МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ФРАНКА
ФАКУЛЬТЕТ ПРИКЛАДНОЇ МАТЕМАТИКИ ТА
ІНФОРМАТИКИ

Індивідуальне завдання №1 з курсу "Теорія ймовірності та математична статистика"

Виконала:
Студентка групи ПМІ-23
Шувар Софія

Кафедра
ДИСКРЕТНОГО АНАЛІЗУ

Викладач:
Квасниця Галина
Андріївна

28 лютого 2021 р.

1 Постановка задачі.

Згенерувати вибірку заданого об'єму (не менше 50) з вказаного проміжку для дискретної статистичної змінної.

1. На підставі отриманих вибіркових даних: побудувати варіаційний ряд та частотну таблицю; представити графічно статистичний матеріал, побудувати емпіричну функцію розподілу; обчислити числові характеристики дискретного розподілу.

2. Для цієї вибірки утворити інтервальний розподіл, побудувати гістограму розподілу, обчислити числові характеристики для згрупованих даних.

2 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 ($n_i \geq 1$), число n_i називають частотою варіанти x_i.

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

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

$$W_i = \frac{n_i}{n} \tag{2}$$

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 елеменітв 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 відповідниї 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

2.1.1 Емпірична функція розподілу.

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

$$F^*(x) = W(X < x) = \frac{n_i}{n} \tag{3}$$

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 $x_2 \geq x_1$.

2.1.2 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.

2.1.3 Числові характеристики дискртеного статистичного матеріалу

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

   $$\bar{x}_B = \frac{\sum_{i=1}^{k} X_i n_i}{n} \tag{4}$$

   називають вибірковою середньою величиною дискретного статистичного розподілу вибірки.

2. **Мода** (Мо*). 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 ітп.

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.

$$dev = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5}$$

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

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{6}$$

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

$$s = \sqrt{s^2} \tag{7}$$

7. **Дисперсія**. Для вимірювання розсіювання варіантв вибірки відносно $\bar{x}_B$ вибирається дисперсія.

   Дисперсія вибірки - це середнє армифметичне квадратів відхилень варіант відносно $\bar{x}_B$, яке обчислюється за формулою:

$$D_B = \frac{\sum_{i=1}^{k}(x_i - \bar{x}_B)^2 n_i}{n} \tag{8}$$

   або

$$D_B = \frac{\sum_{i=1}^{k} x_i^2 n_i}{n} - (\bar{x}_B)^2 \tag{9}$$

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

$$\sigma_B = \sqrt{D_B}, \tag{10}$$

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

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

$$R = x_{max} - x_{min}; \tag{11}$$

10. **Коефіцієнт варіації** V. Для порівняння оцінок варіацій статистичних рядів із різним занчення $\bar{x}_B$, які не дорівнюють нулеві, вводиться коефіцієн варіації, який обчислюється за формулою

$$V = \frac{\sigma_B}{\bar{x}_B} 100\% \tag{12}$$

11. **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).

    При $\alpha < \beta$ різницю між квантилем порядку β і квантилем порядку α називають інтерквантильною широтою порядку $\beta - \alpha$

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

$$\mu_k(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^k, k = 1, 2... \tag{13}$$

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

$$A_S = \gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} \tag{14}$$

    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.

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

$$E_k = \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 \tag{15}$$

    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.

2.2 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.

2.2.1 Гісторграма 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 їхвисоти $\tilde{h}_i = \frac{\tilde{n}_i}{z_i - z_{i-1}}$

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 $\tilde{h}_i = \frac{\tilde{w}_i}{z_i - z_{i-1}}$.

2.2.2 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 $\omega_i = w_1 + w_2 + ... + w_i, \omega_0 = 0$

2.2.3 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

   $$M_0(x) = z_{M_0-1} + \frac{n_{M_0} - n_{M_0-1}}{(n_{M_0} - n_{M_{M_0-1}}) + (n_{M_0} - n_{M_0+1})}(z_{M_0} - z_{M_0-1}) \quad (16)$$

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

   $$M_e = z_{M-1} + \frac{z_M - z_{M-1}}{n_M}(\frac{n}{2} - m_{M-1}) \quad (17)$$

3 Програмна реалізація.

Свою програму я реалізувала за допомогою мови програмування Python, використовуючи середовище JupyterNotebook, а також можливості бібліотек: Pandas, Numpy, Matplotlib та IpyWidgets.

Інтерфейс користувача. Натиснувши відповідні кнопки, користувач може згенерувати рандомну вибірку (задавши її розмір та діапазон можливих

значень) або ввести її вручну у відповідному полі. Після цього, за допомогою кнопок, користувач може вибрати тип вибірки (дискретна або неперервна). Для Інтервального розподілу користувач повинен ще вказати кількість розбиттів на рівні відрізки вибірки. Також можна обирати представлення статистичного матеріалу (таблиці, полігони частот, емпірична функція розподілу, гістограми частот та числові характеристики). Інтерфейс користувача реалізований за допомогою IpyWidgets.

3.1 Дискретний статистичний розподіл вибірки.

Дискретний статистичний розподіл представлений за допомогою класу DiscreteSamplingDistribution з атрибутами: варіаційний ряд (масив NumPy), Дискретний статистичний розподіл представлений в виді Pandas таблиці (відповідність між варіантами та їх частотами або відносними частотами), розміром вибірки та кількістю різних елементів вибірки. В класі реалізовані наступні методи:

- $show\_database$ – демонструє статистичну таблицю;
- $counts\_polygon$ – метод для демонстрації полігону частот (Matplotlib);
- $frequency\_polygon$ – метод для демонстрації полігону відносних частот (Matplotlib);
- $count\_ecdf\_func$ – метод для обчислення значень емпіричної функції розподілу вибірки ;
- $print\_ecdf\_func$ – метод, що представляє емпіричну функцію розподілу у зрозумілому форматі;
- $drow\_ecdf$ – метод для графічного представлення емпіричної функції розподілу.(Matplotlib);
- get_mean, get_mode, get_median, get_range, $get\_deviation$, $get\_variance$, $get\_standard\_deviation$, $get\_variation$, $get\_dispersion$, $get\_standard\_error$, $get\_quantiles$, $get\_initial\_moment$, $get\_central\_moment$, $get\_skewness$, $get\_kurtosis$, - методи для обчислення відповідних числових характеристик дискретного розподілу.

3.2 Інтервальний статистичний розподіл вибірки.

Інтервальний статистичний розподіл представлений за допомогою класу IntervalSamplingDistribution похідного від класу DiscreteSamplingDistribution. Окрім атрибутів попередньо описаного класу, IntervalSamplingDistribution характеризується ще кількістю розбиттів дискретного статистичного розподілу та власне інтервальним розподілом представленим у вигляді Pandas таблиці. В класі реалізовані наступні методи:

- $show\_integral\_database$ - демонструє статистичну таблицю;
- $draw\_histogram\_count$ – метод для побудови гістограми частот(Matplotlib);
- $draw\_histogram\_frequency$ – метод для побудови гістограми відносних частот(Matplotlib);

- $get\_mean, get\_mode, get\_median, get\_range, get\_deviation, get\_variance, get\_standard_deviation, get\_variation, get\_dispersion, get\_standard_error, get\_initial\_moment, get\_central\_moment, get\_skewness, get\_kurtosis,$ - методи для обчислення відповідних числових характеристик інтервального розподілу.

3.3 Робота програми

За допомогою функції *generate_sample* генерується рандомна вибірка із вказаними користувачем параметрами, після чого вибірка передається параметрами конструктора в класи DiscreteSamplingDistribution та IntervalSamplingDistribution. Після виклику відповідних методів отримуємо потрібні результати.

4 Отримані результати та їх аналіз

Розмір вибірки: 60

Діапазон вибірки: $[1 - 20]$

Згенерована вибірка:
[13 10 12 14 19 19 3 1 9 15 19 4 2 18 8 9 20 3 6 3 12 12 10 14 14 17 7 8 11 11 2 2 11 13 7 11 17 20 9 12 4 17 12 11 17 7 2 18 2 4 16 13 8 15 5 2 16 20 5 19]

4.1 Дискретна величина

1. **Дискретний варіаційний ряд:**

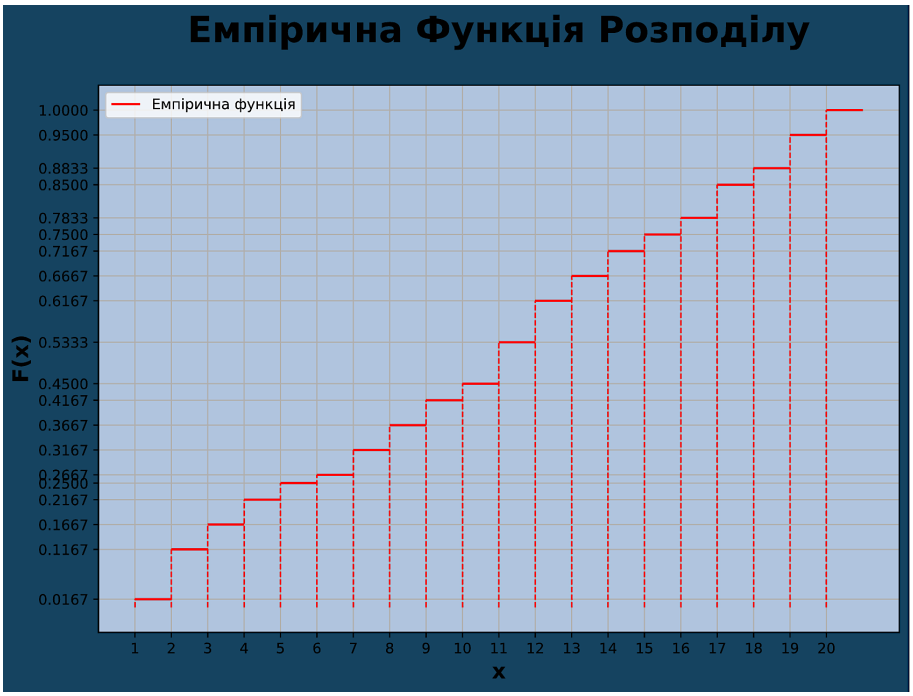
   Варіаційний ряд:
   [1 2 2 2 2 2 2 3 3 3 4 4 4 5 5 6 7 7 7 8 8 8 9 9 9 10 10 11 11 11 11 11 12 12 12 12 12 13 13 13 14 14 14 15 15 16 16 17 17 17 17 18 18 19 19 19 19 20 20 20]

2. **Таблиця частот та відносних частот:**

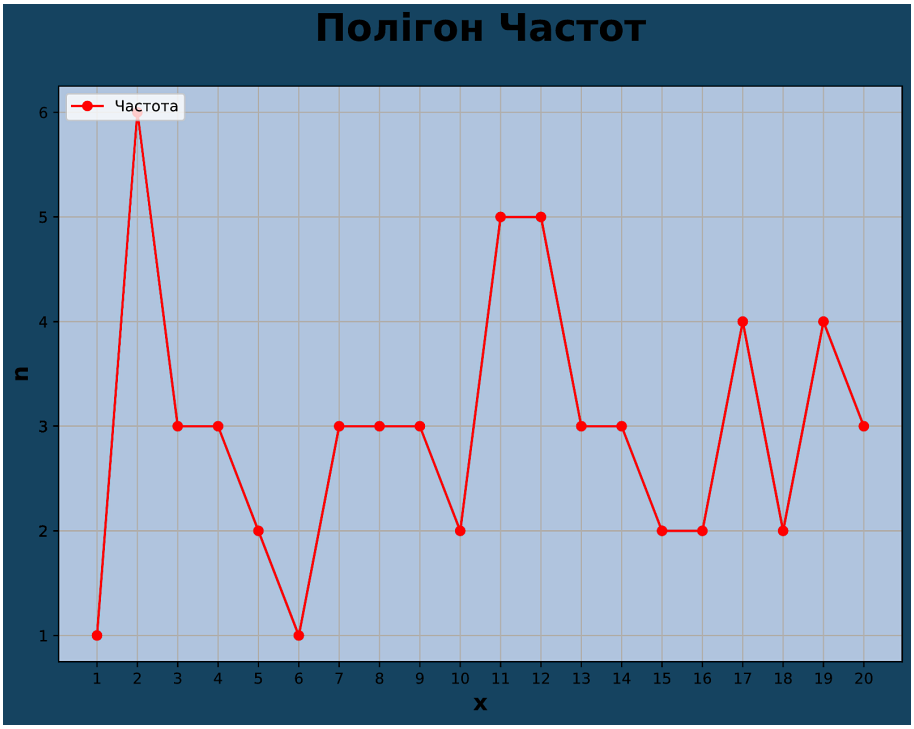
	ni	wi
1	1	0.016667
2	6	0.100000
3	3	0.050000
4	3	0.050000
5	2	0.033333
6	1	0.016667
7	3	0.050000
8	3	0.050000
9	3	0.050000
10	2	0.033333
11	5	0.083333
12	5	0.083333
13	3	0.050000
14	3	0.050000
15	2	0.033333
16	2	0.033333
17	4	0.066667
18	2	0.033333
19	4	0.066667
20	3	0.050000

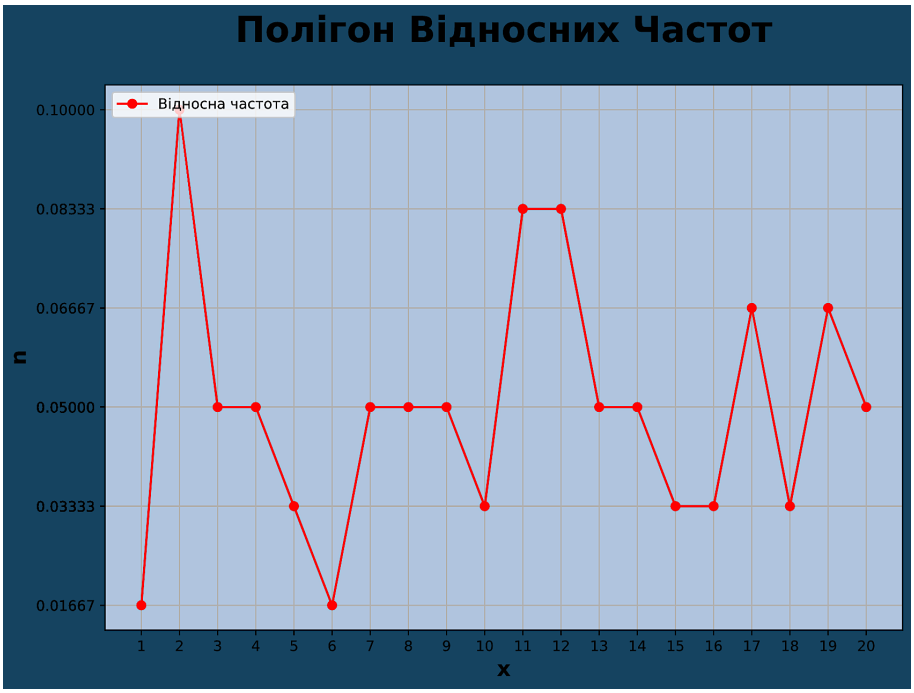
3. **Емпірична функція:**

```
0         x < 1;
0.02      0 <= x < 1;
0.12      1 <= x < 2;
0.17      2 <= x < 3;
0.22      3 <= x < 4;
0.25      4 <= x < 5;
0.27      5 <= x < 6;
0.32      6 <= x < 7;
0.37      7 <= x < 8;
0.42      8 <= x < 9;
0.45      9 <= x < 10;
0.53      10 <= x < 11;
0.62      11 <= x < 12;
0.67      12 <= x < 13;
0.72      13 <= x < 14;
0.75      14 <= x < 15;
0.78      15 <= x < 16;
0.85      16 <= x < 17;
0.88      17 <= x < 18;
0.95      18 <= x < 19;
1.0       x >= 20;
```



4. **Полігон частот та полігон відносних частот:**





5. **Числові характеристики дискретного статистичного розподілу:**

   - Вибіркове середнє: 10.666666666666666
   - Мода: 2
   - Медіана: 11.0
   - Розмах вибірки: 19
   - Девіація: 1971.3333333333335
   - Варіанса: 33.41242937853107
   - Стандарт: 5.780348551647303
   - Варіація: 0.5419076767169347
   - Вибіркова дисперсія: 32.85555555555556
   - Вибіркове середнє квадратичне відхилення: 5.731976583653807
   - Інтерквантильні широти:
     - Квартилі: $[6, 11, 16]$
     - Інтерквартильна широта: 10
     - Децилі: $[2, 4, 7, 9, 11, 12, 14, 17, 19]$
     - Інтердецильна широта: 17
   - Асиметрія: -0.04853646789345575 => Статистичний матеріал скошений вліво.
   - Ексцес: -1.1596239825276036 => Статистичний матеріал низьковершинний.

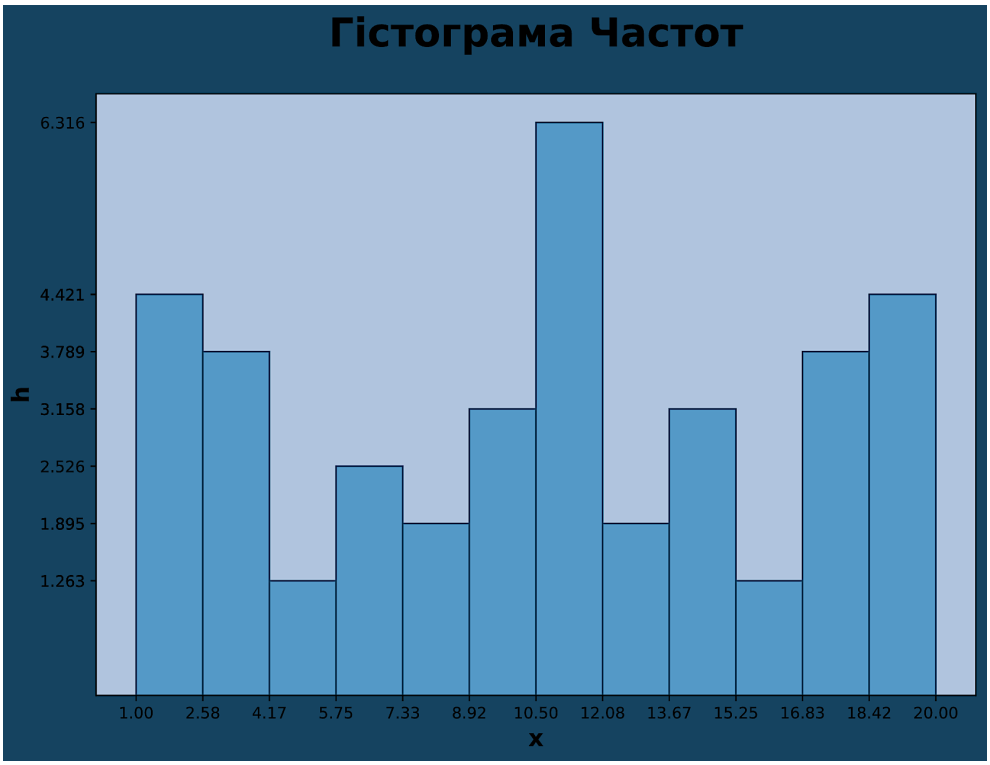
4.2 Неперервна величина

Кількість класів: 12

1. **Таблиця частот та відносних частот:**

	ni	wi
[1.0, 2.58]	7	0.116667
(2.58, 4.17]	6	0.100000
(4.17, 5.75]	2	0.033333
(5.75, 7.33]	4	0.066667
(7.33, 8.92]	3	0.050000
(8.92, 10.5]	5	0.083333
(10.5, 12.08]	10	0.166667
(12.08, 13.67]	3	0.050000
(13.67, 15.25]	5	0.083333
(15.25, 16.83]	2	0.033333
(16.83, 18.42]	6	0.100000
(18.42, 20.0]	7	0.116667

2. **Гістограми частот та відносних частот:**



3. **Числові характеристики інтервального статистичного розподілу:**

   - Вибіркове середнє: 10.631944444444446
   - Мода: 11.159722222222223
   - Медіана: 10.975
   - Розмах вибірки: 19.0
   - Девіація: 1976.9346064814822
   - Варіанса: 33.50736621155055
   - Стандарт: 5.78855476017551
   - Варіація: 0.5444493046801262
   - Вибіркова дисперсія: 32.948910108024705
   - Вибіркове середнє квадратичне відхилення: 5.740114119773639
   - Асиметрія: $-0.08143013321707629 =>$ Статистичний матеріал скошений вліво.
   - Ексцес: $-1.1900560350621816 =>$ Статистичний матеріал низьковершинний.

5 Висновки

У результаті виконання завдання, я навчилась будувати варіаційний ряд, будувати статистичний розподіл вибірки, зображати графічно статистичний матеріал, будувати емпіричну функцію розподілу та обчислювати усі числові характеристики для неперервної та дискретної статистичних змінних (медіану, моду, середнє вибіркове, девіацію, варіансу, дисперсію, стандарт, варіацію, розмах, квантилі, асиметрію та ексцес).