Лабораторна робота №11
Обробка й аналіз даних експерименту. Підбір параметрів розподілів

11.1. Мета роботи

Засвоєння прийомів підбору законів розподілів випадкових величин за експериментальними даними з використанням інструментів середовища системи Python.

11.2. Методичні рекомендації щодо організації самостійної роботи

За темою лабораторної роботи студент повинен: *знати* основні види законів розподілу випадкових величин; *вміти* розв'язувати задачу визначення параметрів розподілів шляхом перевірки статистичних гіпотез; *вміти* застосовувати процедури статистичних бібліотек Python для визначення закону розподілу випадкових величин по вибірках [3,11,18,19].

У деяких випадках імітаційна модель складної системи може бути реалізована у вигляді набору окремих моделей її підсистем. При проведенні експериментів з такою моделлю з метою скорочення витрат часу буває необхідно замінити моделювання роботи однієї з підсистем деяким числовим параметром (принцип параметризації), або випадковою величиною, розподіленою за заданим законом. Щоб така заміна була виконана коректно, дослідник повинен мати у своєму розпорядженні опис залежності даного числового параметра від часу й інших факторів, що фігурують у моделі.

При імітаційному моделюванні підбір законів розподілів виконується на основі статистичних даних, отриманих у ході експерименту.

В основі процедури відшукання закону розподілу деякої випадкової величини за експериментальними даними лежить перевірка статистичних гіпотез.

Перевірка гіпотези полягає в наступному. На підставі вибірки значень випадкової величини (даних експерименту) обчислюється z – часткове значення деякого критерію Z. Якщо $z > z_{кр}$, то від гіпотези H_0 відмовляються. Якщо $z \le z_{кр}$, то говорять, що отримані спостереження не суперечать прийнятій гіпотезі. Тут $z_{кр}$ – граничне (критичне) значення критерію Z, що відповідає обраному рівню значущості α, обумовлене

стандартним чином (за таблицею або через обернену функцію відповідного розподілу).

Зрозуміло, перш ніж висувати гіпотезу щодо значень параметрів розподілу, необхідно визначити вид самого закону розподілу. Найпоширеніший на практиці й досить ефективний метод підбору виду розподілу заснований на використанні графічного представлення експериментальних даних. Вони відображаються у вигляді так званої гістограми відносних частот, яка може бути побудована як вручну, так і за допомогою відповідних інструментальних засобів, що входять до складу більшості пакетів моделювання.

Для ефективного використання графічних засобів системи Python корисно знати **методику побудови гістограми** відносних частот.

Крок 1. Обчислюється величина інтервалу гістограми з наступного співвідношення: $h = (y_{max} - y_{min}) / m$, де (y_{min}, y_{max}) – діапазон зміни спостережуваної змінної *y*; *m* – число інтервалів, обраних дослідником.

Крок 2. За результатами (або в процесі) моделювання визначається число попадань значень *y* в *i*-й інтервал.

Крок 3. Обчислюється відносна частота попадань спостережуваної змінної в кожен інтервал:

$$g_i = \frac{n_i}{N},$$

де n_i – число попадань в *i*-й інтервал;

N – загальне число вимірів (обсяг вибірки).

Крок 4. На кожному *i*-му інтервалі будується прямокутник зі сторонами $h \times g_i$.

Сума площ прямокутників гістограми дорівнює одиниці.

Для найчастіше використовуваних статистичних гіпотез розроблені критерії, що дозволяють проводити їхню перевірку з найбільшою вірогідністю. Розглянемо основні з них.

t-критерій служить для перевірки гіпотези про рівність середніх значень двох нормально розподілених випадкових величин (*x* і *y)* у припущенні, що дисперсії їх рівні (хоча й невідомі). Порівнювані вибірки можуть мати різний обсяг (n_1 і n_2).

Як критерій використовують величину:

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1)Dx + (n_2 - 1)Dy}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}. \qquad (11.1)$$

Величина *T* підкоряється *t-розподілу* Стьюдента.

Критичне значення $t_{кр}$ для *t-критерію* визначається за таблицею для обраного значення α і числа степенів свободи $k = n_1 + n_2 - 2$.

Якщо обчислене за (11.1) значення задовольняє нерівності $T > t_{кр}$, то гіпотезу H_0 відкидають.

Стосовно припущення про "нормальну розподіленість" величин *x* і *y* *t-критерій* не дуже чутливий. Його можна застосовувати, якщо розподіли випадкової величини мають декілька вершин і не занадто асиметричні.

F-критерій служить для перевірки гіпотез про рівність дисперсій *Dx* і *Dy* за умови, що випадкові величини *x* і *y* розподілені нормально.

Гіпотези такого роду мають велике значення в техніці, тому що дисперсія є мірою таких характеристик, як погрішності вимірювальних приладів, точність біологічних процесів, точність наведення при стрілянині тощо.

Як контрольна величина використовується відношення дисперсій $F = Dx/Dy$ (або Dy/Dx – більша дисперсія повинна бути в чисельнику).

Величина *F* підкорюється *F-розподілу* (Фішера) з (m_1, m_2) степенями свободи $m_1 = n_1 - 1; \quad m_2 = n_2 - 1.$ Перевірка гіпотези полягає в наступному.

Для величини $a = \alpha / 2$ і величин m_1, m_2 за таблицею *F-розподілу* вибирають значення $F_{a,m1,m2}$. Якщо обчислене по вибірці *F* більше цього критичного значення, гіпотеза відхиляється з імовірністю помилки α.

Критерії згоди – це критерії, за допомогою яких перевіряють, чи задовольняє розглянута випадкова величина даному закону розподілу.

Критерій згоди Пірсона (χ^2) служить для перевірки гіпотези H_0 про те, що $F_y(y) = F_0(y)$, де $F_y(y)$ – істинний розподіл випадкової величини *y*; $F_0(y)$ – гіпотетичний розподіл.

Перевірка проводиться в такий спосіб:

1. Область значень випадкової величини *y* розбивається (довільно) на *m* непересічних множин ("класів").

2. У результаті *N* експериментів формується вибірка ($y_1, ..., y_N$).

3. Обчислюється контрольна величина χ^2:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - Np_i)^2}{Np_i}, \qquad (11.2)$$

де n_i – число значень *y,* що потрапили в *i*-й клас;

p_i – теоретична ймовірність (для $F_0(y)$) попадання значення *y* в *i*-й клас.

4. За таблицею χ^2-розподілів (або через обернену функцію χ^2-розподілу) знаходять критичне значення χ^2_α для рівня значущості α і $k = m - 1$ степенів свободи. Якщо $\chi^2 > \chi^2_\alpha$, то гіпотеза відкидається.

Зрозуміло, що проведення вручну розрахунків, необхідних для перевірки статистичних гіпотез, вимагає значних витрат часу й сил. Тому багато сучасних математичних пакетів, у тому числі й Python, мають у своєму складі засоби, що дозволяють звести до мінімуму число операцій, виконуваних користувачем уручну.

11.3. Контрольні приклади

Приклад 1. Згенерувати вибірку для випадкової величини, що має нормальний закон розподілу з параметрами: середнє значення дорівнює 3, середнє квадратичне відхилення дорівнює 0.5.

Розв'язання. Виконаємо завдання з використанням пакета **scipy.stats** для Python.

Згенеруємо вибірку V із 100 значень випадкової величини, розподіленої за нормальним законом з параметрами: середнє значення дорівнює 3, середнє квадратичне відхилення дорівнює 0.5:

```
mean = 3.
sigma = 0.5
N = 100
V = stats.norm.rvs(loc=mean, scale=sigma, size=N)
print(V)
```

Output:

```
[2.95871616 3.30905038 3.4252837  2.51502608 1.99064237 3.37200886
 2.50236962 2.98673632 4.07483731 3.19627856 3.28655936 2.83547691
 3.02222143 2.93000184 2.11635738 4.35610835 3.0967367  2.82969193
 3.57489687 2.85328495 3.56821778 3.26420056 2.68754404 3.49462342
 2.84656877 4.19349483 2.95368716 2.46628272 3.13849198 3.09422442
 3.69871039 3.88584888 3.84265356 3.06393752 2.67086956 3.14971549
 3.17499643 2.46782579 3.19008636 3.32452099 2.71549924 3.93675511
 3.33222228 3.84575518 3.05972284 2.26491033 3.16377575 3.24844685
 3.3428852  3.41415907 3.43961845 3.29863015 2.83581087 3.5107348
 3.02135197 2.59530316 3.41371815 2.42866702 4.0318818  3.28888749
 3.24381384 3.37089659 3.481783   2.68610659 2.89611344 3.50990139
 2.83869079 3.37832148 2.59691856 3.06626006 3.08754722 2.24073909
 3.43929512 3.2531051  2.0512648  3.04830235 2.28121672 3.57713521
 3.20028876 3.26373642 3.17020428 2.53243073 3.77550332 2.01761573
 3.42924422 2.61848476 3.80206359 3.23140396 2.78068801 3.76870636
```

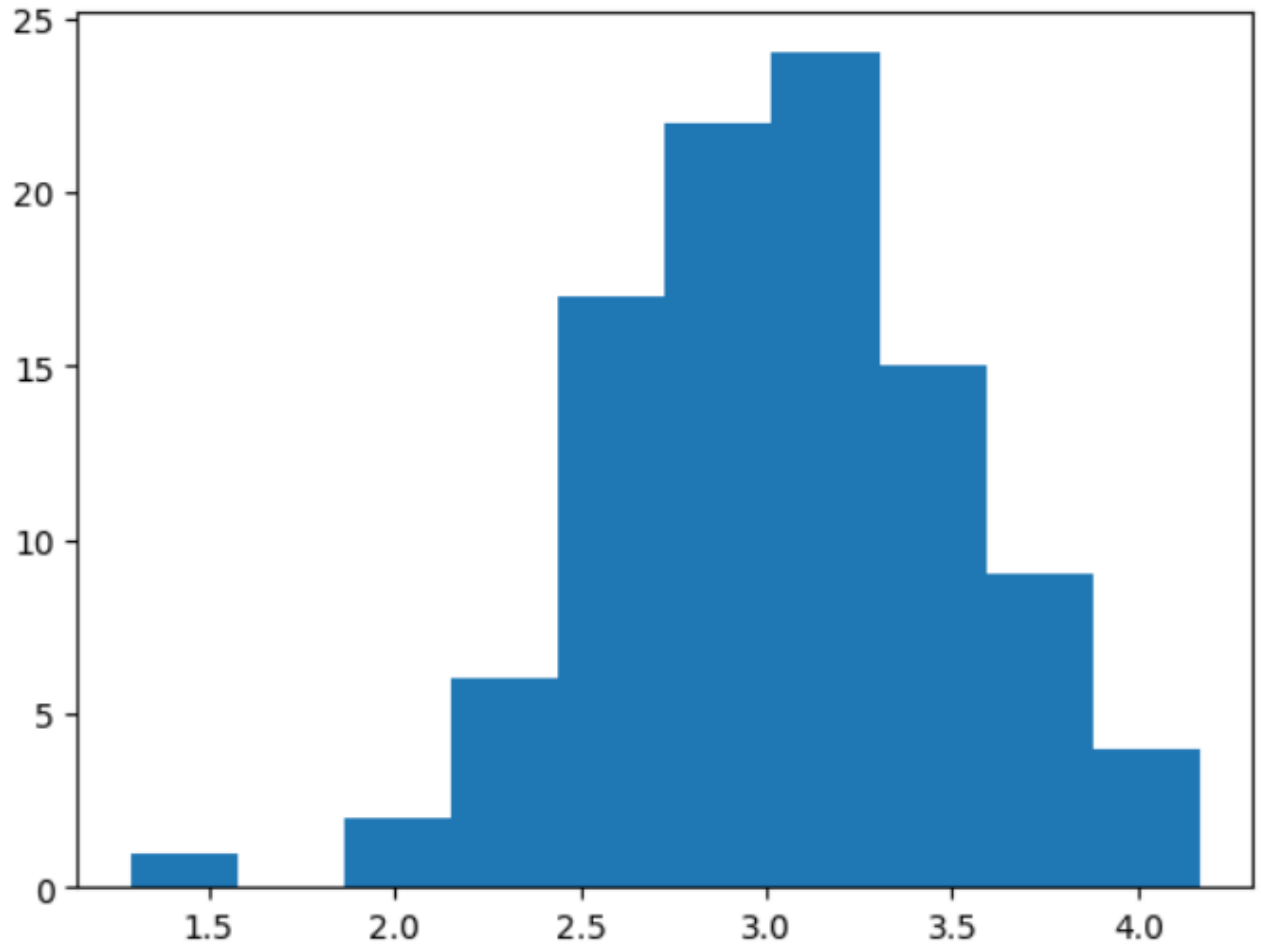
```
1.63439811 2.83817393 2.8430919  2.96887633 2.52985298 2.71722252
3.14641905 2.7567052  3.23464391 3.1118396 ]
```

Приклад 2. Використовуючи згенеровану вибірку з прикладу 1 як вихідні дані експерименту, оцінити вид і параметри закону розподілу відповідної випадкової величини.

Розв'язання. Побудуємо гістограму для вибірки V:

```
import matplotlib.pyplot as plt

plt.hist(V)
```



Тут *hist* – процедура бібліотеки **matplotlib.pyplot**, що підраховує частоту попадання елементів вибірки V в кожний з інтервалів, на які поділена область побудови гістограм від мінімального (min(V)) до максимального (max(V)) значення, та будує стовпцевий графік.

З вигляду гістограми можна зробити висновок, що закон розподілу можливо нормальний. Знайдемо оцінки параметрів цього розподілу. Оскільки передбачається, що закон розподілу є нормальним, а параметрами нормального закону є математичне очікування та середнє

квадратичне відхилення, то знайдемо їх оцінки: вибіркове середнє *Vmean* та вибіркове середнє квадратичне відхилення *Vsd.*

```
import math

Vmean = sum(V)/N
print("Vmean =", Vmean)

S = 0.
for i in range(N):
  S = S + (V[i] - Vmean)**2
Vdisp = S/(N-1)
print("Vdisp =", Vdisp)

Vsd = math.sqrt(Vdisp)
print("Vsd =", Vsd)
```

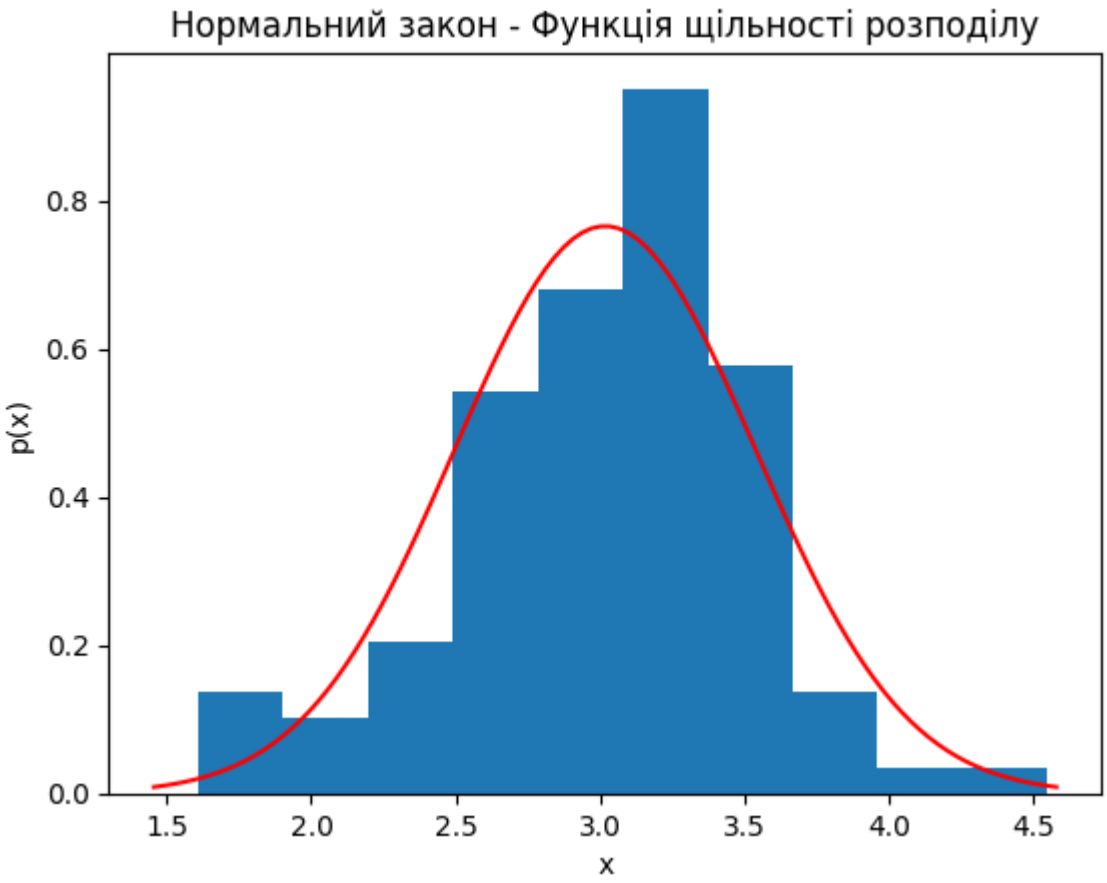
Output:

```
Vmean = 3.0901652643385855
Vdisp = 0.2551927782875111
Vsd = 0.5051660898036517
```

Надалі висуваємо гіпотезу: закон розподілу випадкової величини, для якої в нас є вибірка V, є нормальним з параметрами *a=Vmean, σ=Vsd.*

Спочатку виконуємо візуалізацію цієї гіпотези, тобто порівняємо гістограму з функцією щільності розподілу нормального закону:

```
# Візуалізація гіпотези
plt.hist(V, density=True)
x = np.linspace(Vmean-3*Vsd, Vmean+3*Vsd, 100)
y = stats.norm(Vmean, Vsd).pdf(x)
plt.plot(x, y, color='r')
plt.title('Нормальний закон - Функція щільності розподілу')
plt.xlabel('x')
plt.ylabel('p(x)')
```



Перевіряємо висунуту гіпотезу за критерієм Пірсона з рівнем значущості $\alpha = 0.95$.

```
      import scipy.stats as stats

      m = 10
      int = np.linspace(min(V), max(V), (m+1))
      print(int)

      n = [0]*m
      p = [0]*m
      for i in range(m):
        p[i] = stats.norm.cdf(int[i+1], loc=Vmean, scale=Vsd) -
stats.norm.cdf(int[i], loc=Vmean, scale=Vsd)
        n[i] = 0
        for j in range(N):
          if ((V[j] >= int[i]) and (V[j] < int[i+1])):
            n[i] = n[i] + 1
      print(p)
      print(n)

      # Обчислення значення критерія Пірсона
      HiSq = 0.
      for i in range(m):
        HiSq = HiSq + (n[i] - N*p[i])**2/(N*p[i])
      print(HiSq)

      alfa = 0.95
```

```
    # Обчислення критичного значення для рівня значущості alfa і степенів свободи (m-1)
    HiSqAlfa = stats.chi2.ppf(alfa, m-1)
    print(HiSqAlfa)
```

Output:

```
    [1.63439811 1.90656914 2.17874016 2.45091118 2.72308221 2.99525323
     3.26742425 3.53959528 3.8117663  4.08393732 4.35610835]
    [0.007587790958202869, 0.026034191533313128, 0.06725870032464859,
0.13085970665008972, 0.19176655430178796, 0.2116812471392061,
0.17601236179893454, 0.11023996507917988, 0.05200315957494139,
0.018473626832632983]
    [1, 4, 4, 14, 16, 26, 20, 7, 6, 1]

    HiSq = 5.931014753244127
    HiSqAlfa = 16.918977604620448
```

Тут p_i – теоретична ймовірність попадання значення випадкової величини, розподіленої за нормальним законом з параметрами *a=Vmean, σ=Vsd,* в кожний з *m* інтервалів $(\text{int}_i, \text{int}_{i+1})$ на які поділена область від min(V) до max(V), n_i – частота попадання елементів вибірки в ті ж інтервали $(\text{int}_i, \text{int}_{i+1})$, *punif* – функція розподілу для нормального закону з пакета R, *qchisq* – функція, обернена до функції χ^2-розподілу, *HiSq* – значення критерію Пірсона, *HiSqAlfa* – критичне значення для рівня значущості α і $(m-1)$ степенів свободи.

Оскільки *HiSq* $\leq$ *HiSqAlfa*, то висунута гіпотеза не відкидається. Таким чином, можна з імовірністю 95% вважати, що випадкова величина, для якої в нас є вибірка, розподілена за нормальним законом з параметрами: середнє значення дорівнює 3.0901, середнє квадратичне відхилення дорівнює 0.505.

11.4. Порядок виконання роботи і варіанти завдань

11.4.1. Зміст звіту

У практичній частині роботи необхідно:

- виконати приклади, що наведений в розділі 11.3;
- виконати завдання наведене нижче;
- привести тексти складених програм та результати їх виконання.

11.4.2. Варіанти індивідуальних завдань

1. Згенерувати вибірки для 3-х випадкових величин, що мають наступні закони розподілу: нормальний, рівномірний, пуасонівський із заданими параметрами розподілу (табл. 11.1).

2. Використовуючи згенеровані вибірки як вихідні дані експерименту, визначити вид і параметри закону розподілу для відповідних випадкових величин.

Таблиця 11.1

Варіанти завдань

Варіант	Нормальний	Рівномірний	Пуасонівський
1	μ=10, σ=0.1	a=5, b=15	λ=5
2	μ=11, σ=0.2	a=4, b=14	λ=4
3	μ=12, σ=0.3	a=6, b=13	λ=7
4	μ=13, σ=0.4	a=7, b=14	λ=8
5	μ=14, σ=0.15	a=8, b=15	λ=9
6	μ=15, σ=0.25	a=3, b=12	λ=10
7	μ=16, σ=0.2	a=9, b=17	λ=11
8	μ=17, σ=0.21	a=10, b=20	λ=12
9	μ=18, σ=0.3	a=11, b=22	λ=13
10	μ=19, σ=0.5	a=12, b=23	λ=14
11	μ=20, σ=0.45	a=13, b=25	λ=15
12	μ=21, σ=0.6	a=14, b=26	λ=16
13	μ=22, σ=0.65	a=15, b=25	λ=17
14	μ=9, σ=0.26	a=4, b=10	λ=6
15	μ=23, σ=0.5	a=16, b=27	λ=18

11.5. Контрольні запитання

1. Який закон розподілу випадкової величини називається нормальним?
2. Який закон розподілу випадкової величини називається рівномірним?
3. Який закон розподілу випадкової величини називається експоненціальним?
4. Який закон розподілу випадкової величини називається пуасонівським?
5. Опишіть методику побудови гістограми відносних частот для вибірки.
6. Для чого використовується критерій згоди Пірсона? В чому його суть?