**Лабораторна робота 1: Описові статистики та варіаційний аналіз**

Постановка завдання: Написати ПЗ, що дозволяє користувачу провести первинний статистичний аналіз даних згідно варіанту через реалізацію таких процедур:

1. Попередня обробка даних.
2. Формування варіаційного ряду.
3. Проведення гістограмної оцінки. Передбачити можливість автоматичного та ручного вибору кількості класів для гістограми. Вивести кількість класів, крок розбиття, розмах та кількість даних.
4. Розрахунок точкових характеристик (зсунені та незсунені):
   * Середнє арифметичне
   * Середньоквадратичне відхилення
   * Коефіцієнти асиметрії
   * Ексцес
   * Контрексцес
   * Варіація Пірсона, непараметричний коефіцієнт варіації.
   * MAD
   * MED
5. Розрахунок довірчих інтервалів для всіх характеристик. Забезпечити виведення результатів з відповідною точністю (точність задає користувач).
6. Редагування даних (стандартизація, логарифмування, зсув, вилучення аномальних даних)
7. Побудова функції розподілу
8. Візуалізація результатів.

Написати звіт до лабораторної роботи, що містить теоретичні відомості, приклади роботи ПЗ, висновки. Захистити лабораторну роботу.

Варіант обирається за формулою , де номер в списку підгрупи

**Загальні вимоги до програми**

1. Програма повинна бути незалежна від даних. Вхідний файл має обиратися в діалозі з користувачем. Передбачається, що вхідні дані знаходяться в текстовому файлі, обсяг даних не відомий. Потрібно забезпечити можливість модифікації та збереження даних.

2. Слід уможливити перетворення даних (логарифмування, стандартизація, зсув).

3. Після перетворення або вилучення аномальних значень користувач повинен мати можливість повернутися до початкових даних.

4. Необхідно нанести на одну площину з гістограмою графік статистичної функції щільності, а на площину з графіком емпіричної функції розподілу графік статистичної функції розподілу разом із її довірчими інтервалами.

5. Результатом використання критерію згоди повинні бути як проміжні

результати (статистика критерію та її критичне значення), так і висновок (чи є відтворення розподілу достовірне).

6. Результати виконання всіх обчислень мають виводитись у вигляді таблиць, графіків і текстових коментарів.

7. Для кожного графіка слід виконати автоматичне масштабування, зобразити шкалу й показати одиниці виміру.

8. Відображення результатів повинне відповідати точності обчислень.

**Загальні вимоги до звіту**

Звіт із лабораторної роботи складається з таких частин:

1. Постановка задачі.

2. Теоретична частина.

3. Опис програми (програмні модулі, основні об’єкти, схема взаємодії модулів, інтерфейс, порядок роботи з програмою, опис формату вхідних даних та додаткових можливостей програми).

4. Реалізація (вхідні дані повністю, вихідні результати у вигляді графіків і таблиць, коментарі та пояснення щодо отриманих результатів).

5. Висновки.

Звіт здається в письмовій формі українською мовою.

Варіант 1. Робота call-центру.

Компанія, що надає послуги технічної підтримки, вирішила оптимізувати роботу свого кол-центру. Основна мета — зменшити час очікування клієнтів на відповідь оператора. Для цього менеджер з аналітики попросив стажера (студента) провести аналіз часу очікування клієнтів на лінії, використовуючи дані зібрані за допомогою системи моніторингу кол-центру.

Стажер отримав доступ до "сирих" даних із системи запису дзвінків. Кожен запис містив час дзвінка клієнта і тривалість очікування на з'єднання з оператором.

* + Джерело даних: система автоматичного розподілу дзвінків (Automatic Call Distribution, ACD), яка фіксує інформацію про кожен вхідний дзвінок, включаючи час очікування клієнта.
  + Період збору: дані зібрані протягом одного тижня роботи кол-центру (30 дзвінків щодня, всього 210 записів).

1. Особливості:
   * Середній час очікування становить \_\_\_ хвилин.
   * У системі присутні пропуски (коли дані про час очікування не записалися через технічні збої).
   * Аномальні значення з’являються в години пікових навантажень (наприклад, понад 20 хвилин очікування).
2. Попередня обробка:
   * Дані були експортувані у формат Excel і передані стажеру для аналізу.
   * Стажер виявив пропуски (записи без значень часу очікування) та кілька аномалій (надзвичайно довгий час очікування). Визначити аномалії та вирішити, як із ними працювати (наприклад, чи виключати їх із аналізу).
3. Проведення дослідження:

* Розділення даних на класи (наприклад, час очікування: 0-5 хвилин, 5-10 хвилин, 10-15 хвилин тощо).
* Побудова гістограми розподілу.
* Розрахунок характеристик згідно завданню лабораторної роботи.
* Розрахунок довірчих інтервалів.

Аналіз цих даних допоможе визначити, чи потрібно збільшити кількість операторів у години пікових навантажень. На основі аналізу компанія зможе зменшити час очікування та підвищити задоволеність клієнтів.

Для цього стажеру необхідно зробити висновки та надати свої рекомендації менеджеру. Наприклад:

Якщо середній час очікування перевищує прийнятний рівень (наприклад, 5 хвилин):  
Стажер може запропонувати збільшити кількість операторів у години пікових навантажень.

Якщо середній час очікування в межах норми або нижче - можна залишити поточну кількість операторів або навіть оптимізувати графіки роботи, скоротивши кількість персоналу в тихі години.

Додаткові ідеї.

* Запропонувати впровадження автоматичних відповідей (IVR) для скорочення навантаження на операторів.
* Дослідити залежність часу очікування від типу дзвінків (технічна підтримка, консультації тощо).

Варіант 2. Активні користувачів веб-сайту.

Власник онлайн-платформи вирішив оптимізувати маркетингові кампанії, аналізуючи активність користувачів на вебсайті. Для цього було зібрано дані про кількість унікальних відвідувачів, які заходять на сайт щогодини протягом одного дня. Аналітик – стажер (студент), отримав доступ до системи веб-аналітики (наприклад, Google Analytics), яка автоматично реєструє кількість відвідувачів за кожну годину.

* Дані зібрані протягом однієї доби (24 години) у файл формату Excel. У середньому сайт відвідує 50 користувачів за годину (λ=50). У певні години пікової активності спостерігається більше відвідувачів (наприклад, у робочий час або під час розсилки). У нічні години активність падає до мінімуму.
* Аномалії: значно вищий або нижчий трафік у порівнянні із середнім (λ=50). Наприклад, у годину пікової активності кількість відвідувачів може зрости до 100-120, а в нічний час — зменшитися до 10. Необхідно ідентифікувати аномалії та вирішити, як із ними працювати
* Пропуски: через технічні збої можуть бути відсутні дані про кількість відвідувачів у певні години. Необхідно визначити пропущені значення та замінити їх (наприклад, середнім значенням або інтерполяцією).

Провести дослідження згідно завданню лабораторної роботи.

Рекомендація щодо кількості класів:  
Використовуємо правило Стерджеса для визначення кількості класів:

k=1+3.322⋅lg(n), де n=48 (кількість записів в таблиці).

Зробити висновки, чи потрібно адаптувати маркетингові кампанії для збільшення трафіку в години низької активності. Надати рекомендації щодо оптимізації навантаження на сервери в години пікової активності.

 Години низької активності**.** Якщо відвідувачів у певні години (наприклад, 00:00–06:00) значно менше середнього (<30), рекомендувати запускати нічні акції чи спеціальні розсилки, щоб активізувати користувачів.

 Години пікової активності.Якщо в години пікового навантаження (наприклад, 10:00–14:00) кількість відвідувачів значно перевищує середнє (>80), запропонувати:

* + Оптимізувати роботу серверів.
  + Запускати спеціальні маркетингові кампанії для утримання цих користувачів.

 Загальні рекомендації.Якщо розподіл показує, що трафік нестабільний, оптимізувати план розсилок, оновлень контенту чи запускати рекламу в соціальних мережах у відповідні години.

Варіант 3. Швидкість завантаження веб-сторінки.

Інженери DevOps великої компанії вирішили провести моніторинг тривалості завантаження вебсторінок із сервера, щоб перевірити продуктивність системи. Мета — визначити, чи є збої у швидкості роботи сайту, чи потрібно оптимізувати серверні ресурси.

Інженери DevOps автоматизували збір даних за допомогою внутрішнього інструмента моніторингу (наприклад, New Relic або Pingdom), який реєструє час завантаження сторінки кожні 10 хвилин протягом 24 годин. Дані передано стажеру-аналітику (студенту) у файлі формату Excel для подальшого аналізу.

Особливості:

* У таблиці даних є випадкові пропуски (через тимчасові збої моніторингу).
* Аномалії: різко збільшений час завантаження через пікове навантаження на сервер або збої в роботі (наприклад, понад 4 секунди).

Попередня обробка даних:

* Знайти пропущені значення і запропонувати методи їх заміни (середнє значення, інтерполяція тощо).
* Ідентифікувати аномалії (наприклад, значення > 4 секунд) та вирішити, чи потрібно їх виключати або враховувати у висновках.

Гістограмна оцінка:

* При побудові варіаціного ряду використовувати Правило 3-х сігм або Правило квадратного кореня.
* Побудувати гістограму розподілу часу завантаження

Розрахунок характеристик:

* Обчислити середній час завантаження сторінки.
* Визначити стандартне відхилення та інші незсунені характеристики згідно завдання лабораторної роботи.

Обчислити довірчі інтервали, врахувати можливість внесення даних щодо йомовірності вручну.

Презентація висновків дослідження:

* Розробити рекомендації щодо оптимізації серверних ресурсів.
* Якщо середній час завантаження сторінки перевищує норму (наприклад, 2 секунди), запропонувати рішення (наприклад, масштабування серверів, оптимізація коду сторінки).

Варіант 4. Ефективність рекламних компаній

Маркетинговий відділ компанії провів аналіз ефективності різних рекламних кампаній. Їх мета — з'ясувати, скільки коштує залучення одного користувача залежно від типу реклами (соціальні мережі, контекстна реклама, email-кампанії тощо). Стажер-аналітик (студент) отримав доступ до рекламного кабінету (наприклад, Google Ads, Facebook Ads), де зібрав дані про витрати на рекламу та кількість залучених користувачів для кожного типу кампанії у файл формату Excel.

Розрахунок вартості залучення:

* Для кожної кампанії розрахували середню вартість залучення користувача за формулою:
  + Вартість залучення= Загальні витрати на рекламу/(Кількість залучених користувачів)

Особливості розподілу:

* Вартість залучення варіюється від $2 до $10 залежно від типу кампанії.

Обробка даних:

* Дані зібрані для 5 видів рекламних кампаній.
* Є пропуски в даних для кількох кампаній (через некоректне введення в рекламному кабінеті). Необхідно заповнити пропуски (наприклад середнім значенням або інтерполяцією).

Аномалії: кампанії з надто високою (>$10) або низькою (<$2) вартістю залучення. Необхідно їх визначити та прийняти рішення чи потрібно їх виключати або враховувати у висновках.

Як обробляти аномалії.

* Залишити аномалії для аналізу:  
  Якщо кількість аномалій незначна, залишити їх у наборі для окремого аналізу.
* Виключити аномалії:  
  Якщо аномалії суттєво спотворюють результати, виключити їх перед розрахунком статистичних характеристик.
* Окремий аналіз:

Виділити аномальні кампанії у звіт і з'ясувати причини високої/низької вартості. Наприклад, технічні помилки або неефективне налаштування реклами.

Гістограмна оцінка.

Врахувати діапазон можливих значень вартості залучення. Дані розподілені рівномірно між $2 і $10, тому можна створити класи з однаковим кроком:

* Якщо крок = 1 долар, отримаємо 8 класів (2–3, 3–4, ..., 9–10).
* Якщо крок = 2 долари, отримаємо 4 класи (2–4, 4–6, 6–8, 8–10).

Для детального аналізу краще використовувати менший крок (1 долар, 7–8 класів). Надати можливість користувачеві змінювати кількість класів вручну.

* Побудувати гістограму розподілу вартості.

Розрахунок статистичних характеристик:

* Обчислити середню вартість залучення користувача.
* Розрахувати стандартне відхилення, коефіцієнт варіації та інші основні характеристики згідно завданню лабораторної роботи.

Презентація висновків:

* Розробити рекомендації для зниження вартості залучення користувачів.
* Підказати, які типи кампаній є найефективнішими.

Варіант 5. Час виконання програмного скрипта.

Команда розробників проводить тестування продуктивності програмного забезпечення. В рамках цього тестування вони записують час виконання програмного скрипта під різними умовами, щоб визначити стабільність і продуктивність. Дані збирали DevOps-інженери за допомогою автоматизованої системи моніторингу (наприклад, Prometheus), яка записувала час виконання скриптів у різних середовищах та передала їх стажеру-аналітику (студенту) у файлі формату Excel. Дані записувались кожні 5 хвилин протягом робочого дня (8 годин).

Особливості:

* Пропуски: Через збої в системі моніторингу деякі значення відсутні.
* Аномалії: У випадках, коли система перевантажена або працює неправильно, час виконання значно збільшується (понад 8 секунд) або зменшується (менше 3 секунд).

Стажеру доручено провести аналіз продуктивності програмного скрипта та підготувати висновки для команди розробників. Для цього потрібно:

* 1. Провести попередню обробку даних.
* Виявити пропуски в даних і запропонувати методи їх заповнення (наприклад, середнє значення, медіана, інтерполяція).
* Ідентифікувати аномалії (значення < 3 секунд або > 8 секунд) і вирішити, чи залишати їх для подальшого аналізу, чи вилучати.

Рекомендації що робити з аномаліями.

* Окремий аналіз: Виокремити аномальні значення, щоб зрозуміти, чи вони викликані технічними збоями або нештатними умовами.
* Вилучення: Якщо аномалії не мають реальної причини і лише заважають аналізу, вилучити їх.

2. Провести гістограмне оцінювання.

- Побудувати варіаційний ряд (Рекомендована кількість класів — 8–9).

Розподіл діапазону:

* Основні значення часу виконання скрипта (μ±3σ) знаходяться в діапазоні 2–8 секунд.
* Рекомендується поділ із кроком 0.5 або 1 секунда, наприклад:

2–3, 3–4, 4–5, 5–6, 6–7, 7–8.

* + - Реалізувати можливість зміни кількості класів в ручну.

- Побудувати гістограму розподілу.

3. Розрахувати статистичні характеристики згідно завданню лабораторної роботи.

4. Оцінити, чи стабільно працює скрипт, виходячи з отриманих результатів.

* Надати рекомендації щодо оптимізації коду або системних ресурсів, якщо середній час виконання значно відхиляється від очікуваного (μ=5).
* Запропонувати заходи щодо усунення причин аномалій.

Рекомендації щодо висновків.

Стабільність виконання:

* Якщо час виконання має високий розкид (велике стандартне відхилення), запропонувати оптимізацію коду або системних ресурсів.

Ідентифікація піків:

* Виявити, чи є певні часові інтервали (наприклад, ранок чи вечір), коли час виконання значно змінюється.

Усунення причин аномалій:

* Аналізувати аномальні випадки, щоб уникнути подібних ситуацій у майбутньому.

Варіант 6. Покупки в інтернет магазині

Мережа магазинів проводить аналіз поведінки клієнтів інтернет-магазину, щоб зрозуміти, у які години дня найбільше покупок і як це залежить від часу. Дані будуть використовуватись для оптимізації графіка роботи персоналу та маркетингових акцій. Аналітик (наш студент) отримав дані від системи продажів (POS-системи), яка реєструє кожну покупку в реальному часі.

Дані зібрані за одну добу, розділені на 48 інтервалів (кожні 30 хвилин) знаходяться у файлы формату Excel.

Особливості.

* Пропуски: Через збої у системі частина даних відсутня.
* Аномалії: Надзвичайно високі (наприклад, > 50) або низькі значення (наприклад, < 5), викликані технічними збоями.

Рекомендації щодо обробки аномалій.

* Залишити для аналізу, якщо вони відображають реальні події (наприклад, акції чи незаплановані пікові години).
* Вилучити, якщо це технічні помилки або некоректні дані.

1. Провести попередню обробку даних.

* Заповнити пропущені значення, використовуючи середнє значення для відповідного часу доби.
* Визначити аномалії та надати пояснення їх причин.

2. Гістограмне оцінювання.

* Для визначення кількості класів використати правило Стерджеса:
* k=1+3.322⋅lg10(n) де n=48 (кількість інтервалів).
* k≈7
* Рекомендована кількість класів — 7–8.
* Розподіл діапазону.
* Якщо основні значення варіюються від 5 до 50, то крок можливо покласти рівним 5 або 10 покупок:
  + 5–10, 10–15, ..., 45–50 (для детального аналізу).
  + 5–15, 15–25, ..., 35–50 (для грубого аналізу).
* Побудувати гістограму розподілу покупок.
  1. Розрахунок характеристик.
* Середня кількість покупок за кожен часовий інтервал.
* Розрахувати стандартне відхилення та інші статистичні характеристики що наведено в завданні лабораторної роботи.
* Розрахувати довірчі інтервали, передбачити введення значення ймовірності користувачем.
  1. Висновки. Після аналізу даних стажер може зробити наступні висновки:

#### **1. Аналіз активності клієнтів та оптимізація роботи персоналу.**

* Виявити, в які часові інтервали спостерігається **максимальна** та **мінімальна** активність клієнтів (потрібна максимальна кількість персоналу).
* Порівняти активність у різні періоди доби (ніч, ранок, день, вечір) та визначити, чи відповідає вона очікуваному розподілу (запропонувати скорочення персоналу).

#### **2. Причини аномалій:**

* **Якщо кількість покупок значно більша за 50:**  
  Це може свідчити про проведення акцій чи розпродажів, підвищену активність у святкові дні.
* **Якщо кількість покупок значно менша за 5:**  
  Це може бути викликано технічними збоями, закриттям магазину або низькою клієнтською активністю.

#### **3. Маркетингові можливості:**

* + Якщо активність у певний час доби (наприклад, ранок або вечір) нижча за середню, запропонувати проведення акцій (знижки, купони).

Варіант 7. Моніторінг серверної системи

Системні адміністратори проводять моніторинг серверної системи, щоб оцінити її продуктивність і стабільність. Одним із ключових показників є тривалість обробки запитів на сервері. Дані зібрані через систему моніторингу продуктивності (наприклад, Prometheus або Grafana), яка записувала тривалість обробки кожного запиту та передані аналітику-стажеру (студенту) для проведення аналізу.

Дані зібиралися протягом однієї години з інтервалом у 15 секунд (240 записів).

Особливості.

* Пропуски: Через технічні збої деякі записи можуть бути відсутні.
* Аномалії: Надто короткий (< 100 мс) або надто довгий час (> 300 мс) вважається аномалією.

Рекомендації щодо обробки аномалій.

* + Окремий аналіз: Виокремити ці дані, щоб зрозуміти, чи вони спричинені перевантаженням системи або помилками.
  + Вилучення: Якщо аномалії викликані помилками моніторингу, виключити їх.
  + Короткий час (< 100 мс): Можливо, кешування або тестові запити.
  + Довгий час (> 300 мс): Можливо, перевантаження системи, складні операції або проблеми з мережею.

### **Аналітику-стажеру необхідно провести статистичний аналіз даних. А саме**

1. **Попередня обробка**
   * Виявити пропуски в даних і заповнити їх (наприклад, середнім значенням).
   * Ідентифікувати аномалії та пояснити їх причини.
2. **Гістограмне оцінювання**
   * Зробити розподіл даних на класи. Рекомендовано використовувати правило Стерджеса:

* k=1+3.322⋅lg(n) де n=240.
* k≈10k

Рекомендована кількість класів — 10.

Або правило 3х сігм. В цьому випадку розподіл діапазону може бути наступним.

Якщо основні значення (μ±3σ) варіюються від 50 мс до 350 мс. То крок може становити:

* + 25 мс: 50–75, 75–100, ..., 325–350.
  + 50 мс: 50–100, 100–150, ..., 300–350.
* Побудувати гістограму розподілу тривалості.

1. **Розрахунок характеристик**
   * Середня тривалість обробки запиту.
   * Стандартне відхилення та інші характеристики, що визначені в завданні лабораторної роботи.
   * Довірчі інтервали (передбачити можливість задання ймовірності користувачем).
2. **Висновки**

Після аналізу **продуктивності серверної системи** стажер може зробити наступні висновки та Надати рекомендації щодо усунення причин аномалій та оптимізації роботи сервера.

* Визначити, чи відповідає продуктивність сервера очікуванням. Чи відповідає середня тривалість обробки запиту очікуваному значенню (μ=200 мс)?
* Оцінити стабільність роботи сервера через величину стандартного відхилення (σ):
  + **Мале** σ **-** сервер працює стабільно.
  + **Велике** σ **-** сервер може бути перевантажений або працює нестабільно.
* Запропонувати рішення для оптимізації продуктивності:
  + **Якщо середнє значення > 200 мс:** Зменшити складність операцій, масштабувати сервер.

Варіант 8 Продаж програмного забезпечення

Компанія-виробник програмного забезпечення проводить аналіз місячного прибутку від продажів своїх продуктів. Дані допомагають оцінити ефективність продажів, визначити найприбутковіші місяці та зробити прогнози на майбутнє. Фінансовий аналітик зібрав дані про продажі з CRM-системи компанії, яка фіксує дохід від кожного проданого продукту.

Аналітик-стажер (студент) отримав зібрані дані (фінансові звіти, сформовані на основі транзакцій за кожен місяць) у вигляді таблиці Excel із записами про кожну угоду, але вони були "сирими" і потребували підготовки для подальшого аналізу.

* Дані зібрані за 36 місяців (3 роки).

Особливості.

* Пропуски: Відсутність даних у певні місяці через технічні причини.
* Аномалії: Надто низькі (< $1000) або надто високі (> $5000) значення, які можуть бути викликані коригуванням звітності або помилками в розрахунках.

Аналітик-стажер має підготувати ці дані для презентації, а також виконати їх статистичний аналіз.

* 1. Попередня обробка даних:

Видалити записи, які виглядають нереалістично (наприклад, продажі з нульовою сумою або надто високі значення).

Обробка аномалій:

* Окремий аналіз: Виокремити ці значення та з’ясувати причини (наприклад, великі корпоративні угоди або технічні помилки).
* Вилучення: Якщо значення викликані помилками звітності, їх можна вилучити.
  1. Гістограмна оцінка. Рекомендації -використовувати правило Стерджеса для визначення кількості класів:

k=1+3.322⋅lg(n) де n=36

k≈6k

Рекомендована кількість класів — 6.

* Можливий розподіл діапазону - основний діапазон значень: від $1000 до $5000.
* Крок може бути:
  + 500 доларів: 1000–1500, 1500–2000, ..., 4500–5000 (8 класів).
  + 1000 доларів: 1000–2000, 2000–3000, ..., 4000–5000 (4–5 класів).
* У програмі реалізувати функцію автоматичного та ручного поділу на класи.
* Побудувати гістограму розподілу місячних прибутків.
  1. Провести розрахунок характеристик.
* Знайти середній прибуток від продажу програмного забезпечення.
* Обчислити стандартне відхилення, коефіцієнти асиметрії, ексцес, варіацію тощо.

Довірчі інтервали.

* Використи отримані дані для розрахунку довірчих інтервалів, щоб надати бізнесу більш точну оцінку очікуваних прибутків. Врахувати можливість внесення значень ймовірностей вручну.

#### Зробити висновки. Провести а**наліз прибутків та надати рекомендації щодо покращення.**

* Визначити середній місячний прибуток та оцінити, чи він відповідає очікуваному значенню (μ=3000 доларів).
* Оцінити стабільність прибутків за величиною стандартного відхилення (σ):

**Мале** σ **- с**табільні продажі.

**Велике** σ **- з**начні коливання в обсягах продажів.

**Якщо є значні коливання прибутків, з**апропонувати заходи для стабілізації (наприклад, впровадження нових продуктів, розширення ринків).

* **Для підвищення прибутків у слабких місяцях запропонувати:**
  + Провести маркетингові кампанії.
  + Зосередитися на залученні нових клієнтів або укладенні великих угод.
* Виявити аномально низькі (< $1000) та високі (> $5000) значення та пояснити їх:
  + **Низькі прибутки:** Можуть свідчити про відсутність великих угод, проблеми у продажах або технічні помилки у звітності.
  + **Високі прибутки:** Можуть бути результатом корпоративних угод або пікових продажів.