**Міністерство освіти і науки України**

**Національний технічний університет України «КПІ» імені Ігоря Сікорського**

**Кафедра обчислювальної техніки ФІОТ**

**ЗВІТ**

**з лабораторної роботи №1**

**з навчальної дисципліни «Вступ до технології Data Science»**

**Тема:**

**ДОСЛІДЖЕННЯ СТАТИСТИЧНИХ ХАРАКТЕРИСТИК   
ЕКСПЕРИМЕНТАЛЬНИХ ДАНИХ**

**Виконав:**

Студент X курсу кафедри ОТ ФІОТ,

Навчальної групи ІА-73

Петренко П.П.

**Перевірив:**

Професор кафедри ОТ ФІОТ

Писарчук О.О.

**Київ 2021**

**І. Мета:**

виявити дослідити та узагальнити особливості застосування методів статистичного аналізу для задач визначення статистичних характеристик вхідного потоку експериментальних даних з використанням спеціалізованих пакетів мови програмування Python.

**ІІ. Завдання:**

Лабораторія провідної ІТ-компанії реалізує масштабний проект розробки універсальної платформи з обробки Big Data масиву експериментальних даних поточного спостереження для виявлення закономірностей і прогнозування розвитку контрольованого процесу. Платформа передбачає розташування back-end компоненти на власному хмарному сервері з наданням повноважень користувачам заздалегідь адаптованого front-end функціоналу універсальної платформи.

Замовниками ресурсів платформи є: державні та комерційні компанії валютного трейдінгу для прогнозування динаміки зміни курсу валют та ціни інших товарів; метеорологічні служби для прогнозування параметрів метеоумов; департаменти охорони здоров’я для прогнозування зміни показників епідеміологічних ситуацій.

Вам, як Data Science [Engineer](https://jobs.dou.ua/companies/gravitum/vacancies/147764/) поставлене наступне завдання.

**І. Розробити універсальний скрипт мовою Python що реалізує:**

Технічні умови реалізації завдання наведені у таблиці Д.1 додатку 1.

| **Варіант** (порядковий номер в списку групи) | **І рівень складності** | **ІІ рівень складності** |
| --- | --- | --- |
| 6, 21 | Закон зміни похибки – експонентційний;  Закон зміни досліджуваного процесу – лінійний. | Закон зміни похибки – експонентційний, нормальний;  Закон зміни досліджуваного процесу – лінійний, квадратичний. |

1. Модель генерації випадкової величини – похибки вимірювання за заданим у таблиці Д1 додатку 1 закону розподілу

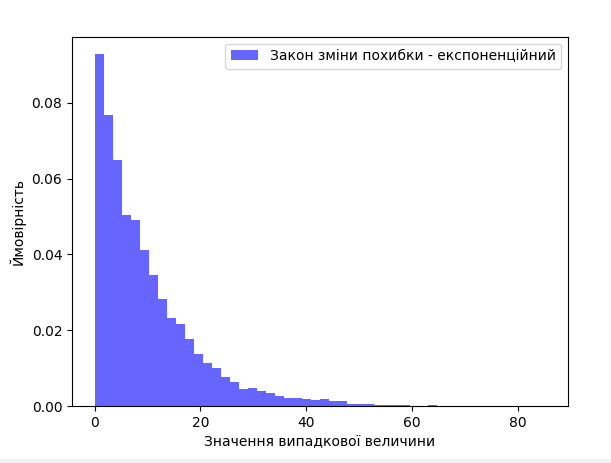


Рис. 1 Розподіл експоненційного закону похибки ( експоненційний розподіл )

2. Модель зміни досліджуваного процесу за заданим у таблиці Д1 додатку 1 закону;

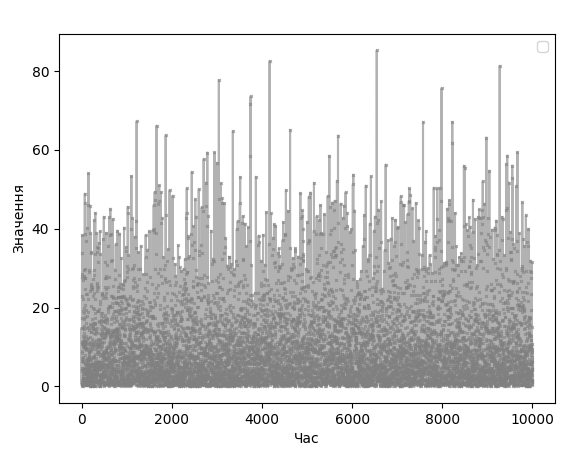


Рис.2 Графік значень моделі зміни досліджуваного процесу

3. Адитивну модель експериментальних даних (вимірів досліджуваного процесу) відповідно до синтезованих в п.1,2 моделей випадкової (стохастична) і невипадкової складових.

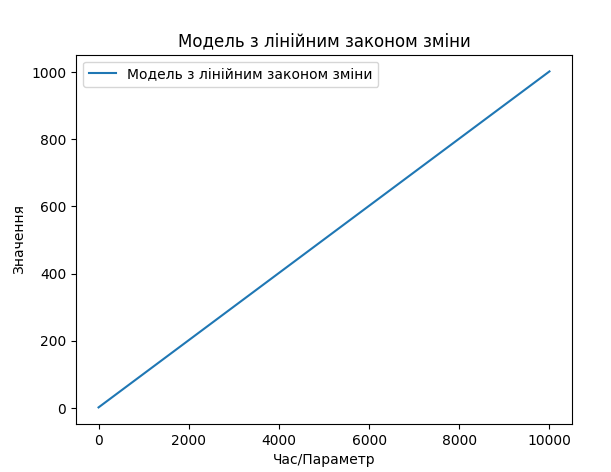


Рис.3 Модель лінійного закону зміни

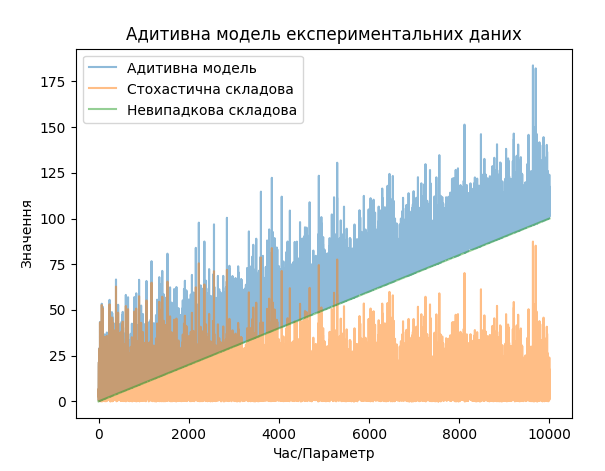


Рис.4 Адитивна модель

4. Метод Монте-Карло для дослідження статистичних характеристик експериментальних даних, сформованих у п.3;

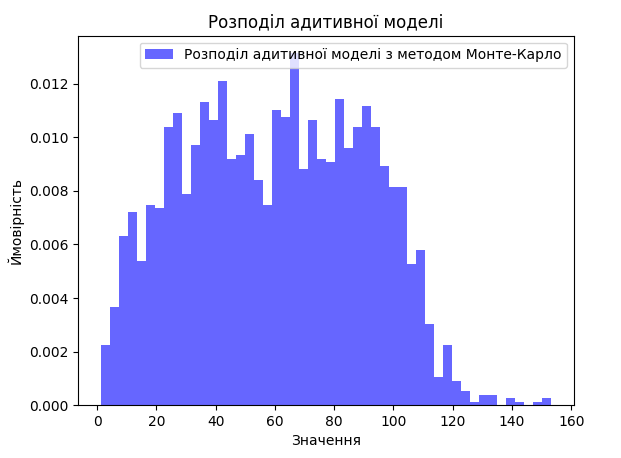
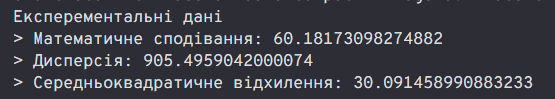


Рис.5 Розподіл адитивної моделі методом монте-карло



5. Визначення статистичних (числових) характеристик експериментальних даних (дисперсію, середньоквадратичне відхилення математичне сподівання, гістограми закону розподілу похибки та експериментальних даних).

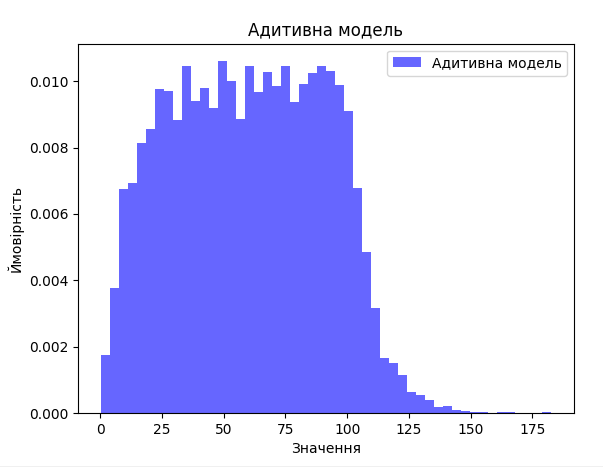


Рис. 6. Адитивна модель

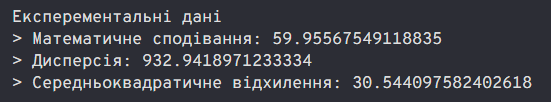


Рис. 7. Статистичні дані адитивної моделі

**ІІ. Провести дослідження зміни статистичних характеристик сформованих моделей при зміні стохастичних характеристик помилки експериментальних даних.**

Обидві моделі мають близькі характеристики, але модель з методом Монте-Карло дає трохи більше значення середнього і стандартного відхилення. Це стається через випадковий характер методу Монте-Карло, який використовує випадкові вибірки для оцінки статистичних показників.

Зміни в параметрах стохастичної складової, таких як середній інтервал між подіями (1/λ) у випадку експоненційного розподілу, можуть призвести до зміни ступеня коливань в експериментальних даних. Якщо значення параметра λ зростає, то коливання в даних стають менш виразними, тобто дані стають менш змінними з часом. У випадку зменшення значення параметра λ, коливання в даних стають більш видимими, і дані стають більш змінними з часом.

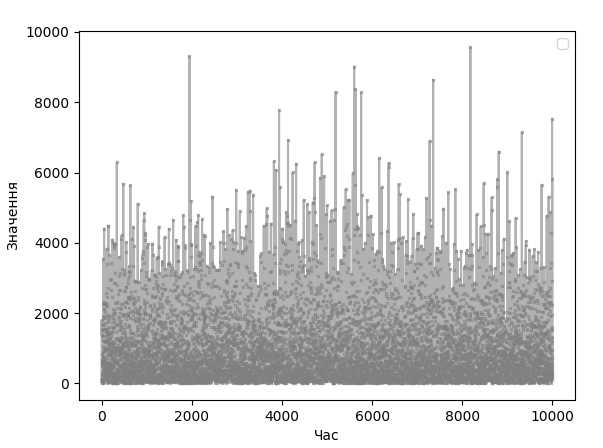


Рис.8 Модель зміни

λ = 0.005 Середньоквадратичне відхилення:288.75 (приблизно)

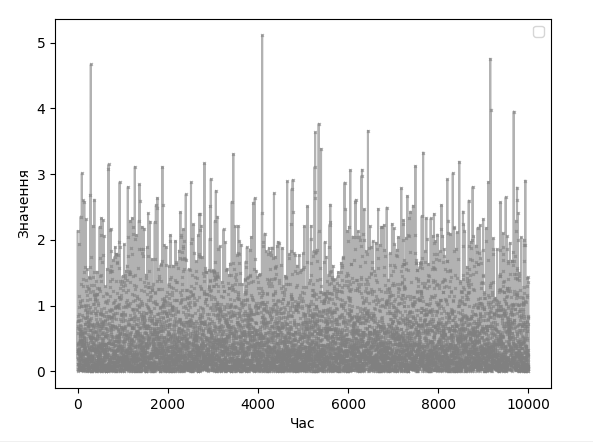


Рис. 9 Модель зміни за значення λ = 2

λ = 2.1 Середньоквадратичне відхилення: 288.89 (приблизно)

Розподіл стохастичної складової також може вплинути на форму розподілу експериментальних даних. Наприклад, експоненційний розподіл має експоненційну форму, яка може вплинути на форму розподілу даних.

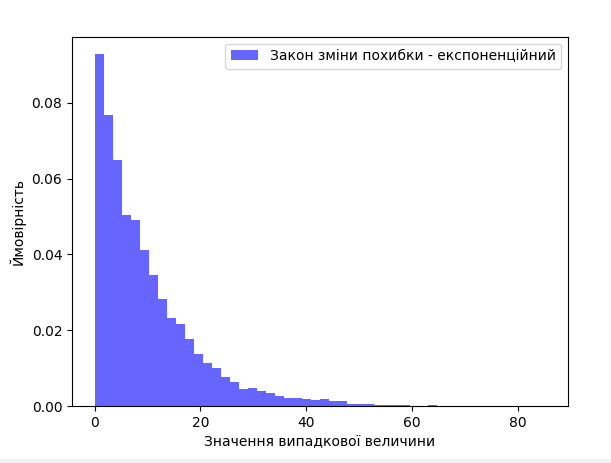


Рис. 10 Експоненційний розподіл

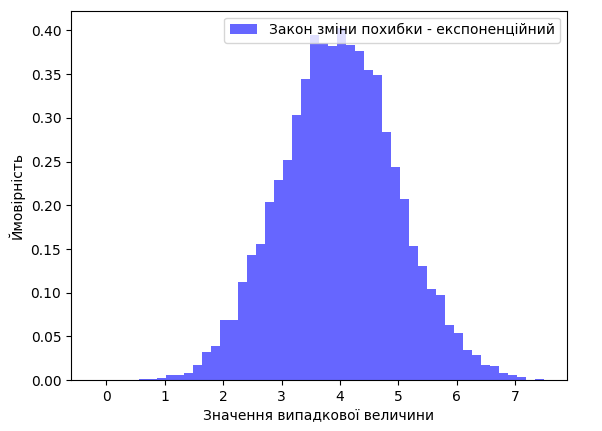


Рис 11. Нормальний розподіл

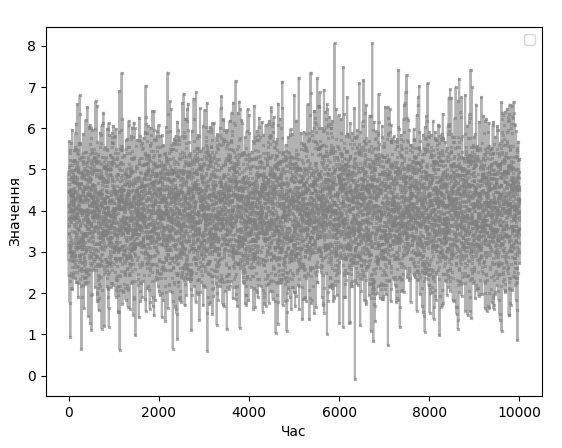


Рис 12. Модель зміни нормального розподілу

Додавання шуму: Стохастична складова вводить випадковий шум в дані, що може призвести до варіацій у значеннях даних. Цей шум може зробити дані менш передбачуваними.

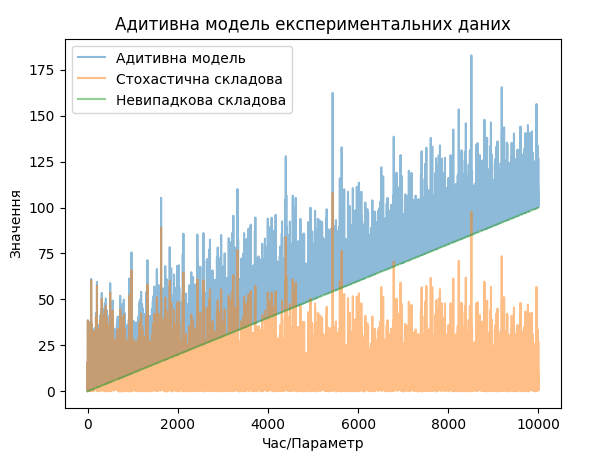


Рис 13. Демонстрація шуму у даних з невипадковою складовою.

Загалом, стохастичні характеристики додають елемент випадковості та непередбачуваності до моделі та даних.

**ІІІ. Довести адекватність сформованих моделей та працездатність розробленого скріпта.**

**Розроблений код повинен бути раціональним та відповідати вимогам до чистого коду.**

**import pandas as pd**

**import numpy as np**

**import matplotlib.pyplot as plt**

**def exp\_d(size=100000, lambda\_val=0.1):**

**return np.random.exponential(scale=1.0 / lambda\_val, size=size)**

**def histogram(data, bins=50, label="", xlabel="", ylabel="", title=""):**

**plt.hist(data, bins=bins, density=True, alpha=0.6, color='b', label=label)**

**plt.xlabel(xlabel)**

**plt.ylabel(ylabel)**

**plt.title(title)**

**plt.legend()**

**plt.show()**

**def plot(data, label="", xlabel="", ylabel="", title=""):**

**plt.plot(data, color="gray", alpha=0.6, marker="x", markersize=2)**

**plt.xlabel(xlabel)**

**plt.ylabel(ylabel)**

**plt.title(title)**

**plt.legend()**

**plt.show()**

**def linear\_law(size=1000, slope=0.1, intercept=2, show=False):**

**data = np.arange(size) \* slope + intercept**

**if show:**

**plt.plot(data, label='Модель з лінійним законом зміни')**

**plt.xlabel('Час/Параметр')**

**plt.ylabel('Значення')**

**plt.title('Модель з лінійним законом зміни')**

**plt.legend()**

**plt.show()**

**return data**

**def constantionous\_law(size=1000, show=False):**

**data = np.arange(size) \* 0.01**

**if show:**

**plt.plot(data, label='Модель з постійним законом зміни')**

**plt.xlabel('Час/Параметр')**

**plt.ylabel('Значення')**

**plt.title('Модель з постійним законом зміни')**

**plt.legend()**

**plt.show()**

**return data**

**def additive\_model(data, show=False):**

**lambda\_val = 0.1**

**size = len(data)**

**stochastic\_component = exp\_d(size, lambda\_val)**

**deterministic\_component = data**

**experimental\_data = stochastic\_component + deterministic\_component**

**if show:**

**plt.plot(experimental\_data, label='Адитивна модель', alpha=0.5)**

**plt.plot(stochastic\_component, label='Стохастична складова', alpha=0.5)**

**plt.plot(deterministic\_component, label='Невипадкова складова', alpha=0.5)**

**plt.xlabel('Час/Параметр')**

**plt.ylabel('Значення')**

**plt.title('Адитивна модель експериментальних даних')**

**plt.legend()**

**plt.show()**

**return experimental\_data**

**def monte\_carlo\_method(data, num\_samples=10000):**

**return np.random.choice(data, size=num\_samples, replace=True)**

**def variance(data):**

**return np.var(data)**

**def mean(data):**

**return np.mean(data)**

**def deviation(data):**

**return (data - np.mean(data))**

**def mean\_squared\_deviations(data):**

**return np.sqrt(np.mean(deviation(data) \*\* 2))**

**def overall\_analysis(data, label="", xlabel="", ylabel="", title=""):**

**print(f" Експерементальні дані \n"**

**+ f" > Математичне сподівання: {mean(data)}\n"**

**+ f" > Дисперсія: {variance(data)}\n"**

**+ f" > Середньоквадратичне відхилення: {mean\_squared\_deviations(data)}\n"**

**)**

**histogram(data, label=label, xlabel=xlabel, ylabel=ylabel, title=title)**

**def normal\_d(size=1000, mean=0, std\_dev=1):**

**return np.random.normal(loc=mean, scale=std\_dev, size=size)**

**if \_\_name\_\_ == '\_\_main\_\_':**

**size, lambda\_ = 10000, 2**

**exponentinal\_distribution\_model = exp\_d(size, lambda\_)**

**histogram(exponentinal\_distribution\_model, 50, 'Закон зміни похибки - експоненційний', 'Значення випадкової величини', 'Ймовірність')**

**plot(exponentinal\_distribution\_model, label='Модель зміни досліджуваного процесу експоненційного розподілу', xlabel='Час', ylabel='Значення')**

**additive\_model\_ = additive\_model(constantionous\_law(size, show=True), show=True)**

**overall\_analysis(monte\_carlo\_method(additive\_model\_, len(additive\_model\_)//4), 'Розподіл адитивної моделі з методом Монте-Карло', 'Значення', 'Ймовірність', 'Розподіл адитивної моделі')**

**overall\_analysis(additive\_model\_, 'Адитивна модель', 'Значення', 'Ймовірність', 'Адитивна модель')**

**IV. Висновки.**

виявив, дослідив та узагальнив особливості застосування методів статистичного аналізу для задач визначення статистичних характеристик вхідного потоку експериментальних даних з використанням спеціалізованих пакетів мови програмування Python.

Виконав: студент Петренко П.П.