




Тема 3. Средние. Распределения. Выбросы*

Анализ данных 2024
(дааа, простите, лекции уходят вперед 🚀)



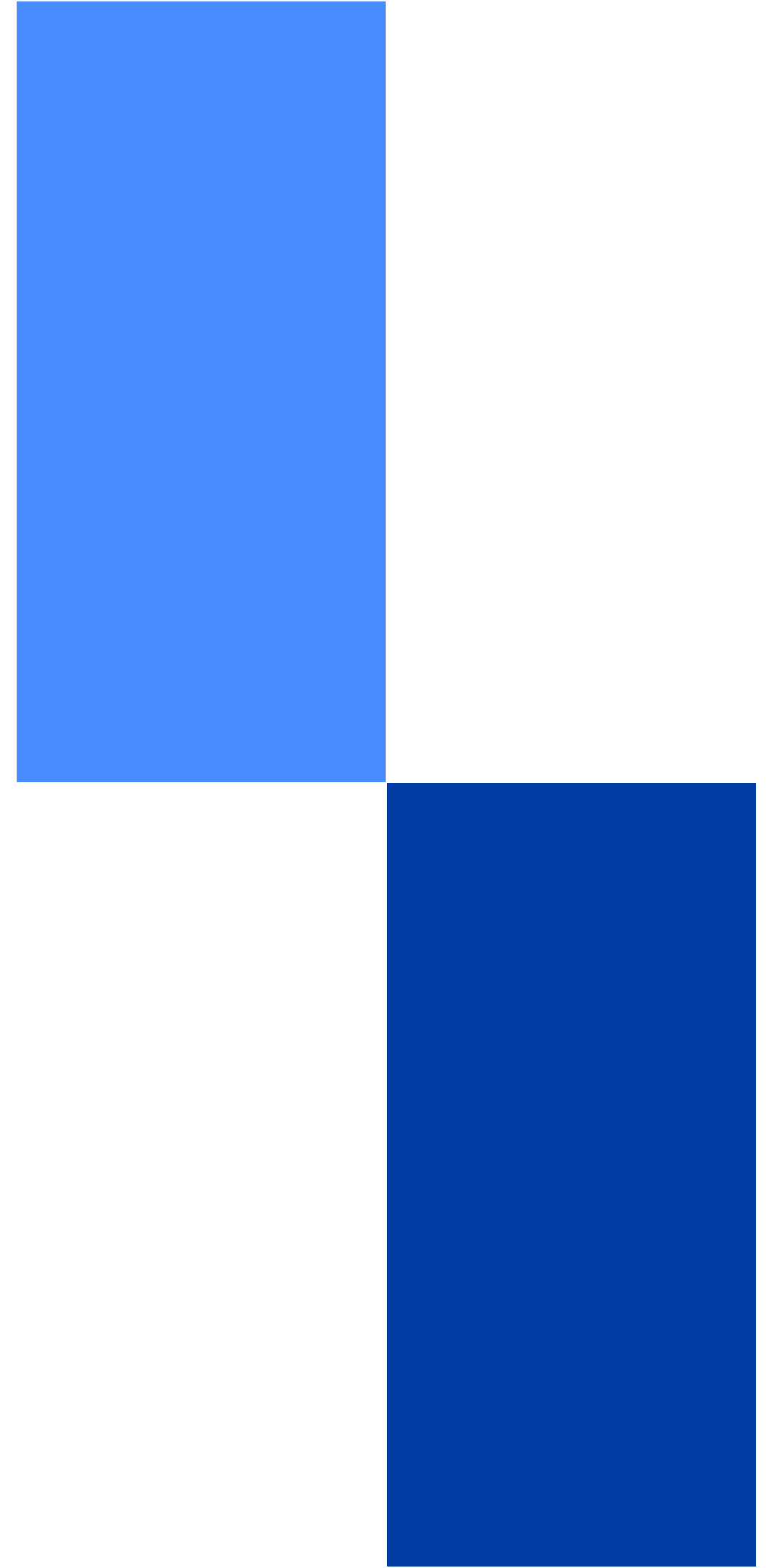
01

Вспоминаем конструкции

`df["столбец"]`

`df["столбец"].mean()`

`df["столбец"].value_counts()`





["столбец2"]

.mean()

.value_counts()...

df[df["столбец"] == 0]

df[(df["столбец"] == 0) | (df["столбец2"] > 0)]

С прошлой лекции

Еще не смотрели на практике, но обязательно!

```
df['длина имен'] = df['имя'].apply(len)  
df
```

[23]:


	имя	должность	длина имен
0	Анна	преподаватель	4
1	Никита	преподаватель	6
2	Илиана	академрук	6

С прошлой лекции

Еще не смотрели на практике, но обязательно!

```
def position(n):  
    if n == 'академрук':  
        return 2  
    else:  
        return 1  
  
df['должность_coded'] = df['должность'].apply(position)  
df
```

[25]:



	имя	должность	длина имен	должность_coded
0	Анна	преподаватель	4	1
1	Никита	преподаватель	6	1
2	Илиана	академрук	6	2

С прошлой лекции

Еще не смотрели на практике, но обязательно!

	order_id	price
0	1	5
1	1	6
2	1	1
3	2	20
4	3	2
5	3	5

```
In [3]: df.groupby('order_id').price.sum()
```

```
Out[3]: order_id
1      12
2      20
3       7
Name: price, dtype: int64
```



02
Новое)

Важная разница (прошлогодний НЭ)

Меры центральной тенденции

типичное,
повторяющееся,
общее в столбце

Меры вариативности

насколько данные
различны,
непохожи,
отличаются

количественные
категориальные

**Меры
центральной
тенденции**

мода

мода,
медиана,
среднее

**Меры
вариативности**

количество
уникальных
категорий

стандартное
отклонение,
дисперсия,
квартили

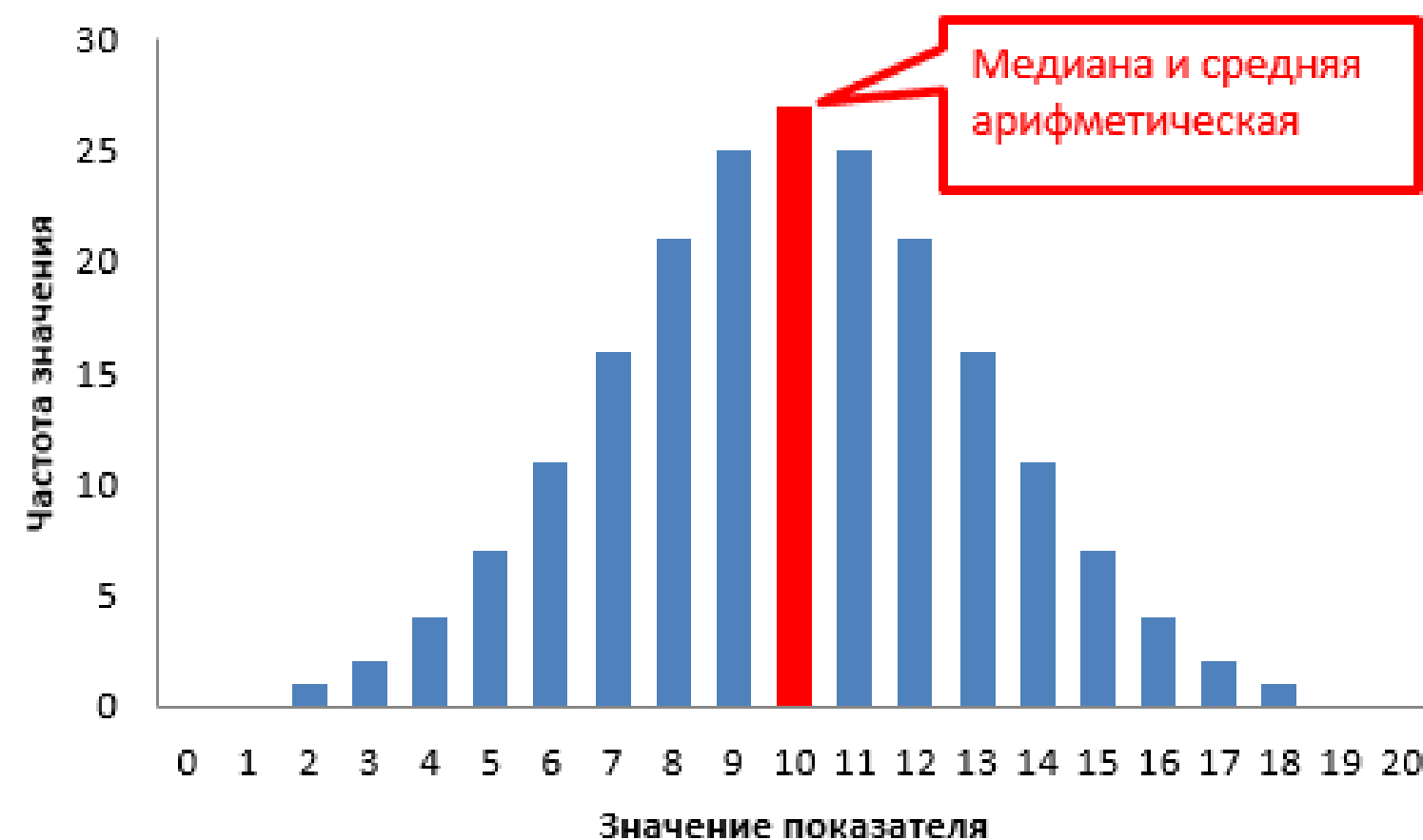
Вопрос из прошлогоднего НЭ

Какие меры ВАРИАТИВНОСТИ применимы к категории машин (зеленая, синяя, желтая)?

- мода
- количество уникальных категорий
- дисперсия
- квартиль

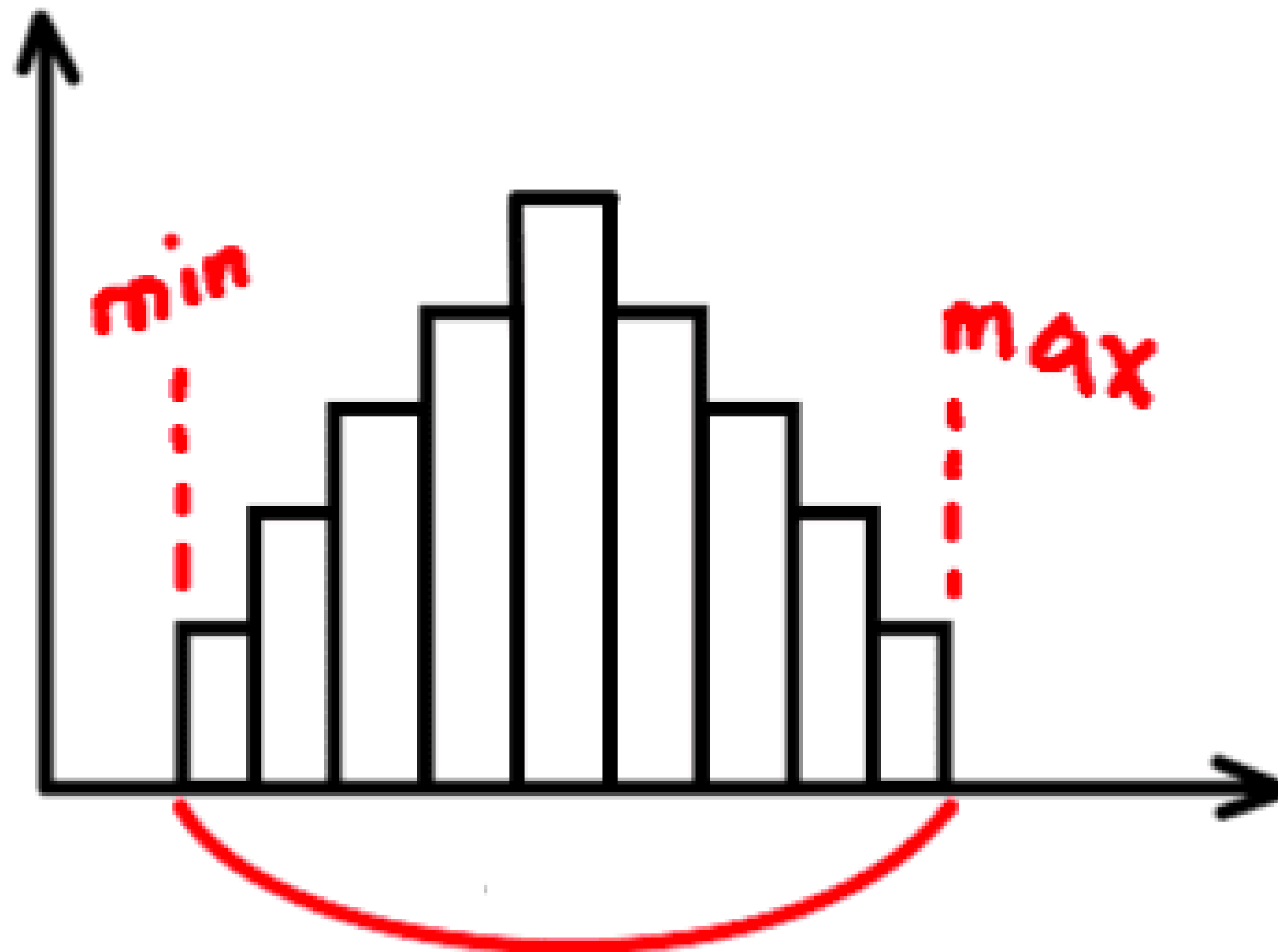
Термины:

- Мода **.mode()** - наиболее часто встречающееся (число, категория)
- Среднее **.mean()** - sum / len (арифметическое)
- Медиана **.median()** - среднее по порядку (устойчивое, честное)



Термины:

- минимум, максимум
- размах = макс - мин

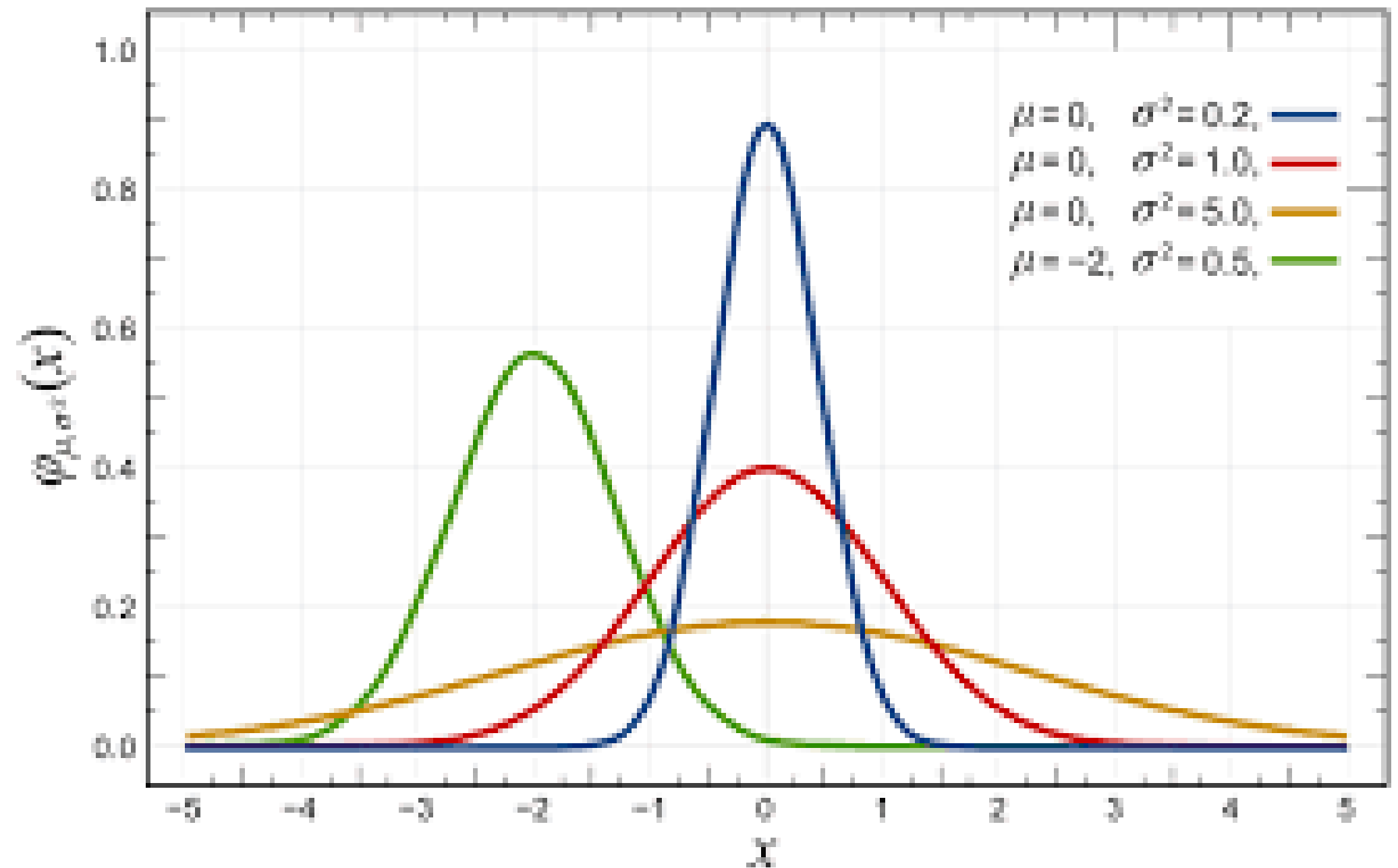


Термины:

- Стандартное (среднеквадратичное) отклонение **.std()**
- Дисперсия **.var()**

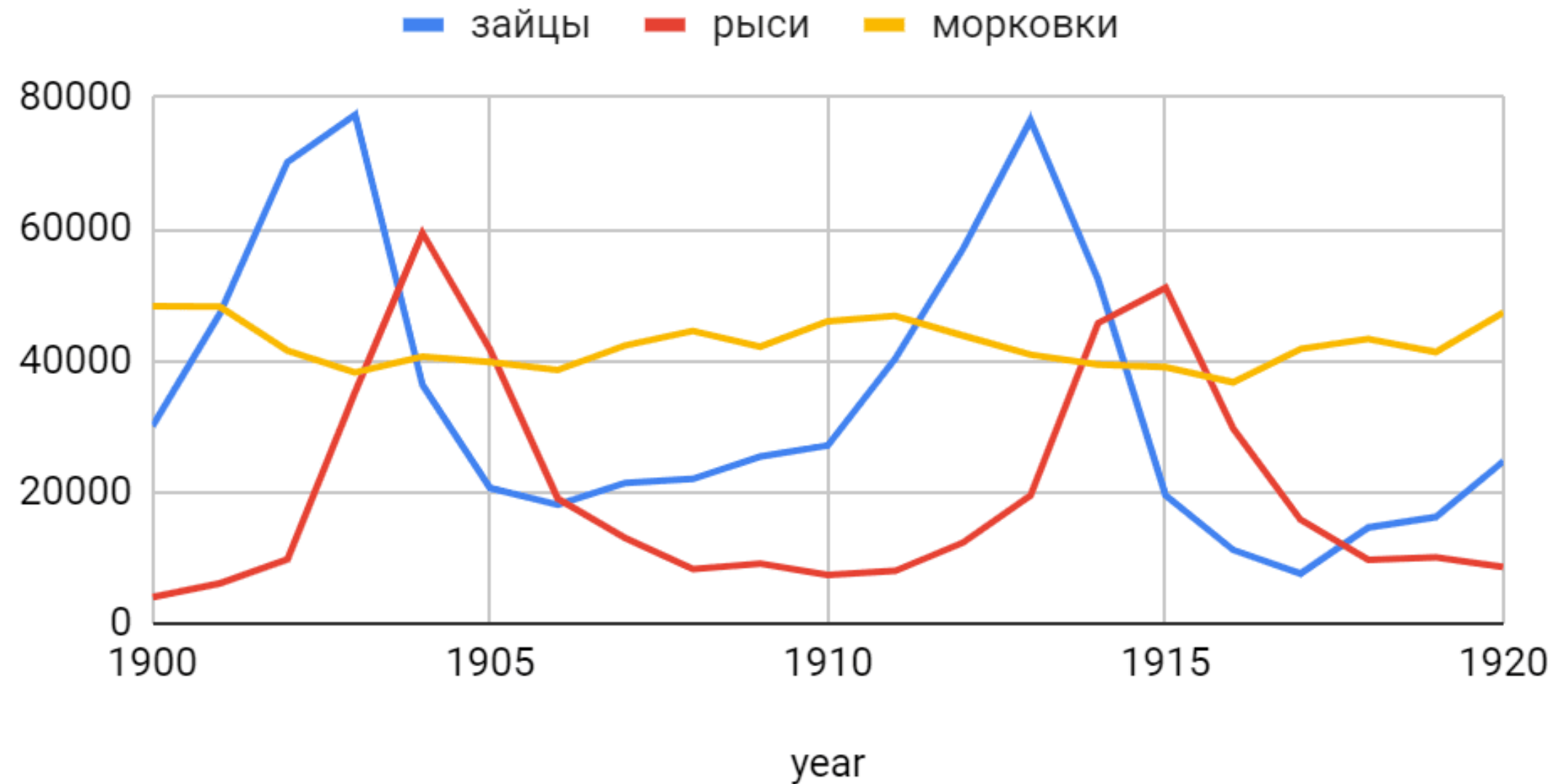
`.std() **2 == .var()`

*(простыми словами,
насколько данные
далеко "разбросаны"
относительно
среднего)*



Задача из НЭ, А2 / А4

Зайцы, рыси, морковки

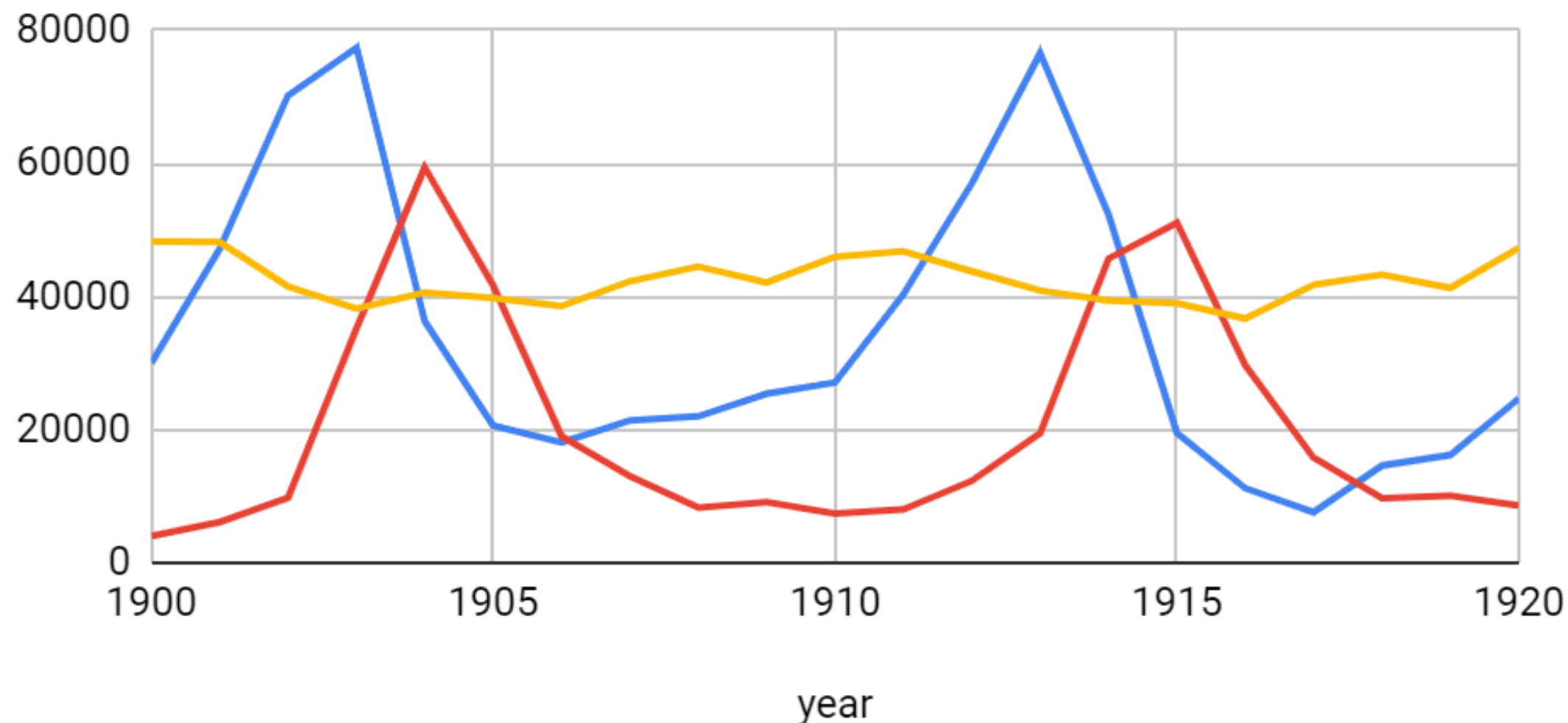


1. В 1910 году популяция зайцев была больше популяции рысей
2. В популяции морковок самая маленькая разница между ее минимальным и максимальным значением
3. Максимум за все время принадлежит популяции рысей
4. В один из годов все три популяции были одинакового размера

Задача из НЭ, А2 / А4

Зайцы, рыси, морковки

зайцы рыси морковки



1. В 1910 году популяция заповней была больше популяции рысей
2. В популяции морковок самая маленькая разница между ее минимальным и максимальным значением
3. Максимум за все время принадлежит популяции рысей
4. В один из годов все три популяции были одинакового размера

Задача из НЭ, А2 / А4

Исследователь Иван собрал данные по численности трёх популяций кальмаров в некотором регионе за 12 лет. Эти данные приведены в таблице ниже.

	Популяция 1	Популяция 2	Популяция 3
Среднее	2002	5401	3048
Медиана	2005	3001	4000
Стандартное отклонение	30	402	350

- 1. В какой-то год количество кальмаров из популяции 3 было аномально высоким
- 2. Если рассматривать промежуток в 9 лет, то среднее и медиана численности кальмаров в популяции 1 обязательно совпадут
- 3. Наибольший разброс имеют наблюдения из популяции 1
- 4. В какой-то год количество кальмаров из популяции 2 было аномально высоким

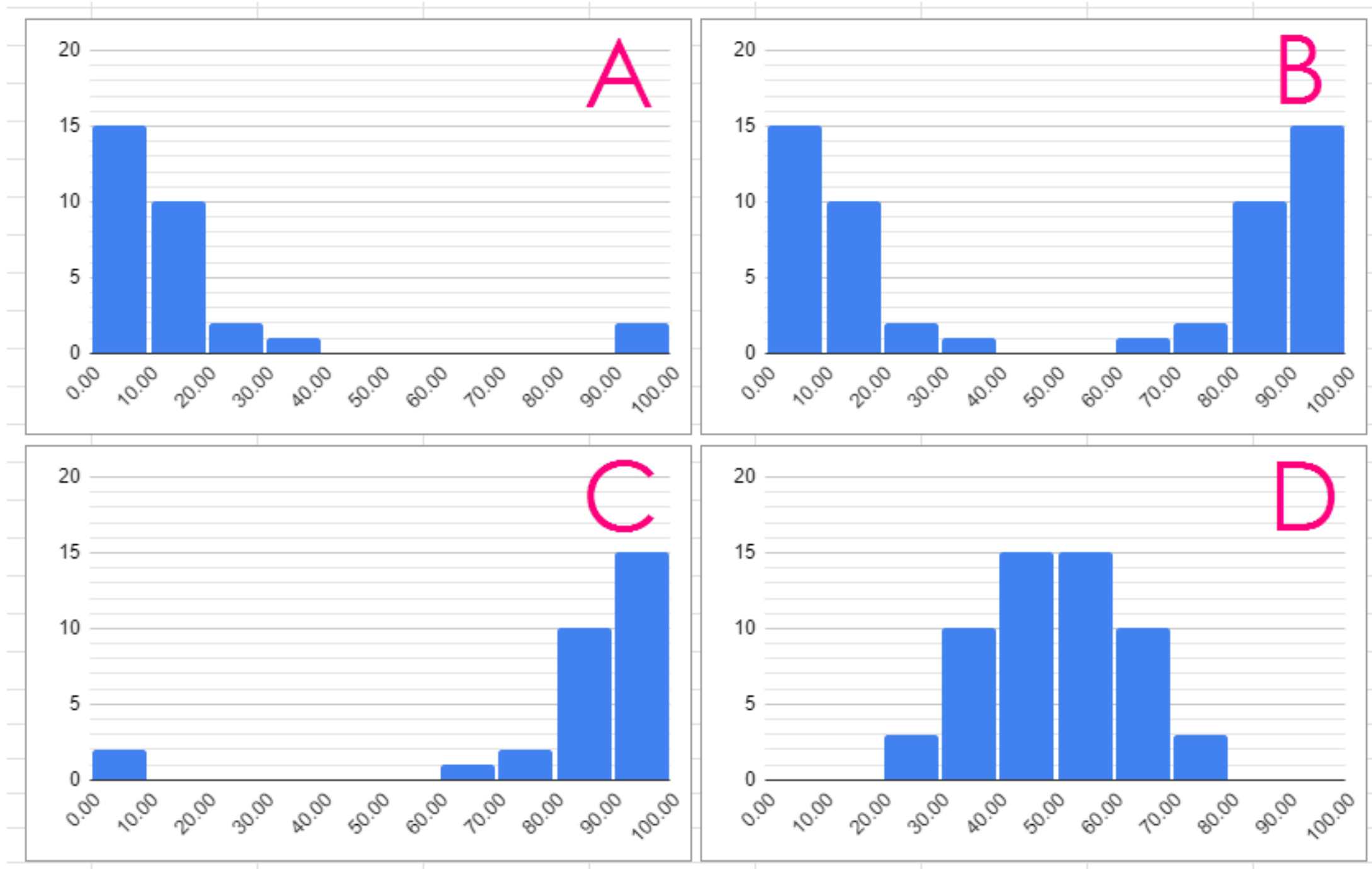
Задача из НЭ, А2 / А4

Исследователь Иван собрал данные по численности трёх популяций кальмаров в некотором регионе за 12 лет. Эти данные приведены в таблице ниже.

	Популяция 1	Популяция 2	Популяция 3
Среднее	2002	5401	3048
Медиана	2005	3001	4000
Стандартное отклонение	30	402	350

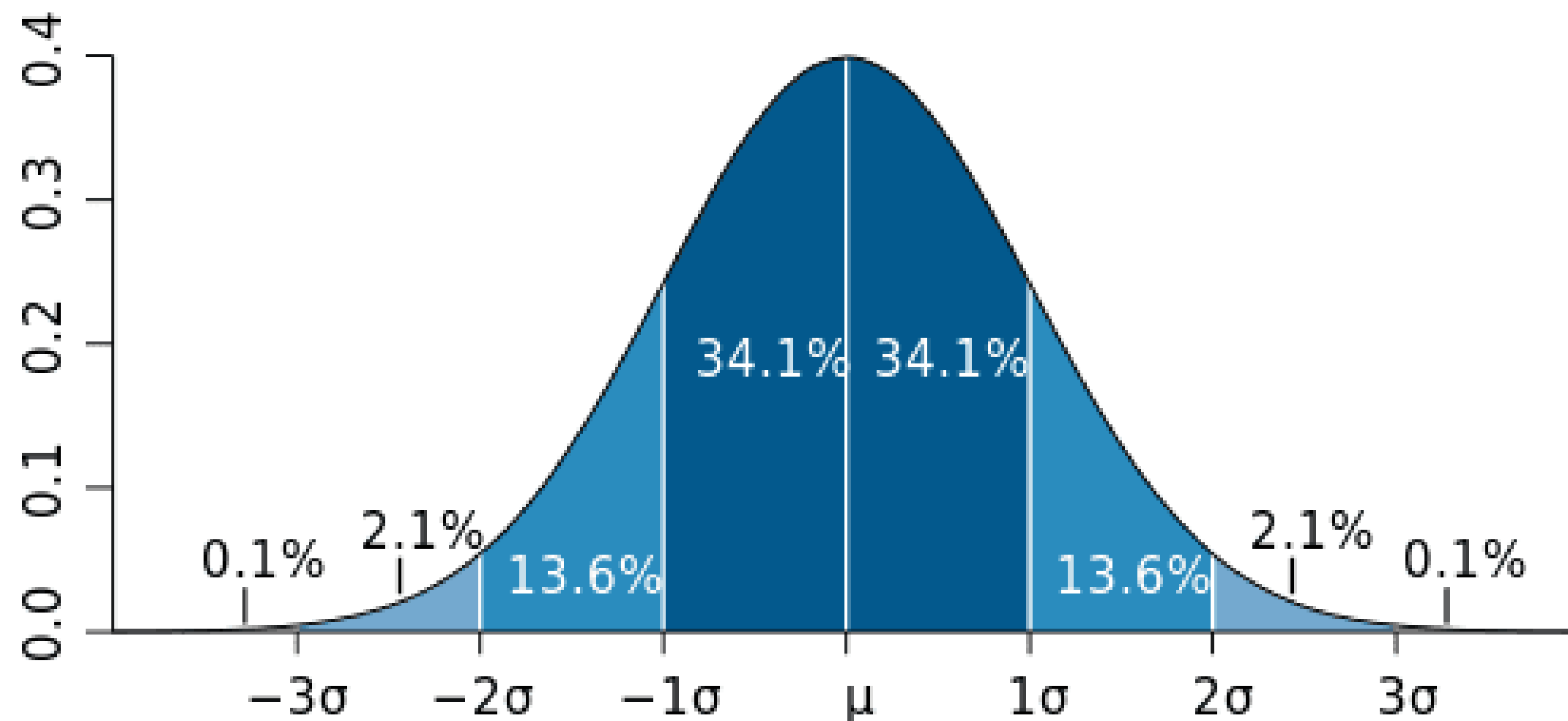
- 1. В какой-то год количество кальмаров из популяции 3 было аномально высоким
- 2. Если рассматривать промежуток в 9 лет, то среднее и медиана численности кальмаров в популяции 1 обязательно совпадут
- 3. Наибольший разброс имеют наблюдения из популяции 1
- 4. В какой-то год количество кальмаров из популяции 2 было аномально высоким

Задача из НЭ, А2 / А4

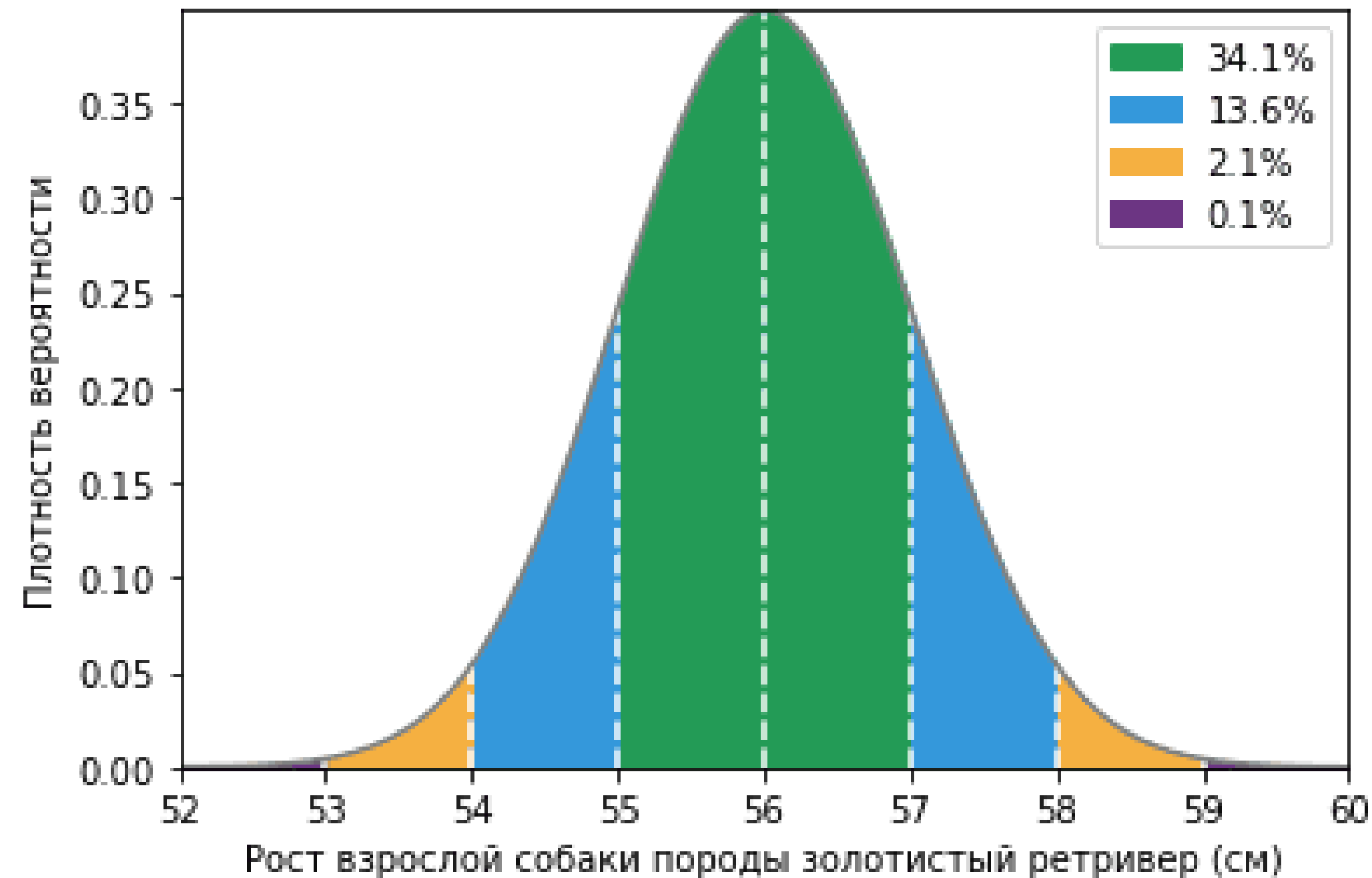


Среди четырех величин, для которых построены гистограммы ниже, выберите величину с **наименьшим** стандартным отклонением.

Закон нормального распределения + стандартное отклонение

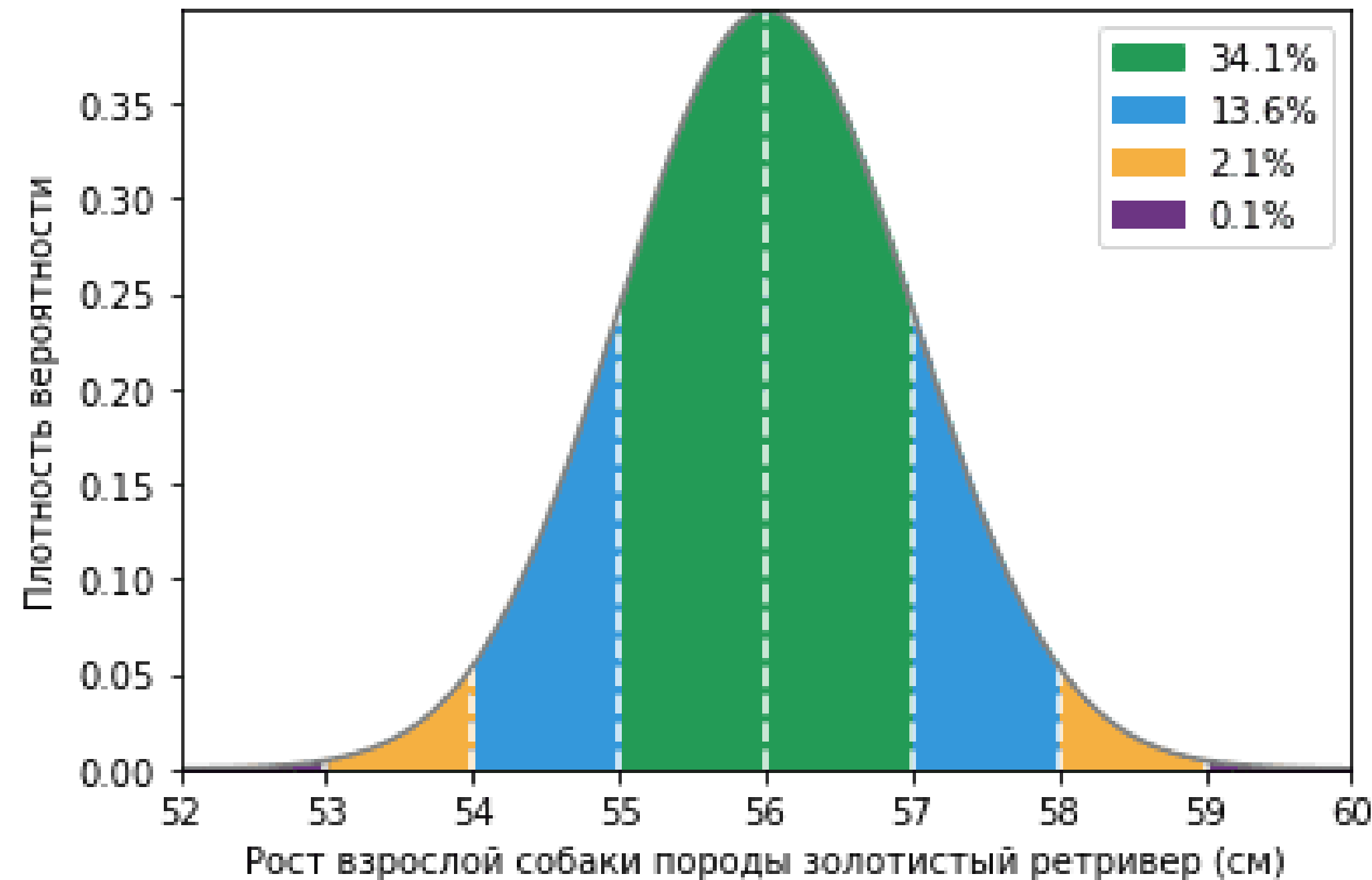


Задача из НЭ, А8



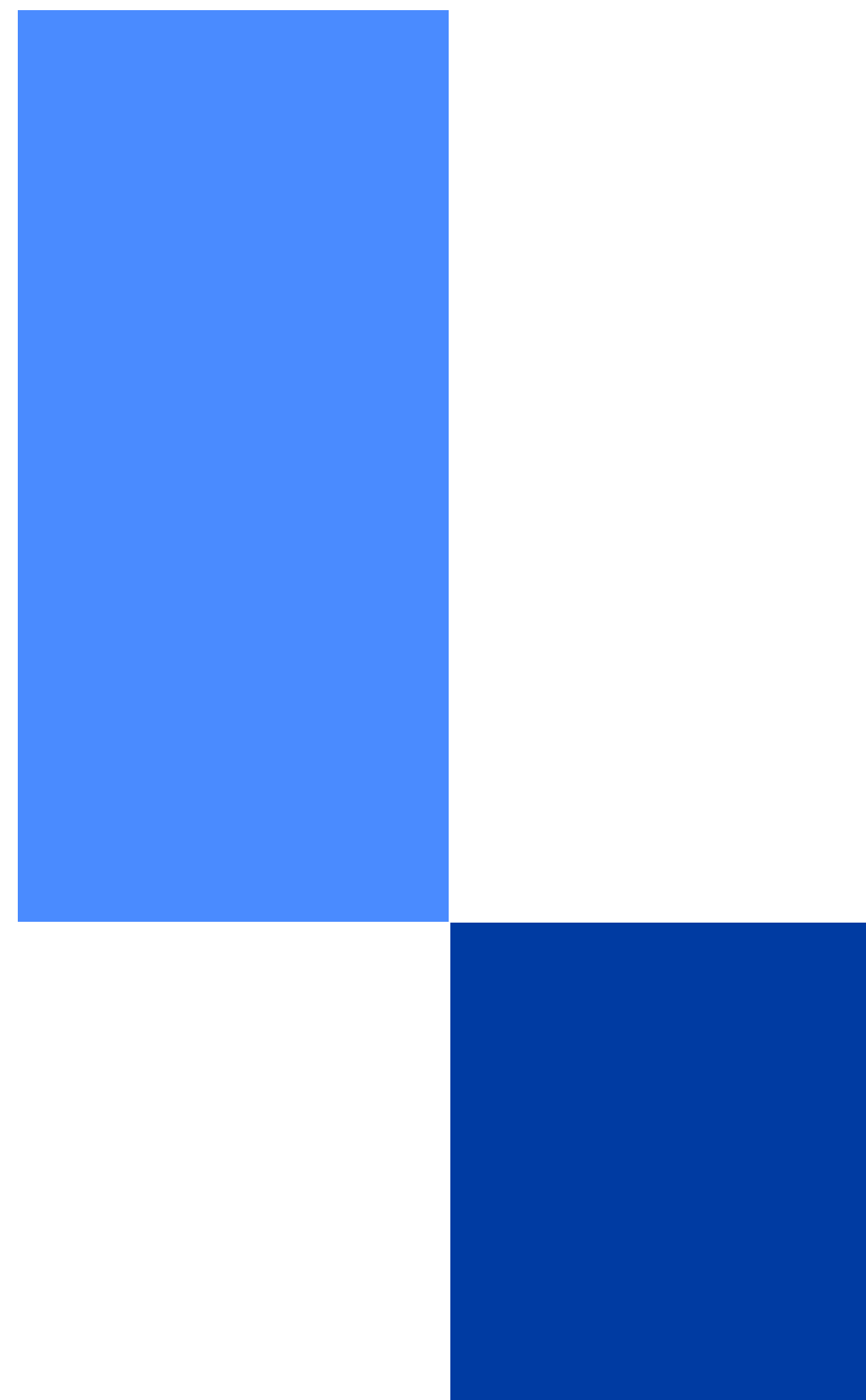
1. Примерно 68.2% взрослых собак породы золотистый ретривер имеют рост от 55 до 57 см
2. Примерно 0.1% взрослых собак породы золотистый ретривер имеют рост менее 53 см
3. 2.1% взрослых собак породы золотистый ретривер имеют рост более 58 см
4. Медиана роста взрослой собаки породы золотистый ретривер равна 59 см

Задача из НЭ, А8



1. Примерно 68.2% взрослых собак породы золотистый ретривер имеют рост от 55 до 57 см
2. Примерно 0.1% взрослых собак породы золотистый ретривер имеют рост менее 53 см
3. 2.1% взрослых собак породы золотистый ретривер имеют рост более 58 см
4. Медиана роста взрослой собаки породы золотистый ретривер равна 59 см

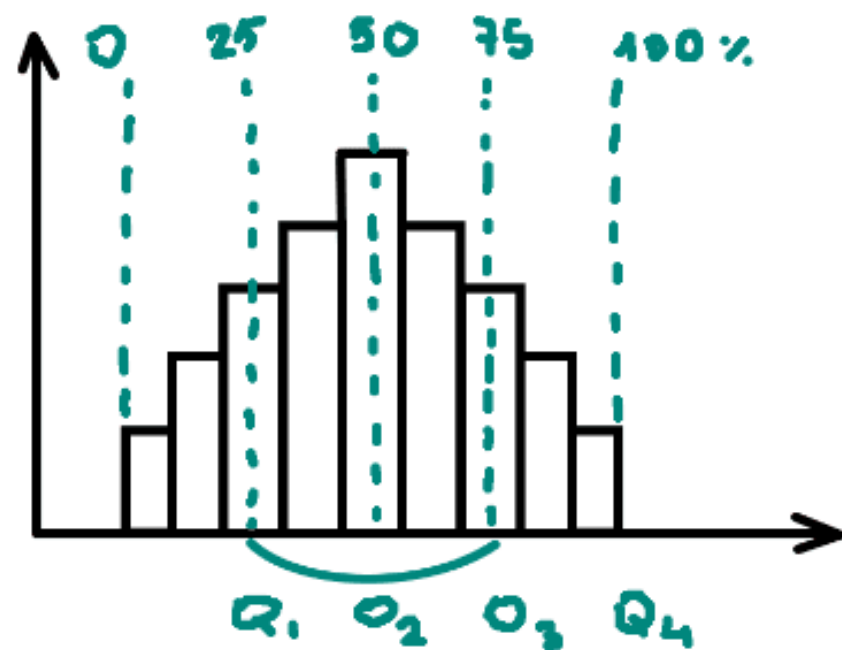
**от гистограмм
к ящикам с
усами**



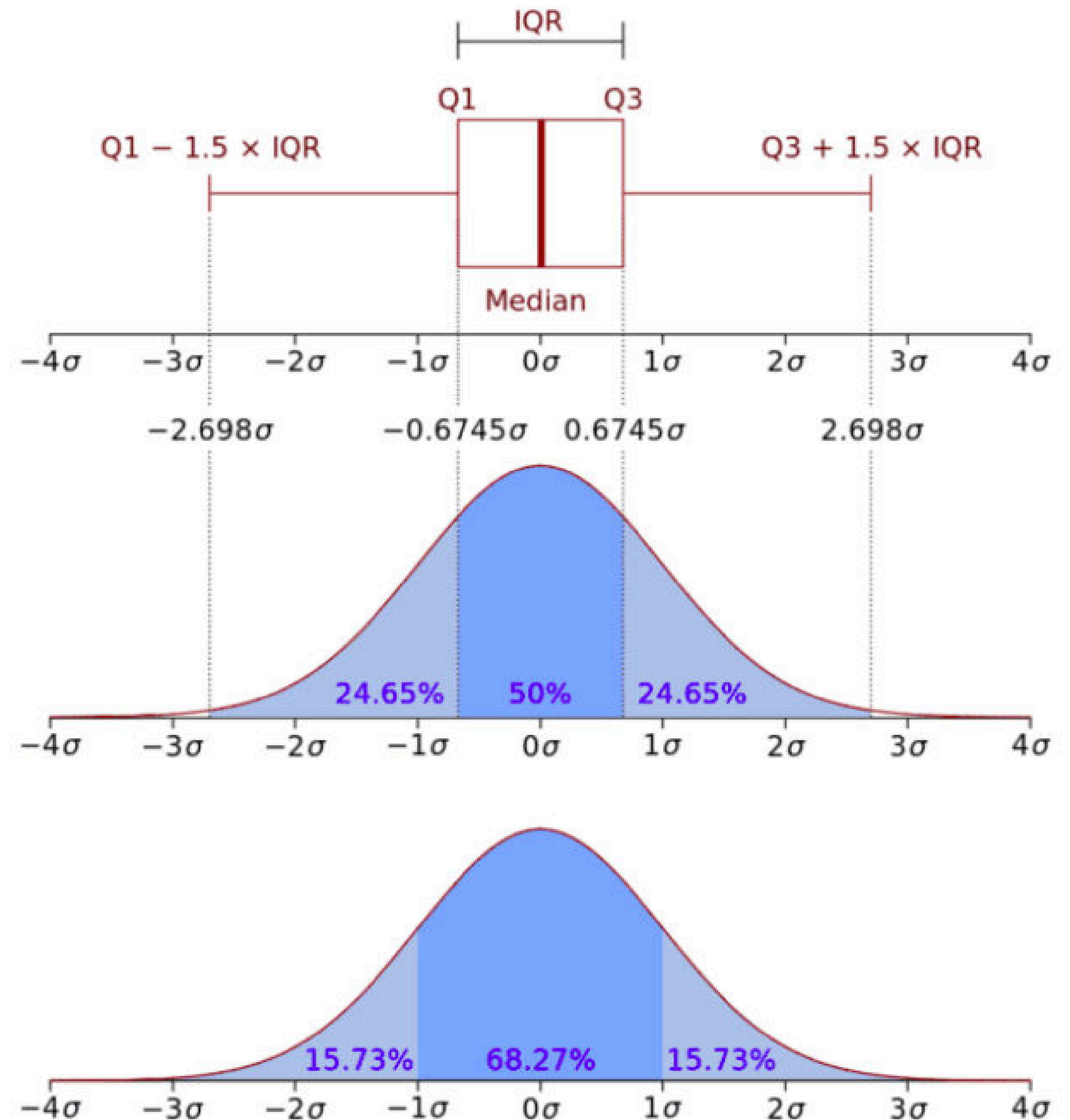
Термины:

- квартиль
- межквартильный размах (интервал)

не путаем с просто размахом (макс - мин)

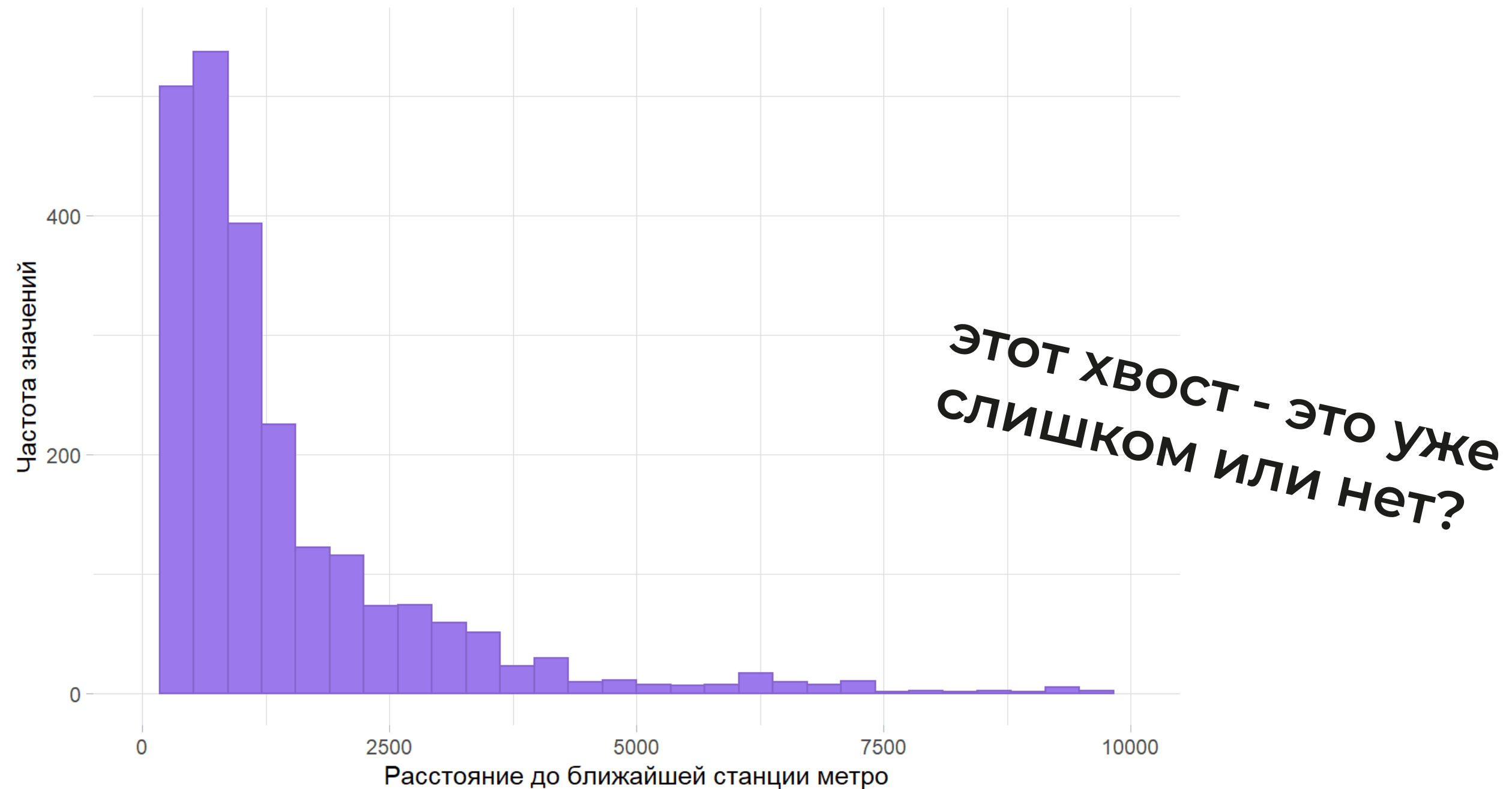


зачем??

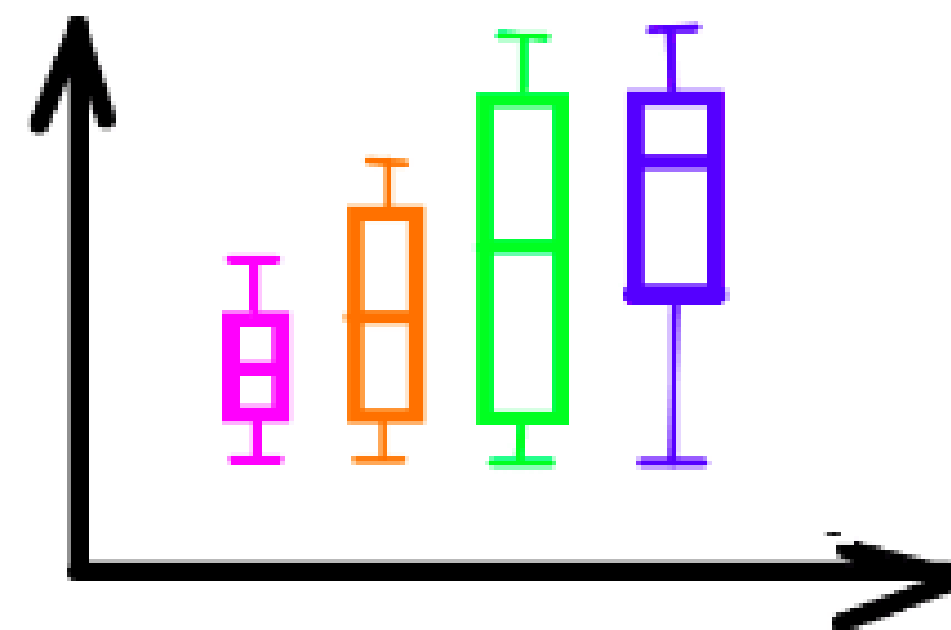
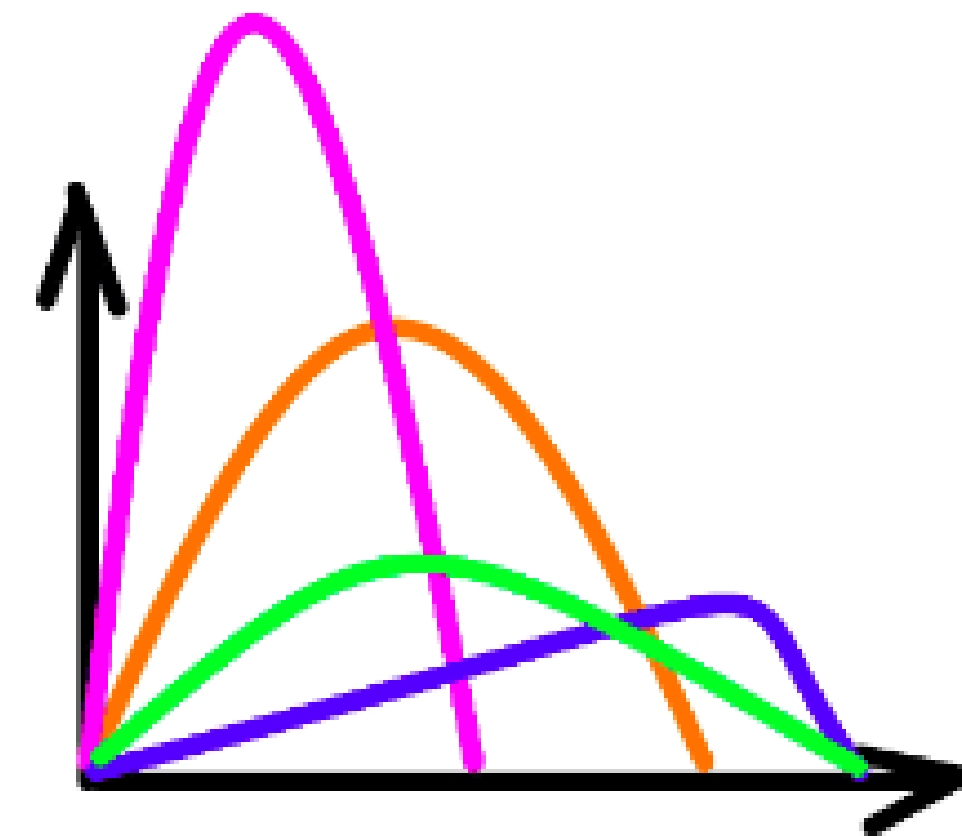
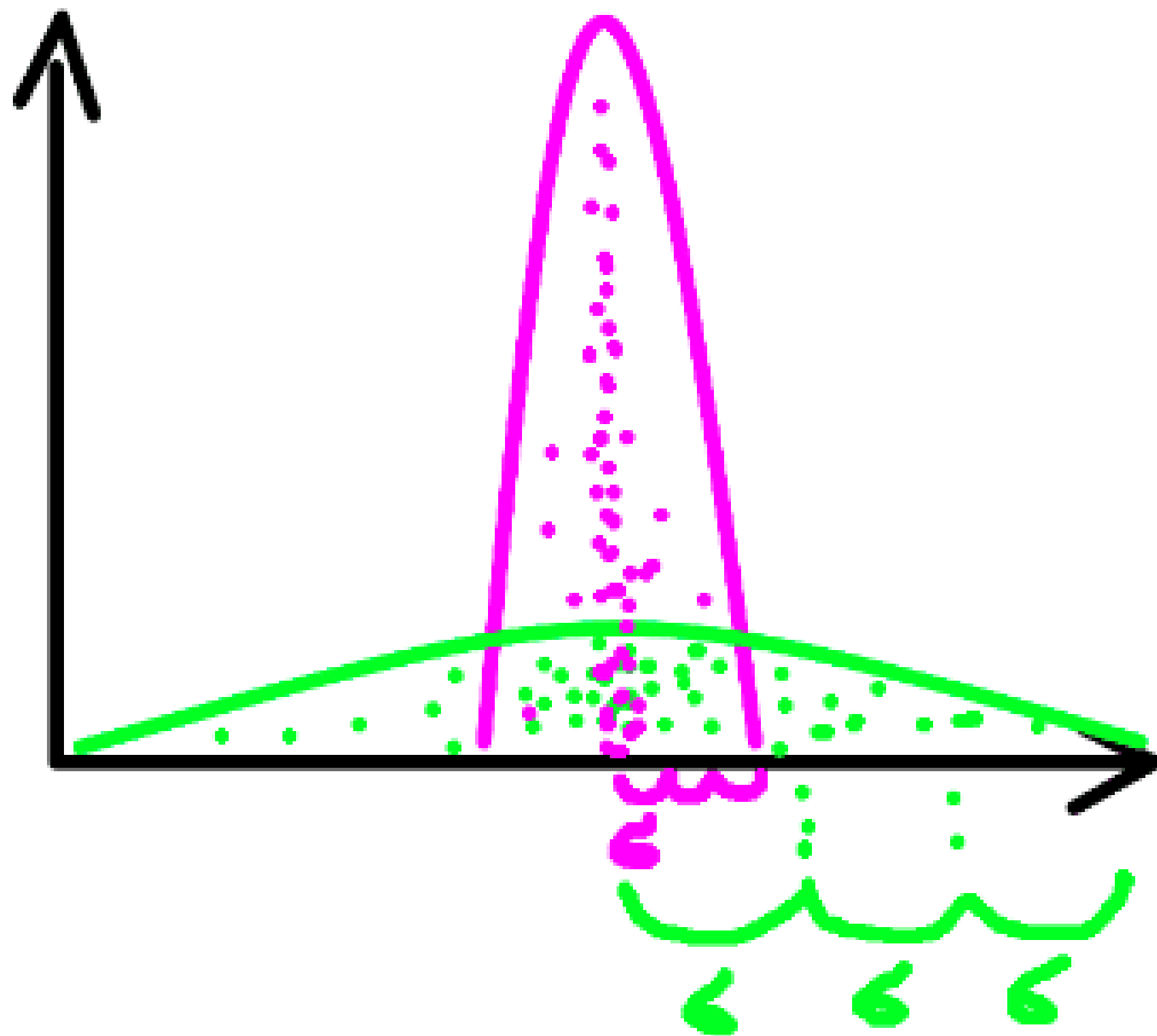


Термины:

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

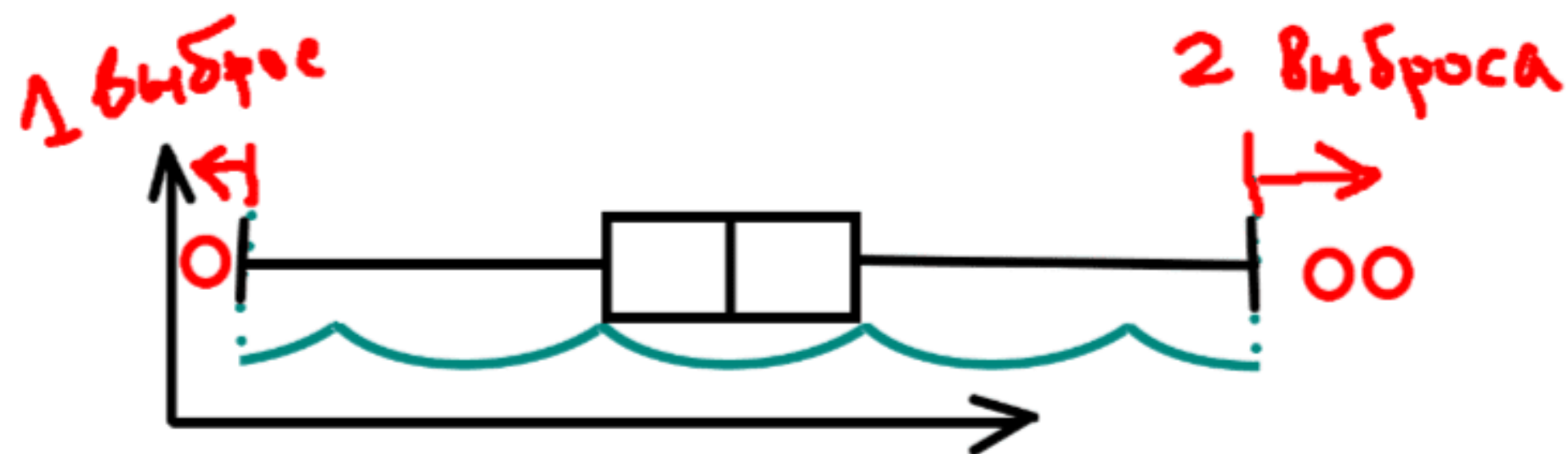
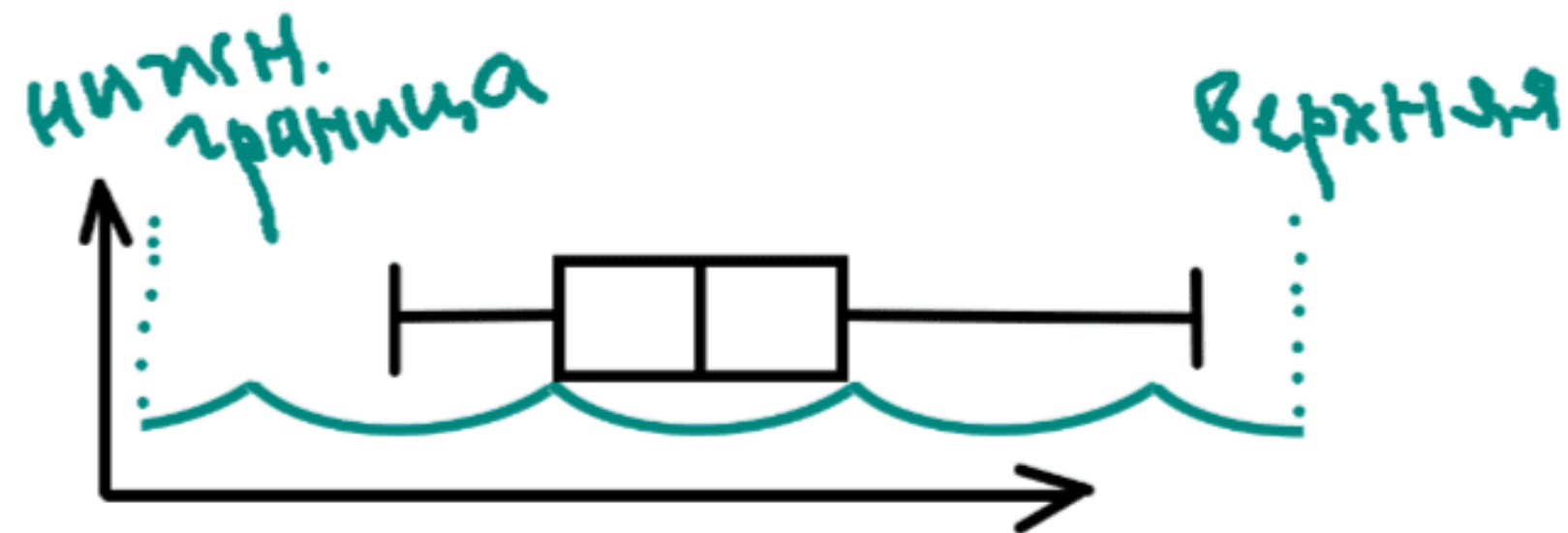


Связь гистограмм и ящиков с усами:



Термины:

- обычно выброс: межквартильный размах $(Q3 - Q1) * 1,5$
- но в НЭ компромисс, чтобы быстрее считалось: $n * .std()$



Задача из НЭ, А6

Студент Михаил решил записывать, сколько времени (в часах) он тратит на выполнение домашних заданий в месяц в течение учебного года. Выберите тип графика, который **меньше всего подойдёт** Михаилу для наглядного изображения динамики количества учебных часов по месяцам.

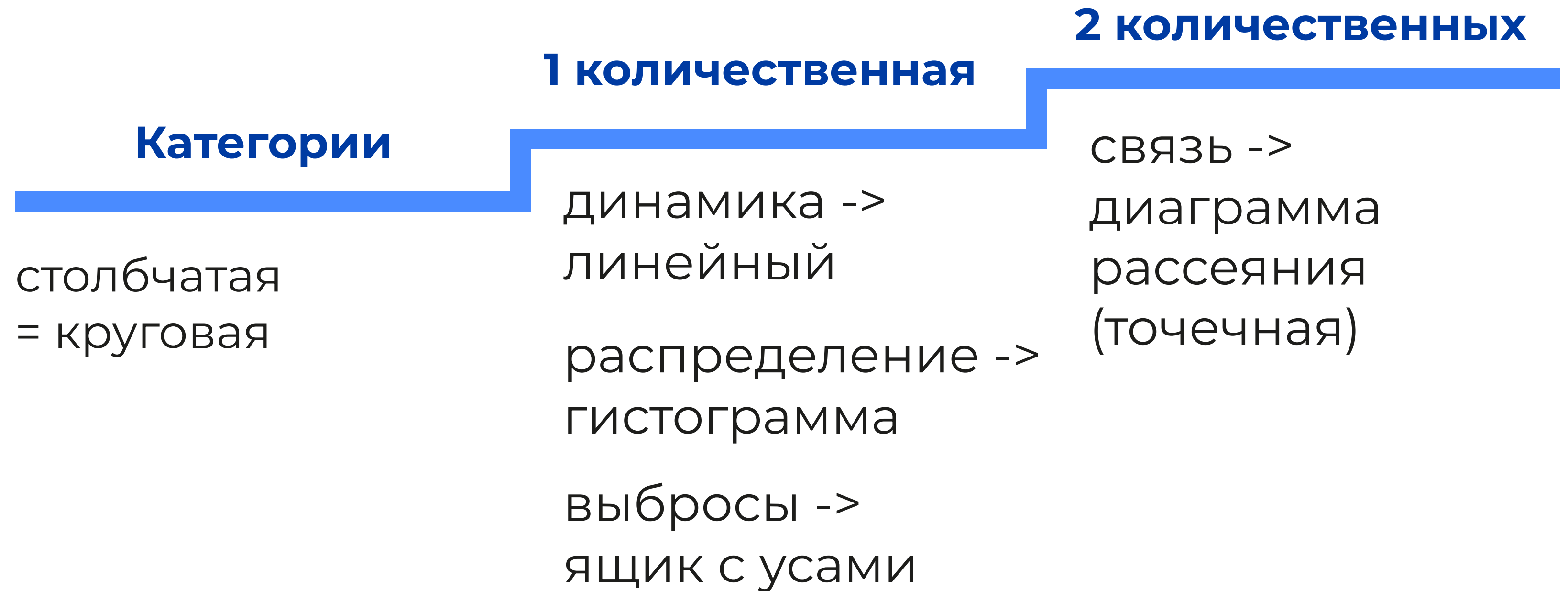
1. Столбчатая диаграмма
2. Ящик с усами
3. Линейный график
4. Круговая диаграмма

Задача из НЭ, А6

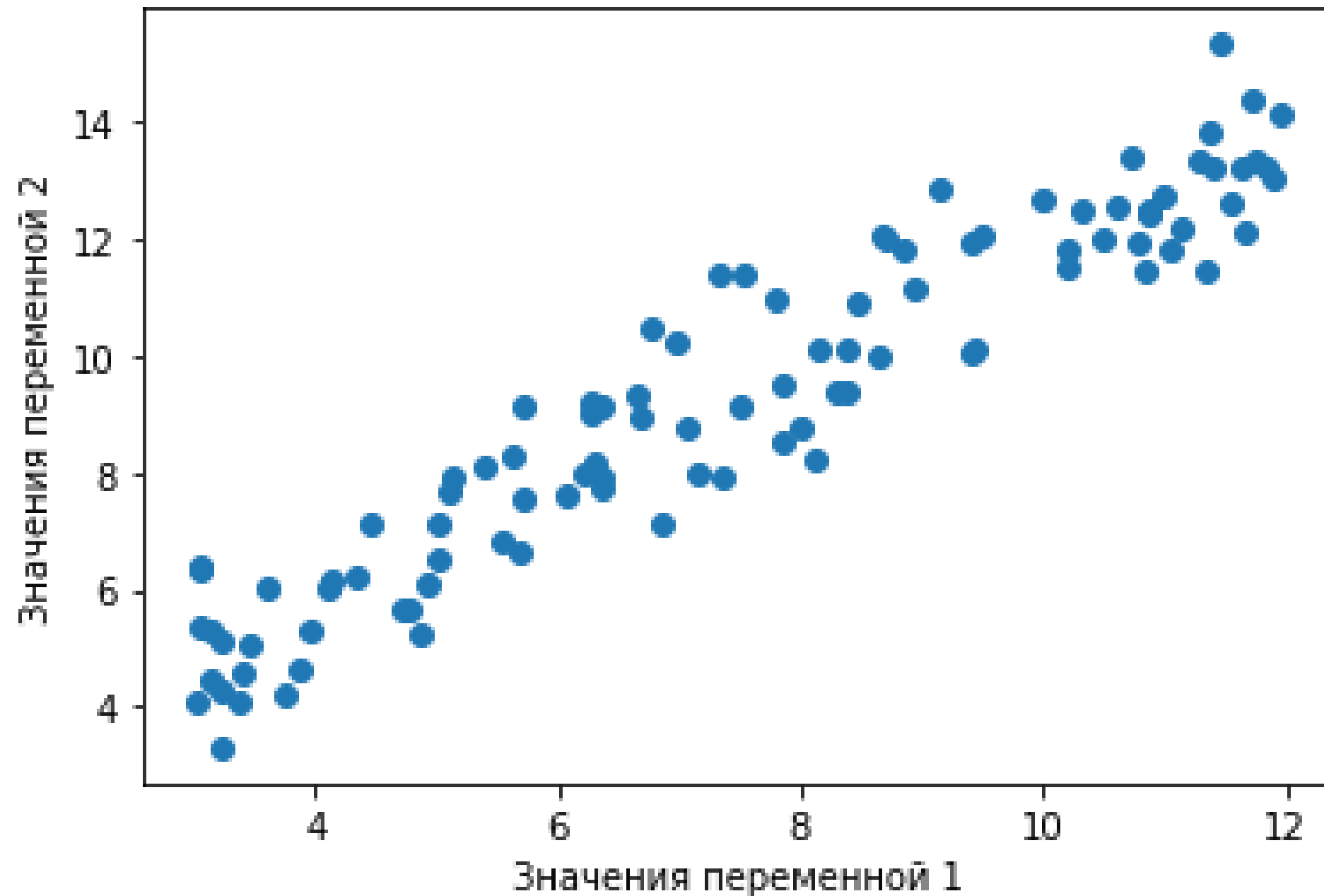
Выберите тип визуализации из предложенных, с помощью которого можно **наиболее корректно** визуализировать распределение выживших и погибших пассажиров «Титаника».

1. Линейный график (line graph)
2. Ящик-с-усами (box plot)
3. Столбчатая диаграмма (bar chart)
4. График рассеяния (scatter plot)

Наивный гайд на А6

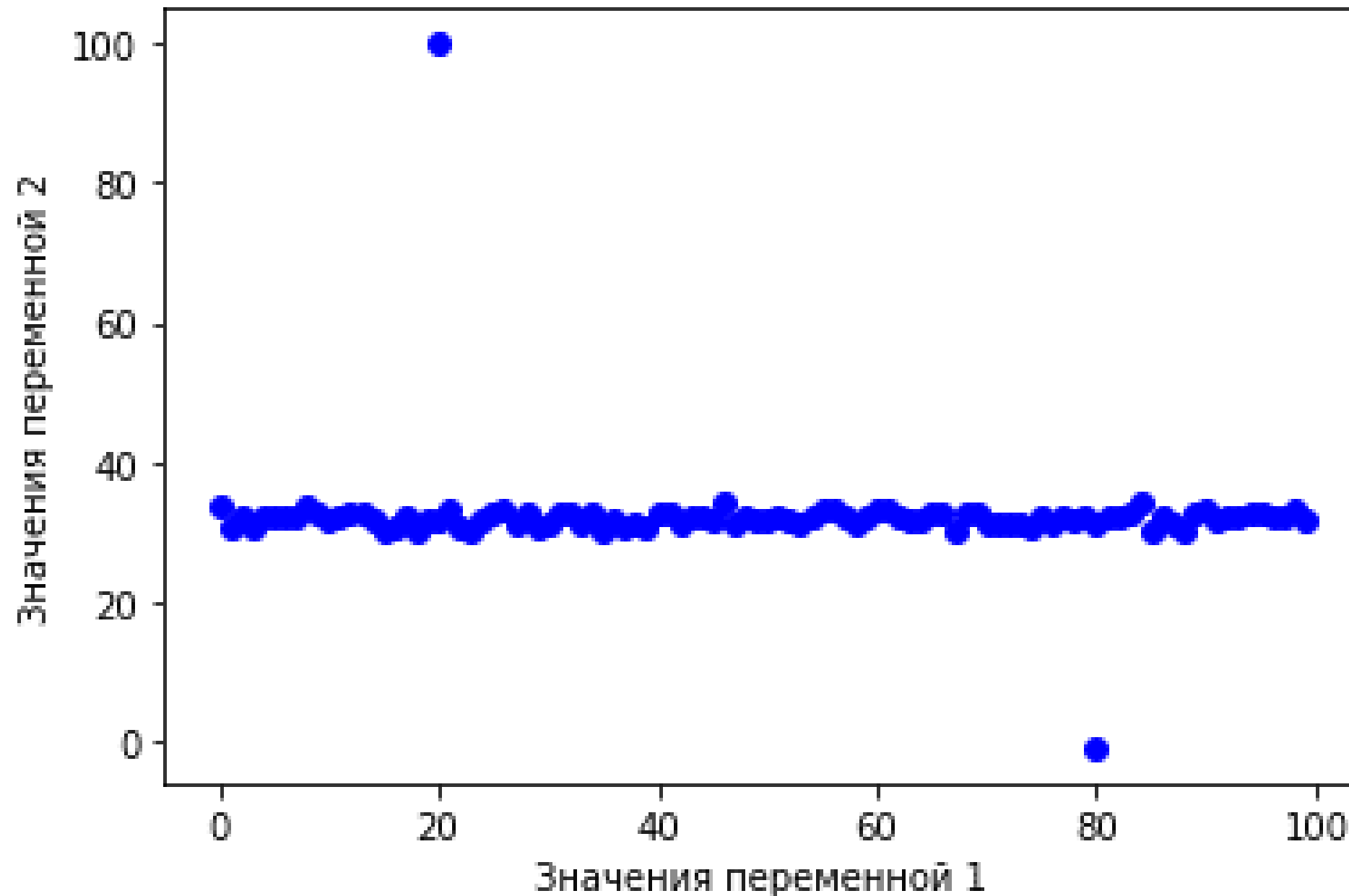


Задача из НЭ, А6 (более новые)



1. Выбросы оказывают сильное влияние на среднее значение переменной 2
- 2.
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.
3. В выборке имеется как минимум 1 выброс
4. Выбросы оказывают большое влияние на медиану переменной 2

Все на питоне:

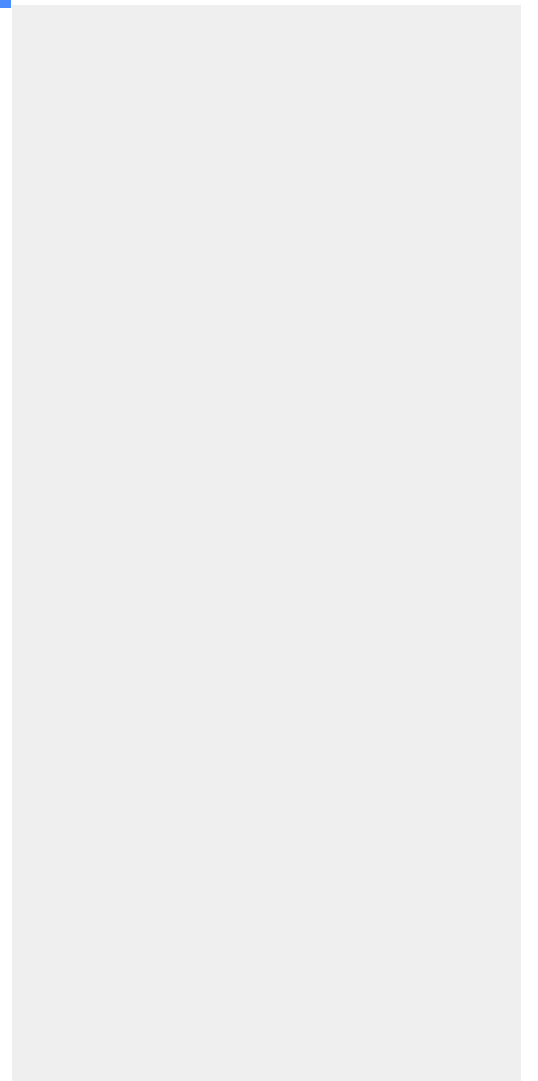
df['столбец']

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

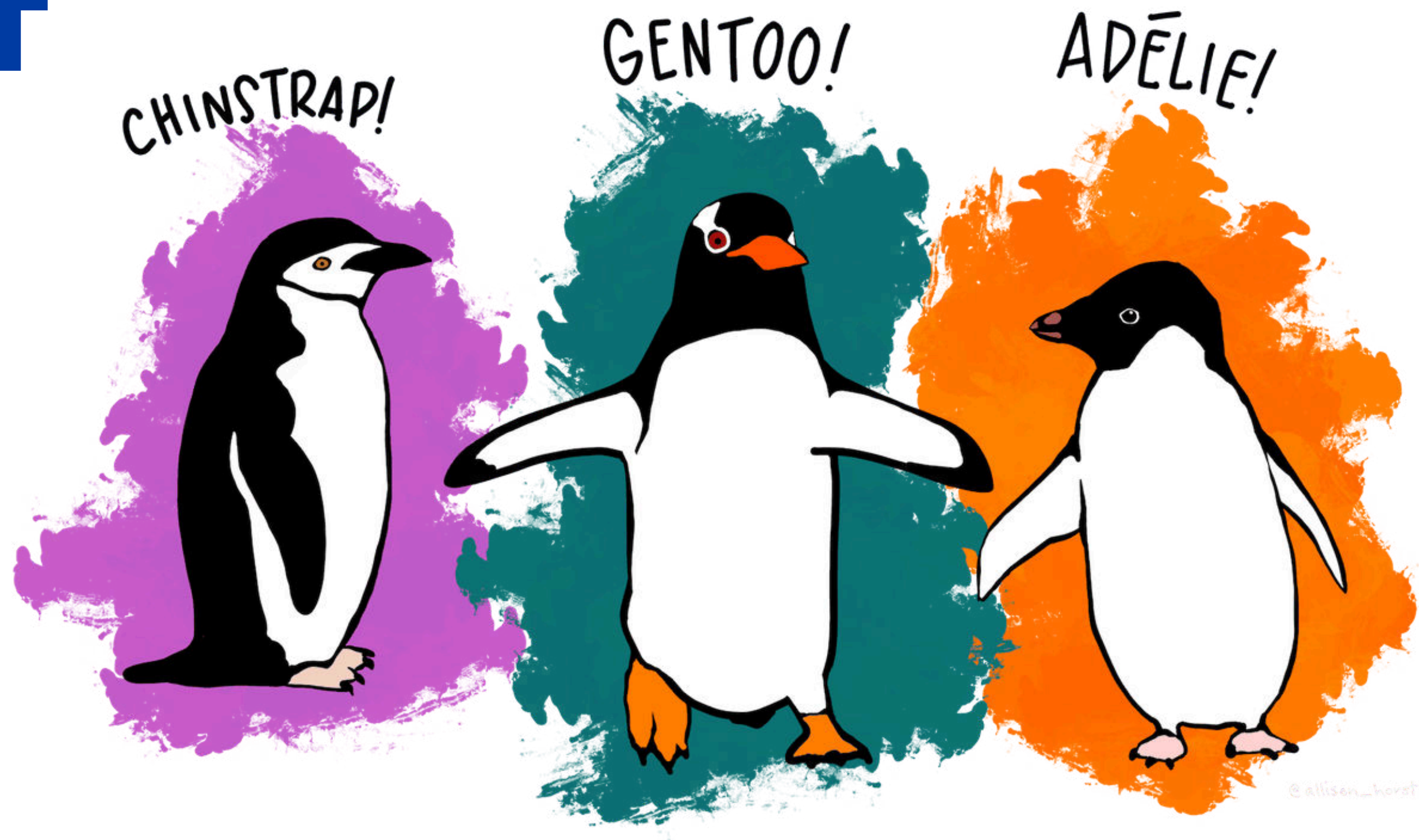
**Только для
генеральной
совокупности!**

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

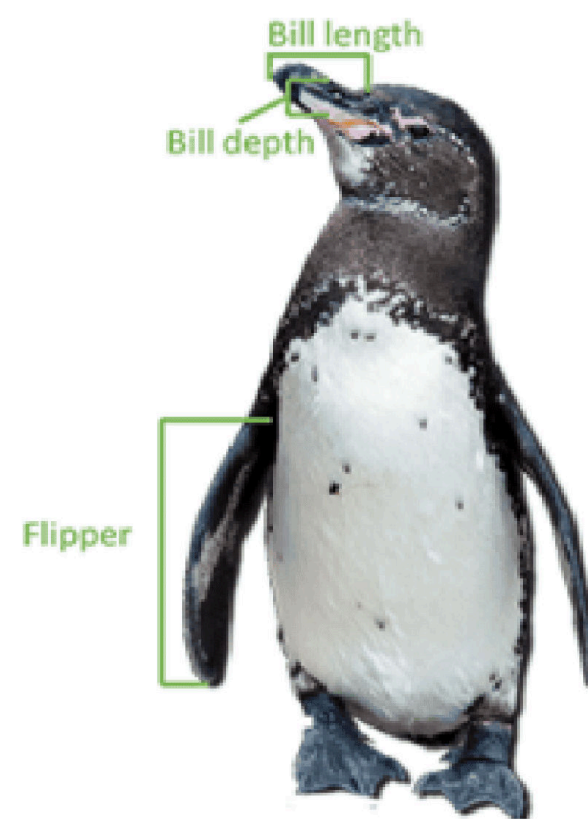
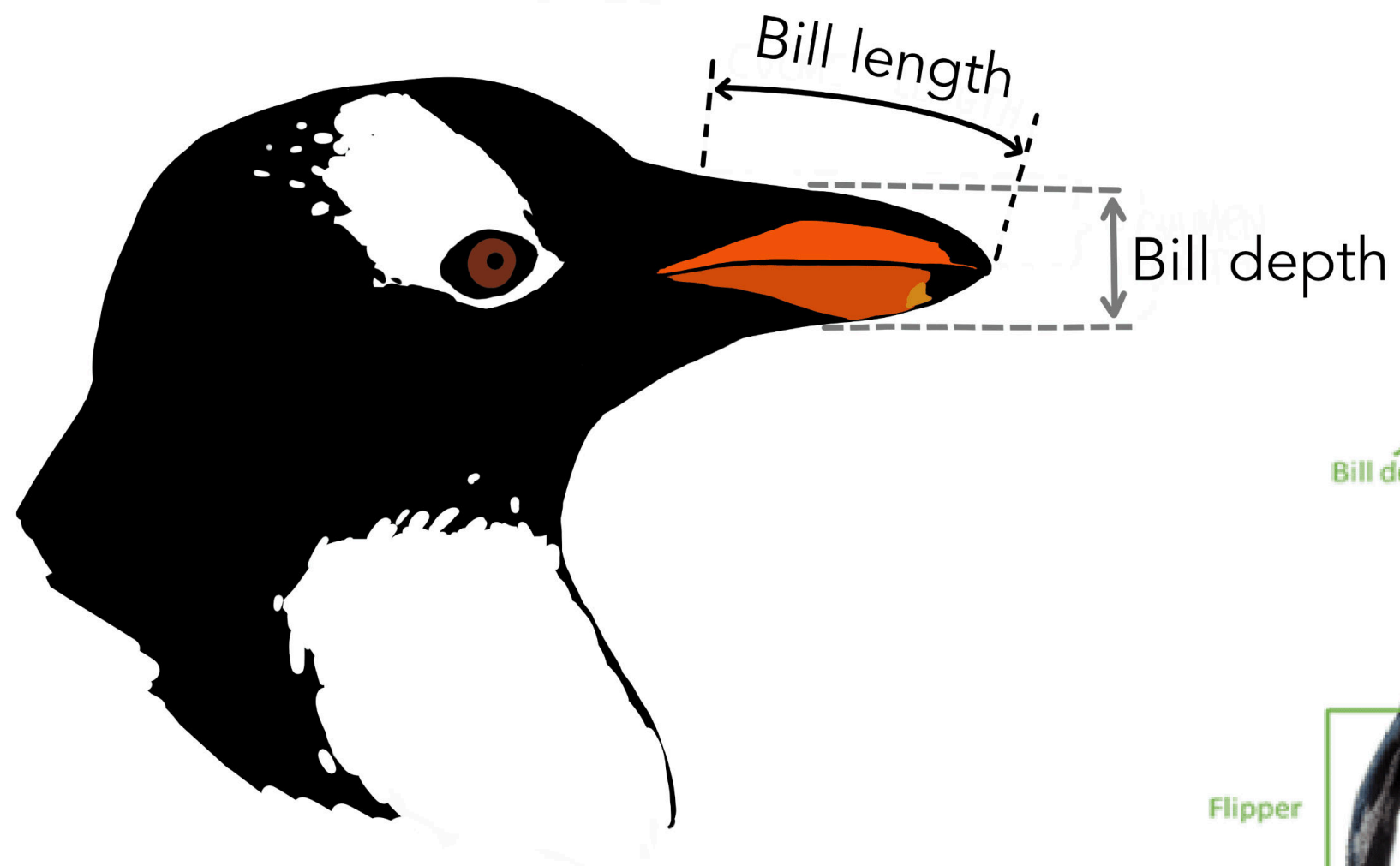
*Это редкий случай, в
условии задачи напишут
обязательно, "если вы
имеете дело с
генеральной
совокупностью"....*



"игрушечный" датасет



"игрушечный" датасет

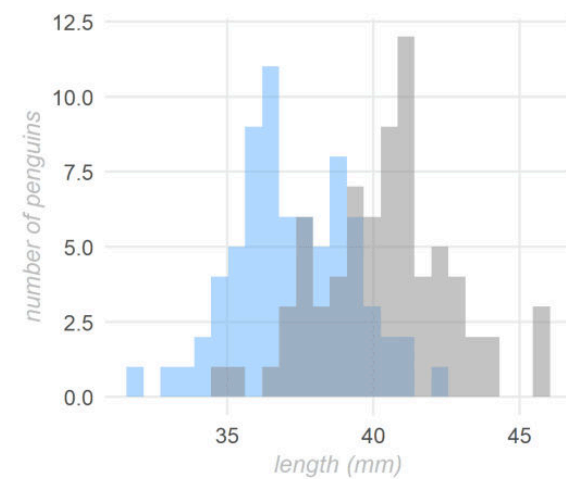
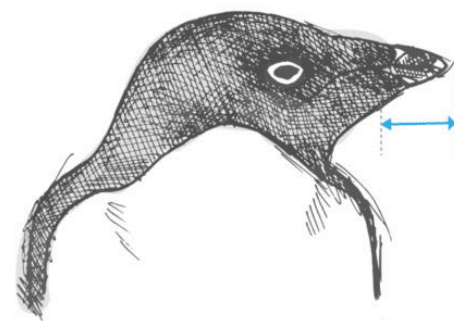


"игрушечный" датасет

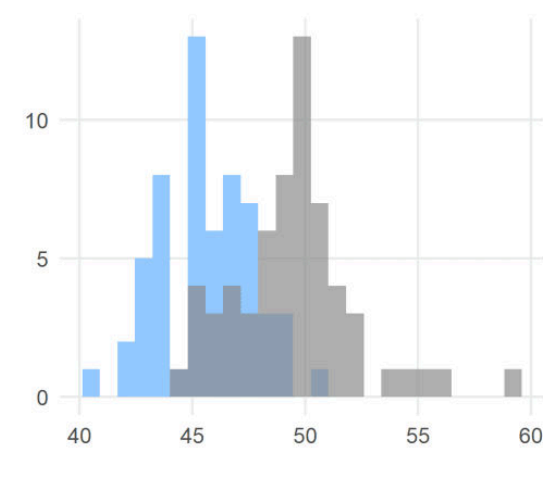
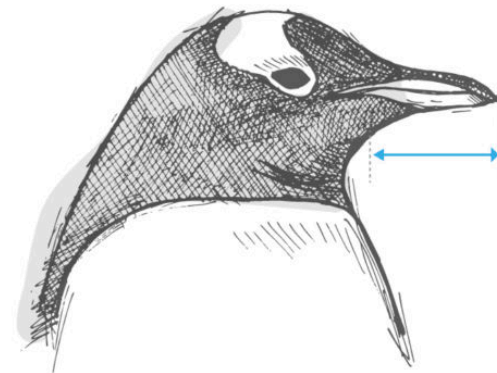
Palmer Penguins Bill Length

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

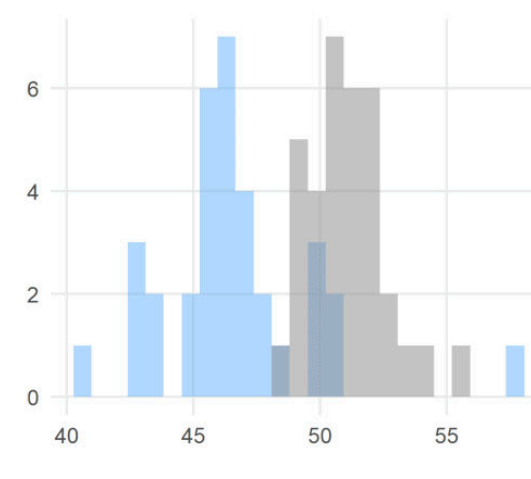
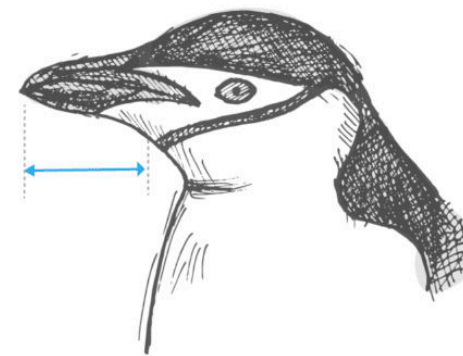
ADELIE



GENTOO





CHINSTRAP



female male

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`)

