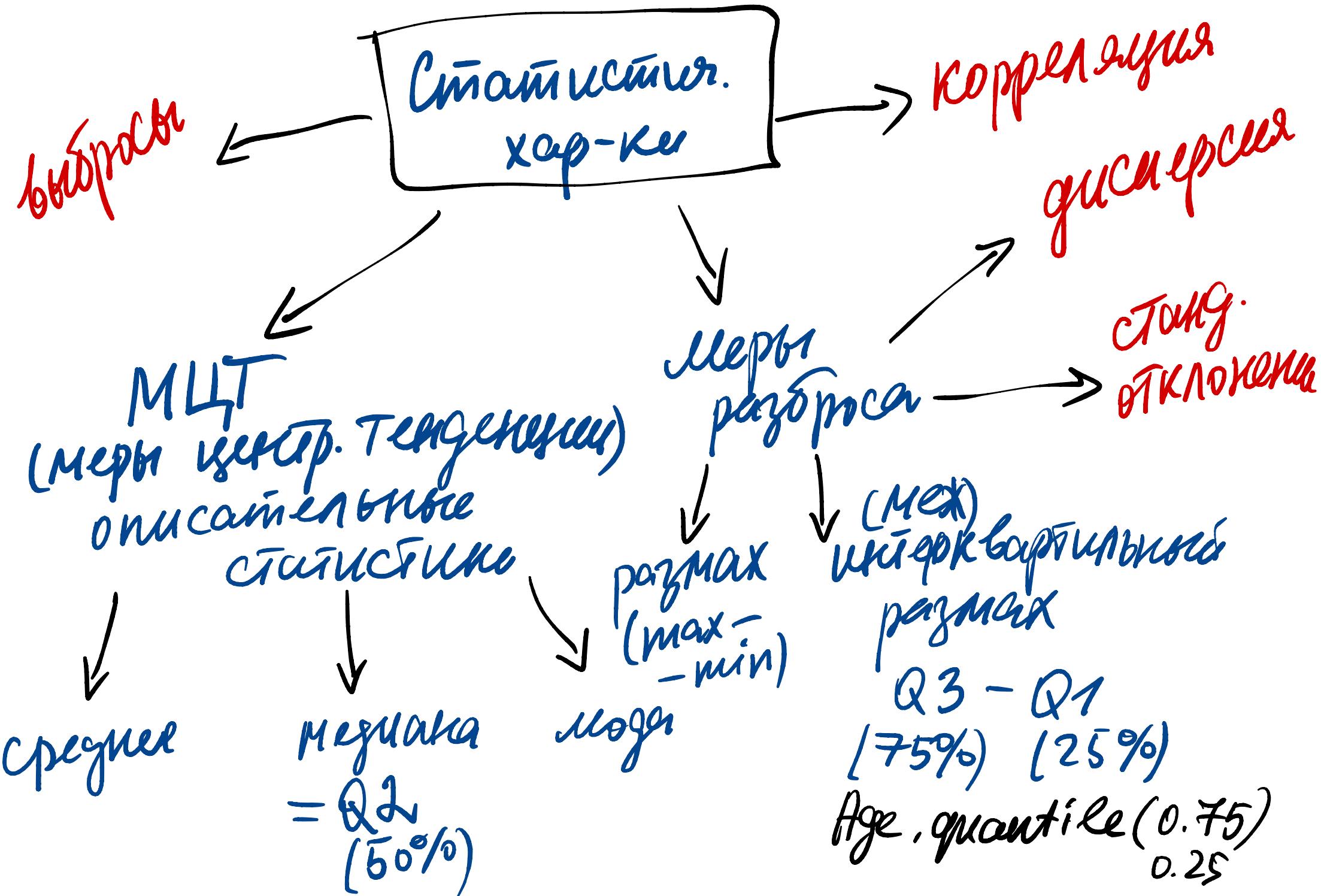
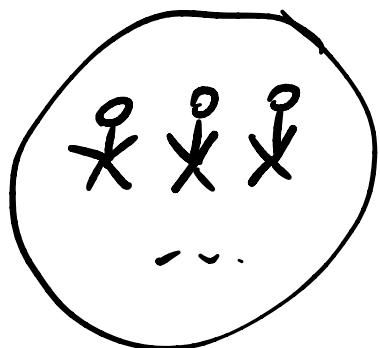


## План

- вспомогательн. осн. стат. характеристики
- дисперсия, стат. оценки
- коррелируются
- выбросы
- Pandas
- Тест на лекции  $12 \frac{18}{-} - 12 \frac{3}{0}$





8 учеников

2, 4, 4, 4, 5, 5, 7, 9

1) среднее  $(2+4+4+\dots+9)/8 = \underline{\underline{5}}$

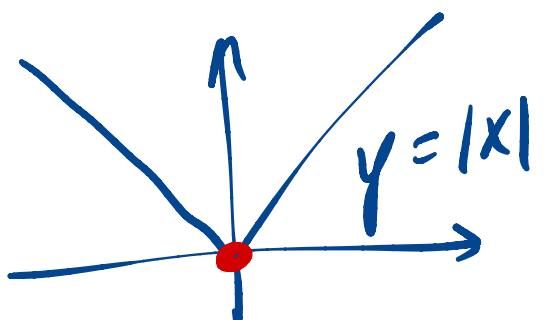
2) Насколько отклоняется оценка  
коиного ученика от среднего? (Берем квадрат отклонений)

$$(2-5)^2$$

$$(4-5)^2$$

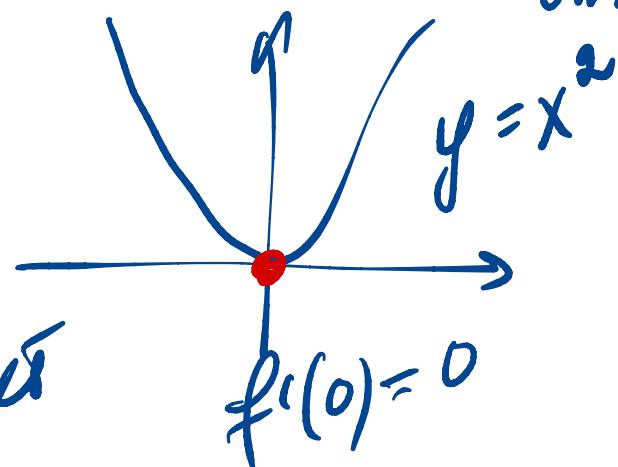
$$\vdots$$

$$(9-5)^2$$



$f'(0)$  - не существует

MAE  
(mean absolute error)



MSE  
(mean squared error)

$$3) \sigma^2 = \frac{9+1+1+1+0+0+16+4}{8}$$

$$\boxed{\sigma^2 = 4}$$

дисперсия  
↑  
"сума" (греч. буква)

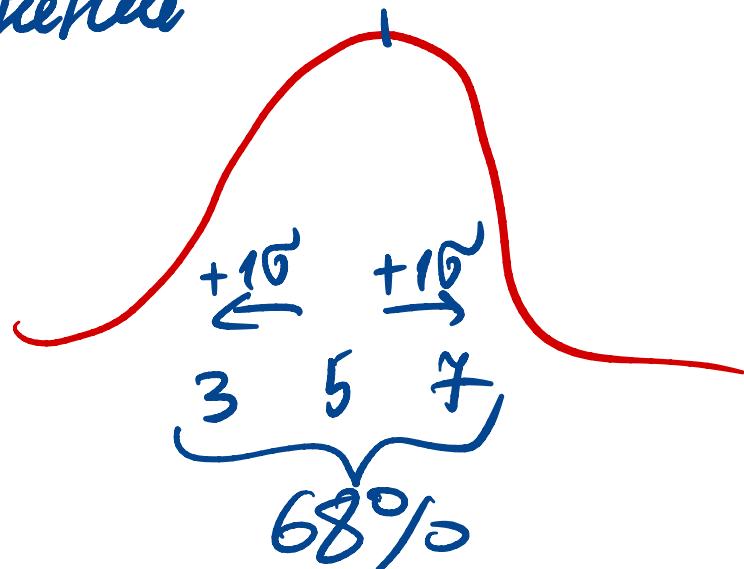
просуммируем  
квадраты откл-ий  
из 2 шага

насколько разбросаны  
даные относительно среднего

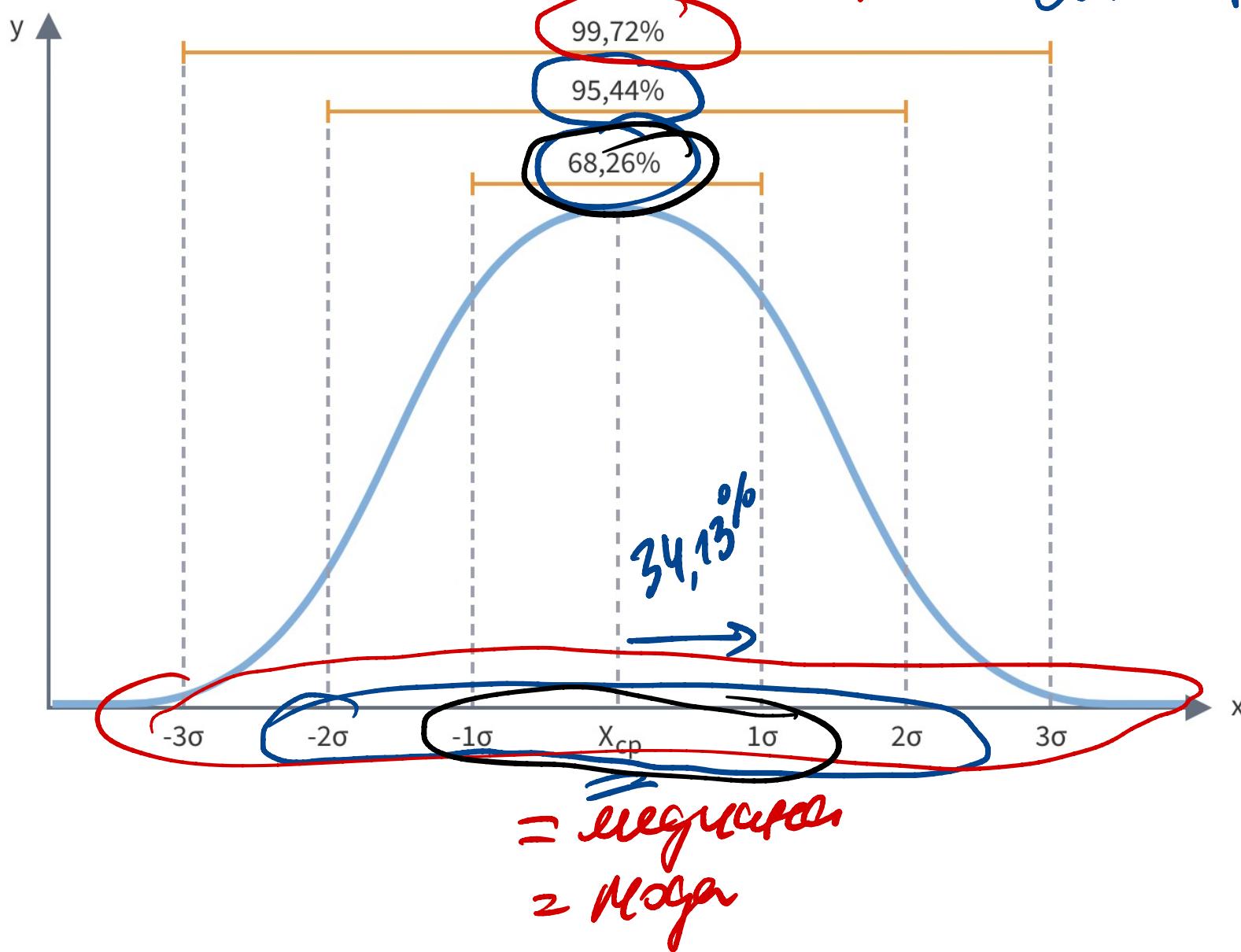
$$\sqrt{\sigma^2} = \sigma \quad +1\sigma, +2\sigma, +3\sigma$$

$\sigma = 2$

сигмаартикал (среднеквадратич.)  
отклонение



норм. расп-ие  
симметричное

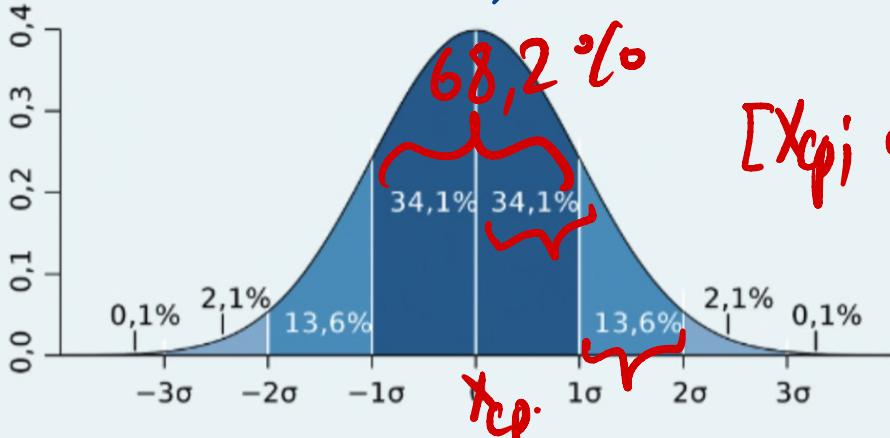


# нормальный / гауссовский

## Какой вывод ( $= 1$ )

Предположим, что распределение веса всех взрослых женщин на планете нормальное. Если среднее — 62 кг, а среднеквадратичное отклонение — 8 кг, то какие выводы можно сделать?

$$(\geq 1)$$



~~✓~~ 47,7 % взрослых женщин весит от 62 до 86 кг.

~~✓~~ 13,6 % взрослых женщин весит от 46 до 62 кг

~~✓~~ 0,1 % взрослых женщин весит больше 86 кг.

~~✗~~ Вес 68,2 % взрослых женщин находится между 58 и 68 кг.

$$x_{ср}$$

$$\sigma$$

$$[54; 70]$$

$$[62; 78]$$

||

$$\begin{aligned} &x_{ср} + 2\sigma \\ &62 + 2 \cdot 8 \\ &62 + 16 \\ &78 \end{aligned}$$

$$\begin{aligned} &x_{ср} + 3\sigma \\ &62 + 3 \cdot 8 \\ &62 + 24 \\ &86 \end{aligned}$$

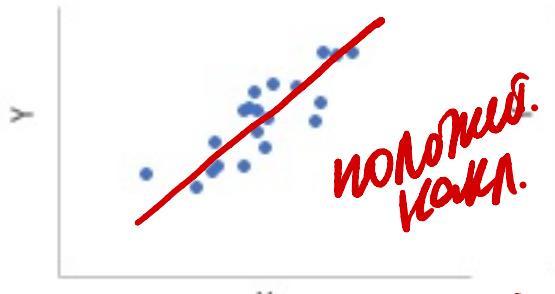
# репрез-сп выборки (отсутствует смещение выборки)

Мы хотим узнать, какой средний уровень образования у совершеннолетних женщин в России. Из вариантов ниже выберите тот, в котором не происходит гарантированного смещения выборки:

- Опрос был проведен посредством социальных сетей для их пользователей.
- Были опрошены только работающие женщины.
- Было опрошено такое количество женщин из каждого региона, которое соответствует доле женщин, проживающих в этом регионе согласно переписи населения.
- Были опрошены жительницы всех городов с населением более 300 000 человек.

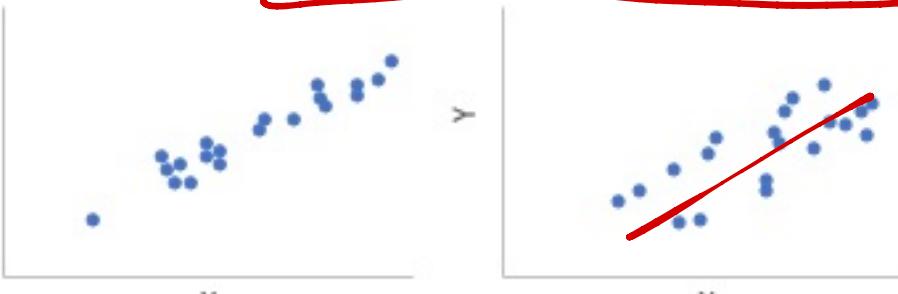
# Корреляция (~ связь)

причино-следственная



Прямая  $r > 0$

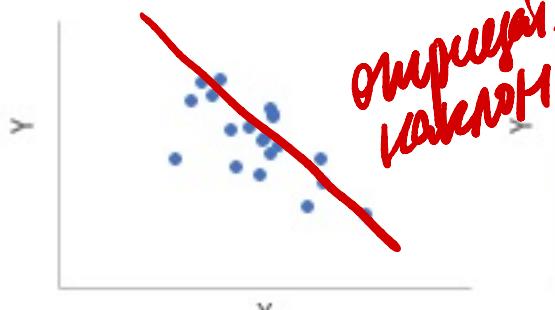
полож. связ.



Сильная

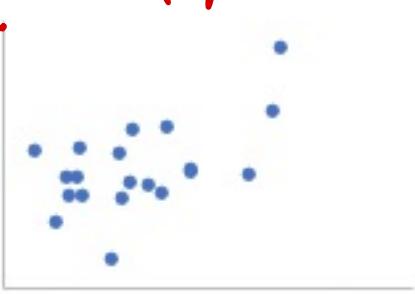
$r \rightarrow 1$

$0.7 < |r| < 1$



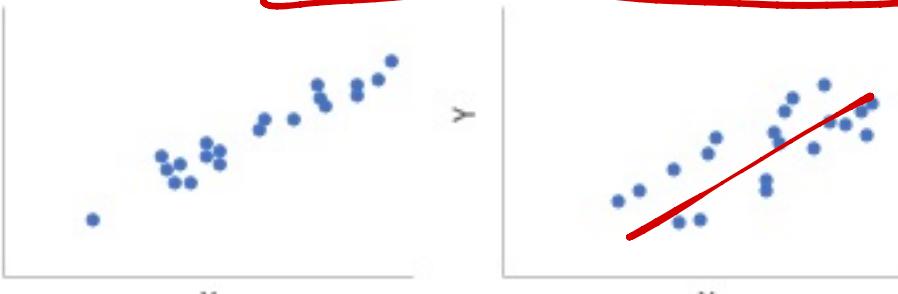
Обратная  $r < 0$

негативные  
корреляционные  
связи



Слабая

$0 < |r| < 0.3$

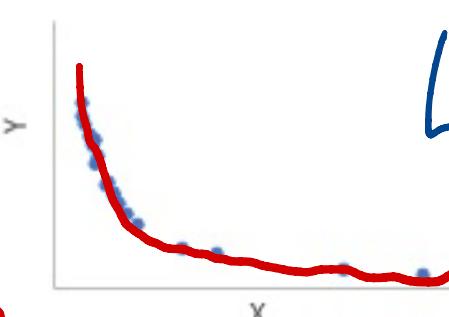


Линейная

Коэффиц. корр-ции  
с corr

$$-1 \leq r \leq 1$$

где

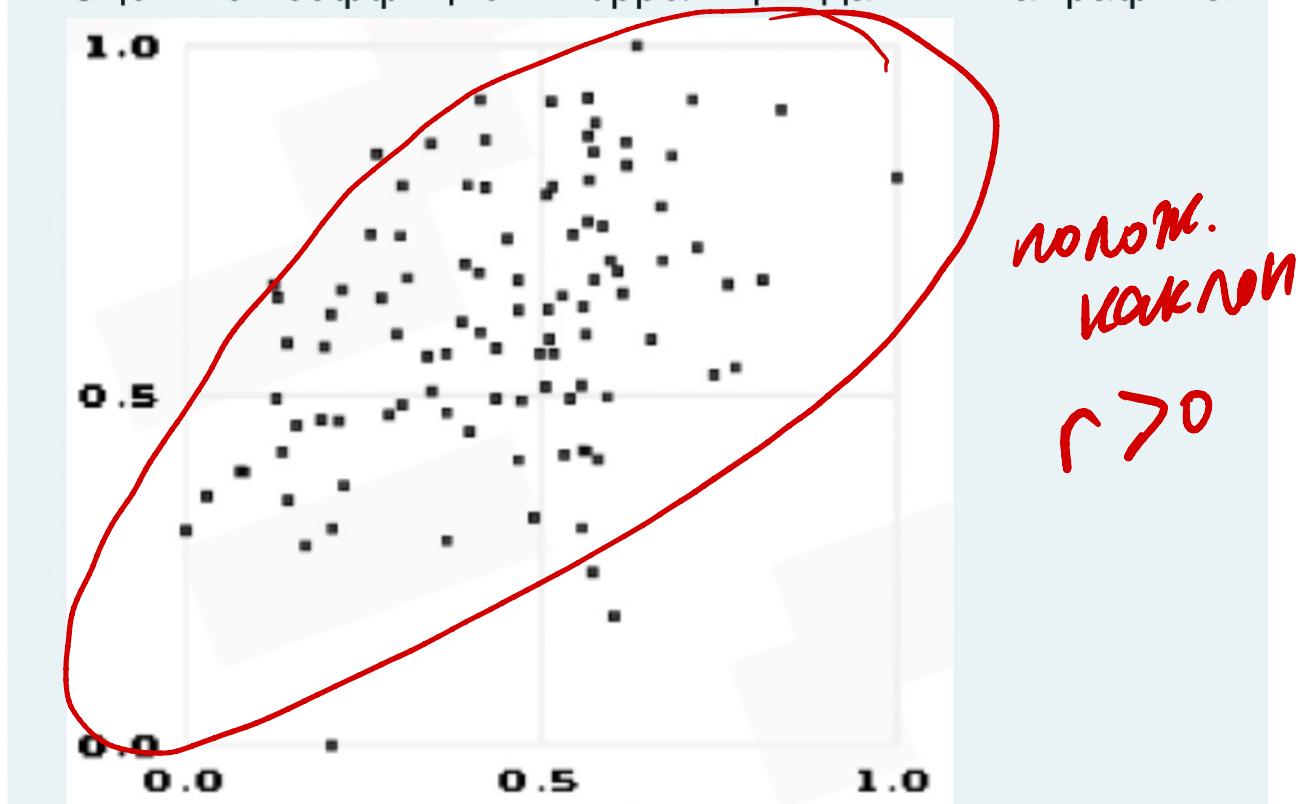


Нелинейная

форма  
связи

0 - отсутствие  
корр-ции

Оцените коэффициент корреляции данных на графике.



X -0.21

X 0.82

✓ 0.41

X -0.54

$0 < |r| < 0.3$  слабая связь

$0.3 < |r| < 0.7$  средняя связь

$0.7 < |r| < 1$  сильная связь

Мы провели исследование и выявили, что у сотрудников компании уровень удовлетворенности трудом коррелирует с их продуктивностью на работе, коэффициент равен 0.81. Какие выводы можно точно сделать из этого наблюдения?

Между уровнем удовлетворенности работой и продуктивностью слабая прямая взаимосвязь.

Если у сотрудника повышается продуктивность, то, скорее всего, повысится и удовлетворенность работой.

Продуктивность на работе влияет на удовлетворенность собственным трудом.

Удовлетворенность собственным трудом напрямую влияет на продуктивность.

*Инженерно-следеб. ке гарантиску*

Кот Матроскин ежедневно в течение недели ходил на рыбалку, чтобы поймать больше рыбок для своего любимого аквариума. Оказалось, что пойманные им рыбки обладают следующими характеристиками:

- Цвет: [серая, серая, золотая, синяя, золотая]
- Длина (см): [15, 12.1, 20, 24.2, 18.7] ← list1
- Масса (г): [800, 495, 302, 1001, 256] ← list2

*import scipy  
scipy.stats.pearsonr(list1,  
list2)[0]*

Выберите две характеристики, между которыми можно корректно рассчитать выборочный коэффициент корреляции Пирсона, и вычислите этот коэффициент по приведённым данным. Ответ округлите до сотых.

Пример ответа: 0.97

Ответ:

# Выбросы Outliers

это наблюдение в анализируемых данных, значение которого сильно отличается от других.

Как опознать выброс? Эвристики.  
Например, расстояние три стандартных  
отклонения от среднего.

С точки зрения исследователя:  
принять решение, как их  
“нейтрализовать”.

С точки зрения журналиста:  
что-то, что нужно расследовать!

# Сила выбросов

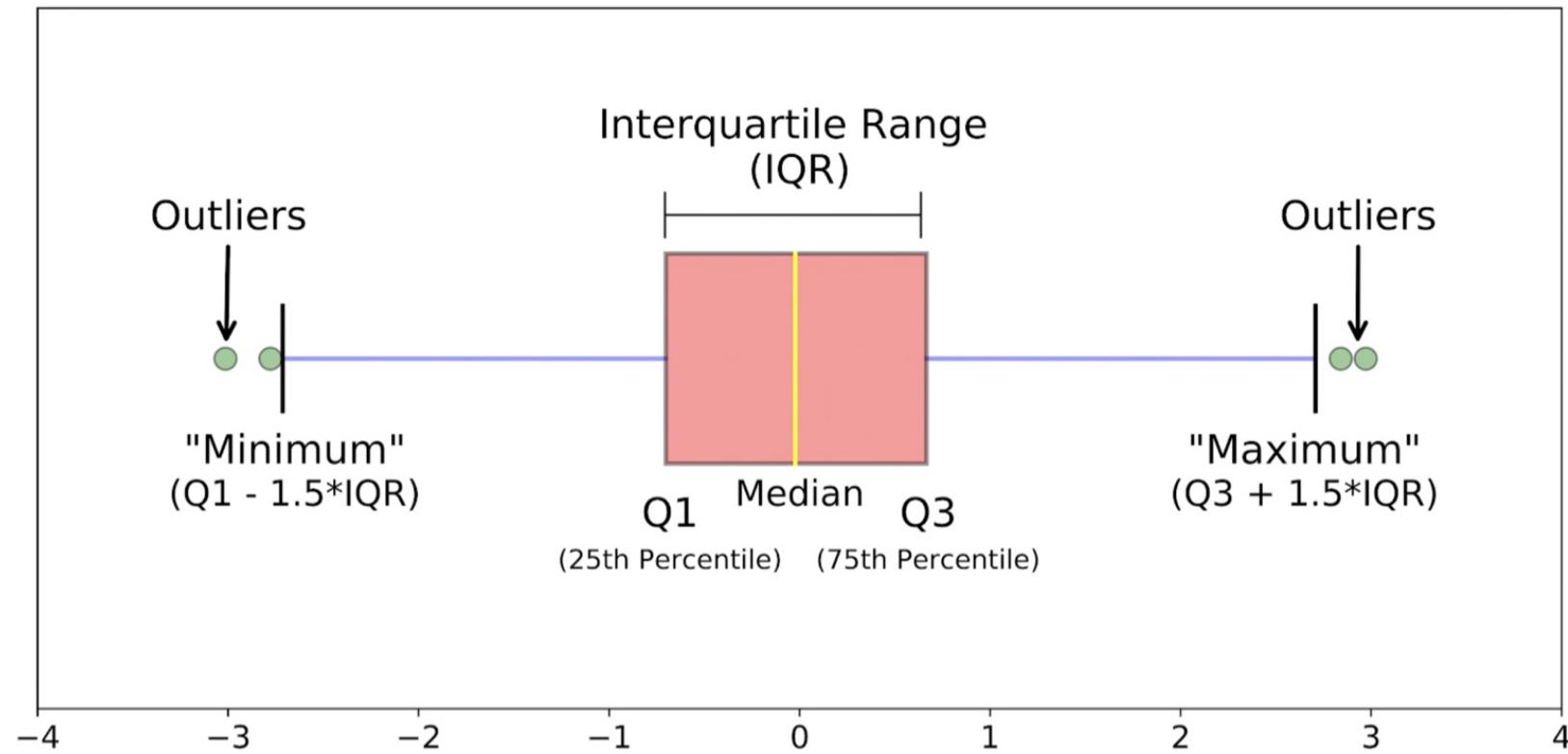
## Слабые выбросы

- Те значения, которые меньше 25% перцентили минус 1,5\*ИКР (интерквартильный размах размах) или больше 75% персентили плюс 1,5\*ИКР.
- Или находятся от среднего на расстоянии от двух до трех стандартных отклонений.

## Сильные выбросы

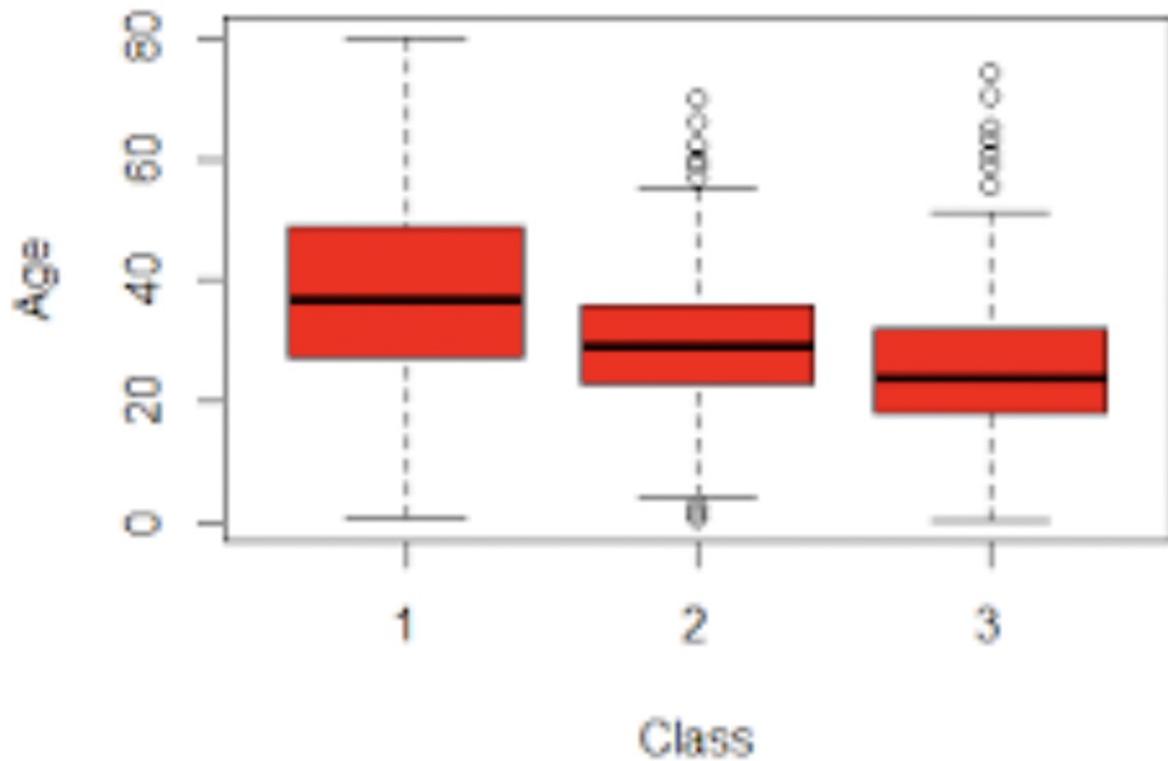
- Те значения, которые меньше 25% перцентили минус 3\*ИКР (межквартильный размах) или больше 75% персентили плюс 3\*ИКР.
- Или дальше от среднего чем три стандартных отклонения.

# Выбросы



В ящике-с-усами, усы как раз ограничивают выборку от выбросов.

# Выбросы



Найдите выбросы  
на графике.



