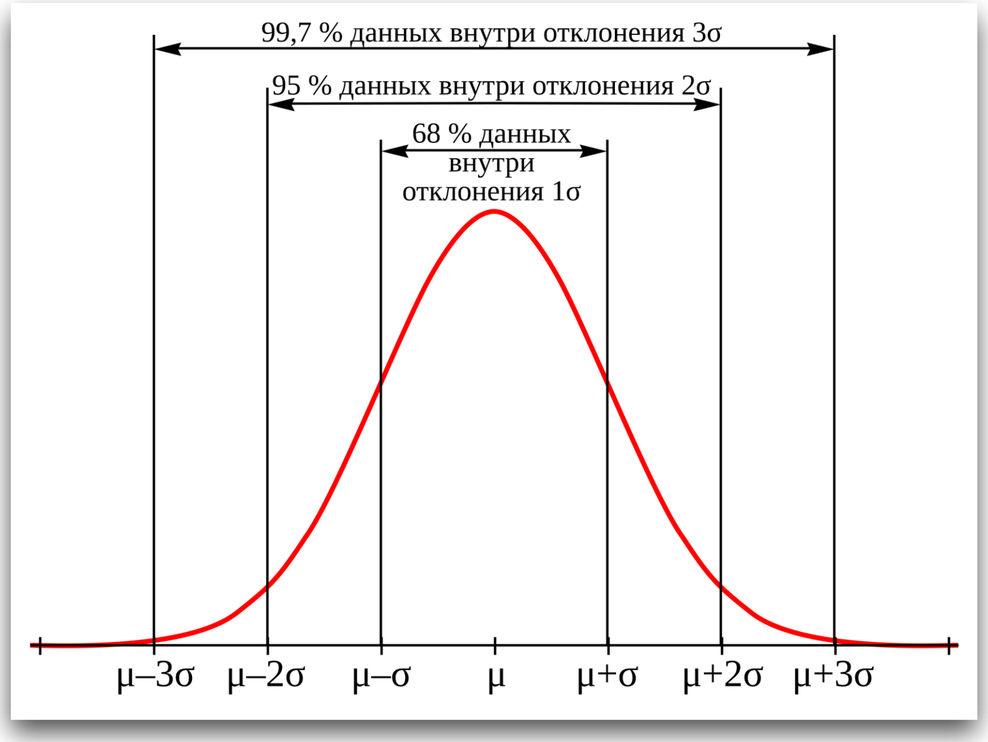
Основы искусственного интеллекта

Введение в описательную статистику.

Описательная статистика

Определение и цель

- Описательная статистика занимается сбором, обработкой, представлением и анализом данных.
- Основная цель описательной статистики – резюмировать данные так, чтобы их можно было легко интерпретировать. Одними из ключевых показателей являются **среднее**, **медиана** и **мода**, которые помогают понять центральную тенденцию данных.



Среднее (Mean)

- **Среднее арифметическое** – это сумма всех значений, деленная на их количество. Оно показывает «среднее» значение данных и чувствительно к выбросам.

Пример: Даны числа

1, 2, 3, 4, 5, 6, 7, 8, 9

$$\bar{x} = \frac{1+2+3+4+5+6+7+8+9}{9} = 5$$

среднее арифметическое

Формула среднего арифметического:

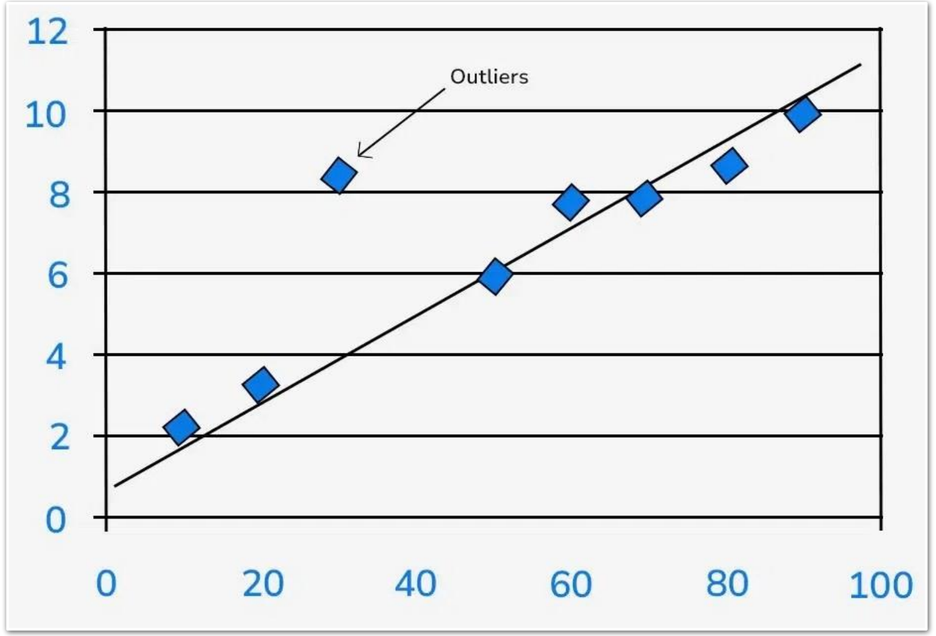
$$\bar{x} = \frac{\sum x_i}{n}$$

где:

- x_i – отдельные значения выборки,
- n – количество значений в выборке.

Выбросы/Посторонние значения (outliers)

- **Выброс/постороннее значение (outlier)** — это точка данных, которая выходит за рамки общей картины набора данных и существенно отличается от других наблюдений.



Выбросы/Посторонние значения (outliers)

Пример

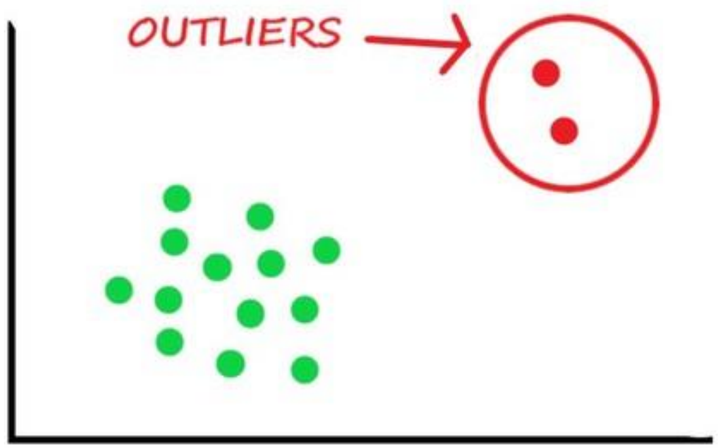
- Набор данных: 1, 2, 4, 6, 8, 3, 1000, 9, 12, 158, 10

  Выбросы 1000 и 158.

Выбросы/Посторонние значения (outliers)

Причины появления

- Ошибки ввода данных
- Изменчивость выборки
- Ошибки измерения
- Настоящие аномалии



Медиана (Median)

- **Медиана** – это значение, которое делит упорядоченный набор данных на две равные части. Если количество значений нечетное, медиана – это центральное значение. Если четное – медиана находится как среднее двух центральных значений.
- **Алгоритм нахождения медианы:**

  1. Упорядочить данные по возрастанию.
  2. Найти центральное значение.
     - Если количество элементов **нечетное**, медиана – это средний элемент.
     - Если **четное**, медиана – это среднее двух центральных элементов.

Медиана (Median)

- **Пример:**
  - Для чисел **3, 4, 5, 7, 8, 9, 11** медиана равна **7** (средний элемент в упорядоченном массиве при нечетном количестве элементов).
  - Для чисел **1, 3, 4, 5, 7, 8, 9, 12** медиана равна (5+7)/2 = **6**. (среднее арифметическое двух центральных элементов в упорядоченном массиве при четном количестве элементов)

Медиана менее чувствительна к выбросам по сравнению со средним.

Мода (Mode)

- **Мода** – это наиболее часто встречающееся значение в наборе данных. Возможны три случая:
  - Одна мода (мономодальное распределение).
  - Две моды (бимодальное распределение).
  - Более двух мод (мультимодальное распределение).

Мода (Mode)

- **Пример:**
  - В наборе данных **2, 3, 3, 4, 5, 5, 6, 6, 7, 7, 8, 2, 3** мода равна **3**, так как это значение встречается чаще всего.
  - В данных **1, 1, 2, 2, 2, 3, 3, 3, 4, 5, 4** моды = **2 и 3** (бимодальное распределение).

Мода особенно полезна для анализа категориальных данных (например, самый популярный тренд, товар, еда и т. д.).

Сравнение среднего, медианы и моды

Показатель	Определение	Чувствительность к выбросам	Когда использовать
Среднее	Средняя величина данных	Высокая	Когда данные распределены нормально
Медиана	Центральное значение в отсортированном ряду	Низкая	Когда есть выбросы или асимметричное распределение
Мода	Самое частое значение	Низкая	Для категориальных данных или определения модального класса

Работа в Pandas

Импорт библиотеки и создание данных

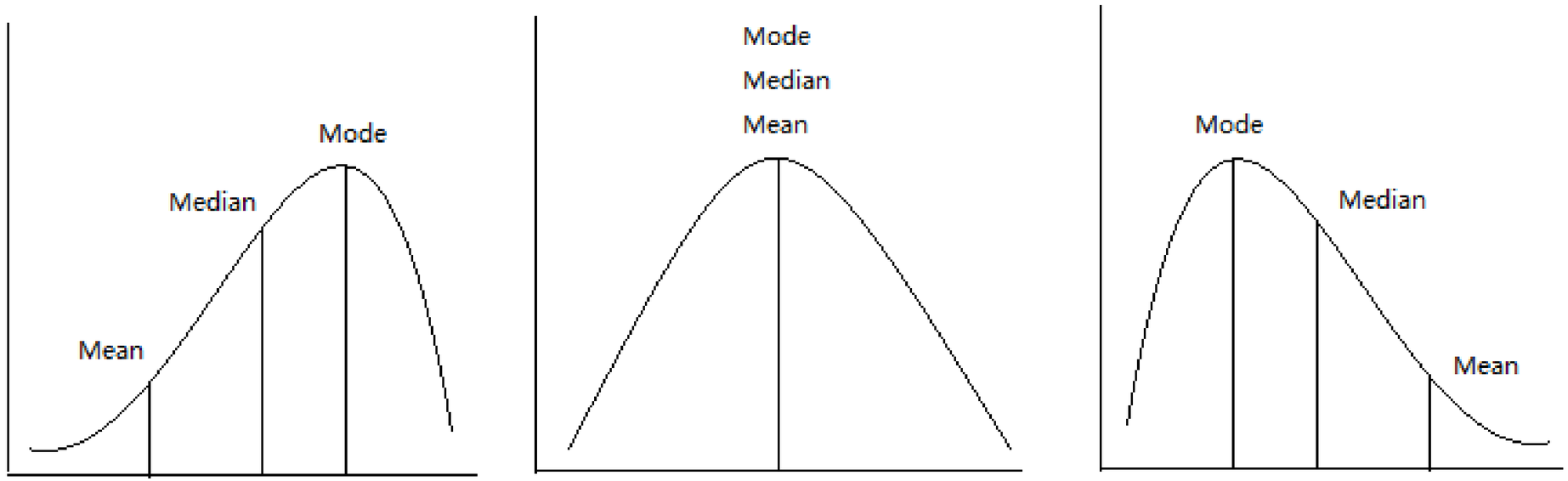
```
import pandas as pd

import numpy as np

# Создадим DataFrame с примерными данными

data = {'Значения': [10, 15, 20, 25, 30, 35, 40, 100]}  # 100 - выброс

df = pd.DataFrame(data)
```

Работа в Pandas

Среднее арифметическое, медиана, мода

```
# Среднее арифметическое
mean_value = df['Значения'].mean()
# Медиана
median_value = df['Значения'].median()
# Мода (может быть несколько значений)
mode_value = df['Значения'].mode().values
print(f"Среднее: {mean_value}")
print(f"Медиана: {median_value}")
print(f"Мода: {mode_value}")
```

Сравнение среднего, медианы и моды



Метрики центральной тенденции

- Они показывают, вокруг каких значений концентрируются данные.
  - **Среднее арифметическое (Mean)** – показывает «типичное» значение.
  - **Медиана (Median)** – полезна при наличии выбросов, так как не зависит от экстремальных значений.
  - **Мода (Mode)** – наиболее часто встречающееся значение, важно в дискретных данных.
- 🛠️ **Применение в ML:**

  Используется для проверки баланса данных (например, классов в задаче классификации).

Размах (Range)

Размах – это разница между максимальным и минимальным значением в наборе данных.

Формула:

$$Range = X_{max} - X_{min}$$

Пример:

Если даны числа **3, 3, 7, 2, 9, 5, 2, 5, 8, 6, 10,** то размах:

10 – 2 = 8

Размах показывает разброс данных, но чувствителен к выбросам.

Работа в Pandas

Размах (range)

```
range_value = df['Значения'].max() - df['Значения'].min()

print(f"Размах: {range_value}")
```

Дисперсия (Variance)

Дисперсия измеряет, насколько сильно данные отклоняются от среднего.

Формула:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Чем больше дисперсия, тем сильнее разброс данных.

Стандартное отклонение

Standard Deviation

Стандартное отклонение - это **квадратный корень из дисперсии.** Оно показывает, насколько типичные значения отклоняются от среднего.

Формула:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Чем больше стандартное отклонение, тем больше разброс данных.

Работа в Pandas

Дисперсия и Стандартное Отклонение

```
# Дисперсия

variance_value = df['Значения'].var()

# Стандартное отклонение

std_dev = df['Значения'].std()

print(f"Дисперсия: {variance_value}")

print(f"Стандартное отклонение: {std_dev}")
```

Метрики разброса (дисперсия и отклонение)

- Они показывают, насколько сильно варьируются данные.
  - **Размах (Range)** – простейший показатель разброса
  - **Дисперсия (Variance)** – средний квадрат отклонений от среднего.
  - **Стандартное отклонение (Standard Deviation)** – корень из дисперсии, показывает типичное отклонение данных от среднего.
- 🛠️ **Применение в ML:**
  Если разброс данных велик, может потребоваться нормализация или стандартизация (например, StandardScaler в sklearn).

Нормализация

Нормализация - это процесс преобразования данных в единую шкалу, чтобы избежать влияния различных диапазонов значений. Она помогает улучшить качество моделей машинного обучения и повышает точность вычислений.

Зачем нужна нормализация?

- Устраняет различие в масштабах данных
- Повышает точность моделей машинного обучения
- Ускоряет процесс обучения
- Улучшает интерпретируемость результатов

Методы нормализации

1. Min-Max нормализация: $X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$

- Переводит значения в диапазон [0,1] или [-1,1]
- Чувствительна к выбросам

2. Z-score стандартизация: $X_{stand} = \frac{X - \mu}{\sigma}$

- Среднее значение = 0, стандартное отклонение = 1
- Подходит для данных с нормальным распределением

Методы нормализации

3. Логарифмическое преобразование: $X_{log} = log(X + 1)$

- Уменьшает влияние выбросов
- Применяется к данным с высокой асимметрией

4. Нормализация по квантилям (RobustScaler):

- Применяется к данным с выбросами
- Использует медиану и межквартильный размах

Метрики для описания данных

- **Коэффициент вариации** (Coefficient of Variation, CV)
- **Квартили** (Quartiles)
- **Межквартильный размах** (IQR - Interquartile Range)
- **Коэффициент асимметрии** (Skewness)
- **Эксцесс** (Kurtosis)

Коэффициент вариации

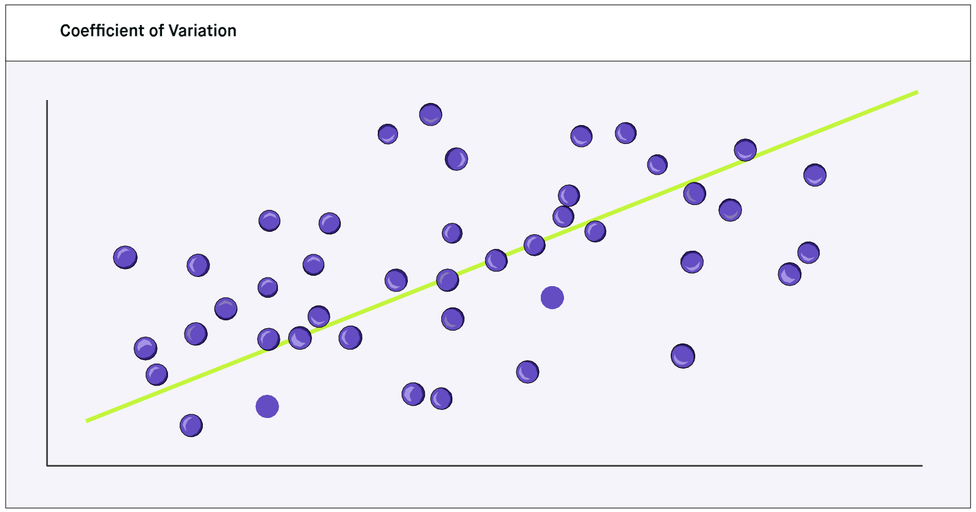
Coefficient of Variation, CV

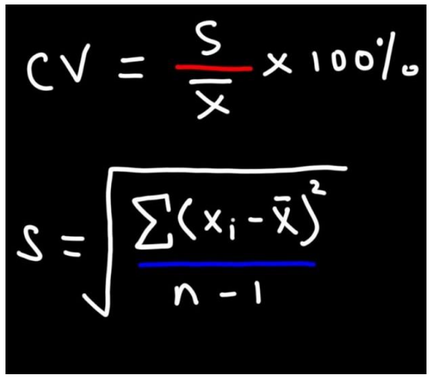
Измеряет относительный разброс данных, выражается в процентах.

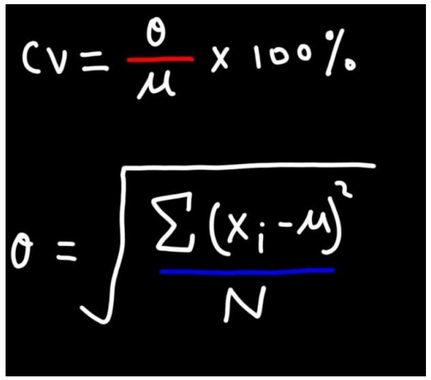
Формула:

$$CV = \left(\frac{s}{\bar{x}}\right) \times 100\%$$

Чем выше CV, тем больше изменчивость данных.







Работа в Pandas

Коэффициент вариации

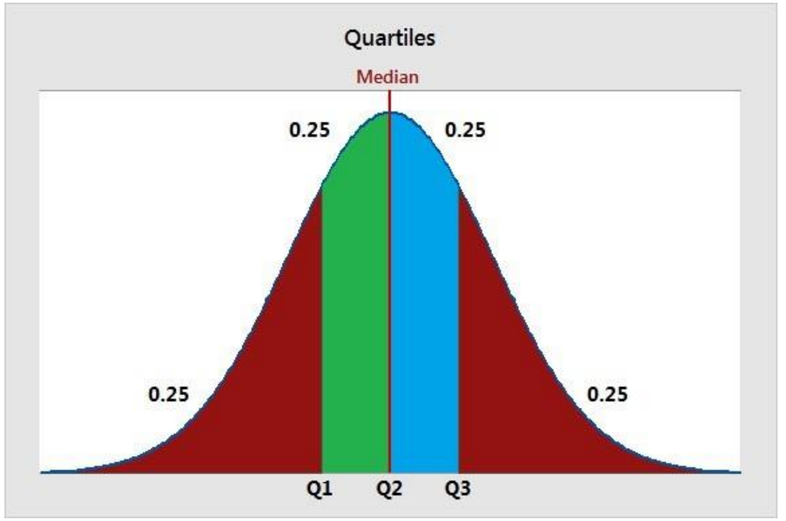
```
cv_value = (std_dev / mean_value) * 100

print(f"Коэффициент вариации: {cv_value:.2f}%")
```

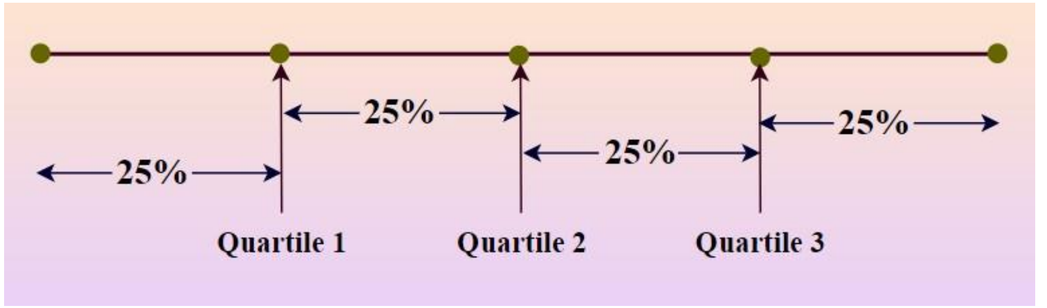
Квартили
Quartiles

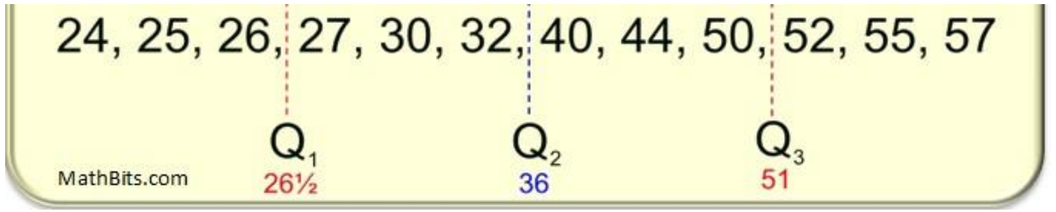
Разделяют данные на 4 равные части.

- **Q1 (первый квартиль)** – 25% данных ниже этого значения.
- **Q2 (медиана)** – 50% данных ниже.
- **Q3 (третий квартиль)** – 75% данных ниже.

- **Пример:**
  Если данные: **2, 4, 6, 8, 10**, то:
  - Q1 = 3
  - Медиана (Q2) = 6
  - Q3 = 9





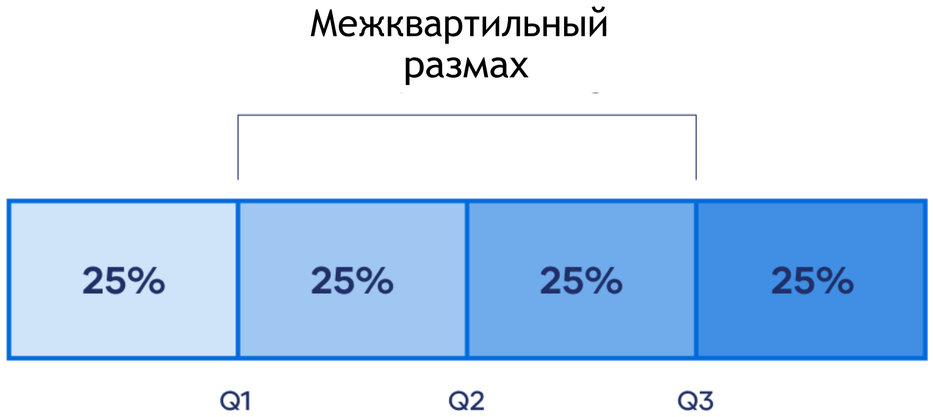


Межквартильный размах

IQR - Interquartile Range

Разница между третьим и первым квартилями:

$$IQR = Q3 - Q1$$



Используется для выявления выбросов (outliers): значения, выходящие за диапазон **[Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]**, считаются выбросами.

Работа в Pandas

Квартили и межквартильный размах

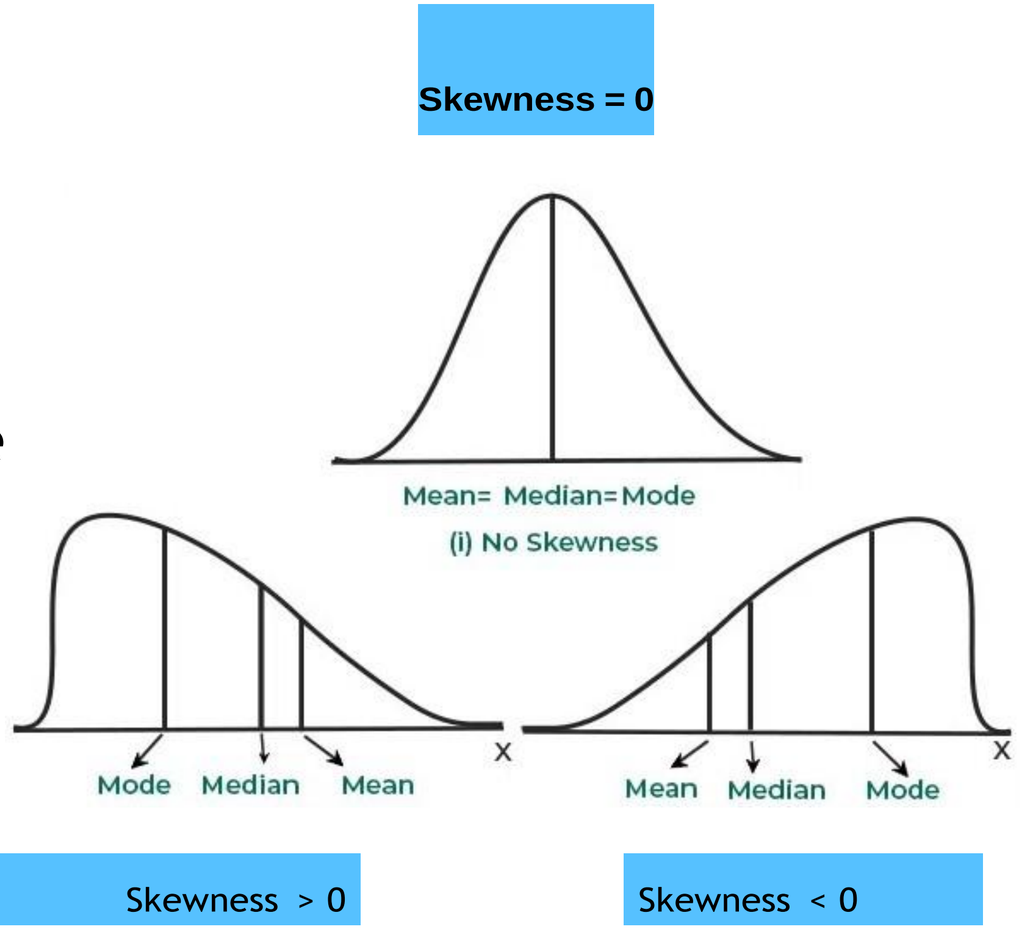
```
q1 = df['Значения'].quantile(0.25)          # Первый квартиль
q3 = df['Значения'].quantile(0.75)          # Третий квартиль


iqr = q3 - q1  # Межквартильный размах

print(f"Q1 (первый квартиль): {q1}")

print(f"Q3 (третий квартиль):
{q3}")print(f"Межквартильный размах (IQR): {iqr}")
```

Коэффициент асимметрии

Skewness

Показывает, насколько распределение данных **симметрично**.

- **Skewness > 0** – правосторонняя (положительная) асимметрия.
- **Skewness < 0** – левосторонняя (отрицательная) асимметрия.
- **Skewness ≈ 0** – симметричное распределение.



$$skewness = \frac{(x - \mu)^3}{\sigma^3}$$

Работа в Pandas

Коэффициент асимметрии (Skewness)

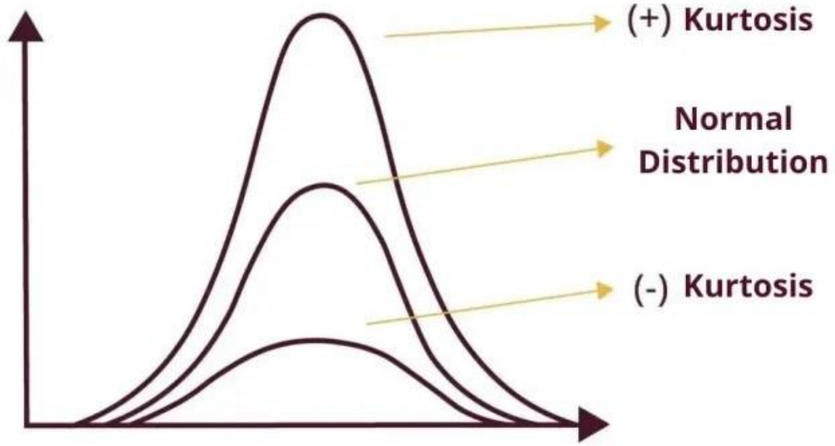
```
skewness = df['Значения'].skew()

print(f"Асимметрия (Skewness): {skewness}")
```

Эксцесс

Kurtosis

Измеряет «пикообразность» распределения.

- **Kurtosis > 3** – распределение с "тяжёлыми хвостами" (острое).
- **Kurtosis < 3** – распределение более плоское.
- **Kurtosis = 3** – нормальное распределение.



$$kurtosis = \frac{(x - \mu)^4}{\sigma^4}$$

Работа в Pandas

Эксцесс (Kurtosis)

```
kurtosis = df['Значения'].kurt()

print(f"Эксцесс (Kurtosis): {kurtosis}")
```

Работа в Pandas

Полная статистика с `.describe()`

```
print(df.describe())
```

Вывод .describe() включает:

- **count** - количество значений
- **mean** - среднее
- **std** - стандартное отклонение
- **min** - минимальное значение
- **25%, 50%, 75%** - квартильные значения
- **max** - максимальное значение

Вывод

- Для комплексного анализа данных полезно использовать **несколько показателей** сразу, так как каждый из них даёт разную информацию о распределении и разбросе данных.
- **Pandas** позволяет легко рассчитывать основные статистические показатели.
- **Среднее, медиана, мода** дают представление о центральной тенденции.
- **Размах, дисперсия, стандартное отклонение** показывают разброс данных.
- **Коэффициент асимметрии и эксцесс** помогают анализировать форму распределения.

Спасибо за внимание!