ПРАКТИЧЕСКАЯ РАБОТА №4.
«ИССЛЕДОВАНИЕ ДАННЫХ НА PYTHON. ОПИСАТЕЛЬНАЯ СТАТИСТИКА»

Область **статистики** можно рассматривать как научную среду для работы с данными. Это определение включает все задачи, связанные со сбором, анализом и интерпретацией данных. Также статистика может относиться к отдельным измерениям, которые представляют собой сводную информацию по данным или определенные их аспекты.

Когда у нас есть набор наблюдений, полезно свести признаки наших данных в одно определение. Этим занимается **описательная статистика**. Как следует из названия, описательная статистика описывает конкретное свойство данных, которые она обобщает. Такую статистику можно разделить на две категории: *меры центральной тенденции* (или *меры центра*) и *меры разброса*.

Описательные статистические величины, или статистики, — числа, которые используются для обобщения и описания данных. Описательные статистики, так называемые *сводные статистики*, представляют собой разные подходы к измерению свойств последовательностей чисел. Они помогают охарактеризовать последовательность и способны выступать в качестве ориентира для дальнейшего анализа.

Всего есть **три меры центральной тенденции**:

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

Но найденная мера центральной тенденции не позволит нам полностью описать признак. Например, мы знаем, что средний чек в ресторане за неделю составил 1000 руб. Это значение может быть получено разными способами: например, все посетители потратили по 1000 руб. Ровно такое же среднее арифметическое будет в ситуации, когда половина посетителей потратила по 500 руб., а другая — по 1500 руб. Конечно, это будут совершенно разные ситуации.

Разница между медианой и средним значением существует из-за **робастности (выбросоустойчивости)**.

Проблема выбросов. Если в данных есть выбросы — значения, которые гораздо выше или ниже остальных, — это может негативно повлиять на среднее значение. Таким образом, среднее значение не робастно, а медиана — напротив, выбросоустойчива.

Выбросы могут отражать интересные события или ошибки в наборе данных, поэтому важно уметь определять их наличие. Сравнение медианы и моды — один из способов определить наличие выбросов, хотя визуализация обычно позволяет сделать это быстрее.

Поэтому ещё один вопрос, который необходимо задать: **Насколько сильно разбросаны друг относительно друга** значения?

Для ответа на такой вопрос существуют меры разброса. Меры разброса отвечают на вопрос: «Как сильно варьируются мои данные?». В мире существует не так много вещей, которые остаются в одном и том же состоянии при каждом наблюдении. Эта изменчивость делает мир нечётким и неопределённым, поэтому полезно иметь показатели, которые могут обобщить эту «нечёткость».

Самые популярные **меры разброса:**

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

Ключевые идеи:

- описательная статистика используется для систематизации и количественного описания данных;
- среднее значение указывает на типичное значение в наборе данных. Оно не робастно;
- медиана является центральным значением в ряду данных. Она робастна;
- мода — значение, которое появляется наиболее часто;
- размах — разность между максимальным и минимальным значениями в наборе данных;
- дисперсия и стандартное отклонение являются средним расстоянием от среднего арифметического значения.

Среднее значение. Наиболее распространенный способ усреднить набор данных — взять его среднее значение. Среднее значение на самом деле представляет собой один из нескольких способов измерения *центра распределения* данных.

Библиотека Pandas содержит функцию **mean()**.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

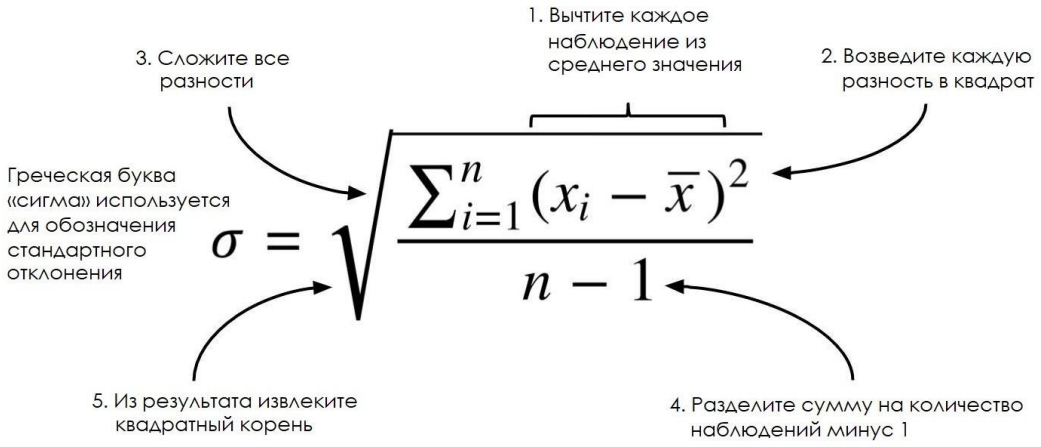
Медиана. Медиана — еще одна распространенная описательная статистика для измерения центра распределения последовательности. Чтобы найти медиану, данные нужно расположить в порядке возрастания. Медианой будет значение, которое совпадает с серединой набора данных. Если в последовательности число точек данных четное, то медиана определяется, как полусумма двух срединных значений.

Библиотека Pandas содержит функцию **median()**.

Мода. Определяется как значение, которое наиболее часто встречается в наборе данных. Мода не так очевидно соответствует понятию «середины» как среднее значение или медиана, но это соответствие абсолютно обосновано: если значение появляется в данных неоднократно, оно приблизит среднее значение к моде. Чем чаще появляется значение, тем сильнее оно влияет на среднее. Таким образом, мода показывает наиболее значимый фактор, формирующий среднее значение.

Размах. Определяется как разность между максимальным и минимальным значениями в наборе данных.

Стандартное отклонение. Помогает узнать, как сильно данные отличаются от типичного значения. Иными словами, оно говорит о том, как сильно данные отличаются от среднего арифметического.



Чем больше стандартное отклонение, тем больше рассеяны данные вокруг среднего значения, и наоборот.

Дисперсия. Часто стандартное отклонение и дисперсию связывают вместе.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

В библиотеке Pandas функции для вычисления дисперсии (варианса) и стандартного отклонения имплементированы соответственно, как **var()** и **std().** При этом последняя по умолчанию вычисляет несмещенное значение, поэтому, чтобы получить тот же самый результат, нужно применить именованный аргумент **ddof=0**, который сообщает, что требуется вычислить смещенное значение стандартного отклонения:

```
load_uk_scrubbed()['Electorate'].std( ddof=0 )
```

Квантили. Медиана представляет собой один из способов вычислить *срединное* значение из списка, т.е. находящееся ровно по *середине*, дисперсия же предоставляет способ измерить разброс данных вокруг среднего значения. Если весь разброс данных представить на шкале от 0 до 1, то значение 0.5 будет медианным.

Для примера рассмотрим следующую ниже последовательность чисел:

```
[10 11 15 21 22.5 28 30]
```

Отсортированная последовательность состоит из семи чисел, поэтому медианой является число 21 четвертое в ряду. Его также называют **0.5-квантилем**. Мы можем получить более полную картину последовательности чисел, взглянув на 0.0 (нулевой), 0.25, 0.5, 0.75 и 1.0 квантили. Все вместе эти цифры не только показывают медиану, но также обобщают диапазон данных и сообщат о характере распределения чисел внутри него. Они иногда упоминаются в связи с *пятичисловой сводкой*.

Квантили можно вычислить непосредственно в Pandas при помощи функции **quantile()**. Последовательность требующихся квантилей передается в виде списка.

```
def ex_1_10():
    '''Вычислить квантили:
       возвращает значение в последовательности xs,
       соответствующее p-ому проценту'''
    q = [0, 1/4, 1/2, 3/4, 1]
    return load_uk_scrubbed()['Electorate'].quantile(q=q)
```

Когда квантили делят диапазон на четыре равных диапазона, как показано выше, то они называются ***квартилями.*** Разница между нижним (0.25) и верхним (0.75) квартилями называется **межквартильным размахом**, или иногда сокращенно **МКР**. Аналогично дисперсии (варианса) вокруг среднего значения, межквартильный размах измеряет разброс данных вокруг медианы.

Выполнить пример.

Для применения статистики следует загрузить библиотеки: math, numpy, pandas, statistics, scipy.stats. Изучите, каким образом можно рассчитать центральные метрики, средневзвешенное, гармоническое среднее, среднее геометрическое, медиану, моду, дисперсию, среднеквадратичное отклонение, смещение, процентили, диапазон. Программный код для расчёта данных показателей приведён ниже.

```
import math
import statistics
import numpy as np
import scipy.stats
import pandas as pd
print("Исходные данные")
x = [8.0, 1, 2.5, 4, 28.0]
x_with_nan = [8.0, 1, 2.5, math.nan, 4, 28.0]
y, y_with_nan = np.array(x), np.array(x_with_nan)
z, z_with_nan = pd.Series(x), pd.Series(x_with_nan)
print(y)
print(y_with_nan)
print(z)
print(z_with_nan)

# Среднее значение
print("Среднее значение")
mean_=sum(x)/len(x)
print(mean_)
mean_=statistics.mean(x)
print(mean_)
m=np.nanmean(y_with_nan)
print(m)
```

```
# Средневзвешенное значение
print("Средневзвешенное значение")
x = [8.0, 1, 2.5, 4, 28.0]
w = [0.1, 0.2, 0.3, 0.25, 0.15]
wmean = sum(w[i] * x[i] for i in range(len(x))) / sum(w)
print(wmean)
wmean = sum(x_ * w_ for (x_, w_) in zip(x, w)) / sum(w)
print(wmean)
# Средневзвешенное значение, использование массивов Numpy и Pandas
x = [8.0, 1, 2.5, 4, 28.0]
y, z, w = np.array(x), pd.Series(x), np.array(w)
wmean = np.average(y, weights=w)
print(wmean)
wmean = np.average(z, weights=w)
print(wmean)

# Гармоническое среднее
print("Гармоническое среднее")
hmean = len(x) / sum(1 / item for item in x)
print(hmean)
hmean==scipy.stats.hmean(y)
print(hmean)

# Среднее геометрическое
print("Среднее геометрическое")
gmean = 1
for item in x:
    gmean *= item
gmean **= 1 / len(x)
print(gmean)


# Медиана
print("Медиана")
n = len(x)
if n % 2:
    median_ = sorted(x)[round(0.5*(n-1))]
else:
    x_ord, index = sorted(x), round(0.5 * n)
    median_ = 0.5 * (x_ord[index-1] + x_ord[index])
print(median_)
print(z.median())
print(z_with_nan.median())

# Медиана
u = [2, 3, 2, 8, 12]
mode_ = max((u.count(item), item) for item in set(u))[1]
print(mode_)


# Дисперсия
print("Дисперсия")
n = len(x)
mean_ = sum(x) / n
var_ = sum((item - mean_)**2 for item in x) / (n - 1)
print(var_)
```

```python
# Среднеквадратическое отклонение
print("Среднеквадратическое отклонение")
std_ = var_ ** 0.5
print(std_)
std_=np.std(y, ddof=1)
print(std_)

# Смещение
print("Смещение")
y, y_with_nan = np.array(x), np.array(x_with_nan)
print(scipy.stats.skew(y, bias=False))
print(scipy.stats.skew(y_with_nan, bias=False))

# Процентили
print("Процентили")
y = np.array(x)
print(np.percentile(y, 5))
print(np.percentile(y, 95))

# Диапазон
print("Диапазон")
print(np.amax(y) - np.amin(y))
print(np.nanmax(y_with_nan) - np.nanmin(y_with_nan))
print(y.max() - y.min())
print(z.max() - z.min())
print(z_with_nan.max() - z_with_nan.min())
```

Результат вычислений:

```
2.7613412228796843
Среднее геометрическое
4.677885674856041
Медианна
4
4.0
4.0
2
Дисперсия
123.19999999999999
Среднеквадратическое отклонение
11.099549540409285
11.099549540409285
Смещение
1.9478432273905927
nan
Процентили
1.3
23.999999999999996
Диапазон
27.0
27.0
27.0
27.0
27.0
```

Задание 1.

Загрузить датасет **forestfires.csv** в среду Google Colab. В файле **forestfires.csv** находятся данные о лесных пожарах.

Выполнить разведочный анализ (EDA).

Исследование данные о пожарах, использую описательную статистику.

Оформить отчет по выполненной работе и сдать преподавателю отчет в формате *.pdf и файл с выполненным домашним заданием.

Состав и описание полей файла forestfires.csv

X пространственная координата по оси x на карте парка Монтесиньо: от 1 до 9
Y пространственная координата по оси Y на карте парка Монтесиньо: от 2 до 9
month (месяц) месяц года: от "янв" до "декабрь"
day (день) день недели: от "пн" до "вс"
FFMC индекс FFMC из системы FWI: от 18,7 до 96,20.
DMC индекс DMC из системы FWI: от 1,1 до 291,3
DC индекс DC из системы FWI: от 7,9 до 860,6
ISI индекс ISI из системы FWI: от 0,0 до 56,10
temp температура в градусах Цельсия: от 2,2 до 33,30
RH относительная влажность в %: от 15,0 до 100
wind (ветер) скорость ветра в км/ч: от 0,40 до 9,40
rain (дождь) дождь снаружи в мм/м2: от 0,0 до 6,4
area сгоревшая площадь леса (в га): от 0,00 до 1090,84