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:	Антонов Олександр Лаврентійович	**Перевірила**:	Марцафей Анна Сергіївна
Група	ІПЗ-24(2)	Дата перевірки	
Форма навчання	денна	Оцінка	
Спеціальність	121		

2022

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

Завдання 1

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 коваріацію.

3. Знайти рівняння лінії регресії y від x.

4. Розрахуйте коефіцієнт кореляції між даними.

Завдання 1

1. Намалюйте діаграму розсіювання для даних. Укажіть, чи існує тренд у даних. Якщо так, то вкажіть, чи є це негативним трендом, чи позитивним.

Діаграма розсіювання (точкова діаграма) — один з типів математичних діаграм, що використовує декартову систему координат для відображення значень двох змінних для набору даних. Дані показані у вигляді набору точок, кожен з яких має значення однієї змінної, тобто визначає її положення на горизонтальній осі та значення іншої змінної — її положення на вертикальній осі.

Тренд - тенденція зміни показників часового ряду. Тренди можуть бути описані різними функціями

Виведення даних з файлу, діаграма розсіювання

```
averagex, averagey = 0.0, 0.0
def connect_txt(nameoffile):
    inputdata = []
    input = open("input_103.txt")
    input.seek(1)
    for line in input:
        inputdata.append(input.read(3))
        input.read(1)
        inputdata.append(input.read(2))
    for i in range(int(len(inputdata))):
        inputdata[i] = inputdata[i].replace(',', '.')
    data = [[0 for i in range(2)] for j in range(int(len(inputdata) / 2))]
    index0 = 0
    index1 = 0
    for i in range(int(len(inputdata))):
        if i % 2 == 0:
            data[index0][0] = float(inputdata[i])
            index0 +=1
        elif i % 2 != 0:
            data[index1][1] = int(inputdata[i])
            index1 += 1
    return data

def dataX(data):
    inputdatadata = []
    for i in range(len(data)):
        inputdatadata.append(data[i][0])
    return inputdatadata
```

визначення тренду

```
def trend(data):
    if max(data) == data[len(data)-1]:
        print("Trend of data is positive")
        f.write("\nTrend of data is positive")
    elif min(data) == data[len(data)-1]:
        print("Trend of data is negative")
        f.write("\nTrend of data is negative")
    else:
        print("The data does not have any trend")
        f.write("\nThe data does not have any trend")
```

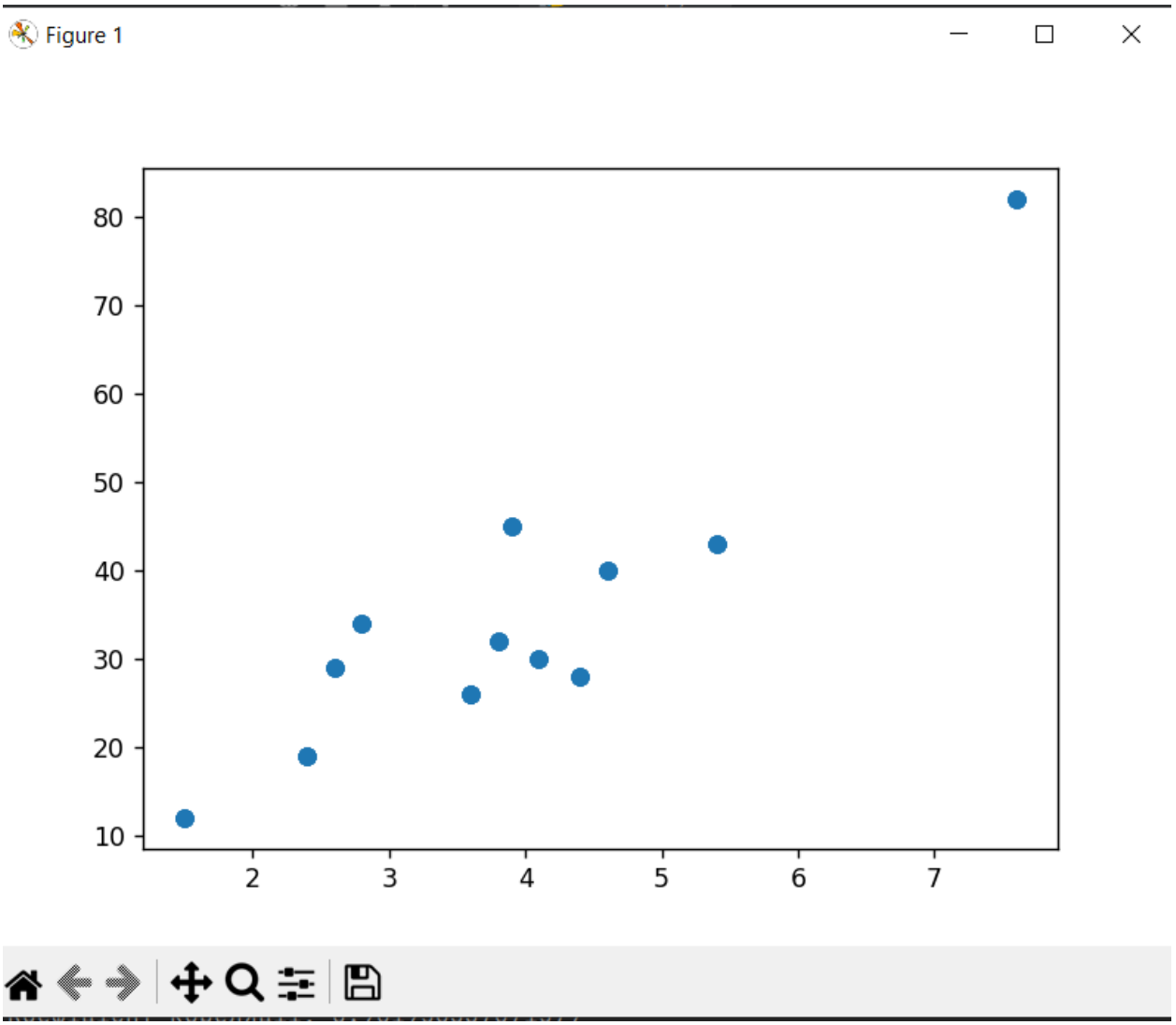
Код створення діаграми розсіювання

```
plt.scatter(infoX, infoY)
plt.show()
```

Результат виконання

```
Input the name of input file: input_10.txt
Sorted data:  [[1.5, 12], [2.4, 19], [2.6, 29], [2.8, 34], [3.6, 26], [3.8, 32], [3.9, 45], [4.1, 30], [4.4, 28], [4.6, 40], [5.4, 43], [7.6, 82]]


Trend of data is positive
```



2. Знайдіть коваріацію.

Коваріація — це міра спільної мінливості двох випадкових змінних. Якщо більші значення однієї змінної здебільшого відповідають більшим значенням іншої, й те саме виконується для менших значень, тобто змінні схильні демонструвати подібну поведінку, то коваріація є додатною. В протилежному випадку, коли більші значення однієї змінної здебільшого відповідають меншим значенням іншої, тобто змінні схильні демонструвати протилежну поведінку, коваріація є від'ємною. Отже, знак коваріації показує тенденцію в лінійному взаємозв'язку між цими змінними.

Коваріацію (рис. 8) було знайдено за допомогою наступної формули (рис. 6) та реалізовано відповідним кодом (рис. 7).

$$\mathrm{Cov}\,(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

Формула знаходження коваріації

Знаходження коваріації

```
def covariance(x, y):
    global averagex, averagey
    covarience = 0.0
    for i in range(len(x)):
        averagex += x[i]
        averagey += y[i]
    averagex = averagex / len(x)
    averagey = averagey / len(y)
    for i in range(len(x)):
        covarience += (x[i] - averagex) * (y[i] - averagey)
    covarience = covarience / (len(x)-1)
    print("Коваріацію: ", covarience)
    sus.write("Коваріацію: " + str(covarience))
```

Результат виконання

```
Коваріацію:  22.820202020202036
```

3. Знайти рівняння лінії регресії y від x.

Рівняння, що відображує зміну середньої величини однієї ознаки (y) в залежності від другої (x), називається рівнянням регресії або рівнянням кореляційного зв'язку.

Рівняння лінії регресії (рис. 12) було знайдено за допомогою наступних формул (рис. 9) (рис. 10) та реалізовано відповідним кодом (рис. 11).

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

Формула знаходження b для рівняння лінії регресії

$$y - \bar{y} = b(x - \bar{x})$$

Формула знаходження лінії регресії

Знаходження рівняння лінійної регресії

```
83  def lineofregression(X, Y):
84      global averagex, averagey
85      byx, sumx, sumy, sumxy, sumx2, sumy2 = 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
86      for i in range(len(X)):
87          sumx += X[i]
88          sumy += Y[i]
89          sumxy += X[i] * Y[i]
90          sumx2 += X[i] * X[i]
91          sumy2 += Y[i] * Y[i]
92      byx = (len(X) * sumxy - (sumx * sumy)) / (len(X) * sumx2 - sumx2)
93      x, y = sp.symbols("x,y")
94      line = sp.Eq(y-averagey, byx*(x-averagex))
95      linex = sp.solve(line, y)
96      liney = sp.solve(line, x)
97      strlinex = str(linex)
98      strliney = str(liney)
99      strlinex = strlinex.replace("[", "")
100     strlinex = strlinex.replace("]", "")
101     strliney = strliney.replace("[", "")
102     strliney = strliney.replace("]", "")
103     print("Лінія регресії y від x. ")
104     print("x = " + strliney)
105     print("y = " + strlinex, "\t(y від x)")
106     sus.write("Лінія регресії y від x.\n")
107     sus.write("x = " + strliney)
108     sus.write("\ny = " + strlinex + "\t(y від x )")
109
```

```
Лінія регресії y від x.
x = 0.750738314447591*y - 22.0444718484419
y = 1.33202206515305*x + 29.3637229167698   (y від x)
```

4. Розрахуйте коефіцієнт кореляції між даними.

Коефіцієнт кореляції Пірсона — в статистиці, показник кореляції (лінійної залежності) між двома змінними X та Y, який набуває значень від −1 до +1 включно. Він широко використовується в науці для вимірювання ступеня лінійної залежності між двома змінними.

наскільки залежні дві величини, що більше значення, то залежні значення. Визначає чи залежить траплення велечини

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Рисунок 13 | Формула розрахунку коефіцієнту кореляції між даними

```python
def correlation(x, y):
    global averagex, averagey
    corcoef, sum1, sum2, sum3 = 0.0, 0.0, 0.0, 0.0
    for i in range(len(x)):
        sum1 += (x[i] - averagex) * (y[i] - averagey)
        sum2 += (x[i] - averagex) * (x[i] - averagex)
        sum3 += (y[i] - averagey) * (y[i] - averagey)
    sum2 = sum2 * sum3
    corcoef = sum1/math.sqrt(sum2)
    print("Коефіцієнт кореляції:", corcoef)
    sus.write("\nКоефіцієнт кореляції:" + str(corcoef))
```

Функція розрахунку коефіцієнту кореляції між даними

```
Коефіцієнт кореляції: 0.901950337071579
```

Приклад виконання, input_10.txt

Висновок: Під час виконання третьої лабораторної роботи було повторено операції з вхідними даними записаних у txt файл, реалізовано потрібні формули для знаходження інформації за завданнями, також побудовано діаграму розсіювання.

Також було проведено дослідження, де було порівняно коваріацію з кореляцією.

Коваріація та кореляція в основному оцінюють зв’язок між змінними. Найближчою аналогією зв'язку між ними є зв'язок між дисперсією та стандартним відхиленням.

Коваріація вимірює загальну варіацію двох випадкових змінних від їхніх очікуваних значень. Використовуючи коваріацію, ми можемо лише оцінити напрямок зв’язку (чи змінні мають тенденцію рухатися в тандемі чи показують зворотний зв’язок). Однак це не вказує ні на силу зв’язку, ні на залежність між змінними.

З іншого боку, кореляція вимірює силу зв’язку між змінними. Кореляція – це масштабна міра коваріації. Він безрозмірний. Іншими словами, коефіцієнт кореляції завжди є чистим значенням і не вимірюється в жодних одиницях.

Зв'язок між двома поняттями можна виразити за допомогою формули нижче:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Рисунок 16 | Формула зображення зв'язку між коваріацією та кореляцією

Де:

- **ρ(X,Y)** – кореляція між змінними X і Y
- **Cov(X,Y)** – коваріація між змінними X і Y
- **σ_X** – стандартне відхилення X-змінної
- **σ_Y** – стандартне відхилення Y-змінної