```
import numpy as np
import pandas as pd
from scipy import stats
```

Bài 1. (4)

a) Đọc dữ liệu và cho biết có bao nhiêu thí sinh trong bảng dữ liệu. Tính điểm trung bình, độ lệch chuẩn, phương sai của điểm môn Ngữ Văn ("ngu_van") (0.5đ)

b) Chọn ngẫu nhiên 25 mẫu, sử dụng 25 mẫu này để ước lượng điểm trung bình của môn Ngữ Văn (sử dụng độ lệch chuẩn tính được bên trên làm độ lệch chuẩn quần thể) với độ tin cậy là 95%. (3đ)

c) Từ kết quả ước lượng bên trên hãy cho nhận xét về giá trị trung bình ước lượng và giá trị thật tính từ dữ liệu. (0.5đ)

```
# 1 (0.5đ)
data = pd.read_csv('dataKHXH.csv')
print(f"Có {data.shape[0]} thí sinh trong bảng dữ liệu.")
mean = data['ngu_van'].mean()
print(f"Điểm trung bình của thí sinh là {mean:.2f}.")
std = data['ngu_van'].std()
var = data['ngu_van'].var()
print(f"Độ lệch chuẩn của điểm là {std:.2f}.")
print(f"Phương sai của điểm là {var:.2f}.")
```

```
Có 565243 thí sinh trong bảng dữ liệu.
Điểm trung bình của thí sinh là 7.02.
Độ lệch chuẩn của điểm là 1.32.
Phương sai của điểm là 1.74.
```

```
# 2 sample 500 thí sinh từ bảng dữ liệu (0.5đ)
n = 25
sample = data.sample(n, random_state=15)
sample_mean = sample['ngu_van'].mean()
print(f"Điểm trung bình của 25 thí sinh là {sample_mean:.2f}.")
```

```
Điểm trung bình của 25 thí sinh là 7.00.
```

Gọi μ là điểm trung bình môn Ngữ Văn của tập dữ liệu. Ta có, $\acute{x}=7.00$ là điểm trung bình môn Ngữ Văn của 25 thí sinh.

Độ tin cậy $95\%=1-\alpha$ => $\alpha=0.05$

Ta có $n=25<30$. => ta sấp sỉ phân phối trên với phân phối **student t** với bậc tự do là n-1, ta có:

$$\left(\acute{x}-t_{n-1;1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}},\acute{x}+t_{n-1;1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right)$$

Trong đó, $\alpha=0.05$, $z_{1-\frac{\alpha}{2}}=1.96$, $\acute{x}=7.44$, $s=0.94$

```
# 1đ
alpha = 0.05
t = stats.t.ppf(1-alpha/2,n-1)
print(f"z-score của phân phối chuẩn với độ tin cậy {1-alpha:.2f} là
{t:.2f}.")
print(sample_mean-t*std/np.sqrt(n), sample_mean+t*std/np.sqrt(n))
```

```
z-score của phân phối chuẩn với độ tin cậy 0.95 là 2.06.
6.4548135223805305 7.5451864776194695
```

(0.5đ) Vậy khoảng giá trị trung bình thật là: (6.45:7.55) với khoảng tin cậy 95%

c) Ta nhận thấy giá trị trung bình của tập dữ liệu nằm trong khoảng ước lượng, điều này cho thấy khoảng ước lượng đáng tin cậy. Tuy nhiên tại số lượng mẫu được lấy ra ít nên sai số lớn

Bài 2

Lấy ngẫu nhiên 1000 mẫu dữ liệu từ dữ liệu gốc, có nhận xét cho rằng: "Có từ 75% thí sinh khối KHXH có điểm Toán trên trung bình". Bạn có thể đưa ra kết luận gì nhận xét trên dựa vào 1000 mẫu vừa lấy được với mức ý nghĩa 5%? (4đ)

```
n = 1000
sample = data.sample(n, random_state=15)

# 0.25đ
gte5 = sample[sample['toan'] >= 5].shape[0]
print("Sô´thí sinh trên điểm trung bình: ",gte5)
p_mu = gte5/n
print("Tỉ lệ thí sinh trên trung bình: ",p_mu)
```

```
Sô´thí sinh trên điểm trung bình:  746
Tỉ lệ thí sinh trên trung bình:  0.746
```

Gọi p là thí sinh có điểm Toán trên trung bình. Ta cần kiểm định giả thuyết:

$H_0: p \geq 0.75$: Có từ 75% thí sinh khối KHXH có điểm Toán trên trung bình

$H_1: p < 0.75$: Tỉ lệ thí sinh khối KHXH có điểm Toán trên trung bình thấp hơn 75%

Mức ý nghĩa là: $5\% = ¿\alpha = 0.05$

```
p = 0.75
alpha = 5/100
```

- Điểm thi của các thí sinh trên trung bình hay không là độc lập với nhau nên tỉ lệ thí sinh đạt điểm trên trung bình tuân theo phân phối nhị thức.
- Ta có
  - $\hat{p} = 0.75$ là tỷ lệ thí sinh có điểm Toán trên trung bình.
  - $n = 1000 > 30$,

- $\hat{p}=0.75, n\hat{p}=1000\times 0.75=750\geq 5$
- $n(1-\hat{p})=1000(1-0.75)=250\geq 5$

=> Vậy phân phối nhị thức của tỷ lệ mẫu có thể xấp xỉ bằng phân phối chuẩn.

$$z=\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$$

```
# 0.25đ
z = (p_mu - p)/np.sqrt(p_mu*(1-p_mu)/n)
print("z =", z)

z = -0.29058535263133406
```

Vì H_1 là ¿ là kiểm định đuôi trái nên ta có

$$p_{values}=P(Z<z)$$

```
p_value = stats.norm.cdf(z)
print("p_value là: ", p_value)
if p_value < alpha:
    print('Reject H0')
else:
    print('Accept H0')

p_value là:  0.3856842319066385
Accept H0
```

Vì $p_{value}=0.386>\alpha=0.05$ nên ta kết luận: không có đủ bằng chứng để bác bỏ nhận xét trên với mức ý nghĩa 5%$

Bài 3 (2đ)

Khảo sát điểm trung bình các bài lab của một môn học X($) và điểm thi cuối kỳ Y($) trong một môn học. Khảo sát ngẫu nhiên 8 sinh viên, ta thu được bảng số liệu sau:

Điểm trung bình các lab	7.5	5.5	3.0	8	9	1	6	6
Điểm thi cuối kỳ	**8**	**6.5**	**4**	**7.5**	**9.5**	**3**	**5**	**6**

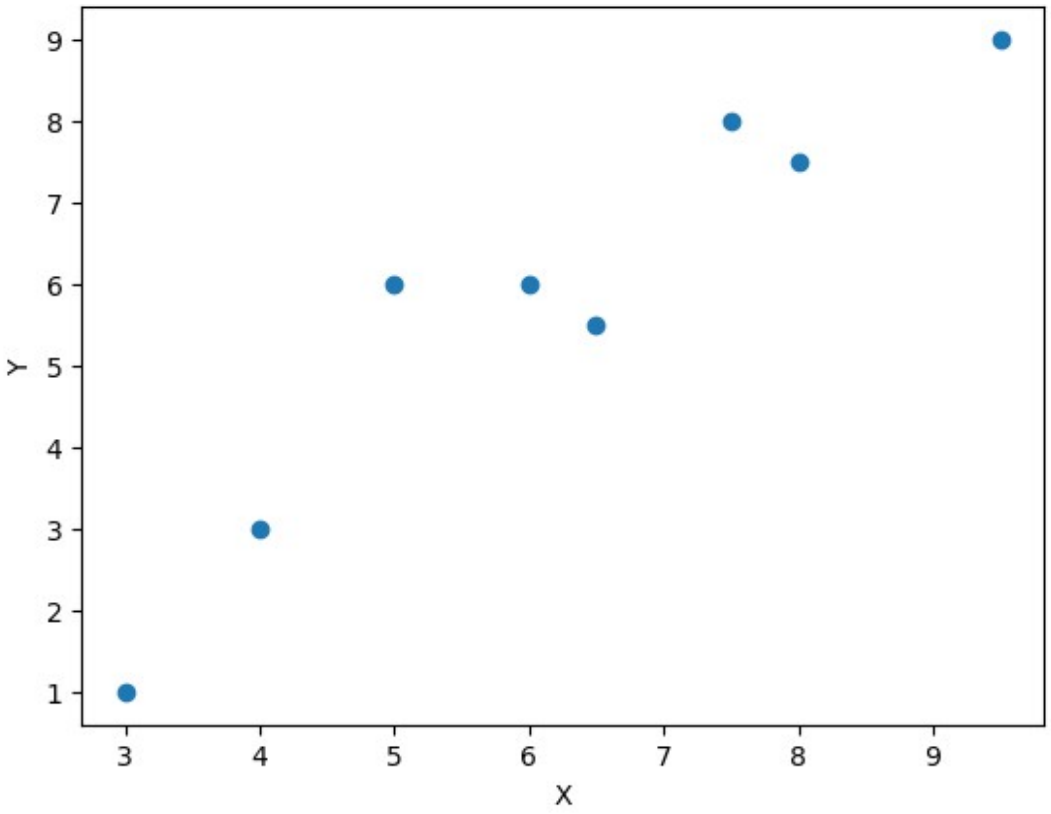
a. Dựa vào bảng dữ liệu trên cho biết có xây dựng được mô hình hồi quy hay không? Nếu có hãy xây dựng mô hình hồi quy để ước điểm trung bình các bài lab dựa theo điểm thi cuối kỳ. (1đ)

b. Dự đoán điểm trung bình các bài lab của một sinh viên có điểm thi cuối kỳ là 6.5. (1đ)

```
X = np.array([8.0, 6.5, 4.0, 7.5, 9.5, 3.0, 5.0, 6.0])
Y = np.array([7.5, 5.5, 3.0, 8.0, 9.0, 1.0, 6.0, 6.0])
```

Ta có biến phụ thuộc là điểm trung bình các bài lab và biến giải thích là điểm cuối kỳ.

```
# plot X,Y
import matplotlib.pyplot as plt
plt.scatter(X,Y)
plt.xlabel('X')
plt.ylabel('Y')
plt.show()
```




```
# covariance và correlation coefficient của X và Y
cov = np.cov(X,Y)
print("Covariance: \n", cov)
corr = np.corrcoef(X,Y)
print("Correlation coefficient: \n", corr)
```

```
Covariance:
 [[4.63839286 5.375     ]
 [5.375      7.        ]]
Correlation coefficient:
 [[1.         0.94329099]
 [0.94329099 1.        ]]
```

(0.5đ)

Ta thấy hệ số tương quan của X và Y là 0.93 nên mối quan thuận, tức là nếu X tăng thì Y cũng sẽ tăng.

Vậy có thể xây dựng mô hình hồi quy giữa X và Y:

$$Y = \beta_0 + \beta_1 X$$

```
x_bar = np.mean(X)
y_bar = np.mean(Y)
print("x_bar = ", x_bar)
print("y_bar = ", y_bar)
beta_1 = np.sum((X-x_bar)*(Y-y_bar))/np.sum((X-x_bar)**2)
beta_0 = y_bar - beta_1*x_bar
print("beta_1 = ", beta_1)
print("beta_0 = ", beta_0)
```

```
x_bar =  6.1875
y_bar =  5.75
beta_1 =  1.1588065447545717
beta_0 =  -1.4201154956689122
```

(0.5đ) Vậy mô hình hồi quy giữ điểm trung bình các bài lab Y và điểm thi cuối kỳ X là:

$$Y \approx -1,42 + 1,15\ X$$

b. Dự đoán điểm điểm thi cuối kỳ biết điểm trung bình các bài lab của một sinh viên là 6.2

(0.5đ) Điểm trung bình các bài lab 6.2 => $Y = 6.2$. Thay Y vào công thức hồi quy ta xây dựng bên trên.

```
Y1 = 6.15
X1 = (Y1-beta_0)/beta_1
print("X1 = ", X1)
```

```
X1 =  6.532682724252492
```

(0.5đ) Vậy điểm trung bình các lab của bạn sinh viên này khoảng 6.5đ