# BÀI HƯỚNG DẪN THỰC HÀNH 3. THỐNG KÊ SUY DIỄN VÀ ANOVA

## 1. Giới thiệu về kiểm định Kiểm Định T-Test 🌵 🌵 🌵 🌵

- **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 |

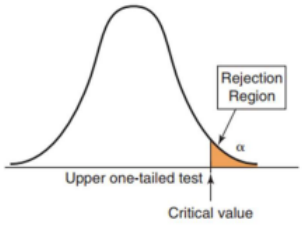
  - Có thể cho rằng cho rằng thời giản sửa chữa trung bình là **lớn hơn** 25 được hay không?
  - Đây là kiểm định **one-ways**
  - **Phát biểu bài toán:**
    - Giả thuyết: $H_0\ \mu_A \leq 25$
    - Đối thuyết: $H_1\ \mu_A > 25$

DAIT002. Thống kê Python Page 27 Sci Eng. Nguyen Minh Nhut

- Thực hành lại với dữ liệu CadSoft kiểm định T-Test

**Yêu cầu như sau:**

- **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




DAIT002. Thống kê Python Page 28 Sci Eng. Nguyen Minh Nhut

In [3]:
```
!pip install scipy
```

```
Defaulting to user installation because normal site-packages is not writeable
Collecting scipy
  Downloading scipy-1.12.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (38.4 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 38.4/38.4 MB 8.0 MB/s eta 0:00:0000:01m00:01
Requirement already satisfied: numpy<1.29.0,>=1.22.4 in /home/nhut/.local/lib/python3.10/site-packages (from scip
y) (1.26.4)
Installing collected packages: scipy
Successfully installed scipy-1.12.0
```

In [6]:
```
# 1.1 Hướng dẫn thực hành kiểm định t-test với bài thực hành trên
from scipy import stats
import numpy as np

# Dữ liệu
data = [
    [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]
]

# Tổng hợp dữ liệu thành một mảng 1D
flat_data = np.concatenate(data)

# Tính toán trung bình mẫu và độ lệch chuẩn
sample_mean = np.mean(flat_data)
sample_std = np.std(flat_data)

# Số lượng quan sát
n = len(flat_data)

# Độ tự do
degrees_of_freedom = n - 1

# Giá trị trung bình đề xuất
mu = 25

# Tính toán t-Statistic và p-value
t_statistic, p_value = stats.ttest_1samp(flat_data, mu)

# In kết quả
print("Trung bình mẫu:", sample_mean)
print("Độ lệch chuẩn mẫu:", sample_std)
print("Giá trị t-Statistic:", t_statistic)
print("Giá trị p-value:", p_value)

# Kiểm tra liệu trung bình có lớn hơn 25 hay không
if p_value >= 0.05:  # Chọn mức ý nghĩa là 0.05
    print("Có bằng chứng để bác bỏ giả thuyết H0: Trung bình không lớn hơn 25.")
else:
    print("Không đủ bằng chứng để bác bỏ giả thuyết H0: Trung bình có thể lớn hơn 25.")
```

```
Trung bình mẫu: 21.90909090909091
Độ lệch chuẩn mẫu: 19.2635055124476
Giá trị t-Statistic: -1.0521681183492575
Giá trị p-value: 0.2985994510452377
Có bằng chứng để bác bỏ giả thuyết H0: Trung bình không lớn hơn 25.
```

In [8]:
```
!pip install pandas
```

```
Defaulting to user installation because normal site-packages is not writeable
Collecting pandas
  Downloading pandas-2.2.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (13.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.0/13.0 MB 43.6 MB/s eta 0:00:0000:0100:01
Collecting tzdata>=2022.7
  Downloading tzdata-2024.1-py2.py3-none-any.whl (345 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 345.4/345.4 KB 54.9 MB/s eta 0:00:00
Requirement already satisfied: python-dateutil>=2.8.2 in /home/nhut/.local/lib/python3.10/site-packages (from pand
as) (2.8.2)
Requirement already satisfied: numpy<2,>=1.22.4 in /home/nhut/.local/lib/python3.10/site-packages (from pandas)
(1.26.4)
Collecting pytz>=2020.1
  Downloading pytz-2024.1-py2.py3-none-any.whl (505 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 505.5/505.5 KB 58.4 MB/s eta 0:00:00
Requirement already satisfied: six>=1.5 in /usr/lib/python3/dist-packages (from python-dateutil>=2.8.2->pandas)
(1.16.0)
Installing collected packages: pytz, tzdata, pandas
Successfully installed pandas-2.2.1 pytz-2024.1 tzdata-2024.1
```

In [10]:
```
#1.2 Hướng dẫn import csv dữ liệu CadSoft để kiểm định T-Test 1 Sample
import pandas as pd
from scipy import stats
import numpy as np

# Đọc dữ liệu từ file CSV
data = pd.read_csv('DAIT002_CadSoft_OneSample_LAB03.csv')

# Lấy dữ liệu từ cột 'sample'
samples = data['Time']

# Tính toán trung bình mẫu và độ lệch chuẩn
sample_mean = np.mean(samples)
sample_std = np.std(samples)

# Số lượng quan sát
n = len(samples)

# Độ tự do
degrees_of_freedom = n - 1

# Giá trị trung bình đề xuất
mu = 25

# Tính toán t-Statistic và p-value
t_statistic, p_value = stats.ttest_1samp(samples, mu)

# In kết quả
print("Trung bình mẫu:", sample_mean)
print("Độ lệch chuẩn mẫu:", sample_std)
print("Giá trị t-Statistic:", t_statistic)
print("Giá trị p-value:", p_value)

# Kiểm tra liệu trung bình có lớn hơn 25 hay không
if p_value >= 0.05:  # Chọn mức ý nghĩa là 0.05
    print("Có bằng chứng để bác bỏ giả thuyết H0: Trung bình không lớn hơn 25.")
else:
    print("Không đủ bằng chứng để bác bỏ giả thuyết H0: Trung bình có thể lớn hơn 25.")
```

```
Trung bình mẫu: 21.90909090909091
Độ lệch chuẩn mẫu: 19.2635055124476
Giá trị t-Statistic: -1.0521681183492575
Giá trị p-value: 0.2985994510452377
Có bằng chứng để bác bỏ giả thuyết H0: Trung bình không lớn hơn 25.
```

In [ ]:
```
# Học viên thực hành thêm F-Critical
```

## 2. Thực hành kiểm định T-Test 2 Sample 🌴🌴

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

DAIT002. Thống kê Python Page 34 Sci Eng. Nguyen Minh Nhut

- Thực hành dữ liệu đăng ký bậc tiểu học của Việt Nam và Thái Lan từ 2013 đến 2020

In [11]:
```
# 2.1 Lập trình kiểm định T-test 2 Sample
from scipy import stats

# Dữ liệu
vietnam_data = [102.1, 105.5, 106.8, 110.5, 109.9, 112.2, 117.4, 119.0, 120.0,122.1]
thailand_data = [102.7, 107.7, 105.4, 106.2, 101.1, 99.2, 99.3, 99.4, 99.5,101.6]

# Kiểm định T-Test 2 mẫu
t_statistic, p_value = stats.ttest_ind(vietnam_data, thailand_data)

# In kết quả
print("Giá trị t-Statistic:", t_statistic)
print("Giá trị p-value:", p_value)

# Kiểm tra liệu có bằng chứng để bác bỏ giả thuyết H0 hay không
alpha = 0.05  # Mức ý nghĩa
if p_value < alpha:
    print("Có bằng chứng để bác bỏ giả thuyết H0: Hai tập dữ liệu có trung bình khác nhau.")
else:
    print("Không đủ bằng chứng để bác bỏ giả thuyết H0: Hai tập dữ liệu có thể có trung bình bằng nhau.")
```

```
Giá trị t-Statistic: 4.360829530228135
Giá trị p-value: 0.000376743792536237
Có bằng chứng để bác bỏ giả thuyết H0: Hai tập dữ liệu có trung bình khác nhau.
```

In [12]:
```
#Học viên thực hành tính thêm F-Critical
```

## 💪Thực hành: Kiểm định T-Test 2 sample bằng cách tạo file CSV và import dữ liệu vietnam_data và thailand_data bằng CSV💪

Học viên tự tạo data csv sau đó import vào bằng thư viện pandas

In [ ]:
```
# Code thực hành tại đây
```

## 3. Thực hành Kiểm Định Chi-Square 🌊 🌊

In [13]:
```
#3.1 Thực hành với dataset Energy Survey
data_energy = pd.read_csv('DAIT002_EnergySurvey_LAB03.csv')
data_energy
```

Out[13]:

| | Respondent | Gender | Brand |
|---|---|---|---|
| **0** | 1 | Male | Brand 3 |
| **1** | 2 | Female | Brand 3 |
| **2** | 3 | Male | Brand 3 |
| **3** | 4 | Male | Brand 1 |
| **4** | 5 | Male | Brand 1 |
| **...** | ... | ... | ... |
| **95** | 96 | Male | Brand 1 |
| **96** | 97 | Male | Brand 3 |
| **97** | 98 | Female | Brand 3 |
| **98** | 99 | Male | Brand 2 |
| **99** | 100 | Female | Brand 1 |

100 rows × 3 columns

In [14]:
```
#3.2 Chi-Square Table
chiqtable_data_energy = pd.crosstab(data_energy.Gender,data_energy.Brand)
chiqtable_data_energy
```

Out[14]:

| Brand | Brand 1 | Brand 2 | Brand 3 |
|---|---|---|---|
| **Gender** | | | |
| **Female** | 9 | 6 | 22 |
| **Male** | 25 | 17 | 21 |

In [21]:
```
#3.3 In ra kết quả ChiSquare
c, p, dof, expected = stats.chi2_contingency(chiqtable_data_energy)

print("Grand Total:", c)
print("p-value:", p)
```

```
Grand Total: 6.4924250792329055
p-value: 0.038921342064441915
```

In [22]:
```
#3.4 Sinh viên thực hành tính F-Criticald để so sánh Grand Total và so sánh p-value với 0.05
```

In [23]:
```
#3.5 Sinh viên đưa ra kết luận
```

## 4. Thực hành phân tích ANOVA (ANOVA Analysis) NG NG NG

**Data thực hành: DAIT002_Insurance_Survey_LAB03.csv**

### 4.1 Kiểm định Levene dữ liệu Insurance Survey CotEducation vaà Satisfaction

In [24]:
```
#4.1 Import dự liệu Insurance Survey
data_insurance_survey = pd.read_csv('DAIT002_Insurance_Survey_LAB03.csv')
data_insurance_survey
```

Out[24]:

| | Education | Marital Status | Years Employed | Satisfaction |
|---|---|---|---|---|
| 0 | Some college | Divorced | 4 | 4 |
| 1 | Some college | Divorced | 2 | 1 |
| 2 | Graduate degree | Widowed | 26 | 3 |
| 3 | Some college | Married | 9 | 4 |
| 4 | Graduate degree | Married | 6 | 4 |
| 5 | Graduate degree | Married | 10 | 5 |
| 6 | College graduate | Married | 4 | 5 |
| 7 | College graduate | Divorced | 9 | 3 |
| 8 | Graduate degree | Married | 6 | 5 |
| 9 | Graduate degree | Married | 1 | 5 |
| 10 | College graduate | Married | 4 | 5 |
| 11 | College graduate | Married | 2 | 3 |
| 12 | Some college | Married | 3 | 2 |
| 13 | Some college | Married | 2 | 3 |
| 14 | Graduate degree | Married | 4 | 4 |
| 15 | College graduate | Married | 5 | 3 |
| 16 | College graduate | Married | 15 | 3 |
| 17 | College graduate | Married | 12 | 3 |
| 18 | Graduate degree | Single | 10 | 5 |
| 19 | Some college | Married | 3 | 4 |
| 20 | Some college | Divorced | 15 | 4 |
| 21 | Graduate degree | Married | 2 | 5 |
| 22 | College graduate | Divorced | 20 | 4 |
| 23 | College graduate | Married | 18 | 2 |

In [26]:
```
#4.2 Group by Statisfaction theo nhóm Education
data_group_by = data_insurance_survey.groupby('Education')['Satisfaction'].apply(list)
data_group_by
```

Out[26]:
```
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 [29]:
```
#4.3 Kiểm định Levene
from scipy.stats import levene
stat, p = levene(*data_group_by,center = 'mean')

print('p-value:',p)
```

```
p-value: 0.40520616699352924
```

In [ ]:
```
#4.4 Học viên Nhận xét kiểm định Levene
```

In [ ]:
```
#4.5 Học viên tính F-Critical và so sánh với STAT nhận xét đối chiếu kết quả p-value
```

### 4.2 Kiểm định ANOVA dữ liệu Insurance Survey

In [32]:
```
#4.6 Kiểm định ANOVA
from scipy import stats
fvalue, pvalue= stats.f_oneway(*data_group_by)
print(fvalue,pvalue)
```

```
3.9246517319277117 0.03563539756488997
```

In [ ]:
```
#4.7 Học viên nhận xét kiểm định ANOVA với p-value
```

In [ ]:
```
#4.8 Học viên tính F-Critical và so sánh với STAT nhận xét đối chiếu kết quả p-value
```

### 4.3 Kiểm định Turkey (ANOVA sâu) dữ liệu Insurance Survey

In [35]:
```
!pip install statsmodels
```

```
Defaulting to user installation because normal site-packages is not writeable
Collecting statsmodels
  Downloading statsmodels-0.14.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (10.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.8/10.8 MB 36.2 MB/s eta 0:00:0000:0100:01
Collecting patsy>=0.5.4
  Downloading patsy-0.5.6-py2.py3-none-any.whl (233 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 233.9/233.9 KB 34.1 MB/s eta 0:00:00
Requirement already satisfied: pandas!=2.1.0,>=1.0 in /home/nhut/.local/lib/python3.10/site-packages (from statsmo
dels) (2.2.1)
Requirement already satisfied: scipy!=1.9.2,>=1.4 in /home/nhut/.local/lib/python3.10/site-packages (from statsmod
els) (1.12.0)
Requirement already satisfied: packaging>=21.3 in /home/nhut/.local/lib/python3.10/site-packages (from statsmodel
s) (23.2)
Requirement already satisfied: numpy<2,>=1.18 in /home/nhut/.local/lib/python3.10/site-packages (from statsmodels)
(1.26.4)
Requirement already satisfied: pytz>=2020.1 in /home/nhut/.local/lib/python3.10/site-packages (from pandas!=2.1.0,
>=1.0->statsmodels) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in /home/nhut/.local/lib/python3.10/site-packages (from pandas!=2.1.
0,>=1.0->statsmodels) (2024.1)
Requirement already satisfied: python-dateutil>=2.8.2 in /home/nhut/.local/lib/python3.10/site-packages (from pand
as!=2.1.0,>=1.0->statsmodels) (2.8.2)
Requirement already satisfied: six in /usr/lib/python3/dist-packages (from patsy>=0.5.4->statsmodels) (1.16.0)
Installing collected packages: patsy, statsmodels
Successfully installed patsy-0.5.6 statsmodels-0.14.1
```

In [38]:
```
#4.9 Kiểm định ANOVA sâu nếu trung bình là khác nhau
from statsmodels.stats.multicomp import pairwise_tukeyhsd
tukey = pairwise_tukeyhsd(endog=data_insurance_survey['Satisfaction'],groups=data_insurance_survey['Education'],alp
print(tukey)
```

```
          Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================================
     group1           group2     meandiff p-adj   lower   upper  reject
---------------------------------------------------------------------
College graduate Graduate degree   1.0556 0.1003 -0.1715  2.2826  False
College graduate    Some college  -0.3016 0.8231 -1.5742  0.9711  False
 Graduate degree    Some college  -1.3571 0.0409 -2.6641 -0.0502   True
---------------------------------------------------------------------
```

In [ ]:
```
#4.10 Nhận xét kiểm định ANOVA sâu
```

In [39]:
```
#4.11 Học viên trình bày cách tính p-adj của từng nhóm và đưa ra cách code
```

In [ ]:
```
#4.12 T-alpha so sánh với MeanDiff như thế nào? T-alpha tính như thế nào? Hãy trình bày Code tính
```

## 5. Bài tập 🌻🌻🌻

### 5.1 Kiểm định T-test Oneway với dữ liệu sau

- Dataset: DAIT002_SalesCaring_LAB03
- Yêu cầu: Thực hành T-test oneway, hãy cho biết trung bình doanh thu của hãng xe có lớn 55 được hay không
- Dùng ngôn ngữ lập trình Python để chứng minh

### 5.2 Kiểm định T-test Oneway với dữ liệu sau

- Dataset: DAIT002_RenewableConsump_LAB03
- Yêu cầu: Thực hành T-test 2 way, hãy cho biết trung bình về lượng tiêu thụ năng lượng tái chế của Viet Nam và Singapore là bằng nhau được không?
- Dùng ngôn ngữ lập trình Python để chứng minh

### 5.4 Phân tích phương sai dữ liệu sau đây:

- Dataset: DAIT002_Freshman_Data_LAB03
- Yêu cầu: Thực hành phân tích ANOVA cột dữ liệu College và HS GPA%
- Nếu phương sai không bằng nhau vẫn hãy kiểm định ANOVA với bài tập này
- Dùng ngôn ngữ lập trình Python để chứng minh

In [ ]: