Міністерство освіти і науки України

Національний технічний університет України

“Київський політехнічний інститут імені Ігоря Сікорського”

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

Лабораторна робота №2

Аналіз даних з використанням мови Python

Тема: Статистичний аналіз даних

Варіант: 1

Виконав

студент групи ІП-11:

Панченко С. В.

Перевірив:

Тимофєєва Ю. С

Київ 2023

ЗМІСТ

1 МЕТА ЛАБОРАТОРНОЇ РОБОТИ

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

2 ЗАВДАННЯ

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?

3 ЗАВДАННЯ

3.1 Знайти середній вік матерів і батьків і порівняти ці середні значення

Для початку створимо датафрейм з датасету Birthweight.csv.

```
In [68]: import pandas as pd
         import numpy as np
         df = pd.read_csv('data/Birthweight.csv')
         df.head(10)
```

Out[68]:

	ID	Length	Birthweight	Headcirc	Gestation	smoker	mage	mnocig	mheight	mppwt	fage	fedyrs	fnocig
0	1360	56	4.55	34	44	0	20	0	162	57	23	10	35
1	1016	53	4.32	36	40	0	19	0	171	62	19	12	0
2	462	58	4.10	39	41	0	35	0	172	58	31	16	25
3	1187	53	4.07	38	44	0	20	0	174	68	26	14	25
4	553	54	3.94	37	42	0	24	0	175	66	30	12	0
5	1636	51	3.93	38	38	0	29	0	165	61	31	16	0
6	820	52	3.77	34	40	0	24	0	157	50	31	16	0
7	1191	53	3.65	33	42	0	21	0	165	61	21	10	25
8	1081	54	3.63	38	38	0	18	0	172	50	20	12	7
9	822	50	3.42	35	38	0	20	0	157	48	22	14	0

Рисунок 3.1 - Датафрейм Birthweight.csv

Оскільки лабораторна робота вимагає саме використання NumPy, а не Pandas, то дістанемо матрицю даних з датафрейма, а індекси колонок будемо знаходити за допомогою словника.

```
In [69]: mat = df.values
         cols = dict([
             (name, index) for index, name in enumerate(df.columns)])
         print(df.loc[0, 'ID'])
         print(mat[0, cols['ID']])
         print(cols["fage"])
```

```
1360
1360.0
10
```

Рисунок 3.2 - Робота з матрицею NumPy

Знайдемо середній вік матері та батька по всіх новонароджених, порівняємо їх. mage - колонка віку матері дитини, fage - колонка віку батька дитини.

Побачимо, що вік батька в середньому вищий.

```
In [84]: mean_mage = mat[:, cols["mage"]].mean()
         mean_fage = mat[:, cols["fage"]].mean()
         print(f'Mean mage: {mean_mage}')
         print(f'Mean fage: {mean_fage}')
         print(f"Delta: {mean_fage - mean_mage}")

         Mean mage: 25.547619047619047
         Mean fage: 28.904761904761905
         Delta: 3.3571428571428577
```

Рисунок 3.3 - Середній вік батьків - матері та батька дитини

3.2 Перевірити чи нормально розподілена вага дітей

Перевіримо нормальність розподілу ваги дітей за допомогою критерію Андерсона. Імпортуємо з пакету scipy модуль stats, використаємо тест Колмогорова-Смірнова. Побачимо, що розподіл ваги дітей не є нормальним. Якщо pvalue < 0.05, то приймається альтернативна гіпотеза.

```
In [71]: import scipy.stats as stats
         res = stats.kstest(mat[cols["Birthweight"]], 'norm')
         print(res)
         print(f'Is pvalue > 0.05? {res.pvalue > 0.05}')
         print(f'H_0 is {res.pvalue > 0.05}')

         KstestResult(statistic=0.7499793424930875, pvalue=8.015167938365836e-10)
         Is pvalue > 0.05? False
         H_0 is False
```

Рисунок 3.4 - Перевірка розподілу ваги на нормальність за допомогою теста Колмогорова-Смірнова

3.3 Перевірити за допомогою статистичних гіпотез чи у матерів, що палять, легші діти

Застосуємо двовибірковий t-критерій з альтернативною гіпотезою того, що у матерів, які палять, легші діти. Як бачимо альтернативна гіпотеза приймається, тому у курців діти мають меншу масу ніж у тих людей, що не палять.

```
In [72]: smoker = mat[mat[:, cols['smoker']] == 1][:, cols['Birthweight']]
         no_smoker = mat[mat[:, cols['smoker']] == 0][:, cols['Birthweight']]
         res = stats.ttest_ind(smoker, no_smoker, alternative='less')
         print('H1: smoker < no_smoker')
         print(f'H1 is {res.pvalue < 0.05}')

         H1: smoker < no_smoker
         H1 is True
```

Рисунок 3.5 - Доведення альтернативної гіпотези, що курці мають новонароджених з меншою масою

3.4 Чи є зв’язок між зростом матері та дитини?

Використаємо тест Пірсона для перевірки даної гіпотези. Функція stats.pearsonr бере за нульову гіпотезу те, що послідовності є некорельованими та нормально розподіленими. 'two-sided' - кореляція є ненульовою, 'less' - кореляція є від'ємною, 'greater' - кореляція є додатною. Бачимо, що поле staticstic показує чітку кореляцію, а оскільки pvalue > 0.05, то кореляція існує точно.

```
In [83]: x = np.arange(10, 20)
         y = x**2
         res = stats.pearsonr(mat[:, cols['Length']],
                              mat[:, cols['fheight']],
                              alternative='two-sided')
         print(res)
         print(f'pvalue > 0.05? {res.pvalue > 0.05}')

         PearsonRResult(statistic=0.20835843471346874, pvalue=0.1854521333515197)
         pvalue > 0.05? True
```

Рисунок 3.6 - Тест Пірсона для перевірки кореляції між зростом матері та дитини

4 ВИСНОВОК

Під час виконання даної лабораторної роботи я ознайомився з основними функціями бібліотеки NumPy та SciPy для описової статистики, перевірки статистичних гіпотез, кореляційного аналізу та лінійної регресії. У першому завданні було отримано, що середній вік батьків вищий ніж матерів. У другому завданні був використаний тест Колмогорова-Смірнова, далі побачили, що вага немовлят не розподілена нормально. За допомогою двовибіркового t-критерію дізналися, що матері, які палять, мають легших новонароджених дітей. І в кінці за допомогою тесту Пірсона вказали, що існує кореляція між зростом матері та дитини.