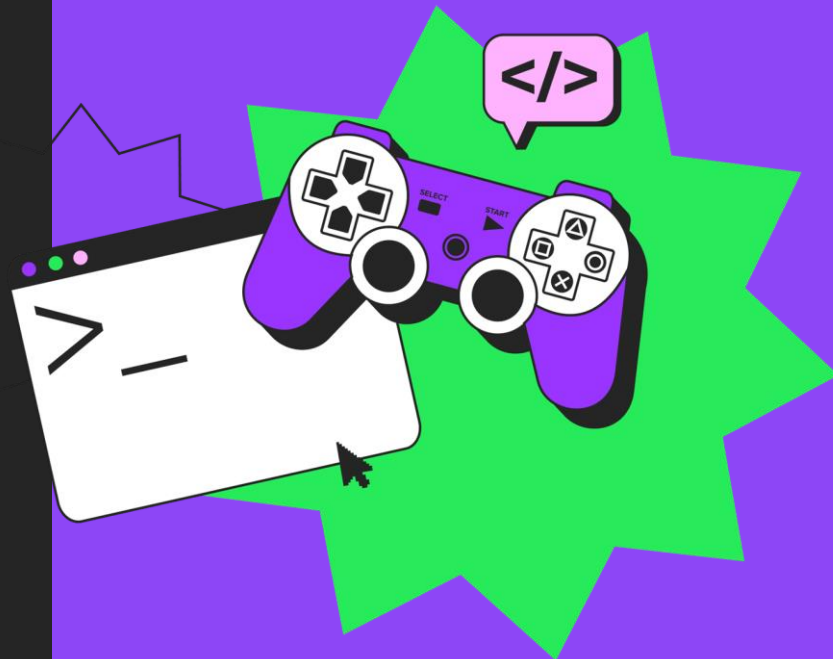


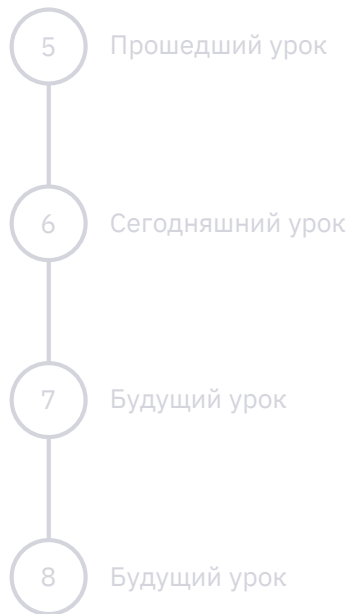
Разведочный анализ или EDA (exploratory data analysis)

Описательная статистика. Графический анализ.





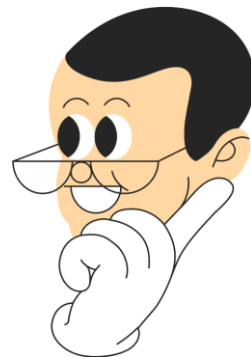
План курса





Что будет на уроке сегодня

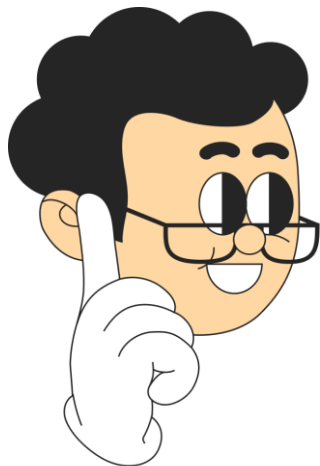
- Понятия генеральной совокупности и выборки
- Математическое ожидание
- Параметры описательной статистики, нечувствительные к выбросам
- Графический анализ: боксплот и гистограмма.





Генеральная совокупность

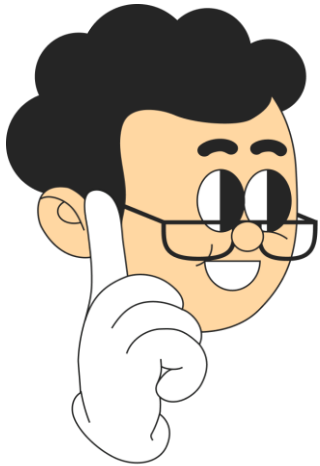
Генеральная совокупность- это множество, которое содержит данные обо всех объектах, соответствующих определенным характеристикам.



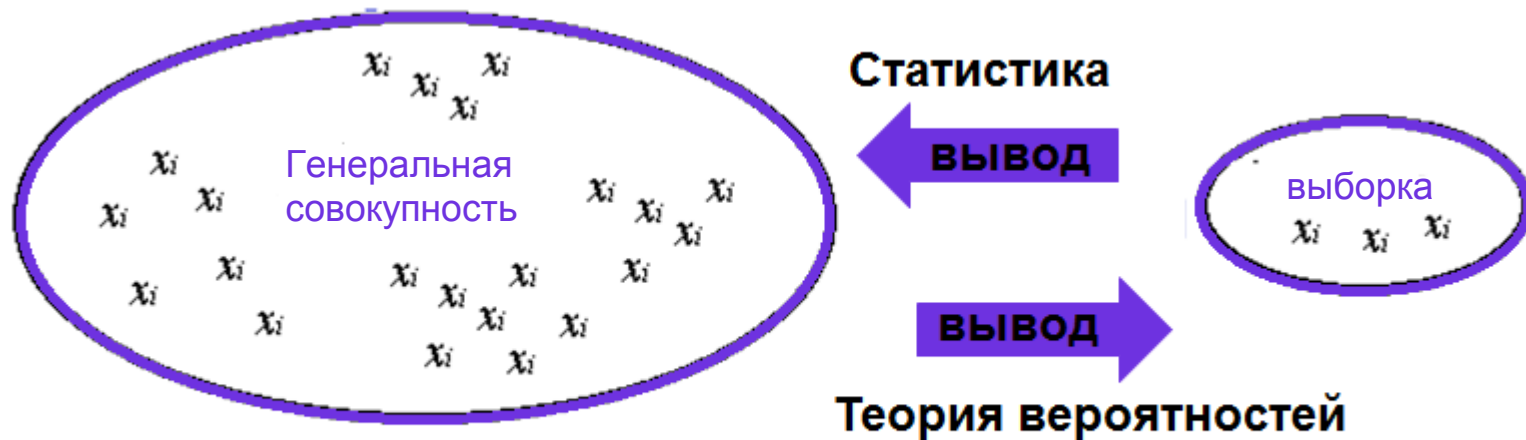


Выборка

Выборка - это случайным образом выбранная часть генеральной совокупности.



Статистика VS теория вероятностей



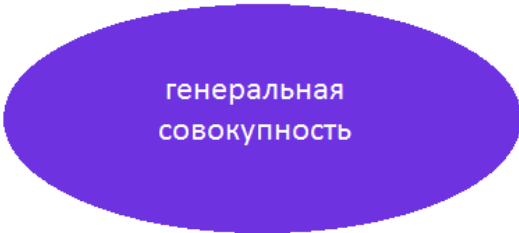
x_i - значения случайной величины



Математическое ожидание и его точечная оценка

Генеральная совокупность

Математическое ожидание — среднее значение случайной величины при стремлении количества выборок или количества измерений к бесконечности.




генеральная
совокупность

$$M(X) = \frac{1}{n} \sum_{i=1}^n x_i$$

Выборка

Оценка математического ожидания - это среднее арифметическое одномерной случайной величины конечного числа испытаний обычно называют оценкой математического ожидания.

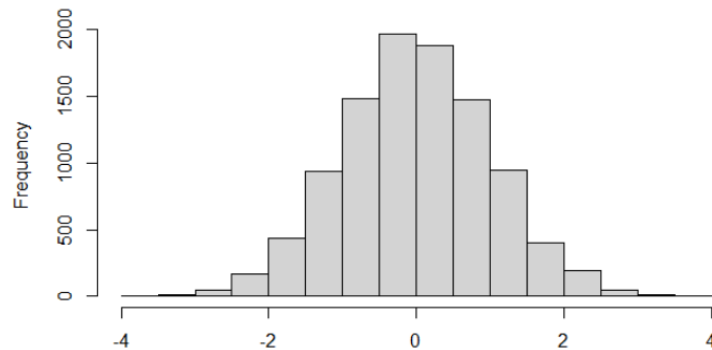
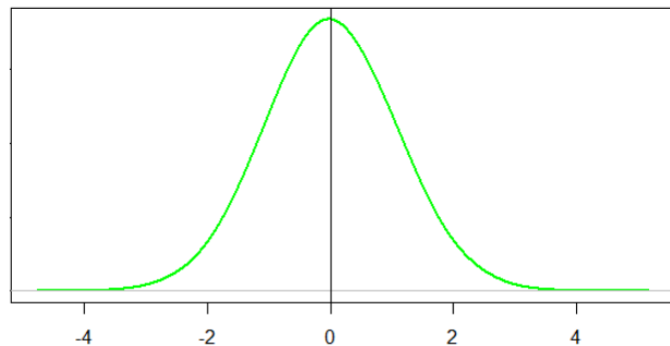


статистические
данные

$$\bar{X} = \frac{1}{m} \sum_{i=1}^m x_i$$

Задача математического ожидания

Основная задача математического ожидания - показать, вокруг какого значения группируется большая доля значений случайной величины.

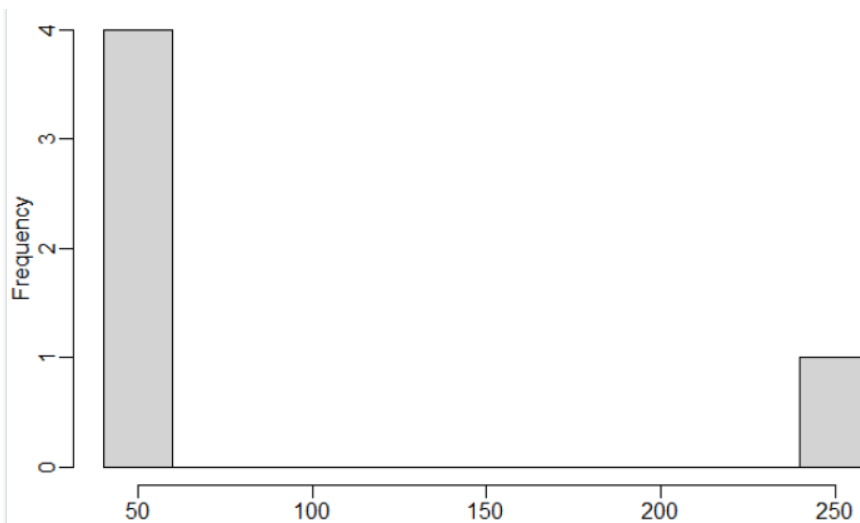


Недостаток математического ожидания

Математическое ожидание очень чувствительно к выбросам

Заработные платы: 50, 52, 51, 50, 257

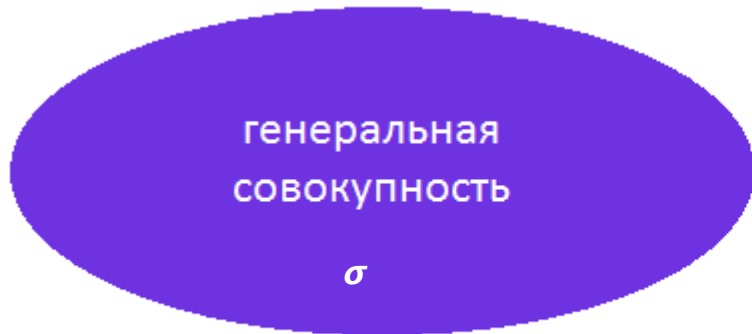
$$\bar{X} = 92$$





Среднее квадратичное отклонение

Среднее квадратичное отклонение показывает, насколько далеко наблюдения могут быть "разбросаны" относительно среднего значения.





Дисперсия

Генеральная совокупность

$$\sigma^2 = \frac{\sum_{i=1}^m (x_i - \mu)^2}{m}$$

генеральная
совокупность

Выборка

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

статистические
данные



Смещенная и несмещенная дисперсия по выборке

Смещенная дисперсия

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

Несмещенная дисперсия

Если объем выборки меньше 100 обязательно применение данной формулы:

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



Функции для описательной статистики

Смещенные стандартное отклонение и дисперсия

```
import numpy as np
x=np.array([167, 181, 174, 178, 175, 164, 182, 178,193, 166, 154, 170, 177])
x
array([167, 181, 174, 178, 175, 164, 182, 178, 193, 166, 154, 170, 177])
np.std(x)
9.382092977531892
np.var(x)
88.02366863905326
np.sqrt(88.02366863905326)
9.382092977531892
```

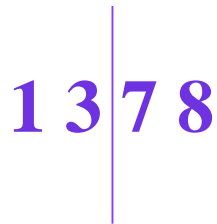
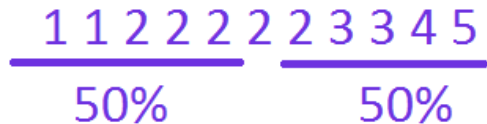
Несмещенные стандартное отклонение и дисперсия

```
import numpy as np
x=np.array([167, 181, 174, 178, 175, 164, 182, 178,193, 166, 154, 170, 177])
x
array([167, 181, 174, 178, 175, 164, 182, 178, 193, 166, 154, 170, 177])
np.std(x, ddof=1)
9.765191977579056
np.var(x, ddof=1)
95.35897435897436
np.sqrt(95.35897435897436)
9.765191977579056
```



Медиана

Медиана – значение, которое делит выборку на две равные части так, что значения, которые меньше медианы, составляют 50% выборки



$$\frac{N_{[\frac{n}{2}]} + N_{[\frac{n}{2}+1]}}{2}$$

Расчет медианы в Python



```
import numpy as np
z= np.array([100, 80, 75, 77, 89, 33, 45, 25, 65, 17, 30, 24, 57, 55, 70, 75, 65, 84, 90, 150])
z
array([100,  80,  75,  77,  89,  33,  45,  25,  65,  17,  30,  24,  57,
        55,  70,  75,  65,  84,  90, 150])
z.shape
(20,)
z.sort()
z
array([ 17,  24,  25,  30,  33,  45,  55,  57,  65,  65,  70,  75,  75,
        77,  80,  84,  89,  90, 100, 150])
(z[9]+z[10])/2
67.5
```




Мода

Мода - наиболее часто встречающееся в выборке значение.

		буквы									
частота		a	b	c	d	e	f	g	h	i	j
		1	2	2	2	5	3	3	2	2	1



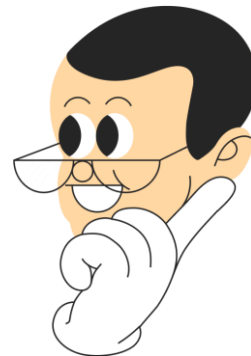
Параметры, нечувствительные к выбросам

Первый квартиль - такое значение, что 25% наблюдений в выборке не превышают эту величину.

Второй квартиль - синоним медианы.

Третий квартиль - такое значение, что 75% наблюдений в выборке не превышают эту величину.

Интерквартильное расстояние - отрезок, равный разности 3-го и 1-го квартиля





Расчет квартилей в Python



```
z= np.array([1, 2, 4, 2, 1, 5, 7, 2, 3, 5, 7, 8, 9])  
z  
array([1, 2, 4, 2, 1, 5, 7, 2, 3, 5, 7, 8, 9])
```

```
z.sort()  
z  
array([1, 1, 2, 2, 2, 3, 4, 5, 5, 7, 7, 8, 9])
```

Если $n \cdot k/100$ целое число, то k -я перцентиль – это среднее значение элементов под номерами $n \cdot k/100$ и $n \cdot k/100 + 1$

Если $n \cdot k/100$ не целое число, то k -я перцентиль совпадает с измерением $j+1$, где j – максимальное целое число, которое меньше, чем $n \cdot k/100$

```
n= len(z)  
n  
13
```

```
k=25  
n*k/100
```

```
3.25  
3+1  
4
```

```
z[3]  
2
```



Вопрос

Посчитать 3 квартиль





Вопрос

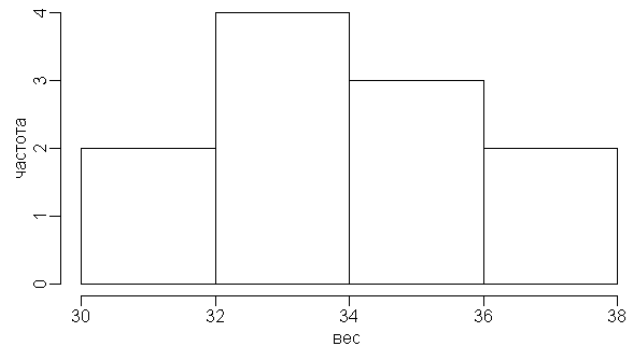
Посчитать межквартильное расстояние





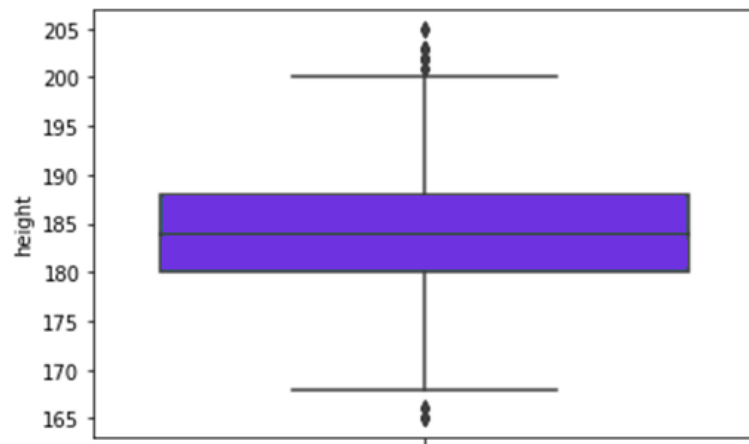
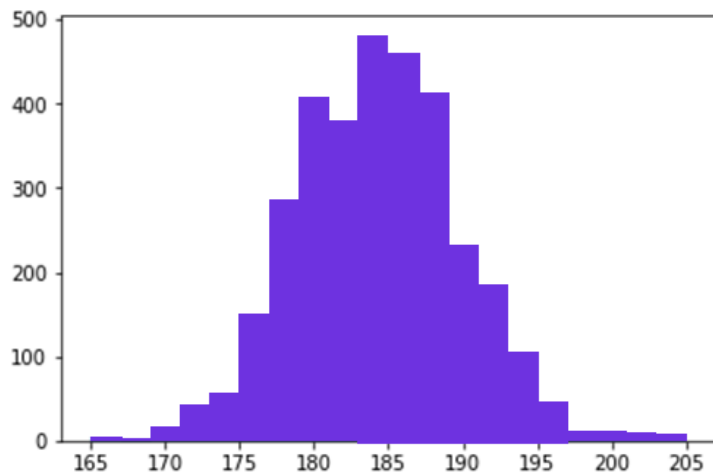
Размах

32.34566,
34.96313,
33.87,
35.61900,
35.60872,
33.11,
32.78,
30.71787,
30.45296 ,
36.41410,
37.86643



$$R = X_{\max} - X_{\min}$$

Графическое представление данных

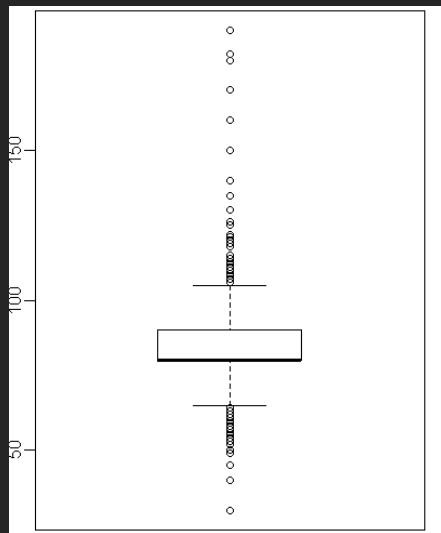


$$X_1 = Q1 - 1.5 \cdot (Q3 - Q1) ;$$
$$X_2 = Q3 + 1.5 \cdot (Q3 - Q1)$$



Вопрос

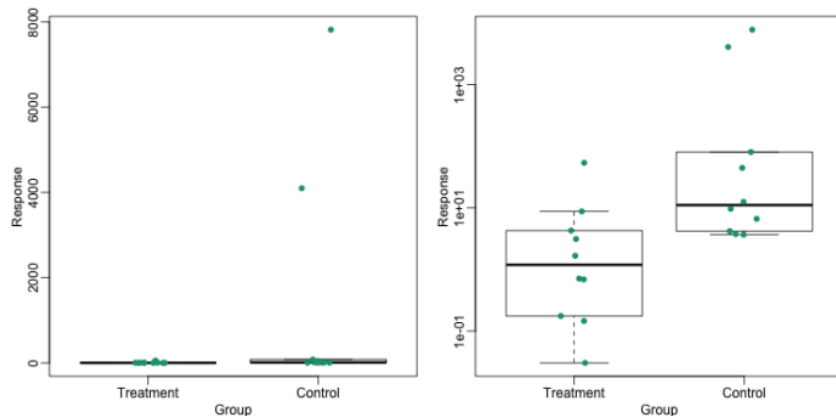
Интерпретировать график



Правила визуализации данных

1. Располагать значения в определенном порядке
2. Избегать круговых диаграмм
3. Не использовать псевдотрехмерную графику
4. Стараться максимально просто изображать данные
5. Использовать одинаковые единицы измерения
6. Не оставлять много знаков после запятой
7. Добавлять легенду на графики
8. При необходимости прибегать к масштабированию данных

для графического анализа





Вопрос 1

Какую функцию использовать для расчета стандартного отклонения при небольших выборках?

- 1) `std (x)`
- 2) `var(x)`
- 3) `std(x, ddof=1)`
- 4) `var(x, ddof=1)`





Вопрос 2

Что показывает 1 квартиль?





Вопрос 3

Выбрать параметры, чувствительные к выбросам.

- 1) медиана
- 2) математическое ожидание
- 3) размах
- 4) 25 перцентиль
- 5) мода





Конец