyết?

1 Đưa ra quyết định khách quan

Thay thế quyết định cảm tính bằng bằng chứng thống kê, giảm thiểu sai lầm chủ quan.

2 Xác định ý nghĩa thống kê

Phân biệt kết quả ngẫu nhiên với hiệu ứng có ý nghĩa thực sự.

3 Đánh giá mối quan hệ và sự khác biệt

Xác định mối liên hệ giữa các biến và phát hiện khác biệt đáng kể giữa các nhóm.

4 Xác minh hiệu quả can thiệp

Đánh giá liệu phương pháp mới có tạo ra khác biệt có ý nghĩa so với phương pháp hiện tại.

Trong kiểm định giả thuyết thống kê:

- H_0 là giả thiết không đưa ra tác dụng hoặc không có sự khác biệt.
- H_1 là giả thuyết cho thấy có sự khác biệt hoặc tác động đáng kể.
- Ví dụ:
 H_0 : Loại thuốc mới không có tác dụng chữa bệnh ung thư.
 H_1 : Loại thuốc mới chữa khỏi bệnh ung thư.

2 Các khái niệm cốt lõi

2.1 Miền bác bỏ

Một trong những cách giải quyết bài toán kiểm định giả thuyết là dùng một thống kê G , được gọi là **tiêu chuẩn thống kê**.

Định nghĩa: Thống kê $T = G(X_1, X_2, \dots, X_n)$ được gọi là một tiêu chuẩn thống kê (*test statistics*) nếu giá trị của nó được dùng để xem xét bác bỏ hay chấp nhận giả thuyết H_0 .

Ứng với mẫu cụ thể quan sát được, giá trị của tiêu chuẩn thống kê T được ký hiệu là t_{qs} . Ta sẽ dựa vào giá trị này để đưa ra kết luận chấp nhận hay bác bỏ giả thuyết đang xét bằng cách so sánh giá trị đó với miền tiêu chuẩn.

Miền W trong \mathbb{R} được gọi là *miền bác bỏ* hay *miền tiêu chuẩn* nếu miền này được dùng cùng với tiêu chuẩn thống kê T và giá trị cụ thể t_{qs} của tiêu chuẩn đó để đưa ra kết luận về giả thuyết H_0 :

- Nếu $t_{qs} \in W$ thì bác bỏ giả thuyết H_0 .
- Ngược lại, nếu $t_{qs} \in W^c$ thì chấp nhận H_0 .

Khi bác bỏ hay chấp nhận giả thuyết H_0 thì ta gặp phải hai loại sai lầm:

- **Sai lầm loại I:** Bác bỏ giả thuyết H_0 nhưng thực tế H_0 là đúng.
- **Sai lầm loại II:** Chấp nhận giả thuyết H_0 nhưng thực tế H_0 là sai.

Quyết định bác bỏ hay chấp nhận giả thuyết hoàn toàn dựa vào thông tin mẫu, do đó ta sẽ có xác suất mắc sai lầm loại I và sai lầm loại II. Ký hiệu α là xác suất mắc sai lầm loại I:

$$\alpha = P(\text{Sai lầm loại I}) = P(\text{bác bỏ } H_0 \mid H_0 \text{ đúng}).$$

Lúc đó α được gọi là **mức ý nghĩa**. Ký hiệu β là xác suất mắc sai lầm loại II:

$$\beta = P(\text{Sai lầm loại II}) = P(\text{chấp nhận } H_0 \mid H_0 \text{ sai}) = P(\text{chấp nhận } H_0 \mid H_1 \text{ đúng}).$$

Trường hợp đặc biệt, khi dùng tiêu chuẩn T và miền bác bỏ W để tiến hành kiểm định giả thuyết, ta sẽ có:

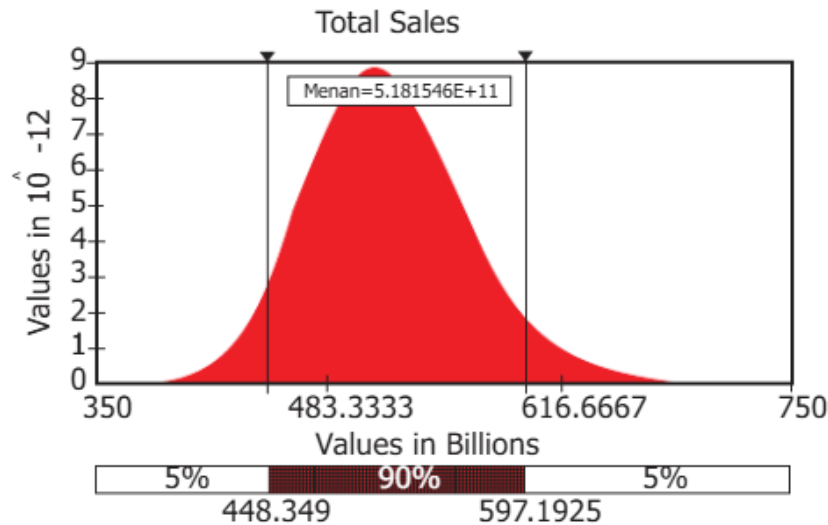
$$\alpha = P(T \in W \mid H_0)$$

$$\beta = P(T \in W^c \mid H_1).$$

Khi tiến hành kiểm định, người ta luôn mong muốn sao cho có thể cực tiểu hoá cả hai loại sai lầm loại I và loại II, tuy nhiên khi cỡ mẫu cố định thì mong muốn trên là không thực hiện được. Vì nói chung sai lầm loại I giảm xuống sẽ kéo theo sai lầm loại II tăng lên. Do đó, trong thực hành người ta thường cố định xác suất mắc sai lầm loại I và tìm cách cực tiểu xác suất mắc sai lầm loại II.

2.2 Các bước làm bài toán kiểm định

Để tiến hành kiểm định giả thuyết, thông thường người ta có thể sử dụng miền tiêu chuẩn, xác suất ý nghĩa hoặc ước lượng khoảng của các tiêu chuẩn hay tham số thống kê, với các bước thực hiện tương ứng.



2.2.1 Sử dụng miền tiêu chuẩn

- Bước 1: Xác định tham số cần kiểm định, đặt giả thuyết và đối thuyết.
- Bước 2: Xác định tiêu chuẩn thống kê và tính giá trị của tiêu chuẩn thống kê đối với giá trị mẫu đã cho.
- Bước 3: Xác định miền bác bỏ W .
- Bước 4: So sánh giá trị của tiêu chuẩn thống kê với miền bác bỏ W và kết luận bác bỏ hay chấp nhận giả thuyết H_0

2.2.2 Sử dụng xác suất ý nghĩa p-value

Trong kiểm định giả thuyết, ta đặt H_0 (không khác biệt) và H_1 (có khác biệt). Sau khi tính được *thống kê kiểm định* (z , t , χ^2 , ...), ta định nghĩa **p-value** là xác suất (giả sử H_0 đúng) để quan sát được kết quả *ít nhất* cực đoan như dữ liệu hiện có.

- Bước 1: Như trên
- Bước 2: Như trên
- Bước 3: Tính xác suất ý nghĩa p-value tương ứng với giá trị cụ thể của tiêu chuẩn thống kê đã có ở Bước 2.
- Bước 4: So sánh p-value với mức ý nghĩa đã định trước, thường là 5%, 1%, 0.1% và 0.5%

2.2.3 Sử dụng ước lượng khoảng

- Bước 1: Như trên.
- Bước 2: Xác định tiêu chuẩn thống kê và tìm *khoảng tin cậy* (ước lượng khoảng) của tiêu chuẩn đó (hoặc của tham số cần quan tâm) ứng với mẫu đã có và độ tin cậy đã định trước.
- Bước 3: So sánh khoảng tin cậy trên với một giá trị đã định.
 - Nếu khoảng tin cậy **không chứa** giá trị đó \Rightarrow **bác bỏ giả thuyết**.
 - Nếu khoảng tin cậy **có chứa** giá trị đó \Rightarrow **chấp nhận giả thuyết**.

2.3 Tổng kết

Trong thống kê, kiểm định giả thuyết (Hypothesis Testing) là công cụ quan trọng để đưa ra quyết định dựa trên dữ liệu. Tuy nhiên, việc nhớ hết các bước, công thức và tình huống áp dụng có thể khiến bạn choáng ngợp.

Để giúp việc ôn tập trở nên nhanh chóng và hiệu quả, **cheat sheet** này tổng hợp các khái niệm cốt lõi, công thức quan trọng và ví dụ minh họa cụ thể. Bạn có thể dùng nó như một bản đồ nhỏ – khi cần kiểm định giả thuyết, chỉ cần mở ra là biết ngay: bắt đầu từ đâu, công thức nào cần dùng, và cách kết luận ra sao.



Hypothesis Testing Cheat Sheet

Hypothesis testing

A hypothesis is a statement made about the value of a population parameter. It can be tested by carrying out an experiment or taking a sample from the population. The statistic calculated from the sample is called the test statistic.

The null hypothesis (H_0) is the hypothesis assumed to be correct. This is rejected if the test statistics is lower than a given threshold, called the significance level.

The alternative hypothesis (H_1) tells us about the parameter if your assumption is shown to be wrong.

Example 1: John wants to see if a coin is unbiased or biased towards coming down heads. He tosses the coin 8 times and counts the number of heads, X , obtained in 8 tosses.

- Describe the test statistic.
The test statistic is X , the number of heads obtained in 8 tosses.
- Write down a suitable null hypothesis.
The probability of landing heads for an unbiased coin is 0.5 so $H_0: p = 0.5$
- Write down a suitable alternative hypothesis.
The probability for heads is greater than 0.5 if the coin is biased towards heads so: $H_1: p > 0.5$

Finding critical values

A critical region is a region of the probability distribution which, if the test statistic falls within it, would cause you to reject the null hypothesis. The critical value is the first value to fall inside of the critical region.

The actual significance level of a hypothesis test is the probability of incorrectly rejecting the null hypothesis.

Example 2: A single observation is taken from the binomial distribution $B(6, p)$. The observation is used to test $H_0: p = 0.35$ against $H_1: p > 0.35$

- Using a 5% significance level, find the critical region for this test.
Assume H_0 is true then $X \sim B(6, 0.35)$
 $P(X \geq 4) = 1 - P(X \leq 3)$
 $= 1 - 0.8826$
 $= 0.1174$
 $P(X \geq 5) = 1 - P(X \leq 4)$
 $= 1 - 0.9777$
 $= 0.0223$
The critical region is 5 or 6.
- State the actual significance level of this test.
 $P(\text{reject null hypothesis}) = P(X \geq 5)$
 $= 0.0223$
 $= 2.23\%$

You can use the cumulative binomial tables or your calculator

This is the same as the probability of X falling within the critical region

One-tailed test

A one-tailed test can be used to test if the probability has increased or decreased.

For one-tailed tests,

$$H_1: p > \dots \text{ or } p < \dots$$

Example 3: The standard treatment for a particular disease has a $\frac{1}{2}$ probability of success. A researcher has produced a new drug which has been successful with 11 out of 20 patients. He claims that the new drug is more effective than the standard treatment. Test, at 5% significance level, the claim made by the researcher.

- Define your test statistic, X and parameter, p .
 X is the number of patients in the trial for whom the drug was successful. p is the probability of success for each patient.
- Formulate a model for the test statistic.
 $X \sim B(20, p)$
- Identify your null and alternative hypotheses.
 $H_0: p = 0.4$
 $H_1: p > 0.4$
- Method 1:
Assume H_0 is true and calculate the probability of 11 or more successful treatments
 $X \sim B(20, 0.4)$
 $P(X \geq 11) = 1 - P(X \leq 10)$
 $= 1 - 0.8725$
 $= 0.1275$
 $= 12.75\%$
- Compare probability with significance level.
 $12.75\% > 5\%$ so, there is not enough evidence to reject H_0
- Write a conclusion in context.
The new drug is no better than the old one.

The researcher claims that the new drug is better so $p > 0.4$

OR

- Method 2:
Work out the critical region and see if 11 lies within it.
 $P(X \geq 13) = 1 - P(X \leq 12)$
 $= 0.021$
 $P(X \geq 12) = 1 - P(X \leq 11)$
 $= 0.0565$
The critical region is 13 or more. Since 11 is not in the critical region, we accept H_0 .
- Write a conclusion in context of the question.
There is no evidence that the new drug is better than the old one.

Edexcel Stats/Mech Year 1

Two-tailed Test

A two-tailed test is used to test if the probability is changed in either direction. The critical region is split at either end of distribution. The significance level at each end is halved.

For two-tailed tests,

$$H_1: p \neq \dots$$

Example 4: In Enrico's restaurant, the ratio of non-vegetarian to vegetarian meals is found to be 2 to 1. In Manuel's restaurant in a random sample of 10 people ordering meals, 1 ordered a vegetarian meal. Using a 5% significance level, test whether the proportion of people eating vegetarian meals in Manuel's restaurant is different from Enrico's restaurant.

- The proportion of people eating vegetarian meals at Enrico's is $\frac{1}{3}$.
- X is the number of people in the sample at Manuel's restaurant who ordered vegetarian meals.
 p is the probability that a randomly chosen person at Manuel's orders a vegetarian meal.
- $H_0: p = \frac{1}{3}$, $H_1: p \neq \frac{1}{3}$
If H_0 is true, $X \sim B(10, \frac{1}{3})$
- Method 1:
 $P(X \leq 1) = P(X = 0) + P(X = 1)$
 $= \binom{10}{0} (\frac{1}{3})^0 (\frac{2}{3})^{10} + 10 \binom{10}{1} (\frac{1}{3})^1 (\frac{2}{3})^9$
 $= 0.01734... + 0.08670...$
 $= 0.104$ (3s.f.)

$0.104 > 0.025$ so insufficient evidence to reject H_0 .

Method 2:
Let c_1 and c_2 be the two critical values.
 $P(X \leq c_1) \leq 0.025$ and $P(X \geq c_2) \leq 0.025$

For lower tail:
 $P(X \leq 0) = 0.017341... < 0.025$
 $P(X \leq 1) = 0.10404... > 0.025$
So $c_1 = 0$

For upper tail:
 $P(X \geq 6) = 1 - P(X \leq 5)$
 $= 0.07656... > 0.025$
 $P(X \geq 7) = 1 - P(X \leq 6)$
 $= 0.01966... < 0.025$
So $c_2 = 7$

Observed value of 1 is not in critical region so H_0 is not rejected.

- Conclusion: There is no evidence that proportion of vegetarian meals at Manuel's restaurant is different from Enrico's.

Hình 1: Nguồn: PMT Tuition

3 Kiểm định giả thuyết thống kê bằng Python

Mặc dù có hàng trăm bài kiểm định giả thuyết thống kê mà bạn có thể sử dụng, nhưng chỉ có một tập hợp con nhỏ mà bạn có thể cần sử dụng trong một dự án học máy.

Các bạn có thể xem thêm tại [Statistical Methods for Machine Learning](#)

3.1 Kiểm định tính chuẩn

Sử dụng để kiểm tra xem dữ liệu của bạn có phân phối chuẩn Gauss hay không.

3.1.1 Shapiro-Wilk Test

Mục tiêu: Kiểm tra xem mẫu dữ liệu có phân phối chuẩn Gauss hay không. **Giả định:** Các quan sát trong mỗi mẫu là độc lập và phân bố giống hệt nhau (iid).

- H_0 : mẫu có phân phối chuẩn Gauss.
- H_1 : mẫu không có phân phối chuẩn Gauss.

```

1 # Example of the Shapiro-Wilk Normality Test
2 from scipy.stats import shapiro
3 data = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
4 stat, p = shapiro(data)
5 print('stat=%.3f, p=%.3f' % (stat, p))
6 if p > 0.05:
7     print('Probably Gaussian')
8 else:
9     print('Probably not Gaussian')

```

3.1.2 D'Agostino's K^2 Test

Mục tiêu: Kiểm tra một mẫu dữ liệu có phân phối chuẩn (Gaussian) hay không.

Giả định:

- Các quan sát độc lập và phân phối giống hệt nhau (i.i.d).

Diễn giải:

- H_0 : mẫu tuân theo phân phối Gaussian.
- H_1 : mẫu không tuân theo phân phối Gaussian.

Python code:

```

1 # Example of the D'Agostino's K^2 Normality Test
2 from scipy.stats import normaltest
3
4 data = [0.873, 2.817, 0.121, -0.945, -0.055,
5         -1.436, 0.360, -1.478, -1.637, -1.869]
6
7 stat, p = normaltest(data)
8 print('stat=%.3f, p=%.3f' % (stat, p))
9
10 if p > 0.05:
11     print('Probably Gaussian')
12 else:
13     print('Probably not Gaussian')

```

3.1.3 Anderson-Darling Test

Mục tiêu: Kiểm tra một mẫu dữ liệu có phân phối Gaussian hay không.

Giả định:

- Các quan sát độc lập và phân phối giống hệt nhau (i.i.d).

Diễn giải:

- H_0 : mẫu tuân theo phân phối Gaussian.
- H_1 : mẫu không tuân theo phân phối Gaussian.

Python code:

```

1 # Example of the Anderson-Darling Normality Test
2 from scipy.stats import anderson
3
4 data = [0.873, 2.817, 0.121, -0.945, -0.055,
5         -1.436, 0.360, -1.478, -1.637, -1.869]
6
7 result = anderson(data)
8 print('stat=%.3f' % (result.statistic))
9
10 for i in range(len(result.critical_values)):
11     sl, cv = result.significance_level[i], result.critical_values[i]
12     if result.statistic < cv:
13         print('Probably Gaussian at the %.1f%% level' % (sl))
14     else:
15         print('Probably not Gaussian at the %.1f%% level' % (sl))

```

3.2 Kiểm định tương quan - Correlation Tests

Mục tiêu: kiểm tra xem hai mẫu có liên quan hay không.

3.2.1 Hệ số tương quan Pearson

Kiểm tra xem hai mẫu có mối quan hệ tuyến tính hay không.

Giả định:

- Các quan sát trong mỗi mẫu là độc lập và phân bố giống hệt nhau (iid).
- Các quan sát trong mỗi mẫu đều có phân phối chuẩn.
- Các quan sát trong mỗi mẫu có cùng phương sai.

Diễn giải:

- H_0 : hai mẫu độc lập.
- H_1 : có sự phụ thuộc giữa các mẫu.

```

1 # Example of the Pearson's Correlation test
2 from scipy.stats import pearsonr
3 data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
4 data2 = [0.353, 3.517, 0.125, -7.545, -0.555, -1.536, 3.350, -1.578, -3.537, -1.579]
5 stat, p = pearsonr(data1, data2)
6 print('stat=%.3f, p=%.3f' % (stat, p))
7 if p > 0.05:
8     print('Probably independent')
9 else:
10    print('Probably dependent')

```

3.2.2 Hệ số tương quan thứ hạng của Spearman

Kiểm tra xem hai mẫu có mối quan hệ đơn điệu hay không.

Giả định:

- Các quan sát trong mỗi mẫu là độc lập và phân bố giống hệt nhau (iid).
- Các quan sát trong mỗi mẫu có thể được xếp hạng.

Diễn giải:

- H_0 : hai mẫu độc lập.
- H_1 : có sự phụ thuộc giữa các mẫu.

```

1 # Example of the Spearman's Rank Correlation Test
2 from scipy.stats import spearmanr
3 data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
4 data2 = [0.353, 3.517, 0.125, -7.545, -0.555, -1.536, 3.350, -1.578, -3.537, -1.579]
5 stat, p = spearmanr(data1, data2)
6 print('stat=%.3f, p=%.3f' % (stat, p))
7 if p > 0.05:
8     print('Probably independent')
9 else:
10    print('Probably dependent')

```

3.2.3 Hệ số tương quan thứ hạng của Kendall

Kiểm tra xem hai mẫu có mối quan hệ đơn điệu hay không.

Giả định:

- Các quan sát trong mỗi mẫu là độc lập và phân bố giống hệt nhau (iid).
- Các quan sát trong mỗi mẫu có thể được xếp hạng.

Diễn giải:

- H_0 : hai mẫu độc lập.
- H_1 : có sự phụ thuộc giữa các mẫu.

```

1 # Example of the Kendall's Rank Correlation Test
2 from scipy.stats import kendalltau
3 data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
4 data2 = [0.353, 3.517, 0.125, -7.545, -0.555, -1.536, 3.350, -1.578, -3.537, -1.579]
5 stat, p = kendalltau(data1, data2)
6 print('stat=%.3f, p=%.3f' % (stat, p))
7 if p > 0.05:
8     print('Probably independent')
9 else:
10    print('Probably dependent')

```

3.2.4 Kiểm định Chi-Squared

Kiểm tra xem hai biến phân loại có liên quan hay độc lập hay không.

Giả định:

- Các quan sát được sử dụng trong tính toán bảng dự phòng là độc lập.
- 25 ví dụ trở lên trong mỗi ô của bảng dự phòng.

Diễn giải:

- H_0 : hai mẫu độc lập.
- H_1 : có sự phụ thuộc giữa các mẫu.

```

1 # Example of the Chi-Squared Test
2 from scipy.stats import chi2_contingency
3 table = [[10, 20, 30], [6, 9, 17]]
4 stat, p, dof, expected = chi2_contingency(table)
5 print('stat=%.3f, p=%.3f' % (stat, p))
6 if p > 0.05:
7     print('Probably independent')
8 else:
9     print('Probably dependent')

```

3.3 Kiểm định giả thuyết thống kê tham số

3.3.1 Kiểm định t-student

Kiểm tra xem giá trị trung bình của hai mẫu độc lập có khác biệt đáng kể hay không.

Giả định:

- Các quan sát trong mỗi mẫu là độc lập và phân bố giống hệt nhau (iid).
- Các quan sát trong mỗi mẫu đều có phân phối chuẩn.
- Các quan sát trong mỗi mẫu có cùng phương sai.

Diễn giải:

- H_0 : giá trị trung bình của các mẫu là bằng nhau.
- H_1 : giá trị trung bình của các mẫu không bằng nhau.

```

1 # Example of the Student's t-test
2 from scipy.stats import ttest_ind
3 data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
4 data2 = [1.142, -0.432, -0.938, -0.729, -0.846, -0.157, 0.500, 1.183, -1.075, -0.169]
5 stat, p = ttest_ind(data1, data2)
6 print('stat=%.3f, p=%.3f' % (stat, p))
7 if p > 0.05:
8     print('Probably the same distribution')
9 else:
10    print('Probably different distributions')

```

3.3.2 Kiểm định t-Paired Student

Kiểm tra xem giá trị trung bình của hai mẫu ghép đôi có khác biệt đáng kể hay không.

Giả định:

- Các quan sát trong mỗi mẫu là độc lập và phân bố giống hệt nhau (iid).
- Các quan sát trong mỗi mẫu đều có phân phối chuẩn.
- Các quan sát trong mỗi mẫu có cùng phương sai.
- Các quan sát trên mỗi mẫu được ghép nối.

Diễn giải:

- H_0 : giá trị trung bình của các mẫu là bằng nhau.
- H_1 : giá trị trung bình của các mẫu không bằng nhau.

```

1 # Example of the Paired Student's t-test
2 from scipy.stats import ttest_rel
3 data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
4 data2 = [1.142, -0.432, -0.938, -0.729, -0.846, -0.157, 0.500, 1.183, -1.075, -0.169]
5 stat, p = ttest_rel(data1, data2)
6 print('stat=%.3f, p=%.3f' % (stat, p))
7 if p > 0.05:
8     print('Probably the same distribution')
9 else:
10    print('Probably different distributions')

```

3.3.3 Kiểm định t-ANOVA

Kiểm tra xem giá trị trung bình của hai hoặc nhiều mẫu độc lập có khác biệt đáng kể hay không.

Giả định:

- Các quan sát trong mỗi mẫu là độc lập và phân bố giống hệt nhau (iid).
- Các quan sát trong mỗi mẫu đều có phân phối chuẩn.
- Các quan sát trong mỗi mẫu có cùng phương sai.

Diễn giải:

- H_0 : giá trị trung bình của các mẫu là bằng nhau.
- H_1 : một hoặc nhiều giá trị trung bình của các mẫu không bằng nhau.

```

1 # Example of the Analysis of Variance Test
2 from scipy.stats import f_oneway
3 data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
4 data2 = [1.142, -0.432, -0.938, -0.729, -0.846, -0.157, 0.500, 1.183, -1.075, -0.169]
5 data3 = [-0.208, 0.696, 0.928, -1.148, -0.213, 0.229, 0.137, 0.269, -0.870, -1.204]
6 stat, p = f_oneway(data1, data2, data3)
7 print('stat=%.3f, p=%.3f' % (stat, p))
8 if p > 0.05:
9     print('Probably the same distribution')
10 else:
11    print('Probably different distributions')

```

3.3.4 Kiểm định giả thuyết thống kê phi tham số

subsectionKiểm định Mann-Whitney U Kiểm tra xem phân phối của hai mẫu độc lập có bằng nhau hay không.

Giả định:

- Các quan sát trong mỗi mẫu là độc lập và phân bố giống hệt nhau (iid).
- Các quan sát trong mỗi mẫu có thể được xếp hạng.

Diễn giải:

- H_0 : phân phối của cả hai mẫu đều bằng nhau.
- H_1 : sự phân phối của cả hai mẫu không bằng nhau.

```

1 # Example of the Mann-Whitney U Test
2 from scipy.stats import mannwhitneyu
3 data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
4 data2 = [1.142, -0.432, -0.938, -0.729, -0.846, -0.157, 0.500, 1.183, -1.075, -0.169]
5 stat, p = mannwhitneyu(data1, data2)
6 print('stat=%.3f, p=%.3f' % (stat, p))
7 if p > 0.05:
8     print('Probably the same distribution')
9 else:
10    print('Probably different distributions')

```

3.3.5 Kiểm định Wilcoxon có dấu hiệu-hạng

Kiểm tra sự phân phối của hai mẫu ghép đôi có bằng nhau hay không.

Giả định:

- Các quan sát trong mỗi mẫu là độc lập và phân bố giống hệt nhau (iid).
- Các quan sát trong mỗi mẫu có thể được xếp hạng.
- Các quan sát trên mỗi mẫu được ghép nối.

Diễn giải:

- H_0 : phân phối của cả hai mẫu đều bằng nhau.
- H_1 : sự phân phối của cả hai mẫu không bằng nhau.

```

1 # Example of the Wilcoxon Signed-Rank Test
2 from scipy.stats import wilcoxon
3 data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
4 data2 = [1.142, -0.432, -0.938, -0.729, -0.846, -0.157, 0.500, 1.183, -1.075, -0.169]
5 stat, p = wilcoxon(data1, data2)
6 print('stat=%.3f, p=%.3f' % (stat, p))
7 if p > 0.05:
8     print('Probably the same distribution')
9 else:
10    print('Probably different distributions')

```

3.3.6 Kiểm định Kruskal-Wallis H

Kiểm tra xem phân phối của hai hoặc nhiều mẫu độc lập có bằng nhau hay không.

Giả định:

- Các quan sát trong mỗi mẫu là độc lập và phân bố giống hệt nhau (iid).
- Các quan sát trong mỗi mẫu có thể được xếp hạng.

Diễn giải:

- H_0 : phân phối của tất cả các mẫu đều bằng nhau.
- H_1 : sự phân phối của một hoặc nhiều mẫu không bằng nhau.


```
1 # Example of the Kruskal-Wallis H Test
2 from scipy.stats import kruskal
3 data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
4 data2 = [1.142, -0.432, -0.938, -0.729, -0.846, -0.157, 0.500, 1.183, -1.075, -0.169]
5 stat, p = kruskal(data1, data2)
6 print('stat=%.3f, p=%.3f' % (stat, p))
7 if p > 0.05:
8     print('Probably the same distribution')
9 else:
10    print('Probably different distributions')
```

3.3.7 Kiểm tra Friedman

Kiểm tra xem sự phân phối của hai hoặc nhiều mẫu ghép đôi có bằng nhau hay không.

Giả định:

- Các quan sát trong mỗi mẫu là độc lập và phân bố giống hệt nhau (iid).
- Các quan sát trong mỗi mẫu có thể được xếp hạng.
- Các quan sát trên mỗi mẫu được ghép nối.

Diễn giải:

- H_0 : phân phối của tất cả các mẫu đều bằng nhau.
- H_1 : sự phân phối của một hoặc nhiều mẫu không bằng nhau.

```
1 # Example of the Friedman Test
2 from scipy.stats import friedmanchisquare
3 data1 = [0.873, 2.817, 0.121, -0.945, -0.055, -1.436, 0.360, -1.478, -1.637, -1.869]
4 data2 = [1.142, -0.432, -0.938, -0.729, -0.846, -0.157, 0.500, 1.183, -1.075, -0.169]
5 data3 = [-0.208, 0.696, 0.928, -1.148, -0.213, 0.229, 0.137, 0.269, -0.870, -1.204]
6 stat, p = friedmanchisquare(data1, data2, data3)
7 print('stat=%.3f, p=%.3f' % (stat, p))
8 if p > 0.05:
9     print('Probably the same distribution')
10 else:
11    print('Probably different distributions')
```

4 Kiểm định giả thiết thống kê bằng Excel

Tổng hợp các hàm và công cụ Excel thường dùng cho kiểm định thống kê:

Loại kiểm định	Hàm/Công cụ Excel	Mô tả
Kiểm định t (t-Test)	Data Analysis > t-Test	So sánh trung bình của 2 tập dữ liệu (3 loại: paired, equal variance, unequal variance)
Kiểm định z (z-Test)	Data Analysis > z-Test	So sánh trung bình với mẫu lớn hoặc khi biết phương sai tổng thể
ANOVA	Data Analysis > ANOVA	So sánh trung bình của nhiều nhóm (single factor, two-factor)
Tương quan	CORREL(), Data Analysis > Correlation	Đo lường mối quan hệ tuyến tính giữa các biến
Hồi quy tuyến tính	Data Analysis > Regression	Phân tích mối quan hệ giữa biến phụ thuộc và biến độc lập
Thống kê mô tả	Data Analysis > Descriptive Statistics	Cung cấp các thống kê cơ bản (mean, median, mode, standard deviation, etc.)
Hàm phân phối chuẩn	NORM.DIST(), NORM.INV(), NORM.S.DIST()	Tính xác suất và giá trị của phân phối chuẩn
Hàm phân phối t	T.DIST(), T.INV(), T.TEST()	Tính xác suất và giá trị của phân phối t-Student
Hàm phân phối F	F.DIST(), F.INV(), F.TEST()	Tính xác suất và giá trị của phân phối F (dùng trong ANOVA)
Hàm phân phối Chi bình phương	CHISQ.DIST(), CHISQ.INV(), CHISQ.TEST()	Kiểm định tính độc lập giữa các biến phân loại
Phân tích p-value	Data Analysis (kết quả có sẵn)	So sánh p-value với mức ý nghĩa alpha để đưa ra kết luận

Revised at AUG 2025

44

III. Tiền Xử Lý Dữ Liệu & Phân Tích Hồi Quy

1 Tiền xử lý dữ liệu là gì?

1.1 Khái niệm

Tiền xử lý dữ liệu là bước đầu tiên trong quy trình xử lý dữ liệu, nhằm chuẩn bị dữ liệu thô để sẵn sàng cho các giai đoạn phân tích và xử lý tiếp theo. Dữ liệu thô thường có thể chứa lỗi, thiếu sót hoặc không đồng nhất, do đó, tiền xử lý dữ liệu giúp cải thiện chất lượng và tính toàn vẹn của dữ liệu.

Quá trình tiền xử lý dữ liệu bao gồm nhiều bước như làm sạch dữ liệu, chuẩn hóa định dạng, xử lý các giá trị bị thiếu và loại bỏ thông tin trùng lặp. Mục tiêu là đảm bảo dữ liệu đầu vào đạt được độ chính xác và nhất quán cao nhất, từ đó hỗ trợ các công cụ và mô hình phân tích hoạt động hiệu quả.

Hiểu rõ tiền xử lý dữ liệu là gì không chỉ giúp tối ưu hóa quy trình xử lý dữ liệu mà còn đóng vai trò quan trọng trong việc đưa ra những kết quả phân tích đáng tin cậy và có giá trị thực tiễn cao.

1.2 Vai trò

Tiền xử lý dữ liệu đóng vai trò thiết yếu trong toàn bộ quy trình xử lý dữ liệu, đảm bảo rằng dữ liệu thô được chuẩn bị tốt nhất để phục vụ các giai đoạn phân tích và xử lý tiếp theo. Dưới đây là những vai trò chính của tiền xử lý dữ liệu:

- Cải thiện chất lượng dữ liệu đầu vào
- Tăng hiệu quả xử lý dữ liệu
- Hỗ trợ phân tích chính xác hơn
- Tối ưu hóa sử dụng tài nguyên
- Tăng cường khả năng tích hợp dữ liệu



Làm sạch dữ liệu thô

Dữ liệu thô thường chứa giá trị thiếu, sai lệch và định dạng không đồng nhất cần được xử lý.



Đảm bảo độ chính xác

Chuẩn hóa và biến đổi dữ liệu thành định dạng phù hợp cho việc phân tích hiệu quả.



Hỗ trợ quyết định chính xác

Loại bỏ giá trị ngoại lệ và xử lý dữ liệu thiếu giúp ngăn chặn sai lệch và tối ưu quyết định kinh doanh.

1.3 Các bước thực hiện trong tiền xử lý dữ liệu

1. Thu thập dữ liệu

Dữ liệu được thu thập từ nhiều nguồn khác nhau như cơ sở dữ liệu, cảm biến, hệ thống trực tuyến hoặc tài liệu vật lý. Đây là bước đầu tiên và đóng vai trò quan trọng trong việc đảm bảo dữ liệu thô có chất lượng và phù hợp với mục tiêu.

2. Làm sạch dữ liệu

Loại bỏ các lỗi, giá trị bị thiếu, dữ liệu trùng lặp hoặc không cần thiết. Quá trình này giúp cải thiện độ chính xác và chất lượng của dữ liệu, đảm bảo tính toàn vẹn cho các bước xử lý sau.

3. Chuẩn hóa dữ liệu

Dữ liệu được chuyển đổi sang định dạng đồng nhất, chẳng hạn như chuẩn hóa đơn vị đo lường, mã hóa dữ liệu danh mục hoặc chuyển đổi kiểu dữ liệu. Điều này giúp giảm sự không nhất quán và dễ dàng hơn trong việc xử lý.

4. Tích hợp dữ liệu

Kết hợp dữ liệu từ nhiều nguồn khác nhau để tạo ra một tập dữ liệu thống nhất, đồng nhất. Đây là bước cần thiết trong các hệ thống phân tích yêu cầu tổng hợp thông tin từ nhiều nguồn.

5. Giảm kích thước dữ liệu

Loại bỏ các thông tin không quan trọng hoặc giảm số lượng biến số nhằm tối ưu hóa tài nguyên xử lý và thời gian thực hiện.

6. Kiểm tra dữ liệu đầu ra

Sau khi hoàn thành các bước trên, dữ liệu được kiểm tra để đảm bảo tính chính xác và sẵn sàng cho quá trình xử lý hoặc phân tích tiếp theo.

1.4 Các phương pháp xử lý dữ liệu phổ biến

1 Xử lý giá trị thiếu (Null values) Sử dụng IF, ISBLANK và IFERROR để phát hiện ô trống, thay thế bằng giá trị trung bình, trung vị hoặc phổ biến nhất.	2 Loại bỏ giá trị ngoại lệ (Outliers) Áp dụng phương pháp IQR với QUARTILE hoặc Z-score (STANDARDIZE) để xác định và xử lý giá trị nằm ngoài khoảng tin cậy.	3 Tạo biến giả (Dummy) cho biến phân loại Chuyển biến phân loại thành biến nhị phân (0/1) bằng IF, SWITCH hoặc IFS để cải thiện phân tích hồi quy.	4 Kiểm tra và đánh giá sau xử lý Sử dụng biểu đồ phân phối, PivotTable và thống kê mô tả để xác nhận dữ liệu đã được xử lý đúng.
--	---	---	---

Xử lý giá trị thiếu (Null values)

Xử lý đúng cách các giá trị thiếu là yếu tố quan trọng ảnh hưởng trực tiếp đến kết quả phân tích dữ liệu.

- **Phát hiện giá trị thiếu:** Sử dụng ISBLANK(), ISNA(), IFERROR() để xác định các ô trống trong dữ liệu.
- **Thay thế bằng giá trị thống kê:** Áp dụng AVERAGE, MEDIAN, hoặc MODE.SNGL để thay thế mà không gây sai lệch phân tích.
- **Lọc hoặc loại bỏ:** Với dữ liệu có ít giá trị thiếu, dùng Filter hoặc Advanced Filter để lọc hoặc loại bỏ chúng.
- **Dự đoán giá trị thiếu:** Dùng FORECAST hoặc phân tích hồi quy để ước tính giá trị dựa trên mối quan hệ với các biến khác.

Phát Hiện Giá Trị Ngoại Lệ Trong Excel

Giá trị ngoại lệ (outliers) là những quan sát có giá trị cực kỳ cao hoặc thấp, có thể làm sai lệch kết quả phân tích và dự báo.

- **Nhận diện giá trị ngoại lệ:** Sử dụng Box Plot, Histogram hoặc Scatter Plot để trực quan hóa outliers. Áp dụng nguyên tắc IQR: Giá trị ngoài khoảng ($Q1 - 1.5 \cdot IQR$) và ($Q3 + 1.5 \cdot IQR$) là ngoại lệ.

- **Phương pháp thống kê phát hiện outliers:** Dùng QUARTILE.EXC để tính Q1 và Q3, sau đó áp dụng công thức. Z-score: Sử dụng hàm STANDARDIZE, giá trị có $|z| > 3$ thường là ngoại lệ IQR.
- **Kỹ thuật xử lý outliers:** Loại bỏ: Dùng Filter để loại trừ giá trị ngoại lệ nếu chắc chắn là sai sót. Thay thế: Áp dụng MEDIAN hoặc phương pháp Winsorization.

Tạo biến giả (Dummy) cho biến phân loại

- **Khái niệm biến giả (Dummy):** Biến nhị phân (0/1) đại diện cho giá trị của biến phân loại, giúp đưa dữ liệu định tính vào mô hình định lượng.
- **Quy tắc tạo biến giả:** Với n giá trị phân loại, tạo (n-1) biến giả để tránh đa cộng tuyến. Giá trị 1 đại diện "có", 0 đại diện "không".
- **Tạo biến giả trong Excel:** Sử dụng IF, IFS hoặc SWITCH kết hợp công thức mảng và Power Query để tự động hóa tạo biến giả cho dữ liệu lớn.

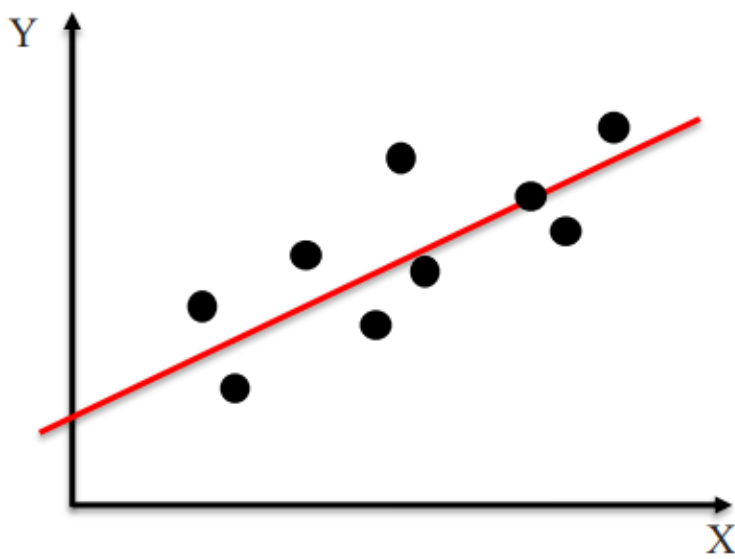
2 Hàm Excel Phổ Biến Trong Tiền Xử Lý Dữ Liệu

Phương pháp tiền xử lý	Hàm Excel/Công cụ	Mô tả
Xử lý giá trị thiếu (Null)	IFERROR(), IF(ISBLANK()), AVERAGE(), MEDIAN()	Phát hiện và điền giá trị thiếu bằng giá trị trung bình, trung vị hoặc giá trị dự đoán
Loại bỏ giá trị trùng lặp	Remove Duplicates, UNIQUE()	Loại bỏ các bản ghi trùng lặp trong tập dữ liệu
Phát hiện giá trị ngoại lệ	QUARTILE(), STDEV(), IF(), AVERAGEIF()	Xác định và xử lý các giá trị nằm ngoài khoảng bình thường
Chuẩn hóa dữ liệu	UPPER(), LOWER(), PROPER(), TRIM()	Thống nhất định dạng văn bản, loại bỏ khoảng trắng thừa
Tạo biến giả (Dummy)	IF(), IFS(), VLOOKUP()	Chuyển đổi biến phân loại thành biến nhị phân (0/1) hoặc mã hóa
Biến đổi dữ liệu	LN(), SQRT(), POWER(), LOG10()	Chuyển đổi phân phối dữ liệu (logarit, căn bậc hai, bình phương...)
Điều chỉnh thang đo	STANDARDIZE(), MIN(), MAX()	Chuẩn hóa dữ liệu về cùng thang đo (ví dụ: 0-1 hoặc z-score)
Tách cột dữ liệu	Text to Columns, LEFT(), RIGHT(), MID()	Phân tách dữ liệu từ một cột thành nhiều cột
Gộp dữ liệu	CONCATENATE(), &, TEXTJOIN()	Kết hợp dữ liệu từ nhiều cột thành một
Chuyển đổi định dạng thời gian	DATE(), DATEVALUE(), TEXT()	Chuẩn hóa các định dạng ngày tháng khác nhau

Phương Pháp	Công Cụ/Hàm Excel	Mô Tả	Ưu Điểm	Hạn Chế
Tra cứu dữ liệu	VLOOKUP(), HLOOKUP()	Tìm kiếm giá trị dựa trên cột khóa và trả về giá trị tương ứng từ bảng khác	Dễ sử dụng, phổ biến	Chỉ tra cứu từ trái sang phải, không linh hoạt với dữ liệu thay đổi
Tra cứu đa chiều	INDEX() + MATCH()	Kết hợp để tìm giá trị từ bảng khác dựa trên dòng và cột	Linh hoạt hơn VLOOKUP, tìm kiếm theo mọi hướng	Cú pháp phức tạp hơn, khó học hơn
Tra cứu nâng cao	XLOOKUP()	Hàm tra cứu hiện đại thay thế VLOOKUP và INDEX-MATCH	Linh hoạt, hỗ trợ tìm kiếm theo nhiều hướng, có giá trị mặc định	Chỉ có trong Excel 365 và các phiên bản mới hơn
Nối dữ liệu	Power Query (Get & Transform)	Kết nối và biến đổi dữ liệu từ nhiều nguồn	Tự động làm mới, xử lý được khối lượng dữ liệu lớn	Yêu cầu học thêm về cách sử dụng Power Query
Tạo bảng động	Pivot Table	Tổng hợp và phân tích dữ liệu từ nhiều bảng	Phân tích linh hoạt, tính toán tự động	Chủ yếu dùng để tổng hợp, không phải kết hợp dữ liệu chi tiết
Hợp nhất dữ liệu	Consolidate	Kết hợp dữ liệu từ nhiều vùng hay sheet	Dễ sử dụng với dữ liệu có cùng cấu trúc	Ít linh hoạt, khó xử lý dữ liệu phức tạp
Tạo quan hệ	Data Model	Thiết lập quan hệ giữa các bảng trong mô hình dữ liệu Excel	Mạnh mẽ, xử lý được khối lượng dữ liệu lớn, quan hệ phức tạp	Yêu cầu hiểu biết về mô hình dữ liệu, phức tạp hơn
Công thức mảng	FILTER(), UNIQUE(), SORT()	Sử dụng các hàm mảng động để kết hợp và lọc dữ liệu	Mạnh mẽ, xử lý được các điều kiện phức tạp	Chỉ có trong Excel 365, đòi hỏi hiểu biết về công thức mảng

3 Mô Hình Hồi Quy Tuyến Tính

Mô hình hồi quy tuyến tính là mô hình dùng để dự đoán giá trị của một biến (Y) dựa trên mối quan hệ với một hoặc nhiều biến độc lập (X).



Phương trình tuyến tính cơ bản:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \varepsilon$$

Các thành phần trong phương trình:

- Y: biến phụ thuộc (doanh thu)
- X: biến độc lập (giá, khuyến mãi,...)
- β : hệ số hồi quy (mức ảnh hưởng)
- ϵ : phần dư (sai số)

Đánh giá mô hình

- R^2 : cho biết mô hình giải thích bao nhiêu % biến động của Y
- p-value: Đánh giá ý nghĩa của từng biến X
- Phân tích phần dư: xem mô hình có vi phạm giả định không

3.1 Tại sao phải Phân Tích Hồi Quy?

Phân tích hồi quy là gì? Kỹ thuật thống kê xác định mối quan hệ giữa biến phụ thuộc (Y) và biến độc lập (X), giúp hiểu cách một biến thay đổi khi biến khác biến động.

- Hồi quy đơn biến: Phân tích quan hệ giữa 1 biến X và 1 biến Y (ví dụ: giá ảnh hưởng đến doanh thu)
- Hồi quy đa biến: Phân tích quan hệ giữa nhiều biến X và 1 biến Y (ví dụ: giá, quảng cáo, thời tiết, v.v.)

Tại sao sử dụng phân tích hồi quy?

- Dự báo: Ước tính giá trị tương lai từ dữ liệu quá khứ
- Phân tích nhân quả: Hiểu mối quan hệ giữa các yếu tố với kết quả
- Kiểm định giả thuyết: Xác minh giả thuyết về quan hệ giữa các biến
- Tối ưu hóa: Xác định giá trị tối ưu của biến đầu vào

3.2 Các Bước Phân Tích Hồi Quy Trong Excel

Quy trình phân tích hồi quy trong Excel bao gồm 4 bước chính:

- Bước 1: Bật Add-in Data Analysis Toolpak
Đảm bảo rằng bạn đã kích hoạt Data Analysis Toolpak trong Excel
- Bước 2: Chọn Regression
Trong Data Analysis, chọn Regression để bắt đầu phân tích.
- Bước 3: Thiết lập vùng dữ liệu Y (doanh thu) và X (biến ảnh hưởng)
Xác định phạm vi dữ liệu cho biến phụ thuộc (Y) và biến độc lập (X).
- Bước 4: Chạy mô hình và diễn giải kết quả
Phân tích các chỉ số quan trọng như R^2 , hệ số, p-value và sai số chuẩn để hiểu ý nghĩa thống kê và mức độ ảnh hưởng của mô hình.

3.3 Thực hành**3.3.1 Hồi quy đơn biến****3.3.2 Hồi quy đa biến**