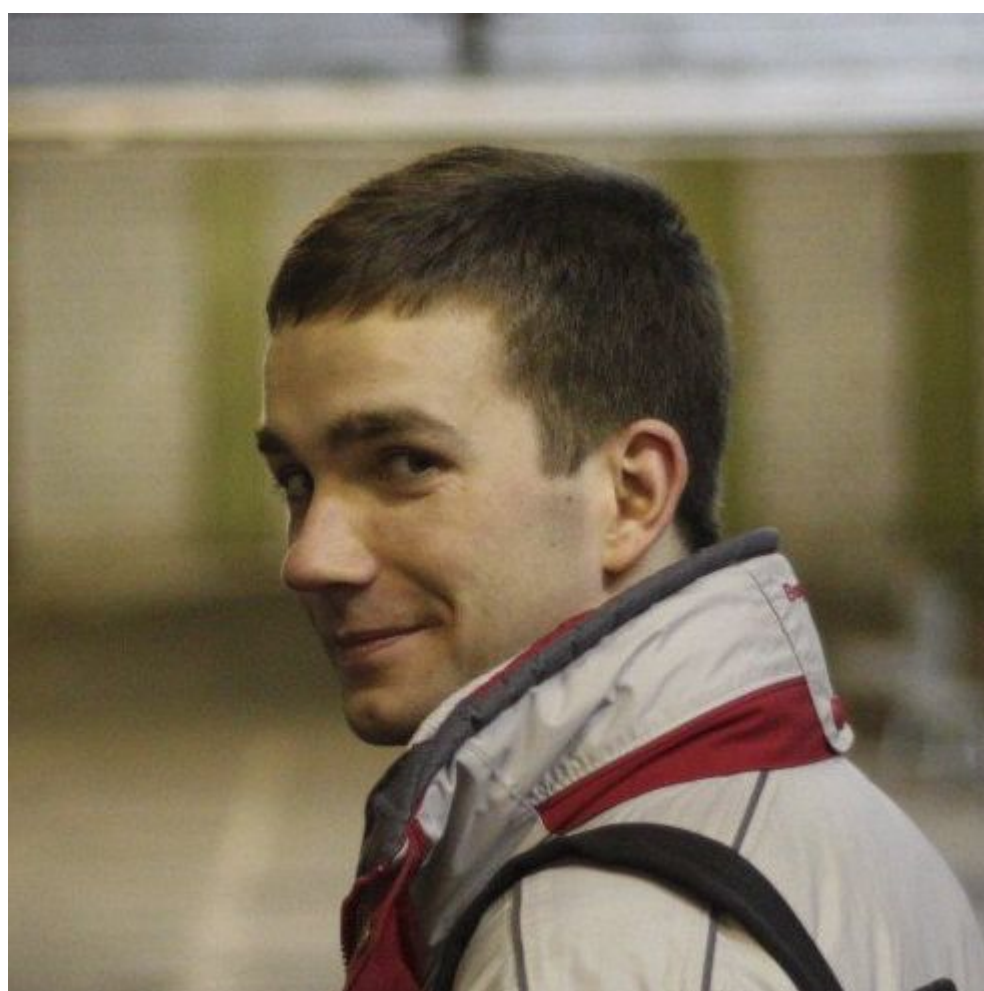


ЗАНЯТИЕ 0.3

ОБЗОР МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ



КОНСТАНТИН БАШЕВОЙ

Старший
аналитик



kbashevoy@gmail.co
m



/konstantin.bashevo
y

ЦЕЛИ ЗАНЯТИЯ

В КОНЦЕ ЗАНЯТИЯ ВЫ:

- будете знать основные распределения случайных величин
- познакомитесь с понятием корреляции и ее смыслом в некоторых задачах
- сможете проверять гипотезы.

О ЧЁМ ПОГОВОРИМ И ЧТО
СДЕЛАЕМ

-
1. Понятия теории вероятностей: основы;
 2. Нормальное распределение и правило трех сигм: углубление в теорию;
 3. Корреляция и ее применение в различных задачах;
 4. Формулировка и проверка гипотез: теория и практика

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ

Множество всех объектов, относительно которых хотим
сделать выводы

Общественные опросы ВЦИОМ

Генеральная совокупность – все жители РФ

Выборка – респонденты, которые согласились ответить на
вопросы

Большинство россиян предпочли официальный брак гражданскому

Добавить в «Мою Ленту» ?



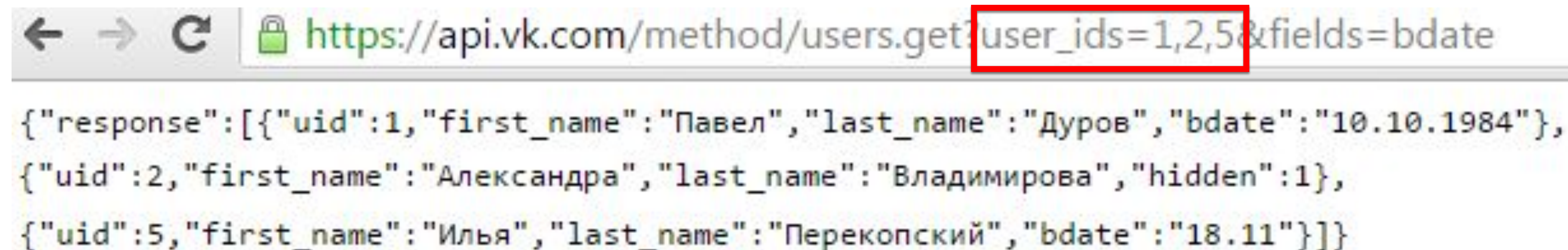
Фото: Антон Белицкий / РИА Новости

Большинство россиян (78 процентов) предпочитают официальный брак гражданскому. Об этом свидетельствуют результаты опроса Всероссийского центра изучения общественного мнения (ВЦИОМ), **опубликованные** на его сайте в понедельник, 28 августа.

За то, что официальный брак — наиболее естественная форма взаимоотношений, высказался даже 51 процент тех, у кого нет штампа в паспорте.

ПРОСТАЯ СЛУЧАЙНАЯ ВЫБОРКА

Оцениваем средний возраст пользователей ВКонтакте,
просто перебирая ID случайным образом



The screenshot shows a web browser's address bar with the URL `https://api.vk.com/method/users.get?user_ids=1,2,5&fields=bdate`. The parameter `user_ids=1,2,5` is highlighted with a red box. Below the address bar, the JSON response is displayed:

```
{"response": [{"uid": 1, "first_name": "Павел", "last_name": "Дуров", "bdate": "10.10.1984"}, {"uid": 2, "first_name": "Александра", "last_name": "Владимирова", "hidden": 1}, {"uid": 5, "first_name": "Илья", "last_name": "Перекопский", "bdate": "18.11"}]}
```

СТРАТИФИЦИРОВАННАЯ ВЫБОРКА

Разбиваем генеральную совокупность на несколько групп.
Из каждой группы набираем выборку для исследования

Сэмплирование в системах аналитики, АБ-тестирование

Cookie ID	Page	Timestamp
Htr5BSS1	http://yandex.ru/	1439704677.13
Bm76BQr	http://lenta.ru/sport	1439717451.08
AQAAABJb	http://livejournal.com/1075498.html	1439744394.56

Не можем отбирать строки случайным образом. Для подсчета аудиторных показателей необходимо группировать по посещениям или уникальным посетителям

ВЕРОЯТНОСТЬ



ВЕРОЯТНОСТЬ СОБЫТИЯ

$$P(A) = \lim_{N \rightarrow \infty} \frac{m}{N}$$

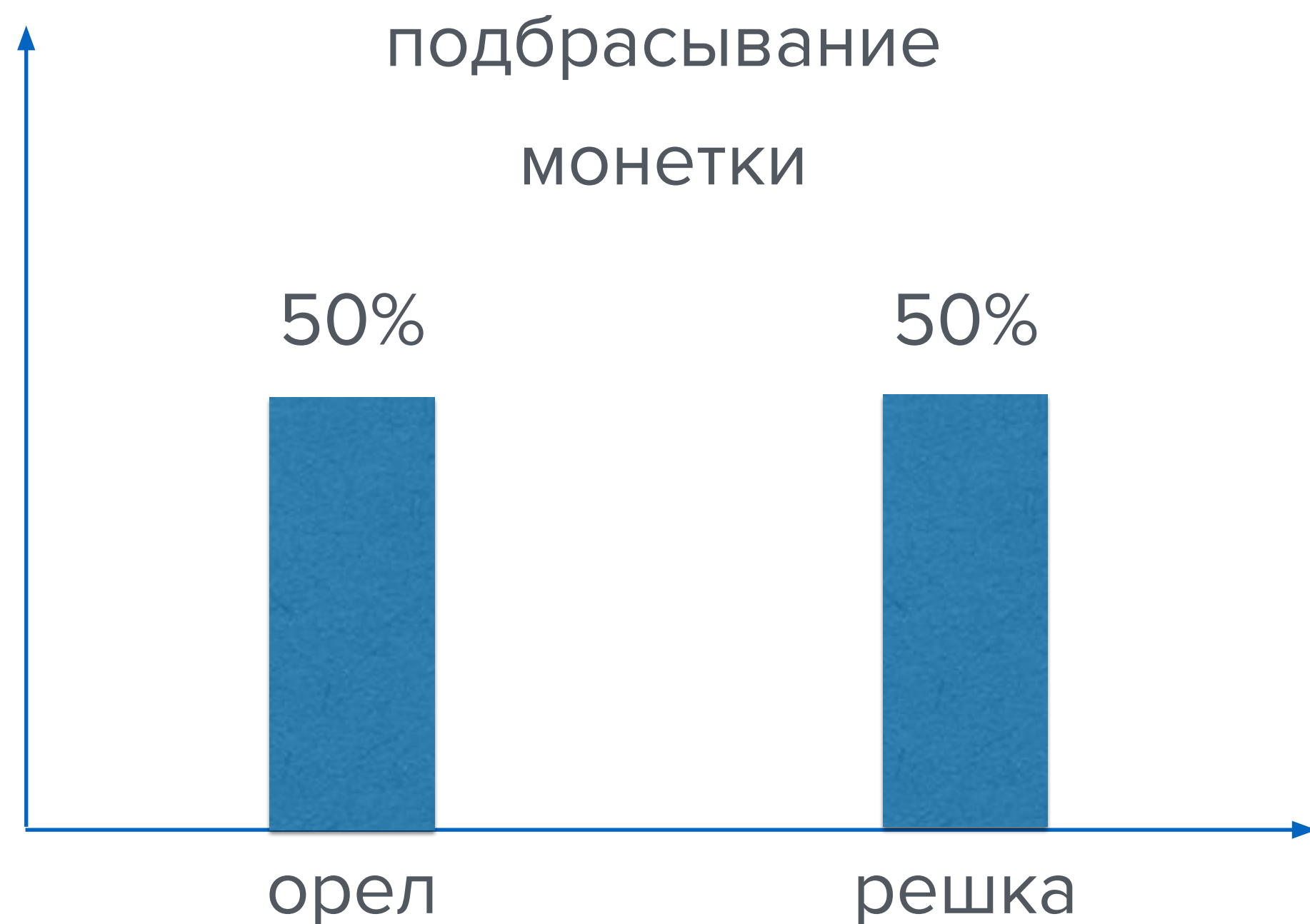
частотное
определение

N – количество наблюдений

m – количество наступлений события A

РАСПРЕДЕЛЕНИЕ

Закон, описывающий область значений случайной величины и вероятности их исхода



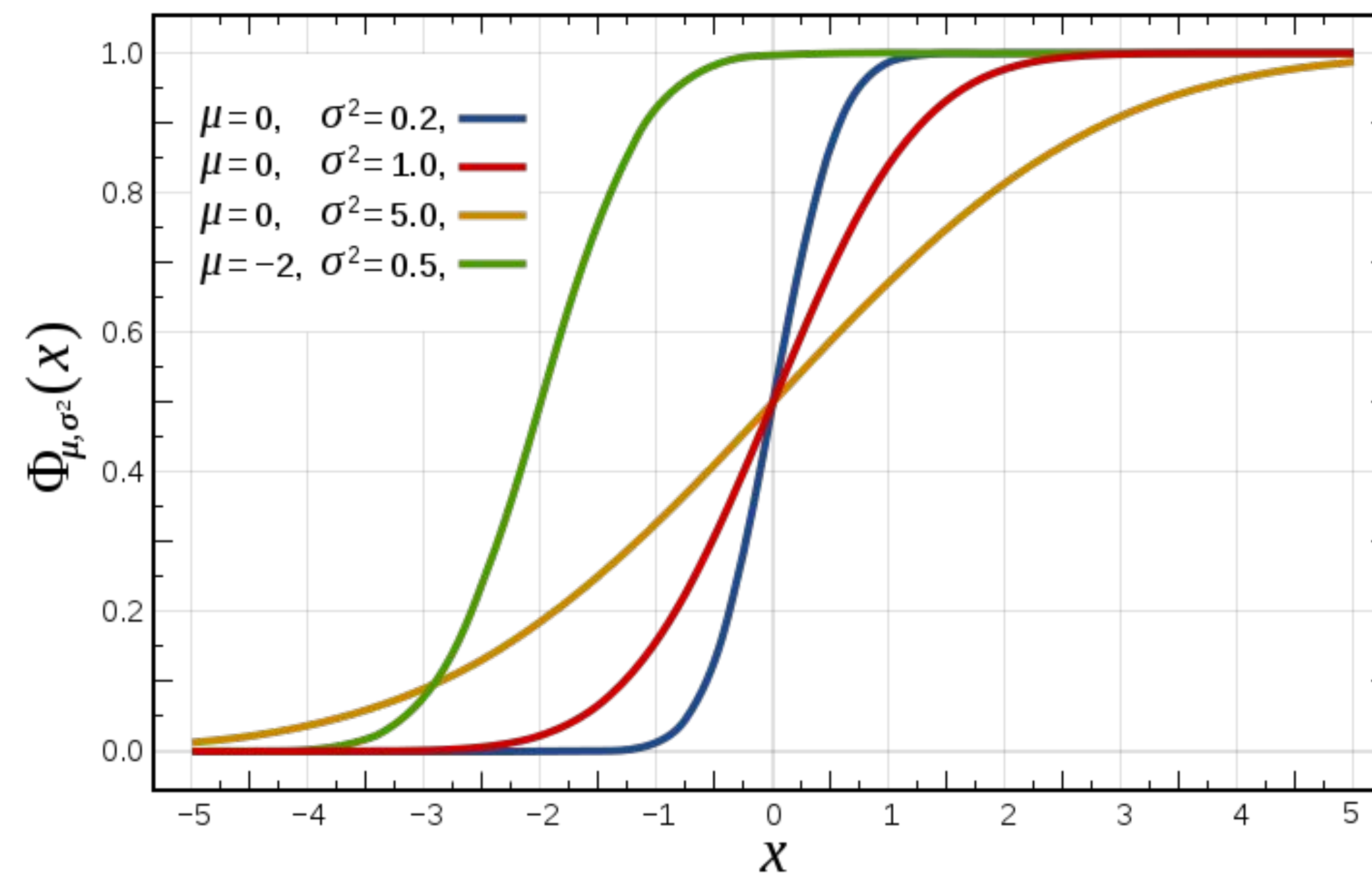
ПРАКТИЧЕСКОЕ ЗАДАНИЕ 1

Распределение для суммы выпадающих очков



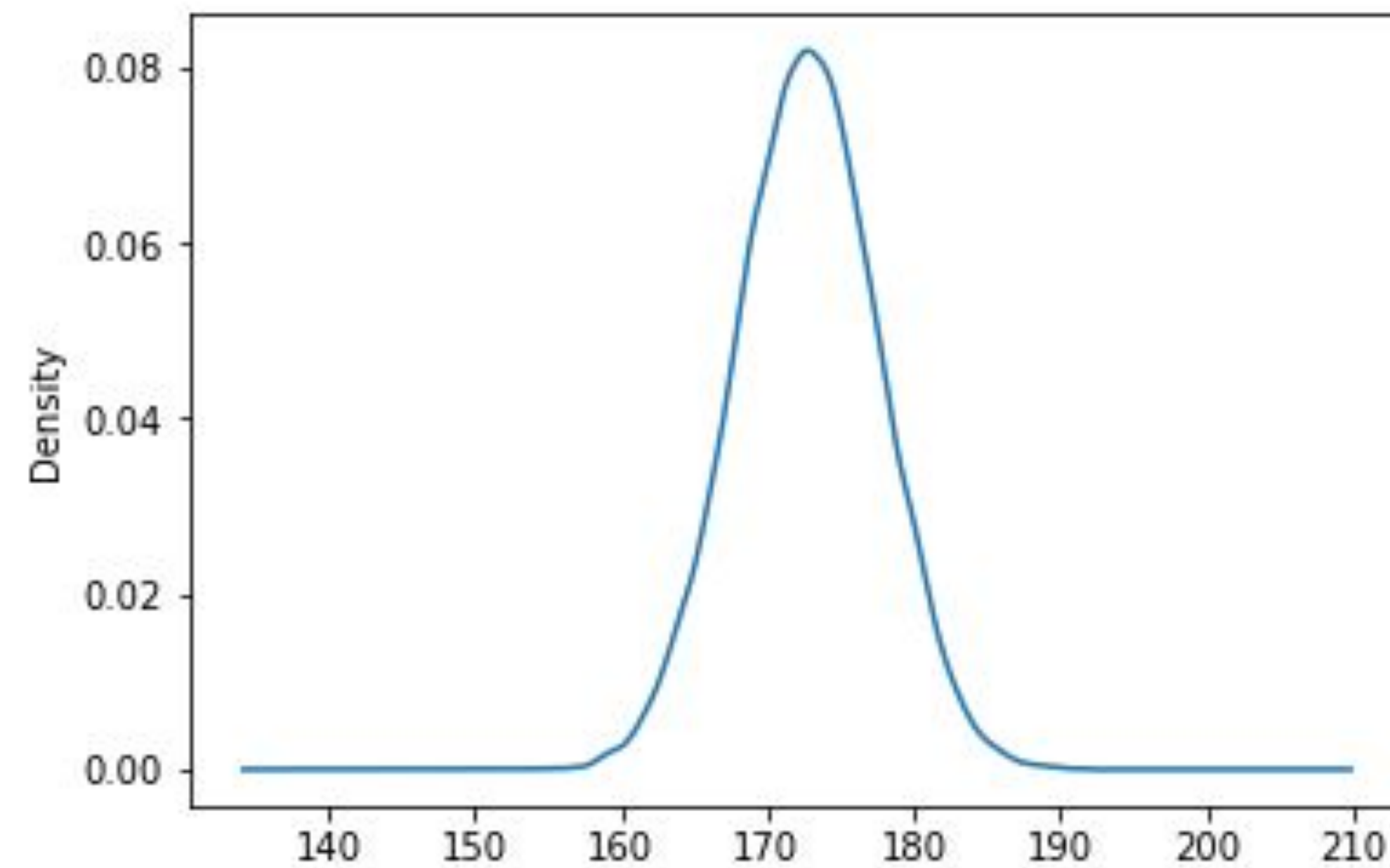
ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

Вероятность $F(x)$, что случайная величина примет значение, меньшее или равное x



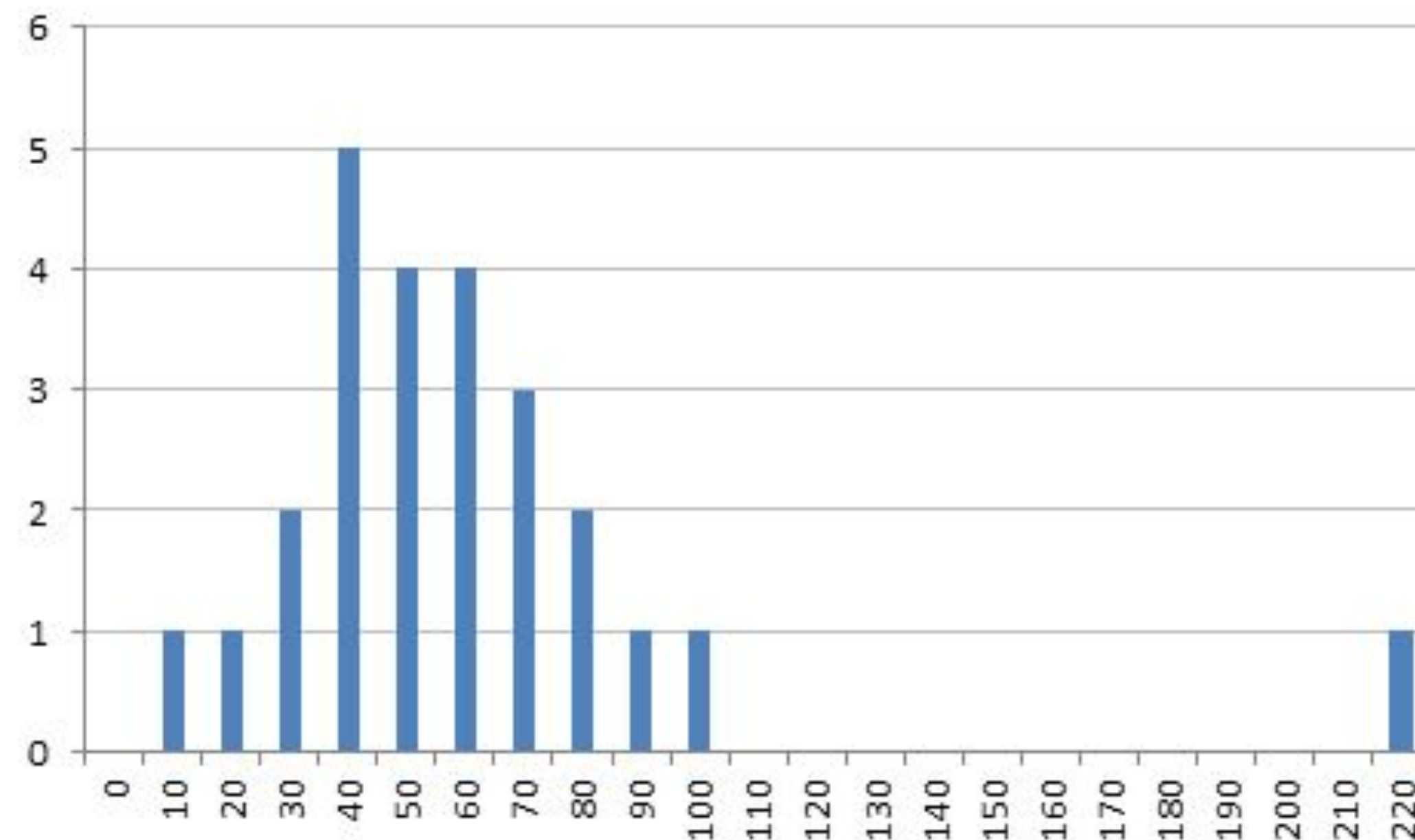
ПЛОТНОСТЬ ВЕРОЯТНОСТИ

Производная функции распределения



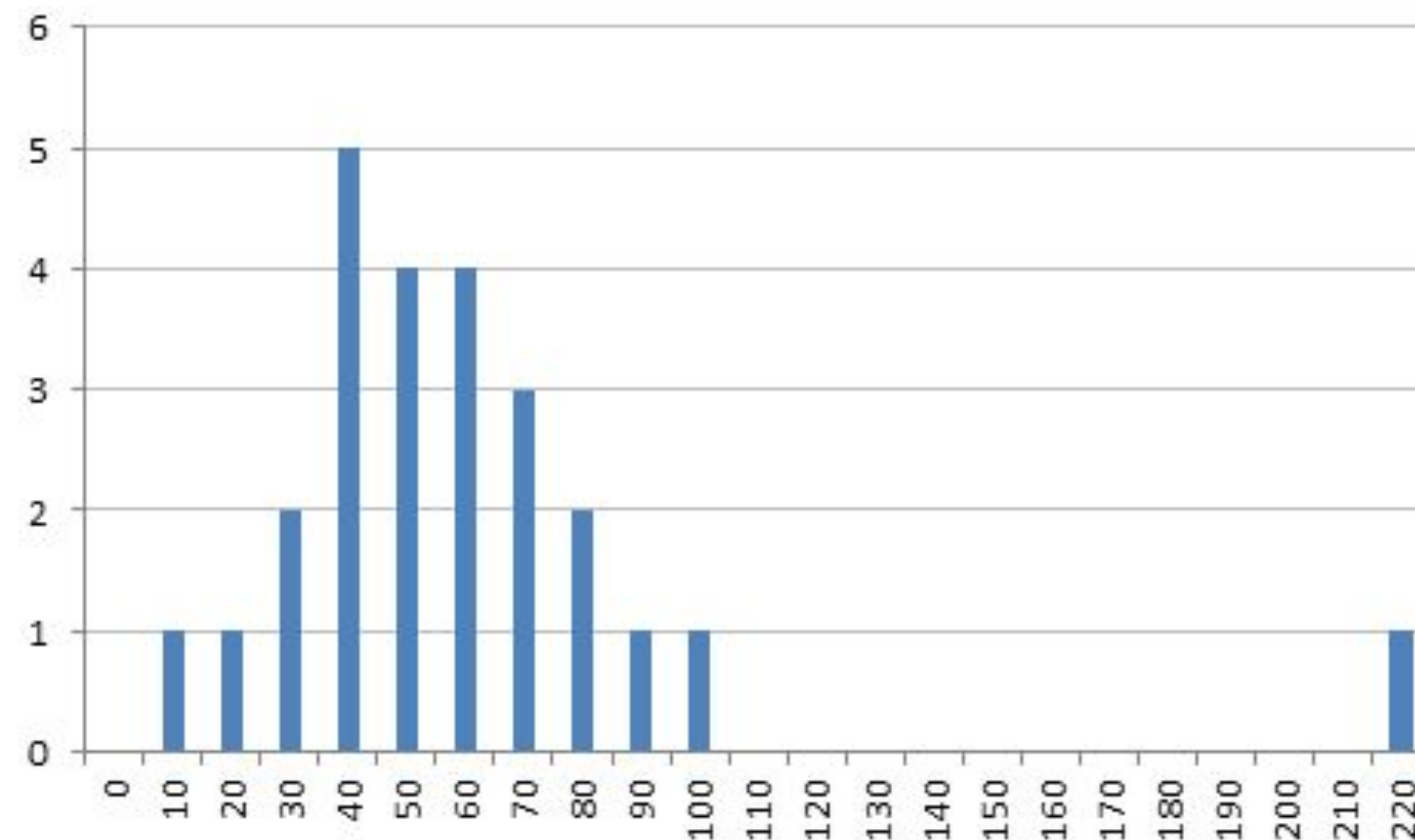
СРЕДНЕЕ И КОМПАНИЯ

Среднее (mean) – отношение суммы значений признака к количеству измеренных значений



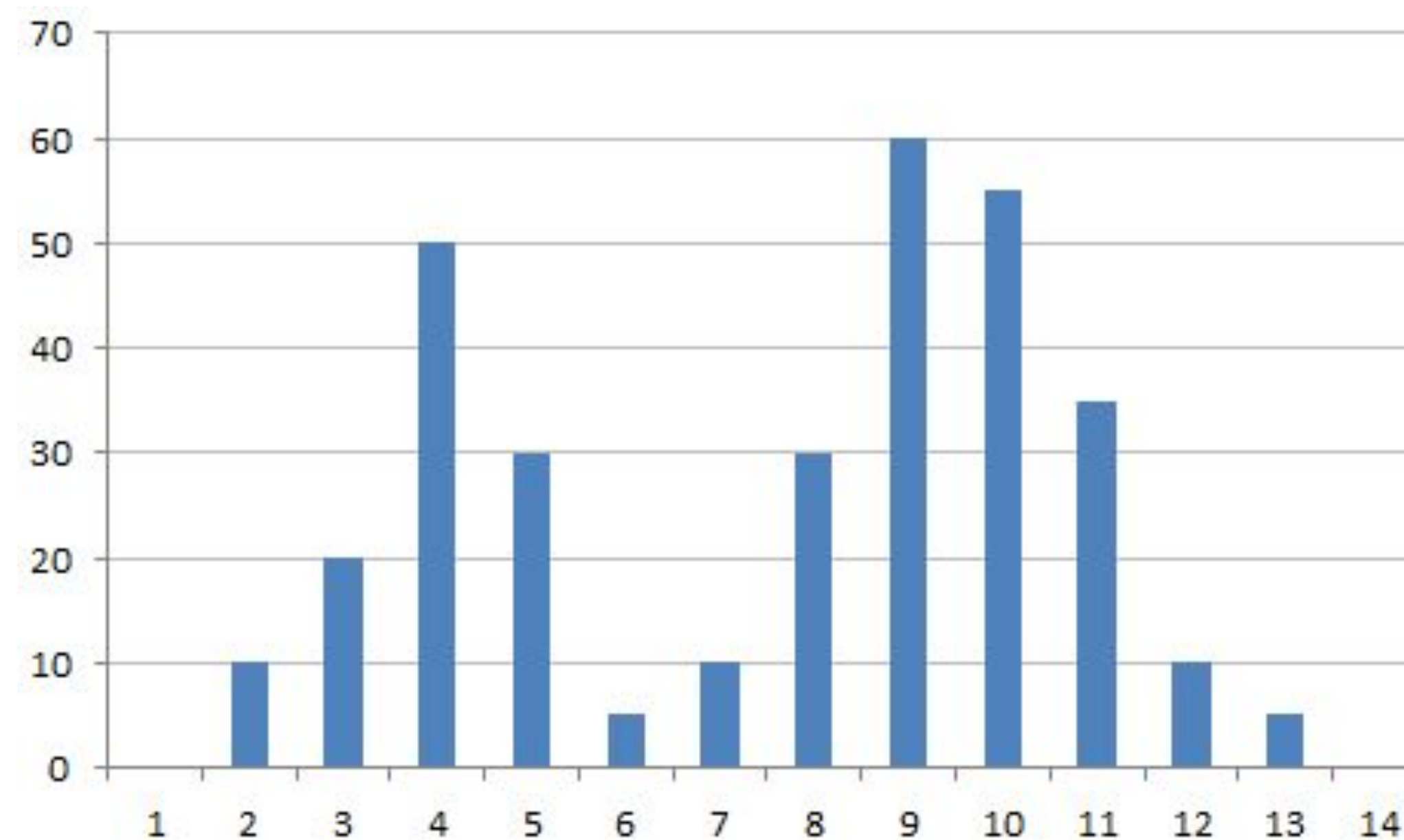
СРЕДНЕЕ И КОМПАНИЯ

Медиана (median) – значение признака, которое делит упорядоченное множество данных пополам



СРЕДНЕЕ И КОМПАНИЯ

Мода (mode) – значение признака, которое встречается наиболее часто



МЕРЫ ИЗМЕНЧИВОСТИ

Размах (Range) – разница между максимальным и минимальным значением

Дисперсия (variance) – средний квадрат отклонений признака от среднего значения

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

МЕРЫ ИЗМЕНЧИВОСТИ

Среднеквадратичное отклонение (std)

$$std = \sqrt{D}$$

Std имеет ту же размерность, что и измеряемая величина.

Показывает отклонение от среднего значения по выборке

РАЗМЕРНОСТЬ

Это важно

$$5 \text{ копеек} = \sqrt{25 \text{ копеек}} = \sqrt{\frac{1}{4}} \text{ рубля} = \frac{1}{2} \text{ рубля} = 50 \text{ копеек}$$

РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

РАСПРЕДЕЛЕНИЕ БЕРНУЛЛИ

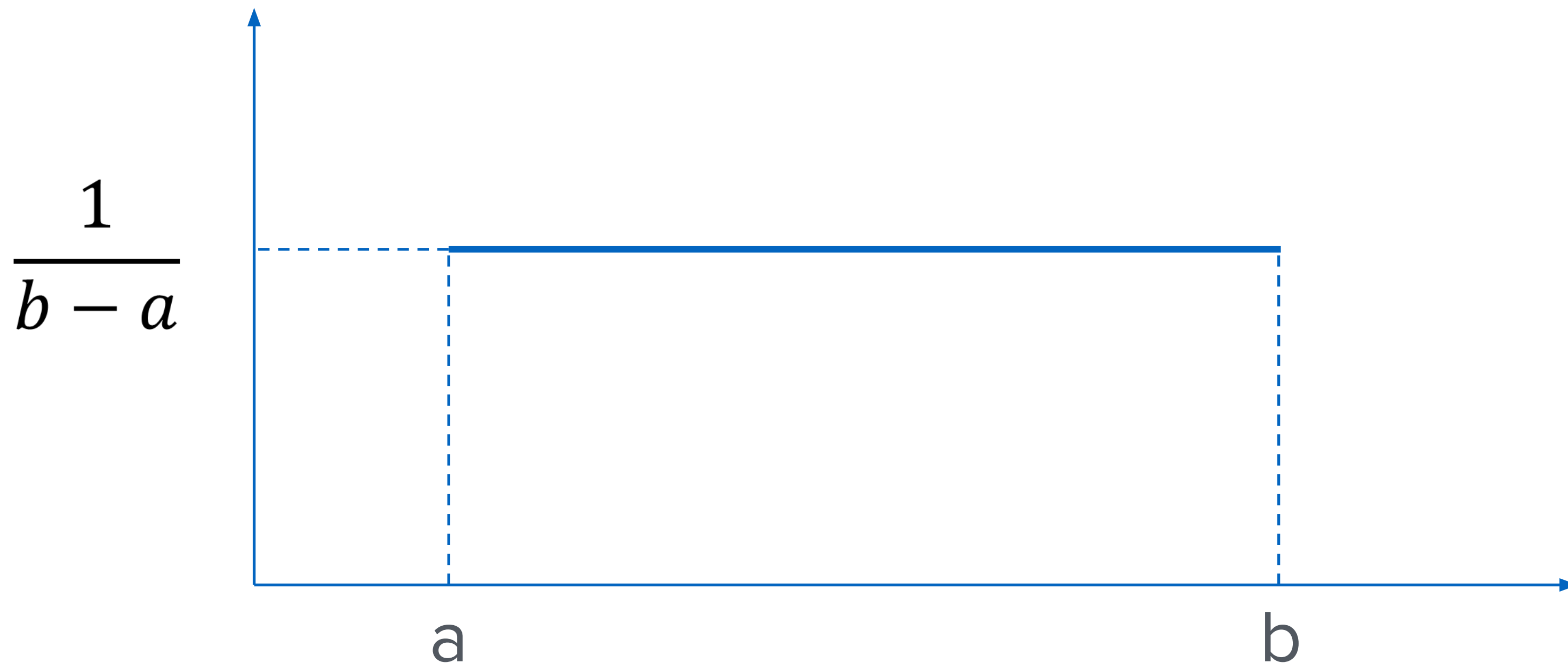
Случайная величина принимает 2 значения 0 и 1 с вероятностями p и q , при этом $p + q = 1$

БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Распределение «успехов» в последовательности из n независимых случайных экспериментов. Вероятность успеха в каждом равна p

РАВНОМЕРНОЕ РАСПРЕДЕЛЕНИЕ

Плотность вероятности принимает постоянное значение на определенном интервале



ГИПЕРГЕОМЕТРИЧЕСКОЕ

Имеем N элементов, D из которых обладают определенным свойством (дефект). Достаем n элементов. Тогда вероятность того, что k элементов из n являются дефектными равна:

$$P(k) = \frac{C_D^k C_{N-D}^{n-k}}{C_N^n}, \text{ где } C_n^k = \frac{n!}{k! (n-k)!}$$

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Плотность вероятности:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

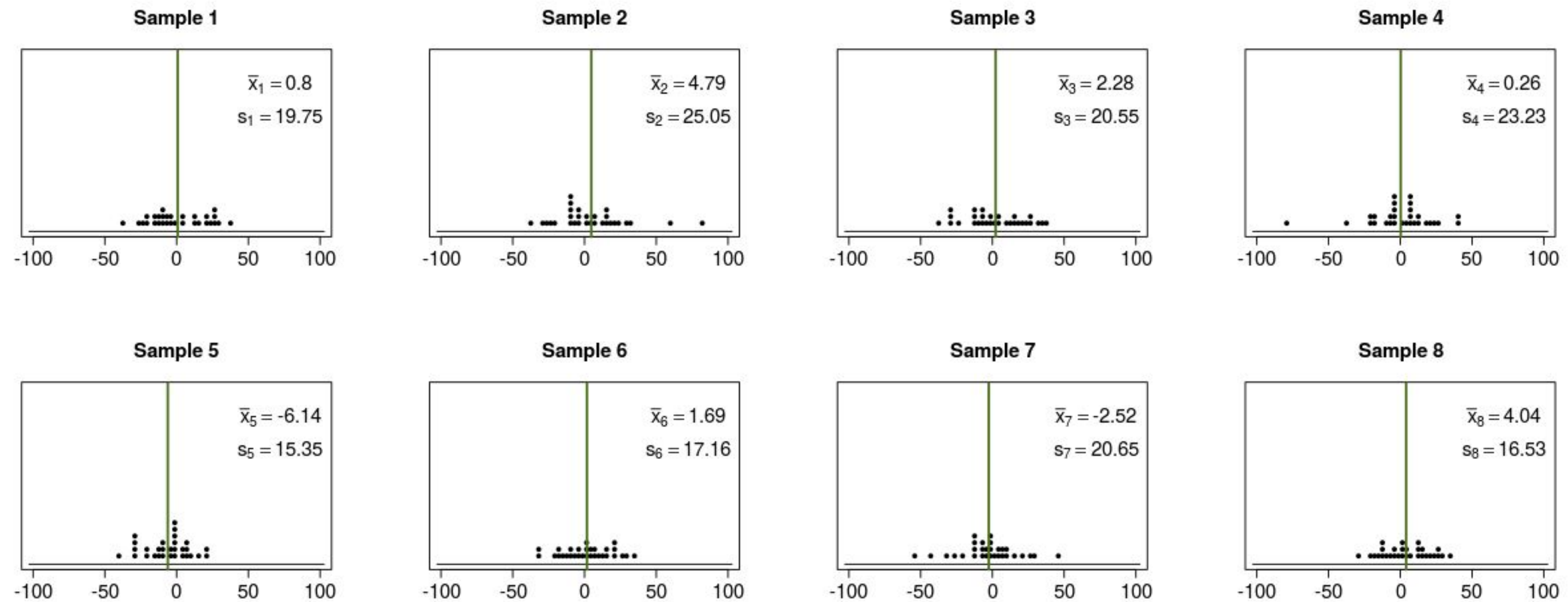
μ — среднее значение, σ — среднеквадратическое отклонение

ЦПТ

Сумма n независимых одинаково распределённых случайных величин имеет распределение, близкое к нормальному

РАСПРЕДЕЛЕНИЯ

ЦПТ

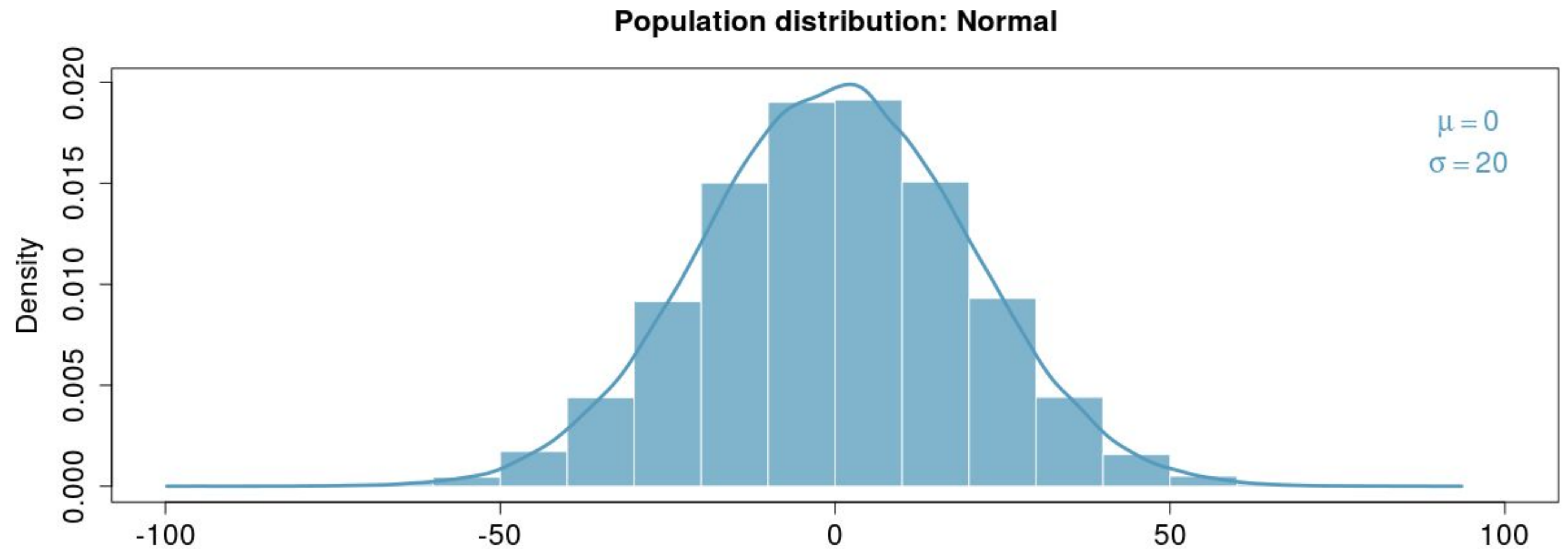


https://gallery.shinyapps.io/CLT_mean/

РАСПРЕДЕЛЕНИЯ



ЦПТ



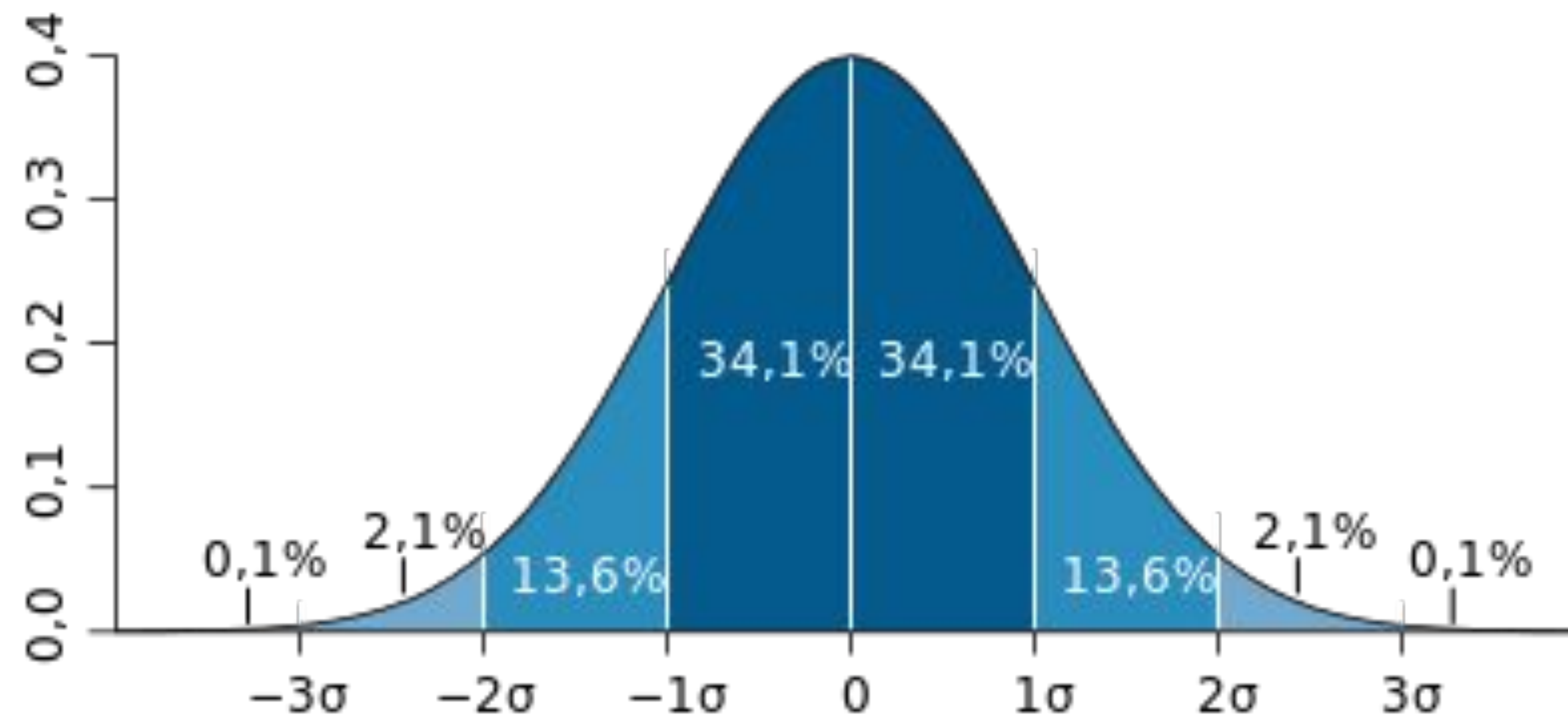
https://gallery.shinyapps.io/CLT_mean/

Z-СТАНДАРТИЗАЦИЯ

Приведение набора данных к нулевому среднему и $STD = 1$

$$Z_i = \frac{X_i - \bar{X}}{\sigma_x}$$

ПРАВИЛО ТРЕХ СИГМ



Доля значений

σ – 68.2%

2σ – 95.4%

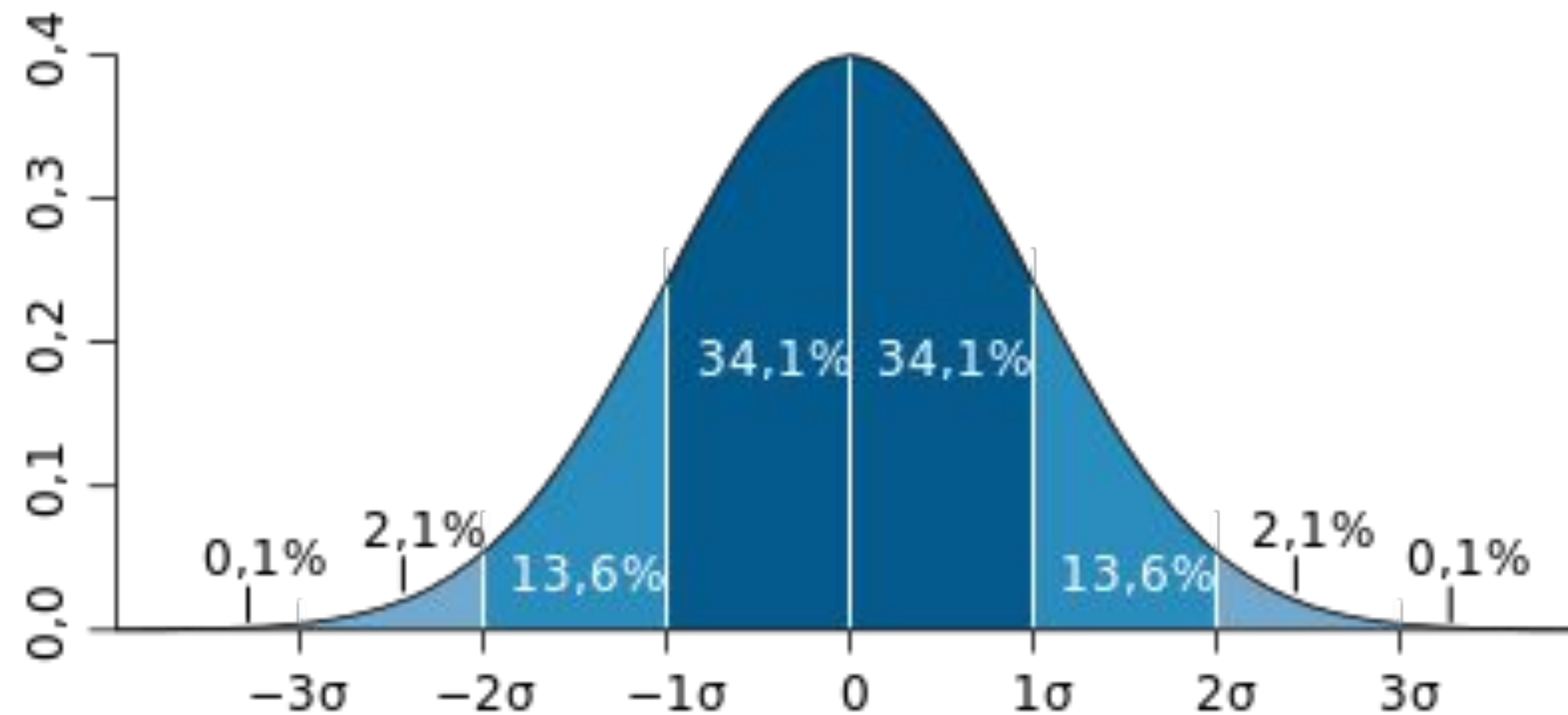
3σ – 99.6%

ПРИМЕР

Средний рост в нашей выборке 173см

Какова вероятность встретить человека выше 185см?
(в генеральной совокупности)

ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ



95% всех наблюдений
лежат в интервале ± 1.96
 σ

P-ЗНАЧЕНИЕ

Какова вероятность отклониться от среднего значения на величину, большую тестового значения?

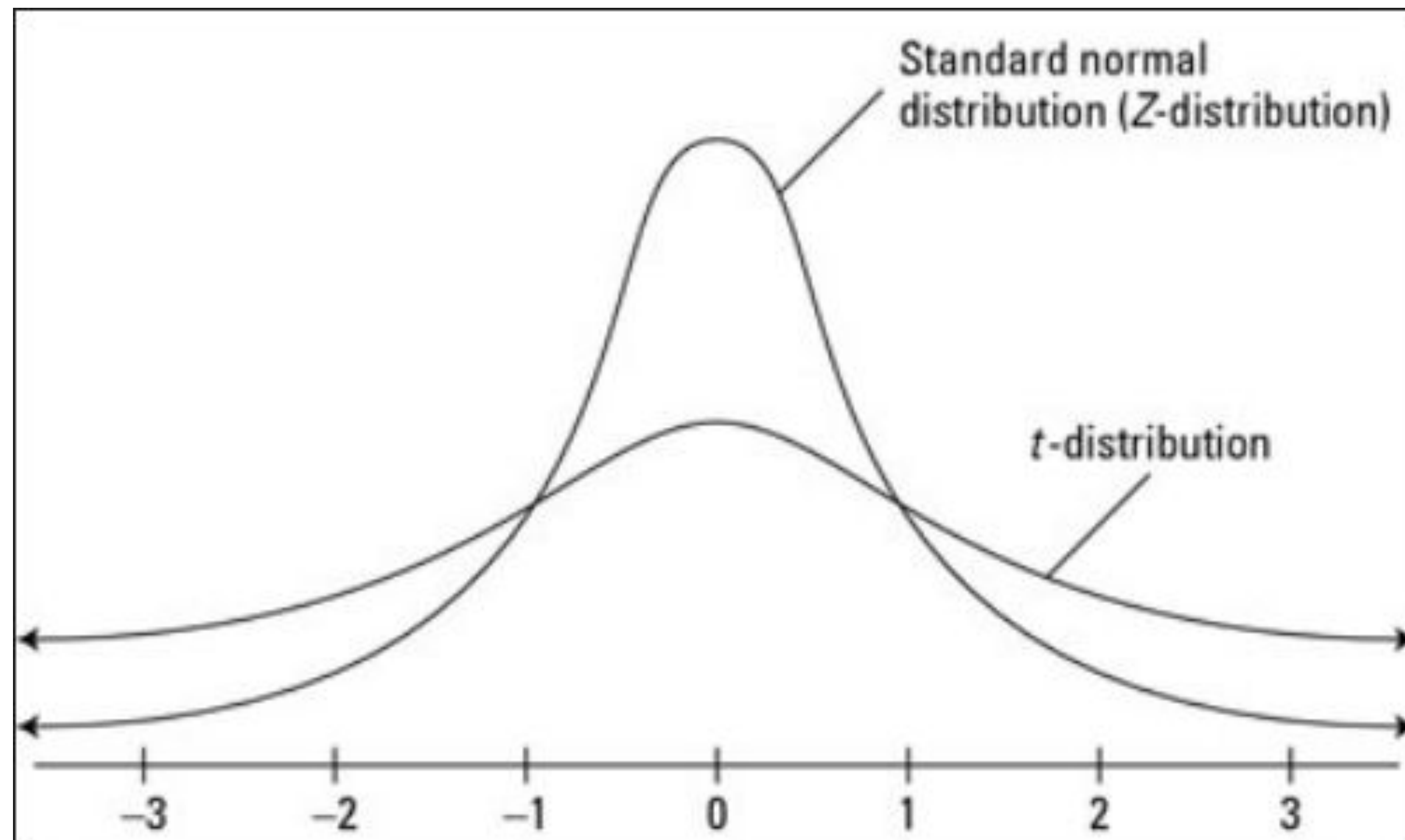
T-КРИТЕРИЙ СТЬЮДЕНТА

Исходные данные должны иметь нормальное
распределение

$N < 30$

Нарушается предположение о том, что выборочные
средние будут вести себя в соответствии с нормальным
законом

Т-РАСПРЕДЕЛЕНИЕ



ПРАКТИЧЕСКОЕ ЗАДАНИЕ 2

T-РАСПРЕДЕЛЕНИЕ

Возьмем из нашей выборки 20 человек ростом от 170 до 180см.

Можно ли утверждать, что их средний вес больше, чем в среднем по всем имеющимся данным?

КОРРЕЛЯЦИЯ

КОРРЕЛЯЦИЯ

Показывает статистическую взаимосвязь двух величин.

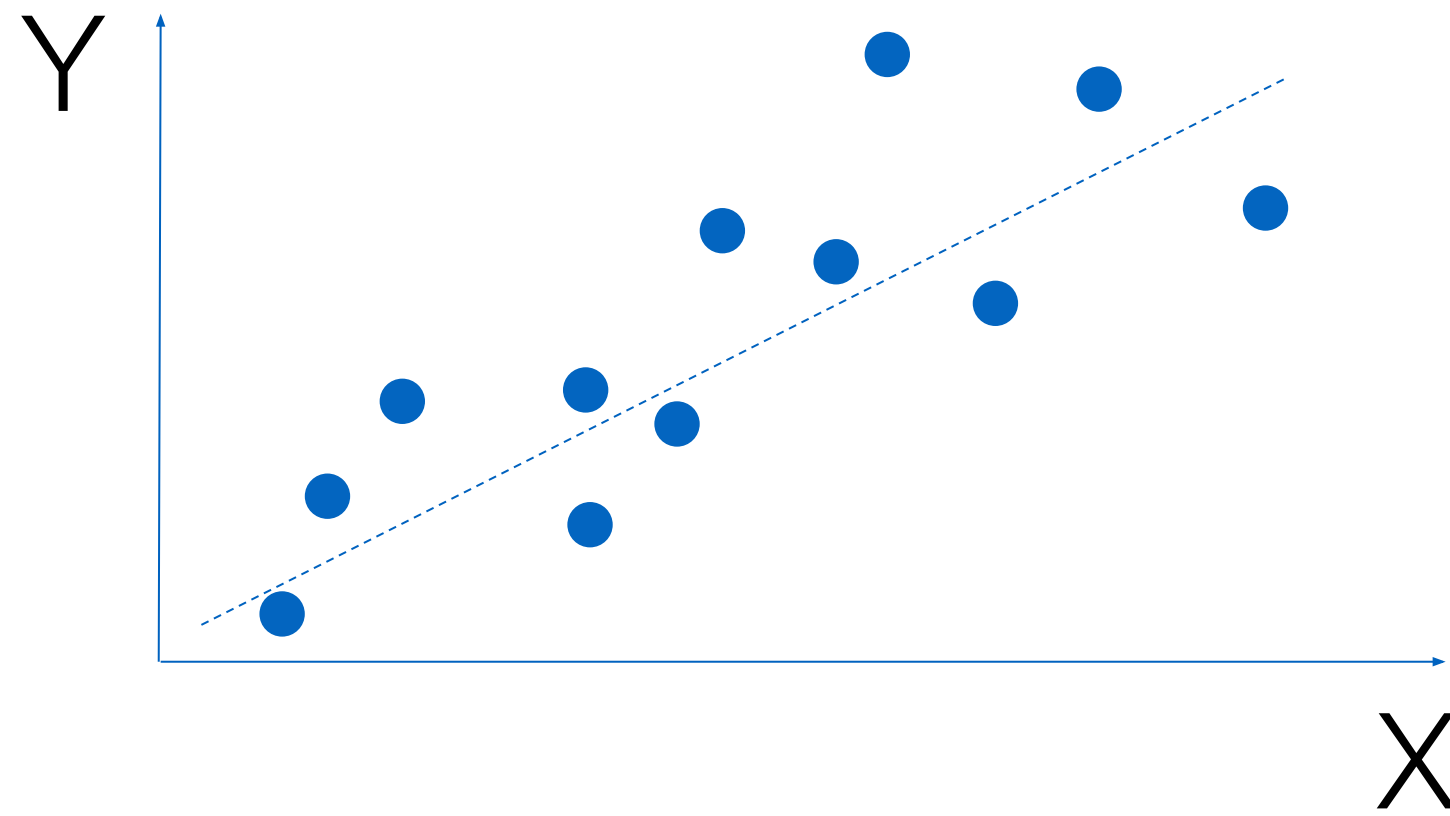
Насколько изменение одной величины связано с изменением другой

$$r_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}}$$

КОРРЕЛЯЦИЯ



Положительная



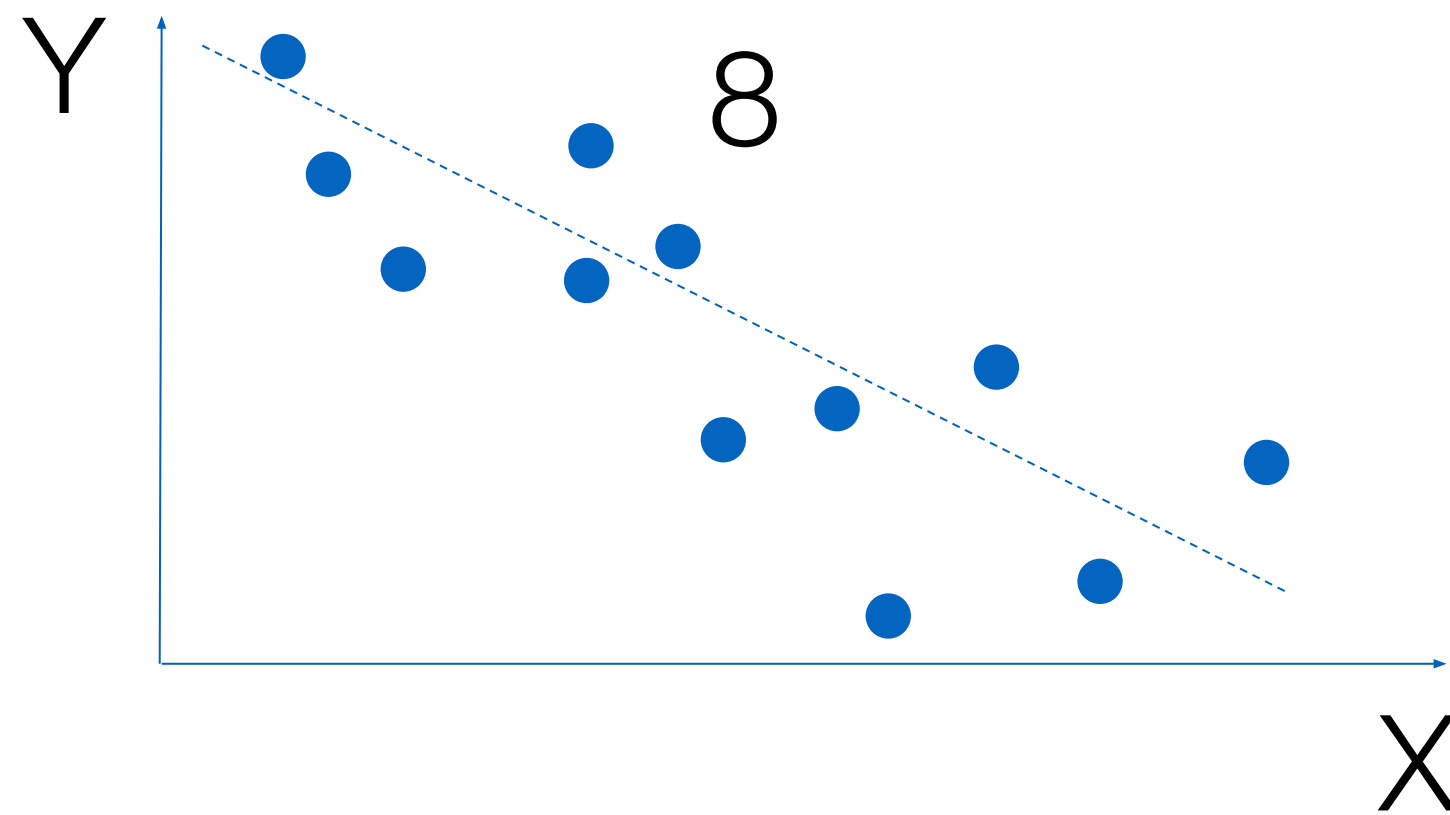
r

\approx

Отрицательная

0.

8

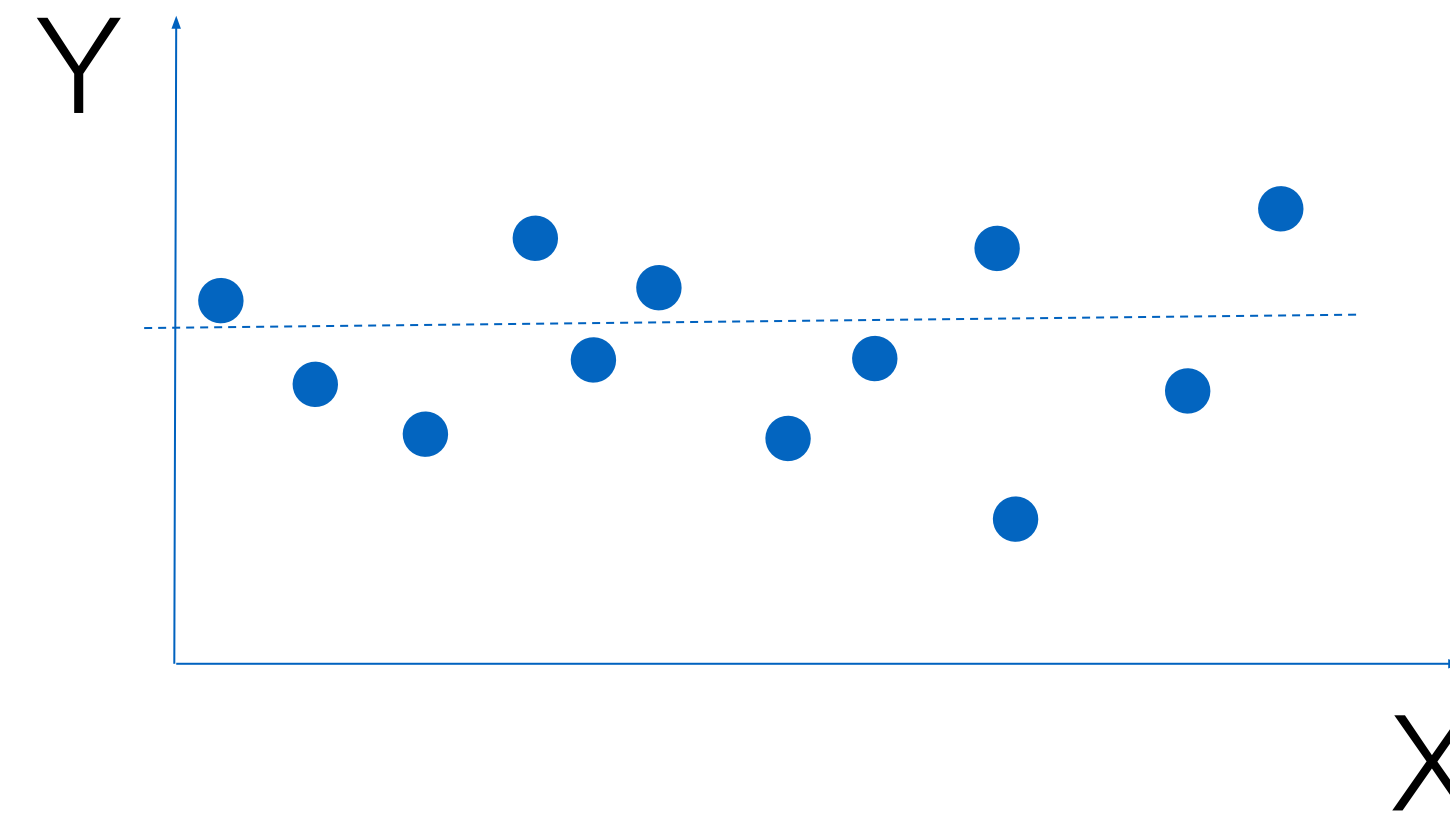


$r \approx -$

0.8

$$-1 \leq r_{XY} \leq 1$$

Отсутствие



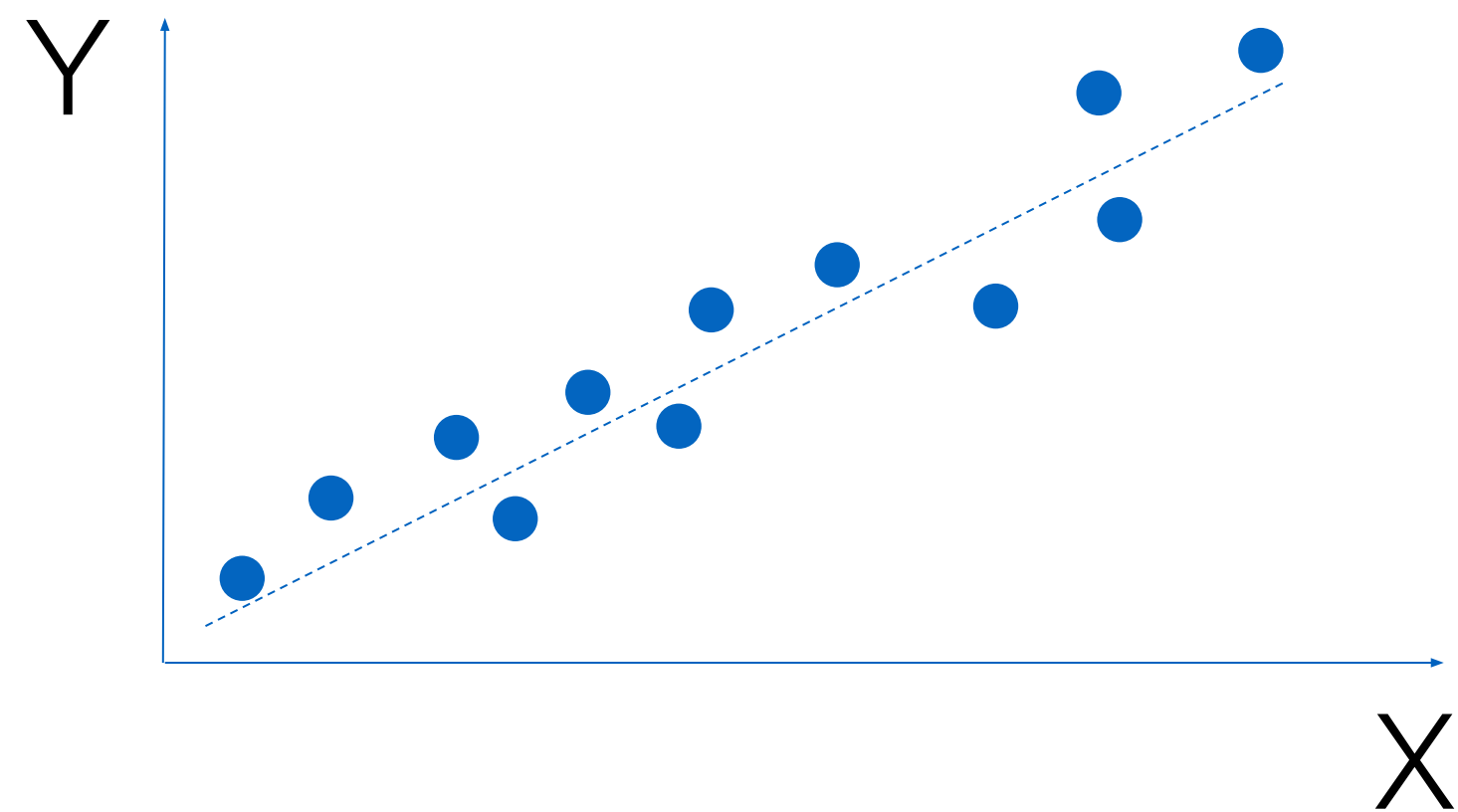
$r \approx 0$

КОРРЕЛЯЦИЯ И ЗАВИСИМОСТЬ

ЛОЖНАЯ КОРРЕЛЯЦИЯ

Коэффициент корреляции между X и Y более 90%

Верно ли, что Y зависит от X ?



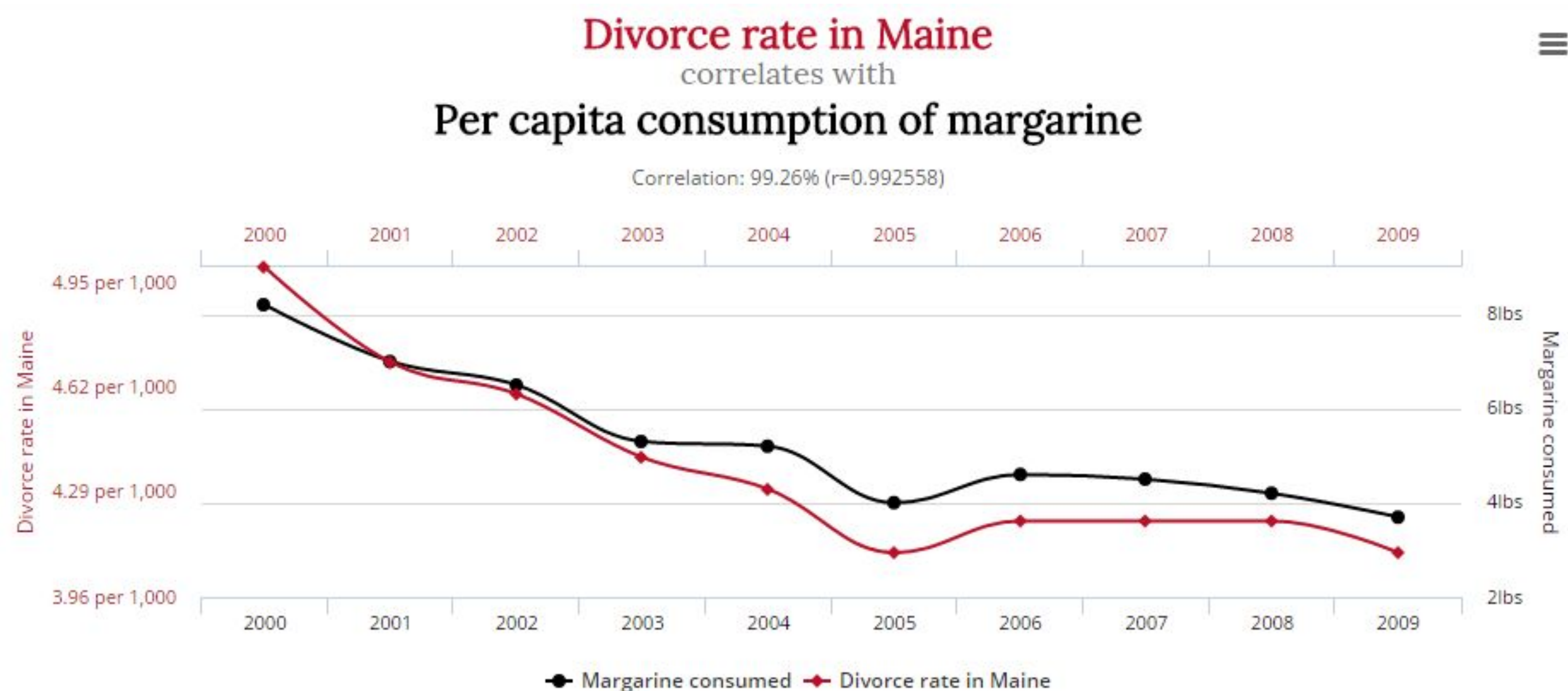
ЛОЖНАЯ КОРРЕЛЯЦИЯ

Примеры

- Количество пожаров пропорционально количеству машин в пожарной части.
- Интеллект школьника хорошо коррелирует с размером его обуви.
- Расходы на науку в США и число самоубийств имеют коэффициент корреляции 99,8%.

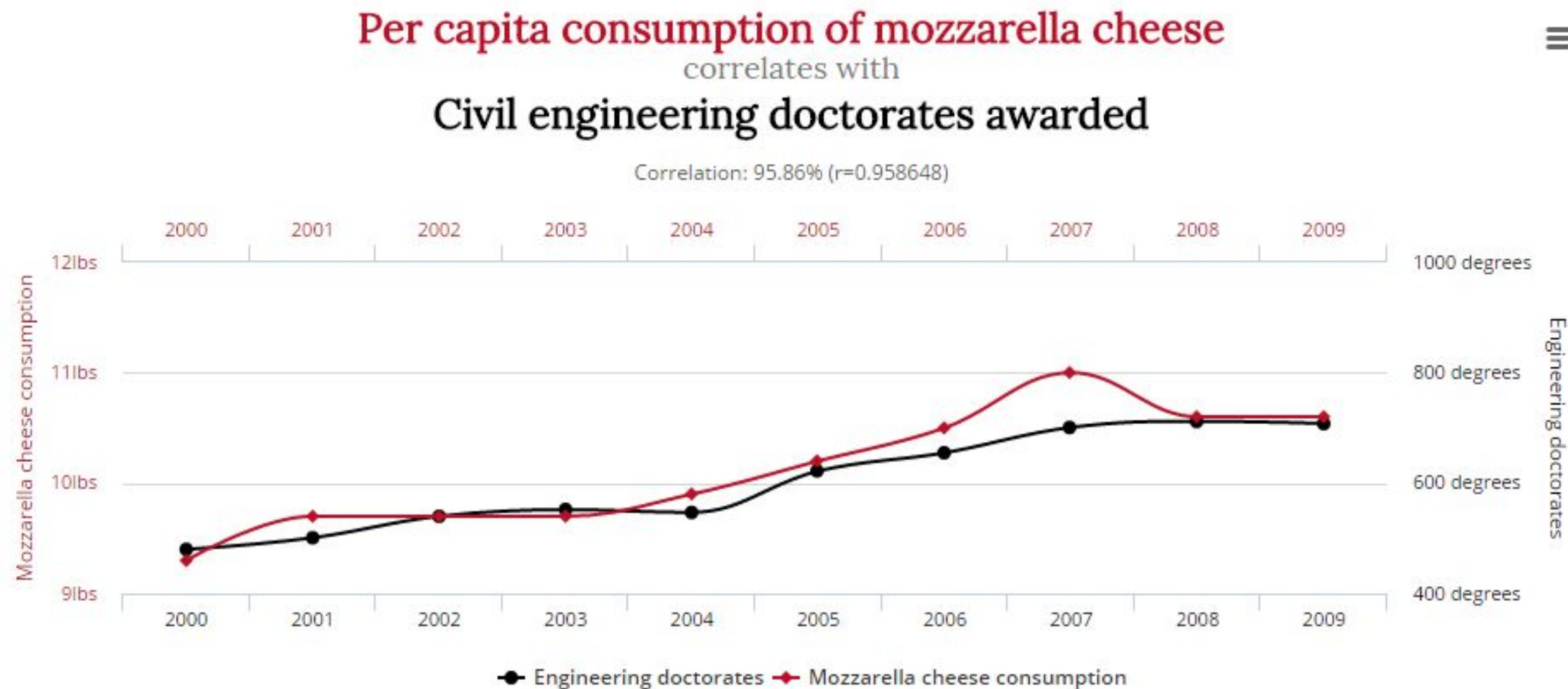
СТРАННЫЕ КОРРЕЛЯЦИИ

<http://www.tylervigen.com/spurious-correlations>



СТРАННЫЕ КОРРЕЛЯЦИИ

<http://www.tylervigen.com/spurious-correlations>





НЕТОЛОГИЯ
групп

Спасибо за внимание!

КОНСТАНТИН БАШЕВОЙ

Habr: @kpi_maker