



# **Тема 4. Выбросы, пропуски, корреляции**

`df["столбец"].apply(функция)`

`df.groupby('категория')['столбец'].mean()`

**количес-  
венные** **категори-  
альные**

**Меры  
центральной  
тенденции**

мода

мода,  
медиана,  
среднее

**Меры  
вариативности**

количество  
уникальных  
категорий

стандартное  
отклонение,  
дисперсия,  
квартили

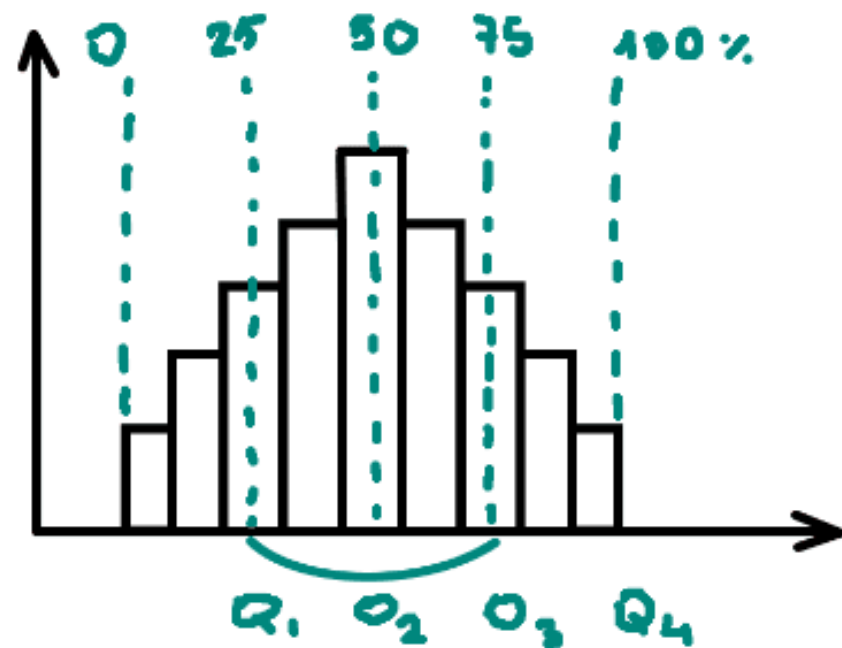
**от гистограмм  
к ящикам с  
усами**



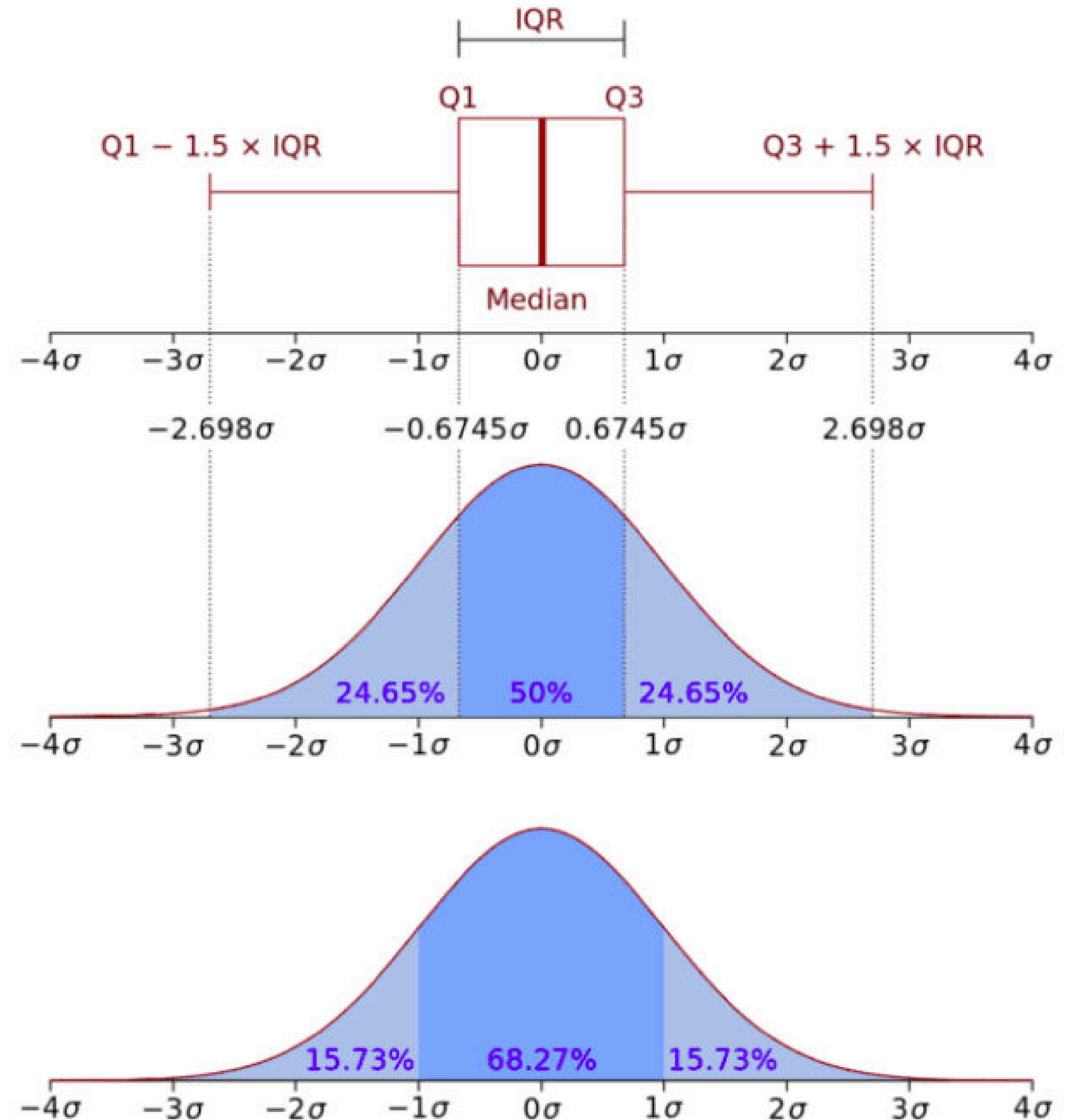
# Термины:

- квартиль
- межквартильный размах (интервал)

*не путаем с просто размахом (макс - мин)*

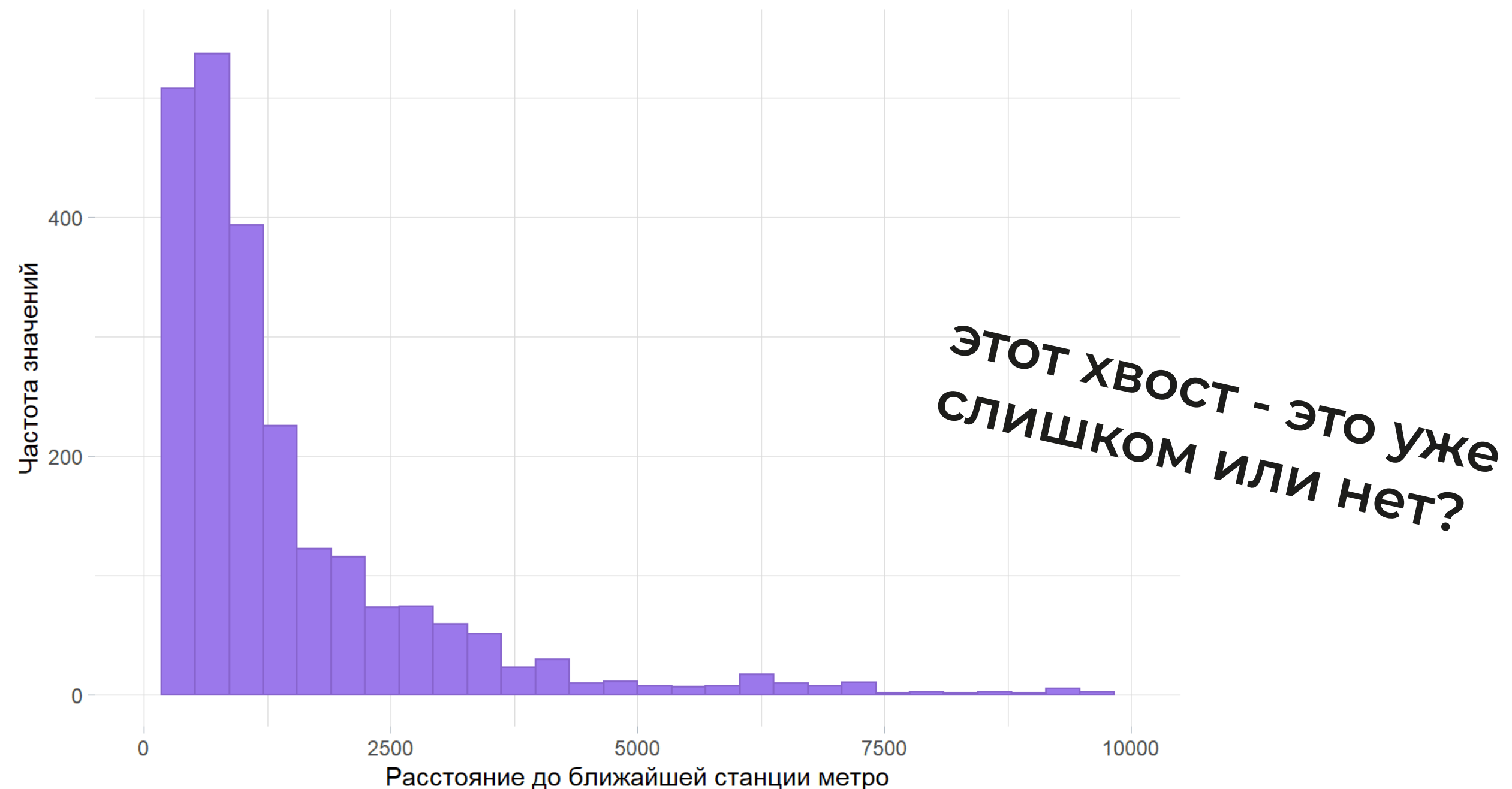


**зачем??**

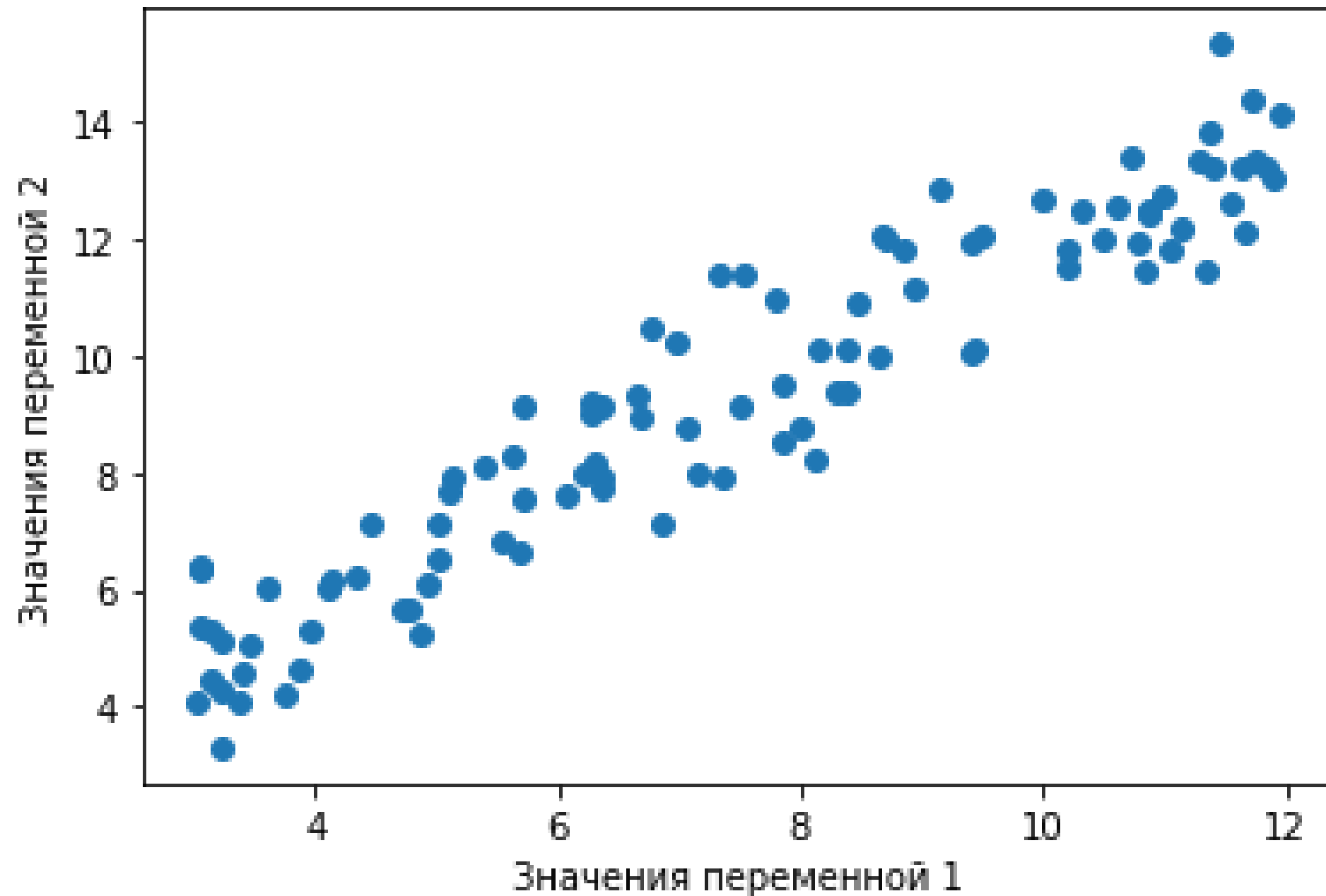


# Термины:

- выброс - отличается от распределения, выделяется (**слишком** маленькое / большое значение)

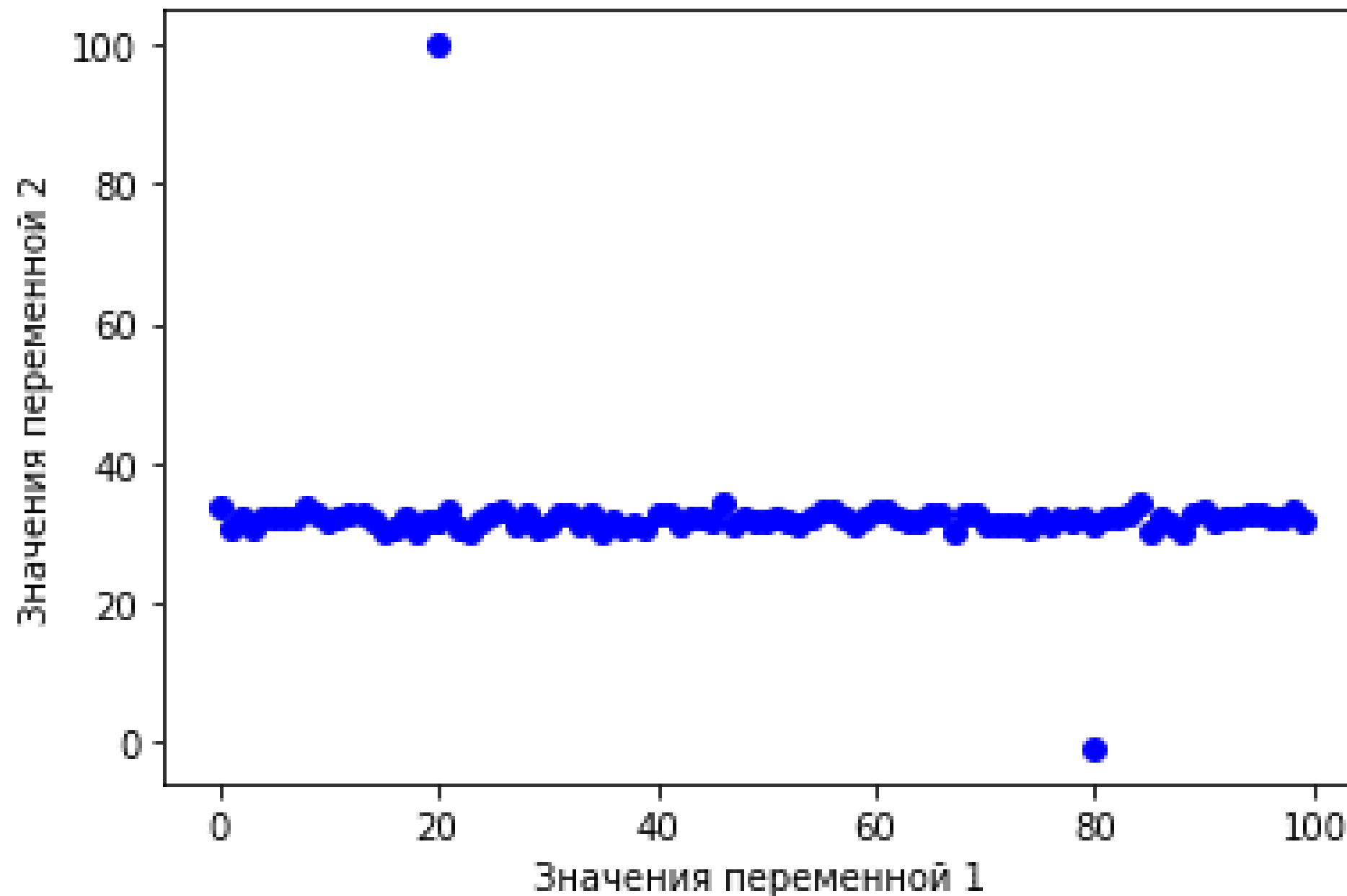


# Задача из НЭ, А6 (более новые)



1. Выбросы оказывают сильное влияние на среднее значение переменной 2
2. *По данной выборке можно построить линию регрессии*
3. В данных, скорее всего, нет выбросов
4. Выбросы оказывают сильное влияние на среднее значение переменной 1

# Задача из НЭ, А6 (более новые)



1. Выбросы оказывают малое влияние на среднее значение переменной 1
2. *По этой выборке нельзя построить линию регрессии*
3. В выборке имеется как минимум 1 выброс
4. Выбросы оказывают большое влияние на медиану переменной 2



# Все на питоне:

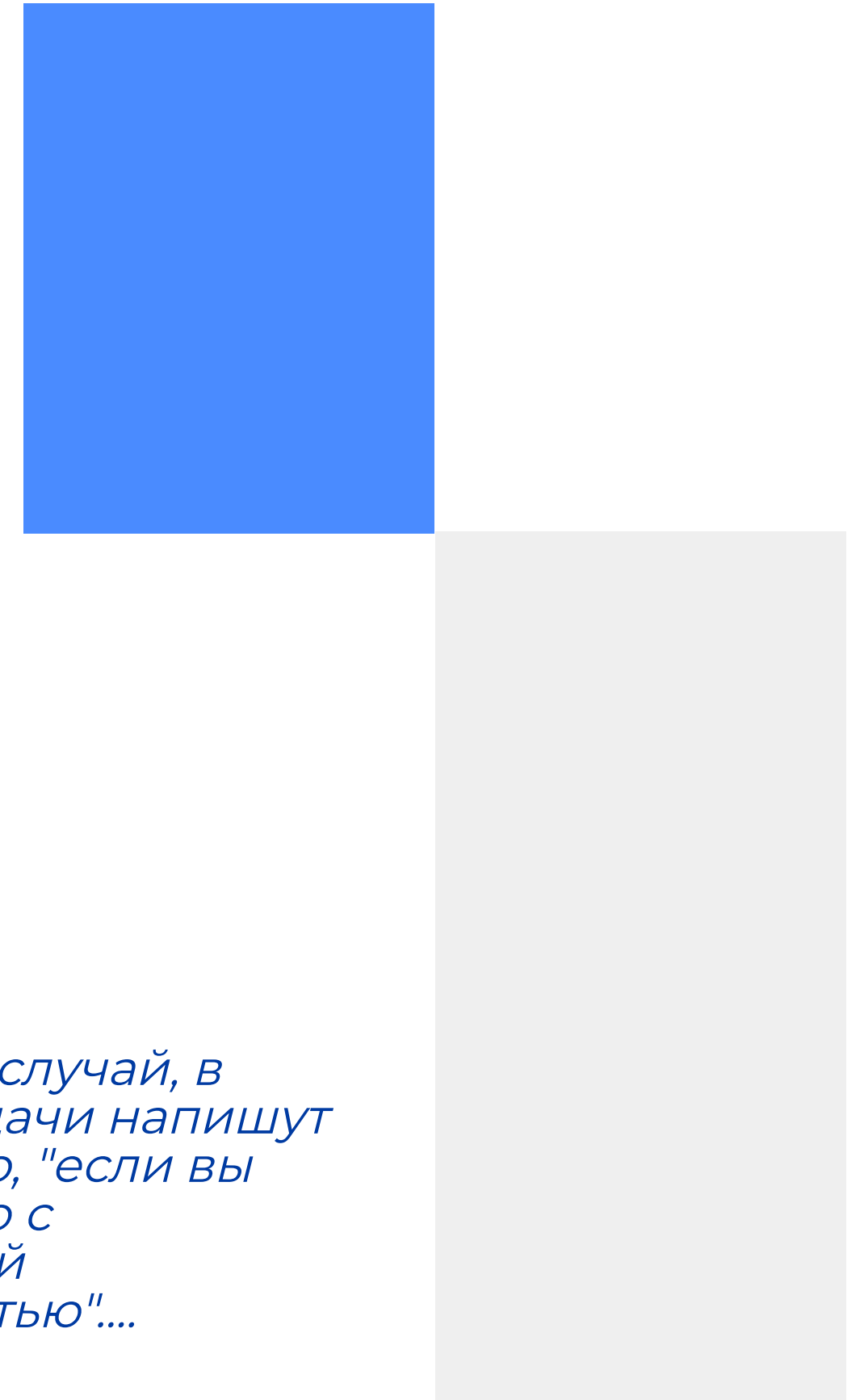
**df['столбец']**

**.min()  
.max()  
.mean()  
.mode()  
.median()  
.std()  
.var()**

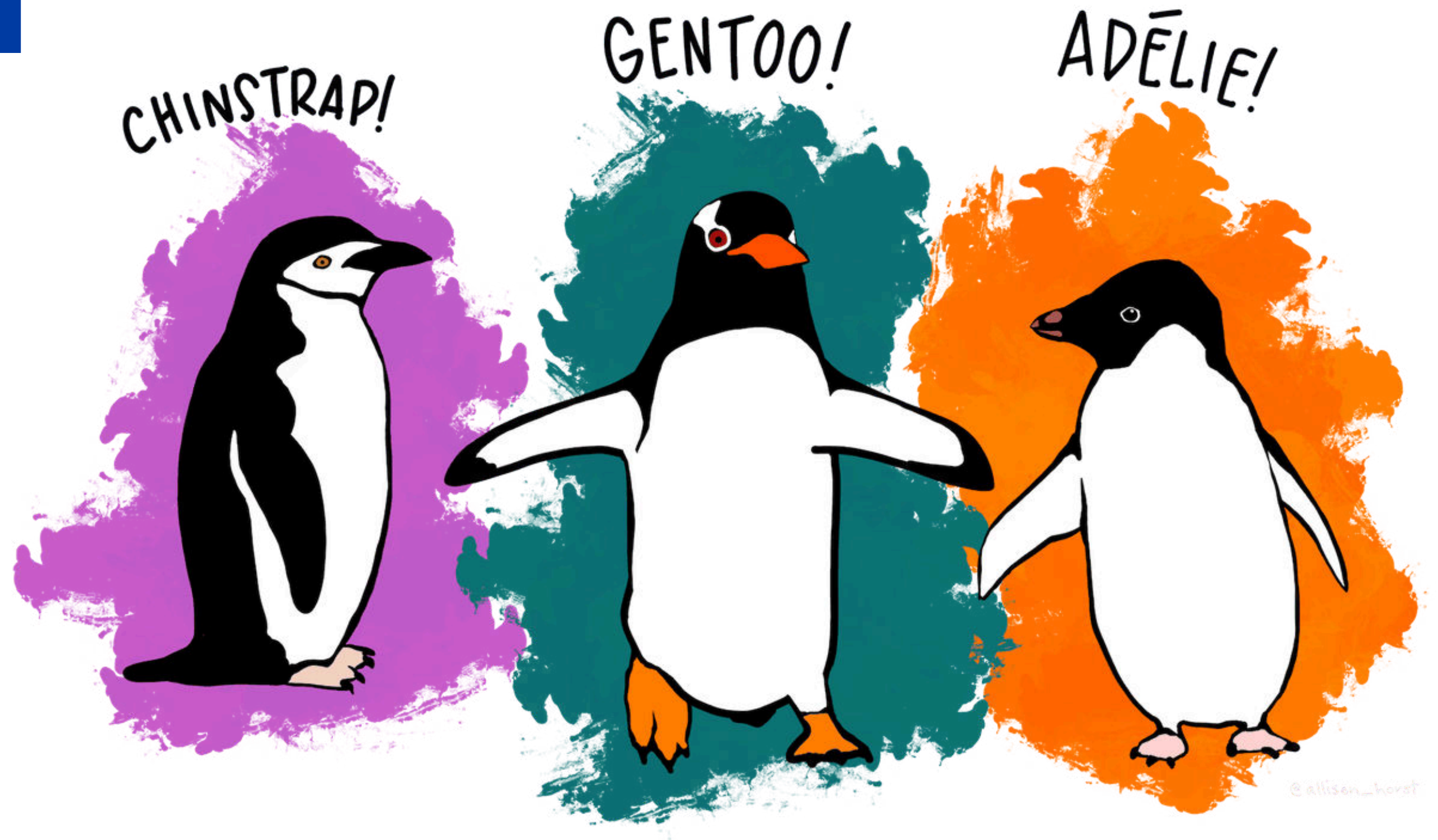
**Только для  
генеральной  
совокупности!**

**.std(ddof=0)  
.var(ddof=0)**

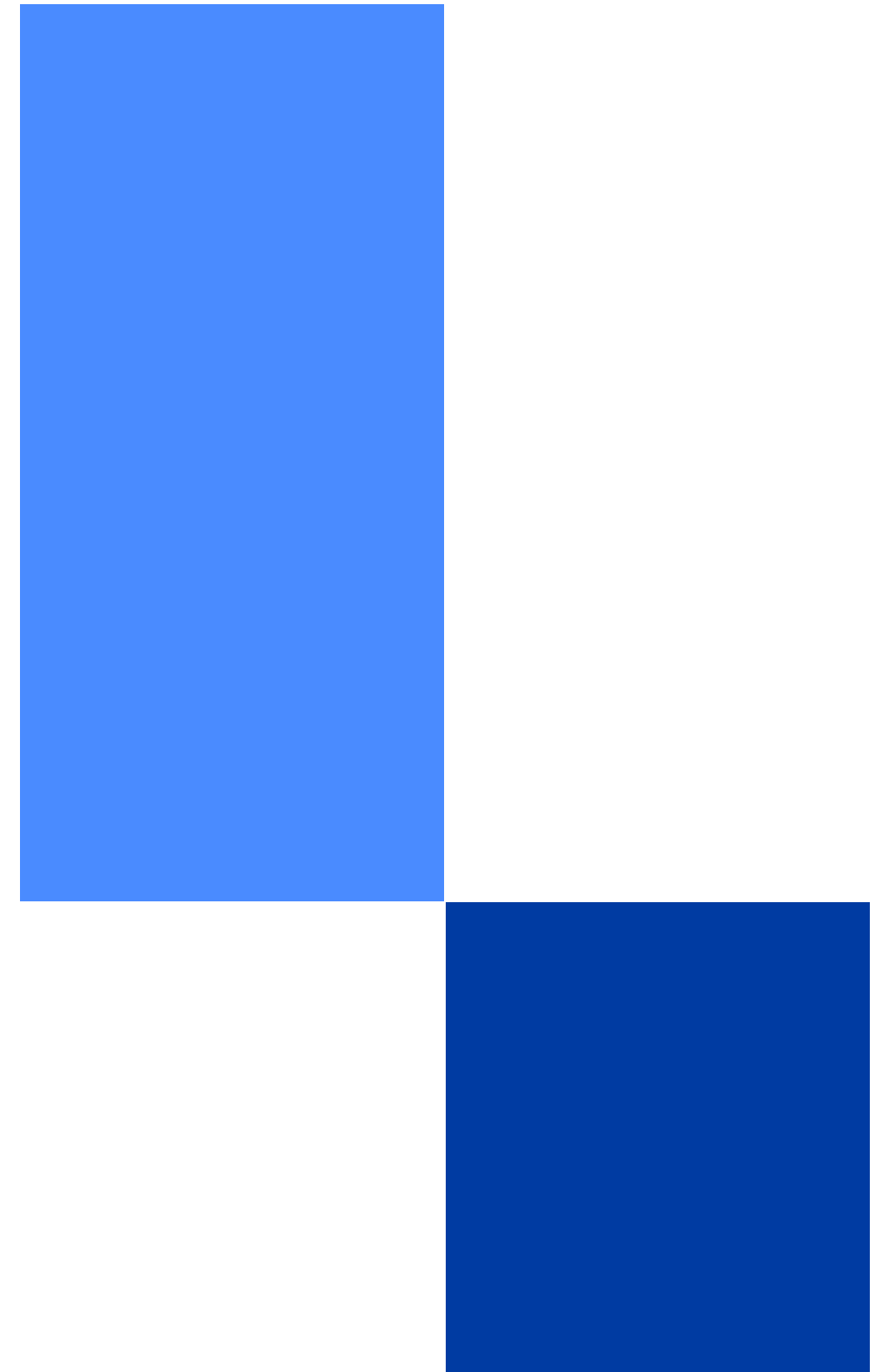
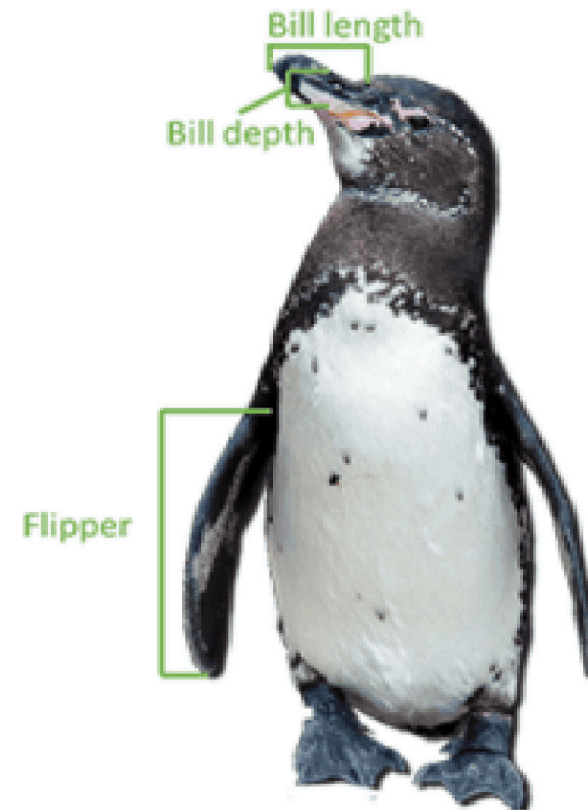
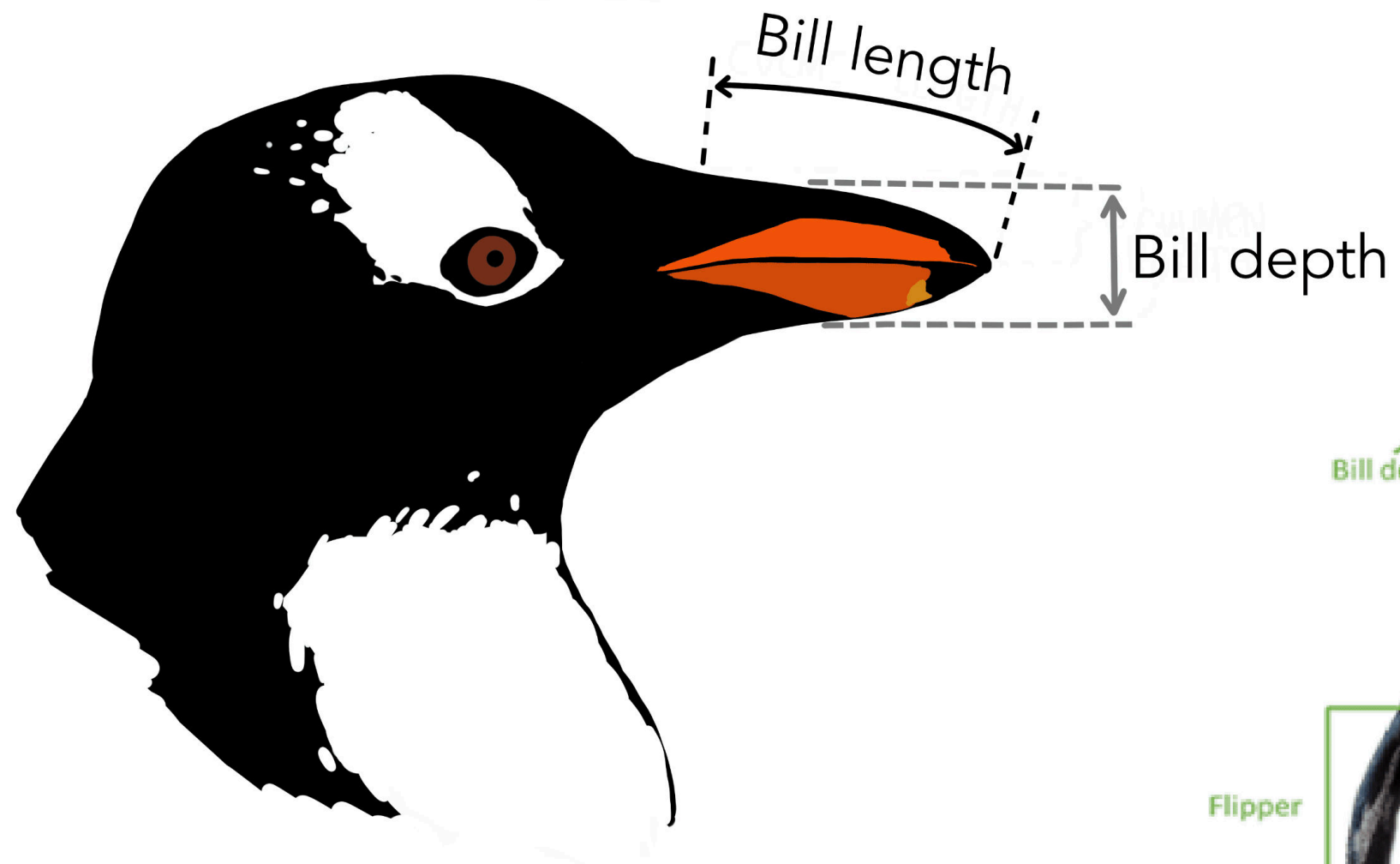
*Это редкий случай, в  
условии задачи напишут  
обязательно, "если вы  
имеете дело с  
генеральной  
совокупностью"....*



# "игрушечный" датасет



# "игрушечный" датасет

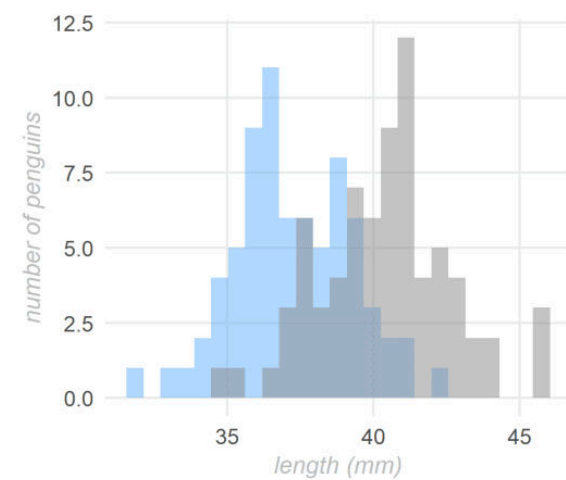
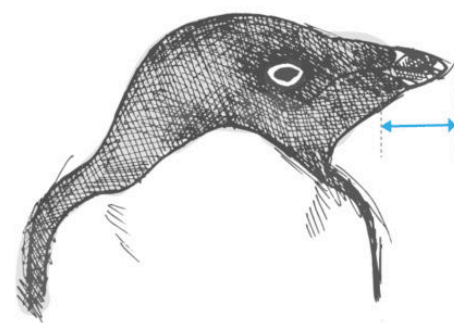


# "игрушечный" датасет

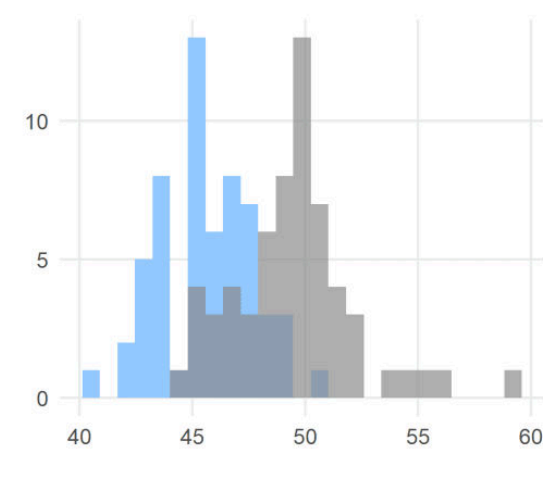
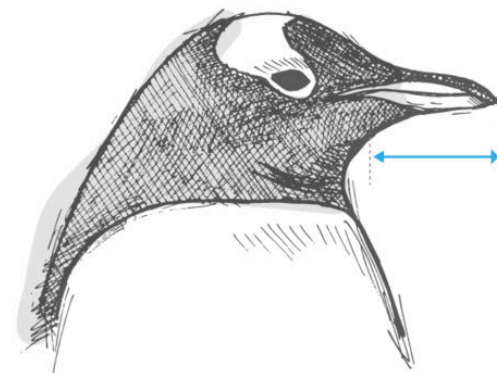
## Palmer Penguins Bill Length

Palmer Archipelago is a group of islands off the northwestern coast of the Antarctic Peninsula. The histograms show that females has shorter bills than males in every species

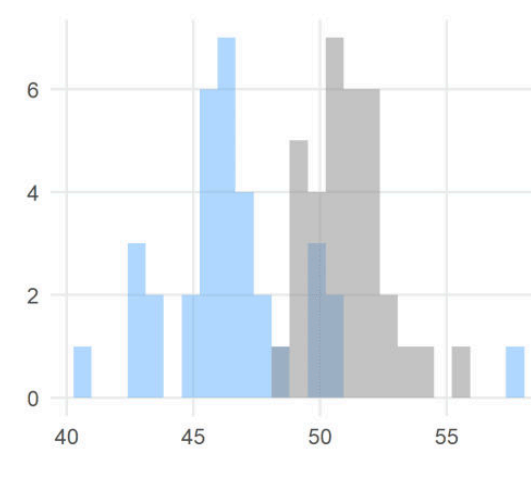
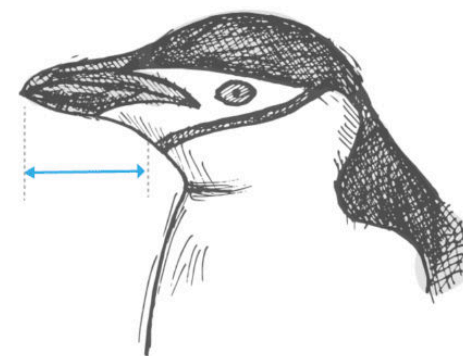
ADELIE



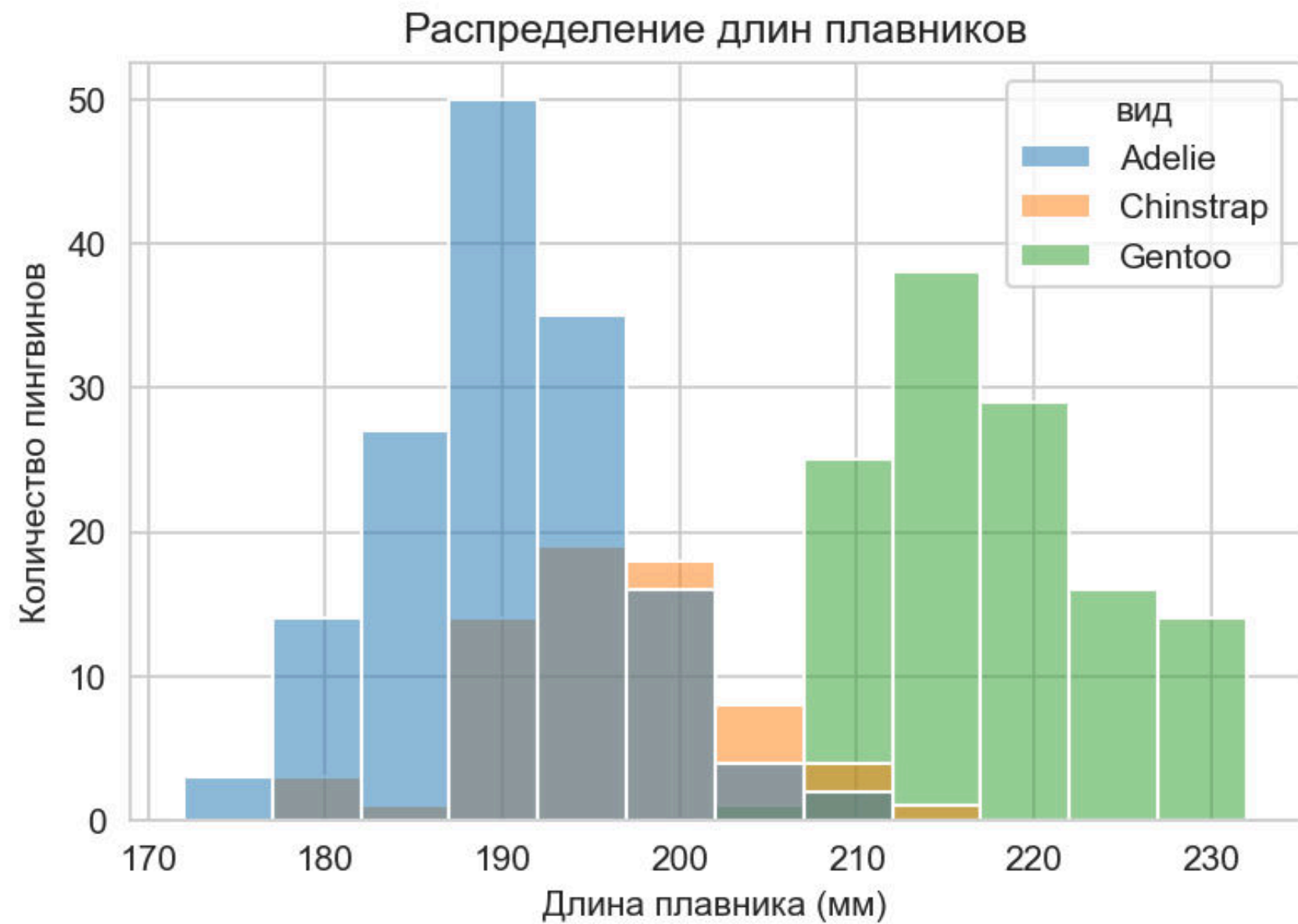
GENTOO



CHINSTRAP

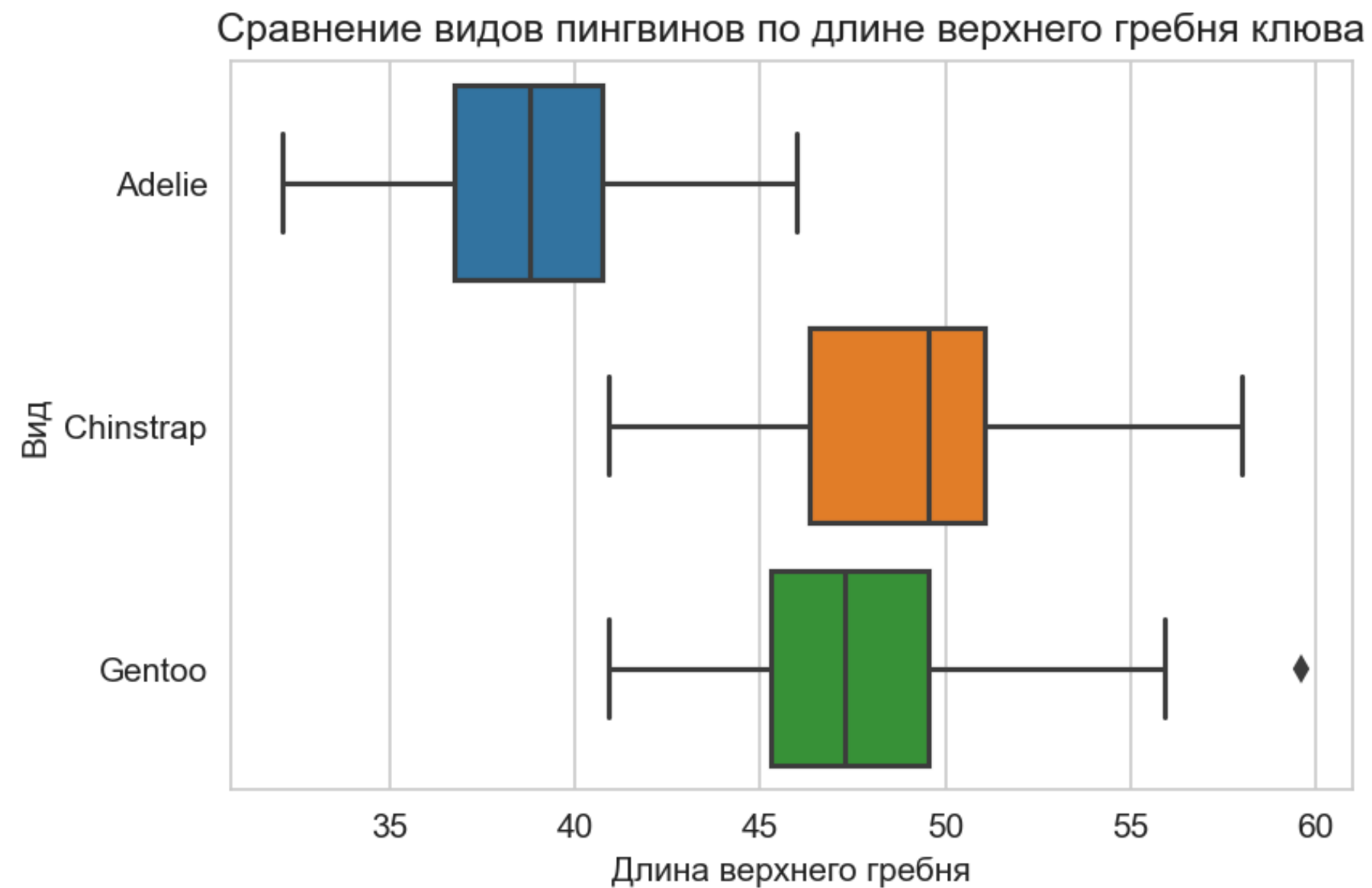


female male





**гистограммы:  
распределение**

**ящики с усами:  
распределение  
+ выбросы (!)**





# df.describe()

	df.describe()			
	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
count	342.000000	342.000000	342.000000	342.000000
mean	43.921930	17.151170	200.915205	4201.754386
std	5.459584	1.974793	14.061714	801.954536
min	32.100000	13.100000	172.000000	2700.000000
25%	39.225000	15.600000	190.000000	3550.000000
50%	44.450000	17.300000	197.000000	4050.000000
75%	48.500000	18.700000	213.000000	4750.000000
max	59.600000	21.500000	231.000000	6300.000000

# df.describe()

какие переменные перед нами  
(категориальные / количественные,  
меры среднего / вариативности?)

```
df.describe()
```



	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
count	342.000000	342.000000	342.000000	342.000000
mean	43.921930	17.151170	200.915205	4201.754386
std	5.459584	1.974793	14.061714	801.954536
min	32.100000	13.100000	172.000000	2700.000000
25%	39.225000	15.600000	190.000000	3550.000000
50%	44.450000	17.300000	197.000000	4050.000000
75%	48.500000	18.700000	213.000000	4750.000000
max	59.600000	21.500000	231.000000	6300.000000

- количество
- среднее
- станд.отклонение
- минимум
- 1 квартиль
- медиана (=2 квартиль)
- 3 квартиль
- максимум (=4 квартиль)

# df.describe()

скорее всего, не понадобится,  
но в 1 задании демоверсии есть

так ищем квантили (25%, 50%, 75%)

```
df.describe()['столбец']['25%']
```

или так (для продвинутых):

```
import numpy as np  
np.quantile(df['столбец'], 0.25)
```



`df[['species', 'island', 'sex']].describe()`

	species	island	sex
count	344	344	333
unique	3	3	2
top	Adelie	Biscoe	MALE
freq	152	168	168

меры среднего  
и вариативности  
категориальных  
переменных

# df.describe(include='all')

меры среднего И вариативности  
категориальных И количественных переменных

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
count	344	344	342.000000	342.000000	342.000000	342.000000	333
unique	3	3	NaN	NaN	NaN	NaN	2
top	Adelie	Biscoe	NaN	NaN	NaN	NaN	MALE
freq	152	168	NaN	NaN	NaN	NaN	168
mean	NaN	NaN	43.921930	17.151170	200.915205	4201.754386	NaN
std	NaN	NaN	5.459584	1.974793	14.061714	801.954536	NaN
min	NaN	NaN	32.100000	13.100000	172.000000	2700.000000	NaN
25%	NaN	NaN	39.225000	15.600000	190.000000	3550.000000	NaN
50%	NaN	NaN	44.450000	17.300000	197.000000	4050.000000	NaN
75%	NaN	NaN	48.500000	18.700000	213.000000	4750.000000	NaN
max	NaN	NaN	59.600000	21.500000	231.000000	6300.000000	NaN

## Важные последние замечания:

в `describe()` :

- НЕТ дисперсии, но `.std() ** 2`
- `.std()` и `.var()` считаются к выборке (БЕЗ `ddof=0`)






02

**2 слова о  
выборках**

# Проблемы с выборками



*Журнал Literary Digest проводил опросы общественного мнения перед выборами президента в 1920, 1924, 1928, 1932, а также и в 1936 году, и каждый раз прогноз, составленный на основе опроса, оказывался верным*

Разосланы 10 млн анкет:

подписчикам журнала;  
людям, выбранным по телефонным книгам;  
по спискам регистрации автомобилей.

Около 2,5 млн анкет заполнены:

57 % –за республиканца Альфа Лэндона,  
40 % –за демократа Франклина Рузвельта.

# Задача из НЭ, А5 / А10

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

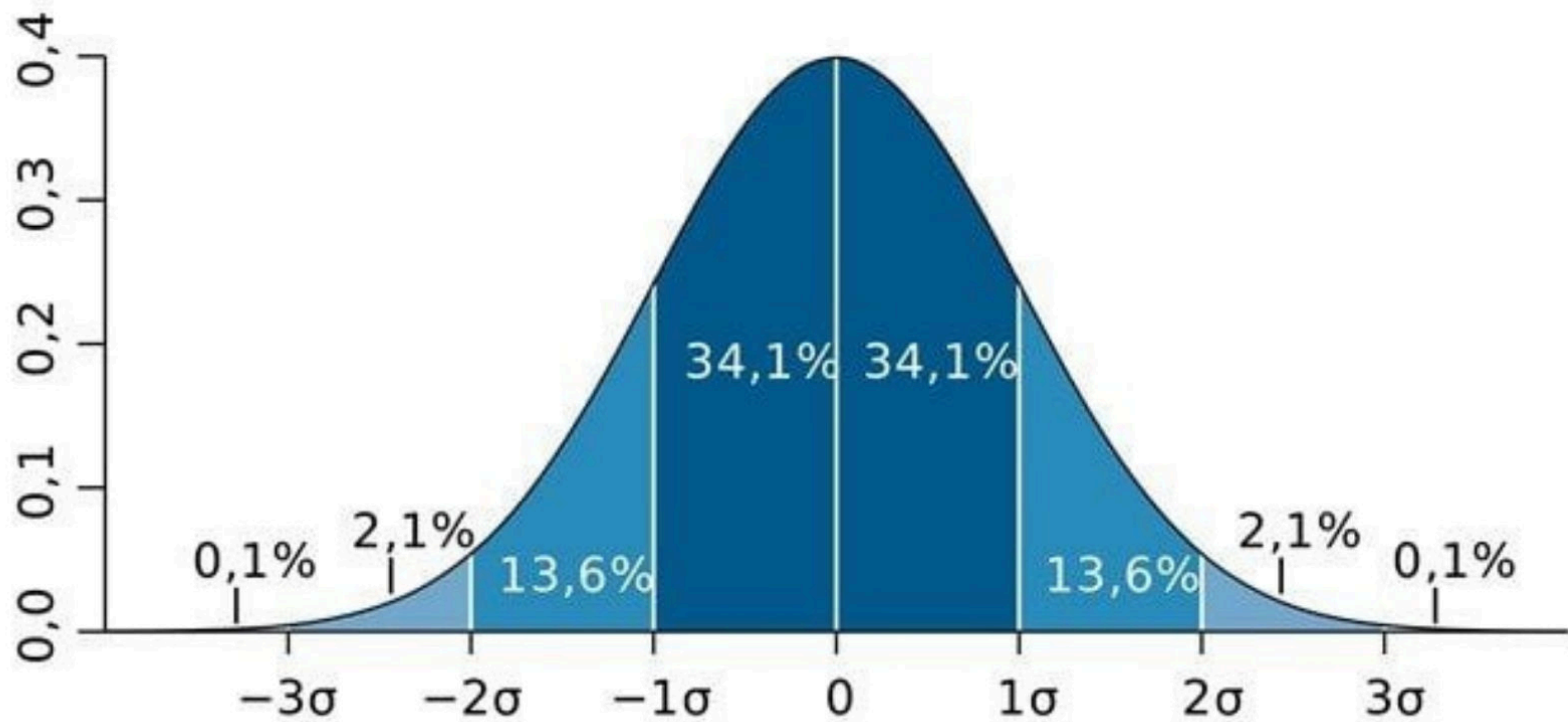
1. Было опрошено такое количество женщин из каждого региона, которое соответствует доле женщин, проживающих в этом регионе согласно переписи населения.
2. Опрос был проведен посредством социальных сетей для их пользователей.
3. Были опрошены жительницы всех городов с населением более 300 000 человек.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

# Задача из НЭ, А5 / А10

Аналитик Алексей занимается исследованием рынка кофеен. Какая из собранных им выборок будет **более репрезентативной**, чем другие?

1. Данные о дневных продажах кофеен разных брендов, собранные по разным городам в разные дни
2. Данные о дневных продажах кофеен города Москва, собранные в разные дни
3. Данные о дневных продажах кофеен одного бренда, собранные в разные дни
4. Данные о продажах кофеен разных брендов, собранные по разным городам и в один день



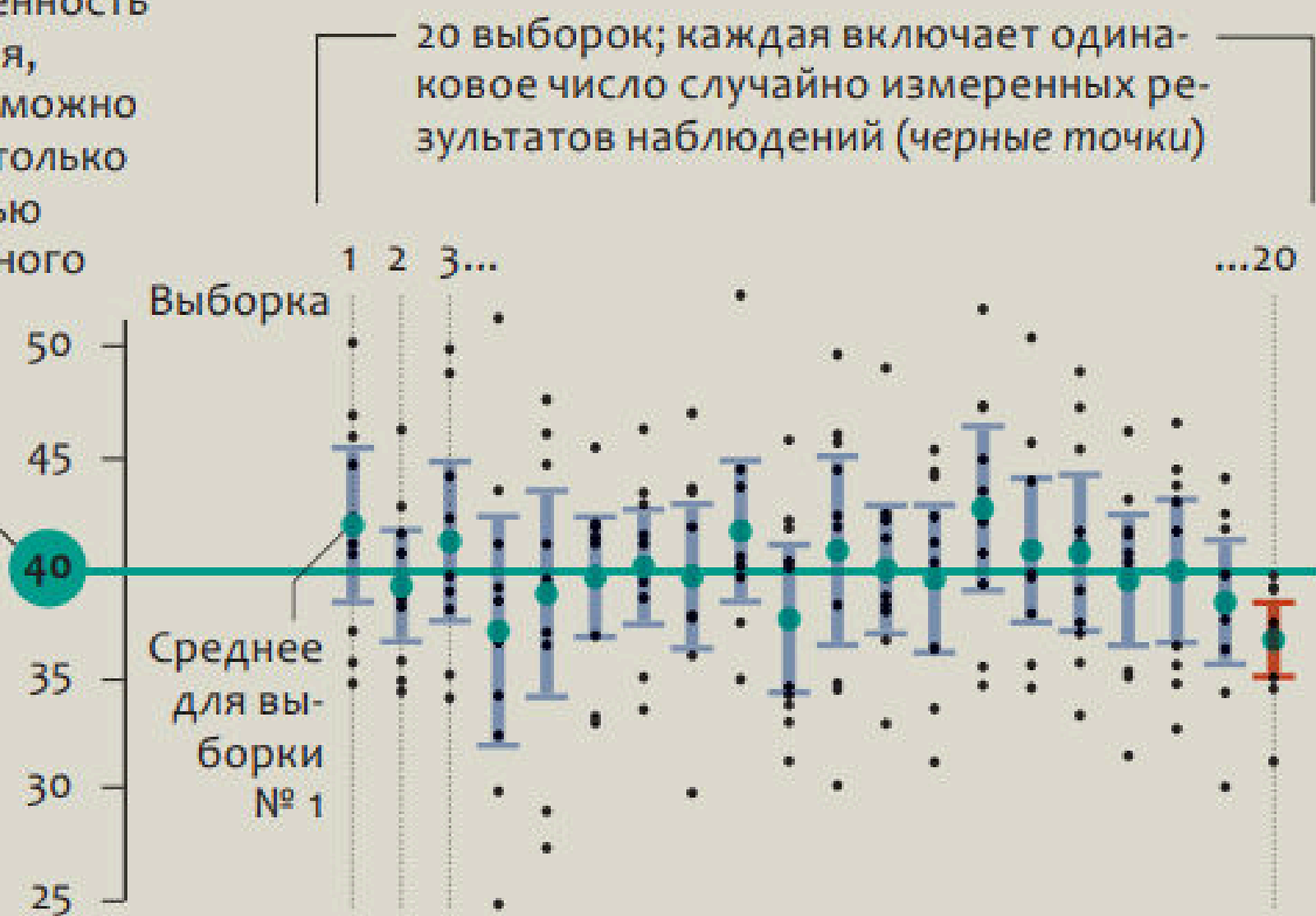




# В реальном мире в выборки закладывают интервал доверия (напр., 95%)



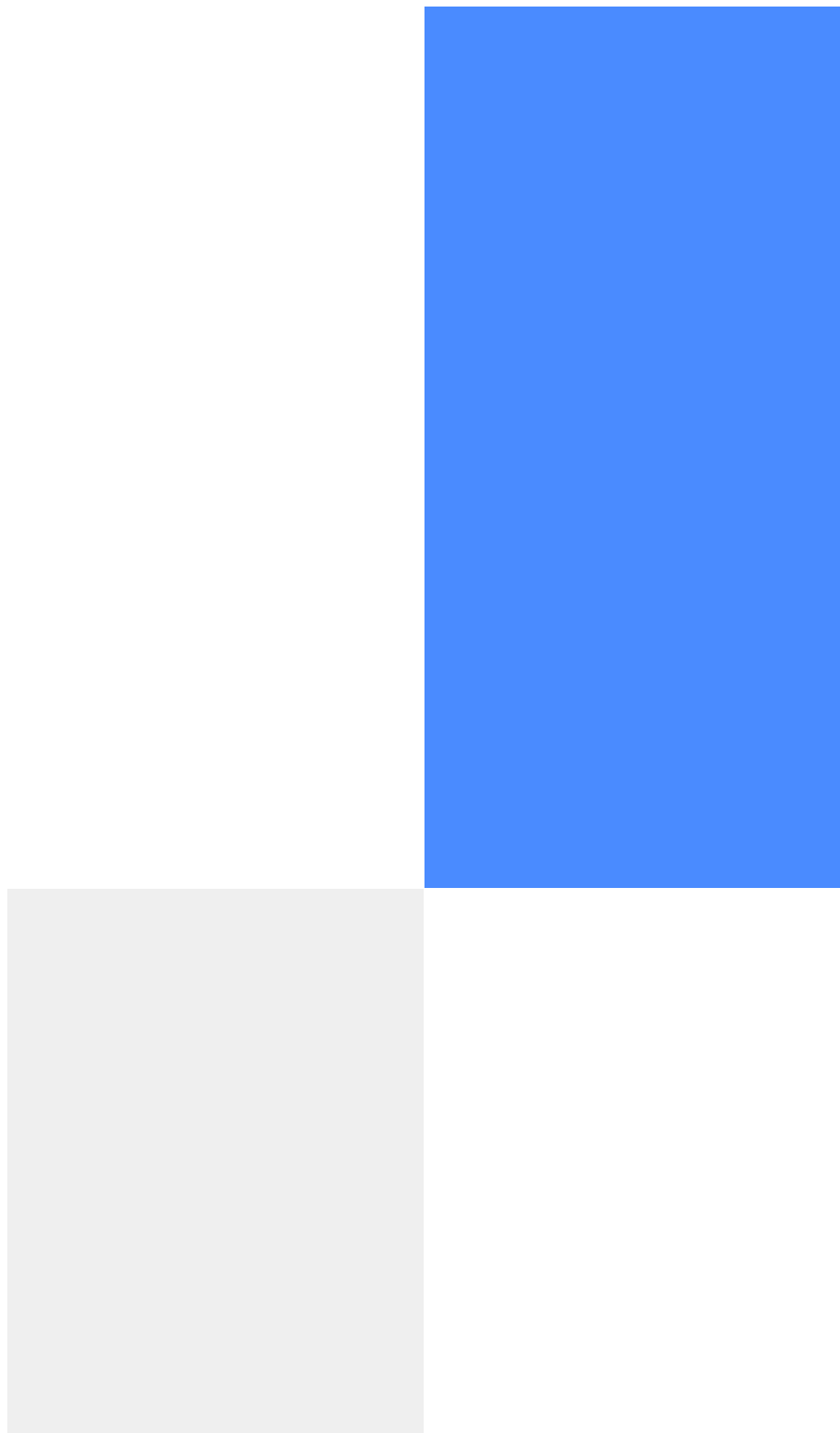
Реальная средняя численность населения, которую можно оценить только с помощью выборочного анализа



## Что означает доверительный интервал?

Если мы берем интервал ошибок или доверительные границы с уровнем доверия 95%, то это означает, что на указанном интервале истинное значение содержится с вероятностью 95%. Однако когда мы говорим, что в пределах доверительного интервала находится истинное значение, то здесь смысл следующий: если мы будем повторять случайные выборки одного и того же размера достаточно большое количество раз и независимо друг от друга, то в 95% случаев истинное значение будет попадать в доверительный интервал. Несмотря на то что на практике такое распространенное ошибочное толкование не может кардинально сказаться на принятии решений, то, что даже ученые неправильно понимают смысл доверительных интервалов, показывает, насколько сложно корректно интерпретировать изображения, в которых отражается неопределенность.

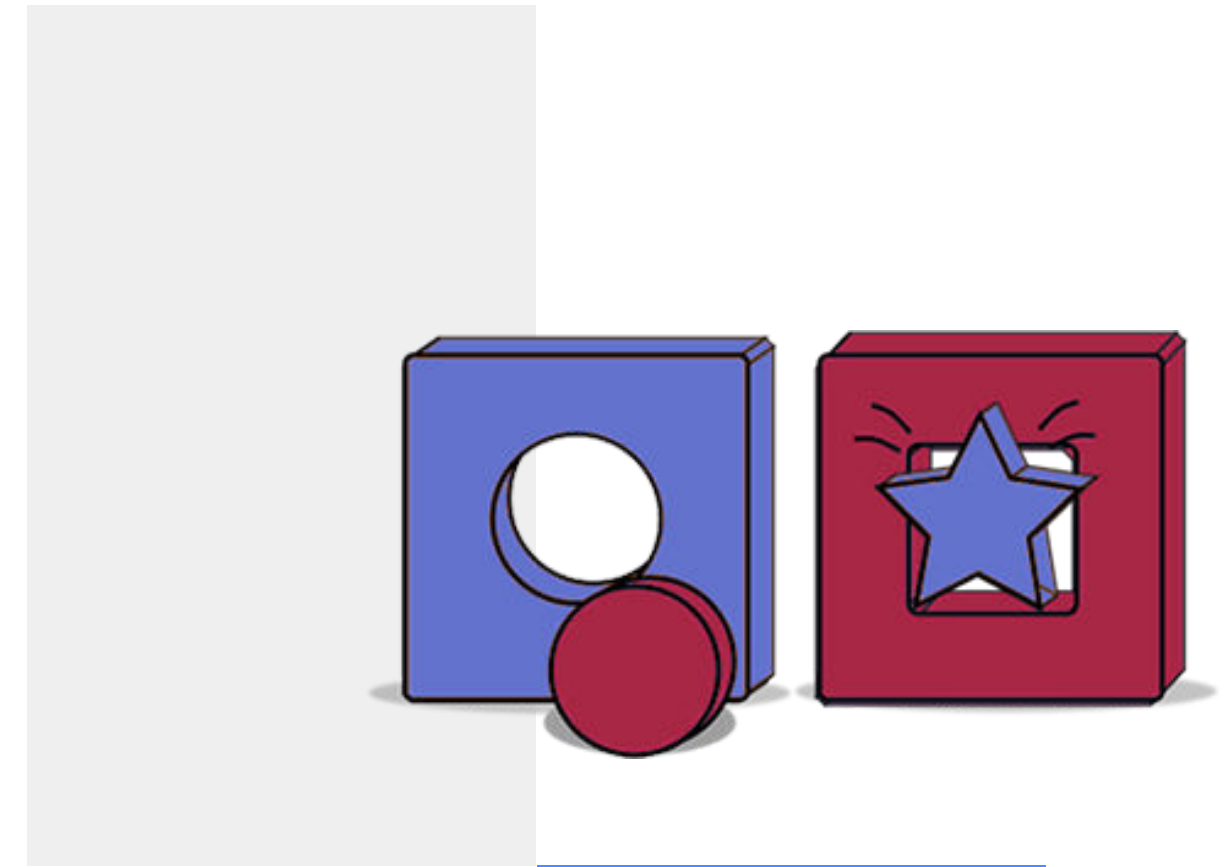
Несмотря на правильные подсчеты, значение реальной средней численности населения все-таки не попало в одну из 20 выборок, у каждой из которых доверительный интервал равнялся 95%.



03

Пропуски  
NaN

# Удаляем пропуски



Радикально)

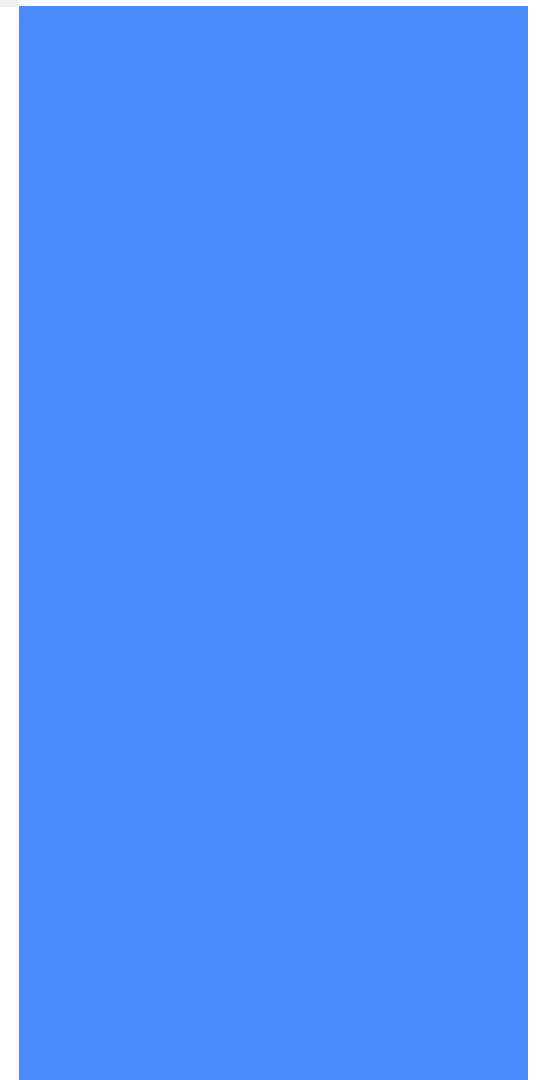
```
df.dropna()
```

По столбцу

```
df['столбец'].dropna()
```

Но! это опять аналог, "покажи, что будет"...

```
df.dropna(inplace=True)  
df = df.dropna()
```



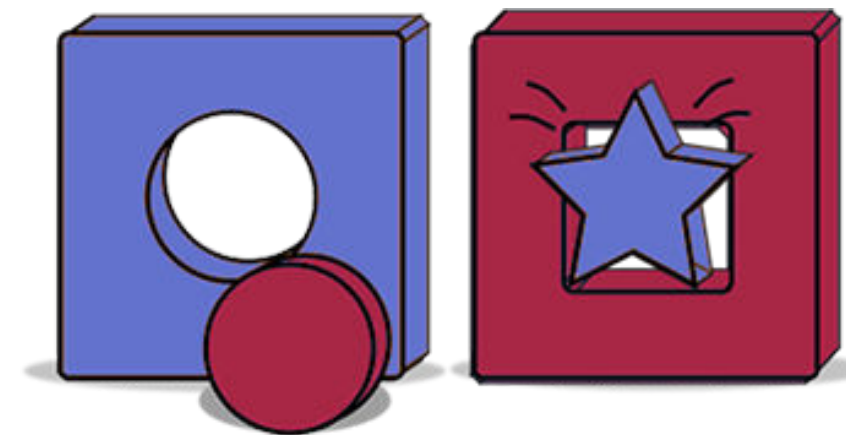
# Заполняем пропуски

## По столбцу

```
df['столбец'].fillna(0)
```

```
my_median = df['столбец'].median()  
df['столбец'].fillna(my_median)
```


```
df['столбец'].fillna(0, inplace=True)
```





04

# Корреляции



# Измерение силы связи между двумя количественными переменными

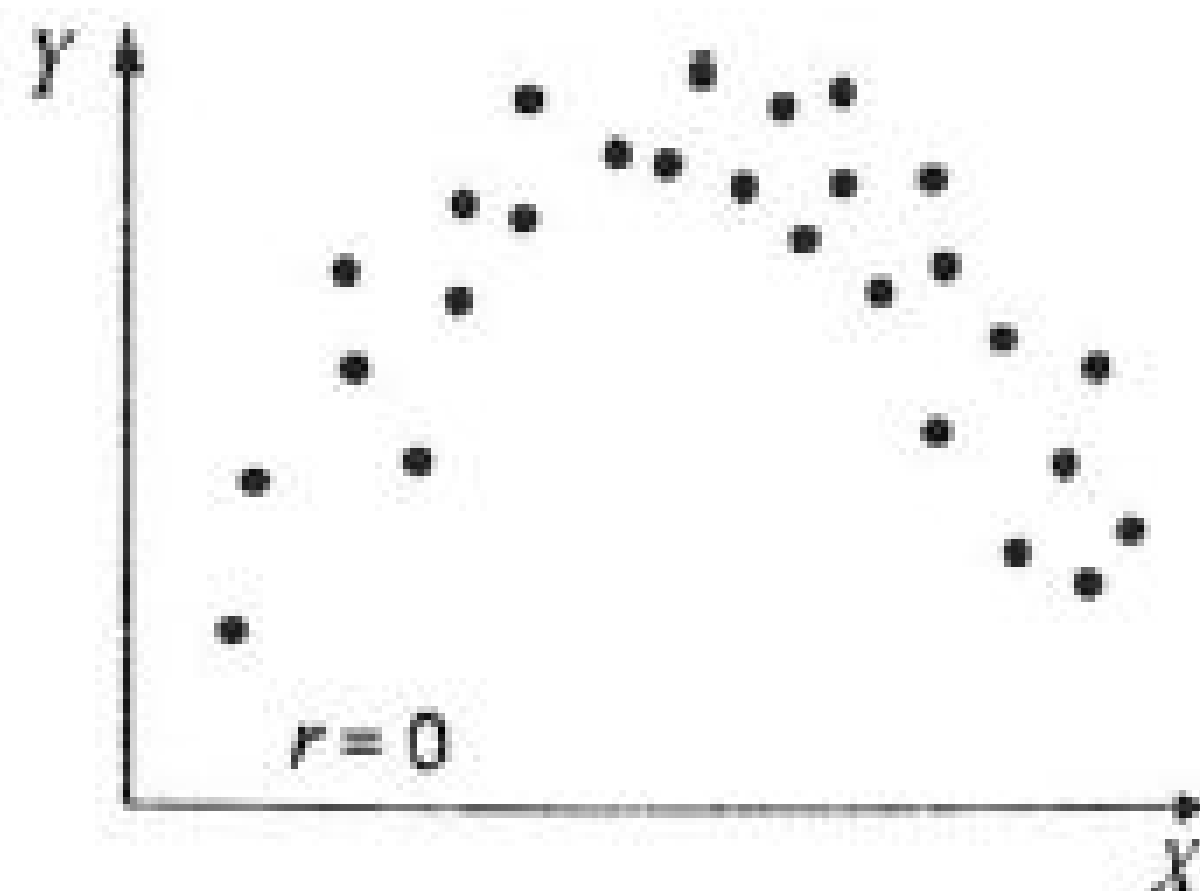
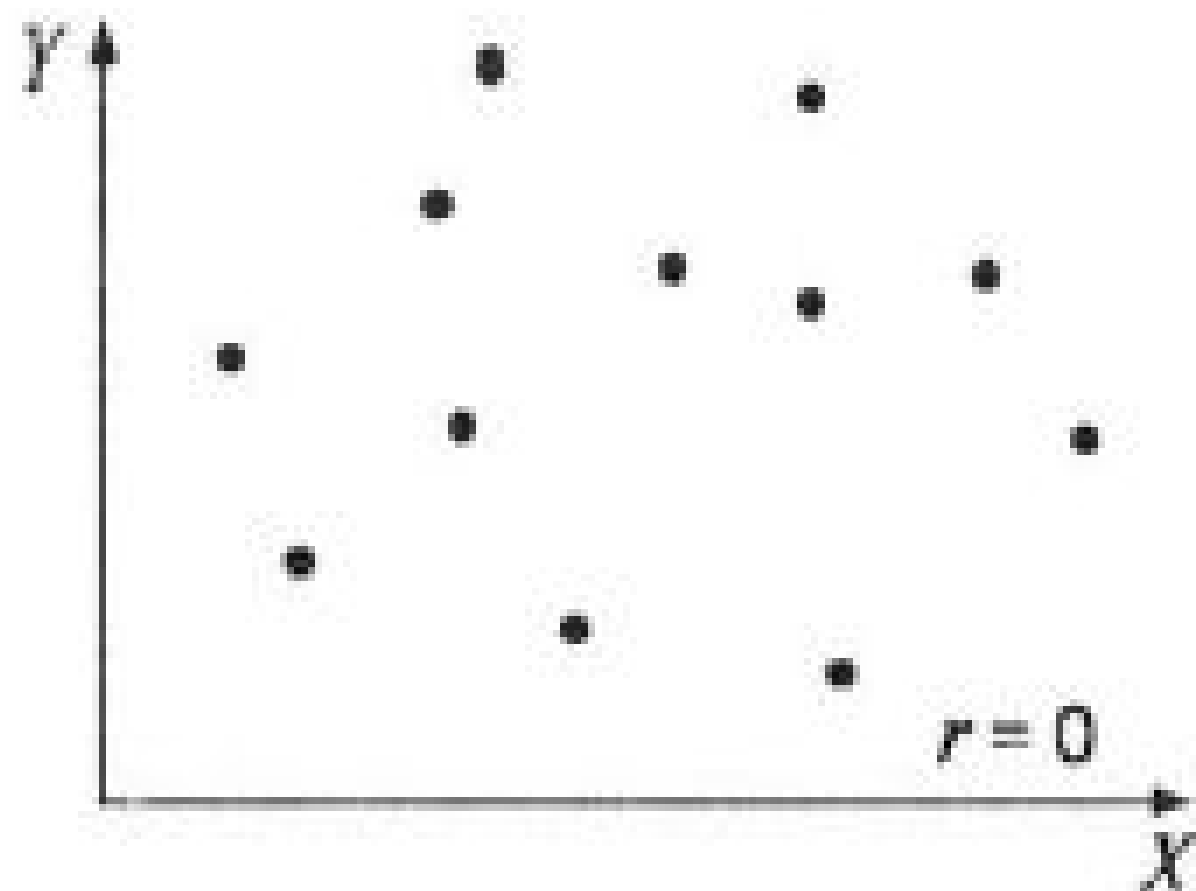
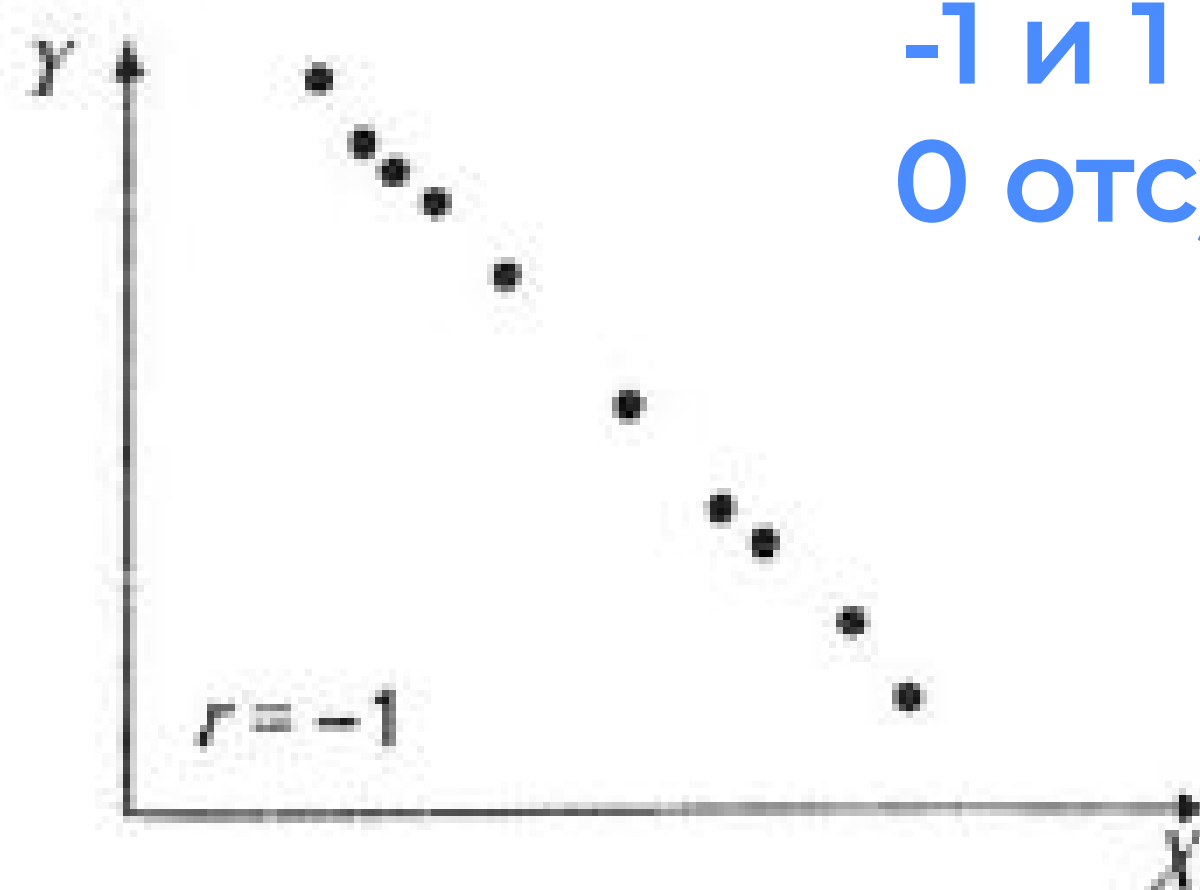
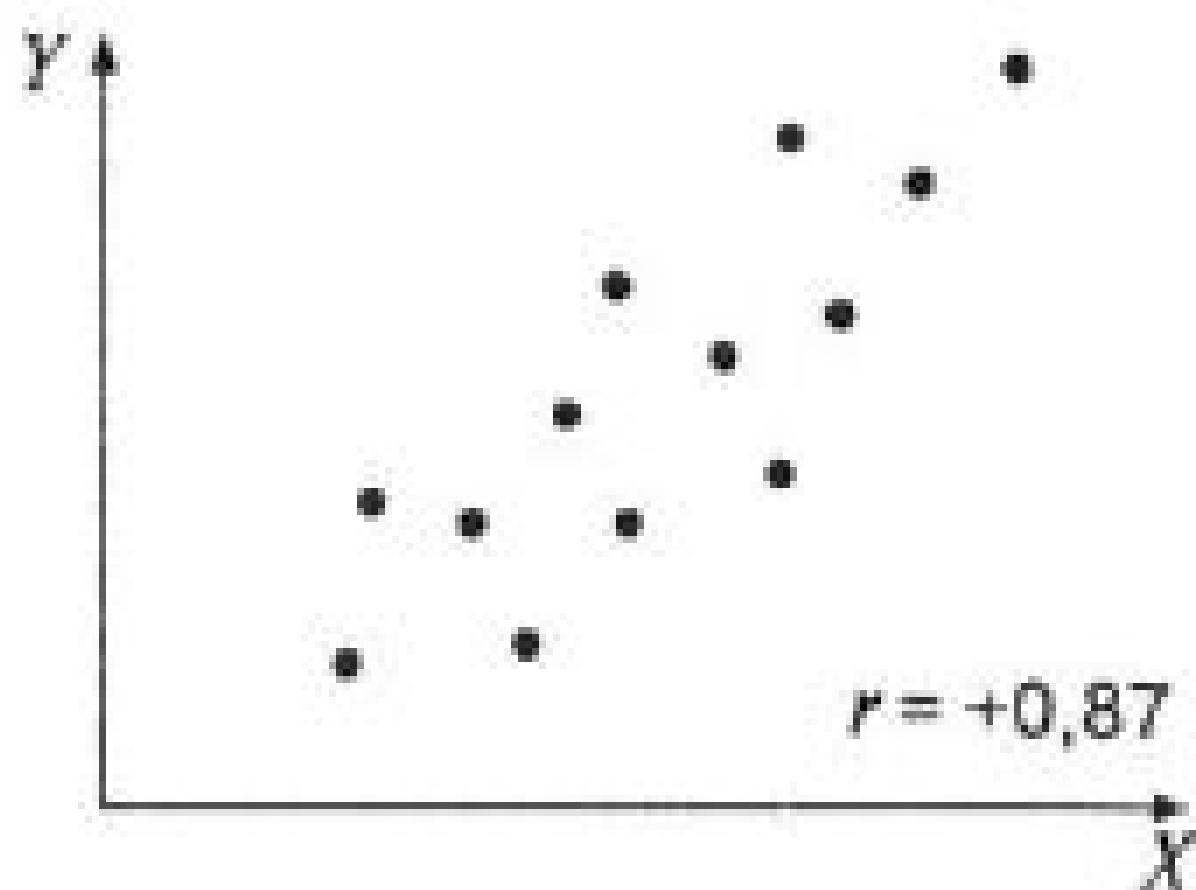
```
import numpy as np  
np.corrcoef(X, Y)
```

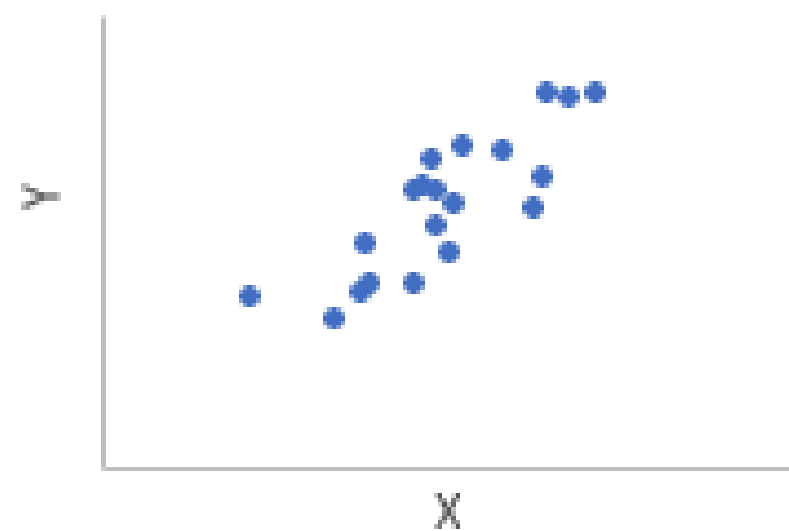
**2 списка  
(вектора\*)**

```
df.corr()
```

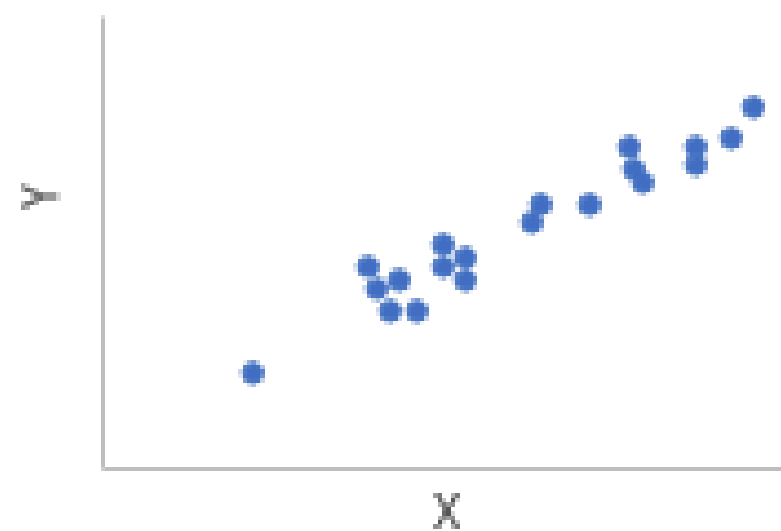
**таблица**

-1 и 1 сильная  
0 отсутствует

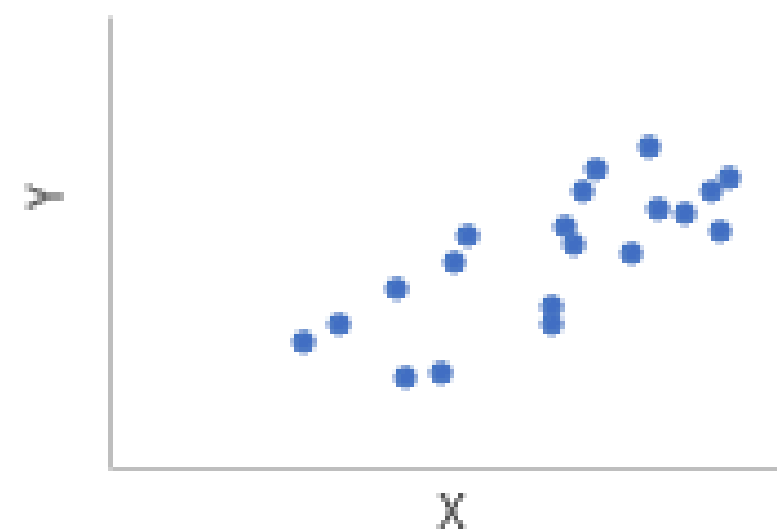




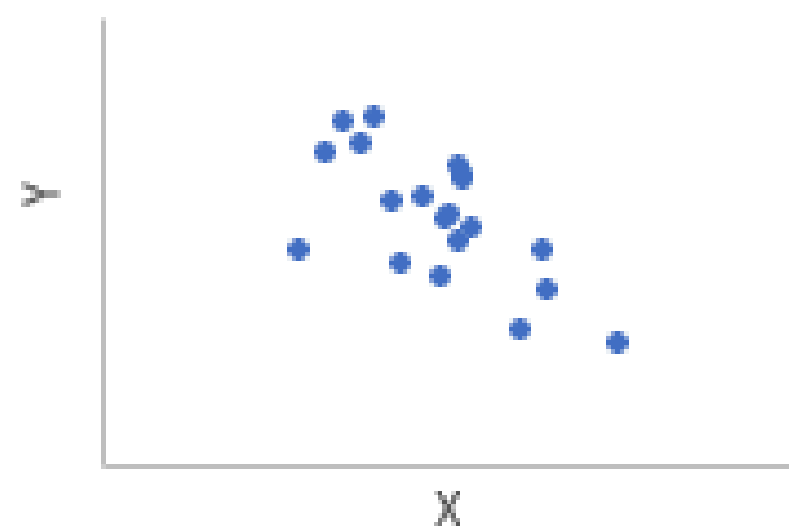
Прямая



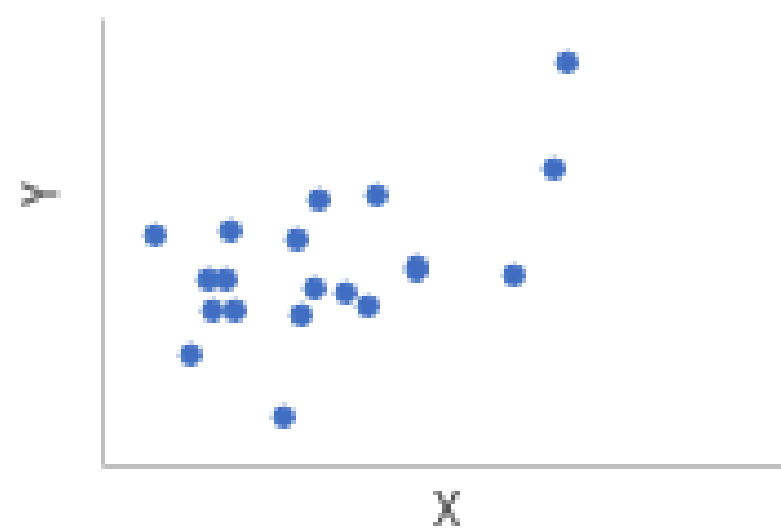
Сильная



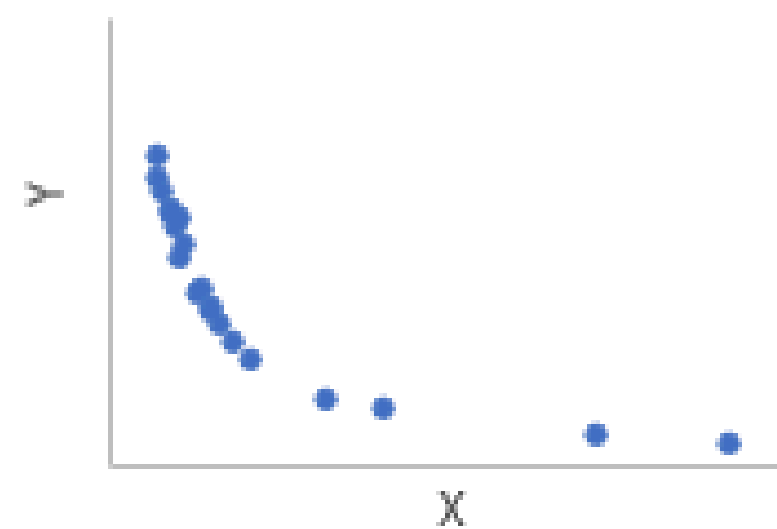
Линейная



Обратная



Слабая

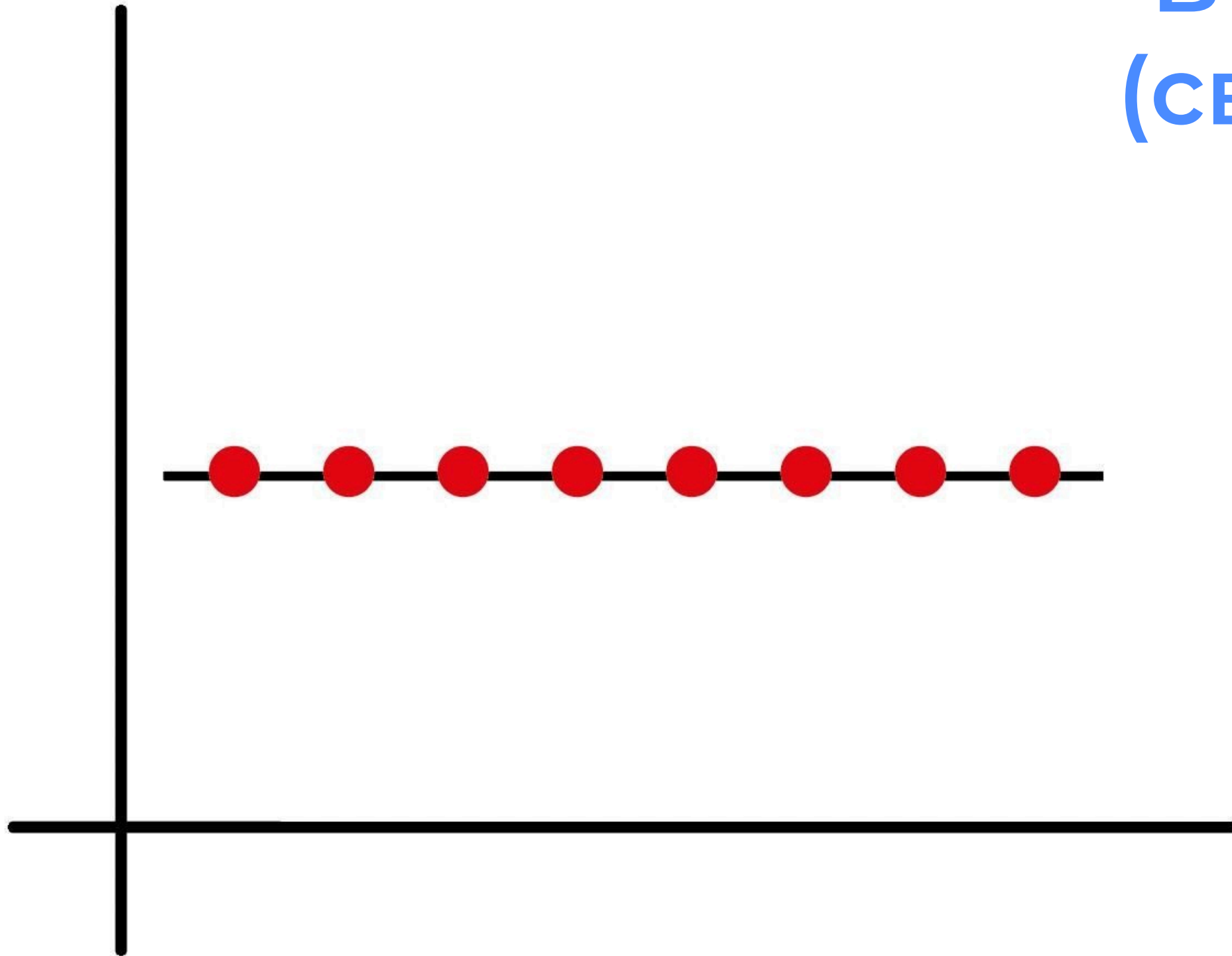


Нелинейная

-1 и 1 сильная  
0 отсутствует



В прошлогоднем НЭ  
(связь здесь равна 0)





# Ограничения

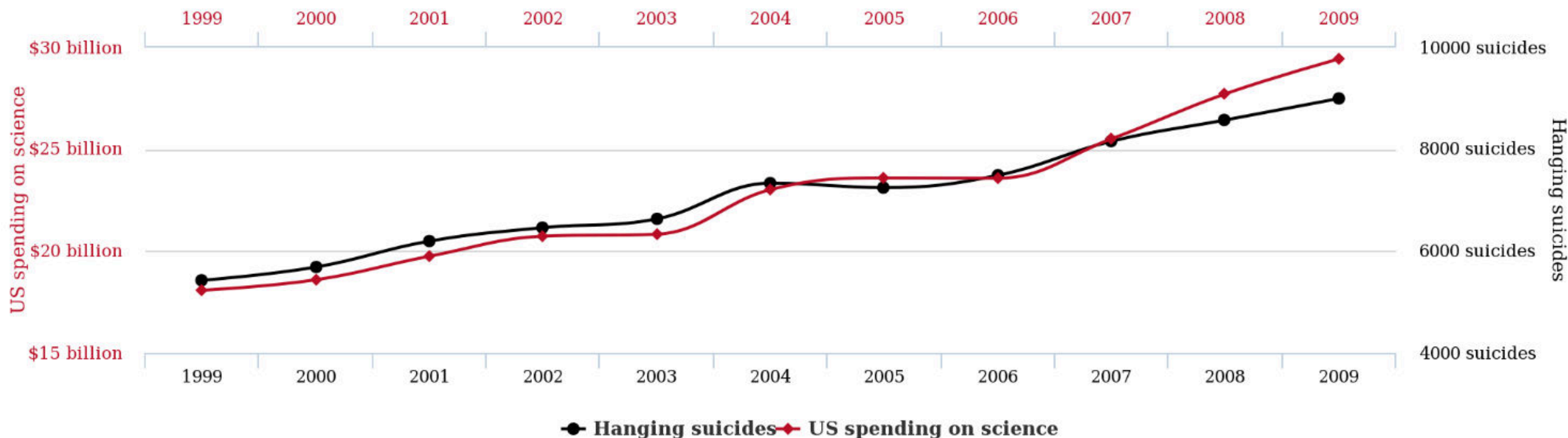
**1) можно измерить  
силу ТОЛЬКО  
линейной связи**

**2) связь != причина**

# Безумные корреляции ★

Нашел серию графиков, которые смешно иллюстрируют разницу между корреляцией данных и причинно-следственной связью.

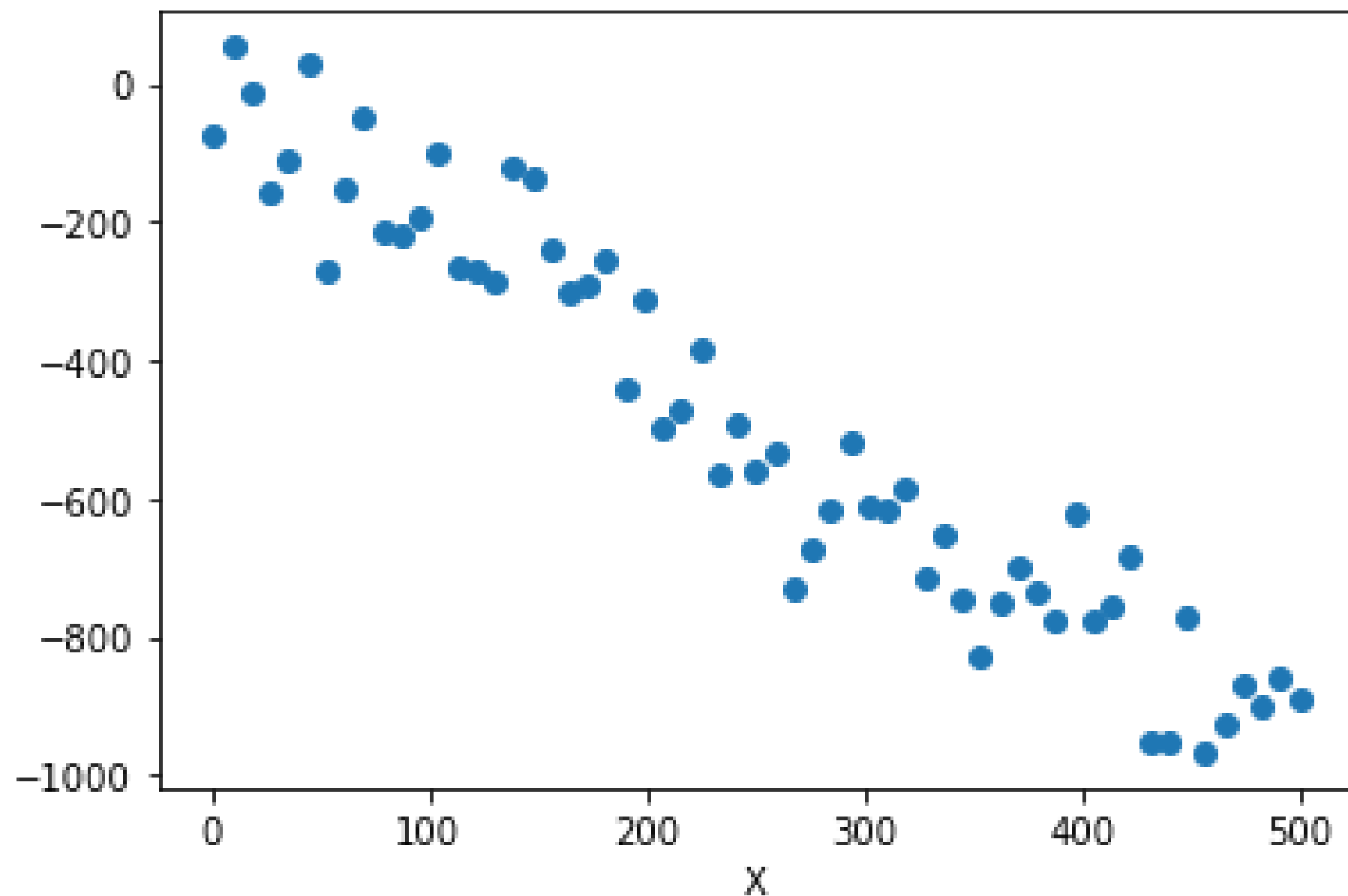
## US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation



tylervigen.com

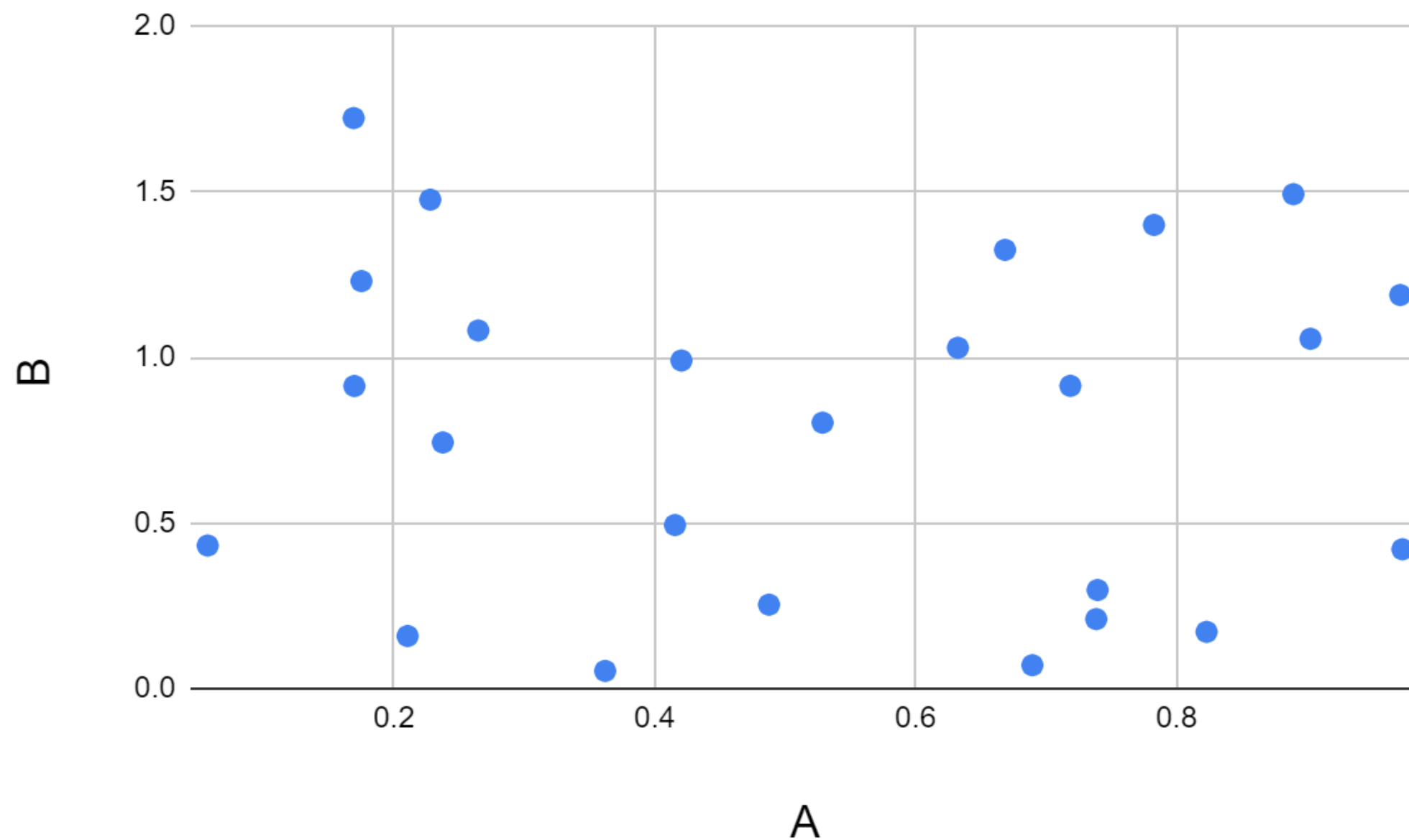
Затраты США на науку, космос и технологии / Суициды путем повешения и удушения. Корреляция 99,79%.

# Задача из НЭ, А10



1. Между переменными  $X$  и  $Y$  существует сильная положительная линейная взаимосвязь
2. Между переменными  $X$  и  $Y$  существует сильная отрицательная линейная взаимосвязь
3. Между переменными  $X$  и  $Y$  существует сильная положительная нелинейная взаимосвязь
4. Между переменными  $X$  и  $Y$  существует сильная отрицательная нелинейная взаимосвязь

# Задача из НЭ, А11



Одно из чисел ниже — коэффициент корреляции между этими двумя переменными. Выберите это число:

- 0.92
- 0.61
- -0.80
- -0.05

# Задача из НЭ, А11

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?

1. Удовлетворенность собственным трудом напрямую влияет на продуктивность.
2. Продуктивность на работе влияет на удовлетворенность собственным трудом.
3. Между удовлетворенностью работой и продуктивностью слабая прямая взаимосвязь.
4. **Если у сотрудника повышается продуктивность, то, скорее всего, повысится и удовлетворенность работой.**

# Задача из НЭ, А11

## Никогда не выбираем:

А (напрямую / косвенно) влияет на В (или В на А?)

А является причиной В

Между А и В нелинейная связь (корреляция ее вообще не чувствует!)

## Всегда выбираем:

Между А и В сильная / слабая положительная / отрицательная связь

Когда А возрастает, В скорее всего возрастает (убывает)



# Задача из НЭ, А11

Всегда выбираем:

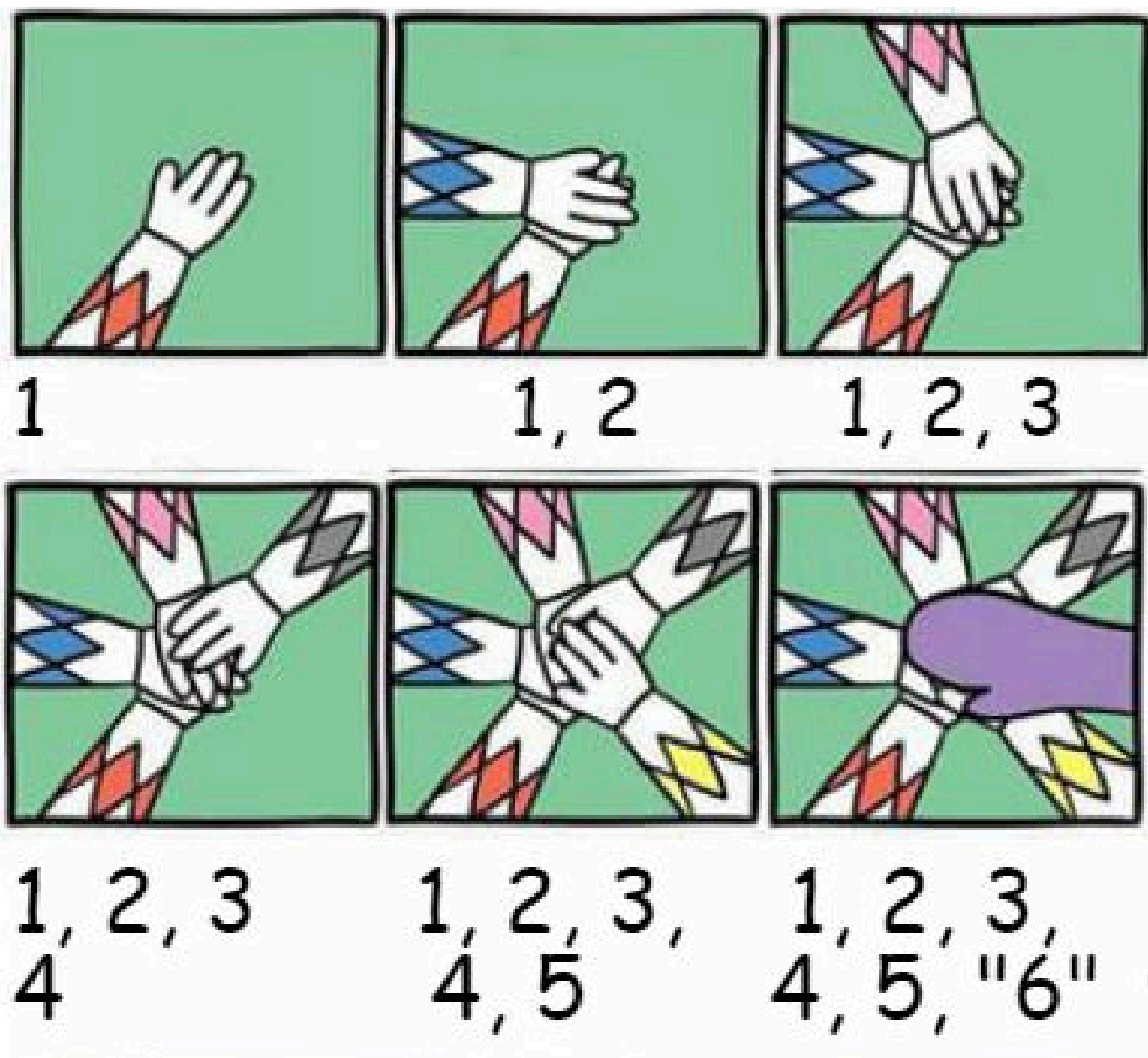
- Между А и В сильная / слабая положительная / отрицательная связь
- Когда А возрастает, В скорее всего возрастает (убывает)





# часть В и С

Остальные задачи на практику:  
рассчитайте корреляцию



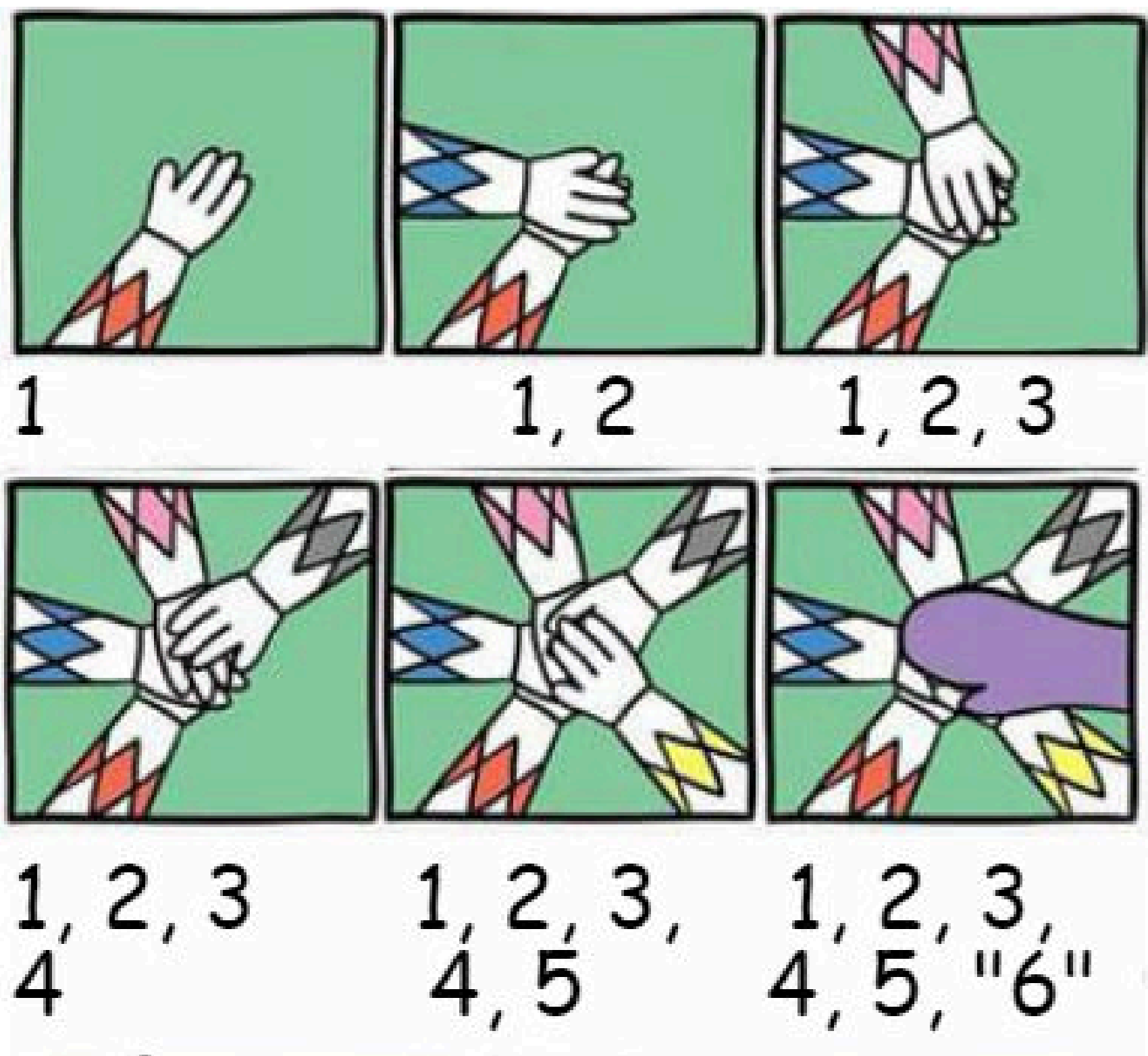
## Вектор (бонусик))

Это как список, но данные должны быть одного типа

```
import numpy as np  
np.array([1, 2, 3, 4, 5, 6])
```

Результат:  
[1 2 3 4 5 6]

**зачем??**

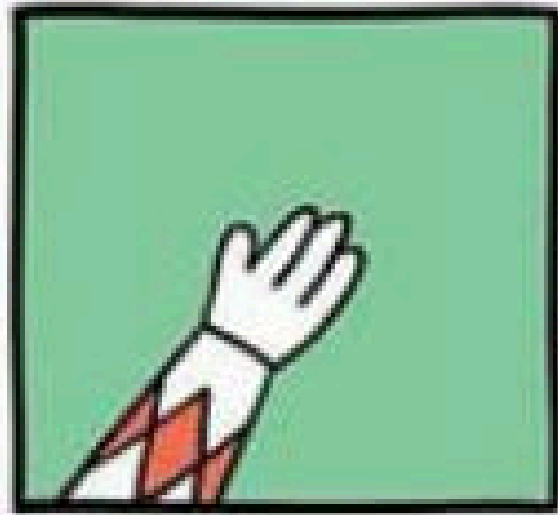


## Вектор

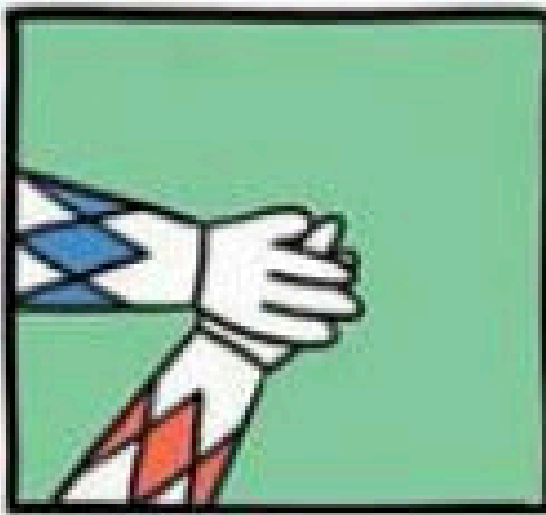
$[1, 2, 3] + [4, 5, 6] == [1, 2, 3, 4, 5, 6]$

`np.array([1, 2, 3]) +`  
`np.array([4, 5, 6]) ==`  
`[5 7 9]`

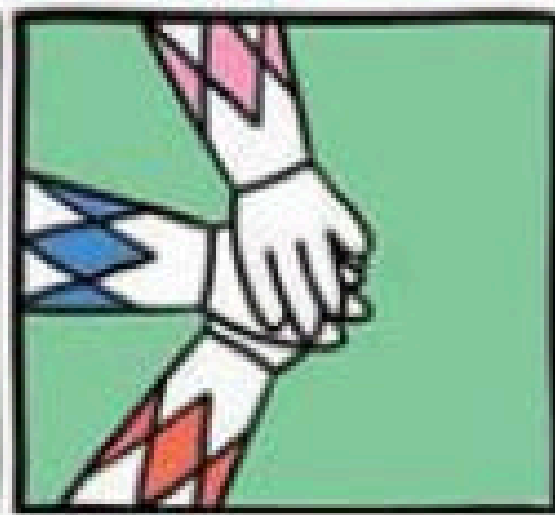
**`pd.Series` тоже вектор**



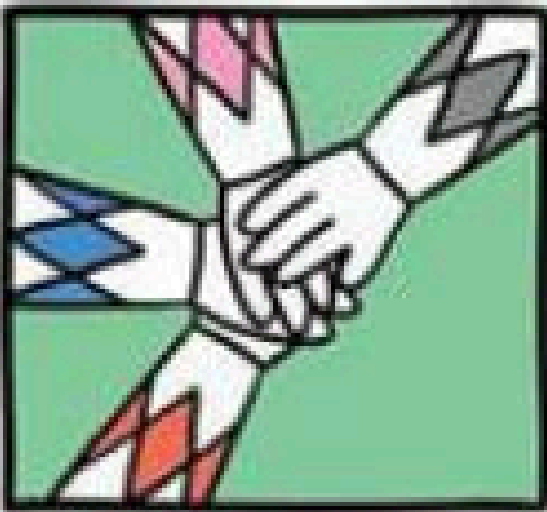
1



1, 2



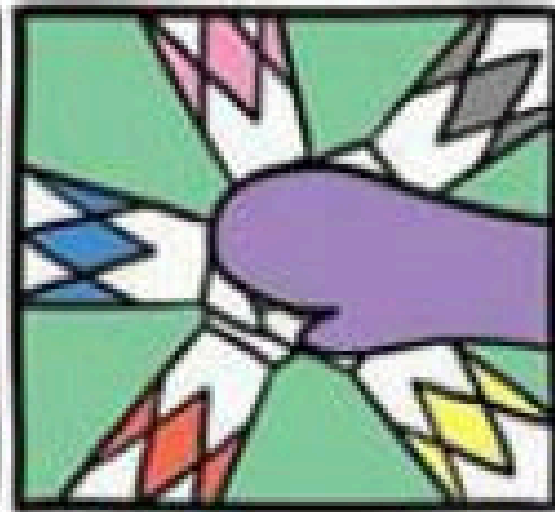
1, 2, 3



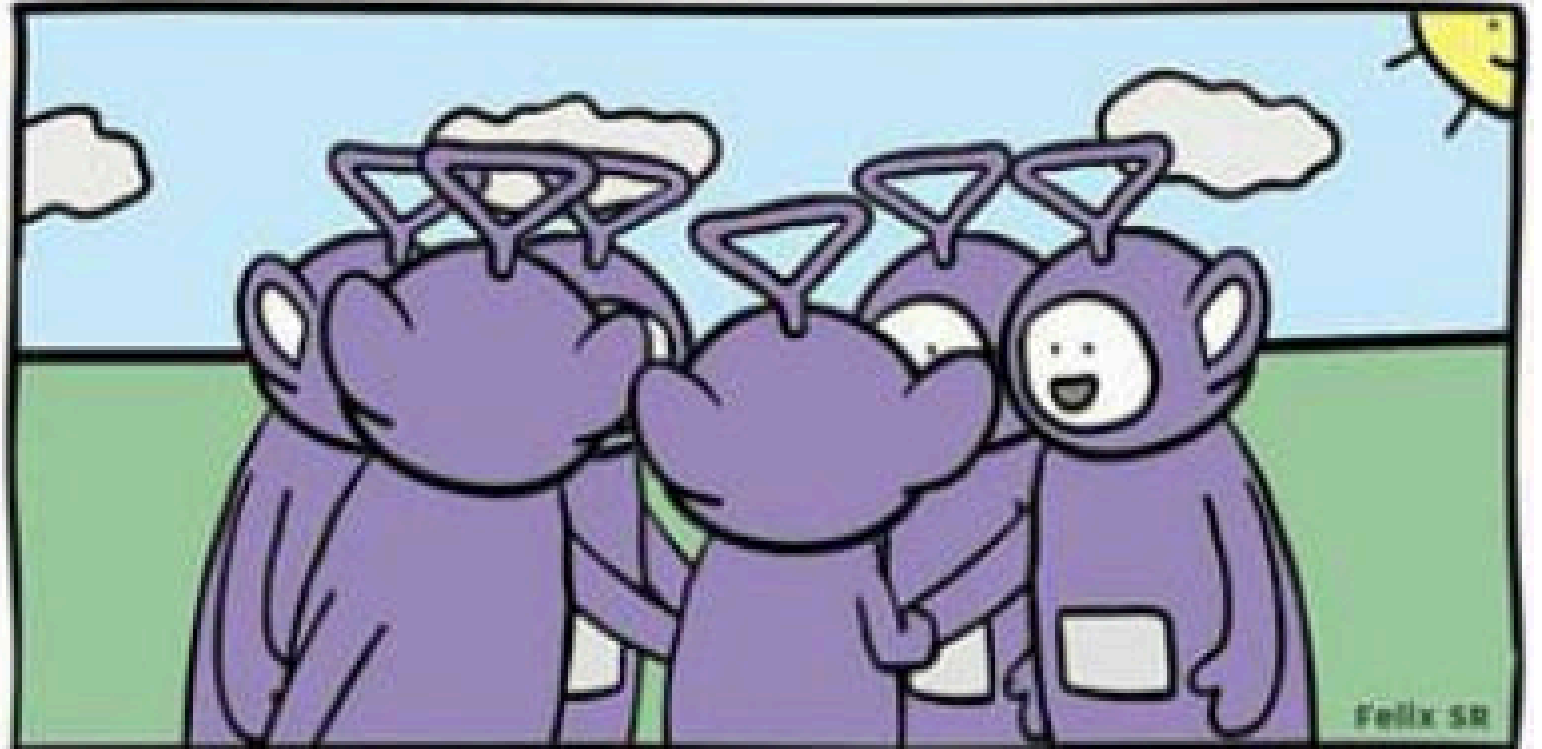
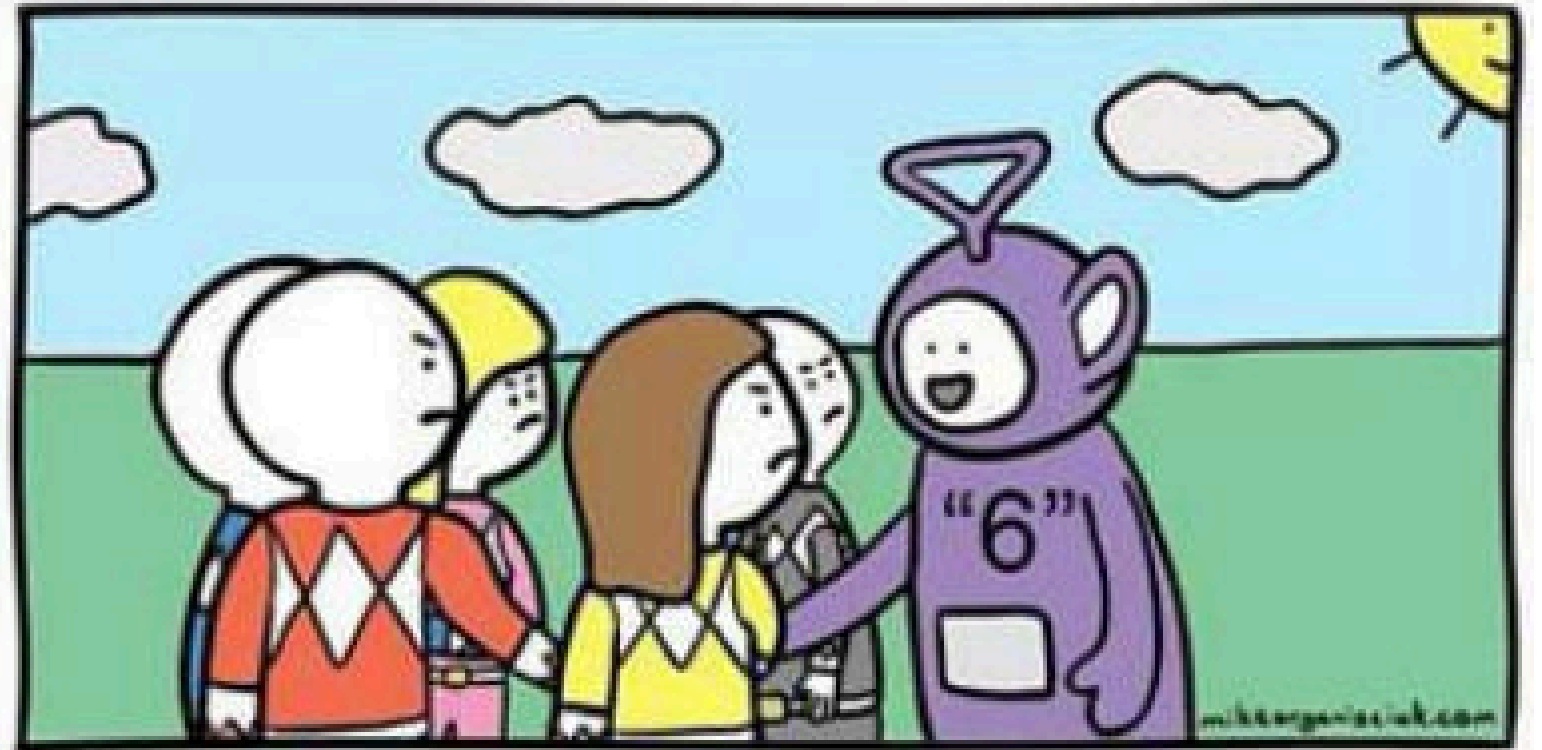
1, 2, 3  
4



1, 2, 3,  
4, 5



1, 2, 3,  
4, 5, "6"



"1", "2", "3", "4", "5", "6"