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
Урбанський Максим

Оцінка
Перевірила:
доц. Квасниця Г.А.

Львів 2024

Постановка задачі

У поданих нижче задачах наведено результати досліджень вибірок з деяких генеральних сукупностей.

- Зчитати дані з текстового файлу, побудувати полігон або гістограму частот;
- на основі графічного представлення сформулювати гіпотезу про закон розподілу досліджуваної ознаки генеральної сукупності (у задачах 1 - 5 рекомендуємо перевіряти вибірки на нормальний закон, а в задачах 6 -10 — на інші, наприклад, рівномірний, показниковий, біномний, закон розподілу Пуассона);
- 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.

ВАРІАНТ 14

Короткі теоретичні відомості:

Гіпотетичні закони розподілу:

1. Біномний закон розподілу.

Імовірності обчислюються

$$p_i = P(\xi = i) = C_N^i p^i (1-p)^{N-i}$$, де $$p = \frac{\overline{x}}{N}$$

2. Нормальний закон розподілу.

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}} dt$$, де

параметри розподілу $a = \overline{x},\ \sigma = s$ оцінюються на основі вибірки.

3. Рівномірний закон розподілу.

$$F(x)=\begin{cases}0, & x<a\\ \dfrac{x-a}{b-a}, & a\le x<b\\ 1, & x\ge b\end{cases}$$

$$a=\bar{x}-\sqrt{3}s\,,\ b=\bar{x}+\sqrt{3}s$$

4. Показниковий закон розподілу.

$$F(x)=\begin{cases}0, & x<0\\ 1-e^{-\lambda x}, & x\ge 0\end{cases}$$, де $$\lambda=\frac{1}{\bar{x}}$$

5. Закон розподілу Пуассона.

$$P(\xi=i)=e^{-\lambda}\frac{\lambda^{i}}{i!}$$, де $\lambda>0$

Статистичною гіпотезою *називають будь-яке твердження про властивості (ознаки) генеральної сукупності, що перевіряється на основі вибірки.*

У математичній статистиці виділяють два основні типи статистичних гіпотез:

- гіпотези про закон розподілу ймовірностей випадкової величини (ознаки генеральної сукупності);
- гіпотези про значення параметрів розподілу випадкової величини (ознаки генеральної сукупності).

Наприклад, твердження "закон розподілу випадкової величини X є нормальний" – гіпотеза про закон її розподілу, твердження "у нормальному розподілі випадкової величини X параметри розподілу a= 20 і σ = 1,5"- гіпотеза про значення параметрів розподілу випадкової величини.

Сформульовану гіпотезу називають ***основною (нульовою)*** і позначають H_0.

У результаті статистичної перевірки гіпотези може бути прийняте одне з двох правильних рішень:

- гіпотеза приймається і вона є правильна;
- гіпотеза відхиляється і вона є неправильна.

Поряд з тим, у результаті статистичної перевірки статистичної гіпотези можуть бути допущені помилки (прийняті неправильні рішення) двох типів:

- гіпотеза відхиляється, але вона правильна (помилка першого роду);
- гіпотеза приймається, але вона неправильна (помилка другого роду).

Виявляється, що помилка першого роду має більш важкі наслідки, ніж помилка другого роду.

Виникає питання: як застрахувати себе від помилки першого роду або, принаймні, як звести до мінімуму ризик допустити помилку першого роду? Для цього вводиться спеціальне число α, яке виражає ймовірність відкинути вірну гіпотезу.

Імовірність допустити помилку першого роду називають ***рівнем значущості*** *і позначають через* α .

Число α задають наперед і найбільш часто його вибирають рівним 0,1; 0,05; 0,01. Якщо $\alpha = 0{,}05$, то це означає, що ймовірність допустити помилку першого роду є мала, а саме, ми ризикуємо її допустити у 5-ти випадках із 100.

Інформацію про випадкову величину, яка міститься у гіпотезі, називають ***гіпотетичною*** *або* ***теоретичною****, а інформацію про неї, яку отримують на основі вибірки, називають* ***статистичною*** *або* ***емпіричною****.*

Перевірка статистичної гіпотези проводиться за такою ***схемою:***

- *формулюють нульову гіпотезу* H_0 *і альтернативну гіпотезу* H_1 *і задають рівень значущості* α *для перевірки гіпотези* H_0;
- *визначають за статистичними даними критерій К для перевірки гіпотези* H_0, *який є випадковою величиною з відомим розподілом її ймовірностей*;
- *визначають критичні області відносно даних критерію К та рівня значущості* α. *Для визначення критичної області достатньо знайти критичні точки* $k_{кр}$ *за допомогою відповідних рівнянь* (1) – (4), *які наведені вище*;
- *знаходять емпіричне (спостережене) значення критерію* $K_{емп}$ *на основі вибірки*;
- *приймають рішення*: *якщо емпіричне значення критерію* $K_{емп}$ *попадає в критичну область, то нульову гіпотезу* H_0 *відхиляють*; *якщо ж значення* $K_{емп}$ *попадає в область допустимих значень, то нульову гіпотезу* H_0 *приймають.*

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

Якщо закон розподілу випадкової величини невідомий, то формулюють нульову гіпотезу про вигляд густини розподілу. Наприклад: „ випадкова величина має густину нормального розподілу ймовірностей".

Для перевірки таких гіпотез часто застосовують критерій ***„ хі-квадрат" Пірсона (критерій узгодження),*** який грунтується на визначенні відхилення емпіричних характеристик від гіпотетичних характеристик.

Критерій Пірсона (критерій узгодження) має вигляд:

$$K = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i} = n \sum_{i=1}^{m} \frac{(w_i - p_i)^2}{p_i}, \qquad (5)$$

де n_i - емпіричні частоти, np_i - теоретичні частоти, w_i – емпіричні відносні частоти, p_i – теоретичні ймовірності, n – обсяг вибірки.

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

- *статистичні дані* (*результати вибірки*) *записують у вигляді дискретного статистичного розподілу*:

x_i	x_1	x_2	...	x_m
n_i	n_1	n_2	...	n_m

- *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, $n \geq 100$. Також мають виконуватись умови $n_i \geq 5, np_i \geq 10$ в окремих групах. Якщо ці умови не виконуються, сусідні групи слід об'єднати.

Схема перевірки гіпотези про вигляд густини розподілу ймовірностей неперервної випадкової величини за критерієм Пірсона:

- *статистичні дані* (*результати вибірки*) *записують у вигляді інтервального статистичного розподілу*:

$(z_{i-1}, z_i]$	$(z_0, z_1]$	$(z_1, z_2]$	...	$(z_{m-1}, z_m]$
n_i	n_1	n_2	...	n_m

де n – обсяг вибірки, n_i *- число варіант вибірки, що попадають в інтервал* $(z_{i-1}, z_i]$;

- *оскільки перевіряється гіпотеза про те, що генеральна сукупність задовольняє певному* (*конкретному*) *закону розподілу з густиною* $p(x)$, *то для кожного інтервалу* $(z_{i-1}, z_i]$ *можна визначити теоретичні ймовірності* p_i *попадання значень випадкової величини в цей інтервал*;
- *для визначення теоретичних ймовірностей* p_i *використовуємо формули*:

$$p_i = P(z_{i-1} < Z \leq z_i) = F(z_i) - F(z_{i-1}), \tag{6}$$

$$p_i = \int_{z_{i-1}}^{z_i} f(x)dx; \tag{7}$$

причому $z_0 = -\infty, z_m = +\infty, \sum_{i=1}^{m} p_i = 1$.

- *одержані результати обчислень зручно записати у вигляді таблиці*:

$(z_{i-1}, z_i]$	$(-\infty, z_1]$	$(z_1, z_2]$	...	$(z_{m-1}, +\infty)$
n_i	n_1	n_2	...	n_m
p_i	p_1	p_2	...	p_m

- *обчислюють емпіричне значення критерію узгодження Пірсона*

$$\chi^2_{емп} = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}. \tag{8}$$

Випадкова величина Z має відомий розподіл „хі-квадрат” з $k = m - s - 1$ ступенями вільності, де m – число часткових інтервалів в інтервальному варіаційному ряді, s – число параметрів густини гіпотетичного розподілу;

- *за даним рівнем значущості α і кількістю $k = m - s - 1$ ступенів вільності знаходимо критичну точку $k_{кр} = \chi^2_{кр}(\alpha, k)$ за таблицею критичних значень розподілу χ^2 (див. „Додаток 5”);*
- *співставляємо значення $\chi^2_{емп}$ і $k_{кр}$: якщо $\chi^2_{емп} \geq k_{кр}$, то гіпотезу H_0 про вигляд густини розподілу відхиляють; якщо ж $\chi^2_{емп} < k_{кр}$, то гіпотезу H_0 приймають.*

ПРОГРАМНА РЕАЛІЗАЦІЯ

Для виконання завдання я використовував мову програмування Python, середовище Jupyter Notebook і бібліотеку matplotlib.pyplot для графіків і бібліотеку pandas для таблиць і модуль math для обчислення факторіалів.

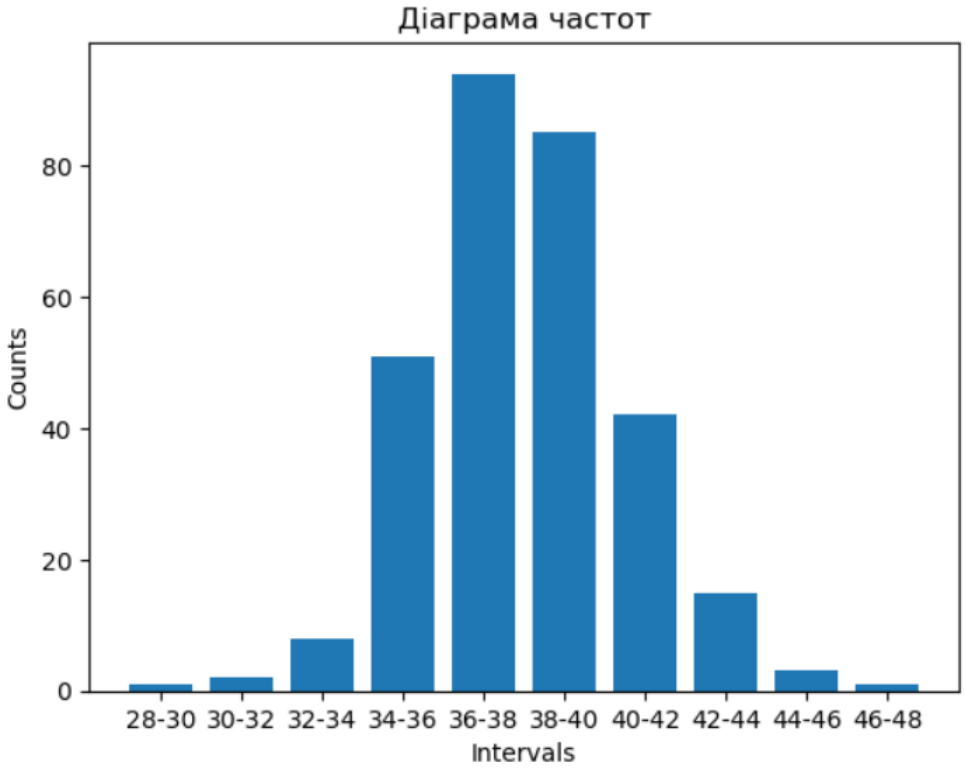
Завдання 1. Варіант 14

n_i(варіант 14)	1	2	8	51	94	85	42	15	3	1

Зчитуємо дані з файлу і знаходимо середину кожного інтервалу і відображаємо дані в таблиці.

X	(28, 30)	(30, 32)	(32, 34)	(34, 36)	(36, 38)	(38, 40)	(40, 42)	(42, 44)	(44, 46)	(46, 48)
ni	1	2	8	51	94	85	42	15	3	1
zi	29.0	31.0	33.0	35.0	37.0	39.0	41.0	43.0	45.0	47.0

З діаграми частот можна припустити що розподіл нормальний тому перевіряємо цю гіпотезу.



Функції get_varianca і get_standart для обчислення варіанси і стандарту вілповідно.

Функція get_p - шукає pi.

Функція union - об’єднює інтервали і ni і pi в яких не виконується умова ni>5 або npi>10.

Отримуємо потрібну вибірку

Знаходимо емпіричне значення за допомогою функції emp а також d.f за формолою. Беремо критичне значення з таблиці L_kr_table. І отримоємо результат що критичне значення більше імперичного, отже гіпотеза правильна

Завдання 2. Варіант 14

Завдання схоже до попереднього більше частина функцій або ідентична або трохи відозміненна, тому що тепер у нас розподіл не інтервальний а дискретний. Припустили що **закон розподілу біномний** тому прийшлось додатково шукати ймовірність p. Також повнісю перероблена функція get_p() адже змінився закон розподілу. Об’єднуємо наші значення в потрібну вибірку так щоб виконувалися умови ni>5 та npi>10. З уже правильною вибіркою щнаходимо емпіричне значення, d.f, і критичне

беремо з таблиці. Порівнюємо критичне й емпіричне. Критичне більше. Отже, гіпотеза правильна.

Отримані результати

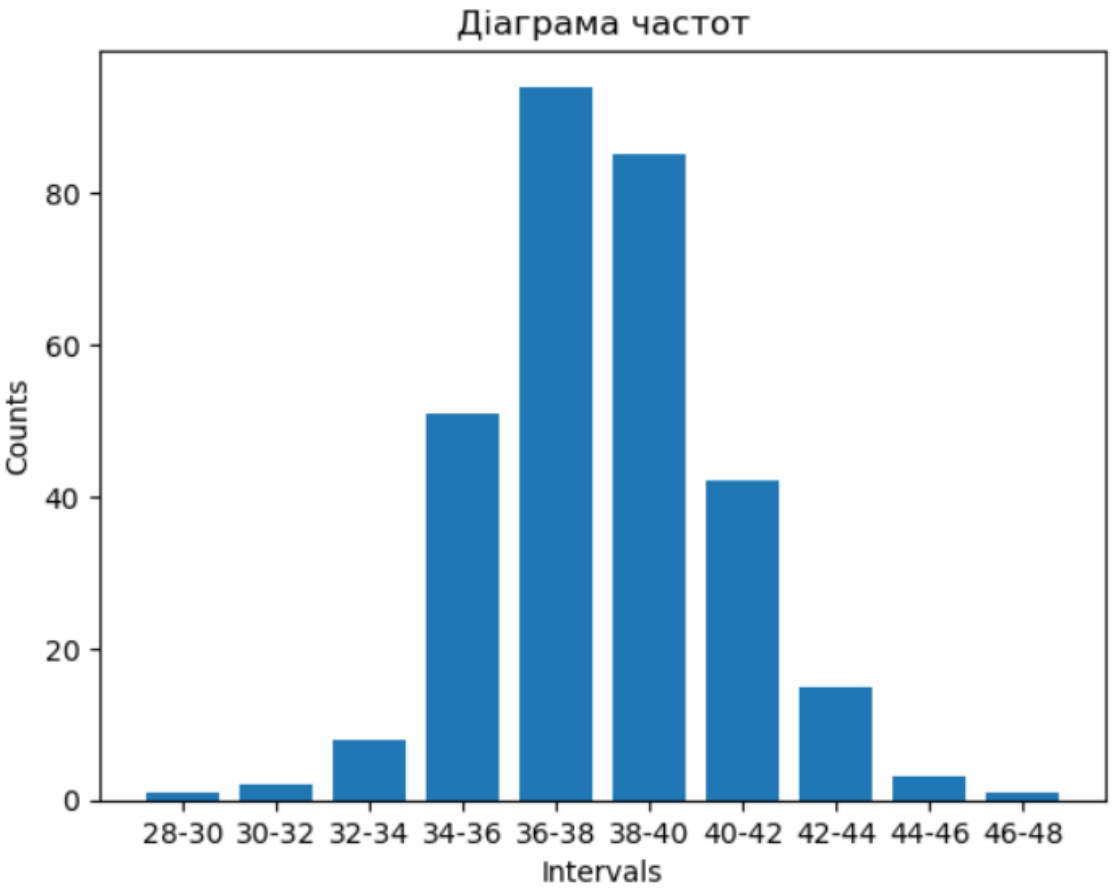
Варіант 14

Умова:

n_i(варіант 14)	1	2	8	51	94	85	42	15	3	1

Таблиця даних

X	(28, 30)	(30, 32)	(32, 34)	(34, 36)	(36, 38)	(38, 40)	(40, 42)	(42, 44)	(44, 46)	(46, 48)
ni	1	2	8	51	94	85	42	15	3	1
zi	29.0	31.0	33.0	35.0	37.0	39.0	41.0	43.0	45.0	47.0



З діаграми можна припустити, що розподіл нормальний.

Знайшов середнє арифметичне **x** = 38.01986 та стандарт **s** = `6.615499`

Знайшов рі

```
[0.0009, 0.0087, 0.0498, 0.1554, 0.2812, 0.2834, 0.16, 0.0507, 0.0089, 0.001]
```

Перевірив вибірку на виконання умов ni>5 та npi>10 та об’єднав інтервали де не виконується

X	(28, 34)	(34, 36)	(36, 38)	(38, 40)	(40, 42)	(42, 48)
ni	11	51	94	85	42	19
pi	0.0594	0.1554	0.2812	0.2834	0.16	0.0606

Емпіричне значення = `4.86444; d.f.= 3; критичне = 7.81.`

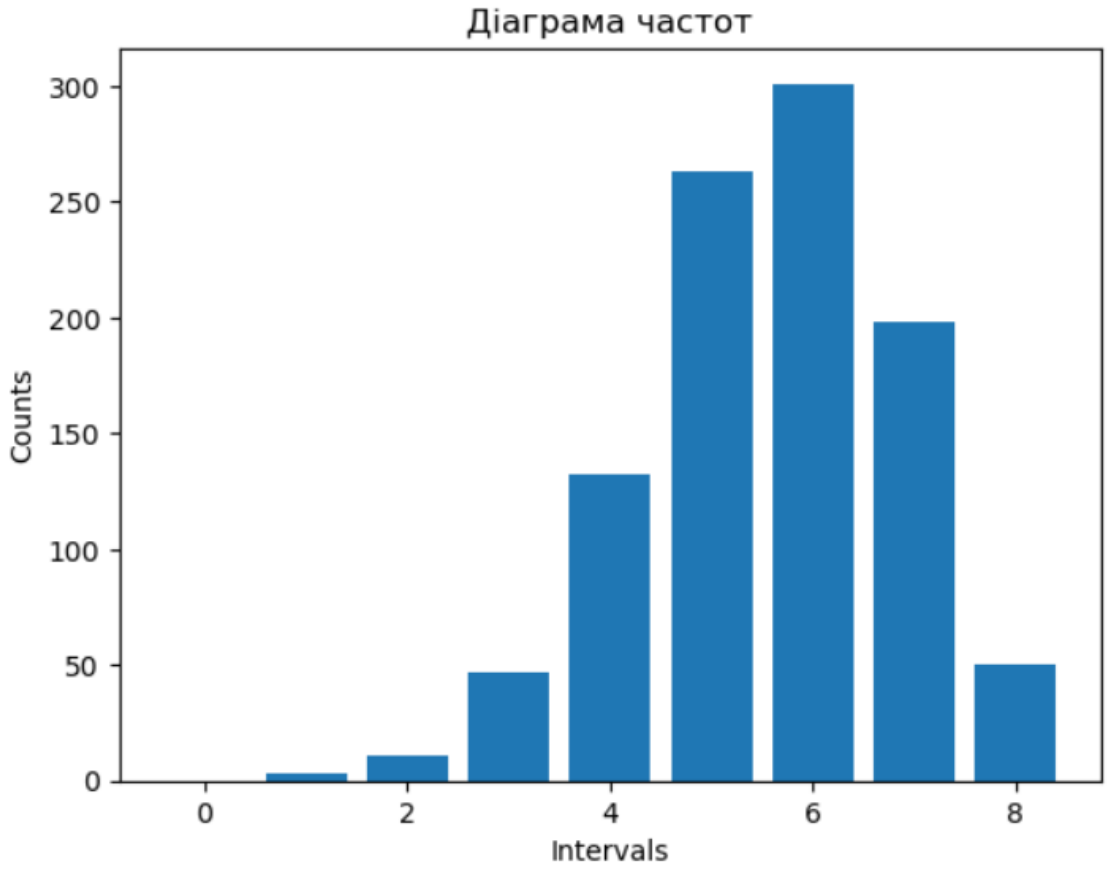
```
Емпіричне - 4.864, критичне - 7.81
Емпіричне 4.864 < критичного 7.81, тому H0 приймаємо
```

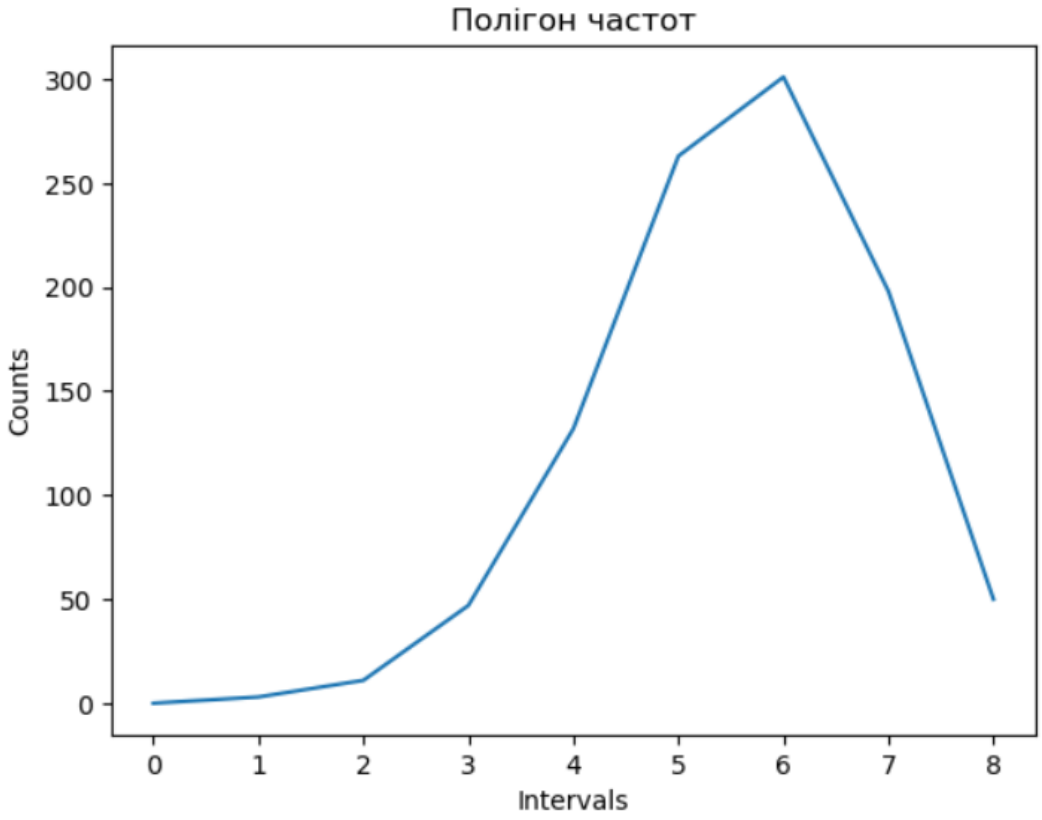
Завдання 2

Умова:

n_i(варіант 14)	0	3	11	47	132	263	301	198	50

X	0	1	2	3	4	5	6	7	8
ni	0	3	11	47	132	263	301	198	50





З діаграми і полігона частот можна пропустити що розподіл біномний.

Середнє значення дорівнює 5.57313

Ймовірність p = `0.69664`

q = 1 – p = `0.30336`

X	0.00000	1.00000	2.00000	3.00000	4.00000	5.00000	6.00000	7.00000	8.00000
ni	0.00000	3.00000	11.00000	47.00000	132.00000	263.00000	301.00000	198.00000	50.00000
pi	0.00007	0.00132	0.01059	0.04864	0.13962	0.25651	0.29453	0.19325	0.05547

Після об’єднання

ni	14.00000	47.00000	132.00000	263.00000	301.00000	198.00000	50.00000
pi	0.01198	0.04864	0.13962	0.25651	0.29453	0.19325	0.05547

Емпіричне значення = `1.74056`

`d.f = 5`

`Критичне дорівнює 11.07`

```
Емпіричне - 1.741, критичне - 11.07
Емпіричне 1.741 < критичного 11.07, тому H0 приймаємо
```

Висновок

Під час виконання цього індивідуального завдання номер два, я застосував знання здобуті на парах для перевірки гіпотези про нормальний закон розподілу. Також сам визначив якому закону підпорядковується розподіл другої задачі – біномний закон, та перевірив чи це правда. Зчитав дані з текстового файлу, побудував полігон і гістограму частот.