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-21
Урдейчук Ростислав Ігорович

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

Кількісні ознаки елементів генеральної сукупності можуть бути одновимірними і багатовимірними, дискретними і неперервними.
Коли реалізується вибірка, кількісна ознака, наприклад X, набуває конкретних числових значень $X = x_i$ які називають варіантою.
Зростаючий числовий ряд варіант називают варіаційним. Кожна варіанта вибірки може бути спостереженою n_i раз ($n_i >= 1$), число n_i називають частотою варіанти x_i.

$$n = \sum_{i=1}^{k} n_i$$

Відношення частоти n_i варіанти x_i називають її відносною частотою і позначають через W_i, тобто

$$W_i = \frac{n_i}{n}$$

Множина всіх можливих значень випадкової величини X називається генеральною сукупністю, а множина значень x_i (i = 1, 2..., k), яка одержана в результаті випробувань, вибіркою з генеральної сукупності або статистичною сукупністю. Число елементів вибірки називається обсягом вибірки.

Послідовність варіант, записаних за зростанням, називається варіаційним рядом (дискретним варіаційним рядом).

Якщо досліджується ознака генеральної сукупності X, яка є неперервною, то варіант буде багато. У цьому разі варіаційний ряд - це певна кількість рівних або нерівних частинних інтервалів чи груп варіант зі своїми частотами.

Такі частинні інтервали варіант, які розміщені у зростаючій послідовності, утворюють інтервальний варіаційний ряд.

2.1. Дискретний статистичний розподіл вибірки та її характеристики

Перелік варіант варіаційного ряду і відповідні їм частот, або відносних частот, називають дискретним статистичним розподілом вибірки.

У табличній формі він має такий вигляд:

$X = x_i$	x_1	x_2	...	x_k
n_i	n_1	n_2	...	n_k
W_i	W_1	W_2	...	W_k

2.2. Емпірична функція розподілу

Функція аргументу x, що визначає відносну частоту події $X < x$, тобто

$$F^*(x) = W(X < x) = \frac{n_i}{n}$$

називається емпіричною, або комулятою. Тут n - обсяг вибірки; n_i - кількість варіант статистичного розподілу вибірки, значення яких менше за фіксовану варіанту x; $F^*(x)$ - називають ще функцією нагромадження відносних частот.

Властивості $F^*(x)$:

1. $0 \leq F^*(x) \leq 1$
2. $F(x_{min}) = 0$, де x_{min} є найменшою варіантою варіаційного ряду;
3. $F(x)|_{x > x_{max}} = 1$, де x_{max} є найбільшою варіантою варіаційного ряду;
4. $F(x)$ є неспадною функцією аргументу x, а саме: $F(x_2) \geq F(x_1)$, при $x_2 \geq x_1$.

2.3 Полігон частот і відносних частот.

Дискретний статичтичний розподіл вибірки можна зобразити графічно у вигляді ламаної лінії, відрізки якої сполучають

координати точкок (x_i; n_i), або (x_i; W_i).
У першому випадку ламану лінію називають полігоном частот, у другому - полігоном відносний частот.

2.4 Числові характеристики дискретного статистичного матеріалу.

1. **Вибіркова середня величина** $\overline{x}_B$. Величину, яка визначається за формулою:

$$\bar{x}_B = \frac{\sum_{i=1}^{k} X_i n_i}{n}$$

   називають вибірковою середньою величиною дискретного статистичного розподілу вибірки.

2. **Мода** (Mo*). Модою дискретного статистичного розподілу вибірки називають варіанту, що має найбільшу частоту появи.

   Мод може бути кілька. Коли дискретний і статистичний розподіл має одну моду, то він називається одномодальним, коли має дві моди - двомодальним і т.д.
3. **Медіана** (Me*). Медіаною дискретного статистичного розподілу вибірки називають варіанту, яка поділяє варіаційний ряд на дві частини, рівні за кількістю варіант;

4. **Девіація** - сума квадратів відхилень елементів статистичного матеріалу від середнього арифметичного.

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

5. **Варіансою** s^2 називається девіація поділена на обсяг статистичного матеріалу без одного.

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

6. **Стандартом** називається арифметийний корінь з варіанси і позначається

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

7. **Дисперсія**. Для вимірювання розсіювання варіантв вибірки відносно $\overline{x}_B$ вибирається дисперсія. Дисперсія вибірки - це середнє армифметичне квадратів відхилень варіант відносно $\overline{x}_B$, яке обчислюється за формулою:

$$D_B = \frac{\sum_{i=1}^{k}(x_i - \bar{x}_B)^2 n_i}{n}$$

$$D_B = \frac{\sum_{i=1}^{k} x_i^2 n_i}{n} - (\bar{x}_B)^2$$

8. **Середнє квадратичне відхилення вибірки** σ_B. При обчисленні D_B відхилення підноситься до квадрата, а отже змінюється одиниця виміру ознаки X, тому на основі дисперсії вводиться середнє квадратичне відхилення.

$$\sigma_B = \sqrt{D_B},$$

   яке вимірює розсіювання варіант вибірки відносно $\overline{x}_B$, але в тих самих одиницях, в яких вимірюється ознака X;

9. **Розмах** (R), Для грубого оцінювання розсіювання варіант відносно $\overline{x}_B$ застосовується величина, яка дорівнює різниці між найбільшою x_{max} і найменшою x_{min} варіантами варіаційного ряду. Ця величина називається розмахом

$$R = x_{max} - x_{min};$$

10. **Квантилем порядку** α, якщо він існує називається цей елемент статистичного матеріалу (відповідного варіаційного ряду), до якого включно маємо α% елементів статистичного матеріалу (відповідного варіаційного ряду).
    При $\alpha < \beta$ різницю між квантилем порядку β і квантилем порядку α
    називають інтерквантильною широтою порядку $\beta - \alpha$

11. **Моментом** порядку k відносно сталої a називається вираз

$$\mu_k(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^k, k = 1, 2...$$

12. **Асиметрією** (γ_1) або скошеністю статистичного матеріалу називається відношення третього центрального моменту до другого центрального моменту в степені півтора

$$A_S = \gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

При $\gamma_1 > 0$ більшість елементів вибірки зосереджено в лівій половині інтервалу (статистичний матеріал скошений вправо).

При $\gamma_1 < 0$ більшість елементів вибірки зосереджено в правій половині інтервалу (статистичний матеріал скошений вліво).

При $\gamma_1 = 0$ статистичний матеріал розташований симетрично відносно середини інтервалу.

13. **Ексцесом** (γ_2) (крутістю, сплющеністю) статистичного матеріалу називається відношення четвертого центрального моменту до другого центрального моменту в квадраті мінус три

$$E_k = \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$$

Він виражає ступінь концентрації елеменів вибірки в околі її середнього:

Якщо $\gamma_2 > 0$, то статистичний матеріал високовершинний; якщо $\gamma_2 < 0$, то статистичний матеріал низьковершинний; якщо $\gamma_2 = 0$, то статистичний матеріал нормально вершинний.

5. Інтервально статистичний розподіл вибірки та його числові характеристики.

Перелік часткових інтервалів і відповідних їм частот, або відносних частот називають інтервальним статистичним розподілом вибірки.

У табличній формі цей розподіл має такий вигляд:

h	$x_0 - x_1$	$x_1 - x_2$	…	$x_{k-1} - x_k$
n_i	n_1	n_2	…	n_k
W_i	W_1	W_2	…	W_k

Тут $h = x_i - x_{i-1}$ є довжиною часткового інтервалу. Як правило цей інтервал береться однаковим.

6. Гістограма частот

Інтервальний статистичний розподіл вибірки можна подати графічно у вигляді гістограми частот або відносних частот, а також, як і для дискретного статистичного розподілу, емпіричною функцією $F^*(x)$ (комулятою).

Гістограмою частот називається східчаста фігура, яка складена з прямокутників, основами яких є частинні інтервали $(x_{i-1}; z_i]$, $i = 1, 2..., m$, а їх висоти $\tilde{h}_i = \frac{\tilde{n}_i}{z_i - z_{i-1}}$

Площа кожного такого прямокутника дорівнює n_i.

Гістограмою відносних частот називається східчаста фігура, яка складена з прямокутників, основами яких є частинні інтервали $(z_{i-1}; z_i]$, а їх висоти $\tilde{h}_i = \frac{\tilde{w}_i}{z_i - z_{i-1}}$.

7. Емпірична функція розподілу.

   Інтервально статистичний розподіл вибірки також характеризується своєю емпіричною функцією розподілу, але, на відміну від дискретного випадку, вона геометрично зображається ламаною лінією, яка з'єднує послідовно точки (z_i, ω_i), де $\omega_i = w_1 + w_2 + ... + w_i$, $\omega_0 = 0$

8. Числові характеристики інтервального статистичного матеріалу.
   1. Мода

      У випадку інтервального статистичного розподілу визначають модальний інтервал, тобто інтервал $[z_{M_0 - 1}, z_{M_0}]$, якому відповідає найбільша частота n_{M_0}.

      Тоді моду обчислюємо у вигляді

$$M_0(x) = z_{M_0-1} + \frac{n_{M_0} - n_{M_0-1}}{(n_{M_0} - n_{M_{M_0-1}}) + (n_{M_0} - n_{M_0+1})}(z_{M_0} - z_{M_0-1})$$

   2. Медіана

      Щоб знайти медіану інтервального статистичного розподілу вибірки, потрібно спочатку виділити медіанний інтервал, тобто той частинний інтервал $[x_{M-1}, z_M]$ зліва і справа від якого розміщені не більше половини варіант спостережень.

      Нехай n_M - відповідна йому частота, а m_{M-1} накопичена частота попереднього інтервалу. Тоді медіана

$$M_e = z_{M-1} + \frac{z_M - z_{M-1}}{n_M}(\frac{n}{2} - m_{M-1})$$

Програмна реалізація:

Для вирішення цієї задачі я використовую мову програмування Python разом з бібліотеками random, **numpy, matplotlib.pyplot** для генерації вибірки, обчислення статистичних характеристик та побудови графіків.

Дискретний випадок:
Користувач вводить мінімальне і максимальне значення для елементів вибірки та її розмір. Генерується масив рандомних значень у вказаних межах. Масив сортується і виводиться варіаційний ряд. Створюються два масиви (значень і кількості повторень) для побудови таблиці частот. Будується діаграма частот та полігон частот на основі вихідних даних. Обчислюються числові характеристики дискретного розподілу, такі як середнє значення, медіана, мода тощо.

Неперервний випадок:
Шукаються інтервали та кількість елементів, що потрапляють в кожен інтервал, для створення інтервальної таблиці. Будується гістограма та емпірична функція розподілу. Обчислюються числові характеристики, такі як середнє значення, медіана тощо, з урахуванням інтервалів. Ця програмна реалізація надає користувачеві можливість аналізувати вибірки для обох типів статистичних змінних, отримувати їх варіаційний ряд, графіки та числові характеристики.

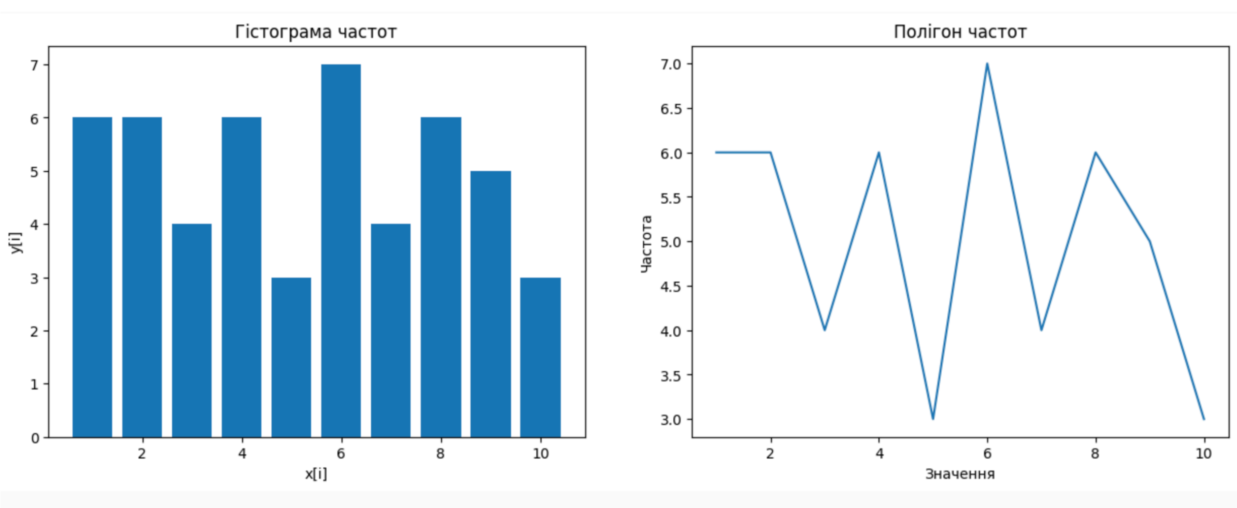
Аналіз отриманих результатів:

[1,2,5,4,4,9,8,6,9,8,6,2,1,1,10,10,6,3,6,6,5,3,9,4,8,10,2,9,1,7,8,4,8,9,4,4,1,2,6,3,1,6,5,7,7,3,7,2,2,8]

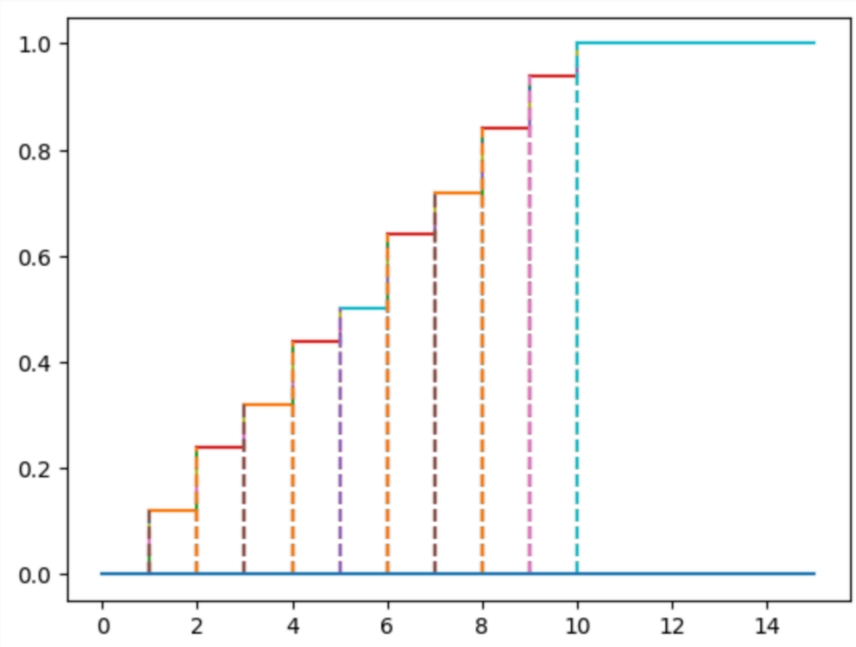
Дискретний розподіл:

x[i]	y[i]
1	6
2	6
3	4
4	6
5	3
6	7
7	4
8	6
9	5
10	3

Діаграма та полігон частот:



Графік емпіричної функції розподілу:



Числові характеристики:

```
Числові характеристики:
Медіана: 6.0
Мода: [7]
Вибіркове середнє значення: 5.35
Розмах: 9
Девіація: 154.55
Варіанса: 8.134210526315 79
Стандарт: 2.852053738328889
Варіація: 0.5330941566969886
Вибіркова дисперсія: 7.727500000000001
Вибіркове середнє квадратичне відхилення: 2.7798381247835278
Центральний момент другого порядку: 7.727500000000001
Центральний момент третього порядку: -4.206749999999992
Центральний момент четвертого порядку: 110.04473125
Асиметрія: -0.1958340397387614
Ексцес: -1.1571447396198629
```

Квантилі:

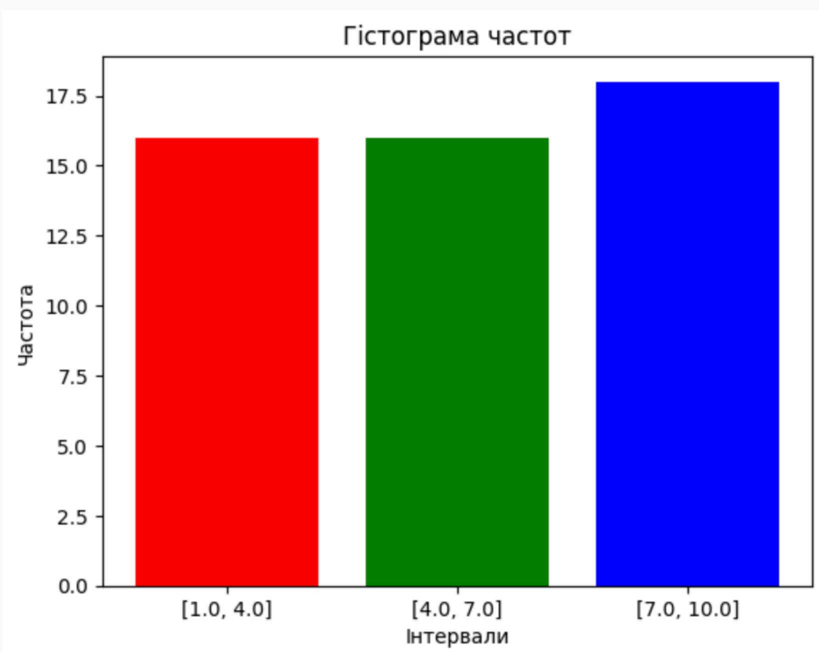
```
Децилі: [1, 2, 3, 4, 6, 6, 7, 8, 9]
Інтердецильна широта: 8
```

Неперервний розподіл:

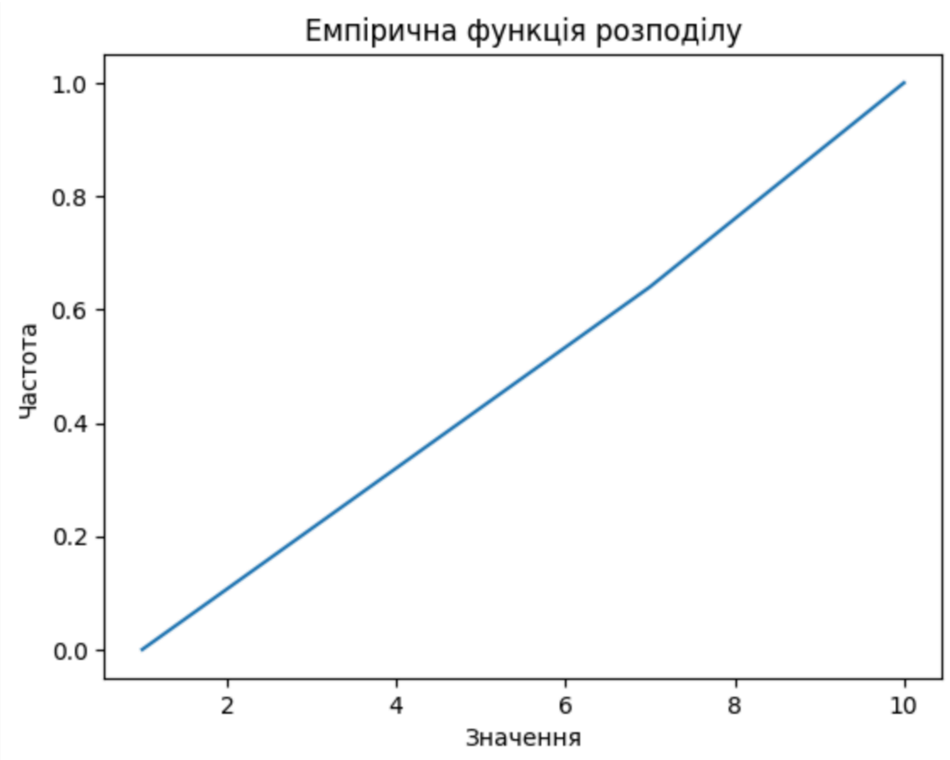
Частотна таблиця:

	Interval	Count
0	[1.0, 4.0]	16
1	[4.0, 7.0]	16
2	[7.0, 10.0]	18

Гістограма:



Графік емпіричної функції розподілу:



Числові характеристики:

```
Числові характеристики:
Медіана: 5.833333333333334
Мода: [7.3]
Вибіркове середнє значення: 5.62
Девіація: 305.28
Варіанса: 6.230204081632652
Стандарт: 2.496037676324749
Варіація: 0.4441348178513788
Вибіркова дисперсія: 1.1651908396946564
Вибіркове середнє квадратичне відхилення: 1.0794400584074395
Центральний момент другого порядку: 6.105599999999999
Центральний момент третього порядку: -1.1197440000000016
Центральний момент четвертого порядку: 55.089745920000006
Асиметрія: -0.0742209008027722
Ексцес: -1.5222054111783547
```

Висновок
Виконуючи це індивідуальне завдання, я глибше вивчив теоретичну частину і вдосконалив свої навички обчислення характеристик статистичного розподілу. Отримав навички представлення статистичних даних у вигляді таблиць, графіків, аналітичних висновків та числових характеристик