Практична робота №3

Статистичне оцінювання та верифікація гіпотез

Мета роботи: Отримати навички переходу від візуальних припущень до математичного доведення гіпотез. Навчитися обирати адекватний статистичний тест залежно від форми розподілу, виявленої на етапі EDA.

Стек технологій:

Pandas *— підготовка та агрегування структурованих даних*

NumPy *— векторні обчислення*

SciPy (stats) *— статистичні тести*

Seaborn *— статистична візуалізація*

Зміст роботи

Завдання 1. Статистичне підтвердження форми розподілу

У ПР №2 було візуально оцінено розподіл по змінній артрит. Зараз необхідно підтвердити це математично.

1. ***Гіпотеза H0:*** показник ARTHRITIS_CrudePrev розподілений нормально. Сформулюйте альтернативну гіпотезу.
2. Проведіть тест ***Шапіро-Вілка*** (*stats.shapiro*) або ***Д'Агостіно*** (*stats.normaltest*) для всієї вибірки.
3. Порівняйте отримане *p-value* з критичним рівнем значущості *(α= 0.05*).
4. Спираючись на результати тесту та значення асиметрії з ПР №2, прийміть або відхиліть гіпотезу про нормальність. Поясніть, як це вплине на вибір тестів у наступних завданнях (параметричні чи непараметричні).

Завдання 2. Математичне доведення впливу розміру міста

У ПР №2 можна побачити різницю на скрипкових графіках, доведіть, що вона не випадкова.

1. ***Гіпотеза H0:*** середній рівень ACCESS2 однаковий для міст типів *Small*, *Medium* та *Large*. Сформулюйте альтернативну гіпотезу.
2. Оберіть тест:
   - Якщо дані розподілені нормально - ***ANOVA*** (*stats.f_oneway*).

– Якщо ні - ***Kruskal-Wallis test*** (stats.kruskal).

3. Розрахуйте *p-value* та зробіть висновок. Чи є різниця між групами статистично значущою?

4. Проведіть *post-hoc* аналіз (наприклад, тест *Манна-Уїтні* для пар *Small-Large*), щоб визначити, між якими саме групами різниця найвідчутніша.

Завдання 3. Верифікація кореляцій та P-value

1. Оберіть пару показників з найвищою кореляцією за результатами ПР №2.

2. Розрахуйте коефіцієнт кореляції та обов'язково його ***p-value*** (*stats.pearsonr* або *stats.spearmanr*).

3. ***Гіпотеза H0:*** кореляція між цими показниками дорівнює нулю (зв'язок випадковий). Сформулюйте альтернативну гіпотезу

4. Обчисліть **коефіцієнт детермінації** (*R^2*) і поясніть, який відсоток варіації одного показника пояснюється іншим.

Завдання 4. Порівняльний аналіз регіонів (A/B testing логіка)

1. Оберіть два штати з контрастними показниками.

2. Сформулюйте гіпотезу про різницю в рівні здоров'я між ними.

3. Виконайте ***T-test*** (для нормальних даних) або ***Mann-Whitney U test*** (для ненормальних).

4. Оцініть, чи є різниця між штатами практично значущою, чи вона просто зафіксована математично через великий обсяг даних?

Завдання 5. Самостійне формулювання та перевірка гіпотез

1. Запропонуйте 3 власні гіпотези.

2. Для кожної гіпотези:

– Чітко запишіть *H0* та *H1*.

– Проведіть тест та візуалізуйте результат (*boxplot/violinplot*).

– Сформулюйте аналітичний висновок (статистичний результат + соціальне значення).

Методичні рекомендації

Робота з гіпотезами

Етап *Data Understanding* у *CRISP-DM* вимагає не просто опису, а критичного погляду. Якщо гіпотеза не підтвердилася (наприклад, населення не впливає на рівень артриту), це також важливий науковий результат. Пам'ятайте, що:

– Кожна гіпотеза повинна мати *чітке предметне обґрунтування*, а не лише статистичний інтерес.

– Ніколи не починайте з тесту - *починайте з візуалізації*.

– Перевірка нормальності є обов'язковою перед параметричними тестами.

Для формулювання гіпотез характерні риси:

– H_0 *завжди містить знак рівності* (=, ≤, ≥).

– H_1 формулюється як *наявність ефекту або відмінності*.

– p-value < 0.05 → підстава для відхилення H_0.

Зверніть увагу на вибір тестів:

– t-test / ANOVA → нормальний розподіл.

– Mann–Whitney / Kruskal–Wallis → асиметричні розподіли.

– χ^2 → зв'язок між категоріальними ознаками.

– Для медичних даних часто доцільні *непараметричні методи*.

Особливу увагу в роботі приділено розмежуванню статистичної значущості та реальної практичної цінності отриманих результатів. Аналіз не обмежується лише формальними показниками (P-value), а обов'язково включає оцінку розміру ефекту та врахування специфічного контексту дослідження. Підсумкові висновки інтегруються у площину предметної області — медицини чи демографії, що дозволяє трансформувати математичні розрахунки у зрозумілі галузеві рекомендації.

В результаті повинно бути:

– Сформульовані та перевірені гіпотези.

– Обґрунтований вибір статистичних тестів.

– Візуалізовані результати.

– Аналітичні висновки з предметної області.

Приклади формулювання гіпотез:

Приклад 1:

H_0: Середній рівень артриту не відрізняється між групами.

H_1: Середній рівень артриту відрізняється статистично значущо.

Приклад 2:

H_0: Кореляція між ACCESS2 та ARTHRITIS відсутня.

H_1: Кореляція між ACCESS2 та ARTHRITIS існує.

Використовуйте візуалізацію з ПР №2 як основу для виконання практичної роботи. Якщо на графіку Jointplot ви бачили нелінійну хмару точок - використовуйте кореляцію Спірмена замість Пірсона.

Правило P-value: якщо p < 0.05 - "Геть нульову гіпотезу!" (результат значущий). Якщо p>= 0.05 - "Немає підстав відхиляти H0" (різниця випадкова).

Вибір тесту. Непараметричні тести (Манна-Уїтні, Краскела-Уолліса) є більш надійними для медичних даних, оскільки вони менш чутливі до викидів (міст-мільйонників), які ви виявили в ПР №2.

Приклади коду

t-test

```
from scipy import stats

stats.ttest_ind(group_a, group_b, equal_var=False)
```

Кореляція

```
stats.pearsonr(x, y)
stats.spearmanr(x, y)
```

ANOVA

```
stats.f_oneway(group1, group2, group3)
```

Перевірка нормальності

```
stats.shapiro(sample)
```

Контрольні питання

1. Чому візуального аналізу недостатньо для прийняття бізнес-рішень?
2. Як викиди, знайдені через *boxplot* у ПР №2, вплинули на результати *T-тесту* в ПР №3?
3. Що таке "помилка першого роду" в контексті аналізу медичних показників?
4. Чому при великій вибірці (500 міст) навіть мізерна кореляція може мати $p < 0.05$?
5. Як результати перевірки гіпотез впливають на етап *Evaluation* у циклі CRISP-DM?