

Excel for Data Analysis – Part II

Visualization - Hypothesis Testing- Modelling

Giảng viên: Nguyễn Thái Hà (Ph.D)

Trợ giảng: Nguyễn Thọ Anh Khoa (Ph.D Candidate)

Ngày: 16/08/2025 (Sat)

Phần 1 - Trục Quan Hóa Dữ Liệu

Nắm vững kỹ năng tạo, chọn và áp dụng các phương pháp hay nhất cho biểu đồ và bảng điều khiển trong Excel.

Phần 2 - Kiểm Định Giả Thuyết & A/B Testing

Hiểu và thực hiện các kiểm định thống kê (t-test, ANOVA) cũng như A/B Testing để đưa ra quyết định kinh doanh, sử dụng Data Analysis ToolPak trong Excel.

Phần 3 - Tiền Xử Lý Dữ Liệu & Phân Tích Hồi Quy

Tiền xử lý dữ liệu (làm sạch, xử lý giá trị thiếu/ngoại lai, tạo biến giả) và xây dựng, diễn giải, cải thiện các mô hình hồi quy tuyến tính trong Excel.

Phần 1: Trục Quan Hóa Dữ Liệu

Tạo biểu đồ và dashboard hiệu quả từ dữ liệu.

Phần 2: Kiểm Định Giả Thuyết (Hypothesis Testing)

Phương pháp đưa ra kết luận dựa trên kiểm định thống kê.

Phần 3: Tiền Xử Lý Dữ Liệu & Phân Tích Hồi Quy

Tiền xử lý dữ liệu, xây dựng các mô hình hồi quy đơn biến và đa biến

Phần 1: Trực Quan Hóa Dữ Liệu

Tạo biểu đồ và dashboard hiệu quả từ dữ liệu.

Phần 2: Kiểm Định Giả Thuyết (Hypothesis Testing)

Phương pháp đưa ra kết luận dựa trên kiểm định thống kê.

Phần 3: Tiền Xử Lý Dữ Liệu & Phân Tích Hồi Quy

Tiền xử lý dữ liệu, xây dựng các mô hình hồi quy đơn biến và đa biến

Vì Sao Cần Trực Quan Dữ Liệu?

Trực quan hóa giúp nhận diện xu hướng nhanh, truyền đạt thông tin hiệu quả và hỗ trợ ra quyết định dựa trên bằng chứng trực quan.

1 Nhận biết xu hướng nhanh chóng

Biểu đồ Excel chuyển số liệu thành hình ảnh, giúp phát hiện ngay xu hướng doanh số và mối tương quan mà bảng số không thể hiện rõ.

2 Truyền đạt thông điệp hiệu quả

Dashboard với biểu đồ trực quan giúp trình bày kết quả phân tích trong 2-3 phút thay vì 15 phút giải thích bảng số.

3 Hỗ trợ ra quyết định

Biểu đồ phân tán và nhiệt thể hiện tương quan dữ liệu, giúp đưa ra quyết định chính xác về phân bổ ngân sách và nguồn lực.

Doanh thu mặt hàng quần áo

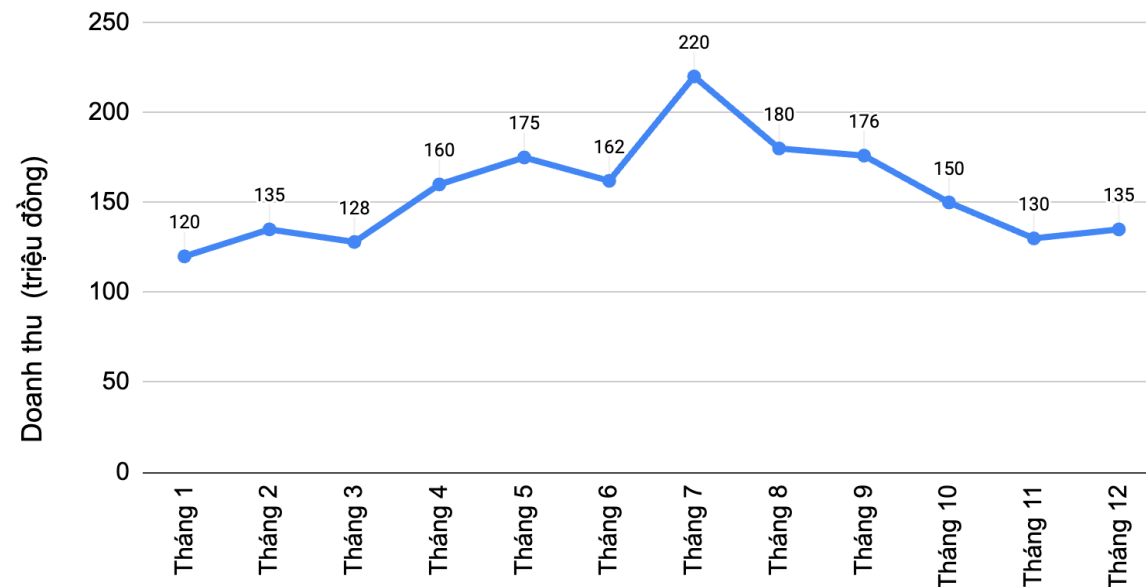
Trực quan hóa dữ liệu giúp chuyển đổi bảng số phức tạp thành biểu đồ dễ hiểu, cho phép nhận biết nhanh xu hướng và hỗ trợ ra quyết định hiệu quả.

✖ 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



Các Loại Biểu Đồ Thường Được Sử Dụng



Biểu đồ cột (Bar Chart)

So sánh giá trị giữa các nhóm



Biểu đồ đường (Line Chart)

Hiển 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

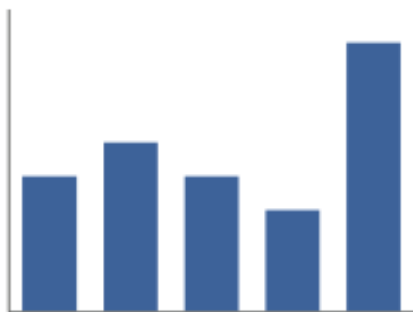


Biểu đồ Water Flow

Thể hiện sự thay đổi giá trị qua các giai đoạn

Biểu Đồ Cột (Bar Plot)

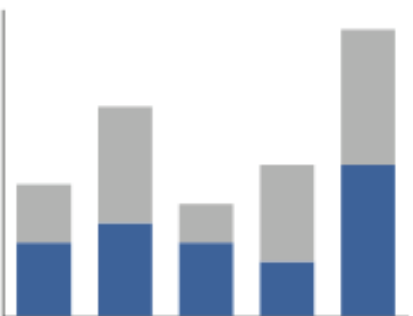
Có 4 loại biểu đồ cột được sử dụng thông dụng: vertical bar, horizontal bar, stacked vertical bar, stacked horizontal bar. Biểu đồ cột được sử dụng để so sánh giá trị giữa các danh mục, với chiều cao cột tương ứng với giá trị.



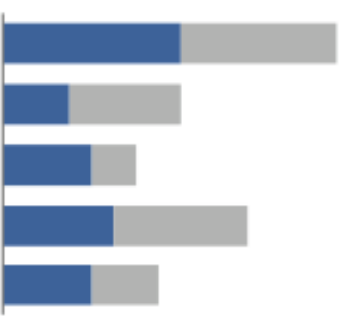
Vertical bar



Horizontal bar



Stacked vertical bar



Stacked horizontal bar

Dễ đọc

Trực quan cho mọi đối tượng

So sánh hiệu quả

Phân biệt rõ giữa các nhóm

Linh hoạt

Có thể so sánh nhiều biến cùng lúc

Tips Để Tạo Biểu Đồ Cột Hiệu Quả

Biểu đồ thiết kế không tốt

Survey results: summer learning program on science

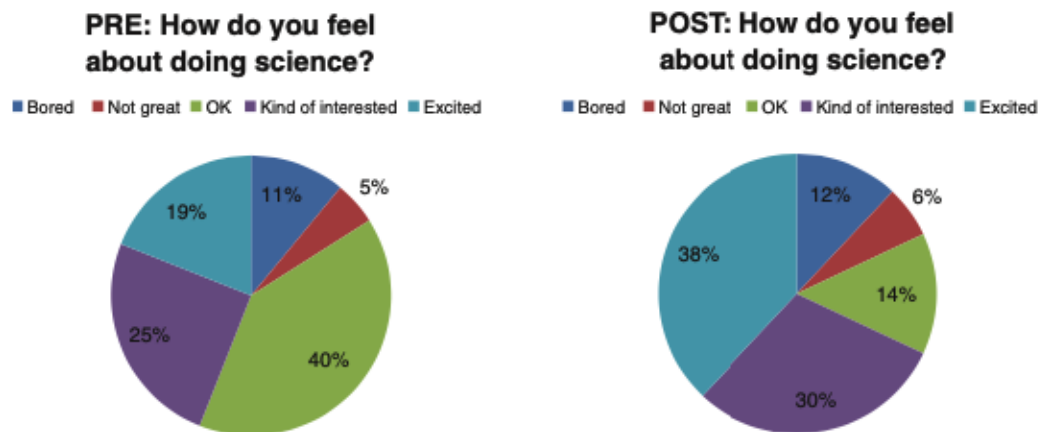


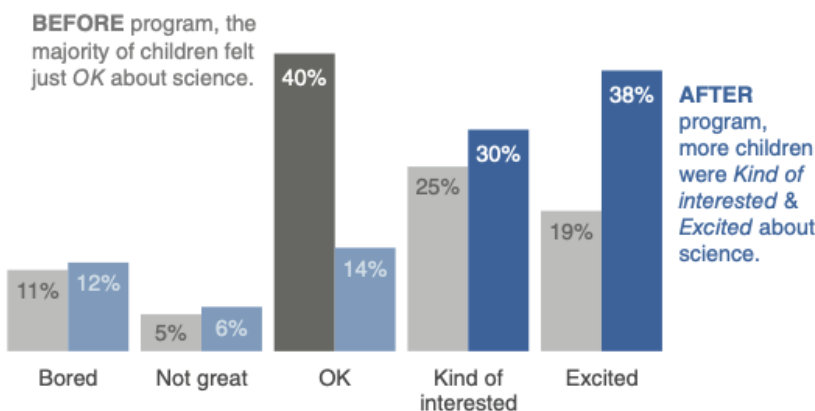
FIGURE 9.28 Original visual

Image source: Storytelling with Data: A Data Visualization Guide for Business Professionals, Cole Nussbaumer Knaflic

Biểu đồ được thiết kế hiệu quả hơn

Pilot program was a success

How do you feel about science?



Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

FIGURE 9.30 Simple bar graph

1. Chọn đúng loại biểu đồ

- Chọn đúng biểu đồ phù hợp

2. Đơn Giản Hóa Nội Dung

- Loại bỏ thông tin dư thừa
- Sắp xếp dữ liệu theo thứ tự logic

3. Gắn Nhãn Rõ Ràng

- Đặt tiêu đề ngắn gọn truyền đạt thông điệp chính
- Thêm chú thích khi cần

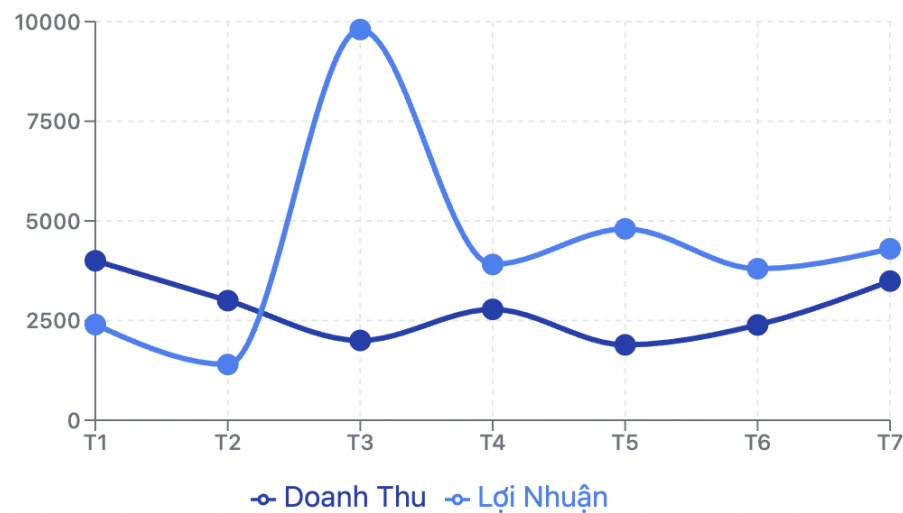
4. Sử Dụng Màu Sắc Phù Hợp

- Chọn màu phù hợp
- Dùng màu tương phản
- Giữ nhất quán màu sắc

Biểu Đồ Đường (Line Chart)

Biểu đồ đường giúp xem sự thay đổi của dữ liệu theo thời gian một cách rõ ràng.

Biểu đồ đường: Thay đổi doanh thu theo tháng



Tips sử dụng biểu đồ đường

Xem xu hướng qua thời gian

- Thích hợp để hiển thị dữ liệu theo thời gian như doanh số theo tháng

So sánh nhiều loại dữ liệu

- Dễ xem xu hướng của nhiều đối tượng trên cùng một biểu đồ

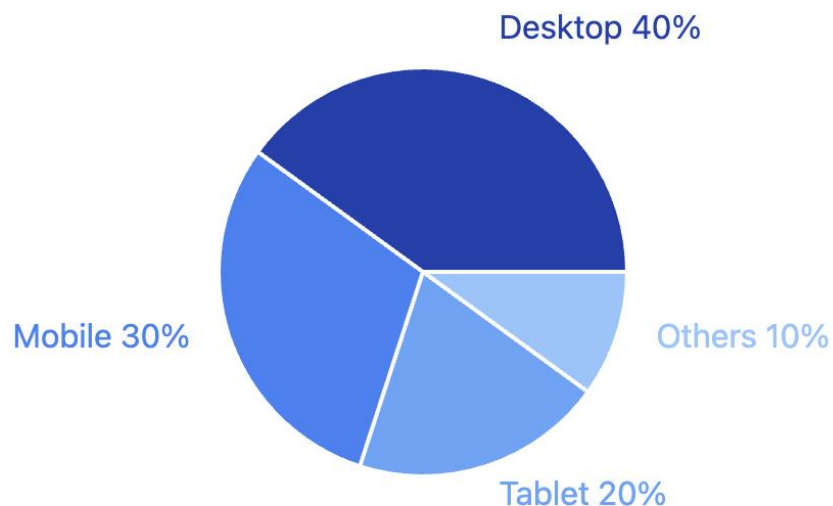
Các kiểu biểu đồ đường trong Excel

- Có nhiều loại: biểu đồ đường đơn giản, biểu đồ có điểm đánh dấu, biểu đồ 2D/3D

Biểu Đồ Tròn (Pie Chart)

Biểu đồ tròn thể hiện tỷ lệ các phần trong một tổng thể. Hiệu quả khi so sánh doanh thu, thị phần hoặc phân bổ ngân sách.

Biểu đồ tròn: Phân bố thiết bị truy cập



Tips sử dụng biểu đồ tròn

Hiển thị tỷ lệ phần trăm

- Giúp người xem nhanh chóng nắm bắt tỷ lệ của từng phần

Tối ưu với dưới 7 phần

- Hiệu quả nhất khi hiển thị không quá 7 phân khúc

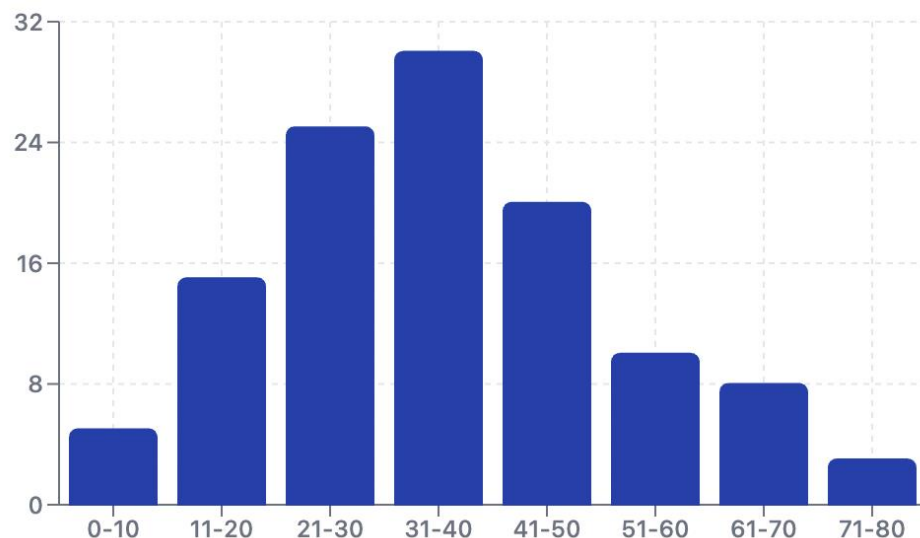
Các biến thể

- Gồm biểu đồ tròn tiêu chuẩn, biểu đồ phân tách và biểu đồ hình khuyên

Biểu đồ Phân Phối (Histogram)

Biểu đồ histogram hiển thị phân phối tần suất của dữ liệu số bằng cách chia thành các khoảng và đếm số lượng điểm dữ liệu trong mỗi khoảng.

Biểu Đồ Histogram: Phân Bố Tần Suất



Tips sử dụng biểu đồ Histogram

Đánh giá hình dạng phân phối

- Nhận biết dữ liệu đối xứng, lệch phải hoặc lệch trái

Phát hiện giá trị ngoại lai

- Xác định outliers không phù hợp với xu hướng chung

Kiểm tra tính chuẩn

- Đánh giá sự phù hợp với phân phối chuẩn

Xác định mô hình thống kê

- Chọn mô hình phù hợp dựa trên dạng phân phối

Heatmap là công cụ trực quan hóa dữ liệu bằng màu sắc, giúp nhận biết nhanh các mẫu và xu hướng thông qua thang màu đậm nhạt.

Heat map đơn giản

Tháng	Doanh thu	Lợi nhuận
Tháng 1	120	-12
Tháng 2	135	-7
Tháng 3	128	13
Tháng 4	160	16
Tháng 5	175	18
Tháng 6	162	16
Tháng 7	220	33
Tháng 8	180	18
Tháng 9	176	18
Tháng 10	150	15
Tháng 11	130	13
Tháng 12	135	14

※ Đơn vị: Triệu đồng

Tips tạo và sử dụng heatmap

Tạo heatmap bằng Conditional Formatting

- Sử dụng Color Scales trong Conditional Formatting để tạo heatmap, giúp nhanh chóng nhận biết xu hướng dữ liệu gradient màu sắc.

Heatmap giúp phân tích hiệu suất nhanh chóng

- So sánh hiệu suất giữa khu vực, sản phẩm hoặc thời gian thông qua màu sắc đậm nhạt, nhanh chóng xác định điểm mạnh và yếu.

Biểu đồ Water Flow (hay còn gọi là biểu đồ thác nước) là dạng biểu đồ trực quan hóa thể hiện sự thay đổi giá trị lũy kế theo từng giai đoạn, giúp người xem hiểu được các yếu tố đóng góp tích cực hoặc tiêu cực vào kết quả cuối cùng.

Biểu đồ Water Flow

2014 Headcount math

Though more employees transferred out of the team than transferred in, aggressive hiring means overall headcount (HC) increased 16% over the course of the year.

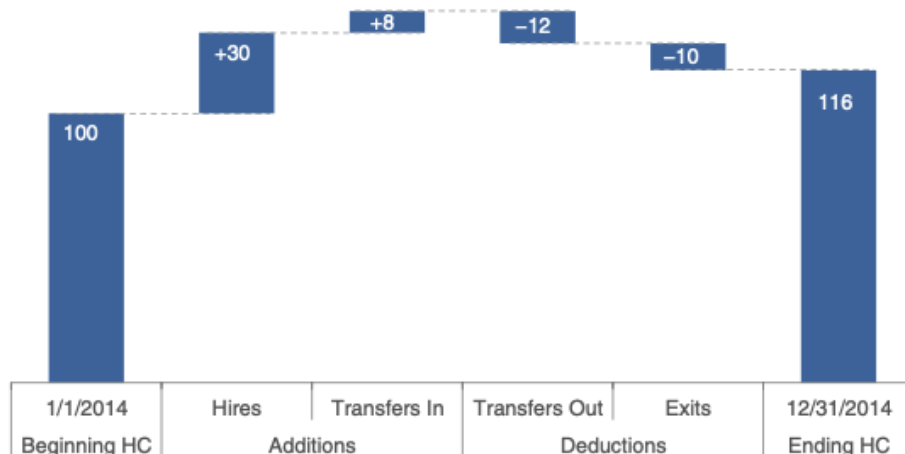


FIGURE 2.17 Waterfall chart

Image source: Storytelling with Data: A Data Visualization Guide for Business Professionals, Cole Nussbaumer Knaflic

Tips tạo và sử dụng Water Flow

Đặc điểm chính

- Hiện thị sự biến động giữa hai điểm dữ liệu - điểm bắt đầu và kết thúc
- Các thanh tăng thường hiển thị màu xanh lá, các thanh giảm thường màu đỏ
- Thanh cuối cùng hiển thị giá trị tổng sau tất cả biến động

Khi nào nên sử dụng

- Phân tích tài chính: theo dõi sự thay đổi doanh thu, chi phí, lợi nhuận
- Phân tích KPI: so sánh chỉ số hiệu suất giữa các giai đoạn
- Phân tích nguyên nhân: xác định các yếu tố tác động tích cực/tiêu cực
- Nghiên cứu thị trường: theo dõi biến động thị phần, khách hàng

Tạo Biểu Đồ Water Flow Trong Excel

Biểu đồ Water Flow (hay biểu đồ thác nước) giúp trực quan hóa sự thay đổi giá trị lũy kế qua các giai đoạn, thể hiện các yếu tố tăng/giảm ảnh hưởng đến kết quả cuối cùng.

1

Chuẩn bị dữ liệu

Tổ chức dữ liệu thành các cột: giá trị ban đầu, các thay đổi (tăng/giảm), và giá trị cuối. Mỗi thay đổi được đặt trong một cột riêng biệt.

2

Tạo biểu đồ cột chồng

Chọn dữ liệu → Insert → Column Chart → Stacked Column. Đây là nền tảng cho biểu đồ Water Flow.

3

Chỉnh sửa hiển thị

Chuyển các giá trị âm thành màu đỏ, giá trị dương thành màu xanh lá. Thêm đường kết nối giữa các cột để tạo hiệu ứng "dòng chảy".

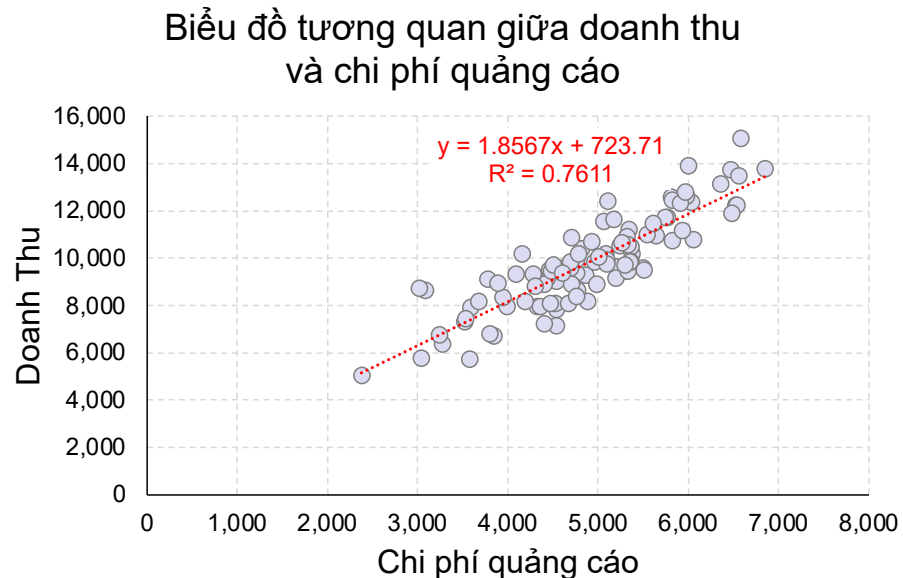
4

Tinh chỉnh định dạng

Thêm nhãn dữ liệu, điều chỉnh màu sắc và căn chỉnh để biểu đồ dễ đọc. Giá trị cuối cùng thường được đánh dấu bằng màu khác biệt.

Biểu đồ phân tán hiển thị mối quan hệ giữa hai biến số, giúp phát hiện mẫu và xu hướng trong dữ liệu.

Biểu Đồ Scatter Plot - Mối Quan Hệ Giữa Các Biến



Tips tạo và sử dụng scatter plot

Tạo scatter plot

- Mỗi điểm trên biểu đồ đại diện cho một cặp giá trị (x,y), giúp nhận diện mẫu và xu hướng.
- Trong Excel: chọn dữ liệu, sử dụng tùy chọn Scatter, thêm nhãn cho các điểm quan trọng.

Thêm trendline và phân tích

- Đường xu hướng làm rõ mối quan hệ giữa các biến. Excel hỗ trợ nhiều dạng: tuyến tính, đa thức, logarithm.
- Hiển thị phương trình và R^2 để đo lường độ mạnh của mối quan hệ và dự đoán giá trị.

3 Tips Để Tạo Scatter Plot Đẹp

1 Cài đặt độ trong suốt khi quá nhiều điểm dữ liệu

- Giảm độ đậm của điểm dữ liệu (opacity) khi visualize dataset lớn giúp nhìn rõ các khu vực tập trung cao và tránh hiện tượng chồng chéo.
- Trong Excel, điều chỉnh này được thực hiện qua Format Data Series.

2 Đặt giới hạn trên và dưới cho dữ liệu

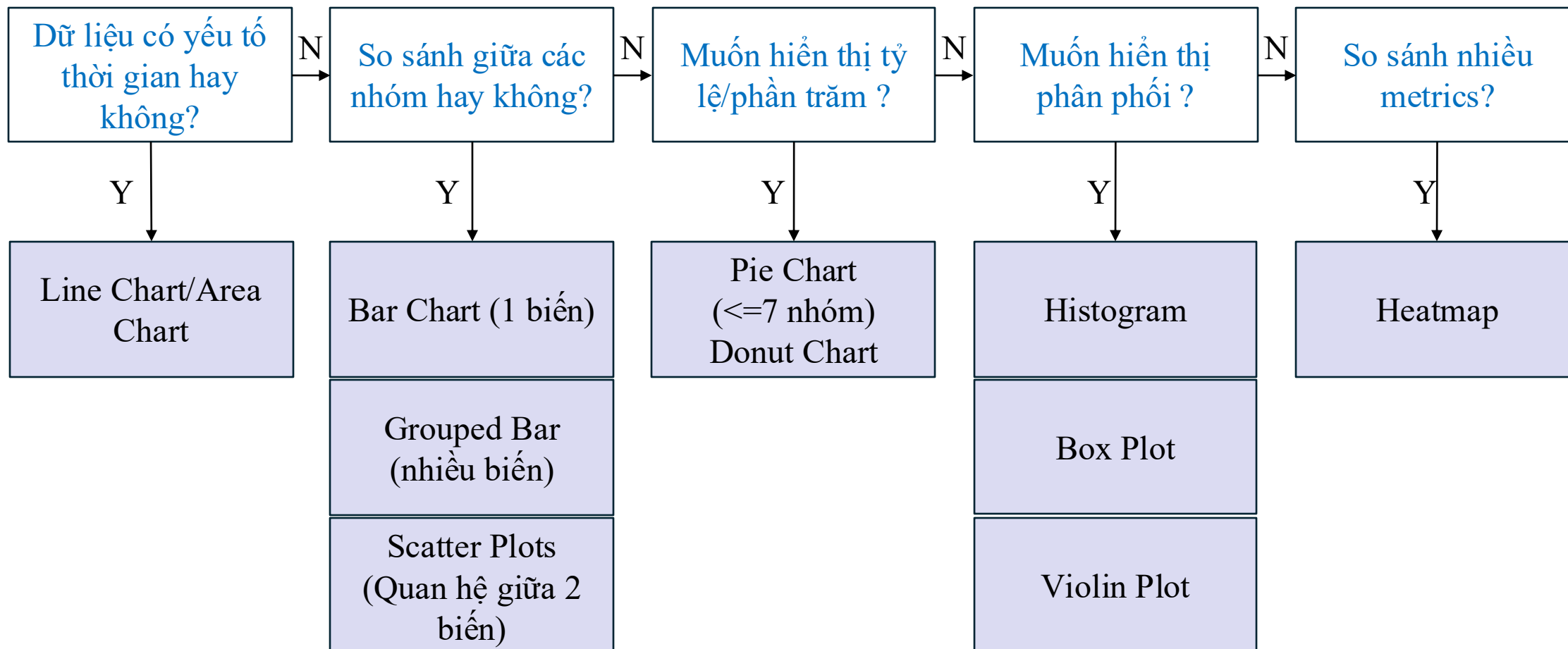
- Thiết lập giới hạn trục x và y phù hợp giúp loại bỏ các giá trị ngoại lệ và tập trung vào phạm vi dữ liệu quan trọng, làm nổi bật xu hướng chính và tăng độ chính xác trong phân tích.

3 Hiển thị đường fitting line

- Thêm đường xu hướng (trendline) để làm rõ mối quan hệ giữa các biến.
- Hiển thị thêm phương trình và giá trị R^2 để đánh giá độ mạnh của mối quan hệ và khả năng dự đoán của mô hình.

Hướng dẫn chọn biểu đồ phù hợp

Decision Tree cho việc chọn biểu đồ



Quick Exercise

Đề Bài: Bạn có dữ liệu bán áo phông trong 1 tuần:

- Thứ 2: 20 chiếc,
- Thứ 3: 35 chiếc,
- Thứ 4: 25 chiếc,
- Thứ 5: 40 chiếc,
- Thứ 6: 60 chiếc,
- Thứ 7: 55 chiếc,
- Chủ nhật: 30 chiếc

Câu hỏi: Nên dùng loại biểu đồ nào và tại sao?

Tầm quan trọng của trực quan hóa dữ liệu:

- Giúp nhanh chóng nhận diện xu hướng, truyền đạt thông tin hiệu quả và hỗ trợ ra quyết định.

Các loại biểu đồ chính đã đề cập:

- Biểu đồ cột, biểu đồ đường, biểu đồ tròn, biểu đồ tần suất, heatmap, biểu đồ thác nước, biểu đồ phân tán.

Các mẹo chính để trực quan hóa hiệu quả:

- Chọn loại biểu đồ phù hợp, đơn giản hóa nội dung, ghi nhãn rõ ràng và sử dụng màu sắc thích hợp.

Công cụ Excel được đề cập:

- Công cụ tạo biểu đồ, định dạng có điều kiện và mẫu biểu đồ.

Phần 1: Trực Quan Hóa Dữ Liệu

Tạo biểu đồ và dashboard hiệu quả từ dữ liệu.

Phần 2: Kiểm Định Giả Thuyết (Hypothesis Testing)

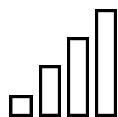
Phương pháp đưa ra kết luận dựa trên kiểm định thống kê.

Phần 3: Tiền Xử Lý Dữ Liệu & Phân Tích Hồi Quy

Tiền xử lý dữ liệu, xây dựng các mô hình hồi quy đơn biến và đa biến

Kiểm Định Giả Thuyết Là Gì?

Là quy trình thống kê để kết luận về giả thuyết từ dữ liệu. Phương pháp kiểm tra xem giả thuyết về tổng thể có được hỗ trợ bởi dữ liệu mẫu hay không.



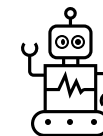
Đánh giá giả thuyết

Đánh giá tính hợp lý của giả thuyết từ dữ liệu mẫu



Phân biệt ý nghĩa

Xác định sự khác biệt có ý nghĩa thống kê hay ngẫu nhiên



Ứng dụng rộng rãi

Công cụ cốt lõi trong khoa học dữ liệu và học máy

Nghiên cứu khoa học, y khoa, phát triển sản phẩm, marketing và nhiều lĩnh vực khác

Tại Sao Phải Kiểm Định Giả Thuyết?

Kiểm định giả thuyết là phương pháp thống kê giúp xác định tính đúng đắn của một giả thuyết dựa trên dữ liệu, cho phép đưa ra kết luận về tổng thể từ mẫu nghiên cứu.

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.

Các thành phần cơ bản của kiểm định

Giả thuyết

Giả thuyết không H_0 : Không có sự khác biệt/ảnh hưởng

Giả thuyết H_1 : Có sự khác biệt/ảnh hưởng

Ví dụ: $H_0: \mu_A = \mu_B$, $H_1: \mu_A \neq \mu_B$

Thống kê kiểm định

t-statistic: So sánh trung bình 2 mẫu nhỏ

z-statistic: So sánh tỉ lệ hoặc mẫu lớn

chi-square: So sánh tần số quan sát và kỳ vọng

Sử dụng Excel

- t-value (T.TEST, T.INV.2T)
- z-value (NORM.S.DIST, NORM.S.INV)
- Chi-square (CHISQ.TEST, CHISQ.INV)

p-value & alpha

Giá trị p: Xác suất quan sát kết quả nếu H_0 đúng

Alpha (α): Ngưỡng bác bỏ H_0 , thường 0.05 (5%).

So sánh p-value và α

- Nếu p-value < α : Bác bỏ H_0
- p-value $\geq \alpha$: Không đủ cơ sở bác bỏ H_0

Population vs Sample

Tập hợp đầy đủ tất cả các đối tượng mà ta muốn nghiên cứu.

- Thường quá lớn hoặc không thực tế để kiểm tra toàn bộ
- Có các thông số đặc trưng gọi là tham số (parameters)
- Ký hiệu: μ (giá trị trung bình), σ (độ lệch chuẩn)



Một phần nhỏ được chọn từ tổng thể để nghiên cứu.

- Dùng để suy luận về tổng thể
- Có các đặc điểm được gọi là thống kê mẫu (statistics)
- Ký hiệu: \bar{x} (giá trị trung bình mẫu), s (độ lệch chuẩn mẫu)

Các Khái Niệm Quan Trọng 1/2

Kiểm định giả thuyết dựa trên bốn khái niệm cơ bản: mức ý nghĩa (alpha), p-value, t-value và cặp giả thuyết H_0/H_1 .

Mức ý nghĩa (alpha)

Ngưỡng quyết định (thường là 0.05) để đánh giá kết quả thống kê. Nếu $p\text{-value} < \alpha$, bác bỏ H_0 . $\alpha = 0.05$ nghĩa là chấp nhận 5% khả năng sai lầm loại I.

p-value

Xác suất quan sát được kết quả cực đoan như dữ liệu mẫu, giả định H_0 đúng. p-value càng nhỏ, bằng chứng chống lại H_0 càng mạnh. Khi $p\text{-value} < \alpha$: bác bỏ H_0 .

t-value

Đo lường khoảng cách giữa trung bình mẫu và giá trị giả định trong H_0 . $|t|$ càng lớn, sự khác biệt càng đáng kể. Công thức: $t = (\bar{x} - \mu) / (s/\sqrt{n})$.

Giả thuyết H_0 và H_1

H_0 (giả thuyết gốc): Không có sự khác biệt/tác động ($\mu_1 = \mu_2$).
 H_1 (giả thuyết thay thế): Có sự khác biệt/tác động ($\mu_1 \neq \mu_2$).
Kiểm định nhằm xác định có đủ bằng chứng bác bỏ H_0 không.

Phân phối xác suất

Mô tả toán học về khả năng xuất hiện của các giá trị của một biến ngẫu nhiên. Phân phối xác suất cho biết mỗi giá trị có thể xảy ra với xác suất bao nhiêu.

Phân phối chuẩn (Normal Distribution)

Phân phối hình chuông đối xứng, được mô tả bởi giá trị trung bình (μ) và độ lệch chuẩn (σ). Nhiều hiện tượng tự nhiên và dữ liệu thực tế tuân theo phân phối này, làm cơ sở cho nhiều kiểm định thống kê.

Định lý giới hạn trung tâm (Central Limit Theorem – CLT)

Khi kích thước mẫu đủ lớn ($n \geq 30$), phân phối của trung bình mẫu sẽ xấp xỉ phân phối chuẩn, bất kể phân phối ban đầu của dữ liệu. Đây là cơ sở cho nhiều phương pháp suy luận thống kê.

Sai lầm loại I, II (Type I, II Error)

Sai lầm loại I (α): Bác bỏ H_0 khi nó đúng (phát hiện dương tính giả).

Sai lầm loại II (β): Không bác bỏ H_0 khi nó sai (phát hiện âm tính giả). Mỗi loại sai lầm có hậu quả khác nhau tùy vào bối cảnh nghiên cứu.

Phân phối xác suất mô tả khả năng xảy ra của các kết quả trong một thử nghiệm.

Hiểu đơn giản về phân phối xác suất

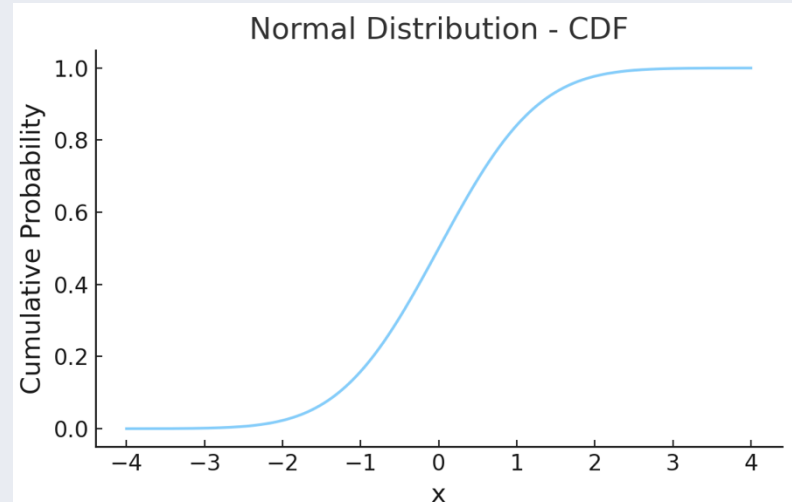
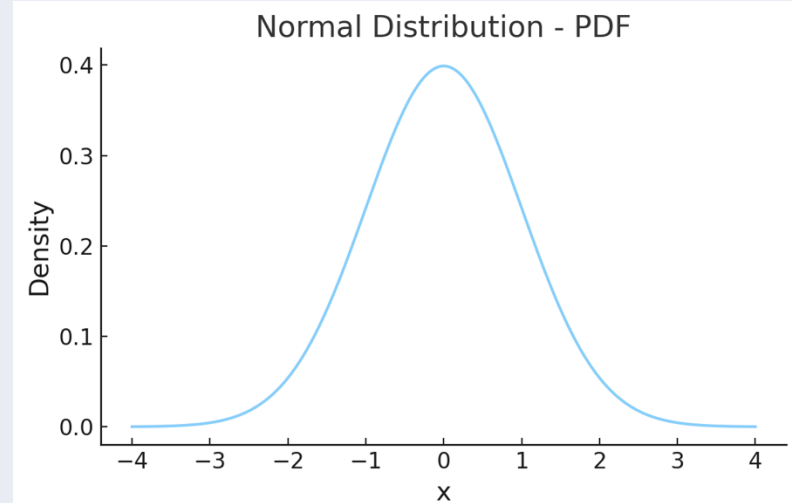
Khi tung đồng xu, xác suất ra mặt ngửa hoặc sấp là 50%. Nhiều sự kiện tự nhiên khi ghi lại (như chiều cao) tạo nên các "hình dạng" đặc trưng trên biểu đồ.

Hàm mật độ xác suất (PDF)

PDF biểu thị khả năng xuất hiện của các giá trị. Diện tích dưới đường cong bằng 1, giúp tính xác suất giá trị trong một khoảng.

Hàm phân phối tích lũy (CDF)

CDF cho biết xác suất một giá trị nhỏ hơn hoặc bằng một mức.
CDF tăng dần từ 0 đến 1.



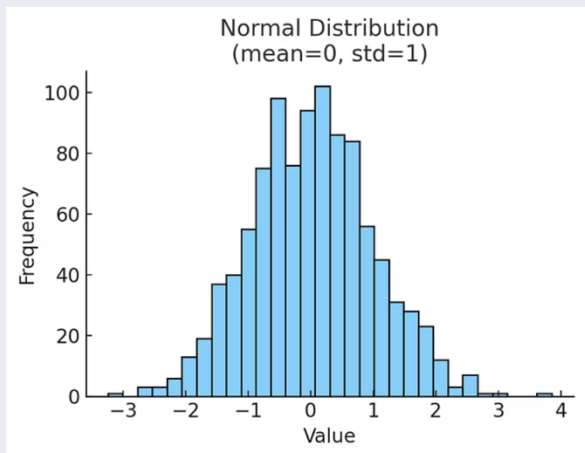
3 Loại Phân Phối Thường Gặp

Phân Phối Chuẩn (Normal)

Dạng hình chuông đặc trưng, thường xuất hiện trong tự nhiên.

- Chiều cao của người trưởng thành
- Điểm thi của học sinh trong lớp đông
- Sai số đo lường trong khoa học

Trong Excel: Sử dụng hàm `NORM.DIST()` để tính xác suất

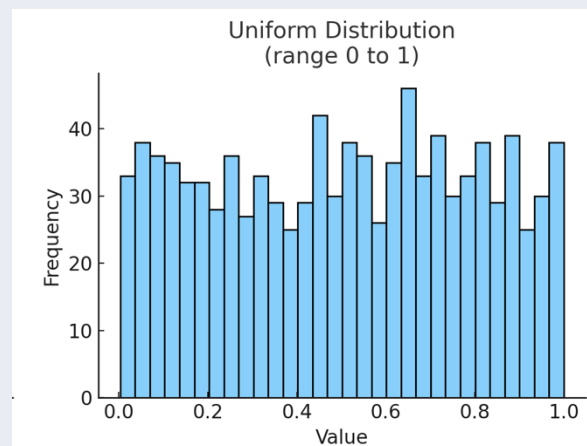


Phân Phối Đồng Đều (Uniform)

Mọi giá trị có xác suất xuất hiện như nhau.

- Số hiển thị khi tung xúc xắc
- Thời gian chờ đợi giữa 2 khách hàng
- Vị trí ngẫu nhiên của điểm trên đoạn thẳng

Trong Excel: Dùng hàm `RAND()` để tạo số ngẫu nhiên đồng đều

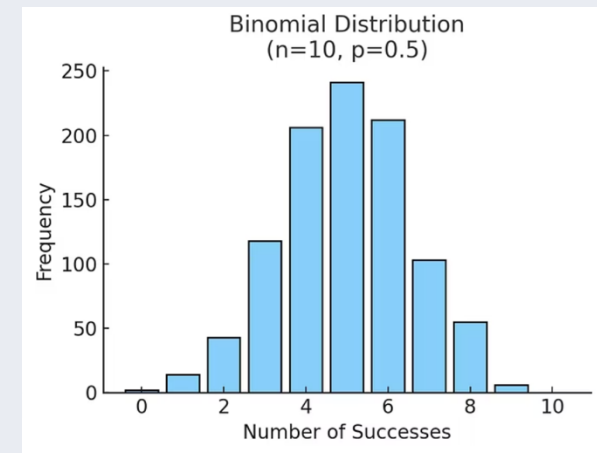


Phân Phối Nhị Thức (Binomial)

Mô tả số lần thành công trong n thử nghiệm độc lập.

- Số lần tung đồng xu được mặt ngửa
- Số sản phẩm lỗi trong dây chuyền sản xuất
- Số khách hàng chấp nhận mua sản phẩm

Trong Excel: Sử dụng hàm `BINOM.DIST()` để tính xác suất



Định lý giới hạn trung tâm khẳng định rằng với mẫu đủ lớn từ bất kỳ quần thể nào, phân phối của giá trị trung bình mẫu sẽ xấp xỉ phân phối chuẩn.

Tại sao định lý này đúng?

Ý tưởng cơ bản: Khi lấy trung bình nhiều biến ngẫu nhiên, các **lệch lạc cá nhân sẽ triệt tiêu**, tạo ra biến mới **ổn định và cân đối hơn** (Điều kiện: Mẫu cần đủ lớn ($n \geq 30$) và các quan sát phải độc lập).

Hệ quả: Phân phối mẫu có trung bình bằng μ và độ lệch chuẩn bằng σ/\sqrt{n} .

Ứng dụng của định lý

- **Xây dựng khoảng tin cậy (Confidence Interval):** Từ mẫu nhỏ, **ước lượng toàn bộ tổng thể** bằng cách dùng công thức chuẩn xác định khoảng giá trị chứa trung bình thực.
- **Kiểm định giả thuyết:** Các kiểm định t, z, chi-square dựa vào **phân phối trung bình mẫu là chuẩn** để đưa ra kết luận về tổng thể.
- **Ứng dụng thực tế:** Dự đoán hành vi hệ thống lớn từ mẫu nhỏ. Kiểm soát quy trình sản xuất bằng giám sát mẫu tính ngẫu nhiên.

Mức Ý Nghĩa (Significance Level)

Là ngưỡng xác suất để quyết định bác bỏ giả thuyết không (H_0).

- Thường được ký hiệu là α (alpha), giá trị phổ biến: 0.05, 0.01, 0.1
- $p\text{-value} < \alpha$: Bác bỏ H_0 , kết quả có ý nghĩa thống kê
- $p\text{-value} \geq \alpha$: Không đủ bằng chứng để bác bỏ H_0

Mức ý nghĩa 5% ($\alpha = 0.05$) nghĩa là chấp nhận rủi ro 5% kết luận sai khi bác bỏ H_0 .

Mối Quan Hệ Giữa Mức Ý Nghĩa và Khoảng Tin Cậy

Mức ý nghĩa $\alpha = 0.05$ tương ứng với khoảng tin cậy 95%. Hai cách tiếp cận bổ sung cho nhau: kiểm định dùng p-value, ước lượng dùng khoảng tin cậy.

Khoảng Tin Cậy (Confidence Interval)

Là khoảng giá trị ước lượng chứa tham số thực của tổng thể với mức độ tin cậy nhất định.

- Công thức: Ước lượng \pm Margin of Error
- CI 95%: Có 95% khả năng khoảng chứa giá trị thực của tham số
- Khoảng hẹp = độ chính xác cao, khoảng rộng = độ chính xác thấp

t-value (Giá trị t)

Là thước đo sự khác biệt giữa trung bình mẫu và giá trị giả định của tổng thể, tính theo đơn vị độ lệch chuẩn.

Công thức tính:

$$t = (\bar{x} - \mu) / (s / \sqrt{n})$$

Trong đó:

- \bar{x} : Giá trị trung bình của mẫu
- μ : Giá trị giả định của tổng thể (từ H_0)
- s : Độ lệch chuẩn của mẫu
- n : Kích thước mẫu

Giá trị t càng lớn, bằng chứng càng mạnh để bác bỏ giả thuyết không (H_0).

p-value (Giá trị p)

Là xác suất quan sát được kết quả cực đoan như dữ liệu mẫu (hoặc cực đoan hơn) với giả định H_0 là đúng.

Cách xác định:

1. Tính t-value từ dữ liệu mẫu
2. Dựa vào phân phối t với bậc tự do phù hợp ($df = n-1$) để tìm xác suất
3. Trong Excel: =T.DIST.2T(|t-value|, df) cho kiểm định hai phía

P-value nhỏ (<0.05) nghĩa là có đủ bằng chứng để bác bỏ H_0 với mức ý nghĩa 5%.

t-value & p-value trong phân phối chuẩn

Probability & Statistical Significance Explained

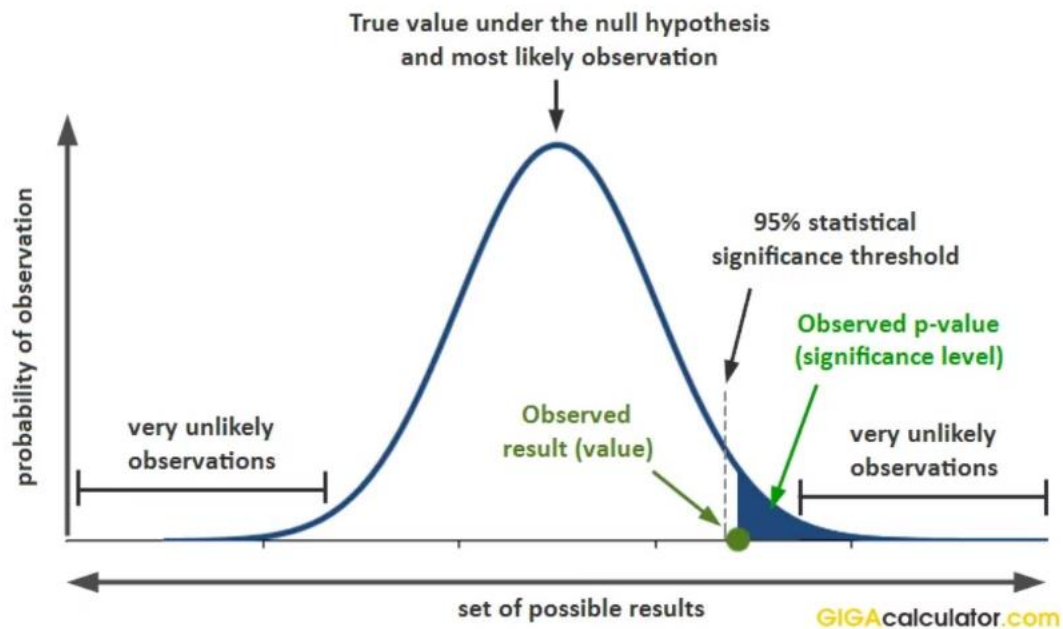


Image Source: <https://medium.datadriveninvestor.com/p-value-t-test-chi-square-test-anova-when-to-use-which-strategy-32907734aa0e>

Trong phân phối chuẩn:

- t-value: Khoảng cách từ giá trị trung bình theo độ lệch chuẩn
- Với phân phối chuẩn, t-value > 1.96 hoặc < -1.96 cho p-value < 0.05 (kiểm định hai phía)
- Vùng bác bỏ H_0 nằm ở hai đuôi của đường cong

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to neurological stimulus, and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats' response times is 1.05 seconds with a sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time?

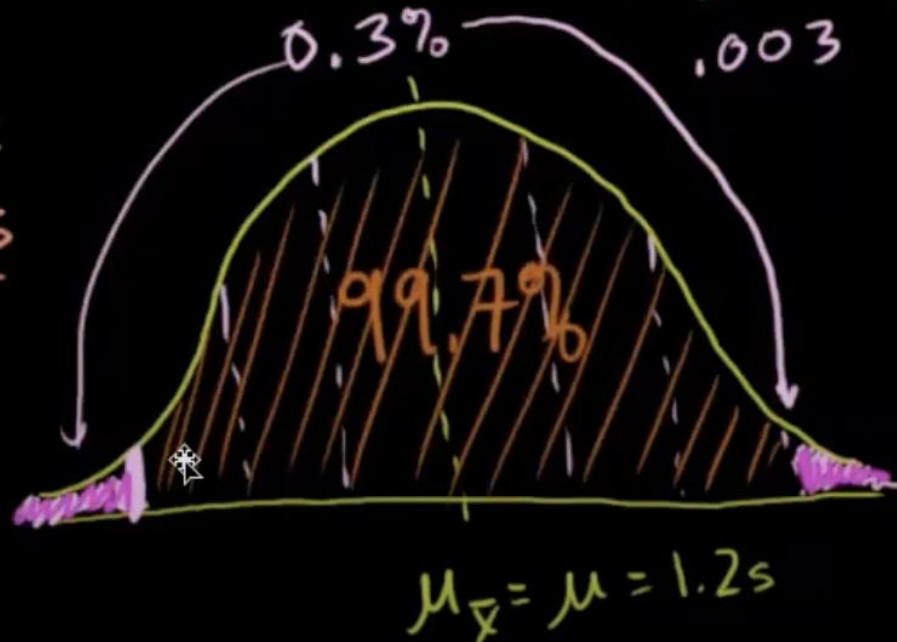
effect H_0 : Drug has no effect $\Rightarrow \mu = 1.2 \text{ s}$ (even w/ drug)

H_1 : Drug has an effect $\Rightarrow \mu \neq 1.2 \text{ s}$ when the drug is given

Assume H_0 :

$$Z = \frac{1.2 - 1.05}{0.05}$$

$$Z = \frac{.15}{.05} = 3$$



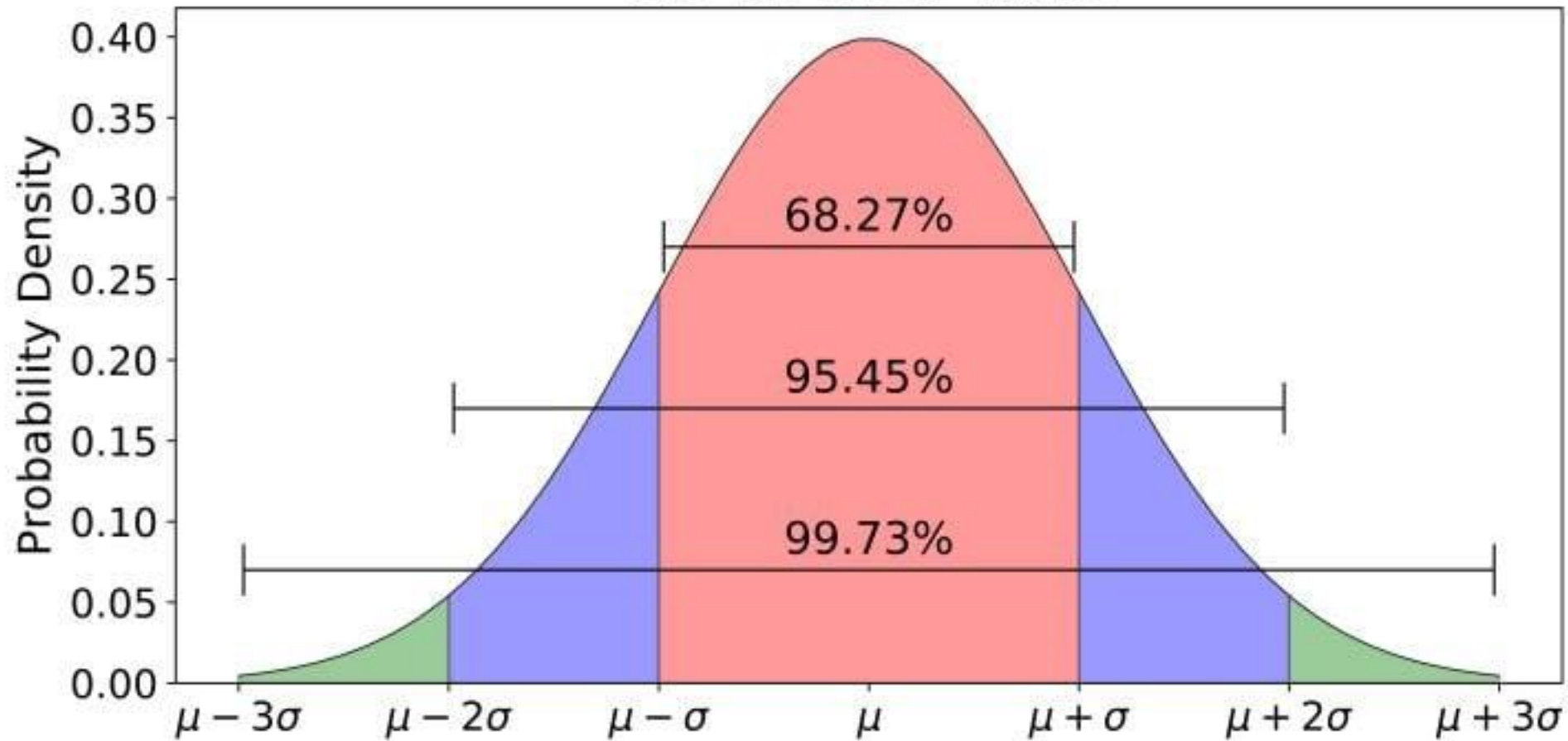
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{100}} \approx \frac{.5}{\sqrt{100}} = \frac{0.5}{10}$$

$$\hat{\sigma}_{\bar{x}} = 0.05$$

p-value = 0.03

If null hypothesis is true, we have 0.3% chance to have $1.05 \pm 0.5 \text{ s}$

68-95-99.7 Rule



Nguồn: <https://builtin.com/data-science/empirical-rule>

So sánh giữa t-value và z-value

t-value

- Dùng khi không biết phương sai tổng thể
- Ước lượng từ phương sai mẫu
- Phụ thuộc vào bậc tự do (df)
- Chịu ảnh hưởng bởi mẫu nhỏ

z-value

- Dùng khi biết phương sai tổng thể
- Độc lập với kích thước mẫu
- Dựa trên phân phối chuẩn
- Phù hợp với mẫu lớn ($n > 30$)

Khi mẫu đủ lớn ($n > 30$), phân phối t xấp xỉ phân phối chuẩn, khiến t-value và z-value trở nên tương đương.

Kích thước mẫu tăng

Khi n tăng, bậc tự do ($df = n - 1$) cũng tăng, làm phân phối t gần với phân phối chuẩn.

Tiệm cận đến phân phối chuẩn

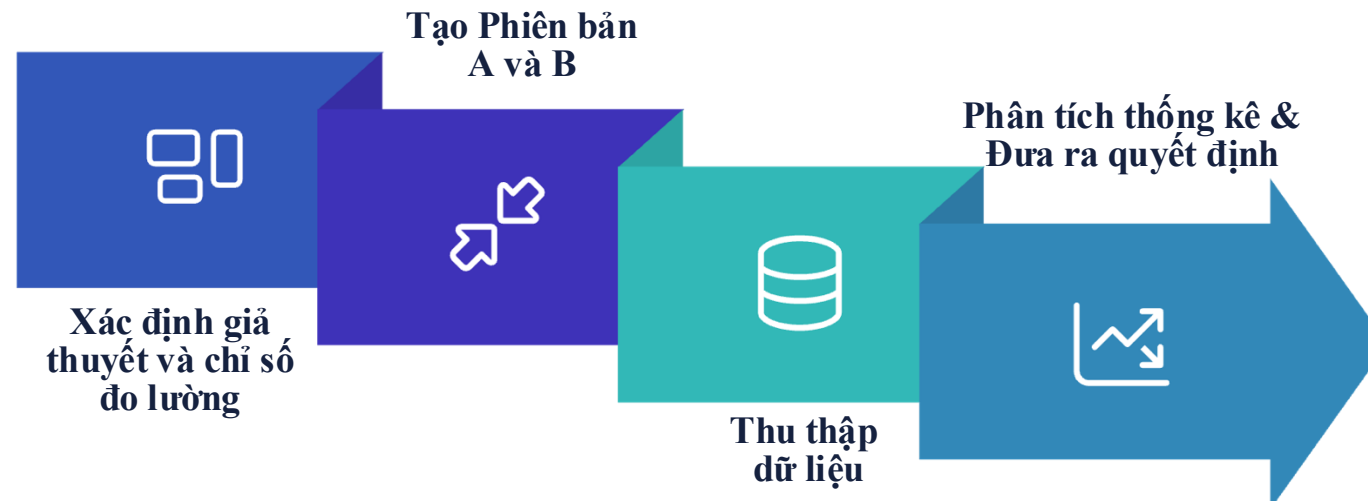
Với $n \rightarrow \infty$, phân phối t tiệm cận phân phối chuẩn, làm t-value gần giống z-value.

Ứng dụng thực tế

Nhiều nhà thống kê coi hai giá trị tương đương khi $n > 30$, cho phép dùng bảng z thay bảng t.

Giới thiệu về A/B Testing

A/B Testing là ứng dụng trực tiếp của kiểm định thống kê, cho phép so sánh hiệu quả của hai hoặc nhiều phiên bản khác nhau (ví dụ: trang web, ứng dụng, chiến dịch) để xác định phiên bản nào hoạt động tốt hơn.



So sánh 2 phiên bản (A/B)

Kiểm tra hiệu quả của các thay đổi về thiết kế, chức năng, nội dung

Phân chia ngẫu nhiên người dùng

Đảm bảo tính khách quan và giảm thiểu sai số hệ thống

Đo lường chỉ số KPI

Tỉ lệ chuyển đổi, doanh thu, thời gian trên trang

So Sánh Hai Trang Chủ Website

Phiên bản A: Banner Lớn Quảng Cáo Khuyến Mãi

- Banner lớn chiếm 70% không gian màn hình đầu tiên
- Tập trung vào khuyến mãi "Giảm giá 50% cho đơn hàng đầu tiên"
- Nút "Mua ngay" màu đỏ nổi bật ở trung tâm banner
- Danh mục sản phẩm hiển thị phía dưới banner

Giả thuyết: Banner khuyến mãi lớn sẽ thu hút sự chú ý và tăng tỉ lệ người dùng nhấp vào nút mua hàng.

Mục tiêu: Xác định phiên bản nào mang lại tỉ lệ chuyển đổi mua hàng cao hơn một cách có ý nghĩa thống kê.

Phiên bản B: Banner Nhỏ, Tập Trung Đánh Giá Khách Hàng

- Banner nhỏ hơn, chỉ chiếm 30% không gian màn hình đầu tiên
- Hiển thị đánh giá 5 sao từ khách hàng thực
- Trích dẫn từ khách hàng về chất lượng sản phẩm
- Sản phẩm bán chạy hiển thị ngay dưới banner

Giả thuyết: Xây dựng niềm tin thông qua đánh giá của khách hàng sẽ tăng tỉ lệ chuyển đổi mua hàng.

Quy trình kiểm định giả thuyết thống kê trên Excel:

1

Xác định giả thuyết

Thiết lập giả thuyết gốc (H_0) và giả thuyết thay thế (H_1) dựa trên vấn đề cần kiểm định

2

Chọn kiểm định phù hợp

Lựa chọn kiểm định thích hợp như t-test, z-test, ANOVA, hoặc kiểm định F tùy thuộc vào dữ liệu và mục tiêu phân tích

3

Thực hiện kiểm định

Sử dụng Data Analysis Tool trên Excel để tiến hành kiểm định đã chọn với dữ liệu của bạn

4

Phân tích kết quả

So sánh p-value với mức ý nghĩa alpha (thường là 0.05). Nếu p-value < alpha, bác bỏ giả thuyết gốc H_0

5

Đưa ra kết luận

Diễn giải kết quả trong ngữ cảnh thực tế và đưa ra quyết định dựa trên kết quả kiểm định

Các Loại Kiểm Định Phổ Biến

Lựa chọn kiểm định phù hợp dựa trên loại dữ liệu và mục tiêu phân tích

t-Test

Mục đích: So sánh trung bình giữa hai nhóm

Ứng dụng: So sánh hiệu quả hai phương pháp, đánh giá trước-sau can thiệp

- One-sample t-test: So sánh với giá trị chuẩn
- Paired t-test: So sánh các cặp dữ liệu liên quan
- Independent t-test: So sánh hai nhóm độc lập

ANOVA

Mục đích: So sánh trung bình giữa nhiều nhóm (>2)

Ứng dụng: So sánh hiệu quả của nhiều phương pháp khác nhau

- One-way ANOVA: Một biến phân loại
- Two-way ANOVA: Hai biến phân loại
- MANOVA: Nhiều biến phụ thuộc

Kiểm Định Phi Tham Số

Mục đích: Dùng khi dữ liệu không tuân theo phân phối chuẩn

Ứng dụng: Phân tích dữ liệu thứ bậc, dữ liệu nhỏ, phân phối lệch

- Mann-Whitney U: Thay thế cho t-test độc lập
- Wilcoxon: Thay thế cho paired t-test
- Kruskal-Wallis: Thay thế cho ANOVA
- Chi-square: Kiểm định tính độc lập giữa các biến phân loại

Kiểm định hiệu quả chiến dịch marketing mới

1. Xác định vấn đề

- Công ty ABC muốn biết liệu chiến dịch quảng cáo mới có làm tăng doanh số bán hàng không

2. Thu thập dữ liệu

- Doanh số trước chiến dịch ($n=30$): trung bình $\mu_1=520$ triệu đồng/tháng
- Doanh số sau chiến dịch ($n=30$): trung bình $\mu_2=580$ triệu đồng/tháng

3. Thiết lập giả thuyết

- $H_0: \mu_1 = \mu_2$ (Doanh số trung bình không thay đổi)
- $H_1: \mu_1 < \mu_2$ (Doanh số trung bình tăng lên)

4. Thực hiện kiểm định

- Sử dụng t-Test: Paired Two Sample for Means từ Data Analysis ToolPak
- Kết quả:** $p\text{-value} = 0.023$, $t\text{-stat} = 2.458$

5. Đưa ra kết luận

- $p\text{-value} = 0.023 < \alpha = 0.05$, nên bác bỏ H_0

Kết luận: Với mức ý nghĩa 5%, có đủ bằng chứng thống kê để kết luận rằng chiến dịch marketing mới đã thực sự làm tăng doanh số bán hàng. Dữ liệu cho thấy doanh số đã tăng trung bình 60 triệu đồng mỗi tháng sau khi triển khai chiến dịch.

<Tham Khảo>

Các Hàm và Công Cụ Thống Kê Trong Excel

AI

AI VIETNAM
@aivietnam.edu.vn

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

Tầm quan trọng của Kiểm định Giả thuyết:

- Đưa ra quyết định khách quan và xác định ý nghĩa thống kê từ dữ liệu.
- Đánh giá mối quan hệ, sự khác biệt giữa các nhóm và hiệu quả can thiệp.

Các khái niệm và phương pháp chính:

- **Khái niệm:** Population vs Sample, Mức ý nghĩa (alpha), p-value, t-value, Phân phối chuẩn, Định lý Giới hạn Trung tâm, Lỗi Loại I & II, Khoảng tin cậy.
- **Kiểm định phổ biến:** t-Test (các loại), ANOVA, kiểm định phi tham số.

Ứng dụng trong A/B Testing:

- Là ứng dụng thực tế để so sánh hiệu quả các phiên bản (H_0 , H_1) dựa trên phân tích p-value và tỷ lệ chuyển đổi.

Công cụ Excel:

- Data Analysis ToolPak.
- Các hàm thống kê (T.TEST, F.TEST, CORREL) và hàm phân phối xác suất.

Phần 1: Trực Quan Hóa Dữ Liệu

Tạo biểu đồ và dashboard hiệu quả từ dữ liệu.

Phần 2: Kiểm Định Giả Thuyết (Hypothesis Testing)




Phương pháp đưa ra kết luận dựa trên kiểm định thống kê.

Phần 3: Tiền Xử Lý Dữ Liệu & Phân Tích Hồi Quy

Tiền xử lý dữ liệu, xây dựng các mô hình hồi quy đơn biến và đa biến

Tại sao cần tiền xử lý dữ liệu?

Tiền xử lý dữ liệu giúp làm sạch dữ liệu thô, nâng cao độ chính xác trong phân tích và tạo nền tảng cho việc phân tích chính xác

-  **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.

Các phương pháp xử lý dữ liệu phổ biến

Excel cung cấp công cụ xử lý dữ liệu thiếu, loại bỏ giá trị bất thường, tạo biến giả và kiểm tra tính chính xác

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.

- ❗ Lựa chọn phương pháp xử lý phụ thuộc vào bản chất dữ liệu, mục đích phân tích và tỷ lệ giá trị thiếu. Luôn ghi chú phương pháp đã sử dụng để đảm bảo tính minh bạch.

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 IQR.
- Z-score: Sử dụng hàm STANDARDIZE, giá trị có $|z| > 3$ thường là ngoại lệ.

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.

 Xử lý outliers cần được thực hiện cẩn thận, có cơ sở khoa học và phù hợp với ngữ cảnh phân tích.

Tạo biến giả (Dummy) cho biến phân loại

Kỹ thuật chuyển đổi biến phân loại thành dạng số học phù hợp cho phân tích định lượng trong Excel.

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.

Kỹ thuật này đóng vai trò quan trọng trong phân tích hồi quy, phân tích phương sai và dự báo khi làm việc với dữ liệu phi số học.

Kiểm tra và đánh giá sau xử lý

Kiểm tra kết quả sau tiền xử lý là bước cần thiết để đảm bảo dữ liệu sẵn sàng cho phân tích.

Kiểm tra tính nhất quán

Kiểm tra các mâu thuẫn trong dữ liệu sau xử lý bằng COUNTIF, AVERAGEIF và các hàm điều kiện.

Đánh giá phân phối

Xem xét phân phối sau chuẩn hóa qua Histogram và QQ-plot. Sử dụng Data Analysis ToolPak để kiểm tra tính chuẩn.

So sánh trước-sau

Tạo biểu đồ so sánh để đánh giá hiệu quả tiền xử lý. Dùng bar charts và scatter plots để trực quan hóa thay đổi.

Lặp lại quy trình

Tối ưu quy trình dựa trên đánh giá. Sử dụng macro và Power Query để tự động hóa, đảm bảo nhất quán và tiết kiệm thời gian.

Đánh giá kỹ lưỡng giúp phát hiện sớm vấn đề tiềm ẩn, cho phép điều chỉnh phương pháp xử lý, từ đó nâng cao độ chính xác của phân tích và dự báo.

Tiền xử lý dữ liệu doanh số bán hàng

Dữ liệu thô (trước xử lý)

Tháng	Doanh số (triệu VND)	Khu vực	Ghi chú
T1/2023	45.2	Miền Bắc	Đầy đủ
T2/2023	NULL	Miền Nam	Thiếu dữ liệu
T3/2023	52.7	Miền Bắc	Đầy đủ
T4/2023	198.5	Miền Bắc	Nghi ngờ sai sót
T5/2023	54.8	MB	Đầy đủ
T6/2023	-12.3	Miền Trung	Giá trị âm

Dữ liệu sau khi xử lý

Tháng	Doanh số (triệu VND)	Khu vực	Miền_Bắc
T1/2023	45.2	Miền Bắc	1
T2/2023	48.95	Miền Nam	2
T3/2023	52.7	Miền Bắc	1
T4/2023	54.8	Miền Bắc	1
T5/2023	54.8	Miền Bắc	1
T6/2023	53.7	Miền Trung	3

- **Điền giá trị thiếu:** T2/2023 được điền bằng giá trị trung bình của T1 và T3 (48.95)
- **Xử lý giá trị ngoại lệ:** T4/2023 (198.5) được thay thế bằng trung vị của dữ liệu hợp lệ (54.8)
- **Chuẩn hóa tên:** "MB" được thống nhất thành "Miền Bắc"
- **Xử lý giá trị âm:** T6/2023 (-12.3) được thay bằng giá trị dự đoán từ xu hướng dữ liệu (53.7)
- **Tạo biến giả:** Thêm cột cho 3 miền với giá trị 1,2,3 đại diện cho khu vực Miền Bắc

<Tham Khảo>

Hàm Excel Phổ Biến Trong Tiền Xử Lý Dữ Liệu

AI VIETNAM

@aivietnam.edu.vn

Bảng dưới đây tổng hợp các phương pháp tiền xử lý dữ liệu phổ biến và các hàm Excel tương ứng để thực hiện chúng.

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

<Tham Khảo>

Phương Pháp Kết Hợp Bảng Dữ Liệu Trong Excel

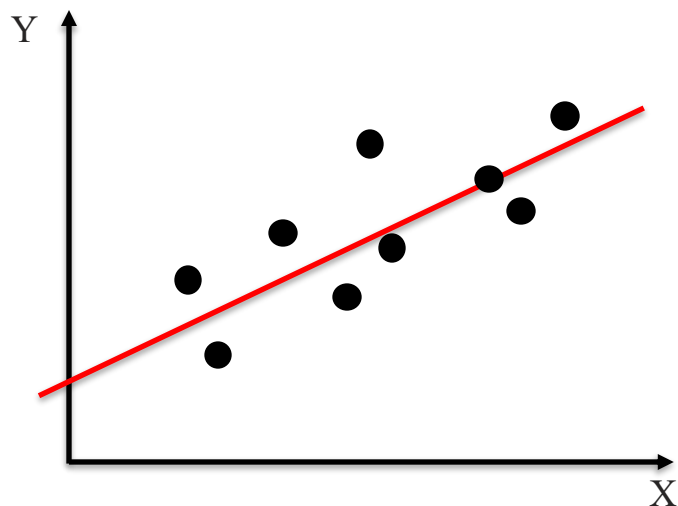
AI VIETNAM
@aivietnam.edu.vn

Dưới đây là các phương pháp phổ biến để kết hợp nhiều bảng dữ liệu thành một bảng trong Excel:

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

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:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$

1

Giải thích:

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)

2

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

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.)

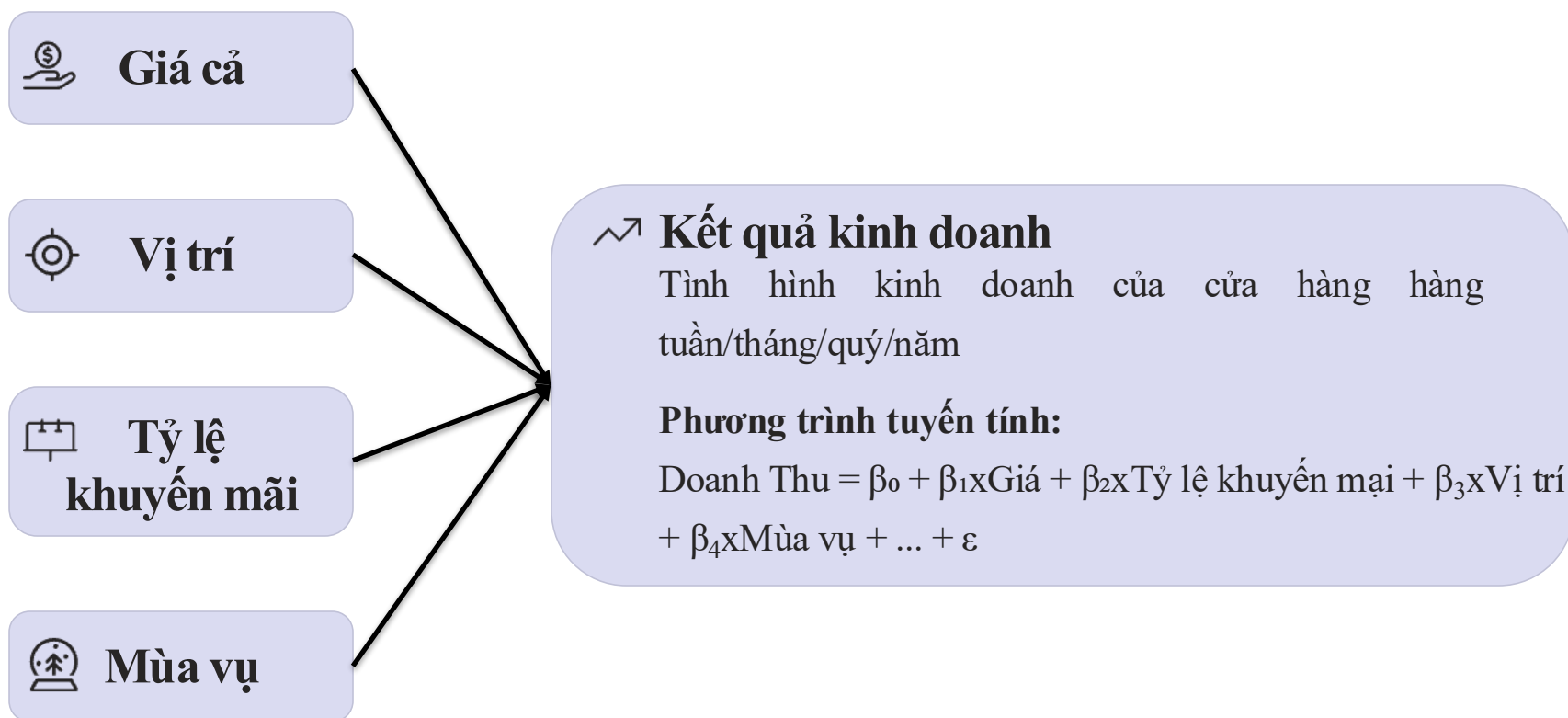
Tại sao sử dụng phân tích hồi quy?

- **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

Yếu Tố Ảnh Hưởng Đến Doanh Thu Cửa Hàng



Phân tích hồi quy giúp xác định mức độ ảnh hưởng của các yếu tố như giá cả, khuyến mãi, vị trí địa lý và mùa vụ đến hiệu quả kinh doanh. Giúp dự đoán doanh thu khi điều chỉnh các yếu tố như giá, khuyến mãi...



Thực Hiện 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.

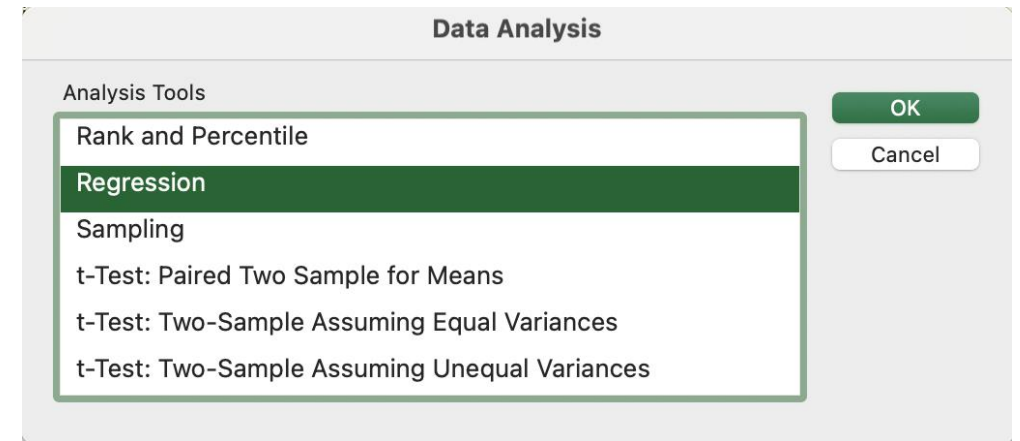
Bước 3: Thiết lập vùng dữ liệu Y (doanh thu) và X (biến ảnh hưởng)

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.

Công cụ Data>Data Analysis trong Excel



Bài toán: Phân tích hồi quy xác định mức độ ảnh hưởng của chi phí đầu tư cho marketing (Marketing Spend) đến hiệu quả kinh doanh (Revenue)

Dữ liệu kinh doanh		
Product_ID	Marketing_Spend	Revenue
SP001	726	42468
SP002	357	17112
SP003	304	27029
SP004	596	37162
SP005	747	32008
SP006	480	24585
SP007	982	63939
SP008	716	50748
SP009	532	24638
SP010	452	31361
SP011	408	36740
SP012	756	51311
SP013	494	34534
SP014	153	6412
SP015	458	30078
SP016	764	41215
SP017	264	24979
SP018	257	23011
SP019	578	35024
SP020	578	34504
SP021	670	37226
SP022	864	57819
SP091	734	45339
SP092	995	50182
SP093	420	34957
SP094	786	42521
SP095	633	44209
SP096	722	35365
SP097	236	16841
SP098	458	13429
SP099	316	13349
SP100	409	28757

Kết quả phân tích hồi quy đơn biến

Regression Result

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.878319475
R Square	0.771445101
Adjusted R Square	0.769112908
Standard Error	6430.57083
Observations	100

ANOVA

	df	SS	MS	F	Significance F
Regression	1	13678536012	13678536012	330.7810077	3.55488E-33
Residual	98	4052519637	41352241.2		
Total	99	17731055649			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	2747.8894	1727.944011	1.590265299	0.114995928	-681.1592532	6176.938053	-681.1592532	6176.938053
Marketing_Spend	52.96027561	2.911923448	18.18738595	3.55488E-33	47.1816583	58.73889292	47.1816583	58.73889292

Marketing_Spend Line Fit Plot

Hồi quy đơn biến – Lý giải kết quả hồi quy

Lý giải ý nghĩa của các giá trị:

1. R Square = 0.77

- Khoảng 77% biến động trong doanh thu có thể được giải thích bởi biến Marketing Spend
- Đây là mức vừa phải – đầu tư vào marketing có ảnh hưởng, nhưng còn các yếu tố khác chưa được đưa vào mô hình

2. Coefficient (Marketing Spend) = 52.9

- Với mỗi 1 đơn vị tăng trong chi phí đầu tư, doanh thu tăng trung bình khoảng 52.9 đơn vị tiền tệ, giữ các yếu tố khác không đổi
- Vì P-value của Marketing Spend rất nhỏ ($\approx 3.6e-33$) → kết quả có ý nghĩa thống kê, tức là mối quan hệ này là thật, không phải do ngẫu nhiên

3. Intercept = 2747.9

- Nếu chi phí đầu tư bằng 0, mô hình dự đoán doanh thu sẽ là 2747.9 (đây là giá trị lý thuyết và không có ý nghĩa thực tế trong kinh doanh – vì giả định là chi phí đầu tư > 0)

4. Significance F = 3.55e-33

- Mô hình tổng thể có ý nghĩa thống kê → đủ tin cậy để sử dụng

Góc nhìn kinh doanh:

- Chi phí đầu tư vào Marketing có ảnh hưởng rõ ràng đến doanh thu: Tăng chi phí có thể dẫn đến doanh thu tăng, nhưng khi đầu tư quá nhiều ta phải cân nhắc đến lợi nhuận và ROI.
- Mô hình này chưa hoàn hảo, vì nó chưa xem xét các yếu tố khác như khuyến mãi, vị trí cửa hàng, gần ga tàu hay không, v.v.

Bài toán: Xây dựng mô hình hồi quy đa biến có hiệu quả tốt hơn cho việc dự báo và đánh giá hiệu quả kinh doanh (Revenue)

Dữ liệu gốc

Product ID	Marketing_Spend	Discount_Ratio	Is_Near_Station	Store_Location	Day_of_Week	Store_Area	Num_Employees	Revenue
SP001	726	0.18	1	Quận 1	Thu	123	8	42480
SP002	357	0.22	1	Quận 3	Thu	198	8	17112
SP003	304	0.08	0	Quận 4	Tue	165	8	27020
SP004	596	0.08	1	Quận 2	Wed	84	9	37163
SP005	747	0.13	0	Quận 1	Fri	131	8	32008
SP006	480	0.22	1	Quận 3	Wed	152	7	24585
SP007	982	0.26	1	Quận 4	Fri	162	8	63839
SP008	716	0.19	1	Quận 4	Fri	145	8	50748
SP009	532	0.28	0	Quận 2	Fri	183	2	24638
SP010	452	0.15	1	Quận 4	Tue	170	4	31381
SP011	408	0.13	1	Quận 4	Thu	134	3	36740
SP012	756	0.14	1	Quận 4	Mon	181	5	51311
SP013	484	0.09	1	Quận 1	Thu	112	8	34534
SP014	153	0.26	1	Quận 3	Wed	61	9	6412
SP015	458	0.13	0	Quận 4	Fri	72	7	30078
SP016	764	0.19	0	Quận 3	Wed	200	4	41215
SP017	264	0.19	1	Quận 4	Wed	80	4	24979
SP018	257	0.18	1	Quận 2	Thu	184	4	23011
SP019	578	0.05	1	Quận 3	Mon	148	7	35024
SP020	578	0.30	1	Quận 4	Fri	55	9	34504
SP021	670	0.28	1	Quận 1	Mon	146	7	37226
SP022	864	0.10	1	Quận 4	Mon	143	6	57818

Dữ liệu đã qua tiền xử lý

Product ID	Marketing_Spend	Discount_Ratio	Is_Near_Station	Day_of_Week	Store_Area	Num_Employees	Quận 1_Quận 2	Quận 1_Quận 3	Quận 1_Quận 4	Revenue
SP001	726	0.18	1	4	123	8	0	0	0	42480
SP002	357	0.22	1	4	198	8	0	1	0	17112
SP003	304	0.08	0	2	165	8	0	0	1	27020
SP004	596	0.08	1	3	84	9	1	0	0	37163
SP005	747	0.13	0	5	131	8	0	0	0	32008
SP006	480	0.22	1	3	152	7	0	1	0	24585
SP007	982	0.26	1	5	162	8	0	0	1	63839
SP008	716	0.19	1	5	145	8	0	0	1	50748
SP009	532	0.28	0	5	183	2	1	0	0	24638
SP010	452	0.15	1	2	170	4	0	0	0	31381
SP011	408	0.13	1	4	134	3	0	0	1	36740
SP012	756	0.14	1	1	181	5	0	0	1	51311
SP013	484	0.09	1	4	112	8	0	0	0	34534
SP014	153	0.26	1	3	61	9	0	1	0	6412
SP015	458	0.13	0	5	72	7	0	0	1	30078
SP016	764	0.19	0	3	200	4	0	1	0	41215
SP017	264	0.19	1	3	80	4	0	0	0	24979
SP018	257	0.18	1	4	184	4	1	0	0	23011
SP019	578	0.05	1	1	148	7	0	1	0	35024
SP020	578	0.30	1	5	55	9	0	0	1	34504
SP021	670	0.28	1	1	146	7	0	0	0	37226

Kết quả phân tích hồi quy đa biến

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.975675
R Square	0.951943
Adjusted R Square	0.947137
Standard Error	3076.991
Observations	100

ANOVA					
	df	SS	MS	F	Significance F
Regression	9	1.7E+10	1875438536	198.0844024	2.5146E-55
Residual	90	8.5E+08	9467875.878		
Total	99	1.8E+10			

	Coefficients	Standard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	2096.855	1877.35	1.116925033	0.266999321	-1632.8213	5826.53203	-1632.8213	5826.53203
Marketing_Spend	50.93728	1.42945	35.63429006	7.45842E-55	48.0974344	53.7771206	48.0974344	53.7771206
Discount_Ratio	-36378.91	4398.69	-8.270403963	1.12703E-12	-45117.669	-27640.154	-45117.669	-27640.154
Is_Near_Station	6974.563	632.495	11.0270685	2.15543E-18	5718.00161	8231.12399	5718.00161	8231.12399
Day_of_Week	22.99946	230.706	0.099691702	0.920810815	-435.33802	481.336946	-435.33802	481.336946
Store_Area	6.884375	7.0745	0.973125068	0.333099374	-7.170358	20.939109	-7.170358	20.939109
Num_Employees	90.35045	140.489	0.643114251	0.521785173	-188.75537	369.456265	-188.75537	369.456265
Quận 1_Quận 2	5390.506	892.08	6.042626524	3.36726E-08	3618.23317	7162.77796	3618.23317	7162.77796
Quận 1_Quận 3	-2682.171	975.191	-2.750405413	0.007195105	-4619.5587	-744.78362	-4619.5587	-744.78362
Quận 1_Quận 4	7118.014	843.126	8.442405329	4.96189E-13	5442.99608	8793.03098	5442.99608	8793.03098

Hồi quy đa biến – Lý giải kết quả hồi quy

Tổng quan mô hình

- **R Square = 0.9519**: Gần **95.2% biến động của doanh thu** được giải thích bởi các biến đầu vào → mô hình có tính giải thích cao
- **Adjusted R Square = 0.9471**: Sau khi điều chỉnh cho số lượng biến, mô hình vẫn rất mạnh
- **Significance F = 2.51e-55**: Mô hình tổng thể có ý nghĩa thống kê rất cao → đáng tin cậy

Góc nhìn kinh doanh

Những yếu tố doanh nghiệp nên quan tâm:

- Chi phí đầu tư vào Marketing là yếu tố tăng doanh thu rất rõ ràng
- Khuyến mãi mạnh làm giảm doanh thu, có thể do làm giảm giá trị đơn hàng
- Vị trí gần ga tàu là lợi thế rõ rệt về doanh thu
- Cửa hàng ở Quận 2 và Quận 4 hoạt động hiệu quả hơn đáng kể so với Quận 1
- Quận 3 nên xem xét lại chiến lược – doanh thu thấp hơn đáng kể

Những yếu tố có thể bỏ qua:

- Ngày trong tuần, diện tích cửa hàng, và số nhân viên không có mối liên hệ đáng kể với doanh thu → có thể là nhiễu hoặc ảnh hưởng gián tiếp

Ý nghĩa từng biến

Biến	Hệ số	P-value	Ý nghĩa thống kê?	Giải thích
Intercept	2096.86	0.267	Không	Không đáng kể – không cần quan tâm quá
Marketing_Spend	50.94	7.46E-55	Có	Mỗi đơn vị tăng giá → doanh thu tăng ~51
Discount_Ratio	-36,378.91	1.13E-12	Có	Giảm giá mạnh → doanh thu giảm nhiều
Is_Near_Station	6974.56	2.15E-18	Có	Gần ga tàu giúp tăng doanh thu gần 7,000
Day_of_Week	23	0.92	Không	Không có ảnh hưởng đáng kể (nhiều)
Store_Area	6.88	0.33	Không	Diện tích không rõ ảnh hưởng
Num_Employees	90.35	0.52	Không	Số nhân viên không có ý nghĩa
Quận 2 (so với Quận 1)	5390.51	3.37E-08	Có	Quận 2 cao hơn Quận 1 ~5390
Quận 3 (so với Quận 1)	-2682.17	0.007	Có	Quận 3 doanh thu thấp hơn Quận 1 ~2682
Quận 4 (so với Quận 1)	7118.01	4.96E-13	Có	Quận 4 vượt trội hơn Quận 1 ~7118

Phương pháp nâng cao hiệu quả mô hình hồi quy trong Excel:

So sánh mô hình bằng R^2 hiệu chỉnh

R^2 thông thường tăng khi thêm biến mới, kể cả khi biến không có ý nghĩa. R^2 hiệu chỉnh (Adjusted R^2) khắc phục vấn đề này bằng cách tính đến số lượng biến độc lập.

- Xem giá trị "Adjusted R Square" trong kết quả phân tích
- Chọn mô hình có R^2 hiệu chỉnh cao nhất
- Cân bằng giữa độ phức tạp và độ chính xác

Kiểm tra đa cộng tuyến

Đa cộng tuyến xảy ra khi các biến độc lập có tương quan cao, làm giảm độ tin cậy của mô hình.

- Dùng Data Analysis → Correlation tạo ma trận tương quan
- Phát hiện biến có tương quan cao (>0.7)
- Áp dụng VIF qua hàm tùy chỉnh
- Loại bỏ hoặc kết hợp biến tương quan cao

Biến đổi dữ liệu trong Excel

Biến đổi dữ liệu cải thiện độ chính xác bằng cách làm cho mối quan hệ giữa các biến tuyến tính hơn.

- Logarithm: `=LN(cell)` hoặc `=LOG10(cell)`
- Căn bậc hai: `=SQRT(cell)`
- Nghịch đảo: `=1/cell`
- Kiểm tra dư bằng biểu đồ phân tán sau biến đổi

1 Xác định điểm ngoại lệ (outliers)

Sử dụng biểu đồ phần dư để phát hiện các điểm dữ liệu nằm xa khỏi phần còn lại, có thể ảnh hưởng đến mô hình.

3 Xem xét loại bỏ hoặc biến đổi biến

Nếu phát hiện biến gây nhiễu, hãy cân nhắc loại bỏ hoặc biến đổi biến để cải thiện mô hình.

2 Phát hiện mô hình sai lệch

Biểu đồ phần dư có thể cho thấy các mẫu hình như quan hệ phi tuyến, phương sai không đồng nhất, hoặc sự phụ thuộc giữa các phần dư, gợi ý rằng mô hình không phù hợp.

4 Thêm biến tương tác hoặc dùng mô hình nâng cao hơn

Ngoài việc loại bỏ biến, bạn cũng có thể thử thêm biến tương tác hoặc sử dụng các mô hình hồi quy nâng cao hơn như phân tích thành phần chính (PCA).

Tiền xử lý dữ liệu:

- Làm sạch, chuẩn hóa dữ liệu để đảm bảo độ chính xác.
- Bao gồm xử lý giá trị thiếu, loại bỏ ngoại lệ, và tạo biến giả.

Phân tích hồi quy:

- Dự báo, phân tích mối quan hệ nhân quả, và kiểm định giả thuyết.
- Sử dụng hồi quy tuyến tính (đơn/đa biến), đánh giá bằng R^2 , p-value, và hệ số.

Công cụ & Cải thiện:

- Ứng dụng Excel qua các hàm tiền xử lý (IFERROR, AVERAGE...) và Data Analysis ToolPak.
- Mô hình được cải thiện bằng cách so sánh R^2 điều chỉnh, xử lý đa cộng tuyến và ngoại lệ.

Phần 1: Trục Quan Hóa Dữ Liệu

Tập trung vào việc chọn và thiết kế các loại biểu đồ phổ biến (cột, đường, tròn, histogram, heatmap, waterfall, scatter) để trục quan hóa dữ liệu hiệu quả bằng Excel, sử dụng các công cụ như Insert Charts và Conditional Formatting.

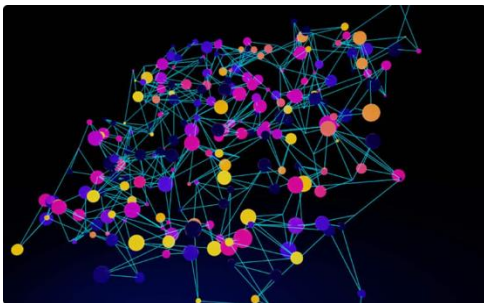
Phần 2: Kiểm Định Giả Thuyết & A/B Testing

Tìm hiểu các khái niệm kiểm định giả thuyết (p-value, t-value) và các loại kiểm định (t-test, ANOVA) - Giới thiệu về A/B Testing.

Phần 3: Tiền Xử Lý Dữ Liệu & Hồi Quy

Học cách tiền xử lý dữ liệu (làm sạch, chuẩn hóa, tạo biến giả) và xây dựng mô hình hồi quy tuyến tính (đơn/đa biến) để dự báo, phân tích mối quan hệ, và cải thiện độ chính xác.

Thank you for your attentions!



Hiểu sâu về dữ liệu

Nắm vững cách đọc và diễn giải dữ liệu để đưa ra các phân tích có giá trị.



Kỹ năng phân tích

Phát triển khả năng sử dụng các công cụ và phương pháp để phân tích dữ liệu hiệu quả.



Ứng dụng thực tế

Vận dụng kiến thức đã học vào các tình huống kinh doanh thực tế để giải quyết vấn đề.



Phát triển nghề nghiệp

Nâng cao năng lực chuyên môn, mở ra cơ hội phát triển trong lĩnh vực phân tích dữ liệu.

Next Steps

- Tiếp tục thực hành với dữ liệu thực tế để củng cố kiến thức.
- Khám phá thêm các tính năng nâng cao của Excel và các công cụ phân tích khác.
- Áp dụng những kỹ năng đã học để đưa ra các quyết định sáng suốt và hiệu quả hơn trong công việc.