МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені Тараса Шевченка
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
Кафедра програмних систем і технологій

Дисципліна
«Ймовірнісні основи програмної інженерії»

Лабораторна робота №3
«ДВОВИМІРНА СТАТИСТИКА»

Виконав:	Булава Геннадій Юрійович	**Перевірила**:	Вечерковська Анастасія Сергіївна
Група	ІПЗ-21(2)	Дата перевірки	
Форма навчання	денна	Оцінка	
Спеціальність	121		

2022

Назва теми: Двовимірна статистика

Мета: навчитись використовувати на практиці набуті знання про міри в двовимірній статистиці.

Постановка задачі

1. Намалюйте діаграму розсіювання для даних. Укажіть, чи існує тренд у даних. Якщо так, то вкажіть, чи є це негативним трендом, чи позитивним.
2. Знайдіть центр ваги і коваріацію.
3. Знайти рівняння лініїї регресії у від x.
4. Розрахуйте коефіцієнт кореляції між даними.
5. Зробити висновок про залежності.

Математична модель

1. Для побудови діаграми розсіювання наносяться відповідні точки на діаграму.
   Знаходимо рівняння лінії тренду за формулами:

$$\bar{Y} = m\bar{X} + b, where$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \ (the\ average\ of\ x)$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i \ (the\ average\ of\ y)$$

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

$$b = \bar{Y} - m\bar{X}$$

   І залежно від коефіцієнта m робимо висновок: Якщо m > 0, то тренд позитивний, якщо m < 0, то тренд негативний.
2. Центр ваги визначається як точка G($\bar{x};\bar{y}$)
   Коваріація знаходиться за формулою:
   $$cov(x;y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_j - \bar{y})$$
3. Рівняння лінії регресії: $y = b_0 + b_1 x$, де
   $$b_1 = \frac{cov(x,y)}{var(x)}$$
   $$b_0 = \bar{y} - b_1\overline{x}$$
4. Коефіцієнт кореляції визначається за формулою:
   $$\rho_{xy} = \frac{cov(x;y)}{\sigma_x \sigma_y}$$

Випробування алгоритму

Вхідні файли: input_10.txt, input_100.txt

Результат для input_10.txt

Завдання 1

Тренд позитивний

Завдання 2

Центр ваги: G(3.892, 35.0)

Коваріація: 23.0

Завдання 3

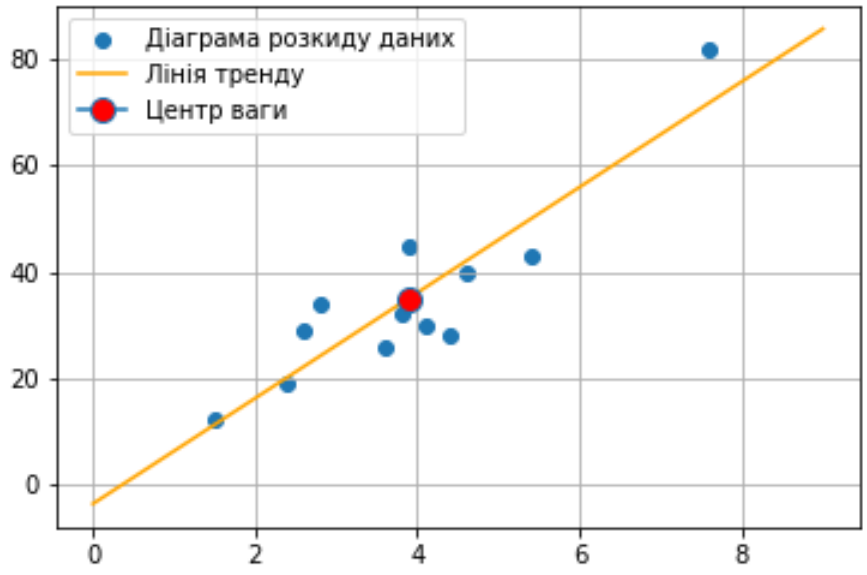
Рівняння лінії регресії: y = 9.953x + -3.735

Завдання 4

Коефіцієнт кореляції між даними: 0.901

Завдання 5

Коефіцієнт кореляції наближується до 1, значить дані майже співпадають з лнінією регресії.



Результат для input_100.txt

Завдання 1

Тренд позитивний

Завдання 2

Центр ваги: G(3.856, 34.5)

Коваріація: 22.592

Завдання 3

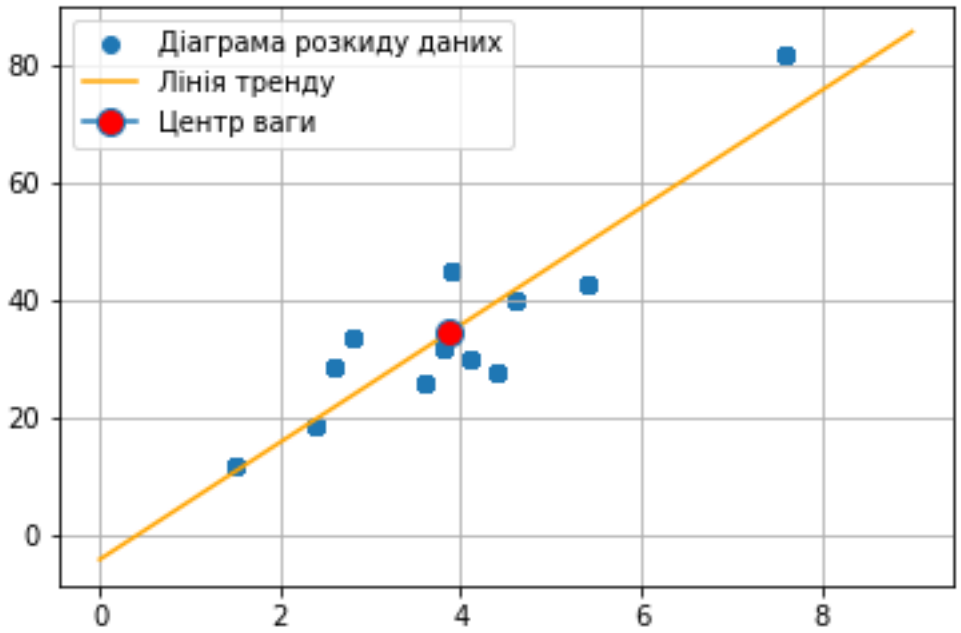
Рівняння лінії регресії: y = 9.982x + -3.991

Завдання 4

Коефіцієнт кореляції між даними: 0.902

Завдання 5

Коефіцієнт кореляції наближується до 1, значить дані майже співпадають з лнінією регресії.



Псевдокод

```
def inputting():
    global a, m
    print('Введіть назву файлу')
    name = input()
    file = open(name, mode="r")
    i = int(0)
    m = int(-1)
    with open(name, "r") as file:
        for line in file.readlines():
            if (i > 0):
                a[0].append(float(line.split('   ')[0].split(',')[0]+'.'+line.split('    ')[0].split(',')
[1]))
                a[1].append(int(line.split('    ')[1]))
            else:
                m = int(line)
            i = i + 1
    file.close()
    return 'output_'+str(m)+'.txt'

def task1():
    global xav, yav, b, b1, cov
    plt.scatter(a[0], a[1], label = 'Діаграма розкиду даних')
    with open(output_name, "w") as f:
        f.write('Завдання 1\n')
        b = float(0)
        sum = float(0)
        xav = float(0)
        yav = float(0)
        for i in range(0, m):
            xav = xav + a[0][i]
            yav = yav + a[1][i]
        xav = xav/m
        yav = yav/m
        for i in range(0, m):
            sum = sum + (a[0][i]-xav)*(a[1][i]-yav)
        sum2 = float(0)
        for i in range(0, m):
            sum2 = sum2 + (a[0][i] - xav)*(a[0][i] - xav)
        b = (sum)/sum2
        a1 = yav - b*xav
        if (b > 0):
            f.write('Тренд позитивний\n')
        elif (b < 0):
            f.write('Тренд негативний\n')
        else:
            f.write('Немає тренду\n')
        x = np.arange(0, 10)
        y = a1 + b * x
        plt.plot(x, y, color = 'orange', label = 'Лінія тренду')
        f.write('\n')

def task2():
    global xav, yav, b, b1, cov
    with open(output_name, "a") as f:
        f.write('Завдання 2\n')
        f.write('Центр ваги: G(' + str(round(xav, 3)) + ', ' + str(yav) + ')\n')
        plt.plot(xav, yav, marker="o", markersize=10, markerfacecolor="red", label='Центр ваги')
        f.write('Коваріація: ')
        cov = float(0)
        for i in range(0, m):
            cov = cov + ((a[0][i]-xav)*(a[1][i]-yav))
        cov = round(cov/m, 3)
        f.write(str(cov) + '\n')
        f.write('\n')
        varx = float(0)
        for i in range(0, m):
            varx = varx + (a[0][i]*a[0][i]-xav*xav)
        varx = varx/m
        b1 = cov/varx
        b = yav - b1*xav
        x = np.arange(0, m)
        y = b + b1 * x

def task3():
    global xav, yav, b, b1, cov
    with open(output_name, "a") as f:
        f.write('Завдання 3\n')
        f.write('Рівняння лінії регресії: y = ' + str(round(b1, 3)) + 'x + ' + str(round(b, 3)) + '\n')
        f.write('\n')

def task4():
    global xav, yav, b, b1, cov, r
    with open(output_name, "a") as f:
        f.write('Завдання 4\n')
        f.write('Коефіцієнт кореляції між даними: ')
        sx = float(0)
        sy = float(0)
        for i in range(0, m):
            sx = sx + (a[0][i]-xav)*(a[0][i]-xav)
        sx = sx/(m)
        sx = math.sqrt(sx)
        for i in range(0, m):
            sy = sy + (a[1][i]-yav)*(a[1][i]-yav)
        sy = sy/(m)
        sy = math.sqrt(sy)
        r = cov/(sx*sy)
        f.write(str(round(r, 3)))
        f.write('\n')
        f.write('\n')

def task5():
    global r
    with open(output_name, "a") as f:
        f.write('Завдання 5\n')
        if (r > 0.5 and r < 1):
            f.write('Коефіцієнт кореляції наближується до 1, значить дані майже співпадають  з лінією
регресії. \n')
        elif (r == 1):
            f.write('Коефіцієнт кореляції дорівнює 1, значить дані співпадають з лінією регресії. \n')
        elif (r == -0):
            f.write('Коефіцієнт кореляції дорівнює 0, значить дані незалежні лінійно. \n')
        elif (r == -1):
            f.write('Коефіцієнт кореляції дорівнює -1, значить дані лінійно залежні або існує сильний
лінійний зв`язок між даними. \n')
        elif (r > 0 and r <= 0.5):
            f.write('Коефіцієнт кореляції наближується до 0, значить зв`язок між даними слабкий. \n')
        elif (r > -1 and r <= -0.5):
            f.write('Коефіцієнт кореляції наближується до -1, значить існує сильний лінійний зв`язок між
даними. \n')
        elif (r > -0.5 and r < 0):
            f.write('Коефіцієнт кореляції наближується до 0, значить зв`язок між даними слабкий. \n')
```

Висновок

В ході виконання лабораторної роботи було опрацьовано тренди (в обох вхідних файлах тренд позитивний), знаходження центра ваги та коваріації. Було опрацьовано знаходження рівняння лінії регресії. Також було опрацьовано знаходження коефіцієнта кореляції між даними.