Лабораторна робота №2

Статистичний аналіз даних

Мета роботи: Ознайомитись з основними функціями бібліотеки NumPy та SciPy для описової статистики, перевірки статистичних гіпотез, кореляційного аналізу та лінійної регресії.

Короткі теоретичні відомості

Статистичними називають гіпотези про вигляд розподілу генеральної сукупності або про параметри відомих розподілів.

Основною (нульовою) називають висунуту гіпотезу і позначають H_0.

Альтернативною (конкурентною) називають гіпотезу, що суперечить основній, її позначають H_1. Альтернативні гіпотези бувають двосторонніми та односторонніми.

Алгоритм перевірки гіпотез:

1) Визначити параметр, стосовно якого потрібно перевірити гіпотезу
2) Визначити основну гіпотезу H_0
3) Визначити гіпотезу H_1 , двосторонню чи односторонню, альтернативну до гіпотези H_0
4) Обрати статистичний критерій для перевірки
5) Визначити критерій відхилення основної гіпотези, наприклад, p-значення нижче за певний рівень значущості α (найчастіше 0,05)
6) Використати відповідну функцію і отримати значення статистичного критерію та p-значення
7) Зробити висновки: чи потрібно відхилити основну гіпотезу.

Коваріація та коефіцієнт кореляції Пірсона показують міру лінійного зв’язку між випадковими величинами. Коефіцієнт кореляції рангу Спірмена є непараметричним показником монотонності зв'язку між двома наборами даних. На відміну від кореляції Пірсона, кореляція Спірмена не передбачає нормального розподілу обох наборів даних. Коефіцієнт кореляції рангу Кендала або тау-коефіцієнт є непараметричним показником подібності впорядкування даних.

Проста лінійна регресійна модель:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Завдання до лабораторної роботи

Створити програму, яка за даними файлу виконує завдання відповідно до варіанту.

Оформити звіт. Звіт повинен містити:

— титульний лист;
— код програми;
— результати виконання коду;
— висновки щодо результатів (відповіді на завдання, прийняття/відхилення гіпотез, висновки про наявність зв’язку і т.д.)

Продемонструвати роботу програми та відповісти на питання стосовно теоретичних відомостей та роботи програми.

Варіант 1.
Файл Birthweight.csv.

1. Знайти середній вік матерів і батьків і порівняти ці середні значення.
2. Перевірити чи нормально розподілена вага дітей.
3. Перевірити за допомогою статистичних гіпотез чи у матерів, що палять, легші діти.
4. Чи є зв’язок між зростом матері та дитини?

Варіант 2.
Файл Crime.csv.

1. Знайти середню частоту злочинів (зараз і десять років тому).
2. Перевірити чи нормально розподілена частота злочинів (зараз і десять років тому).
3. Перевірити за допомогою статистичних гіпотез чи зросла частота злочинів за 10 років.
4. Який зв’язок між частотою злочинів і витратами на поліцію (коефіцієнт Спірмена)?

Варіант 3.
Файл possum.csv.

1. Знайти розмір, який не перевищують очі 25% опосумів.
2. Перевірити чи нормально розподілена довжина тіла.
3. Чи є зв’язок між довжиною тіла і віком опосума?
4. Чи відрізняється загальна довжина тіла опосумів Вікторії та інших провінцій? (за допомогою статистичних гіпотез)

Варіант 4.
Файл frogs.csv.

1. На якій середній відстані від поселення спостерігаються жаби, а на якій – ні, порівняти ці середні значення.
2. Перевірити чи нормально розподілена середня кількість опадів.
3. Чи є зв’язок між кількістю місць для розмноження та відстанню до поселення?
4. Перевірити за допомогою статистичних гіпотез чи однакові середні висоти, на яких спостерігаються жаби і на яких – ні?

Варіант 5.
Файл Budget.csv.

1. Який середній вік найстаршого в сім’ї і його середньоквадратичне відхилення?
2. Перевірити чи нормально розподілені витрати на одежу.
3. Чи є зв’язок між кількістю дітей і витратами на алкоголь?
4. Перевірити за допомогою статистичних гіпотез чи менші витрати на їжу в сім’ях з однією дитиною, ніж з двома.

Варіант 6.
Файл Birthweight.csv.

1. Знайти середній зріст дітей і його медіану.
2. Перевірити чи нормально розподілена кількість сигарет в день
3. Перевірити за допомогою статистичних гіпотез чи у матерів, що старші 35, легші діти.
4. Чи є зв'язок між тривалістю вагітності та вагою дитини?

Варіант 7.
Файл Crime.csv.

1. Знайти середнє та медіанне значення витрат на поліцію десять років тому.
2. Перевірити чи нормальні витрати на поліцію (зараз).
3. Перевірити чи в південних штатах вища частота злочинів (зараз).
4. Побудувати лінійну регресійну модель залежності частоти злочинів від доходу (зараз).

Варіант 8.
Файл possum.csv.

1. Знайти середній вік опосумів та його дисперсію.
2. Перевірити чи нормально розподілена довжина ноги.
3. Перевірити за допомогою статистичних гіпотез чи довжина хвоста самок менша за довжину хвоста самців.
4. Побудувати лінійну регресійну модель залежності довжини голови від довжини всього тіла.

Варіант 9.
Файл frogs.csv.

1. В умовах якого середньоквадратичного відхилення мінімальних температур від середнього значення спостерігаються жаби?
2. Перевірити чи нормально розподілена відстань до поселення.
3. Чи є зв'язок між висотою та кількістю опадів?
4. Перевірити за допомогою статистичних гіпотез чи відрізняється середня кількість опадів там, де спостерігаються жаби і де ні?

Варіант 10.
Файл Budget.csv.

1. Яка середня кількість дітей в сім'ї і її відхилення?
2. Перевірити чи нормально розподілені доходи.
3. Чи є зв'язок між витратами на пальне та витратами на транспорт?
4. Побудувати лінійну регресійну модель залежності витрат на їжу від доходу.

Варіант 11
Файл Birthweight.csv.

1. Знайти середній вік батька, його медіану і моду.
2. Перевірити чи нормально розподілена вага матерів.

3. Перевірити за допомогою статистичних гіпотез чи у чоловіків, що палять, легші діти.
4. Чи є зв’язок між вагою матері та вагою дитини?

Варіант 12.

Файл Crime.csv.

1. Знайти середню кількість років навчання в північних штатах (десять років тому і зараз).
2. Перевірити чи нормально розподілена кількість безробітних молодих чоловіків (десять років тому і зараз).
3. Перевірити за допомогою статистичних гіпотез чи зменшилась кількість безробітних чоловіків віком 25-39 років за 10 років.
4. Чи є лінійний зв’язок між частотою злочинів і кількістю бідних сімей?

Варіант 13.

Файл possum.csv.

1. Знайти середню довжину хвоста та її середньоквадратичне відхилення.
2. Перевірити чи нормально розподілена ширина голови.
3. Перевірити за допомогою статистичних гіпотез чи відрізняється довжина хвоста опосумів з 1-го місця і 7-го.
4. Побудувати лінійну регресійну модель залежності довжини тіла від віку.

Варіант 14.

Файл frogs.csv.

1. Яка середня кількість місць для розмноження жаб?
2. Перевірити чи нормально розподілена висота.
3. Чи є зв’язок між кількістю місць для розмноження та середньою максимальною температурою?
4. Побудувати лінійну регресійну модель залежності кількості можливих груп для розмноження від висоти.

Варіант 15.

Файл Budget.csv.

1. Які середні витрати та їх середньоквадратичне відхилення?
2. Перевірити чи нормально розподілений вік.
3. Чи є зв’язок між витратами на алкоголь та витратами на їжу?
4. Перевірити за допомогою статистичних гіпотез чи більші витрати на транспорт в сім’ях з двома дітьми, ніж з однією.