**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 «Технології Data Science»**

**Тема:**

**ПІДГОТОВКА ТА АНАЛІЗ ДАНИХ ДЛЯ СТАТИСТИЧНОГО НАВЧАННЯ**

**Виконав:**

Студент 4 курсу кафедри ФІОТ,

Навчальної групи ІП-11

Олександр Головня

**Перевірив:**

Професор кафедри ОТ ФІОТ Олексій Писарчук

**Київ 2024**

**І. Мета:**

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

**ІІ. Завдання:**

|  |  |
| --- | --- |
| **Варіант**  (порядковий номер в списку  групи) | **ІІ рівень складності 8 балів** |
| 5 | Закон зміни похибки – нормальний, рівномірний;  Закон зміни досліджуваного процесу (тренду) – лінійний, квадратичний.  Комбінаторика похибка / тренд – довільна.  Реальні дані – 3 показники. |

**Завдання ІІ рівня – максимально 8 балів.**

Розробити програмний скрипт мовою Python що забезпечує аналіз властивостей і характеристик вихідних даних відповідно до етапів:

1. Модель генерації випадкової величини за заданим у табл.1 додатку 1 закону розподілу;
2. Модель зміни (ідеальний тренд) досліджуваного процесу за заданим у табл.1 додатку 1 законом;
3. Адитивна модель статистичної вибірки відповідно до синтезованих в п.1,2 моделей випадкової (стохастична) і невипадкової складових. Параметри закону розподілу та закону зміни досліджуваного процесу обрати самостійно.
4. Визначення статистичних (числових) характеристик сформованих в п.1,3 вибірок (дисперсія, середньоквадратичне відхилення, математичне очікування, гістограма

закону розподілу).

1. Визначення статистичних характеристик реальних даних, заданих файлом Oschadbank (USD).xls за умов табл. 1 додатку 1.
2. Провести аналіз отриманих результатів та верифікацію розробленого скрипта.

**ІІІ. Результати виконання лабораторної роботи.**

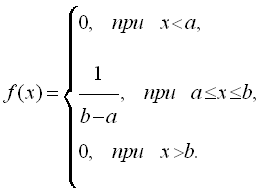
* 1. Синтезована математична модель перетворень графічних об’єктів відповідно до індивідуального завдання.

Синтезована математична модель в даній лабораторній роботі включає в себе дві основні компоненти:

1. Модель генерації випадкової величини: Ця модель використовується для створення випадкових чисел згідно з вказаними законами розподілу, такими як нормальний та експоненційний закони. Вона служить для створення помилок або шуму, які додаються до ідеального тренду.

**Рівномірний закон розподілу:**

1. ***Щільність розподілу ймовірностей*:**

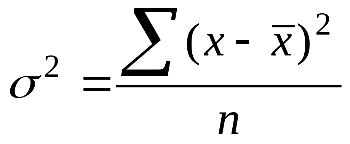
****

де: 𝑥 – випадкова величина; 𝑎, 𝑏 – межі реалізації ВВ, параметри закону розподілу ВВ.

1. ***Числові характеристики:***

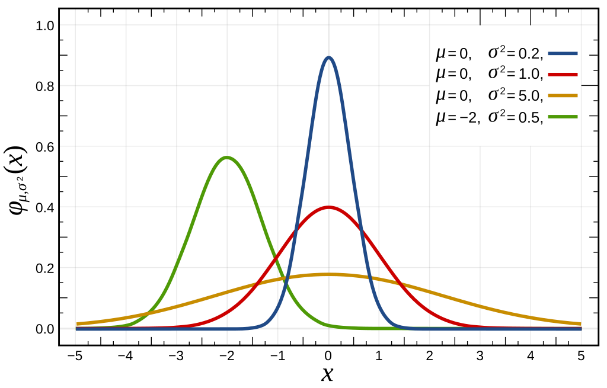
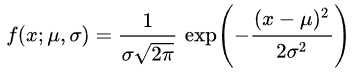
Математичне сподівання:

Дисперсія:



**Нормальний закон розподілу:**

1. ***Щільність розподілу ймовірностей*:**

де � — математичне сподівання, �2 — дисперсія випадкової величини. Параметр � також відомий, як стандартне відхилення.

1. ***Числові характеристики:***

Математичне сподівання:

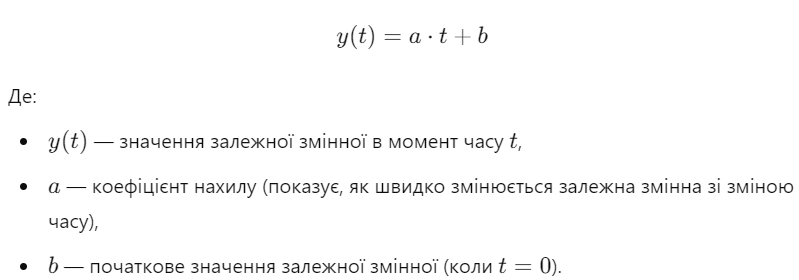


Дисперсія:



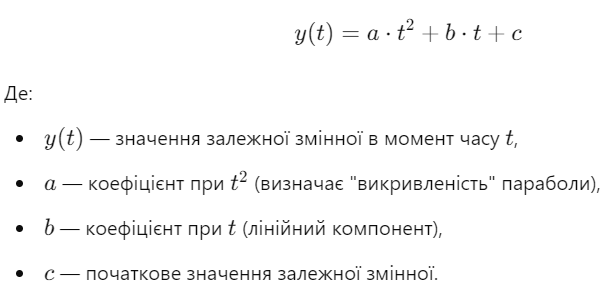
1. Модель ідеального тренду: Ця модель представляє собою ідеальний сценарій або закон зміни досліджуваного процесу. У лабораторній роботі реалізовані два види трендів: постійний (константний) та квадратичний. Модель ідеального тренду використовується для створення основної залежності вихідних даних.

Формула лінійної моделі виглядає так:



Ця модель описує пряму лінію, де зміна значення відбувається з постійною швидкістю.

Формула квадратичної моделі виглядає так:



Квадратична модель описує параболічну залежність, тобто процес з прискоренням або уповільненням.

Ці моделі можна використовувати для різних типів трендів у залежності від характеру змін досліджуваного процесу.

***Методика визначення статистичних характеристик стохастичних даних.***

1. Виділення систематичної складової (тренду).
2. Обчислення середнього значення (математичного очікування)
3. Обчислення Дисперсії. Дисперсія показує, наскільки дані розкидані відносно середнього значення. Чим більша дисперсія, тим більше варіативність даних.
4. Середньоквадратичне відхилення. Це квадратний корінь із дисперсії. Воно показує, наскільки значення в середньому відхиляються від математичного очікування.
5. Коефіцієнт асиметрії (скос). Асиметрія вимірює ступінь "нерівності" розподілу даних. Якщо асиметрія дорівнює нулю, то розподіл симетричний.
6. Гістограма — це графічне представлення розподілу даних, яке показує, як часто зустрічаються певні значення. Її можна використовувати для оцінки форми розподілу й визначення, чи відповідає він, наприклад, нормальному розподілу.

Визначення статистичних характеристик стохастичних даних дозволяє зробити висновки про властивості процесу та його випадкову природу. Такий аналіз допомагає краще зрозуміти поведінку даних та прогнозувати їх у майбутньому.

* 1. Блок схема алгоритму та її опис.

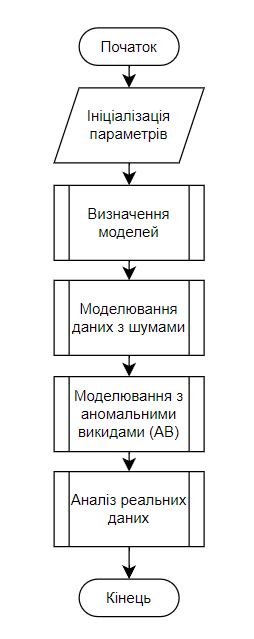


Рис.1 - Блок-схема алгоритму програми

Опис алгоритму:

Початок програми.

1. Ініціалізація параметрів:

Визначення обсягу вибірки (кількість даних) n. Задаються значення констант: обсяг вибірки, кількість реалізацій, коефіцієнт аномальних викидів, параметри нормального розподілу тощо.

1. Визначення моделей:

Викликаються функції моделей

Модель тренду:

* Квадратична модель.
* Лінійна модель.

Шуми:

* Нормальні помилки.
* Рівномірні помилки.

1. Моделювання даних з шумами:

Створюються моделі з трендом та шумами:

* Нормальні помилки + тренд.
* Рівномірні помилки + тренд.

Для цих моделей будуються графіки та проводиться статистичний аналіз.

1. Моделювання з аномальними викидами :

Створюються моделі тренду з нормальними помилками та АВ.

Створюються моделі тренду з рівномірними помилками та АВ.

Будуються графіки та виконується статистичний аналіз цих моделей.

1. Аналіз реальних даних:

Завантажуються реальні дані з архіву Ощадбанку, проводиться обробка та виведення результатів:

* Коливання курсу USD в 2022 році.
* Аналіз статистичних характеристик реальних даних.

1. Завершення програми.
   1. **Опис структури проекту програми в середовищі PyCharm.**

Для реалізації розробленого алгоритму мовою програмування Python сформовано проєкт.

Проєкт базується на лінійній бізнес-логіці функціонального програмування та має таку структуру.

|  |
| --- |
|  |
| Рис.2. Структура проекту. |

ds1.ipynb – файл програмного коду лабораторної роботи;

ds1.doc – файл звіту лабораторної роботи.

Oschadbank (USD).xls – файл xls з реальними даними.

* 1. **Результати роботи програми відповідно до завдання.**

Результатом роботи програми є:

Діаграми: Програма генерує різні діаграми, які візуалізують дані.

Числові характеристики: Програма розраховує числові характеристики для вибірок, такі як математичне сподівання (середнє значення), дисперсія (середньоквадратичне відхилення) та інші. Ці характеристики надають кількісні оцінки розподілу даних.

Вивід результатів: Результати виводяться на екран у вигляді текстових повідомлень, діаграм і графіків. Вони можуть бути використані для аналізу та порівняння різних даних та моделей.

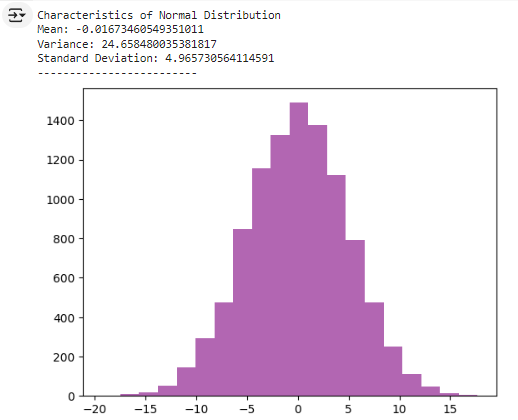


Рис.3. – Статистичні характеристики Нормального розподілу.

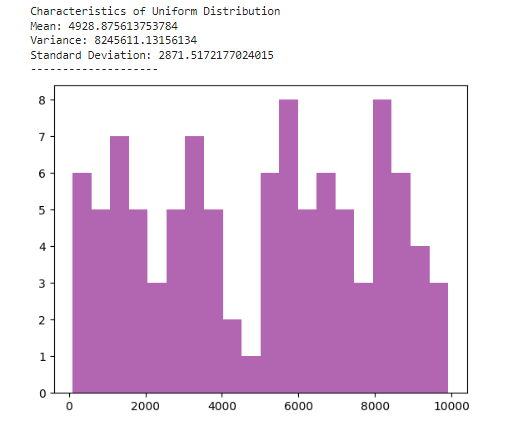


Рис.4. – Статистичні характеристики Рівномірного розподілу.

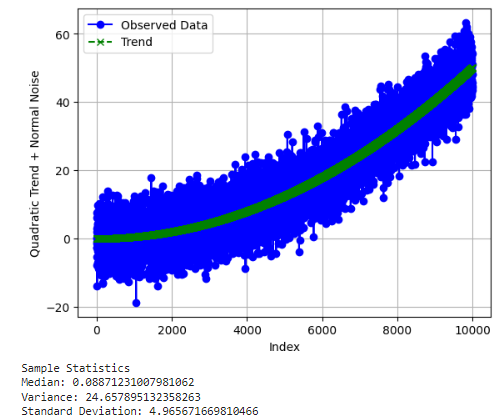


Рис.5. – Квадратична модель з Норм. шумом

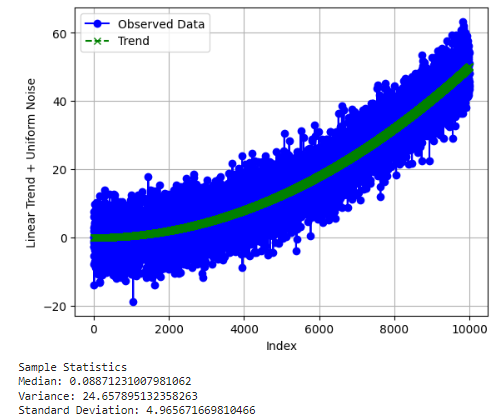


Рис.6. – Лінійна модель з рівномірним шумом

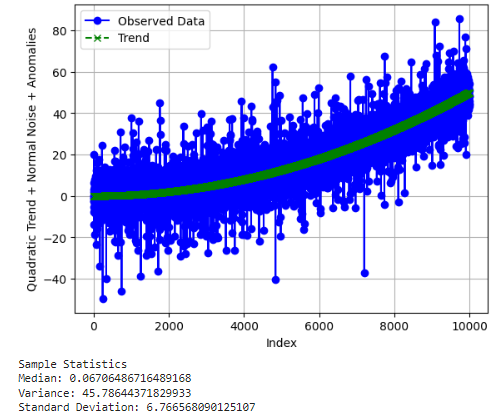


Рис.5. – Квадратична модель з Норм. шумом + АВ

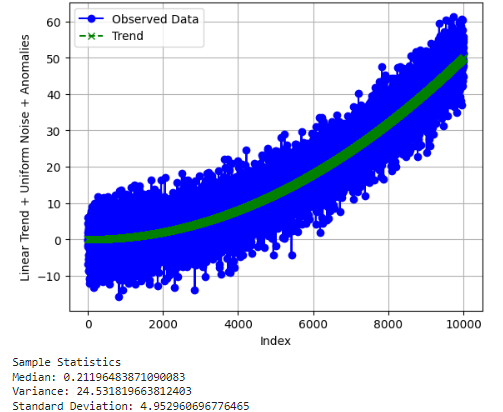


Рис.6. – Лінійна модель з Рівномірним шумом + АВ

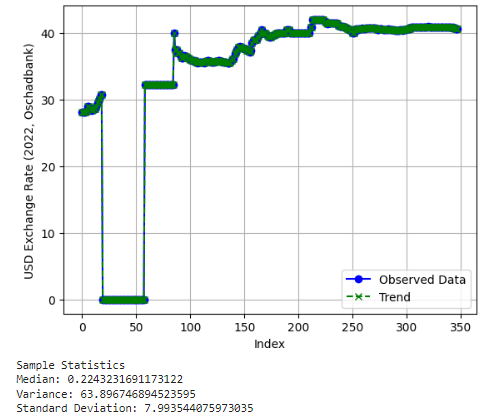
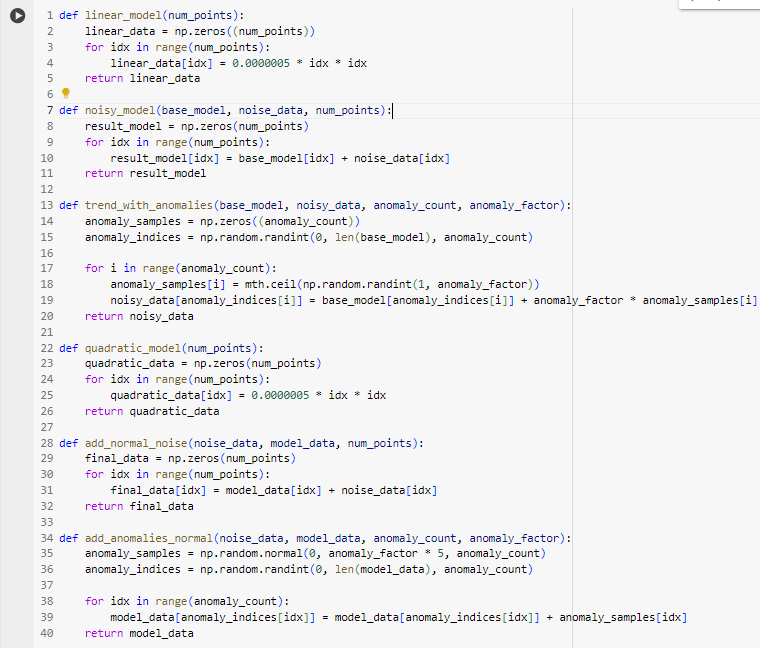


Рис. 7. – Визначення статистичних характеристик реальних даних, заданих файлом Oschadbank (USD).xls

* 1. **Програмний код.**

Програмний код послідовно реалізує алгоритм рис.1 та спрямовано на отримання результатів, поданих вище. При цьому використано можливості Python бібліотек: pip; pandas; numpy; xlrd; matplotlib. Контексні коментарі пояснюють сутність окремих скриптів наведеного коду програми. (Повний лістинг коду у файлі ds1.py або [github](https://github.com/YeaLowww/))







* 1. **Аналіз результатів відлагодження та верифікації результатів роботи програми.**

Результати відлагодження та тестування довели працездатність розробленого коду. Це підтверджується результатами розрахунків, які не суперечать теоретичним положенням.

Верифікація функціоналу програмного коду, порівняння отриманих результатів з технічними умовами завдання на лабораторну роботу доводять, що усі завдання виконані у повному обсязі.

**IV. Висновки.**

В результаті виконаної лабораторної роботи був розроблений програмний скрипт мовою Python який забезпечує аналіз різних характеристик та властивостей вибірок даних, включаючи синтез випадкових величин, моделювання тренду та визначення статистичних параметрів.

Синтезовані дані є корисними для тестування і верифікації алгоритмів їх обробки і будуть використані в подальших дослідженнях часових рядів.

Виконав: студент Олександр Головня