МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ
ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені ІВАНА ФРАНКА

Кафедра дискретного аналізу
та інтелектуальних систем

Індивідуальне завдання №1

з курсу "Теорія ймовірності та математична статистика"

Виконав:
студент групи ПМі-23
Гуменюк Станіслав

Оцінка

Перевірила:
доц. Квасниця Г.А.

Львів 2024

Постановка задачі:

1. Згенерувати вибірку заданого об'єму (не менше 50) з вказаного проміжку для дискретної статистичної змінної. На підставі отриманих вибіркових даних:

- побудувати варіаційний ряд та частотну таблицю; представити графічно статистичний матеріал, побудувати графік емпіричної функції розподілу; обчислити числові характеристики дискретного розподілу.

2. Згенерувати вибірку заданого об'єму (не менше 50) з вказаного проміжку для неперервної статистичної змінної. На підставі отриманих вибіркових даних:

- утворити інтервальний статистичний розподіл, побудувати гістограму та графік емпіричної функції розподілу, обчислити числові характеристики.

Короткі теоретичні відомості

1) **Дискретний випадок**

Нехай серед спостережень (1) зустрічаються такі можливі значення одновимірної дискретної варіанти x, впорядковані за величиною:

$$x_{(1)} < x_{(2)} < .... < x_{(k)}$$

і нехай ці значення зустрічаються відповідно часто:

$$n_1, n_2, ..., n_k, \quad (n_1 + n_2 + ... + n_k = n)$$

Число n_i називається частотою значення $x_{(i)} (i = 1,2,...k)$. Тоді статистичний матеріал (1) зручно записати в формі таблички з двома рядками у першому рядку виписуємо в зростаючому порядку можливі значення варіанти, а в другому – відповідні їм частоти. Дістанемо частотну таблицю

$$\begin{array}{c|c} x_{(1)} \quad x_{(2)} \; \ldots\ldots\ldots\ldots \; x_{(k)} & \sum \\ \hline n_1 \quad n_2 \; \ldots\ldots\ldots\ldots\ldots \; n_k & n \end{array} \qquad (2)$$

Частотна таблиця (2) називається ще **статистичним розподілом дискретної варіанти** x.

Для графічного представлення частотної таблиці на вісь абсцис наносимо можливі значення дискретної мінливої величини та відкладемо в цих точках відповідні частоти $n_i (i = 1, 2, ...)$. Отримаємо **діаграму частот**.

Якщо з'єднати відрізками сусідні пункти $(x_{(i)}, n_i)$, то дістанемо **полігон** частот.

2) **Неперервний випадок**.

а) **не згруповані дані**.

Якщо статистичний матеріал малий або середній, то спостереження (1) над одновимірною неперервною варіантою впорядкуємо за величиною: від найменшого до найбільшого. Нехай $x_{(1)}$ буде найменше зі спостережень (1) і т.д., в кінці $x_{(n)}$ буде найбільше зі спостережень (1)

$$x_{(1)} = \min \quad (x_1, ..., x_n)$$

....

$$x_{(n)} = \max \quad (x_1, ..., x_n)$$

В силу обмеженої точності деякі спостереження можуть бути одинакові. так упорядковані спостереження (1) записуємо у формі ряду:

$$x_{(1)} \leq x_{(2)} \leq .... \leq x_{(n)} \qquad (3)$$

Ряд (3) називається **варіаційним рядом** для спостережень (1) над одновимірною неперервною мінливою величиною.

Приклад 1. Періодична система Менделєєва, з огляду на атомну вагу елементів, утворює варіаційний ряд.

Для графічного представлення варіаційного ряду наносимо на вісь абцис елементи варіаційного ряду $x_{(i)} (i = 1,2,.....,n)$ та пов'яжемо з кожною точкою $x_{(i)}$ масу $\frac{1}{n}$.

Нарисуємо східчасту лінію зі стрибками вверх у пунктах $x_{(i)}$ на $\frac{1}{n}$. Від $-\infty$ до $x_{(1)}$ маємо лінію на рівні нуль. У точці $x_{(1)}$ маємо стрибок на $\frac{1}{n}$ і відрізок на висоті $\frac{1}{n}$ до точки $x_{(2)}$ у точці $x_{(n)}$ останній стрибок на $\frac{1}{n}$ і лінія на висоті 1буде продовжуватися до безмежності. Якщо б зустрілося два одинакові $x_{(i)}$, або більше, то в цій точці був би стрибок на $\frac{2}{n}$, або відповідно більше.

Одержане графічне представлення варіаційного ряду називаються **емпіричною функцією розподілу** або **емпіричною кумулятою.**

Таким чином, емпірична функція розподілу

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \frac{k}{n}, & x_{(k)} \leq x < x_{(k+1)} \quad (k = 1,....,n-1) \\ 1, & x_{(n)} \leq x \end{cases}$$

Для кожного x емпірична функція розподілу $F_n(x)$ є випадковою змінною з розподілом $P\left\{F_n(x) = \frac{k}{n}\right\} = C_n^k \left[F_n(x)\right]^k . \left[1 - F_n(x)\right]^{n-k}$

б) **<u>Згруповані дані</u>**. Якщо статистичний матеріал середній або великий, то знайдемо найменше та найбільше зі спостереження

$$x_{(1)} = \min \quad (x_1,......., x_n), \quad x_{(n)} = \quad \max \quad (x_1,..........., x_n)$$

Означення. Різниці між найбільшим і найменшим елементами статистичного матеріалу називається **<u>розмахом статистичного матеріалу</u>**

$$\rho = \quad x_{(n)} \quad - \quad x_{(1)} \qquad\qquad 2^r < \quad n \quad \le \quad 2^{r+1}$$

Інтервал розмаху ділимо досить довільним способом на $(r+1)$ однакові або неоднакові інтервали, де $r-$натуральне , $r = 1, \quad 2,......$

Центри одержаних інтервалів позначимо в зростаючому порядку через $z_1,......., z_i,.........z_{r+1.}$

Нехай на інтервалі з центром в точці z_i попадає n_i спостережень.

Очевидно, що $n_1 + \quad n_2 + ... \quad n_{r+1} \quad = n$

Тоді статистичний матеріал представимо у вигляді таблиці з двох рядків:

1-й в зростаючому порядку- центри інтервалів

2-й - відповідні частоти

$z_1 \quad z_2 ....z_i.......................z_{r+1}$	$\sum$
$n_1 n_2 ....n_i.......................n_{r+1}$	n

У цій таблиці замість кожного з n_i індивідуальних значень статистичного матеріалу (1), що попадають у інтервал з центром в т. z_i, розглядається $n_i -$кратно повторений центр $i-$го інтервалу z_i .

Одержана таблиця – частотна. При такому представлені дальша математична обробка статистичного матеріалу значно спрощується.

Для графічного представлення одержаної частотної таблиці наносимо на абцису центри інтервалів. В точці z_i ставимо ординату n_i. Одержимо **графік частот**

Якщо з'єднати верхушки сусідніх вершин графіка частот відрізками, то одержимо многокутник частот або **<u>полігон частот</u>** .

Якщо над інтервалом з центром в т. z_i поставити прямокутник висотою n_i, то одержимо **<u>гістограму частот.</u>**

Ми розглянули лише такі графічні представлення, які нагадують функцію розподілу або густину.

Числові характеристики поділяються на три групи:

1. Числові характеристики **<u>центральної тенденції</u> (локації)**. До них відноситься:

а) медіана (M_e)

б) мода (M_o)

в) середнє арифметичне ($\bar{x}$)

2. Числові характеристики **розсіяння**. До них відноситься:

а) варіанса (s^2)

б) стандарт (s)

в) розмах (ρ)

г) варація (v)

д) інтерквантильність широт

3. Числові характеристики **форми**: До них відноситься:

а) асиметрія (γ_1) *(Ac)*

б) ексцес (γ_2) *(Ек)*

<u>Статистики центральної тенденції</u>

<u>Медіана.</u>

<u>Означення.</u> Медіаною називають цей елемент статистичного матеріалу, який ділить відповідний варіаційний ряд (3) на дві рівні за обсягом частини. Медіану позначаємо M_e.

Якщо обсяг статистичного матеріалу непарний, то медіана визначається однозначно. Наприклад, якщо варіаційний ряд статистичного матеріалу буде ($n = 2k + 1$)

$$\underbrace{x_{(1)} \le x_{(2)} \le \cdots \le x_{(k)}}_{k} \le x_{(k+1)} \le \underbrace{\cdots \le x_{(2k+1)}}_{k} \text{ , то } M_e = x_{(k+1)}$$

Якщо обсяг статистичного матеріалу парний, то медіаною може бути інтервал. Наприклад, якщо варіаційний ряд статистичного матеріалу буде ($n = 2k$)

$$\underbrace{x_{(1)} \le x_{(2)} \le \cdots \le}_{k-1} x_{(k)} \le x_{(k+1)} \le \underbrace{\cdots \le x_{(2k)}}_{k-2}$$

То $M_e = \lfloor x_{(k)}, x_{(k+1)} \rfloor$ $\quad M_e = \dfrac{x_{(k)} + x_{(k+1)}}{2}$

<u>Мода.</u>

Означення. Модою називають цей елемент статистичного матеріалу, який найчастіше зустрічається. Моду позначаємо M_o.

Не виключено, що декілька значень статистичного матеріалу зустрічаються найчастіше та однаково часто, тоді всі вони модні. Мода типове значення статистичного матеріалу. Мода широко використовується в демографії. У демографії, при багатовершинних розподілах, краще вказати моди, ніж середнє арифметичне.

5	4	3	2	1	Σ	$M_o = 4$ (найчастіше зустрічається)
7	12	5	0	1		

Середнє арифметичне.

Означення. Середнім арифметичним називається сума всіх елементів статистичного матеріалу, поділена на обсяг статистичного матеріалу, позначається $\bar{x}$:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Статистики розсіяння

Варіанса.

Означення. **Варіансою** називається девіація (сума квадратів відхилень елементів статистичного матеріалу від середнього арифметичного) поділена на обсяг статистичного матеріалу без одного і позначається s^2.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Стандарт.

Означення. Стандартом (флуктуацією, середнім квадратичним відхиленням) називається арифметичний корінь з варіанси і позначається

$$s = +\sqrt{s^2}\ .$$

Розмах.

Означення. Розмахом називається різниця між найбільшим і найменшим елементами статистичного матеріалу і позначається ρ.

$$\rho = \max(x_1, \cdots, x_n) = \min(x_1, \cdots, x_n) = x_{(n)} - x_{(1)}$$

Варіацією вибірки називається відношення стандарту цієї вибірки до середнього арифметичного

$$v = \frac{s}{\bar{x}}.$$

Інтерквантильні широти.

Означення. Квантилем порядку α, якщо він існує, називається цей елемент статистичного матеріалу (відповідного варіаційного ряду), до якого включно маємо $\alpha\%$ елементів статистичного матеріалу (відповідного варіаційного ряду).

Статистичний матеріал (1) має квантилі тільки порядків кратних $\cdot\frac{100}{n}$, інші квантилі не існують; елемент $x_{(i)}$ є кантилем порядку $i\cdot\frac{100}{n}$ $(i = 1,\cdots,n)$.

Квантилі порядку 25, 50, 75 називаються **квартилями**: першим Q_1; другим Q_2; третім Q_3. Різниця між третім і першим квартилем $Q_3 - Q_1$ називається **інтерквартильною широтою** (інтерквартильний розмах).

Очевидно, що

$$Q_1 = x_{\left(\frac{n}{4}\right)},\ Q_3 = x_{\left(\frac{3n}{4}\right)}.$$

Від Q_1 виключно до Q_3 включно розташовано 50% центральних елементів статистичного матеріалу.

Квантилі порядку 12,5; 25,0;…, 87,5 називаються **октилями**: першим O_1; другим O_2; …сьомим O_7. Різниця між сьомим і першим октилем $O_7 - O_1$ називається **інтероктильною широтою**. Очевидно, що

$$O_1 = x_{\left(\frac{n}{8}\right)},\ldots,\ O_7 = x_{\left(\frac{7n}{8}\right)}$$

Від O_1 виключно до O_7 включно розташовано 75% центральних елементів статистичного матеріалу.

Квантилі порядку 10; 20;…, 90 називаються **децилями**: першим $Д_1$; другим $Д_2$; …дев'ятим $Д_9$. Різниця між дев'ятим і першим децилем $Д_9 - Д_1$ називається **інтердецильною широтою**. Очевидно, що

$$Д_1 = x_{\left(\frac{n}{10}\right)},\ldots,Д_9 = x_{\left(\frac{9n}{10}\right)}$$

Від $Д_1$ виключно до $Д_9$ включно розташовано 80% центральних елементів статистичного матеріалу.

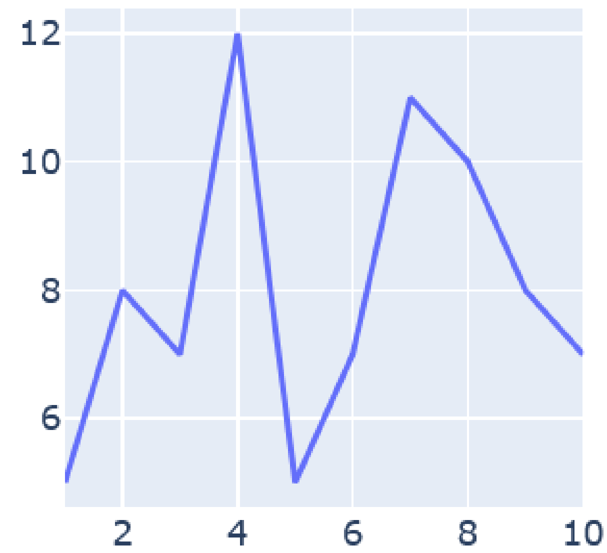
Програмна реалізація

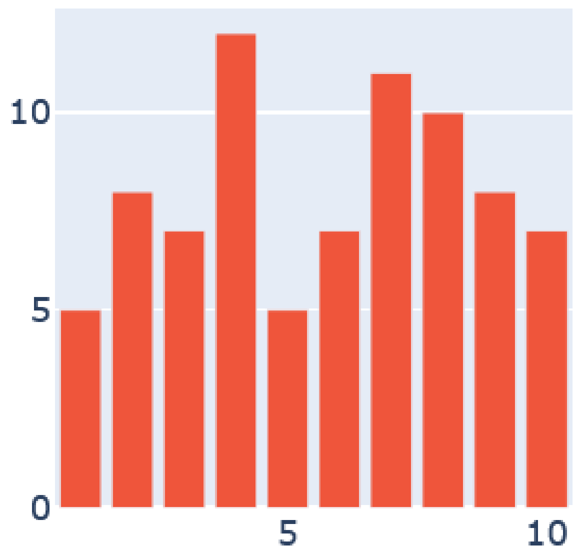
Програма реалізована на платформі .NET мовою програмування C#. Для математичних функцій використав бібліотеку Math, для побудови графіків використав бібліотеку Plotly.NET.

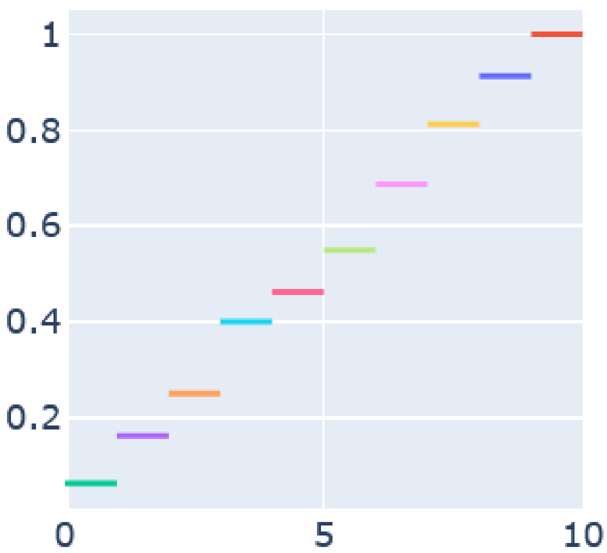
Отримані результати в дискретному випадку

```
Введіть початок проміжку:
1
Введіть кінець проміжку:
10
Введіть об'єм вибірки:
80
Вибірка:
1 3 7 8 8 3 2 2 6 8 4 1 10 9 4 5 6 6 8 9 4 9 5 2 10 4 6 4 4 1 9 7 4 5 4 10 2 10 6
 4 8 8 8 7 5 7 4 2 3 9 7 7 7 1 3 3 9 3 3 10 10 7 9 6 7 7 1 4 6 9 2 5 4 8 7 8 8 2
10 2
Варіаційний ряд:
1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 6 6 6 6
 6 6 6 7 7 7 7 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 8 8 9 9 9 9 9 9 9 9 10 10 10 10 10 1
0 10
Центральної тенденції
Медіана 6, 6
Середнє арифметичне: 5,7
Мода: 4
Розсіювання
Девіація: 596,7999999999998
Варіанса: 7,554430379746833
Стандарт: 2,748532404711073
Розмах: 9
Коефіцієнт варіації: 0,4821986674931707
Вибіркова дисперсія: 7,459999999999998
Вибіркове середнє квадратичне відхилення: 2,7313000567495322
```

```
Квантилі:
Q1 квартиль: 3
Q2 квартиль: 6
Q3 квартиль: 8
Інтерквартильна широта: 5

Q1 октиль: 2
Q2 октиль: 3
Q3 октиль: 4
Q4 октиль: 6
Q5 октиль: 7
Q6 октиль: 8
Q7 октиль: 9
Інтероктильна широта: 7

Q1 дециль: 2
Q2 дециль: 3
Q3 дециль: 4
Q4 дециль: 4
Q5 дециль: 6
Q6 дециль: 7
Q7 дециль: 8
Q8 дециль: 8
Q9 дециль: 9
Інтердецильна широта: 7


Коефіцієнти асиметрії та ексцесу
Коефіцієнт асиметрії: -0,08142132111471263
Коефіцієнт ексцесу: -1,1883234264603346
```







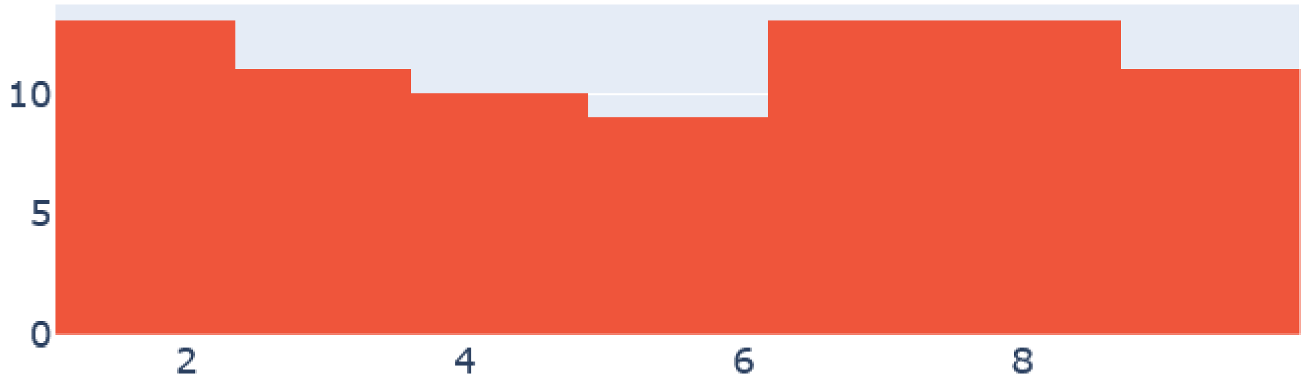
1	2	3	4	5	6	7	8	9	10
5	8	7	12	5	7	11	10	8	7

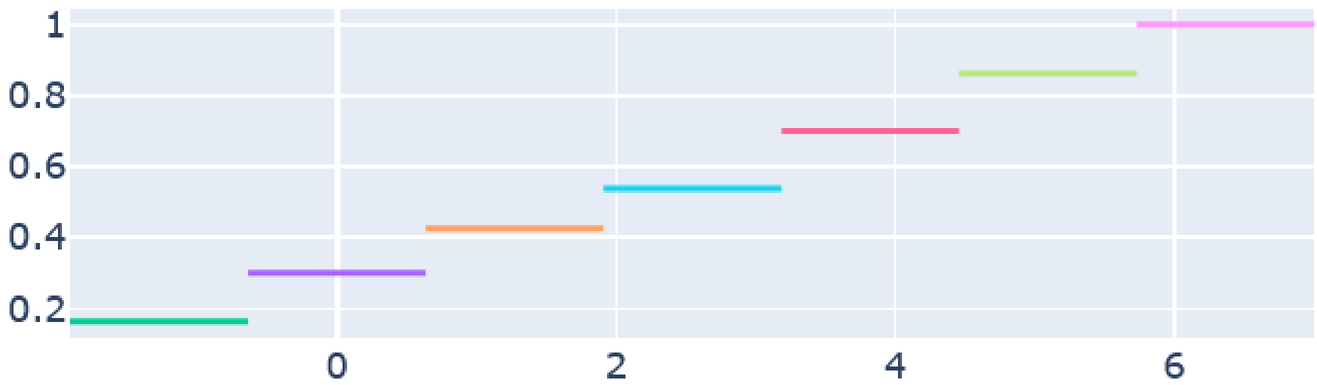
Неперервний випадок

```
Введіть початок проміжку:
1
Введіть кінець проміжку:
10
Введіть об'єм вибірки:
80
Вибірка:
6,885 4,365 2,014 9,186 5,793 2,921 6,648 7,49 9,118 9,164 1,643 8,382 6,865 2,095 5,166 8,479 1,509 9,928 4,014 4,285 6,355 7,295 1,
21 5,081 9,429 9,753 1,137 4,571 8,194 3,599 8,214 6,661 1,143 4,52 7,009 1,062 7,706 6,948 2,402 8,86 8,64 4,934 1,749 8,617 7,051 9
,98 6,687 9,117 3,288 2,045 9,705 2,521 8,7 3,093 5,491 8,565 9,145 3,228 4,539 3,217 8,205 4,325 7,935 5,194 1,911 4,867 3,945 6,093
 5,694 7,414 3,116 1,384 7,587 7,306 1,429 4,566 3,579 4,951 3,384 7,083
Центральної тенденції
Середнє арифметичне: 5,592237499999998
Медіана:
Медіана 5,491, 5,694
Мода: 2,336
Розсіювання
Девіація: 574,4153264875
Варіанса: 7,271080082120253
Стандарт: 2,696494035246556
Розмах: 8,918000000000001
Коефіцієнт варіації: 0,482185178159289
Вибіркова дисперсія: 7,18019158109375
Вибіркове середнє квадратичне відхилення: 2,6795879498709776
```

```
Квантилі:
Q1 квартиль: 3,228
Q2 квартиль: 5,491
Q3 квартиль: 7,935

Q1 октиль: 1,911
Q2 октиль: 3,228
Q3 октиль: 4,52
Q4 октиль: 5,491
Q5 октиль: 6,948
Q6 октиль: 7,935
Q7 октиль: 8,86

Q1 дециль: 1,643
Q2 дециль: 2,921
Q3 дециль: 3,599
Q4 дециль: 4,566
Q5 дециль: 5,491
Q6 дециль: 6,865
Q7 дециль: 7,414
Q8 дециль: 8,382
Q9 дециль: 9,118


Коефіцієнти асиметрії та ексцесу
Коефіцієнт асиметрії: -0,10321782718440806
Коефіцієнт ексцесу: -1,2392670114370796
```

(1,062; 2,336]	(2,336; 3,61]	(3,61; 4,884]	(4,884; 6,158]	(6,158; 7,432]	(7,432; 8,706]	(8,706; 9,98]
13	11	10	9	13	13	11





Висновки

Під час лабораторної з обробки статистичних даних на C# вивчено та успішно застосовано основні методи аналізу даних. З використанням функцій та структур C# реалізовано підрахунки, визначено середні значення, варіації та інші статистичні характеристики. Алгоритми підтвердили потужність та гнучкість C# у роботі з великим обсягом даних. Лабораторна робота поліпшила розуміння принципів статистичної обробки даних та їх використання в програмуванні