МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені Тараса Шевченка
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра програмних систем і технологій

Дисципліна

«Ймовірнісні основи програмної інженерії»

Звіт з лабораторної роботи № 3

на тему:

«Двовимірна статистика»

Виконала:	Дрозд Єлизавета Андріївна	**Перевірила:**	Марцафей А. С.
Група	ІПЗ-12(2)	Дата перевірки	
Форма навчання	денна	Оцінка	
Спеціальність	121		

2022

Мета роботи:

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 лінії регресії y від x.
4. Розрахуйте коефіцієнт кореляції між даними.
5. Зробити висновок про залежності.

Побудова математичної моделі:

Формули знаходження центру ваги та коваріації:

2. Regression Lines

Consider a sample of bivariate data (x_1, y_1), (x_2, y_2), …, (x_n, y_n) of two related variables X and Y. Let $\bar{x}$ denote the mean of $x_1, x_2, \ldots, x_n$ and $\bar{y}$ denote the mean of $y_1, y_2, \ldots, y_n$.

- The point $G(\bar{x}, \bar{y})$ is called the center of gravity of the data.
- The covariance of the data is defined by: $\text{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.
- An alternative formula for the covariance is: $\text{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}$.

Формули знаходження рівняння лінії регресії:

Consider all lines with equations given by: $y = k + mx$. Let $d_i = y_i - (k + mx_i)$.

Set $D = \sum_{i=1}^{n} d_i^2$. Among all lines $y = k + mx$, consider the line that minimizes D. Such a line is called the *least-squares regression line* that best fits the data. Its coefficients are given by $m \equiv b_1 = \frac{\text{cov}(X,Y)}{\text{Var}(X)}$ and $k \equiv b_0 = \bar{y} - b_1\,\bar{x}$.

Note that, the regression line, passes through the center of gravity G of the data as $y = \bar{y} - b_1\bar{x} + b_1 x$ or $y - \bar{y} = b_1(x - \bar{x})$.

Формули знаходження коефіцієнта кореляції:

Consider a sample of size n, (x_i, y_i), $i = 1, 2, 3, \ldots, n$, for measured values of two related variables X and Y. Let $\bar{x}$, $\bar{y}$, s_x, and s_y denote their means and their standard deviations. The Pearson's sample correlation coefficient r of X and Y is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Note that the values $\frac{x_i - \bar{x}}{s_x}$ and $\frac{y_i - \bar{y}}{s_y}$ are the z-scores for x_i and y_i, respectively.

Moreover, the above definition can be expressed in any one of the following two forms:

$$r = \frac{1}{n-1}\sum_{i=1}^{n} z_x z_y \quad \text{or} \quad r = \frac{\operatorname{cov}(X,Y)}{\sigma_x \sigma_y}$$

We may also verify that the slope of the regression line of y on x is given by:

$$b_1 = r\frac{\sigma_y}{\sigma_x} = r\frac{s_y}{s_x}$$

Псевдокод алгоритму:

```
import matplotlib.pyplot as plt
import sympy as sp
from math import *

s = input('Enter a file name: ')
f1 = open(s, 'r')
f2 = open('output.txt', 'w')
items = f1.read().split()
items.remove(items[0])
time = []
cost = []
i = 0
j = 1
k = 0
while i < len(items):
    time.append(items[i])
    cost.append(items[j])
    time[k] = time[k].replace(",", ".")
    i += 2
    j += 2
    k += 1
time = [float(i) for i in time]
cost = [int(i) for i in cost]
```

```
def task():
    x_mean = sum(time) / len(time)
    y_mean = sum(cost) / len(cost)
    index = 0
    covariance = 0
    variance_x = 0
    variance_y = 0
    while index < len(time):
        covariance += (time[index] - x_mean) * (cost[index] - y_mean)
        variance_x += pow(time[index] - x_mean, 2)
        variance_y += pow(cost[index] - y_mean, 2)
        index += 1
    covariance /= len(time)
    variance_x /= len(time)
    variance_y /= len(cost)
    standart_deviation_x = sqrt(variance_x)
    standart_deviation_y = sqrt(variance_y)
    cor_coef = covariance/(standart_deviation_x*standart_deviation_y)
    m = covariance / variance_x
    x = sp.Symbol('x')
    s = sp.expand(y_mean - m * x_mean + m * x)
    f2.write("Center of gravity: (%s, %s)\nCovariance = %s" % (x_mean, y_mean, covariance))
    f2.write("\n-------------------------------------------------------------------------------\n")
    f2.write("Regression line equation: y = %s" % s)
    f2.write("\n-------------------------------------------------------------------------------\n")
    f2.write("Correlation coefficient = %s" % cor_coef)
```

```
task()
plt.scatter(time, cost)
plt.grid(True)
plt.xlabel('time, in minutes')
plt.ylabel('amount, in dollars')
plt.savefig('scatter.png')
f1.close()
f2.close()
```

Випробування алгоритму:

Результат роботи програми при введенні даних із файлу input_10.txt:

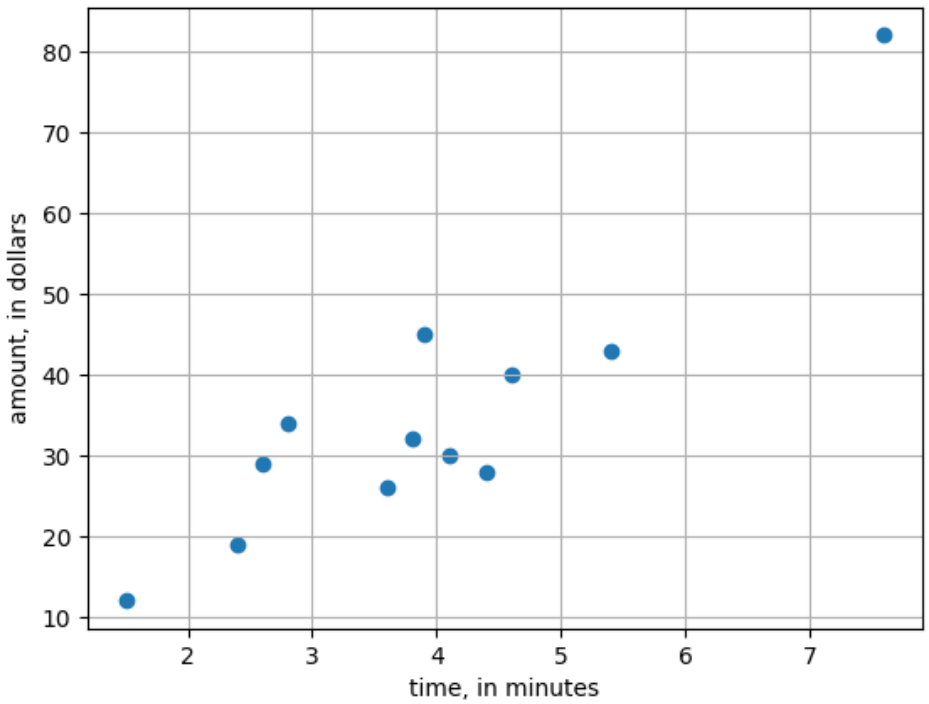
```
"C:\Users\admin\Desktop\2 курс\2 курс 1 семестр\ЙОПІ\LAB3\venv\Scripts\python.exe" "C:/Users/admin/Desktop/2 курс/2 курс 1 семестр/ЙОПІ/LAB3/main.py"
Enter a file name: input_10.txt

Process finished with exit code 0
```

output.txt: Блокнот

Файл Редагування Формат Вигляд Довідка

```
Center of gravity: (3.891666666666667, 35.0)
Covariance = 22.999999999999996
---------------------------------------------------------------------------
Regression line equation: y = 9.9534184823441*x - 3.73538692712247
---------------------------------------------------------------------------
Correlation coefficient = 0.9010014623100245
```

Рд 1, ствп 1 | 100% | Windows (CRLF) | UTF-8



Результат роботи програми при введенні даних із файлу input_100.txt:

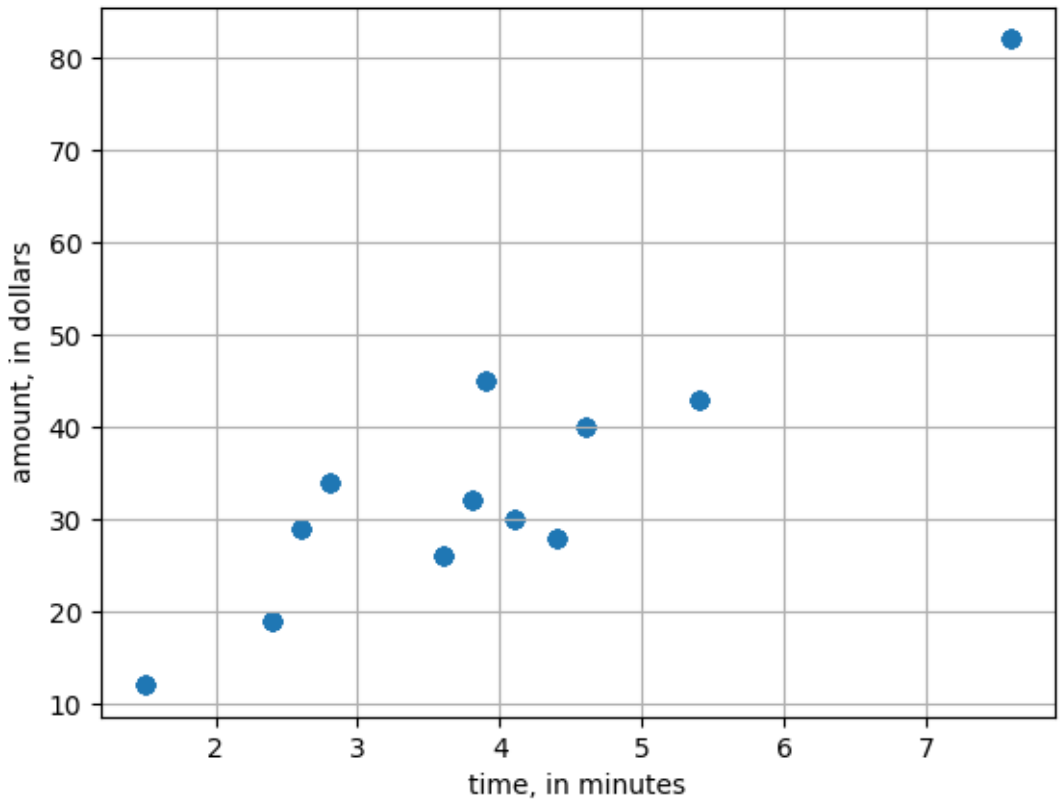
```
"C:\Users\admin\Desktop\2 курс\2 курс 1 семестр\ЙОПІ\LAB3\venv\Scripts\python.exe" "C:/Users/admin/Desktop/2 курс/2 курс 1 семестр/ЙОПІ/LAB3/main.py"
Enter a file name: input_100.txt

Process finished with exit code 0
```

output.txt: Блокнот

Файл Редагування Формат Вигляд Довідка

```
Center of gravity: (3.856000000000001, 34.5)
Covariance = 22.591999999999995
---------------------------------------------------------------------------
Regression line equation: y = 9.98204363255898*x - 3.99076024714746
---------------------------------------------------------------------------
Correlation coefficient = 0.9019503370715775
```

Рд 1, ствп 1 | 100% | Windows (CRLF) | UTF-8



Висновки:

Під час виконання цієї лабораторної роботи я навчилася використовувати здобуті знання про міри в двовимірній статистиці на практиці за допомогою мови програмування Python.