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

Індивідуальне завдання №2 з курсу "Теорія ймовірності та математична статистика"

Виконала:
Студентка групи ПМІ-23
Шувар Софія

Кафедра
ДИСКРЕТНОГО АНАЛІЗУ

Викладач:
Квасниця Галина
Андріївна

4 квітня 2021 р.

1 Постановка задачі.

Зчитати дані з текстовго файлу, побудувати полігон або гістограму частот.

На основі графічного представлення сформулювати гіпотезу про закон розподілу досліджуваної ознаки генеральної сукупності. (У задачай 1-5 рекомендуємо перевіряти вибірки на нормальний закон, а в задачах 6-10 - на інші, на приклад рівномірний показниковий біномний закон розподілу Пуссона);

Передбачити можливість користувачу задати параметри розподілу вручну або оцінити на основі даних вибірки;

для заданого користувачем рівня значущості перевірити сформульовану гіпотезу за критерієм χ^2.

2 Короткі теоретичні відомості.

Однією з найбільш важливих задач математичної статистики є задача про визначення закону розподілу ймовірностей випадкової величини (ознаки генеральної сукупності) за даними вибірки.

Будь-які припущення чи передбачення того чи іншого закону розподілу випадкових величин завжди є статистичними гіпотезами. Об'єктивні дані про них можна отримати за допомогою спеціальних статистичних правил, які називаються критеріями узгодження. Це дає можливість охопити проблему в цілому, оцінити рівень статистичної достовірності між різними показниками, об'єктивно виміряти в межах імовірнісних підходів ступінь ризику, а також з'ясувати, за яких умов гіпотези можна прийняти, а за яких – відхилити .

Оскільки усі припущення щодо законів розподілу є гіпотезами, то їх необхідно перевіряти . Перевірка гіпотез здійснюється за допомогою статистичних критеріїв, що поділяють множину їх можливих значень на дві протилежні підмножини A і $\neg A$, в одній з яких нульова гіпотеза приймається, а в іншій – відхиляється .

Якщо розбіжність між емпіричним і теоретичним розподілами виявиться випадковою, то дані спостережень (вибірка) узгоджуються з гіпотезою про закон розподілу генеральної сукупності й гіпотеза приймаєьсятся. Якщо ж розбіжність виявиться суттєвою, то дані спостережень не узгоджуються з гіпотезою і вона відхиляється.

Для перевірки статистичних гіпотез відомо близько шести десятків різних критеріїв узгодження. З їх допомогою можна отримати об'єктивні дані про те, за яких умов розбіжності між емпіричними і теоретичними розподілами є випадковими (несуттєвими), а за яких – принциповими (суттєвими). Зупинимося на критерії χ^2 Пірсона(критерій узгодження),який грунтується на визначенні відхилення емпіричних характеристик від гіпотетичних характеристик.

Критерій Пірсона (критерій узгодження) має вигляд:

$$K = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i} = n \sum_{i=1}^{m} \frac{(w_i - p_i)^2}{p_i}, \tag{1}$$

де n_i - емпіричні частоти, n_ip - теоретичні частоти, w_i - емпіричні відносні частоти, p_i - теортетичні імовірності, n - обсяг вибірки.

Пірсон показав, що при збільшенні n закон розподілу величини R наближається до розподілу χ^2, тому значення $R_0 = \chi^2_{kp}$ для заданого рівня значущості α і числа степенів свободи k вибирається з таблиці критичних точок розподілу χ^2.

Число степенів свободи обчислюється за формулою $k = m - s - 1$, де m - кількість інтервалів статистичного розподілу вибірки, s - кількість невідомих параметрів теоретичного розподілу, які оцінюються за даними вибірки. Оскільки дані вибірки підлягають обов'язковій умові

$$\sum_{i=1}^{l} \omega_i^* = 1 \tag{2}$$

то число k зменшується ще на 1.

Далі застосування критерію χ^2 виконується за загальною схемою перевірки гіпотези про вигляд густини розподілу ймовірностей випадкової величини за критерієм Пірсона:

- статистичні дані (результативибірки)записують у вигляді інтервального статистичного розподілу;
- оскільки перевіряється гіпотеза про те, що генеральна сукупність задовольняє певному(конкретному) закону розподілу, то для кожного інтервалу визначаєм теоретичні ймовірності попадання значень випадкової величини в цей інтервал;
- обчислюють емпіричне значення критерію узгодження Пірсона;

$$\chi^2_{emp} = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}. \tag{3}$$

- за даним рівнем значущості α і кількістю $k = m-s-1$ ступенів вільності занходимо критичну точку $k_{kp} = \chi^2_{kp}(\alpha, k)$ за таблицею критичних значень розподілу χ^2_{kp};
- співставляємо значенні χ^2_{emn} і k_{kp}: якщо $\chi^2_{emn} \geq k_{kp}$, то гіпотезу H_0 про вигляд густини розподілу відхиляють. Якщо ж $\chi^2_{emn} < k_{kp}$, то гіпотезу H_0 приймають.

Схема перевірки гіпотези про вигляд закону розподілу ймовірностей дискретної випадкової величини має незначні відмінності:

- статистичні дані (результати вибірки)записують у вигляді дискретногостатистичного розподілу;
- на підставі гіпотетичного закону розподілу знаходимо теоретичні ймовірності p_i того, що випадкова величина приймає значення x_i.

Зауваження.Критерій Пірсона застосовують для великих обсягів вибірок, $n \leq 100$. Також мають виконуватись умови $n_i \leq 5$, $np_i \leq 10$ в окремих групах. Якщо ці умови не виконуються, сусідні групи об'єднують.

Розглянемо застосування критерію χ^2 для перевірки гіпотез про розподіл ознаки X генеральної сукупності за найбільш відомими і часто вживаними розподілами.

1. **Нормальний розподіл.**
   Для перевірки гіпотези про нормальний розподіл ознаки X за оцінки невідомих параметрів a і σ приймається вибіркова середня x_B і вибіркове середнє квадратичне відхилення s, Тому ймовірності p_i для i-го інтервалу статистичного розподілу обчислюються за формулою:

$$p_i = \Phi(\frac{x_{i+1} - \bar{x}_B}{s}) - \Phi(\frac{x_i - \bar{x}_B}{s}) \tag{4}$$

2. **Показниковий розподіл.**
   Для перевірки гіпотези про показниковий розподіл ознаки X за оцінку невідомого параметра λ приймається величина $\lambda^* = \frac{1}{\bar{x}}$, тому число p_i для інтервалів (x_{i-1}, x_i) статистичног розподілу знаходиться за формулою

$$p_i = e^{-\lambda^* x_{i-1}} - e^{-\lambda^* x_i}, \tag{5}$$

3. **Рівномірний розподіл.**
   Для перевірки гіпотези про рівномірний розподіл ознаки X за оцінки невідомих параметрів a і b - меж інтервалу - приймаються величини a^* і b^*, обчислювані за формулами

$$a^* = \bar{x}_B - \sqrt{3}s, b^* = \bar{x}_B + \sqrt{3}s. \tag{6}$$

   Тому ймовірності p_i для інтервалів (x_{i-1,x_i}) статистичного розподілу знаходиться за формулою:

$$p_i = \frac{x_i - x_{i-1}}{b^* - a^*} \tag{7}$$

4. **Біноміальний розподіл.**
   Для перевірки гіпотези про біноміальний розподіл ознаки X використовується дискретний статистичний розподіл вибірки, оскільки біноміальний розподіл застосовний до дискретної випадкової величини. Розглядається вибірка n серій по N випробувань, в кожному з яких подія A може відбутися із сталою ймовірністю p. Якщо ця ймовірність (як параметр розподілу) невідома, то вона оцінюється за даними вибірки величиною

$$\tilde{p} = \frac{1}{Nn} \sum_{i=1}^{m} x_i n_i, \tag{8}$$

   Де m - число варіант у вибірці, яке може бути меншим N. В цьому випадку число ймовірностей p_i обчислюється за формулою Бернуллі.

$$p_i = C_N^i \tilde{p}^i (1 - \tilde{p})^{N-i}. \tag{9}$$

5. **Розподіл Пуассона.**
   Для перевірки гіпотези про розподіл дискретної ознаки X за законом Пуассона також застосовується дискретний статистичний розподіл вибірки. Оцінкою параметра λ є величина $\lambda^* = \bar{x}_B$, тому ймовірності p_i для кожної варіанти обцичслюється заформулою Пуассона:

$$p_i = \frac{\lambda^{*i} e^{-\lambda^*}}{i!} \tag{10}$$

3 Програмна реалізація.

Свою програму я реалізувала за допомогою мови програмування Python, використовуючи середовище *JupyterNotebook*, а також можливості бібліотек: *Pandas*, *Numpy*, *Matplotlib*, *Scipy* та *IpyWidgets*. Умови завдань я зберегла у *csv* файлах.
Інтерфейс користувача. Інтерфейс користувача реалізований за допомогою *IpyWidgets*. Натиснувши відповідні кнопки, користувач може запустити виконання завдань запропонованих в умові 4 або 8 відповідно. Після цього, за допомогою випадаючого списку обрати свій варіант, за допомогою слайдера - рівень значущості та за бажання вказати невідомі параметри розподілу. Після цього користувач може побачити графічне представлення статистичного матеріалу(для завдання 4 – гістограму частот, для завдання 8 – діаграму та полігон частот) відповідного завдання та варіанту. Також можна побачити всі необхідні для виконання завдання числові характеристики (значення параметрів розподілу, рівень значущості, число часткових інтервалів, число параметрів густини гіпотетичного розподілу, число ступенів вільності, емпіричне та критичне значення критерію узгодження Пірсона), а також висновок щодо обраної гіпотези.

3.1 Завдання 8 (Дискретний сатистичний розподіл)

Дискретний статистичний розподіл представлений за допомогою класу *DiscreteSamplingDistribution* реалізованим на основі Pandas таблиці. В класі реалізовані наступні методи:

- *show_database* – демонструє статистичну таблицю;
- *counts_polygon* – метод для демонстрації полігону частот (*Matplotlib*);
- *counts_diagram*– метод для демонстрації діаграми частот (*Matplotlib*);
- *get_mean*, *get_deviation*, *get_dispersion*, *get_standard_error* – методи для обчислення відповідних числових характеристик;
- *Binomial_distribution_probabilities* – метод для обчислення ймовірностей для Біномного розподілу.
- *Normalise* – метод для об'єднання інтервалів у випадку, якщо не виконуються умови нормування ($n_i > 5$ та $np_i > 10$);
- *Pearson_cumulative_test_statistic* – метод для обчислення емпіричного значення критерію узгодження Пірсона;

- *Pearson_cumulative_test_critical* - метод для обчислення критичного значення критерію узгодження Пірсона;
- *print_results* – метод для виведення всіх результатів отриманих підчас роботи програми.

3.2 Завдання 3 (Інтервальний сатистичний розподіл)

Інтервальний статистичний розподіл представлений за допомогою класу *IntervalSamplingDistribution* похідного від класу *DiscreteSamplingDistribution*. В класі реалізовані наступні методи:

- *show_database* – демонструє статистичну таблицю;
- *draw_histogram*– метод для демонстрації гістограми частот (*Matplotlib*);
- *get_mean*, *get_deviation*, *get_dispersion*, *get_standard_error* – методи для обчислення відповідних числових характеристик;
- *Normal_distribution_probabilitis* – метод для обчислення ймовірностей для Нормального розподілу;
- *Normalise* – метод для об'єднання інтервалів у випадку, якщо не виконуються умови нормування ($n_i > 5$ та $np_i > 10$);
- *Pearson_cumulative_test_statistic* – метод для обчислення емпіричного значення критерію узгодження Пірсона;
- *Pearson_cumulative_test_critical* - метод для обчислення критичного значення критерію узгодження Пірсона;
- *print_results* – метод для виведення всіх отриманих результатів отриманих підчас роботи програми;
- *interval_x_toString* - статичний метод для виведення інтервалів у належному форматі.

3.3 Робота програми.

За допомогою методу *pd.read_csv*() бібліотеки *pandas* зчитуються дані з *csv* файлів, отримана таблиця транспонуються і передаються аргументами у відповідні класи *DiscreteSamplingDistribution* і *IntervalSamplingDistribution*. Після виклику відповідних методів отримуємо потрібні результати.

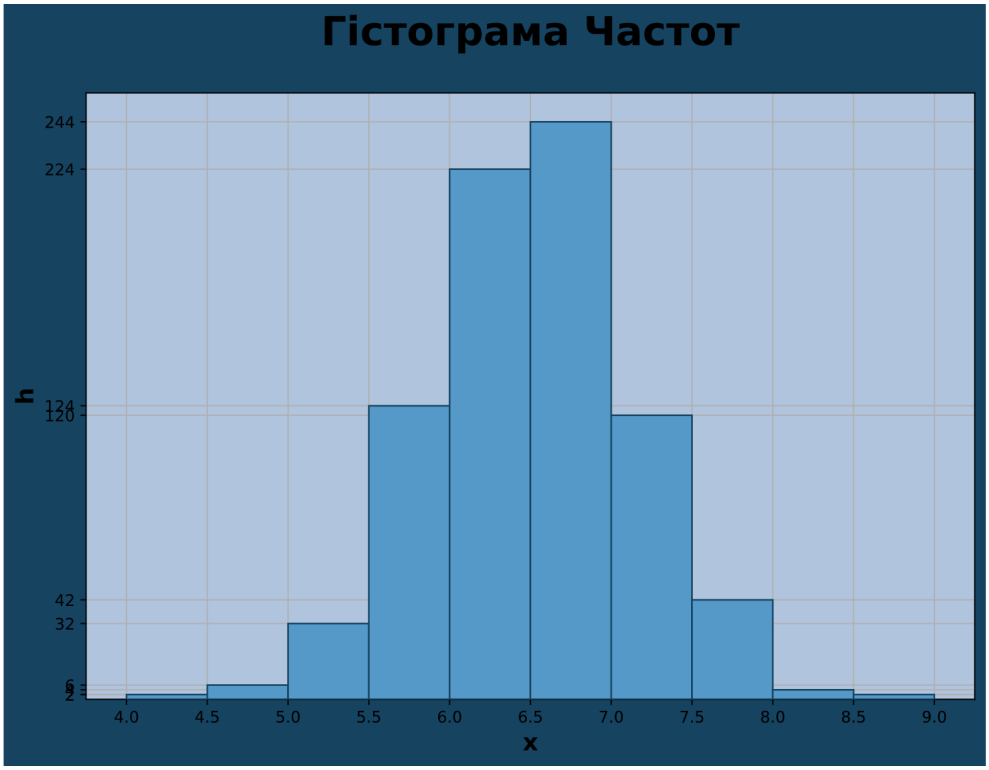
4 Отримані результати (графічні та числові) та їх аналіз.

4.1 ЗАДАЧА 4(Варіант 18)

З метою пошуку шляхів підвищення продуктивності на підприємстві реєструвався час, витрачений робітниками на виготовленя однотипних деталей. Розподіл кількості робітників n_i залежно від часу T, витраченого ними на виконання трудових операцій, наведено в таблиці

T	4.0-4.5	4.5-5.0	5.0-5.5	5.5-6.0	6.0-6.5	6.5-7.0	7.0-7.5	7.5-8.0	8.0-8.5	8.5-9.0
17	1	3	16	66	112	116	60	23	2	1
18	1	3	16	62	112	122	60	21	2	1
19	1	3	14	66	118	116	56	23	2	1
20	1	3	18	66	106	116	64	23	2	1
21	1	3	16	70	112	112	60	25	2	1

Базуючись на гістограму частот я вирішила перевірити гіпотезу про про Нормальний закон розподілу:



Статистичні дані до та після перевірки умов нормування:

T	ni	pi
4.0–4.5	1	0.000898
4.5–5.0	3	0.008569
5.0–5.5	16	0.048550
5.5–6.0	62	0.154860
6.0–6.5	112	0.278619
6.5–7.0	122	0.283032
7.0–7.5	60	0.162340
7.5–8.0	21	0.052525
8.0–8.5	2	0.009569
8.5–9.0	1	0.001037

	ni	pi
[-inf, 5.5]	20	0.058017
(5.5, 6.0]	62	0.154860
(6.0, 6.5]	112	0.278619
(6.5, 7.0]	122	0.283032
(7.0, 7.5]	60	0.162340
(7.5, inf]	24	0.063132

Числові Характеристики:

- a: 6.51375;
- s: 0.6450278579255317;
- Рівень значущості: 0.05;
- Число часткових інтервалів в інтервальному варіаційному ряді: 6;
- Число параметрів густини гіпотетичного розподілу: 2;
- Число ступенів вільності: 3;
- ЕМПІРИЧНЕ значення критерію узгодження Пірсона: 1.5653213697179609;
- КРИТИЧНЕ значення критерію узгодження Пірсона: 7.814727903251178;

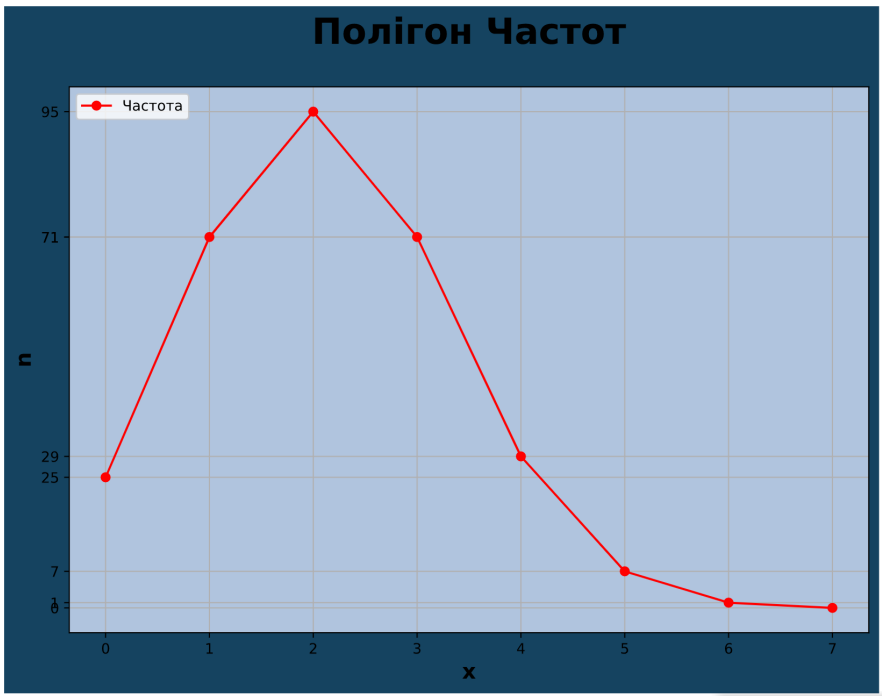
ВИСНОВОК: H0(Нормальний розподіл) - ПРИЙМАЄМО.

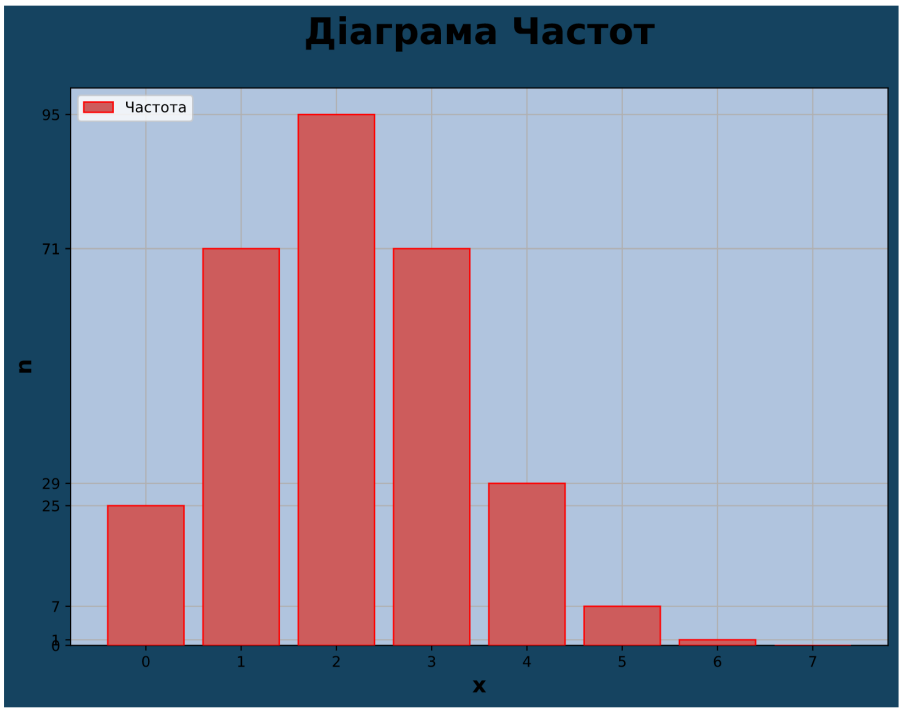
4.2 ЗАДАЧА8 (Варіант 18)

У деякій місцевості було зібрано дані про народжуваність дітей впродовж року. Розподіл кількості новонароджених n_i Залежно від місяця року відображено в таблиці.

devices	0	1	2	3	4	5	6	7
17	25	74	95	68	29	8	1	0
18	25	71	95	71	29	7	1	0
19	24	74	100	68	27	8	1	0
20	26	74	90	68	30	8	1	0
21	25	77	95	65	29	9	1	0

Базуючись на полігон та діаграму частот я вирішила перевірити гіпотезу про про Біномний закон розподілу:





Статистичні дані до та після перевірки умов нормування:

devices	ni	pi
0.0	25	0.081142
1.0	71	0.245147
2.0	95	0.317417
3.0	71	0.228329
4.0	29	0.098547
5.0	7	0.025520
6.0	1	0.003671
7.0	0	0.000226

	ni	pi
0.0	25	0.081142
1.0	71	0.245147
2.0	95	0.317417
3.0	71	0.228329
4.0 5.0 6.0 7.0	37	0.127964

Числові Характеристики:

- p: 0.30148113;
- Рівень значущості: 0.05;
- Число часткових інтервалів: 5;
- Число параметрів густини гіпотетичного розподілу: 1;
- Число ступенів вільності: 3;
- ЕМПІРИЧНЕ значення критерію узгодження Пірсона: 0.24539905628982692;
- КРИТИЧНЕ значення критерію узгодження Пірсона: 7.814727903251178;

ВИСНОВОК: H0(Біномний розподіл) - ПРИЙМАЄМО.

5 Висновки.

Під час виконання цього індивідуального завдання я застосувала свої знання для перевірки гіпотези про рівномірний розподіл генеральної сукупності за критерієм Пірсона. Я перевірла результати виконання моєї програми вручну та переконалася в правильності її обчислень.