#### **Проєкт по аналізу даних**

з предмету “Управління проєктами”

Студента четвертого курсу

механіко-математичного факультету

**Копиці Олексія Сергійовича**

#### 

#### **Опис проєкту**

Мета цього проєкту полягала в аналізі даних з опитувальників, представлених в Excel-таблиці “Data\_Excel”, побудові відповідних гістограм, графіків залежностей для певних пар величин та перевірці статистичної значущості по впливу цих параметрів.

#### **Математичні методи**

Для оцінки впливу параметрів і перевірки їх статистичної значущості були використані наступні математичні методи:

* **Кореляція Пірсона**: Використовується для оцінки лінійної залежності між двома змінними.
* **t-тест для однієї вибірки**: Використовується для перевірки, чи середнє значення вибірки дорівнює заданому значенню (у нашому випадку 0).

**Коефіцієнт кореляції Пірсона (r) -** це число від -1 до 1, яке показує ступінь лінійної залежності між двома змінними, вимірює силу та напрямок цієї залежності.  
Приклади того, як саме за значеннями r інтерпретувати результат, наведені нижче.

#### 

#### 

#### 

#### 

#### 

#### **Позитивна лінійна залежність (0 <= r <= 1)**

* r ≈ 1: Дуже сильна позитивна лінійна залежність.  
  Це означає, що коли значення однієї змінної зростає, значення іншої змінної також зростає.  
  Приклад: Коли температура підвищується, кількість проданих напоїв теж збільшується.
* r ≈ 0.7: Сильна позитивна лінійна залежність.  
  Значна частина варіації однієї змінної пояснюється варіацією іншої змінної.
* r ≈ 0.5: Помірна позитивна лінійна залежність.  
  Є позитивна залежність, але вона не дуже сильна.
* r ≈ 0.3: Слабка позитивна лінійна залежність.  
  Є невелика позитивна залежність між змінними.
* r ≈ 0: Відсутність лінійної залежності.  
  Змінні не мають лінійної залежності. Можливо, є інші види залежності (наприклад, нелінійна).

#### **Негативна лінійна залежність (-1 <= r <= 0)**

* r ≈ -1: Дуже сильна негативна лінійна залежність.  
  Це означає, що коли значення однієї змінної зростає, значення іншої змінної зменшується.  
  Приклад: Коли кількість тренувань збільшується, вага тіла зменшується.
* r ≈ -0.7: Сильна негативна лінійна залежність.  
  Значна частина варіації однієї змінної пояснюється варіацією іншої змінної.
* r ≈ -0.5: Помірна негативна лінійна залежність.  
  Є негативна залежність, але вона не дуже сильна.
* r ≈ -0.3: Слабка негативна лінійна залежність.  
  Є невелика негативна залежність між змінними.
* r ≈ 0: Відсутність лінійної залежності.  
  Змінні не мають лінійної залежності. Можливо, є інші види залежності (наприклад, нелінійна).

**P-значення** (p-value) є важливим показником у статистичних тестах, який допомагає нам оцінити, наскільки сильні докази проти нульової гіпотези (H0). Нульова гіпотеза зазвичай стверджує, що немає ефекту або різниці між групами.

#### **Інтерпретація p-значення**

* Велике p-значення (наприклад, > 0.05) вказує на слабкі докази проти нульової гіпотези, тому ми не можемо відкинути нульову гіпотезу. Це означає, що спостережуваний ефект не є статистично значущим.
* Мале p-значення (наприклад, <= 0.05) вказує на сильні докази проти нульової гіпотези, тому ми відкидаємо нульову гіпотезу. Це означає, що спостережуваний ефект є статистично значущим.

#### **Поріг значущості (α)**

Поріг значущості (α) є значенням, з яким ми порівнюємо p-значення для прийняття рішення про нульову гіпотезу. Найчастіше використовувані значення порогу:

* α = 0.05: Стандартний поріг значущості в багатьох наукових дослідженнях.
* α = 0.01: Більш строгий поріг значущості.
* α = 0.10: Менш строгий поріг значущості.

#### **Як p-значення впливає на рішення**

1. Якщо p-значення <= α:
   * Відкидаємо нульову гіпотезу. Спостережувані дані є статистично значущими, і існує ефект або різниця між групами.
2. Якщо p-значення > α:
   * Не можемо відкинути нульову гіпотезу. Спостережувані дані не є статистично значущими, і ми не можемо зробити висновок про існування ефекту або різниці.

У нашому проекті p-значення обчислюється для кожної пари змінних для оцінки значущості кореляції між ними. Якщо p-значення менше вибраного порогу значущості (наприклад, 0.05), то ми можемо вважати, що існує значуща кореляція між змінними.

### **t-тест**

t-тест - це статистичний тест, який використовується для порівняння середніх значень і визначення, чи є статистично значущі відмінності між ними. Існує кілька видів t-тестів, в цьому проекті використовувався одновибірковий t-тест.

#### **Одновибірковий t-тест**

Одновибірковий t-тест використовується для перевірки, чи середнє значення вибірки значно відрізняється від відомого або гіпотетичного середнього значення генеральної сукупності.

#### **Інтерпретація результатів**

Після обчислення t-статистики порівнюємо отримане значення з критичним значенням t (про те, як визначати критичне значення, згадано в розділі “Визначення критичного значення t”) для відповідного рівня значущості (α) і ступенів свободи (df=n−1, де n - розмір вибірки; детальніше про це в розділі “Визначення критичного значення t”). Якщо абсолютне значення t-статистики більше або дорівнює критичному значенню t, то нульова гіпотеза відхиляється.

Ми також можемо використовувати p-значення, яке повертається функцією t-тесту, для прийняття рішення.

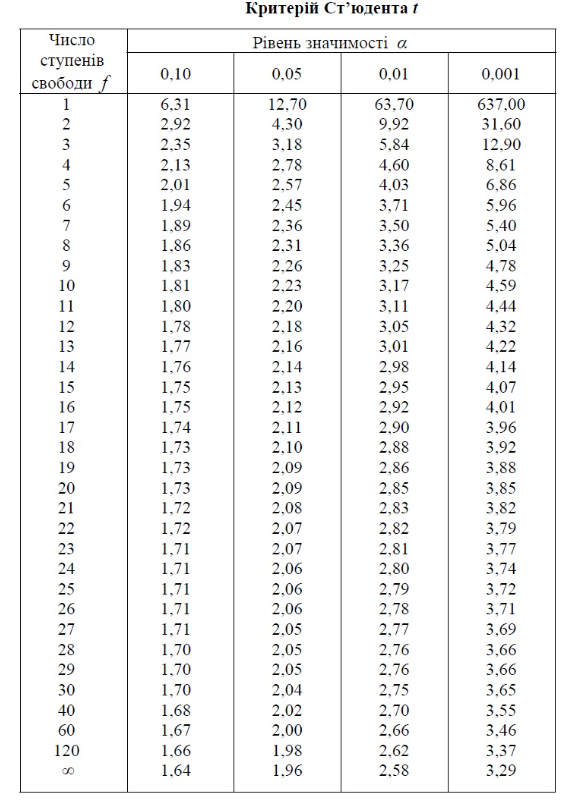
* Якщо p-значення <= α: Відкидаємо нульову гіпотезу. Це означає, що середнє значення вибірки значно відрізняється від гіпотетичного середнього значення.
* Якщо p-значення > α: Не можемо відкинути нульову гіпотезу. Це означає, що немає достатніх доказів для ствердження, що середнє значення вибірки відрізняється від гіпотетичного середнього значення.

### **Критичне значення t**

Критичне значення t - це значення, яке визначає межу, за якою результат t-тесту вважається статистично значущим.

**Визначення критичного значення t**

Критичне значення t залежить від:  
1. Рівня значущості (α): Це ймовірність відхилити нульову гіпотезу, коли вона є істинною. Найчастіше використовуються значення α = 0.05, 0.01, або 0.10.  
2. Ступенів свободи (df): Це кількість незалежних значень у вибірці, які можуть варіювати. Для одновибіркового t-тесту df = n - 1, де n - розмір вибірки.

Маючи ці значення, критичне значення t визначається за допомогою таблиці Критерію Ст’юдента:  


У нашому проекті t-тест використовується для перевірки, чи середня різниця між двома змінними значно відрізняється від нуля.

#### **Опис програми**

Програма складається з п’яти основних функцій:

1. **data\_preparation()**: Підготовка даних з Excel-файлу.
2. **histogram\_creation()**: Створення гістограм для кожного параметра в кожному аркуші таблиці.
3. **get\_arrays\_of\_data()**: Зчитування даних з Excel-файлу та перетворення їх у масиви Python.
4. **construction\_of\_graphs()**: Створення графіків залежностей для певних пар величин з різних аркушів таблиці.
5. **correlation\_calculation()**: Перевірка статистичної значущості впливу параметрів за допомогою кореляції Пірсона та t-тесту.

#### **data\_preparation()**

Ця функція виконує наступні дії:

* Зчитує дані з оригінального Excel-файлу Data\_Excel.xlsx.
* Замінює певні текстові значення в клітинках на числові значення згідно зі словником values\_to\_replace.
* Якщо в непорожньому стовпці є порожня клітинка, вписує в неї "I can't answer".
* Зберігає оновлений Excel-файл як Data\_Excel(1).xlsx.

#### **histogram\_creation()**

Ця функція виконує наступні дії:

* Зчитує дані з підготовленого Excel-файлу Data\_Excel(1).xlsx.
* Створює гістограми для кожного параметра в кожному аркуші таблиці.
* Зберігає гістограми у вигляді зображень у теці histograms.
* Використовує паузу в 2с між побудовою кожної гістограми для уникнення блокування.

#### 

#### 

#### 

#### 

#### **get\_arrays\_of\_data()**

Ця функція виконує наступні дії:

* Зчитує дані з підготовленого Excel-файлу Data\_Excel(1).xlsx.
* Створює словник all\_arrays для збереження масивів даних з кожного аркуша.
* Створює словник all\_num\_of\_cols\_before\_empty для збереження кількості стовпців до того, як зустрінеться перший порожній стовпець, з кожного аркуша (це використовуватиметься у функціях construction\_of\_graphs() та correlation\_calculation()).
* Перебирає кожен аркуш у файлі та зберігає значення непорожніх стовпців у масиви.
* Повертає результат у вигляді словника, який містить масиви даних і кількість стовпців до першого порожнього стовпця для кожного аркуша.

#### **construction\_of\_graphs()**

Ця функція виконує наступні дії:

* Використовує результати, отримані функцією get\_arrays\_of\_data().
* Використовує словник letters\_of\_columns, щоб замінити заголовки деяких стовпців на відповідні до них букви, щоб уникнути помилок через завеликі назви файлів.
* Створює графіки залежностей для пар величин з лівих та правих стовпців з усіх аркушів таблиці (різниця між лівими та правими стовпцями визначається за допомогою масиву all\_num\_of\_cols\_before\_empty, здобутого функцією get\_arrays\_of\_data()). В залежності від кількості повторів, кожна точка на графіку має відповідні розмір, колір та підпис з числом повторів.
* Зберігає графіки у вигляді зображень у теці graphs.
* Використовує паузу в 3с між побудовою кожного графіка для уникнення блокування.

#### 

#### 

#### 

#### 

#### 

#### **correlation\_calculation()**

Ця функція виконує наступні дії:

* Використовує результати, отримані функцією get\_arrays\_of\_data().
* Використовує словник numerical\_equivalent для заміни певних текстових значень з таблиці на відповідні числові.
* Використовує словник letters\_of\_columns, щоб замінити заголовки деяких стовпців на відповідні до них букви.
* Виконує обчислення кореляції Пірсона для кожної пари параметрів з лівих та правих стовпців кожного аркуша (різниця між лівими та правими стовпцями визначається за допомогою масиву all\_num\_of\_cols\_before\_empty, здобутого функцією get\_arrays\_of\_data()).
* Виконує t-тест для оцінки статистичної значущості різниці значень.
* Записує результати розрахунків у текстовий файл correlation\_results.txt.

Ці функції забезпечують повний цикл обробки даних, від підготовки та візуалізації до статистичного аналізу.

#### **Використані бібліотеки**

Для реалізації завдання використовувалися наступні бібліотеки:

* **pandas**: Для зчитування даних з Excel-файлу та роботи з ними у вигляді таблиць.
* **openpyxl**: Для обробки Excel-файлів.
* **numpy**: Для числових обчислень.
* **matplotlib**: Для побудови графіків та гістограм.
* **scipy**: Для проведення статистичних тестів, таких як кореляція Пірсона та t-тест.

**SciPy  
  
SciPy (Scientific Python) -** це відкрита бібліотека Python для наукових і технічних обчислень. Вона базується на бібліотеці NumPy і надає широкий спектр функцій для обробки та аналізу даних. SciPy забезпечує модулі для оптимізації, інтеграції, інтерполяції, власних значень, швидких перетворень Фур'є, сигналів і зображень, розв'язування звичайних диференціальних рівнянь та інших завдань.  
  
У цьому проекті використовувалися наступні модулі та функції з бібліотеки SciPy:  
  
**scipy.stats.pearsonr**: Використовується для обчислення коефіцієнту кореляції Пірсона (детальніше про математичну складову даної функції в розділі “Математичні методи”) та p-значення (детальніше про математичну складову даної функції в розділі “Математичні методи”) для тестування гіпотези про відсутність кореляції. Кореляція Пірсона вимірює лінійну залежність між двома змінними.  
  
from scipy.stats import pearsonr  
correlation\_coefficient, p\_value = pearsonr(x, y)

**pearsonr(x, y)**Input:  
x - перший набір даних (в нашому випадку масив даних з правих стовпців)  
y - другий набір даних (в нашому випадку масив даних з лівих стовпців)Output:  
2 значення: коефіцієнт кореляції Пірсона та p-значення.

**scipy.stats.ttest\_1samp**: Використовується для проведення одновибіркового t-тесту (детальніше про математичну складову даної функції в розділі “Математичні методи”). Цей тест перевіряє, чи середнє значення вибірки суттєво відрізняється від заданого значення (в нашому випадку 0).

from scipy.stats import ttest\_1samp  
t\_statistic, p\_value\_ttest = ttest\_1samp(difference, 0)

**ttest\_1samp(a, popmean)**  
Input:  
**a** - масив вибірки (в нашому випадку масив різниць значень між правими і лівими стовпцями)  
**popmean** - гіпотетичне середнє значення в генеральній сукупності (в нашому випадку 0)  
Output:  
t-статистика та p-значення

Ці функції дозволяють оцінити лінійну залежність між змінними та перевірити гіпотезу про нормальний розподіл різниці значень з середнім, що дорівнює нулю.

### **Висновок**

Цей проєкт продемонстрував, як можна використовувати Python для обробки даних з Excel-файлів, побудови графіків та проведення статистичних аналізів. Було використано кілька важливих бібліотек, таких як pandas, matplotlib, openpyxl, numpy та scipy, для ефективного виконання цих завдань.