

LABO2

BUSINESS ANALYSIS

NGUYEN MINH NHUT

INFORMATION SYSTEM ENGINEERING 1

statistics

plural noun

US  /stə'tistikz/

(infml **stats**, US /stæts/)

a collection of numerical facts or measurements, as about people, business conditions, or weather:

Add to word list 

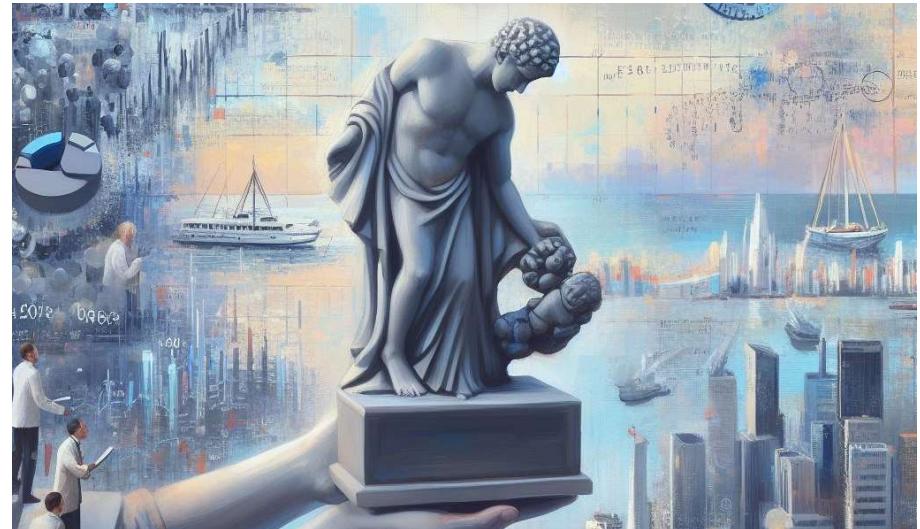


Statistics in daily life



2

Utilizing Statistics to Measure Results



Trong chương này chúng ta sẽ học những nội dung:

- Ý nghĩa của phân tích thống kê diễn giải
- Phương pháp kiểm định trung bình
- T-Test, Levene Test, ANOVA one way test
- ANOVA two-way test, Turkey Test
- Chi-Square Test
- Post Hoc Test

LAB02

THỐNG KÊ DIỄN GIẢI VÀ PHÂN TÍCH PHƯƠNG SAI



PHẦN 1

KHÁI NIỆM VỀ THỐNG KÊ DIỄN GIẢI

Thống kê suy diễn là quá trình sử dụng dữ liệu từ một mẫu nhỏ để đưa ra nhận định về toàn bộ dân số. Thông qua phương pháp này, chúng ta có thể đưa ra kết luận và dự đoán với mức độ tin cậy, giúp hiểu rõ hơn về các đặc điểm của tập dữ liệu lớn.



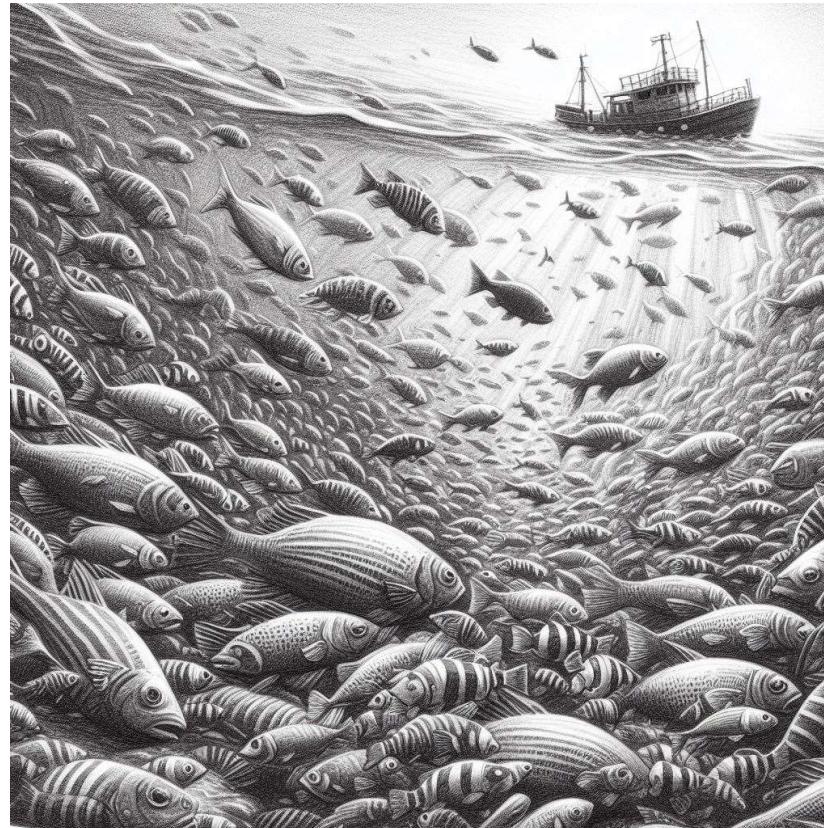
Phần 1. Khái niệm về Thống kê diễn giải



- **Đặt vấn đề về suy luận hoặc suy diễn**
 - Có **2 trường hợp chính** sử dụng suy diễn hoặc suy luận:
 - **Trường hợp 1:** Muốn ước lượng trung bình của tổng thể, nhưng chỉ biết trung bình của mẫu. → Có phương pháp nào để kiểm tra điều này?
 - **Trường hợp 2:** Muốn xem biến nào ảnh hưởng đến biến khảo sát nhiều nhất. (Chương 3 đã học cách tính hệ số tương quan)
 - **Hầu hết chúng ta thường lấy mẫu để thực nghiệm, việc lấy tổng thể là vô cùng tốn kém dẫn đến không khả thi**

Phần 1. Khái niệm về Thống kê diễn giải

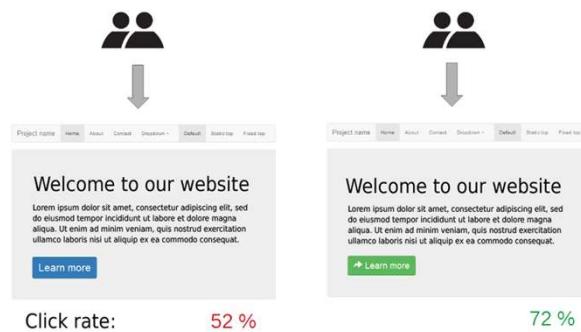
- Đặt vấn đề về suy luận hoặc suy diễn



Phần 1. Khái niệm về Thống kê diễm giải



- **Thiết kế thực nghiệm (Design Experiment)**
 - Thiết kế thực nghiệm là một nền tảng quan trọng trong tất cả các lĩnh vực nghiên cứu.
 - Việc chấp nhận hay từ chối giả thuyết phải có minh chứng trước khi đưa vào mô hình dự báo/dự đoán.
 - Mỗi khi nói về ý nghĩa thống kê (statistics significants) hay t-test hoặc p-values thì nó đóng vai trò là “**pipeline**” trong thống kê suy diễn.
 - Quy trình bắt đầu bằng một giả thuyết “Website A đạt tiêu chuẩn hơn mong đợi”

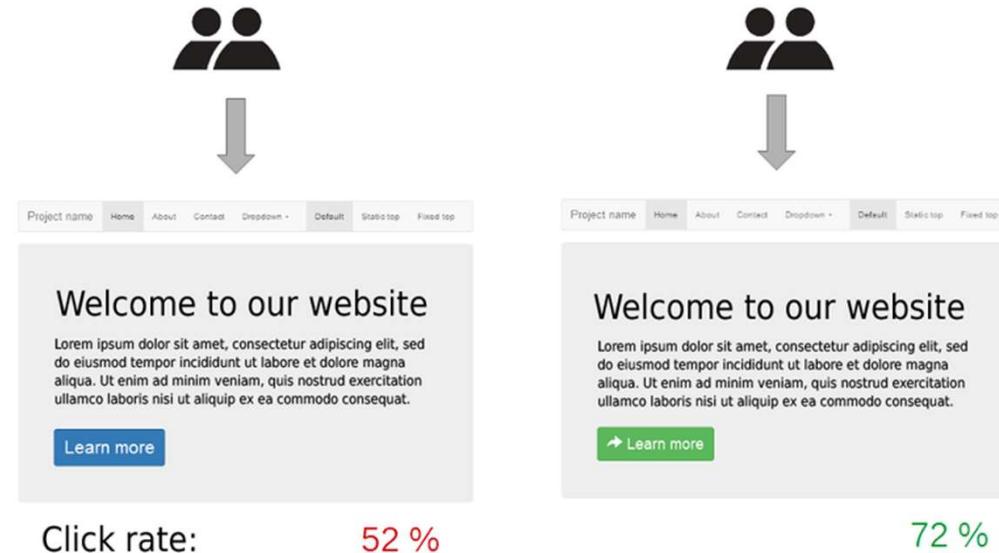


Phần 1. Khái niệm về Thống kê diễm giải



• Thiết kế thực nghiệm (Design Experiment)

– Quy trình bắt đầu bằng một giả thuyết “Website B có tỉ lệ lượt click nhiều hơn Website A”. Một thí nghiệm nhỏ như vậy được gọi là kiểm định A/B.



Phần 1. Khái niệm về Thống kê diễn giải



- **Thiết kế thực nghiệm (Design Experiment)**
 - Quy trình bắt đầu bằng một giả thuyết “Website B có tỉ lệ lượt click nhiều hơn Website A”. Một thí nghiệm nhỏ như vậy được gọi là kiểm định A/B.
 - Kiểm định A/B là được thiết kế sao cho hy vọng sẽ mang lại kết quả kết luận. Dữ liệu được thu thập và phân tích, sau đó rút ra một kết luận.
 - Kiểm định A/B trên hai groups

Phần 1. Khái niệm về Thống kê diễn giải



- **So sánh thống kê cổ điển và khoa học về thống kê diễn giải**

- Trong thống kê cổ điển sẽ thường đặt giả thuyết như sau:

Sự khác nhau giữa giá A và giá B có ý nghĩa thống kê hay không?

- Trong khoa học dữ liệu thì sẽ ít đặt giả thuyết trên hơn mà sẽ đặt như sau:

Trong số nhiều loại giá A, B, C, D,... giá nào là tốt nhất.

Phần 1. Khái niệm về Thống kê diễn giải



• Bài học về sự đồng ý (Getting Permission)

Trong nghiên cứu khoa học và y học liên quan đến đối tượng nghiên cứu là **con người**, thường cần phải có **sự đồng ý của họ**, cũng như phải **có sự chấp thuận từ một ủy ban đánh giá tổ chức**. Những thí nghiệm trong lĩnh vực kinh doanh thường được thực hiện như một phần của các hoạt động điều hành **không bao giờ thực hiện điều này**. Trong hầu hết các trường hợp (ví dụ như thí nghiệm giá cả hoặc thí nghiệm về việc hiển thị tựa đề nào hoặc đề xuất nào nên được thực hiện), thực hành này được rộng rãi chấp nhận. Tuy nhiên, vào năm 2014, Facebook đã vi phạm sự chấp nhận chung này khi thực hiện thí nghiệm với tâm trạng cảm xúc trong bảng tin tin tức của người dùng. **Facebook** đã sử dụng phân tích tâm trạng để phân loại bài đăng trên bảng tin tin tức là tích cực hoặc tiêu cực, sau đó **điều chỉnh cân bằng tích cực/tiêu cực** trong những gì nó hiển thị cho người dùng. Một số người dùng được chọn ngẫu nhiên trải qua nhiều bài đăng tích cực hơn, trong khi người khác trải qua nhiều bài đăng tiêu cực hơn. **Facebook** phát hiện rằng những người dùng trải qua bảng tin tích cực hơn có khả năng cao hơn để đăng bài tích cực, và ngược lại. Tuy nhiên, hiệu ứng này nhỏ, và **Facebook** đã phải đổi mặt với nhiều chỉ trích vì thực hiện thí nghiệm mà **không có sự biết đến của người dùng**. Một số người dùng còn đưa ra ý kiến cho rằng **Facebook** có thể đã đẩy mạnh một số người dùng cảm xúc rất buồn bã nếu họ nhận được phiên bản tiêu cực của bảng tin của mình.

Phần 1. Khái niệm về Thống kê diễn giải



- **Phân biệt khái niệm giả thuyết và giả thiết [4]**
 - **Giả thuyết (Hypothesis)**: là nhận định sơ bộ, cần được chứng minh qua một câu hỏi nghiên cứu nhất định.
Ví dụ 1: “Trái Đất chuyển động xung quanh mặt trời”.
Ví dụ 2: “Giá tiền ảo BTC tháng 12/2023 đang giảm mạnh”.
Ví dụ 3: “Chất lượng không khí đang giảm dần đều”.
- **Giả thiết (Assumption)**: Là một điều kiện giả định trong quan sát và thực nghiệm.
Ví dụ 4: Một xe đang chuyển động với **vận tốc 58km/h** thì **hãm phanh**.
Ngầm hiểu hãm phanh: **Từ vận động 58km/h → 0km/h**

Phần 1. Khái niệm về Thống kê diễn giải



- **Một số khái niệm về kiểm định giả thuyết (Hypothesis Testing)**
 - **Kiểm định giả thuyết (Hypothesis Testing)** hay một tên khác là kiểm định ý nghĩa, là một **phương pháp thống kê** được sử dụng để đưa ra quyết định về một giả thuyết nghiên cứu dựa trên dữ liệu mẫu. Mục tiêu chính của **kiểm định giả thuyết** là xác định xem có **đủ bằng chứng từ dữ liệu** để **bắc bỏ một giả thuyết** (null hypothesis) hay không.

Phần 1. Khái niệm về Thống kê diễn giải

- Một số khái niệm về kiểm định giả thuyết (Hypothesis Testing)



- Bài toán kiểm tra xem phân bón X có ảnh hưởng đến sự tăng trưởng cây trồng không?
- **Giả thuyết:** Phân bón X không làm cây trồng tăng trưởng
- **Đối thuyết:** Phân bón X làm cây trồng tăng trưởng

Phần 1. Khái niệm về Thống kê diễn giải



- Một số khái niệm về kiểm định giả thuyết (Hypothesis Testing)

	Giả thuyết H_0	Đối thuyết H_1
Phát biểu	Trường hợp không ảnh hưởng đến tổng thể	Trường hợp ảnh hưởng đến tổng thể
Keyterms	Không ảnh hưởng, không thay đổi, không khác, không quan hệ	Ảnh hưởng, Thay đổi, Khác nhau, Quan hệ
Ký hiệu	$=, \geq, \leq$	$\neq, >, <$

Phần 1. Khái niệm về Thống kê diễn giải



• **Bài tập xác định giả thuyết, đối thuyết**

Bài tập 1: Một nhóm nghiên cứu muốn xác định xem liệu việc tập thể dục hàng ngày có ảnh hưởng đến cân nặng của người trưởng thành hay không?

Bài tập 2: Một doanh nghiệp muốn kiểm tra xem việc giảm giá sản phẩm có tăng doanh số bán hàng hay không?

Bài tập 3: Một trường học đang thử nghiệm một phương pháp giảng dạy mới và muốn xác định xem nó có cải thiện hiệu suất học tập hay không?

Bài tập 4: Một tổ chức muốn xác định xem việc giảm lượng rác thải có dẫn đến sự cải thiện về chất lượng môi trường hay không?

Bài tập 5: Một nhóm nghiên cứu muốn biết liệu việc sử dụng phần mềm chống virus mới có giảm nguy cơ bị tấn công malware hay không?

Phần 1. Khái niệm về Thống kê diễn giải



• Kiểm định One Ways

- **Đặc điểm chính:** Thường được sử dụng khi bạn quan tâm đến sự khác biệt ở một hướng cụ thể.
 - Thường đối thuyết là: $>$, $<$
 - **Giả thuyết không thay đổi (null hypothesis):** Không có sự khác biệt giữa các nhóm hoặc điều kiện.
 - **Giả thuyết thay thế (alternative hypothesis):** Có sự khác biệt ở một hướng cụ thể (lớn hơn hoặc nhỏ hơn).
 - **Ví dụ 1:** Trung bình học sinh giỏi của tỉnh A lớn hơn tỉnh B.
 - **Giả thuyết:** $\mu_A \leq \mu_B$
 - **Đối thuyết:** $\mu_A > \mu_B$

Phần 1. Khái niệm về Thống kê diễn giải



• Kiểm định Two-way

- **Đặc điểm chính:** Thường được sử dụng khi bạn quan tâm đến sự khác biệt mà không giới hạn ở một hướng cụ thể. Nó cũng được gọi là kiểm định hai phía.
- Thường đối thuyết là: \neq
 - **Giả thuyết không thay đổi (null hypothesis):** Không có sự khác biệt giữa các nhóm hoặc điều kiện.
 - **Giả thuyết thay thế (alternative hypothesis):** Có sự khác biệt giữa các nhóm hoặc điều kiện.
- **Ví dụ 2:** Trung bình học sinh giỏi của tỉnh A khác tỉnh B.
 - **Giả thuyết:** $\mu_A = \mu_B$
 - **Đối thuyết:** $\mu_A \neq \mu_B$

Phần 1. Khái niệm về Thống kê diễn giải



• Mức ý nghĩa alpha

- Mức ý nghĩa alpha α
- Mức ý nghĩa thường là: **5% và 1%**
- Là một ngưỡng quyết định được chọn trước trong quá trình thực hiện một kiểm định thống kê.
- Nó xác định **mức độ chấp nhận được của rủi ro** loại bỏ giả thuyết không đổi khi thực sự có sự khác biệt.
- Nếu giả sử rằng **không có sự khác biệt** thực sự (giả thuyết không đổi), mức ý nghĩa là **xác suất tối đa** mà chúng ta sẽ chấp nhận kết quả thống kê mà theo đó giả thuyết **không đổi bị loại bỏ**. Phổ biến nhất là sử dụng các **mức ý nghĩa như 0.05 hoặc 0.01**, tương ứng với 5% hoặc 1%. Điều này có **nghĩa là nếu giá trị p** (xác suất của kết quả thống kê) nhỏ hơn **mức ý nghĩa** đã chọn, chúng ta sẽ bác bỏ giả thuyết không đổi.

Phần 1. Khái niệm về Thống kê diễn giải



- Test statistics Formula (Unknown Sd):

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

- \bar{x} : Là giá trị trung bình của tập dữ liệu.
- μ_0 : Là trung bình giả thuyết bài toán đề ra.
- s : Là độ lệch chuẩn mẫu của tập dữ liệu
- n: Là số lượng phần tử của tập dữ liệu

- F-Critical

- $F = f(1-\alpha; n - 1)$ cho trường hợp kiểm định 1 phía
- $F = f(1 - \frac{\alpha}{2}; n - 1)$ cho trường hợp kiểm định 2 phía

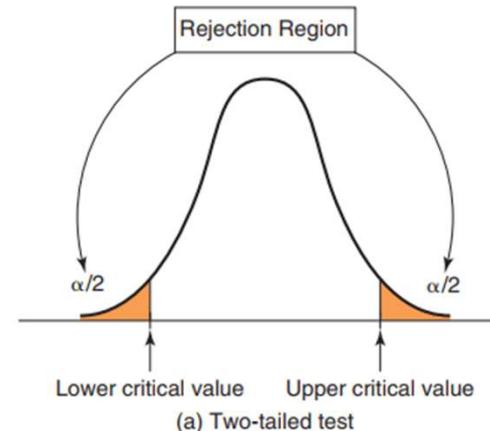
- P-Value (T-Distribution CDF) = CDF(t,n-1)

Phần 1. Khái niệm về Thống kê diễn giải



• Mức ý nghĩa alpha

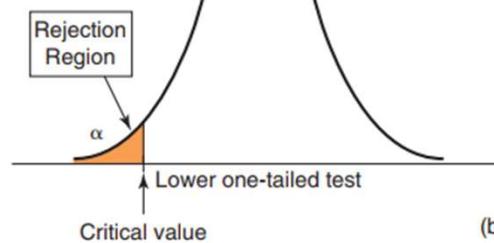
Trường hợp Đôi thuyết là !=



Từ chối H_0 khi
 $t > +F$ hoặc
 $t < -F$

Trường hợp Đôi thuyết $H_1 <$

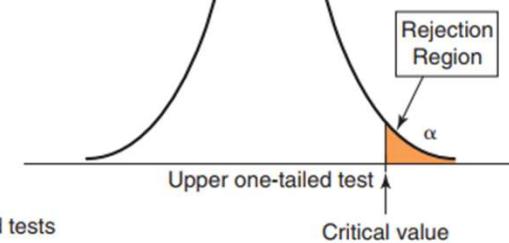
Từ chối H_0 khi
 $t < F$



(b) One-tailed tests

Trường hợp Đôi thuyết $H_1 >$

Từ chối H_0 khi
 $t > F$



Phần 1. Khái niệm về Thống kê diễn giải



- **F-Critical Value của kiểm định t-test**
 - Tra bảng **F-Distribution**
 - Hoặc sử dụng các phần mềm tra giá trị **F-Critical Value**
 - Trong ngôn ngữ lập trình Python kiểm định 1 phía

```
from scipy.stats import t

# Define the significance level (alpha)
alpha = 0.05

# Define the degrees of freedom (n - 1)
degrees_of_freedom = 43 # Replace with your actual degrees of freedom

# Find the t-critical value
t_critical = t.ppf(1 - alpha, degrees_of_freedom)

# Print the result
print(f"t-Critical Value: {t_critical}")
```

Phần 1. Khái niệm về Thống kê diễn giải



- **F-Critical Value của kiểm định t-test**
 - Tra bảng **F-Distribution**
 - Hoặc sử dụng các phần mềm tra giá trị **F-Critical Value**
 - Trong ngôn ngữ lập trình Python kiểm định 2 phía

```
from scipy.stats import t

# Define the significance level (alpha)
alpha = 0.05

# Define the degrees of freedom (n - 1)
degrees_of_freedom = 34 # Replace with your actual degrees of freedom

# Find the t-critical value
t_critical = t.ppf(1 - alpha/2, degrees_of_freedom)

# Print the result
print(f"t-Critical Value: {t_critical}")
```

Phần 1. Khái niệm về Thống kê diễn giải



• Giá trị p-value

- Giá trị p (p-value) là một đại lượng thống kê được sử dụng trong các phân tích thống kê để **đánh giá mức độ hỗ trợ cho** hoặc **chống lại giả thuyết không đổi** (null hypothesis). Nó đo lường xác suất của việc nhận được kết quả thống kê hoặc một kết quả càng "cực kỳ" như vậy (hoặc "cực kỳ khác biệt") so với giả thuyết không đổi, dựa trên giả định rằng không có sự khác biệt thực sự giữa các nhóm hoặc điều kiện.
 - Nếu giá trị **p < alpha**: Bác bỏ giả thuyết không đổi và kết luận rằng có đủ bằng chứng để nói rằng có sự khác biệt đáng kể.
 - Nếu giá trị **p >= alpha**: Không đủ bằng chứng để bác bỏ giả thuyết không đổi, và kết luận rằng không có đủ chứng cứ để nói rằng có sự khác biệt đáng kể

Phần 1. Khái niệm về Thống kê diễn giải



- **p-value cho của kiểm định t-test**

- Bác bỏ giả thuyết khi $p < \alpha$

```
from scipy.stats import t

# Define the value for which you want to find the cumulative distribution
x = -1.05

# Define the degrees of freedom
degrees_of_freedom = 43

# Calculate the cumulative distribution function (CDF)
t_distribution_cdf = t.cdf(x, degrees_of_freedom)

# Print the result
print(f"T-Distribution CDF: {t_distribution_cdf}")
```



PHẦN 2

PHƯƠNG PHÁP KIỂM ĐỊNH TRUNG BÌNH

Kiểm định trung bình sử dụng t-test để đối chiếu giá trị trung bình giữa các nhóm. P-value đánh giá mức độ khác biệt và xác định tính ý nghĩa thống kê.



Phần 2. Phương pháp kiểm định trung bình



- **Các loại trường hợp trong kiểm định giả thuyết**
 - Giả thuyết H_0 thật sự là **đúng**, và kết luận là **chấp nhận** H_0 .
 - Giả thuyết H_0 thật sự là **sai**, và kết luận là **từ chối** H_0 .
 - Giả thuyết H_0 thật sự là **đúng**, và kết luận là **từ chối** H_0 . (Sai lầm loại I – Type Error I)
 - Giả thuyết H_0 thật sự là **sai**, và kết luận là **chấp nhận** H_0 . (Sai lầm loại II – Type Error II)
 - Loại sai lầm nào là **nguy hiểm nhất**?

Phần 2. Phương pháp kiểm định trung bình



- **Bài toán kiểm định trung bình t-test**

- Cho dữ liệu **CabSoft** về thời gian phản hồi sửa chữa của dịch vụ sửa máy tính 44 mẫu như sau:

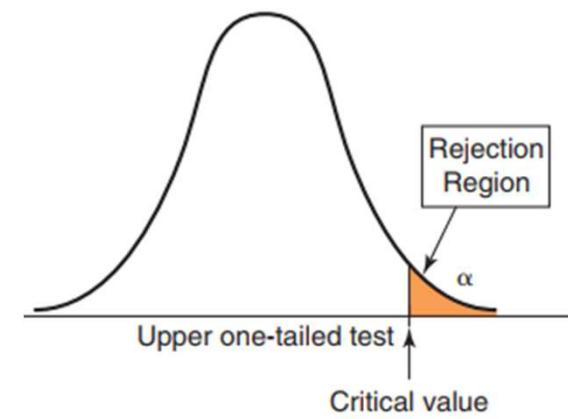
20	12	15	11	22	6	39	19	12	13	13
19	47	24	19	17	13	8	33	21	28	13
2	25	25	48	12	118	27	11	21	5	33
29	2	25	61	15	11	2	31	20	2	15

- Có thể cho rằng cho rằng thời gian sửa chữa trung bình là **lớn hơn** 25 được hay không?
 - Đây là kiểm định **one-ways**
 - **Phát biểu bài toán:**
 - Giả thuyết: $H_0 \mu_A \leq 25$
 - Đối thuyết: $H_1 \mu_A > 25$

Phần 2. Phương pháp kiểm định trung bình



- **Bài toán kiểm định trung bình**
 - **Phát biểu bài toán:**
 - Giả thuyết: $H_0 \mu_A \leq 25$
 - Đối thuyết: $H_1 \mu_A > 25$
 - Trung bình thời gian sửa chữa:
 - Phương sai:
 - Giá trị t-test:
 - Giá trị F-Critical của t-test:
 - Giá trị p-value của t-test:
 - Kết luận



Phần 2. Phương pháp kiểm định trung bình



- **Giới thiệu về Degree of Freedom là gì?**
 - Trong ngữ cảnh của kiểm định thống kê, "degrees of freedom" (độ tự do) là một khái niệm quan trọng và được sử dụng trong nhiều phương pháp thống kê, bao gồm cả kiểm định t-test và phân tích phương sai (ANOVA).
 - Độ tự do là số lượng giá trị trong quá trình thống kê mà có thể thay đổi độc lập mà không làm thay đổi giá trị cuối cùng của thống kê. Trong kiểm định t-test, degree of freedom thường được tính dựa trên kích thước của mỗi mẫu (số lượng quan sát) và số mẫu.
 - $df = n - k$, với ***n*** số lượng quan sát và ***k*** là số mẫu

Phần 2. Phương pháp kiểm định trung bình



- **Bài toán kiểm định trung bình hai biến (t-Test hai mẫu)**
 - Giá trị t trong **t-Test hai biến (Independent Samples T-Test)**

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- **Degree of Freedom:**
 - $df = n_1 + n_2 - 2$
- **F-Critical**
 - $F = f(1-\alpha/2; df)$ cho trường hợp kiểm định
- **P-Value (T-Distribution CDF) = CDF(t,df)**

Phần 2. Phương pháp kiểm định trung bình



- **Bài toán kiểm định trung bình hai biến (t-Test hai mẫu)**
 - Một cách viết khác của kiểm định t-Test hai mẫu

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Phần 2. Phương pháp kiểm định trung bình



- F-Critical Value của kiểm định t-test kiểm định tính độc lập
 - Trong ngôn ngữ lập trình Python **kiểm định 1 phía độc lập**

```
▶ import scipy.stats as stats

# Set the desired level of significance (alpha)
alpha = 0.05

# Degrees of freedom (df) for the independent t-test
# df = n1 + n2 - 2, where n1 and n2 are the sample sizes of the two groups
n1 = 20 # sample size of group 1
n2 = 25 # sample size of group 2
df = n1 + n2 - 2

# Find the critical value for a two-tailed test
critical_value = stats.t.ppf(1 - alpha / 2, df)

print(f'Critical Value: {critical_value}')

Critical Value: 2.0166921941428133
```

Phần 2. Phương pháp kiểm định trung bình

- p-Value của kiểm định t-test kiểm định tính độc lập
 - Trong ngôn ngữ lập trình Python **kiểm định 1 phía độc lập**

```
from scipy.stats import t

t_value = 4.3
df = 18
p_value = 2 * (1 - t.cdf(abs(t_value), df))
p_value

0.0004310971346184189
```

Phần 2. Phương pháp kiểm định trung bình



- **Bài toán kiểm định trung bình t-test Two-Sample**
 - Tỉ lệ đăng ký bậc tiểu học của **Việt Nam** từ năm 2013 – 2020 được thống kê như sau:

102.1	105.5	106.8	110.2	109.9
112.2	117.4	119.0	120.0	123.1

- Tỉ lệ đăng ký bậc tiểu học của **Thái Lan** từ năm 2013 – 2020 được thống kê như sau:

102.7	107.7	105.4	106.2	101.1
99.2	99.3	99.4	99.5	101.6

- Có thể cho rằng trung bình *tỉ lệ đăng ký* bậc tiểu của hai quốc gia từ năm 2013 đến 2020 là **bằng nhau** được hay không?

Phần 2. Phương pháp kiểm định trung bình



- **Bài toán kiểm định trung bình t-test Two-Sample**
 - Lượng khí thải CO₂ (tấn bình quân đầu người) của **Việt Nam** từ năm 2013 – 2020 được thống kê như sau:

1.820	1.981	2.186	2.384
2.445	3.015	3.568	3.676
 - Lượng khí thải CO₂ (tấn bình quân đầu người) của **Singapore** từ năm 2013 – 2019 được thống kê như sau:

3.804	3.736	3.825	3.849
3.794	3.768	3.718	
 - Có thể cho rằng trung bình **Lượng khí thải CO₂ (tấn bình quân đầu người)** của hai quốc gia là **bằng nhau** được hay không? Giải thích

Phần 2. Phương pháp kiểm định trung bình



• Bài toán kiểm định trung bình t-test Two-Sample

- Gọi **X1** là tỉ lệ đăng ký tiểu học Việt Nam
- Gọi **X2** là tỉ lệ đăng ký tiểu học Thái Lan
- Ta có trung bình tỉ lệ của từng quốc gia
 - $\bar{X}_1 = 112.6$
 - $\bar{X}_2 = 102.2$
 - $s_1^2 = 48.61969132$
 - $s_2^2 = 10.02333921$
 - $n_1 = 10$
 - $n_2 = 10$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 4.30839103$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

– Phát biểu bài toán

- Giả thuyết: $H_0 \mu_{VietNam} = \mu_{ThaiLan}$
- Đối thuyết: $H_1 \mu_{VietNam} \neq \mu_{ThaiLan}$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Phần 2. Phương pháp kiểm định trung bình



• Bài toán kiểm định trung bình t-test Two-Sample

– Ta có trung bình tỉ lệ của từng quốc gia

- $\bar{X}_1 = 112.6$
- $\bar{X}_2 = 102.2$
- $s_1^2 = 48.61969132$
- $s_2^2 = 10.02333921$
- $n_1 = 10$
- $n_2 = 10$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 4.30839103$$

– Phát biểu bài toán

- Giả thuyết: $H_0: \mu_{VietNam} = \mu_{ThaiLan}$
- Đối thuyết: $H_1: \mu_{VietNam} \neq \mu_{ThaiLan}$

– F = t-Critical value = 2.101 < t (Từ chối H0)

Phần 2. Phương pháp kiểm định trung bình



• Bài toán kiểm định trung bình t-test Two-Sample

– Ta có trung bình tỉ lệ của từng quốc gia

- $\bar{X}_1 = 112.6$
- $\bar{X}_2 = 102.2$
- $s_1^2 = 48.61969132$
- $s_2^2 = 10.02333921$
- $n_1 = 10$
- $n_2 = 10$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 4.30839103$$

– Phát biểu bài toán

• Giả thuyết: $H_0: \mu_{VietNam} = \mu_{ThaiLan}$

• Đốii thuyết: $H_1: \mu_{VietNam} \neq \mu_{ThaiLan}$

– p-value = 0.0004 < 0.005 (Từ chối H0)



PHẦN 3

KIỂM ĐỊNH CHI-SQUARE

39

Kiểm định chi-square là một phương pháp thống kê được sử dụng để kiểm tra sự tương quan giữa hai biến phân loại. Nó dựa trên so sánh giữa tần suất quan sát và tần suất mong đợi của các sự kiện trong một bảng tần suất.



Phần 3. Kiểm định Chi-Square là gì?



- **Giới thiệu về kiểm định Chi-Square**

- Kiểm định trên hai nhóm dựa trên tần suất xuất hiện
- Giả thuyết H_0 : Hai nhóm là độc lập nhau
- Đối thuyết H_1 : Hai nhóm là phụ thuộc nhau

A	B	C	D	E	F	G	H	I
1	Energy Drink Survey							
2	Respondent	Gender	Brand Preference	Count of Respondent	Column Labels			
3	1	Male	Brand 3	Count of Respondent	Column Labels			
4	2	Female	Brand 3	Row Labels	Brand 1	Brand 2	Brand 3	Grand Total
5	3	Male	Brand 3	Female		9	6	22
6	4	Male	Brand 1	Male		25	17	63
7	5	Male	Brand 1	Grand Total		34	23	43
8	6	Female	Brand 2					100
9	7	Male	Brand 2					
10								

Phần 3. Kiểm định Chi-Square là gì?



- **Giới thiệu về kiểm định Chi-Square**

- Kiểm định trên hai nhóm dựa trên tần suất xuất hiện
- Giả thuyết H_0 : Hai nhóm là độc lập nhau
- Đối thuyết H_1 : Hai nhóm là phụ thuộc nhau

A	B	C	D	E	F	G	H	I
1	Energy Drink Survey							
2	Respondent	Gender	Brand Preference	Count of Respondent	Column Labels			
3	1	Male	Brand 3	Count of Respondent	Column Labels			
4	2	Female	Brand 3	Row Labels	Brand 1	Brand 2	Brand 3	Grand Total
5	3	Male	Brand 3	Female		9	6	22
6	4	Male	Brand 1	Male		25	17	63
7	5	Male	Brand 1	Grand Total		34	23	43
8	6	Female	Brand 2					100
9	7	Male	Brand 2					
10								

$$f_e = \text{expected}_{Female, Brand1} = \frac{\text{SumFemale} * \text{SumBrand1}}{\text{SumAll}} = 12.58$$

Phần 3. Kiểm định Chi-Square là gì?



• Giới thiệu về kiểm định Chi-Square

- Kiểm định trên hai nhóm dựa trên tần suất xuất hiện
- Giả thuyết H_0 : Hai nhóm là độc lập nhau
- Đối thuyết H_1 : Hai nhóm là phụ thuộc nhau

Chi-Square Test					
Count of Respondent	Column Labels				
Row Labels	Brand 1	Brand 2	Brand 3	Grand Total	
Female		9	6	22	37
Male	25	17	21		63
Grand Total	34	23	43	100	
Expected Frequency					
Female	12.58	8.51	15.91		37
Male	21.42	14.49	27.09		63
Grand Total	34	23	43	100	

Expected frequency of Female and Brand 1 = $37 \times 34 / 100$

Phần 3. Kiểm định Chi-Square là gì?



- **Giới thiệu về kiểm định Chi-Square**

- Kiểm định trên hai nhóm dựa trên tần suất xuất hiện
- Giả thuyết H_0 : Hai nhóm là độc lập nhau
- Đối thuyết H_1 : Hai nhóm là phụ thuộc nhau
- **Công thức tính Chi-Square:**

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

- Nếu $\chi^2 > F_{critical}$
- Từ chối H_0
- Hoặc có thể dùng p-value để so sánh với 0.05

Phần 3. Kiểm định Chi-Square là gì?



- **F-Critical của Chi-Square**

- Trong ngôn ngữ lập trình Python **kiểm định Chi-Square**

```
▶ from scipy.stats import chi2

# Độ tự do (degrees of freedom)
df = (2-1)*(3-1)

# Giá trị alpha (mức ý nghĩa)
alpha = 0.05

# Tính giá trị critial
critical_value = chi2.ppf(1 - alpha, df)

# In kết quả
print(f"Critical Value at alpha={alpha} for {df} degrees of freedom: {critical_value}")

Critical Value at alpha=0.05 for 2 degrees of freedom: 5.991464547107979
```

Phần 3. Kiểm định Chi-Square là gì?



- p-Value của kiểm định Chi-Square kiểm định tính độc lập
 - Trong ngôn ngữ lập trình Python **kiểm định Chi-Square**



```
import numpy as np
from scipy.stats import chi2

chi_square = 6.49
number_group_a = 2
number_group_b = 3
# Tính p-value
p_value = 1 - chi2.cdf(chi_square, (number_group_a-1)*(number_group_b-1))

# In kết quả
print(f"P-Value: {p_value}")
```

```
P-Value: 0.03896856435728524
```



PHẦN 4 PHÂN TÍCH PHƯƠNG SAI

Phân tích phương sai (ANOVA) là phương pháp thống kê kiểm tra sự khác biệt trung bình giữa ba hoặc nhiều nhóm. Nó đo lường độ biến động giữa các nhóm để đánh giá tính đáng kể.



Phần 4. Phân tích phương sai



- **Phân tích phương sai là gì?**
 - *Phân tích phương sai (ANOVA)* là phương pháp thống kê kiểm tra sự khác biệt **trung bình giữa ba hoặc nhiều nhóm**. Nó đo lường độ biến động giữa các nhóm để đánh giá tính đáng kể.
 - Mặc dù mục tiêu của ANOVA thường là **so sánh trung bình** giữa các nhóm, nhưng nó không chỉ đơn thuần là "**phân tích trung bình**" vì phương pháp này cũng kiểm tra phương sai để đảm bảo rằng sự khác biệt không chỉ là do biến động ngẫu nhiên. Do đó, tên "**Phân tích phương sai**" là một cách tổng quát và chính xác hơn để mô tả kỹ thuật này.
 - Kỹ thuật phân tích **phương sai ANOVA** viết tắt của từ **Analysis Of Variance**.

Phần 4. Phân tích phương sai

• Các bước tiến hành một thuật toán kiểm định ANOVA



Algorithm 1 ANOVA with Levene's Test and Tukey's Post Hoc Test

```
1: Collect Data:
2: for each group in {Group1, Group2, Group3} do
3:   Collect data for the group
4:   Store data in Pandas DataFrame
5: end for
6: Levene's Test for Homogeneity of Variances:
7: Use Levene's test to check for homogeneity of variances
8: if p-value from Levene's test is significant then
9:   Adjust data or use Welch's ANOVA
10: end if
11: ANOVA:
12: Perform one-way ANOVA on the data
13: Tukey's Post Hoc Test:
14: Perform Tukey's post hoc test for pairwise group comparisons
15: Provide ANOVA Results:
16: Output ANOVA results, including F-statistic and p-value
17: Provide Tukey's Results:
18: Output Tukey's post hoc test results, including adjusted p-values
```

Phần 4. Phân tích phương sai

• Tập dữ liệu Insurance Survey ANOVA



Insurance Survey						
Age	Gender	Education	Marital Status	Years Employed	Satisfaction*	Premium/Deductible**
36	F	Some college	Divorced	4	4	N
55	F	Some college	Divorced	2	1	N
61	M	Graduate degree	Widowed	26	3	N
65	F	Some college	Married	9	4	N
53	F	Graduate degree	Married	6	4	N
50	F	Graduate degree	Married	10	5	N
28	F	College graduate	Married	4	5	N
62	F	College graduate	Divorced	9	3	N
48	M	Graduate degree	Married	6	5	N
31	M	Graduate degree	Married	1	5	N
57	F	College graduate	Married	4	5	N
44	M	College graduate	Married	2	3	N
38	M	Some college	Married	3	2	N
27	M	Some college	Married	2	3	N
56	M	Graduate degree	Married	4	4	Y
43	F	College graduate	Married	5	3	Y
45	M	College graduate	Married	15	3	Y
42	F	College graduate	Married	12	3	Y
29	M	Graduate degree	Single	10	5	N
28	F	Some college	Married	3	4	Y
36	M	Some college	Divorced	15	4	Y
49	F	Graduate degree	Married	2	5	N
46	F	College graduate	Divorced	20	4	N
52	F	College graduate	Married	18	2	N

*Measured from 1-5 with 5 being highly satisfied.

**Would you be willing to pay a lower premium for a higher deductible?

F	G	H	I	J
College graduate	Graduate degree	Some college		
5	3	4		
3	4	1		
5	5	4		
3	5	2		
3	5	3		
3	4	4		
3	5	4		
4	5			
2				

Phần 4. Phân tích phương sai



• Kiểm định Levene's Test

- Kiểm định Levene (Levene's test) được sử dụng để kiểm tra xem **k nhóm** có phương sai bằng nhau hay không? [\[Tham khảo\]](#)
- Nếu phương sai của các mẫu là bằng nhau sẽ được gọi là **đồng nhất phương sai**.
- Một số kiểm thống kê, ví dụ như phân tích phương sai ANOVA, **giả định rằng phương sai là bằng nhau giữa** các nhóm hoặc mẫu. Kiểm định Levene giúp **xác nhận** giả định này.
- **Phát biểu bài toán:**
 - Giả thuyết: $H_0 \sigma_0^2 = \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$
 - Đối thuyết: $H_0 \sigma_i^2 \neq \sigma_j^2$ (i và j là một cặp bất kỳ)
 - Nếu chấp nhận H_0 (Giả thuyết) thì ta có thể nói rằng phương sai các nhóm là bằng nhau → **Có thể kiểm định ANOVA**

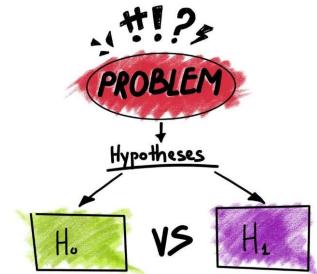
Phần 4. Phân tích phương sai



- Công thức kiểm định Levene

- Phát biểu bài toán:

- Giả thuyết: $H_0 \ \sigma_0^2 = \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$
 - Đối thuyết: $H_0 \ \sigma_i^2 \neq \sigma_j^2$ (i và j là một cặp bất kỳ)



- Dạng công thức kiểm định Levene

$$W = \frac{(N-k) \sum_{i=1}^k N_i (\bar{Z}_i - \bar{Z}_{..})^2}{(k-1) \sum_{i=1}^k \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_i)^2}$$

- N : là số lượng phần tử tổng thể, N_i : Số lượng phần tử nhóm thứ I
 - k là số lượng phần tử;
 - **Ta có** $Z_{ij} = |Y_{ij} - \bar{Y}_i|$ và ta có thêm

$$\bar{Z}_{i.} = \frac{\sum_{j=1}^{N_i} Z_{ij}}{N_i}$$

$$\bar{Z}_{..} = \frac{\sum_{i=1}^k \sum_{j=1}^{N_i} Z_{ij}}{N}$$

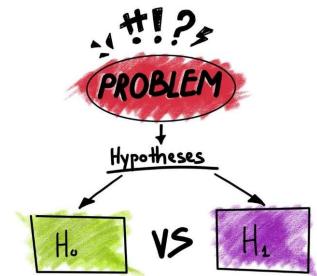
Phần 4. Phân tích phương sai



- Công thức kiểm định Levene

- Phát biểu bài toán:

- Giả thuyết: $H_0 \ \sigma_0^2 = \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$
 - Đối thuyết: $H_0 \ \sigma_i^2 \neq \sigma_j^2$ (i và j là một cặp bất kỳ)



- Dạng công thức kiểm định Levene

$$W = \frac{(N-k) \sum_{i=1}^k N_i (\bar{Z}_i - \bar{Z}_{..})^2}{(k-1) \sum_{i=1}^k \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_i)^2}$$

- Nếu $W > F(1 - \alpha; k - 1; N - k)$ thì bác bỏ giả thuyết H_0
 - Giá trị p-value của kiểm định Levene

$$\text{P-Value} = 1 - \text{CDF}(W; k - 1; N - k)$$

Phần 4. Phân tích phương sai



- **F-Critical của Levene**

- Trong ngôn ngữ lập trình Python **kiểm định Levene**

```
from scipy.stats import f

alpha = 0.05
k = 3 # số nhóm
n = 20 # tổng số quan sát

# Tính giá trị critical
critical_value = f.ppf(q=1-alpha, dfn=k-1, dfd=n-k)

print("Giá trị critical:", critical_value)
```

Phần 4. Phân tích phương sai



- P-Value của kiểm định Levene
 - Trong ngôn ngữ lập trình Python **kiểm định Levene**

```
▶ from scipy.stats import f

# Thay thế giá trị W, N, K bằng giá trị thực tế của bạn
W_value = 0.9433
N = 24
K = 3

# Tính p-value từ W, N, K
df_between = K - 1
df_within = N - K

# Giả thuyết không có sự phân kỳ (H0)
null_hypothesis = f.cdf(W_value, df_between, df_within)

# Giả thuyết có sự phân kỳ (H1)
alternative_hypothesis = 1 - null_hypothesis

print(f'p-value: {alternative_hypothesis}')

p-value: 0.40522773484156815
```

Phần 4. Phân tích phương sai



- **Kiểm định Levene nhanh bằng hàm levene**
 - Thực hành kiểm định Levene trong Python

```
In [3]: sep=ex02.groupby('Education')['satisfaction'].apply(list)
print(sep)
```

```
Education
College graduate    [5, 3, 5, 3, 3, 3, 3, 4, 2]
Graduate degree      [3, 4, 5, 5, 5, 4, 5, 5]
Some college         [4, 1, 4, 2, 3, 4, 4]
Name: Satisfaction, dtype: object
```

```
In [4]: from scipy.stats import levene
stat, p = levene(*sep,center='mean')
print(stat,p)
```

```
0.9433580072525427 0.40520616699352924
```

Phần 4. Phân tích phương sai



• Kiểm định ANOVA một phía

- Là một kỹ thuật thống kê tham số được sử dụng để **phân tích sự khác nhau giữa giá trị trung bình** của các biến phụ thuộc với nhau (Ronald Fisher, 1918).
- Câu hỏi: **Trung bình các nhóm là bằng nhau hay không bằng nhau?**
- Các loại kiểm định ANOVA: Oneway, **Twoway, MANOVA**
- Phát biểu bài toán:
 - Giả thuyết: $H_0 \mu_0 = \mu_1 = \mu_2 = \dots = \mu_k$
 - Đối thuyết: $H_0 \mu_i \neq \mu_j$ (i và j là một cặp bất kỳ)
- Nếu từ chối H_0 (Giả thuyết) thì ta có thể nói rằng ít nhất trung bình 2 nhóm là khác nhau → **Có thể kiểm định ANOVA sâu (Turkey)**

Phần 4. Phân tích phương sai



- Các bước tính kiểm định ANOVA

- Cho dataset như hình minh họa gồm 3 nhóm $N1, N2, N3$
- **Bước 1: Tính trung bình mỗi nhóm**

$$\bar{N}_i = \frac{\sum_{j=1}^{N_i} x_{ij}}{N_i}$$

- **Bước 2: Tính trung bình tổng thể**

$$\bar{N} = \frac{\sum_{i=1}^k \sum_{j=1}^{N_i} x_{ij}}{N}$$

- **Bước 3: Tính biến thiên nội bộ nhóm i**

$$SS_i = \sum_{j=1}^{N_i} (x_{ij} - \bar{N}_i)^2$$

$N1$	$N2$	$N3$
x1	y1	z1
x2	y2	z2
x3	y3	z3
x4	y4	z4
x6	y5	z5
\bar{N}_1	\bar{N}_2	\bar{N}_3

Phần 4. Phân tích phương sai



- Các bước tính kiểm định ANOVA

- Cho dataset như hình minh họa gồm 3 nhóm $N1, N2, N3$
- **Bước 4: Tính biến thiên trong nội bộ các nhóm (Within groups sum of square)**

$$SSW = SS_1 + SS_2 + \dots + SS_k$$

Là những yếu tố do phân tích gây ra

- **Bước 5: Tổng bình phương độ lệch của Các nhóm SSG (Between groups sum of square)**

$$SSG = \sum_{i=1}^K (n_i)(\bar{N}_i - \bar{N})$$

$N1$	$N2$	$N3$
x1	y1	z1
x2	y2	z2
x3	y3	z3
x4	y4	z4
x6	y5	z5
\bar{N}_1	\bar{N}_2	\bar{N}_3

- **Bước 6: Tổng Square tổng thể (Total Groups sum of square)**

$$SST = SSW + SSG$$

Phần 4. Phân tích phương sai



- Các bước tính kiểm định ANOVA

- Bước 6: Tổng Square tổng thể (Total Groups sum of square)

$$SST = SSW + SSG$$

- Biến thiên SSW (được gọi là biến thiên các yếu tố khác), biến thiên SSG là biến thiên giữa các nhóm trong kiểm định

Nhận xét: Nếu $SSG > SSW$ thì phần biến thiên các yếu tố giữa các nhóm nhiều hơn bẩn than chính nó thì tăng khả năng bác bỏ H_0 trong kiểm định ANOVA

- Phương sai do các yếu tố khác tạo ra $MSW = \frac{SSW}{N-k}$
- Phương sai do các nhóm tạo ra $MSG = \frac{SSG}{k-1}$
- Kiểm định ANOVA tính F

$$F = \frac{MSG}{MSW}$$

- Nếu $F > F(\alpha; k - 1; N - k)$ thì bác bỏ giả thuyết H_0

Phần 4. Phân tích phương sai



- **Summary kiểm định ANOVA**
 - Bảng tóm tắt kiểm định ANOVA

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F-statistic
Between Groups	SSB	$k - 1$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Groups	SSW	$N - k$	$MSW = \frac{SSW}{N-k}$	
Total	SST	$N - 1$		

Phần 4. Phân tích phương sai



• Kiểm định ANOVA sâu

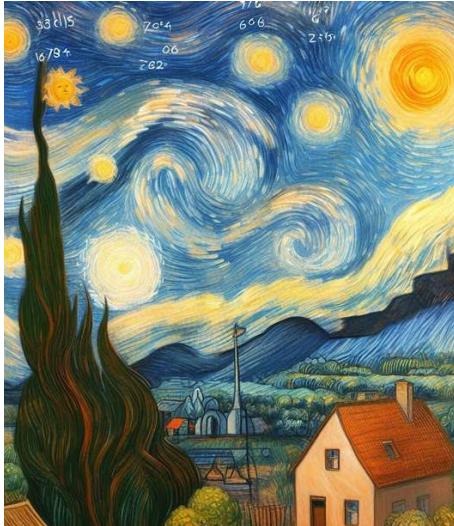
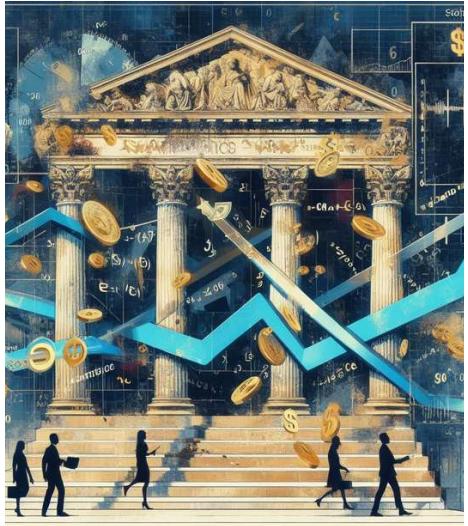
- Khi bác bỏ giả thuyết H0 của kiểm ANOVA ta có thể kiểm định Turkey Post Hoc Test để xem xét 2 nhóm nào bị ảnh hưởng bởi ANOVA sâu:
 - Cách giải quyết bài toán kiểm định ANOVA sâu (Trường hợp $k = 3$)
 - TH 1: $H_0 \mu_0 = \mu_1$ TH 2: $H_0 \mu_1 = \mu_2$ TH 3: $H_0 \mu_0 = \mu_2$
 $H_1 \mu_0 \neq \mu_1$ $H_1 \mu_1 \neq \mu_2$ $H_1 \mu_0 \neq \mu_2$
 - Tính khoảng biến thiên giữa hai nhóm: $D_{ij} = |\bar{Y}_i - \bar{Y}_j|$
 - Tính chỉ số Turkey
- $$T = q_\alpha(k; N - 1) \sqrt{\frac{MSW}{n_{min}}}$$
- Nếu giá trị: $D_{ij} > T$ (Bác bỏ giá trị H0)

Reference

**Python
Statistics**

Nguyen Minh Nhut
UIT – Information System

- [1] Peter Bruce (Author), Andrew Bruce (Author), Peter Gedeck (Author), Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python 2nd Edition
- [2] James R Evans, Business Analytics-Pearson (2017)
- [3] <http://thongke.cesti.gov.vn/dich-vu-thong-ke/tai-lieu-phan-tich-thong-ke/861-thong-ke-mo-ta-trong-nghien-cuu-dai-luong-tuong-quan#:~:text=C%E1%BA%A3%20hai%20thu%E1%BA%ADt%20ng%E1%BB%AF%20%C4%91%E1%BB%81u,tuy%E1%BA%BFn%20t%C3%ADnh%20gi%E1%BB%AFa%20hai%20bi%E1%BA%BFn.>
- [4] <https://bachkhoaluat.vn/cam-nang/9486/-gia-thuyet-va-gia-thiet-trong-nghien-cuu-khoa-hoc>



CẢM ƠN ĐÃ THEO DÕI



ftisu.vn



minhnhut.ftisu@gmail.com



0939013911