

# Описательная статистика

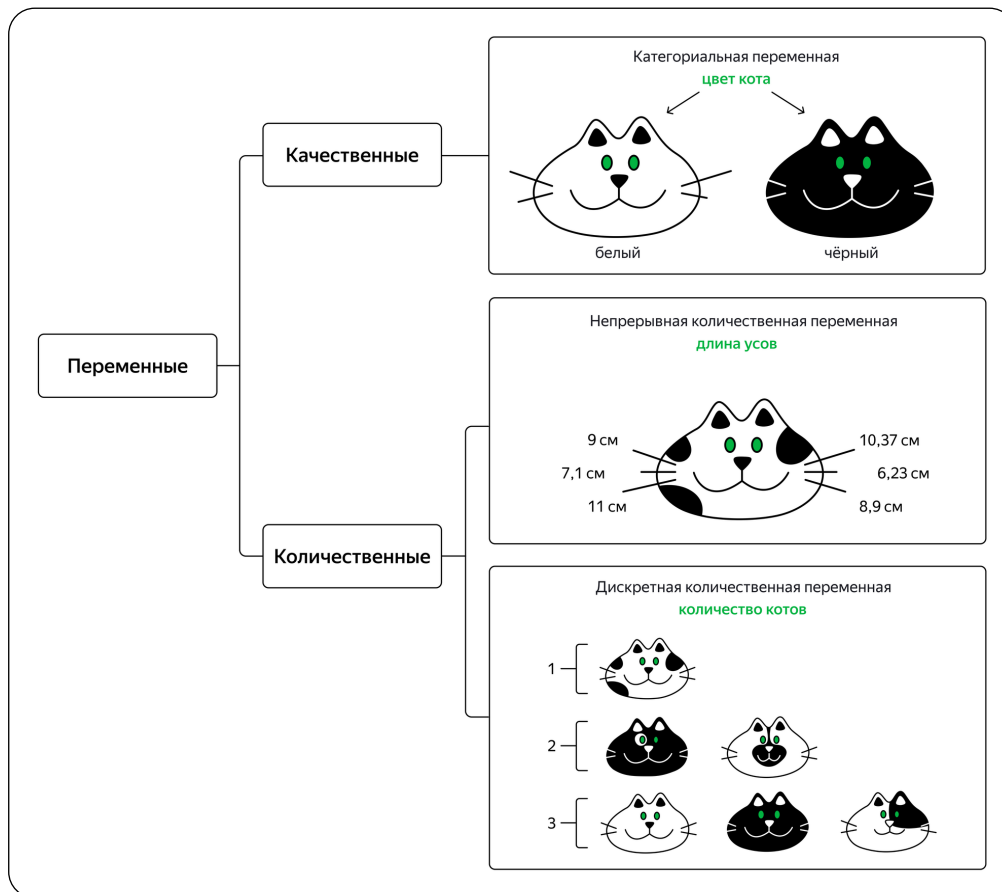
## Типы переменных

Переменные бывают **категориальными** и **количественными**.

Переменные, которые выражают категорию или характеристику, называются **категориальными**.

Переменные, которые описывают численную характеристику, называются **численными**.

- Количественные переменные, которые могут принимать любое численное значение в определенном числовом промежутке, называются **непрерывными**.
- Количественные переменные, которые могут принимать только числовые значения из определённого набора значений, называются **дискретными**.



## Характеристики для описания положения центра набора данных

**Мода** — это значение, которое величина принимает наиболее часто.

## Код: вычисление на Python

```
import pandas as pd

data = pd.Series([1, 4, 1, 6, 7, 1]) # Анализируемый набор данных
print(f'Мода набора данных равна {data.mode()[0]}') # Вывод значения
```

## Описательная статистика

**Медиана** — значение, которое делит набор данных на две равные части: в первой части находятся значения, которые меньше самой медианы, во второй — значения, которые больше неё.

### Код: вычисление на Python

```
import pandas as pd

data = pd.Series([2.3, 2.1, 2.4, 2.5, 2]) # Анализируемый набор данных
print('Медиана исходного набора данных равна', data.median()) # Вывод значения
```

**Среднее арифметическое** — значение, равное сумме всех элементов набора данных, разделенной на количество этих элементов.

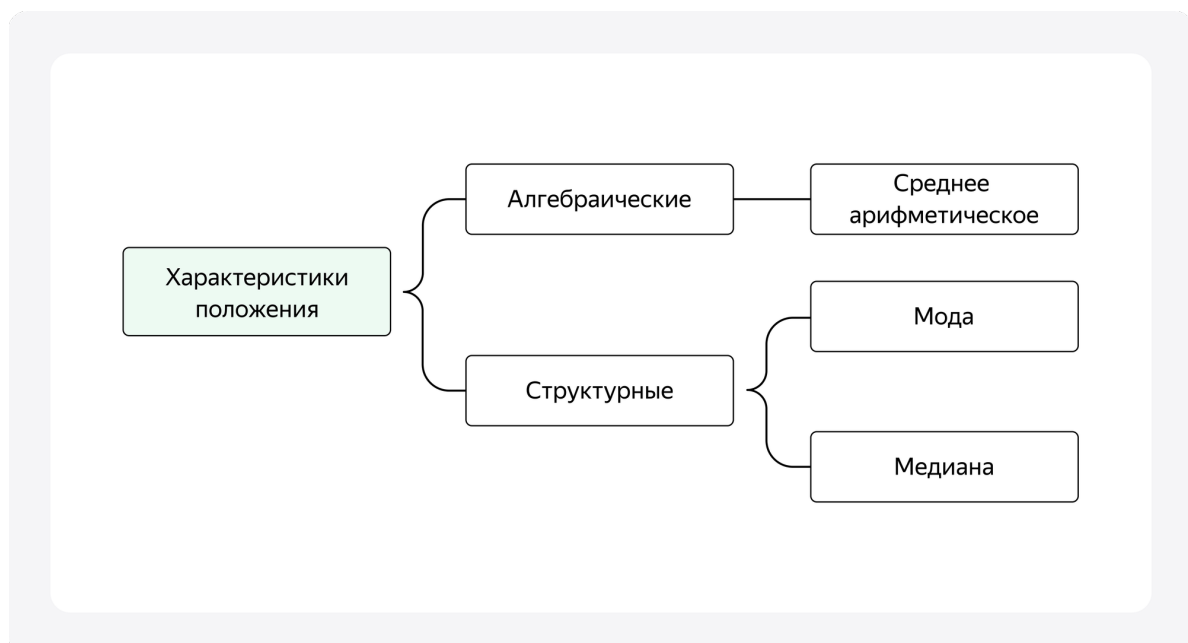
### Код: вычисление на Python

```
import pandas as pd

data = pd.Series([2, 3, 1, 5, 1]) # Анализируемый набор данных
print('Среднее арифметическое исходного набора данных равно', data.mean()) # Вывод значения
```

Мода и медиана являются **структурными** характеристиками положения, так как их вычисляют, исходя из структуры набора данных.

Среднее арифметическое является **алгебраической** характеристикой положения, так как его вычисляют с помощью алгебраических операций над элементами набора данных.



## Описательная статистика

### Характеристики для описания разброса набора данных

**Дисперсия**  $s^2$  для набора данных из  $n$  элементов:  $x_1, x_2, \dots, x_n$  вычисляется по формуле:

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

где  $\bar{x}$  — это среднее значение набора данных  $x_1, x_2, \dots, x_n$ .

### Код: вычисление на Python

```
import pandas as pd

data = pd.Series([2, 3, 1, 5, 1]) # Анализируемый набор данных
print('Дисперсия исходного набора данных равна', data.var()) # Вывод значения
```

**Стандартное отклонение** для выборки  $x_1, x_2, \dots, x_n$  — это корень из дисперсии набора данных:

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

где  $\bar{x}$  — это среднее значение набора данных  $x_1, x_2, \dots, x_n$ .

### Код: вычисление на Python

```
import pandas as pd

data = pd.Series([2, 3, 1, 5, 1]) # Анализируемый набор данных
print('Стандартное отклонение исходного набора данных равно', data.std()) # Вывод значения
```

**Размах набора данных** — это разница между максимальным и минимальным значениями набора данных.

### Код: вычисление на Python

```
import pandas as pd

data = pd.Series([1, 2, 3, 4, 5, 6, 7, 8]) # Анализируемый набор данных

range_value = data.max() - data.min() # Вычисление размаха
print(range_value) # Вывод значения
```

## Описательная статистика

**Межквартильный размах** — это размах 50% значений в центре набора данных.

**Код: вычисление на Python**

```
import pandas as pd
from scipy import stats

data = pd.Series([3, 1, 2, 5, 6, 0, 10]) # Анализируемый набор данных

print(stats.iqr(data)) # Вывод значения
```

**Процентилем** некоторого уровня (p%) называют значение из набора данных, меньше которого ровно p% элементов всего набора данных.

**Код: вычисление на Python**

```
import pandas as pd
from scipy import stats

data = pd.Series([3, 1, 2, 5, 6, 0, 10]) # Анализируемый набор данных

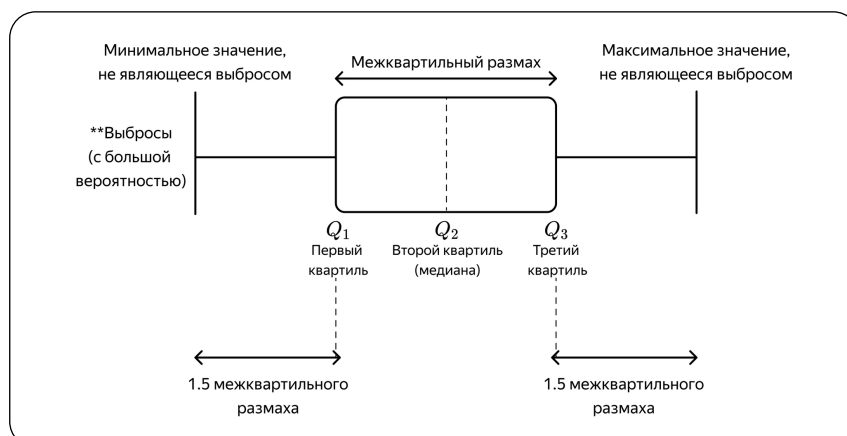
p = 0.1 # Уровень процентиля
print(data.quantile(p)) # Вывод значения
```

Межквартильных размах и процентиль являются **структурными** характеристиками положения, так как их вычисляют, исходя из структуры набора данных.

Дисперсия, стандартное отклонение и размах являются **алгебраическими** характеристиками положения, так как их вычисляют с помощью алгебраических операций над элементами набора данных.

## Диаграммы для изображения набора данных

Для отображения медианы и межквартильного размаха набора данных непрерывной величины используют **диаграмму размаха**. Ее описание приведено ниже на изображении:



## Описательная статистика

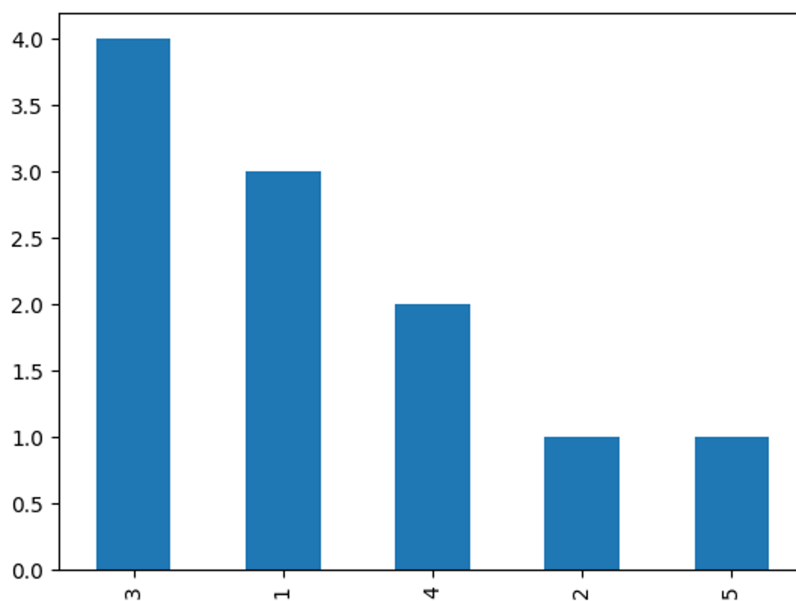
### Код: изображение в Python

```
import pandas as pd
import matplotlib.pyplot as plt

data = pd.Series([-10, 1, 2, 3, 4, 5, 6, 7, 8, 10]) # Анализируемый набор данных

ax = data.plot.box() # Построение графика
plt.show() # Изображение графика
```

Для отображения частоты каждого элемента в наборе данных дискретной величины используют столбчатую диаграмму. По оси X лежат значения из набора данных, а по оси Y количество повторений каждого элемента в наборе. Пример такого графика и способ его получить приведены ниже:



### Код: изображение в Python

```
import pandas as pd
import matplotlib.pyplot as plt

data = pd.Series([1, 1, 1, 2, 3, 3, 3, 3, 4, 4, 5]) # Анализируемый набор данных

ax = data.value_counts().plot.bar() # Построение графика
plt.show() # Изображение графика
```