Практическая работа № 8.1.

Описательные статистики и графики на Python.

Задание 8.1. Описательные статистики.

8.1.1. Основные понятия: случайная величина, наблюдение, генеральнаясовокупность и выборка.
8.1.2. Меры центра: выборочное среднее, истинное среднее, медиана, мода.
8.1.3. Квартили. Эксклюзивный метод подсчета.
8.1.4. Меры разброса: межквартильный размах, стандартное отклонение.
8.1.5. Составьте отчет.

Всего есть **три меры центральной тенденции**:

- Среднее (среднее арифметическое всех значений).
- Медиана (серединное значение).
- Мода (наиболее частое наблюдение).

Самые популярные меры разброса:

- Дисперсия.
- Стандартное отклонение.
- Размах.
- Межквартильный размах

КВАРТИЛИ

Квартили — это такие значения, которые делят наш набор данных **на четыре равные части** (по 25 % в каждой).

- Нижний квартиль Q1 отделяет 25 % наблюдений с наименьшими значениями от остальных 75 %.

- Верхний квартиль Q3 отделяет 25 % наблюдений с наибольшими значениями от остальных 75 %.

- Второй квартиль Q2— это медиана.

КАК НАХОДИТЬ КВАРТИЛИ?

1. Упорядочить элементы набора данных по возрастанию.
2. Найти медиану, разделить данные на две части относительно нее.
3. Набор данных:

Для примера рассмотрим упорядоченный по возрастанию ряд
4,11,12,20,23,23,30,31,32,33,34,36,38,40,41,44,44,44,45,47,48,49,54,56

В ряду четное количество (24) значений, центральные числа 36 и 38, поэтому медиана равна 37 (среднее арифметическое между 36 и 38). Когда в ряду количество нечетное, медиана равна значению в самой середине этого ряда. Со значением медианы всегда совпадает Q_2.

Первый квантиль Q_1 это медиана первой (нижней) половины ряда (12 значений). В нашем случае это среднее арифметическое между шестым (23) и седьмым (30) элементами ряда.

$$Q_1 = (23 + 30)/2 = 26,5$$

Третий квантиль Q_3 это медиана верхней половины ряда (12 значений). В нашем наборе данных это среднее арифметическое между семнадцатым (44) и восемнадцатым (45) элементами ряда.

$$Q_3 = (44 + 45)/2 = 44,5$$

У quantile() существует необязательный параметр, который указывает метод интерполяции, который нужно использовать, когда требуемый квантиль лежит между двумя точками данных. По умолчанию используется линейная интерполяция.

Пример 8.1.1

```
DataFrame.quantile(q=0.5, axis=0,
            numeric_only=True, interpolation='linear')
```

Чтобы вычислять квантиль как среднее между двумя точками в части данных, необходимо сменить вид интерполяции на следующий:

Пример 8.1.2

```
DataFrame.quantile(q=0.5, axis=0,
          numeric_only=True, interpolation='midpoint')
```

Размах

Размах — это разность между максимальным и минимальным наблюдением

$$range = x_{max} - x_{min}$$

В упомянутом выше наборе данных
4,11,12,20,23,23,30,31,32,33,34,36,38,40,41,44,44,44,45,47,48,49,54,56
Размах составляет 52.

$$range = 56 - 4 = 52$$

Межквартильный размах (*InterquartileRange. IQR*)

Межквартильный **р**азмах — это разность между третьим и первым квартилем. Для нашего набора составит 18.

$$IQR = Q_3 - Q_1 = 44{,}5 - 26{,}5 = 18$$

Стандартное отклонение

Стандартное отклонение — это мера стандартного (типичного) отклонения от среднего. Бывает истинным и выборочным. Истинное считается на генеральной совокупности, а выборочное — на выборке.

Пример 8.1.3

Вычисление основных статистических показателей с применением интерполяции.

```
import  pandas as pd
myData=pd.Series([4,11,12,20,23,23,30,31,32,33,34,36,
38,40,41,44,44,44,45,47,48,49,54,56])
mean = myData.mean() # Среднее значение
var = myData.var() # Дисперсия
std = myData.std()# Стандартное отклонение
mode = myData.mode()# Мода – наиболее частое наблюдение
median = myData.median()# Медиана (Q2)
perc25 = myData.quantile(0.25) # Q1
perc75 = myData.quantile(0.75) # Q3
IQR = perc75 - perc25
print(mean,var,std,perc25,median,perc75,IQR,mode)

34.958333333333336                    189.51992753623188
13.766623679618466 28.25 37.0 44.25 16.0 0    44
```

Выбросы

Для нахождения выбросов нужно знать квартили и межквартильный размах. Мы считаем наблюдение выбросом в следующих случаях:

- Если наблюдение меньше, чем значение нижнего квартиля Q1 минус 1.5 межквартильного размаха.
- Если наблюдение больше, чем значение верхнего квартиля Q3 плюс 1.5 межквартильного размаха.

Что делать с выбросами? Их можно удалить перед подсчетом описательных статистик и отдельно упомянуть в отчёте, что такие наблюдения были.