Лабораторна робота №1

Описова статистика (Descriptive statistics)

Мета роботи: Набути практичних навичок обробки даних і аналізу отриманих результатів з метою прийняття коректних обґрунтованих рішень технічного, ділового, політичного, персонального або іншого характеру, використовував можливості Python.

Література

statistics: https://docs.python.org/3/library/statistics.html
numpy.std: https://numpy.org/doc/stable/reference/generated/numpy.std.html
Statistics: https://numpy.org/doc/stable/reference/routines.statistics.html#

Зміст роботи

Завдання 1. Провести аналіз на наборі даних flats.csv.

– **З бібліотек знадобляться NumPy і Pandas:**

```
import numpy as np
import pandas as pd
```

– Завантажте файл і отримайте загальну інформацію про файл:

```
df = pd.read_csv('flats.csv')
```

– Використовував функції head() і tail() подивимося 30 перших і 10 останніх записів.

– Перевіримо тип даних загальної площі:

```
df.info()
```

– В результаті перевірки, бачимо, що змінна «Загальна площа» має формат *object,* змінемо тип:

```
df['Загальна_площа'] = pd.to_numeric(df['Загальна_площа'].astype(
'str').str.replace(',','.'))
```

– Дізнаємося скільки квартир продається у кожному місті, посортуємо дані у зручному порядку:

```
df.groupby('Місто')
print(df['Місто'].value_counts(sort=False))
```

– Знайдемо скільки продається трикімнатних квартир у кожному міст, окрім «Києво-Святошинського району»:

```
df = df[(df['Кімнат']==3) & (df['Місто']!='Києво-Святошинський')]
print(df['Місто'].value_counts(sort=True))
```

– Знайдемо середнє значення площі квартир в кожному регіоні:

```
df.groupby('Місто')['Загальна_площа'].mean()
```

– Знайдемо скільки продається двокімнатних квартир загальною площею більше 60 м.кв.:

```
df1 = df['Місто'][(df['Кімнат']==2) & (df['Загальна_площа'] > 60)
]
print(df.value_counts(sort=True))
```

– Обчислимо середнє значення площі квартир і середньоквадратичне відхилення:

```
print('СЕРЕДНЄ ЗНАЧЕННЯ ПЛОЩІ')
print(df['Загальна_площа'].mean())
print('\nСЕРЕДНЬОКВАДРАТИЧНЕ ВІДХИЛЕННЯ')
print( df['Загальна_площа'].std())
```

Знайти:

- В кожному місті саму маленьку за площею квартиру;
- В кожному місті квартиру з максимальною кількістю кімнат;
- Скільки продається трикімнатних квартир у Львові і Одесі
- Саму дорогу і саму дешеву квартиру в місті.

– Зробити висновки щодо набору даних.

Завдання 2. Провести аналіз статистичних характеристик на наборі даних flast.csv.

– Використайте метод

```
describe().
```

– Використайте прискорений розвідувальний аналіз даних з використанням бібліотеки pandas-profiling.

– Звіт можна експортувати в інтерактивний HTML файл:

```
profile = pandas_profiling.ProfileReport(df)
profile.to_file(outputfile="AAA data profiling.html")
```

– Зробіть висновки щодо набору даних.

Методичні рекомендації

Статистика – наука про збір, організацію та трактування даних.

Розрізняють *описову та вивідну статистику*. ***Описова статистика*** – вивчає властивості спостережуваних даних. ***Вивідна статистика*** – виводимо припущення про властивості розподілу даних з яких походять спостережувані дані

Завдання *описової статистики* (descriptive statistics) полягає в тому, щоб з використанням математичних інструментів звести сотні значень вибірки до кількох підсумкових показниках, які дають уявлення про вибірку. В якості таких статистичних показників використовуються: *середнє, медіана, мода, дисперсія, стандартне відхилення* тощо.

Для чого потрібні ці показники? Ці показники дозволять зробити певні статистичні висновки про розподіл, з якого була взята вибірка.

Для того щоб охарактеризувати кількісні особливості розподілу досліджуваних величин застосовується описова статистика. Основні завдання якої:

- Опис груп об'єктів дослідження.
- Статистична оцінка параметрів розподілу.
- Компактне візуальне уявлення даних про показниках.

Середнє арифметичне (**Mean**) (формула 1) є параметром вибору для опису математичного очікування для симетричних одновимірних розподілів.

$$\mu = \frac{1}{n}\sum_{i=1} x_i, \qquad (1)$$

де n - кількість випадків;

x_i - значення і-го випадку.

Для середнього арифметичного є одна чудова властивість: сума квадратів відхилень значень ознаки від значення міри центральної тенденції мінімальна щодо аналогічного показника для інших мір центральної тенденції (мода, медіана, середнє геометричне та ін.). Тим не менш, важливість середнього арифметичного для опису даних знижується пропорційно наростанню асиметрії розподілу досліджуваного параметра. Середнє арифметичне ідеально підходить для опису математичного очікування нормального розподілу (і його різновидів), добре для логістичних розподілів.

Мода (**Mode**) - це значення, що найбільш часто зустрічаєтеся у наборі даних (вибірці). Як міра центральної тенденції застосовується вкрай рідко, а як самостійний параметр - практично ніколи. Розподіл в залежності від кількості мод підрозділяються на мономодальні (один локальний максимум) і мультимодальні (декілька локальних максимумів).

Квантиль (**Quantile**) - це таке число, що задана випадкова величина не перевищує його з певною ймовірністю.

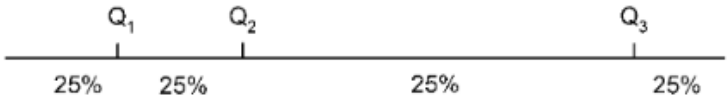

Рис.1. Квантилі

Медіана (**Median**) - це можливе значення ознаки, яка ділить ранжирувану сукупність на дві рівні частини: половина значень лежить вище медіани, половина - нижче. Якщо говорити більш звичною мовою, медіана - середня позиція у впорядкованому ряду значень. Вона добре підходить як міра центральної тенденції для одновимірних розподілів навіть в умовах вираженої асиметрії розподілу або при наявності виражених викидів значень. Якщо медіана ділить ранжирувану сукупність на дві рівні частини, то процентилі (**Percentile**) - це такі значення ознаки, які ділять її на 100 рівних частин.

Рис.2. Процентилі

Децілі - ділять ранжирувану сукупність на десять рівних частин.

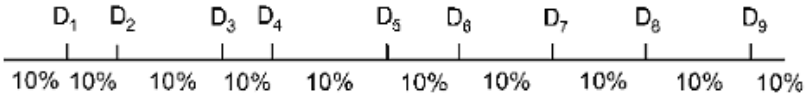

Рис.3. Децілі

Квартилі (**Quartile**) - ділять ранжирувану сукупність на чотири рівні частини. Заходи центральної тенденції при їх коректному використанні дають уявлення про стан розподілу і допомагають встановити його вид для подальших розрахунків.

Величина відсотка (рис.1, 2, 3), зазначена під інтервалом, означає частку об'єктів вибірки, що потрапили в цей інтервал.

Якщо медіана ділить дані порівну, то квартилі ділять їх на чотири частини. Вони позначаються Q1, Q2, Q3, Q4.

- Q1 - 25%
- Q2 - 50% (співпадає з медіаною)
- Q3 - 75%
- Q4 - 100%

Інтерквартильний розмах (IQR) =Q3 - Q1

Приклад

Нехай маємо ряд 62, 81, 63, 77, 63, 81, 65, 72, 72, 76.

Спочатку відсортуємо дані в зростаючому порядку:

62, 63, 63, 65, 72, 72, 76, 77, 81, 81.

Медіана(та Q2)): (72+72)/2 =72

Для обрахунку Q1 та Q3 значення медіани включаються до інтервалу.

Q1: (63+65)/2 =64 Q3: (76+77)/2 =76.5

Інтерквартильний розмах(ІКР): Q3 - Q1 = 76.5 - 64 = 12.5

Дисперсія та середньоквадратичне відхилення

Середньоквадратичне відхилення(standard deviation) дає розуміння, наскільки далеко знаходиться типове спостереження від середнього значення.

Дисперсія(variance) - обчислюється як середнє значення відстаней від всіх спостережень до середнього значення у квадраті.

$$var = \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad (2)$$

Середньоквадратичне відхилення (standard deviation) σ- обчислюється як корінь квадратний з дисперсії.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{N}} \quad (3)$$

numpy.std (a, axis = None, dtype = None, out = None, ddof = 0, keepdims = <no value>)

Функція std() обчислює середньоквадратичне (стандартне) відхилення значень елементів масиву.

Приклад

Нехай маємо ряд 5, 3, 2, 8, 2. Обрахуємо дисперсію та середньоквадратичне відхилення.

Середнє значення μ=4. Обчислимо дисперсію та середньоквадратичне відхилення:

Дисперсія (формула 2):

var=((5−4)^2+(3−4)^2+(2−4)^2+(8−4)^2+(2−4)^2)/5=(1^2+(−1)^2+(−2)^2 +4^2+(−2)^2)/5=(1+1+4+16+4)/5=5.2

Середньоквадратичне відхилення (формула 3): σ=2.28

Для ряду 3, 5, 5, 3, 4, середнє значення μ=4, var=0.8, середньоквадратичне відхилення: 0.89

Якщо поглянути на обидва ряди, бачимо що значення другого більш тісно розташовані навколо середнього. Значення дисперсії та середньоквадратичного відхилення дозволяють це виразити чисельно.

Приклад

Нехай Степан має 800 друзів у Facebook, Аня 1000 послідовників в Instagram. Хто більш популярний?

Для відповіді на це питання потрібні будуть дані про середнє значення та середньоквадратичне відхилення для кількості друзів у Facebook та послідовників у Instagram.

Facebook: μ=649, середньоквадратичне відхилення σ=50

Instagram: μ=843, середньоквадратичне відхилення σ=60

Обчислимо відстань до середнього значення в середньоквадратичних відхиленнях:

Степан: (x−μ)/σ=(800−649)/50 =3.02.

Аня: (x−μ)/σ =(1000−843)/60 =2.61.

Можемо вважати, що Степан більш популярний, оскільки значення кількості його друзів знаходиться далі від середнього значення.

Процес перетворення даних з допомогою формули $\frac{x-\mu}{\sigma}$ має назву ***z-стандартизація***, а отримані значення - z-значення ймовірності.

Розмах (**Range**) - це різниця між мінімальним і максимальним значеннями кількісного показника. Є досить грубою мірою розсіювання і застосовується досить рідко. В даний час часто використовується поняття «Non-Outlier Range», яке фактично є не різницю між максимумом і мінімумом, а різницю між 99 і 1-м процентилями. Використання даного параметра дозволяє знизити вплив одиничних викидів на уявлення про розсіяння.

Інтерквартильний розмах (**Interquartile Range**) - це різниця між значеннями верхньої і нижньої квартилі. Власне існують і інші варіанти квантильних «розмахів»: доцільний, інтерперсентільний тощо. Всі вони мають аналогічний сенс.

Метод describe

Метод **describe** показує основні статистичні характеристики даних по кожній числовій ознаці (типи int64 і float64): число непропущених значень, середнє, стандартне відхилення, діапазон, медіану, 0.25 і 0.75 квартилі.

```
df.describe()
```

Щоб подивитися статистику по нечисловим ознаками, потрібно явно вказати параметрі include.

```
df.describe(include=['object', 'bool'])
```

Щоб подивитися на розподіл користувачів по змінній XXX. Необхідно зазначити значення параметра normalize = True, щоб подивитися не абсолютні частоти, а відносні.

```
df[XXX].value_counts(normalize=True)
```

DataFrame можна впорядкувати за значенням будь-якої ознаки. Наприклад, за XXX (ascending = False для сортування за спаданням):

df.sort_values (by = 'XXX', ascending = False) .head ()

Сортувати можна і по групі стовпців:

df.sort_values (by = ['XXX', 'YYY'], ascending = [True, False]). head ()

Профілювання

Профілювання - процес, який допомагає зрозуміти дані, а Pandas Profiling - Python бібліотека, яка робить це. Простий і швидкий спосіб виконати попередній аналіз даних Python Pandas DataFrame.

Функція Pandas Profiling відображає багато інформації (значення, які найбільш часто зустрічаються; пропущені значення; кореляції; квантільна і описова статистика; довжина даних та інше.) за допомогою одного рядка коду і в інтерактивному HTML-звіті. Ця інформація необхідна для того, щоб знати, чи корисні дані для майбутнього використання.

pandas_profiling.ProfileReport(df)

Для встановлення бібліотеки скористайтеся наступним кодом:

```
! pip install https://github.com/ydataai/ydata-profiling/archive/refs/heads/master.zip
```

Далі потрібно перегрузити ядро.

Потім імпортувати бібліотеки:

```
from pandas_profiling import  ProfileReport
import pandas_profiling
```

Потім виконати профілювання дата сету.

Контрольні запитання

1. Що таке генеральна сукупність і вибірка з неї?
2. Що таке простий випадковий вибір?
3. Що таке варіаційний ряд?
4. Що таке статистичний ряд?

5. Дана вибірка (-1, 2, 0, -2, 3). Обчисліть mode.
6. Дана вибірка (-3, 2, 1, -2). Обчисліть розмах.
7. Дана вибірка (–2,1,2,0). Обчисліть σ^2
8. Дана вибірка (-4,2,1,0). Обчисліть μ.
9. Дана вибірка (-3,1,2,3,1,4, -5). середньоквадратичне відхилення
10. Дана вибірка (-5,1, -4,0,3,1). Обчисліть med.
11. Дана вибірка (1, -2,3,1,0,2,1, -3, -2). Складіть варіаційний ряд.
12. Дана вибірка (-3,0,3,5, -1,1,2, -1). Вкажіть нижню квартиль
13. Дана вибірка (1, -2,3,1,0,2,1, -3, -2). Обчисліть медіану.
14. Дана вибірка (1,2,1,2,2,3). Складіть статистичний ряд.
15. Дана вибірка (-2,1, -1,0,2,0,0,1). Знайдіть різницю x- med