

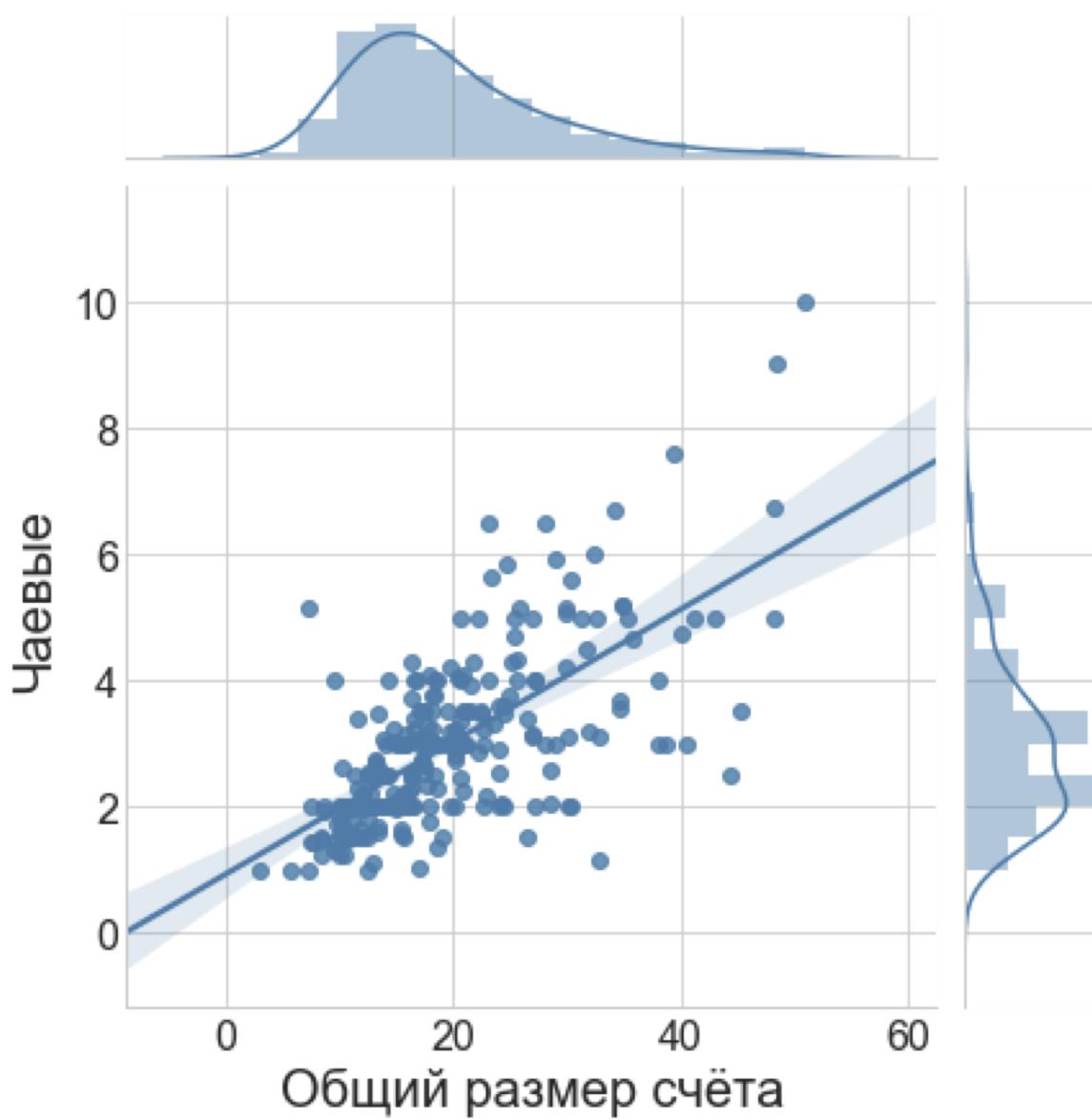
# Особенности в данных

# План

- В реальности случайные величины часто зависят друг от друга: поговорим как эту зависимость можно измерить
- На практике часто встречается нормальное распределение: поговорим про его свойства и как с ним работать
- Данные не бывают идеальными: поговорим про то, какие проблемы в них могут возникать и как с ними бороться
- В том числе о том, как описательные статистики помогают в этой борьбе

# Зависимые и независимые случайные величины

# Зависимые случайные величины



- Случайные величины часто взаимосвязаны между собой
- Нужен какой-то способ измерять взаимосвязь между ними

# Независимость

Говорят, что события  $A$  и  $B$  **независимы**, если

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

Говорят, что случайные величины  $X$  и  $Y$  **независимы**, если

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y) =$$

$$\mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq y) = F_X(x) \cdot F_Y(y)$$

Можно сформулировать это же определение в терминах плотностей:

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

# Ковариация

**Ковариация** – мера линейной зависимости двух случайных величин, вычисляется как

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))]$$

По аналогии с дисперсией, раскрыв скобки, получаем более простую формулу для вычисления:

$$\begin{aligned} Cov(X, Y) &= \mathbb{E}[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))] = \\ &= \mathbb{E}(X \cdot Y - X \cdot \mathbb{E}(Y) - \mathbb{E}(X) \cdot Y + \mathbb{E}(X) \cdot \mathbb{E}(Y)) = \\ &= \mathbb{E}(X \cdot Y) - \mathbb{E}(X \cdot \mathbb{E}(Y)) - \mathbb{E}(\mathbb{E}(X) \cdot Y) + \mathbb{E}(\mathbb{E}(X) \cdot \mathbb{E}(Y)) = \\ &= \mathbb{E}(X \cdot Y) - \mathbb{E}(Y) \cdot \mathbb{E}(X) - \mathbb{E}(X) \cdot \mathbb{E}(Y) + \mathbb{E}(X) \cdot \mathbb{E}(Y) = \\ &= \mathbb{E}(X \cdot Y) - \mathbb{E}(X) \cdot \mathbb{E}(Y) \end{aligned}$$

# Свойства ковариации

$X, Y, Z$  – случайные величины       $a$  – константа

1.  $Cov(X, Y) = Cov(Y, X)$
2.  $Cov(a, b) = 0$
3.  $Cov(a \cdot X, Y) = a \cdot Cov(X, Y)$
4.  $Cov(X + a, Y) = Cov(X, Y)$
5.  $Cov(X + Z, Y) = Cov(X, Y) + Cov(Z, Y)$
6.  $Cov(X, X) = Var(X)$

# Свойства ковариации

$X, Y, Z$  – случайные величины       $a$  – константа

7. Если случайные величины независимы,

$$Cov(X, Y) = 0$$

8. Обратное неверно. Если ковариация равна нулю, случайные величины могут быть зависимы.
9. Если  $X$  и  $Y$  зависимы, тогда

$$\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y) + Cov(X, Y)$$

$$Var(X + Y) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$$

# Корреляция Пирсона

Ковариация имеет размерность равную произведению размерностей случайных величин

**Пример:** если  $X$  – рост,  $Y$  – вес, ковариация измеряется в  $\text{рост} \cdot \text{вес}$

Это неудобно  $\Rightarrow$  вводится безразмерный коэффициент корреляции:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma(X) \cdot \sigma(Y)}$$

Коэффициент корреляции характеризует тесноту и направленность линейной связи между случайными величинами и принимает значение от  $-1$  до  $1$ .

# Выборочные аналоги:

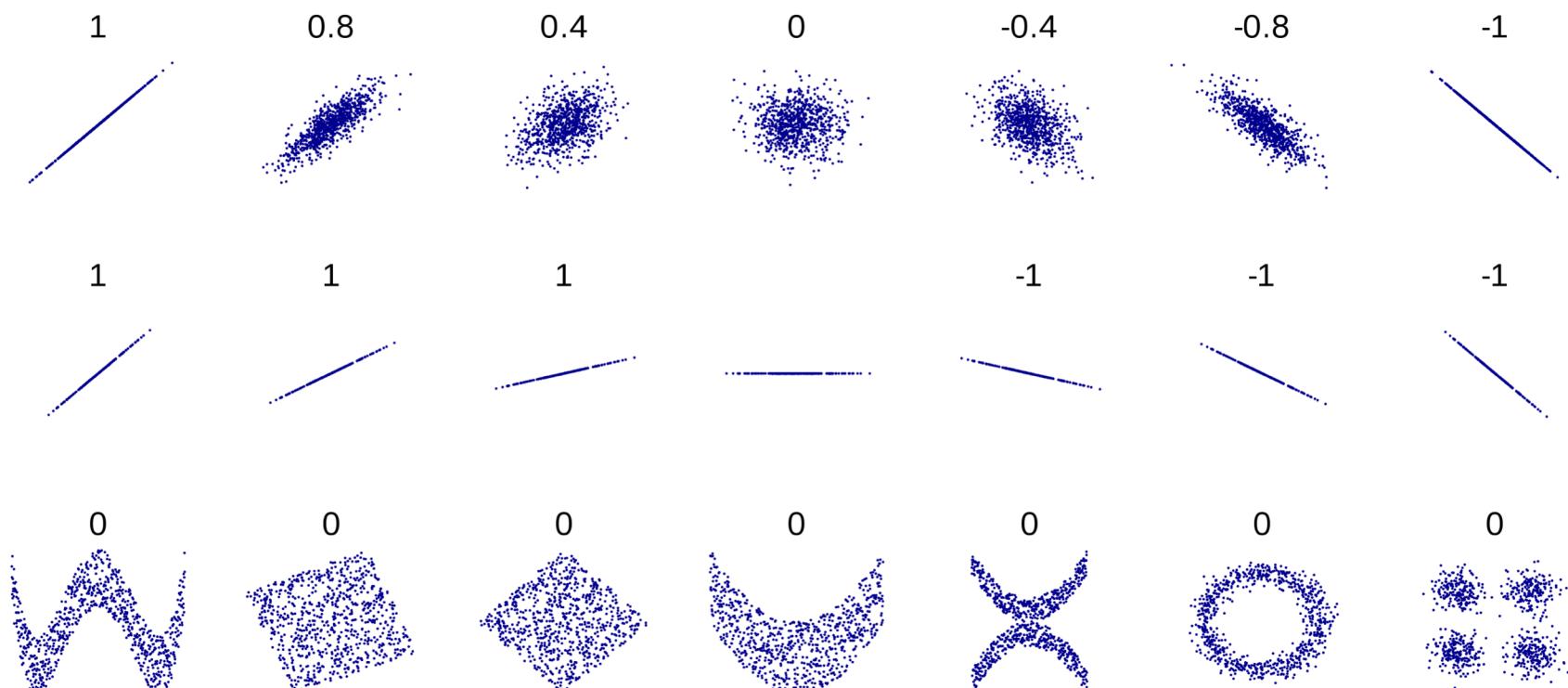
Выборочная ковариация:

$$\widehat{Cov}(X, Y) = \bar{xy} - \bar{x} \cdot \bar{y} = \frac{1}{n} \sum_{i=1}^n x_i \cdot y_i - \left( \frac{1}{n} \sum_{i=1}^n x_i \right) \cdot \left( \frac{1}{n} \sum_{i=1}^n y_i \right)$$

Выборочная корреляция (корреляция Пирсона):

$$\hat{\rho}(X, Y) = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\hat{\sigma}_x \cdot \hat{\sigma}_y}$$

# Корреляция Пирсона



# Корреляция Пирсона



Корреляция Пирсона улавливает только линейную взаимосвязь и чувствительна к выбросам

► Угадай корреляцию: <http://guessthecorrelation.com/>

# Корреляция Спирмена

**Корреляция Спирмена** – мера силы монотонной взаимосвязи. Вычисляется как корреляция Пирсона между **рангами наблюдений**.

$x_1, x_2, \dots, x_n$  – выборка

$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

Упорядочим

по возрастанию

**Правила выставления ранга:**

1. Порядковый номер наблюдения – ранг
2. Если встречаются несколько одинаковых значений, им присваивается одинаковое значение ранга, равное среднему арифметическому их порядковых номеров

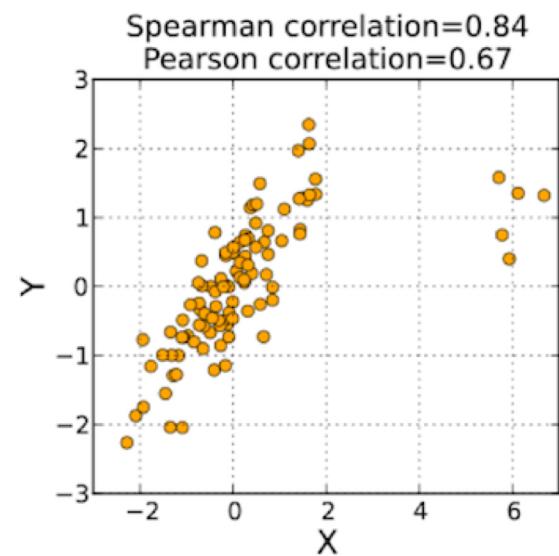
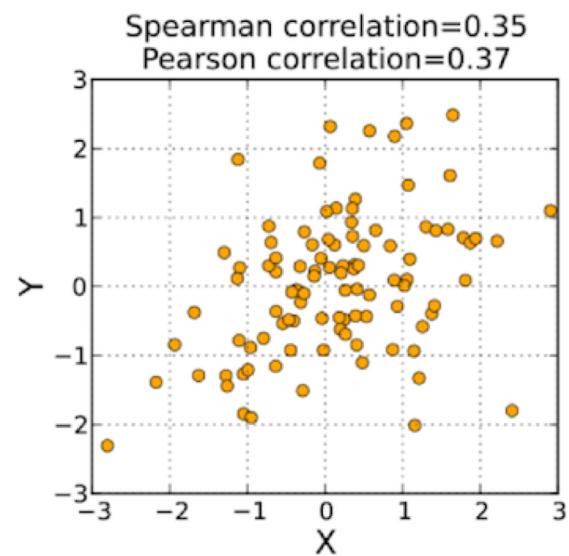
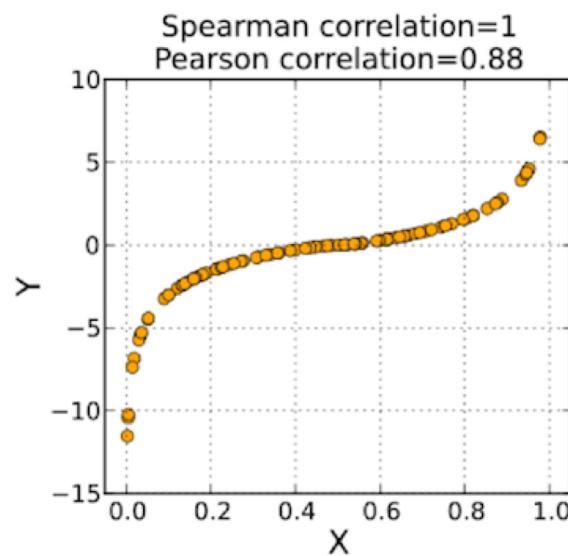
# Корреляция Спирмена

Пример:

	$X$	$Y$
Выборка:	10, 8, 6, 7, 4, 10, 9, 5	9, 9, 4, 5, 6, 8, 10, 7
Порядок:	7, 5, 3, 4, 1, 8, 6, 2	6, 7, 1, 2, 3, 5, 8, 4
Ранг:	7.5, 5, 3, 4, 1, 7.5, 6, 2	6.5, 6.5, 1, 2, 3, 5, 8, 4
	$r_x$	$r_y$

$$\hat{\rho}_s(X, Y) = \hat{\rho}_p(r_x, r_y) \approx 0.645$$

# Корреляция Спирмена

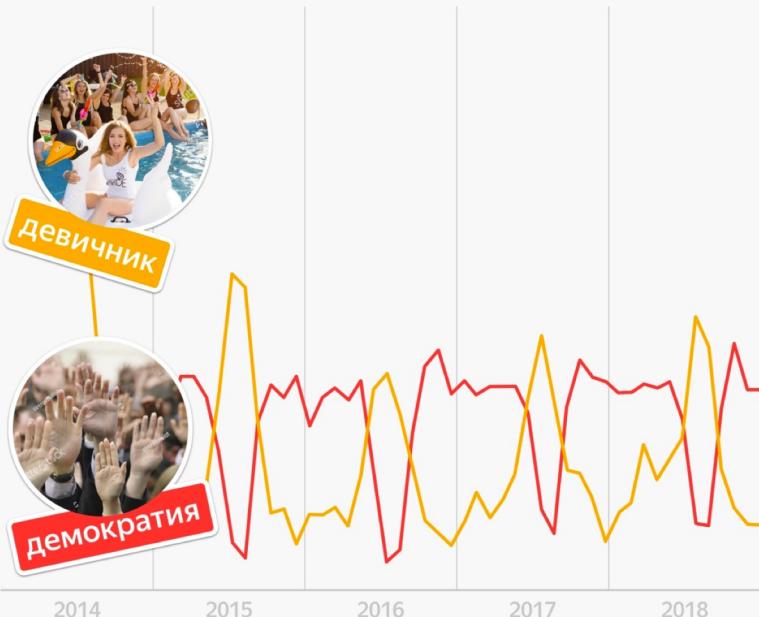


Корреляция Спирмэна пытается уловить  
в данных монотонность

# Корреляция не означает причинность

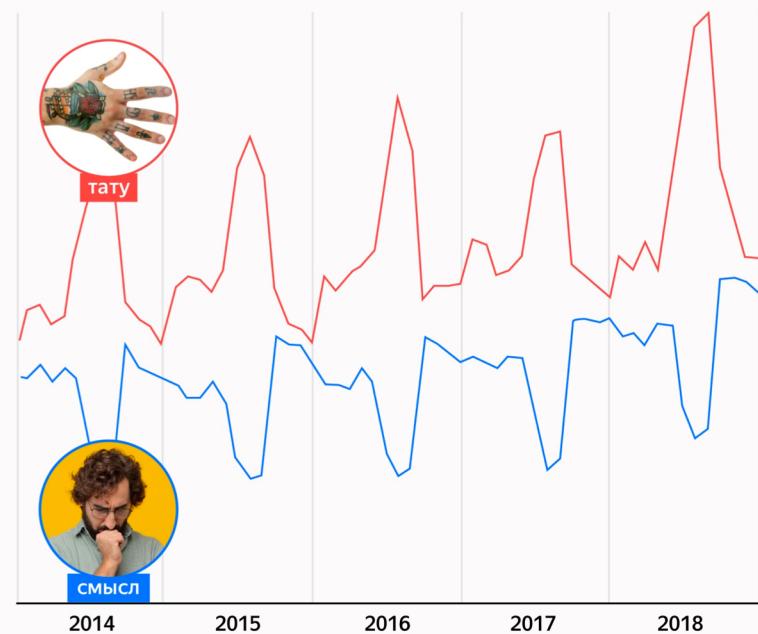
Парадоксы в Поиске — Яндекс

Когда в Поиске снижается доля запросов со словом **демократия**, становится больше запросов со словом **девичник**



Парадоксы в Поиске — Яндекс

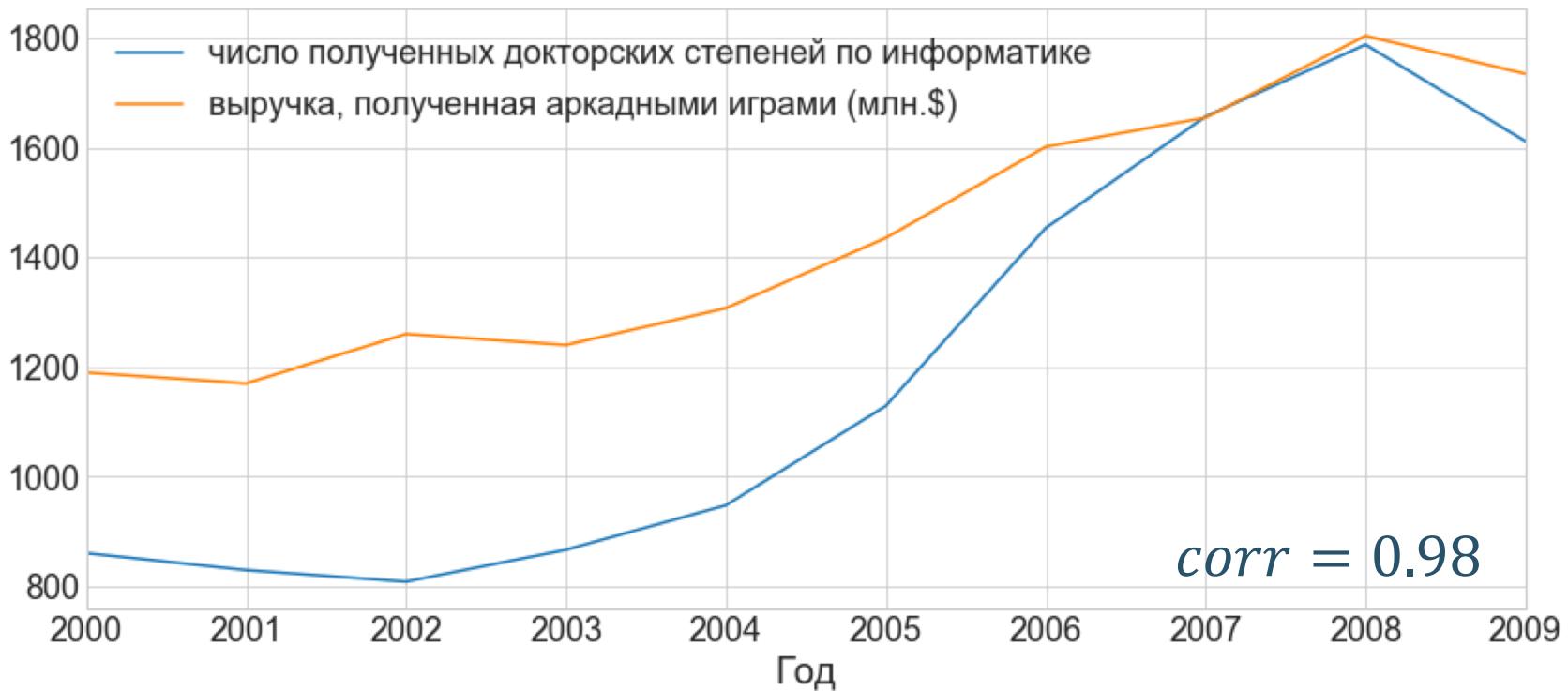
Когда в Поиске растёт интерес к **тату**, снижается доля запросов со словом **смысл**



- Больше примеров в пабликах Яндекса в социальных сетях по тегу #ПарадоксыВПоиске

# Ложная корреляция

Это связь, которая не имеет содержательного смысла



**Разгадка:** в динамике обоих рядов присутствует тренд. Если провести очистку от него, корреляция пропадёт.

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

# Ложная корреляция

Корреляция между величинами может быть вызвана общей причиной:

- Общий тренд в данных
- Спрос на мороженое и число грабежей коррелируют из-за погоды
- Цены на различные продукты могут коррелировать из-за инфляции

# Резюме

- Ковариация и корреляция Пирсона задают меру линейной связи между случайными величинами
- Корреляция Спирмена пытается измерить меру монотонной взаимосвязи
- Корреляции между переменными не достаточно для наличия причинно-следственной связи
- Иногда в данных присутствует ложная корреляция
- При работе с корреляцией надо быть осторожным: про то, как делать свои выводы аккуратно, мы поговорим на будущих неделях нашей специализации

# Нормальное распределение и его свойства

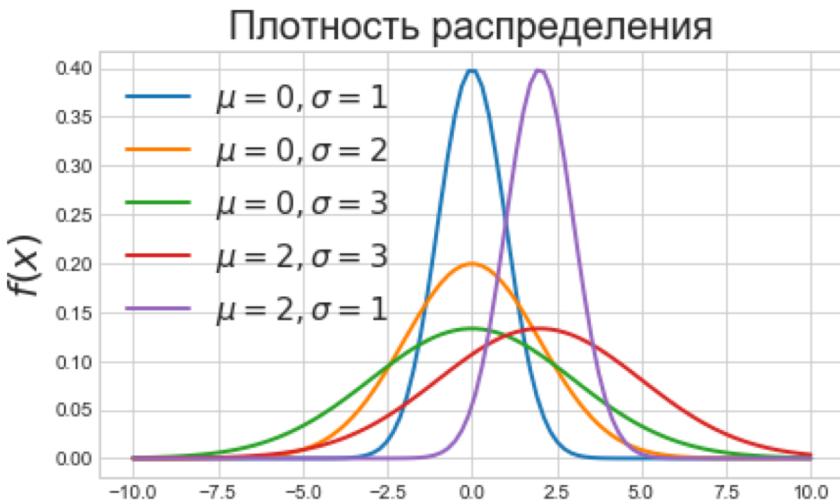
# Нормальное распределение

- В статистике часто встречается нормальное распределение
- Оно используется для проверки гипотез и для того, чтобы понимать насколько точными у нас получаются прогнозы и оценки
- Его обычно используют, когда у нас есть в распоряжении большая выборка, это разрешает делать Центральная Предельная Теорема (о ней поговорим на будущих неделях)
- Давайте познакомиться с нормальным распределением поближе

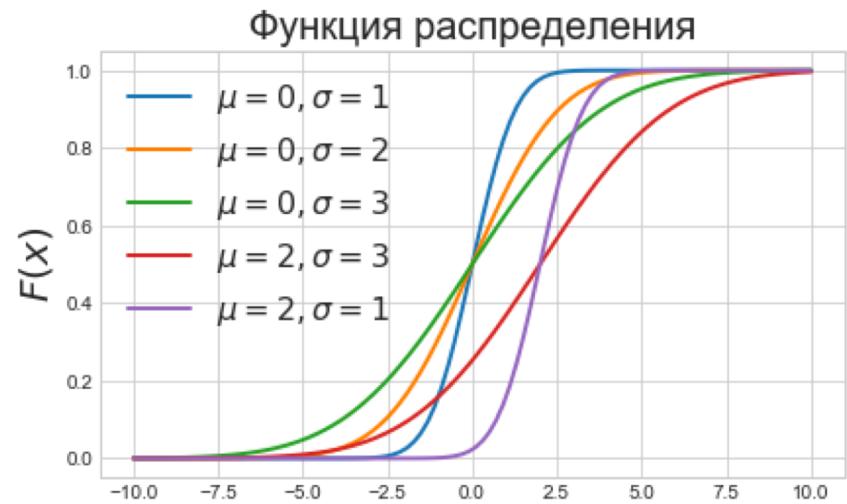
# Нормальное распределение

Нормальная случайная величина:  $X \sim N(\mu, \sigma^2)$

$$\mathbb{E}(X) = \mu, \text{Var}(X) = \sigma^2$$



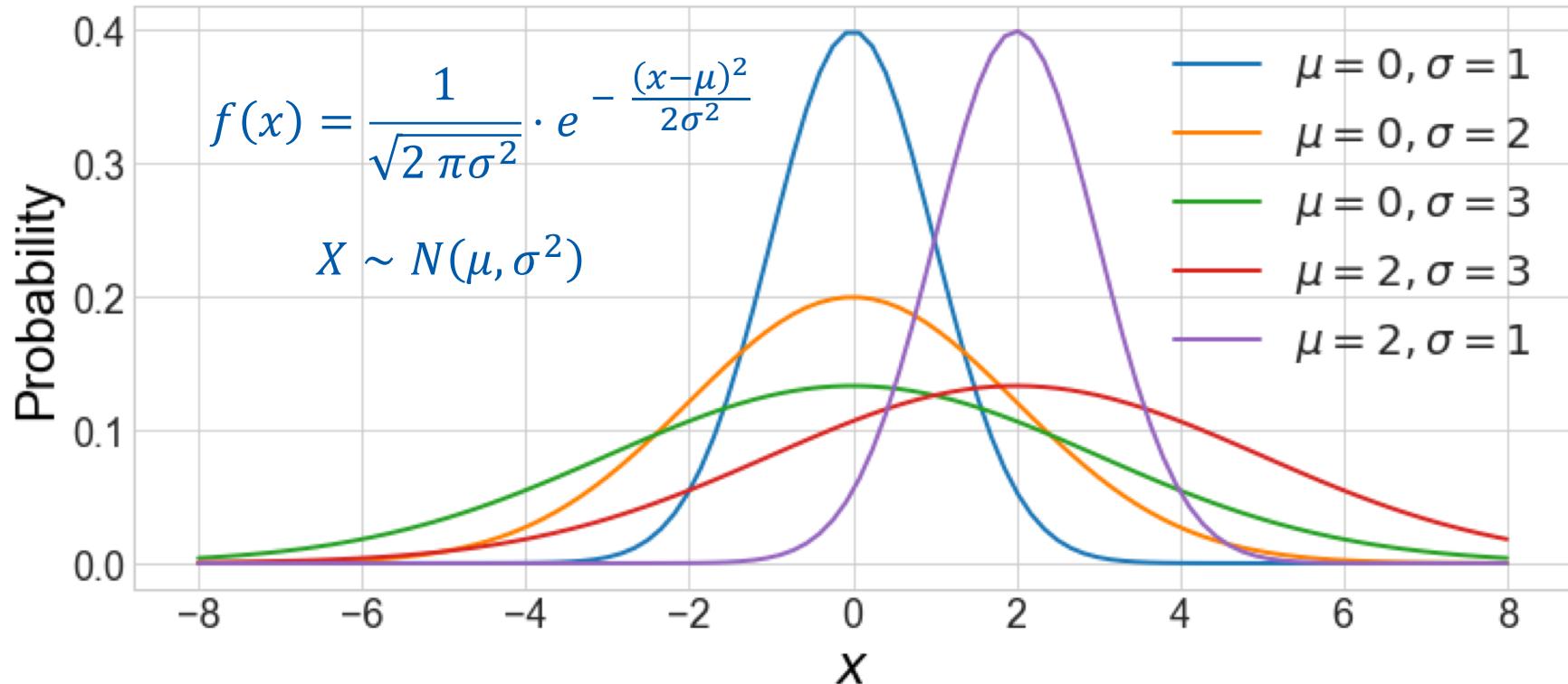
$$f(x) = \frac{1}{\sqrt{2 \pi \sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$F(x) = \int_{-\infty}^x f(x) \, dx$$

Функцию распределения нельзя найти в аналитическом виде, интеграл не берётся

# Свойства нормального распределения



1. Распределение симметрично относительно точки  $\mathbb{E}(X) = \mu$
2. Параметр  $\mu$  не влияет на форму кривой и отвечает за её сдвиг кривой вдоль оси  $x$ , параметр  $\sigma$  определяет степень “размытости” кривой

# Свойства нормального распределения

$$X \sim N(\mu_x, \sigma_x^2)$$

$$Y \sim N(\mu_y, \sigma_y^2)$$

$a$  – константа

3.  $X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$

4.  $X + a \sim N(\mu_x + a, \sigma_x^2)$

5.  $a \cdot X \sim N(a \cdot \mu_x, a^2 \cdot \sigma_x^2)$

Нормальная случайная величина устойчива  
к суммированию и линейным преобразованиям

# Центрирование и нормирование

$$X \sim N(\mu, \sigma^2)$$

центрирование

$$X - \mu \sim N(0, \sigma^2)$$

нормирование

$$\frac{X - \mu}{\sqrt{\sigma^2}} \sim N(0, 1)$$

- Распределение  $N(0, 1)$  называется **стандартным нормальным распределением**

# Стандартное нормальное распределение

- Функцию распределения для нормального распределения нельзя найти в аналитическом виде
- Для функции распределения случайной величины  $N(0, 1)$  составлены таблицы

# Как найти вероятность

$$X \sim N(7, 16)$$

$$\mathbb{P}(X \leq 15)$$

Искать такую вероятность неудобно, нужны были бы таблицы для всех возможных  $\mu$  и  $\sigma$

# Как найти вероятность

$$X \sim N(7, 16)$$

$$\mathbb{P}(X \leq 15) = \mathbb{P}\left(\frac{X - 7}{4} \leq \frac{15 - 7}{4}\right)$$

$$= \mathbb{P}(N(0, 1) \leq 2) = F_{N(0,1)}(2) = \Phi(2) \approx 0.98$$



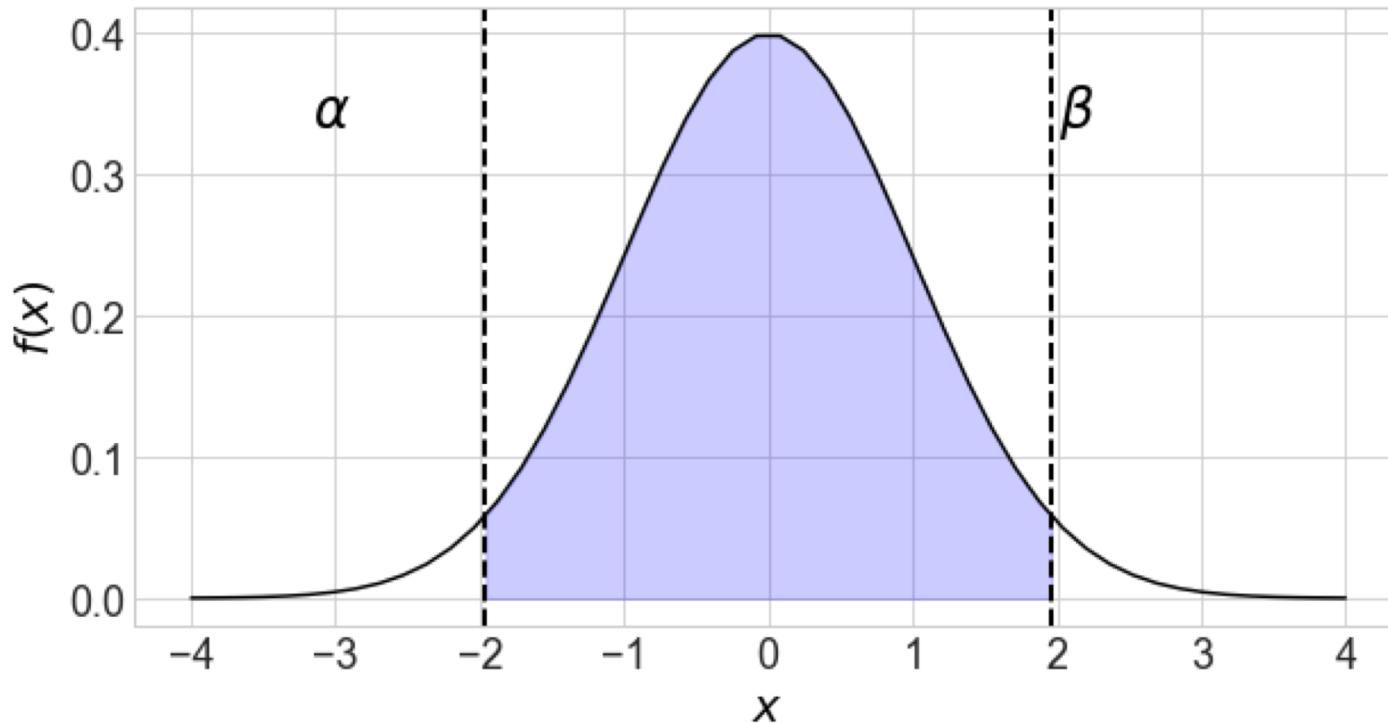
Обозначение  
для функции  
распределения  $N(0,1)$

Раньше активно пользовались таблицами для распределения  $N(0, 1)$ , сегодня для любого распределения расчёты делает компьютер

# Как найти вероятность

$$X \sim N(\mu, \sigma^2)$$

$$\begin{aligned}\mathbb{P}(\alpha \leq X \leq \beta) &= \mathbb{P}\left(\frac{\alpha - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{\beta - \mu}{\sigma}\right) = \\ &= \Phi\left(\frac{\beta - \mu}{\sigma}\right) - \Phi\left(\frac{\alpha - \mu}{\sigma}\right)\end{aligned}$$

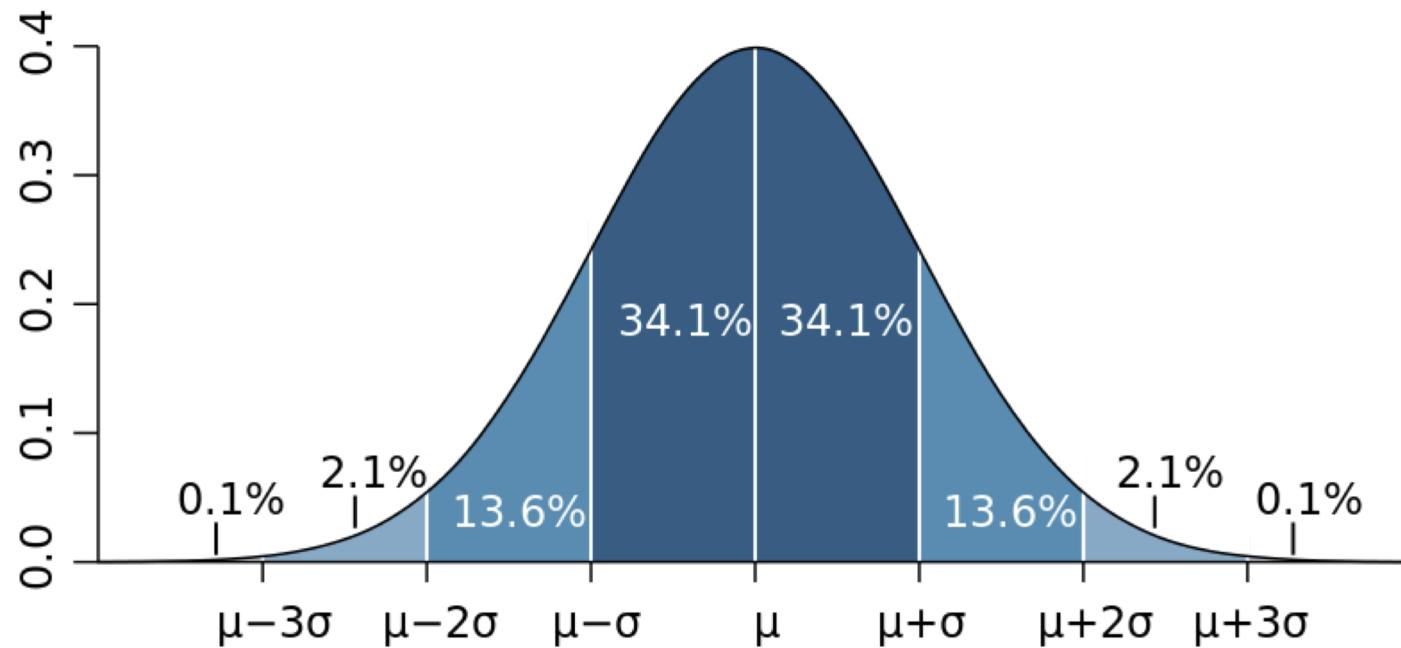


# Правила сигм

$$X \sim N(\mu, \sigma^2)$$

Правило сигмы:

$$\mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$$

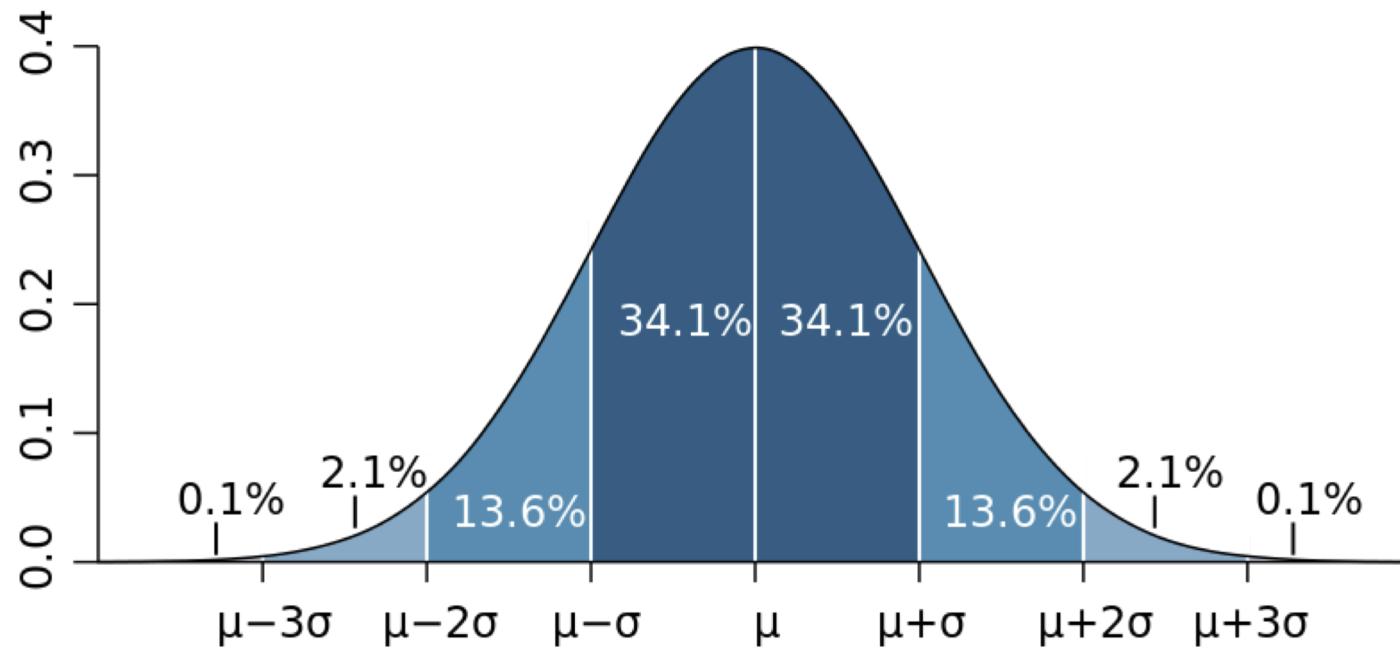


# Правила сигм

$$X \sim N(\mu, \sigma^2)$$

**Правило двух сигм:**

$$\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

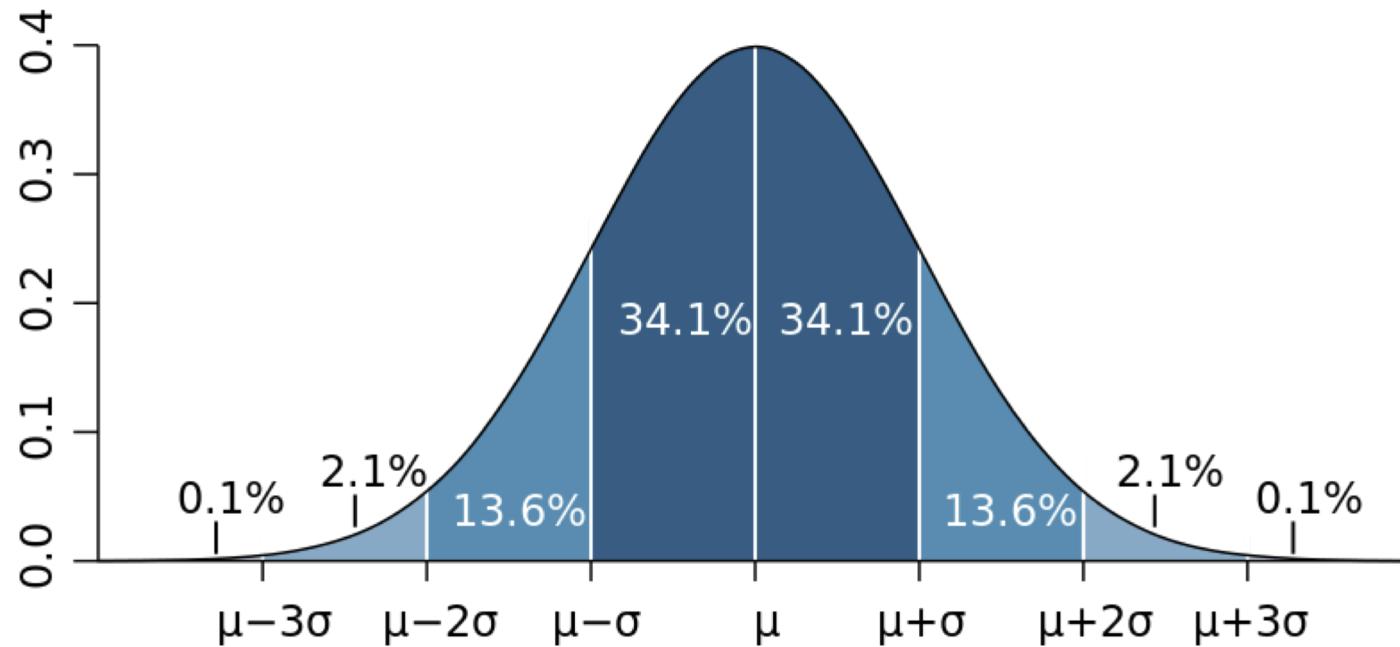


# Правила сигм

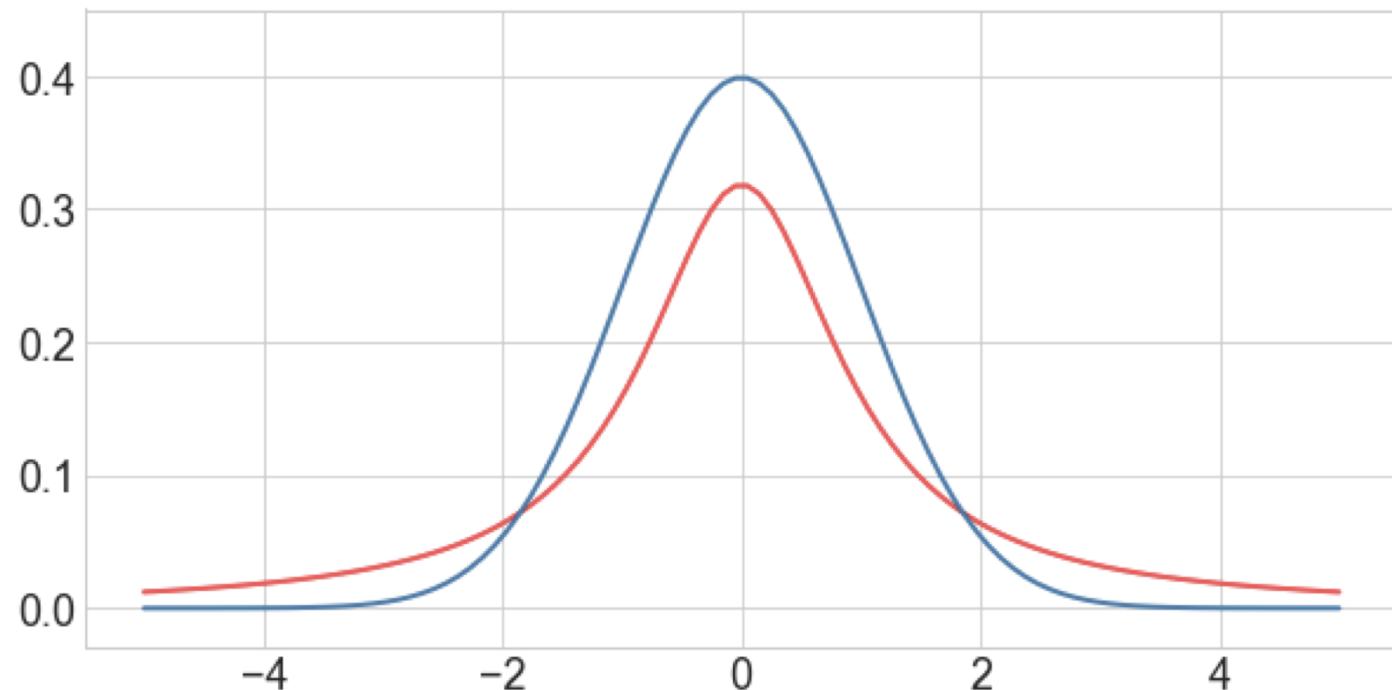
$$X \sim N(\mu, \sigma^2)$$

Правило трех сигм:

$$\mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$$



# Тяжёлые хвосты



- Хвосты красного распределения тяжёлые
- Под ними сосредоточена большая вероятностная масса
- События из-под них (выбросы) более вероятны

# Эксцесс и куртосис

Эксцессом случайной величины  $X$  называют величину

$$\beta_X = \frac{\mathbb{E}[(X - \mathbb{E}(X))^4]}{\sigma^4} - 3$$

Куртосис

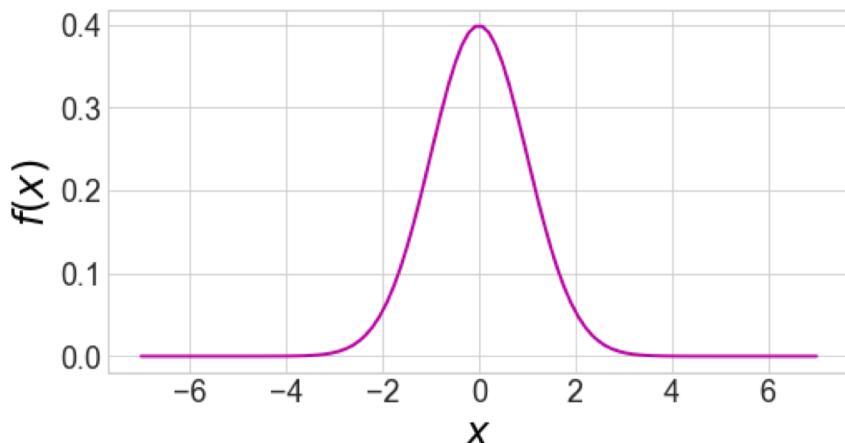
- Число 3 вычитается из куртосиса, чтобы эксцесс нормального распределения был равен нулю
- Если хвосты распределения легче, а пик острее, чем у нормального распределения, тогда  $\beta_X > 0$
- Если хвосты распределения тяжелее, а пик более приплюснутый, тогда  $\beta_X < 0$

# Эксцесс и куртосис

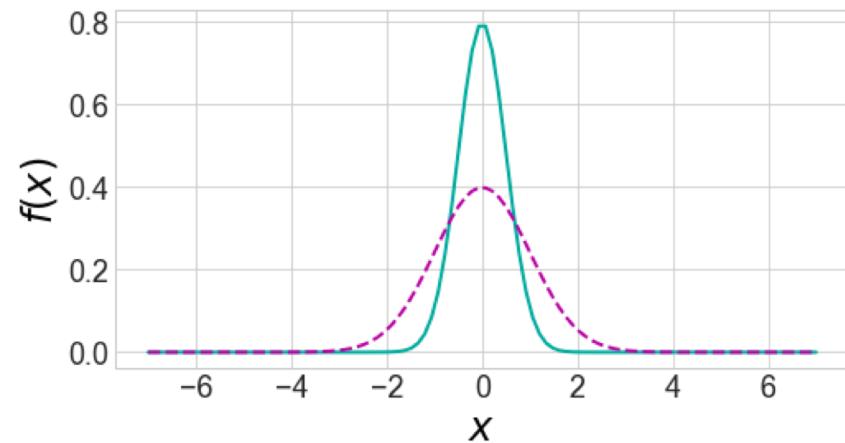
Эксцессом случайной величины  $X$  называют величину

$$\beta_X = \frac{\mathbb{E}[(X - \mathbb{E}(X))^4]}{\sigma^4} - 3$$

Куртосис



Нормальное распределение  
с нулевым эксцессом



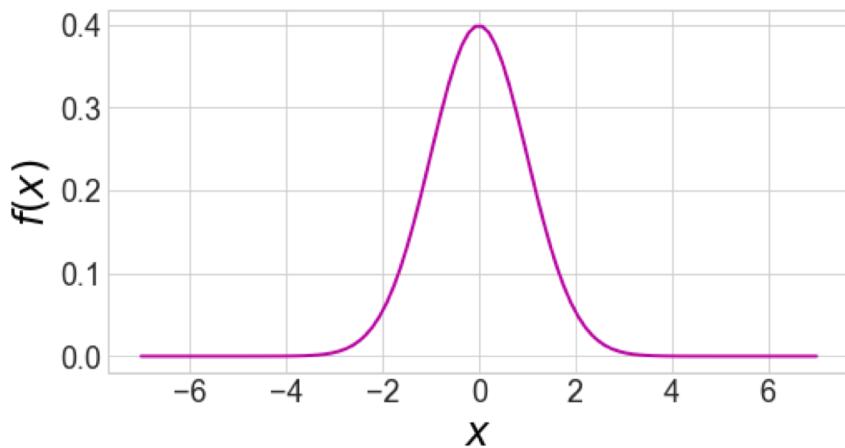
Положительный эксцесс

# Эксцесс и куртосис

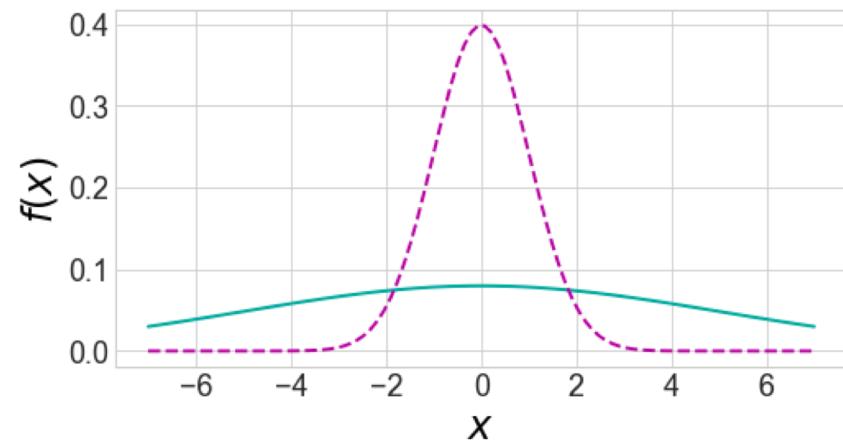
Эксцессом случайной величины  $X$  называют величину

$$\beta_X = \frac{\mathbb{E}[(X - \mathbb{E}(X))^4]}{\sigma^4} - 3$$

Куртосис



Нормальное распределение  
с нулевым эксцессом



Отрицательный эксцесс

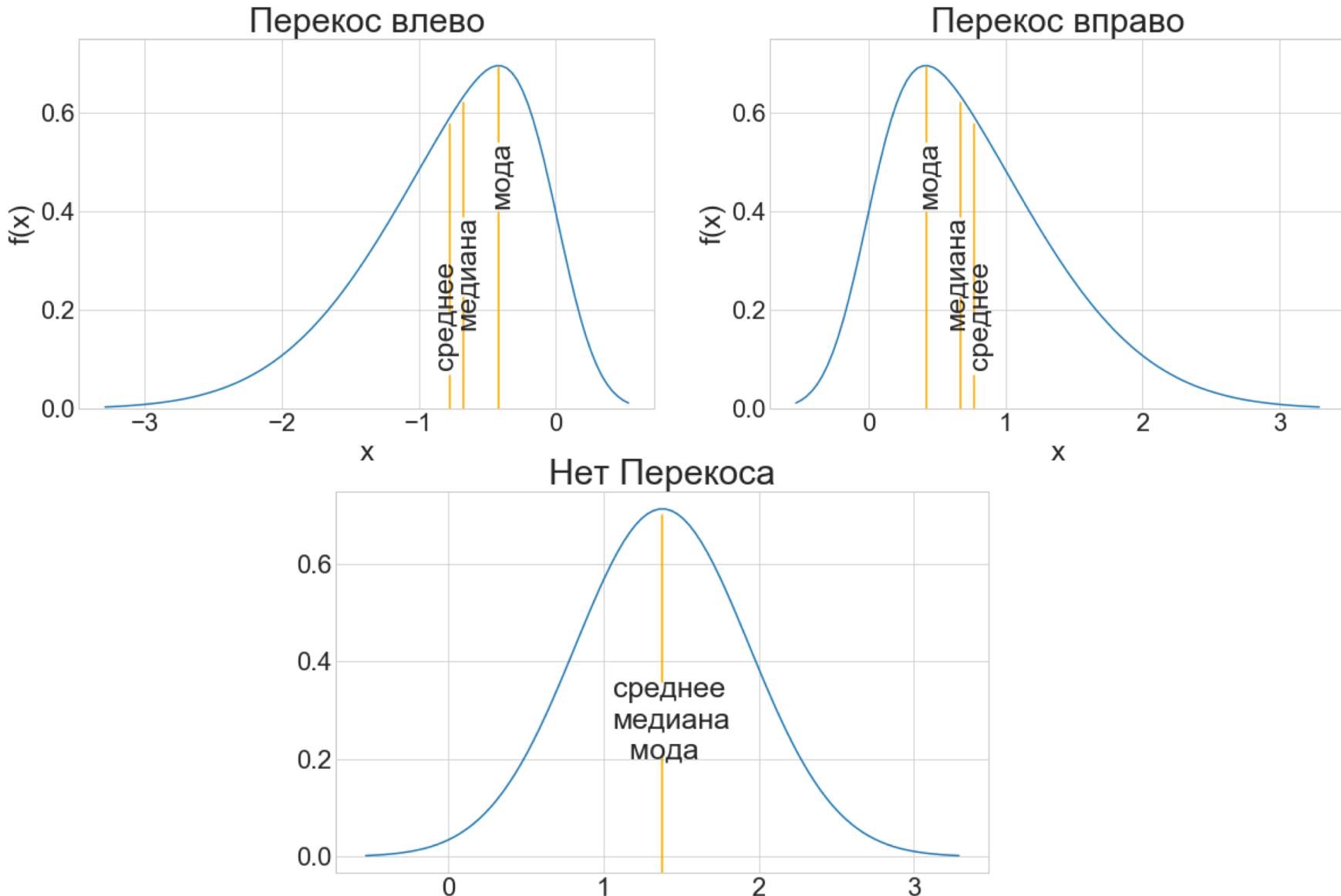
# Коэффициент асимметрии (skewness)

Коэффициентом асимметрии случайной величины  $X$  называют величину

$$A_X = \frac{\mathbb{E}[(X - \mathbb{E}(X))^3]}{\sigma^3}$$

- Если плотность распределения симметрична, то  $A_X = 0$
- Если левый хвост тяжелее, то  $A_X > 0$
- Если правый хвост тяжелее, то  $A_X < 0$

# Коэффициент асимметрии (skewness)



# Эксцесс и асимметрия

- Эксцесс оказывается полезным при поиске тяжёлых хвостов
- Большое значение эксцесса сигнализирует о наличии тяжёлых хвостов и выбросов в данных
- Коэффициент асимметрии характеризует перекос в распределении
- Если у распределения сильный перекос, с применением стандартных статистических методов возникают сложности

# Многомерное нормальное

# Многомерное нормальное

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix} \right]$$

$$X \sim N(\mu, \Sigma)$$



Математическое  
ожидание

$$\mathbb{E}(X_1) = \mu_1$$

$$\mathbb{E}(X_2) = \mu_1$$

# Многомерное нормальное

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix} \right]$$

$$X \sim N(\mu, \Sigma)$$



Ковариационная  
матрица

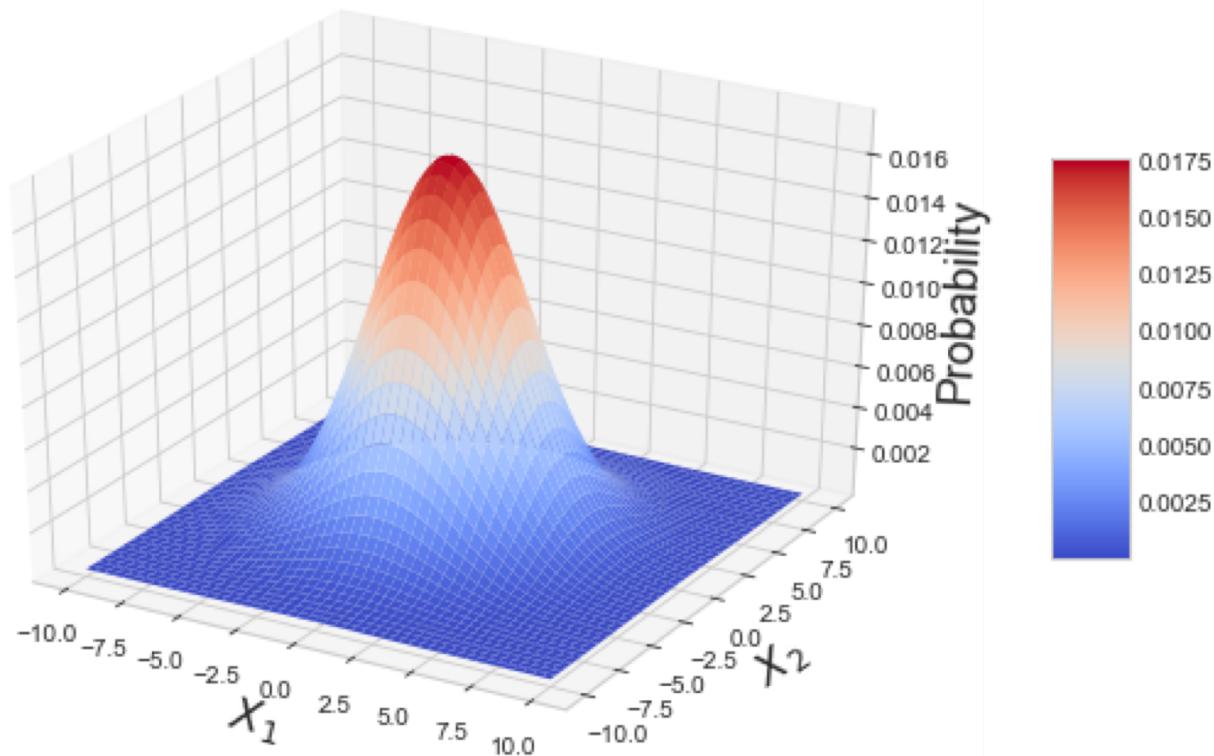
$$Var(X_1) = \sigma_1^2$$

$$Var(X_2) = \sigma_2^2$$

$$Cov(X_1, X_2) = \rho$$

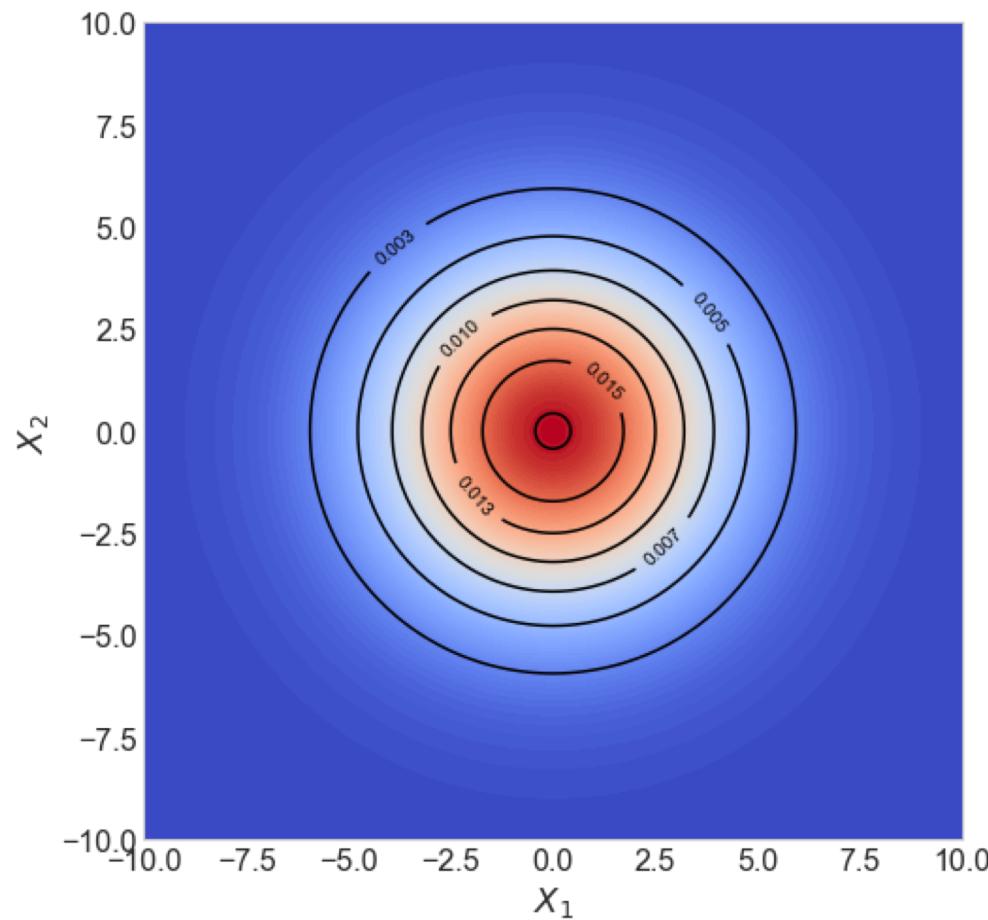
# Многомерное нормальное

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix} \right]$$



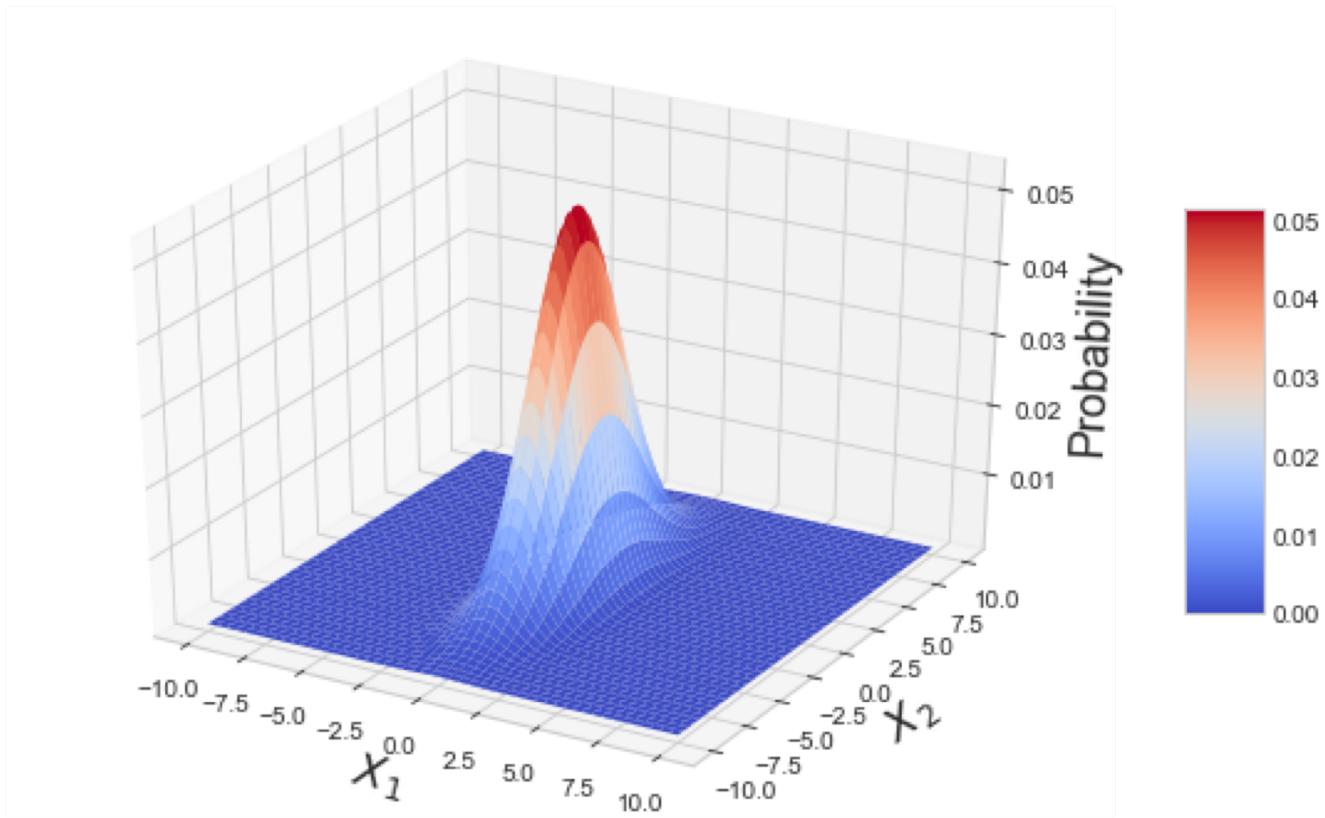
# Многомерное нормальное

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix} \right]$$



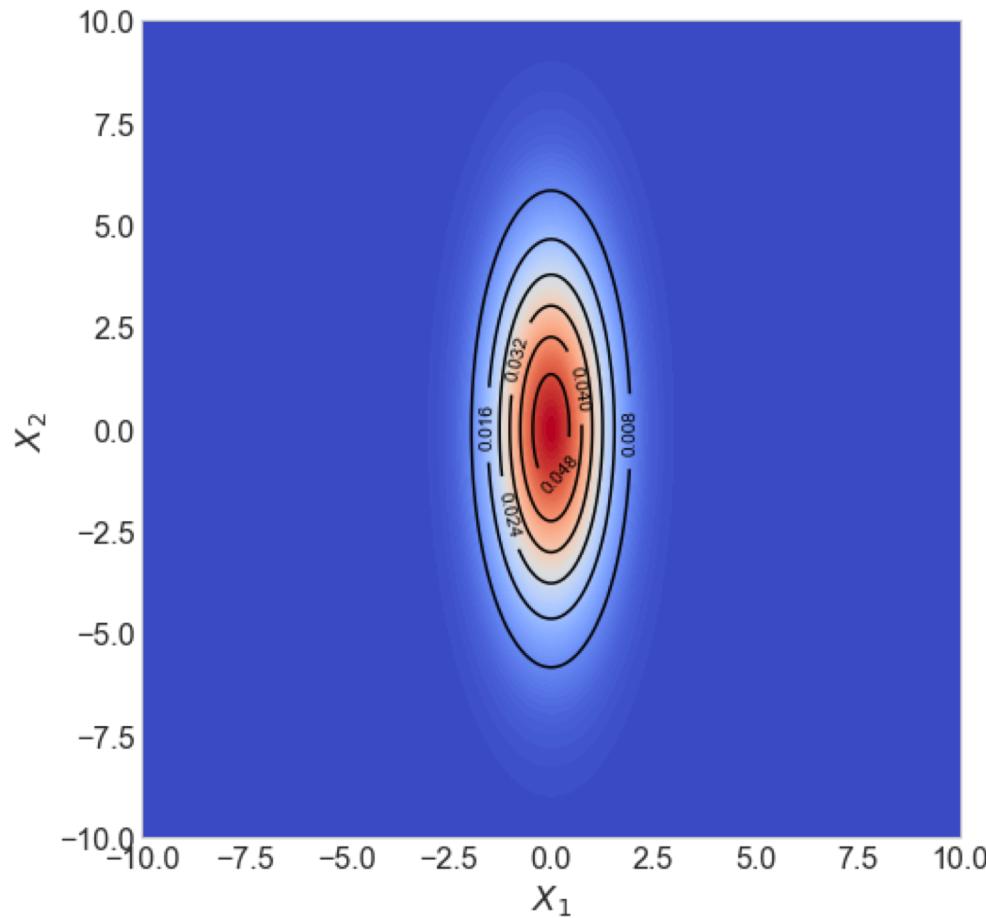
# Многомерное нормальное

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} \right]$$



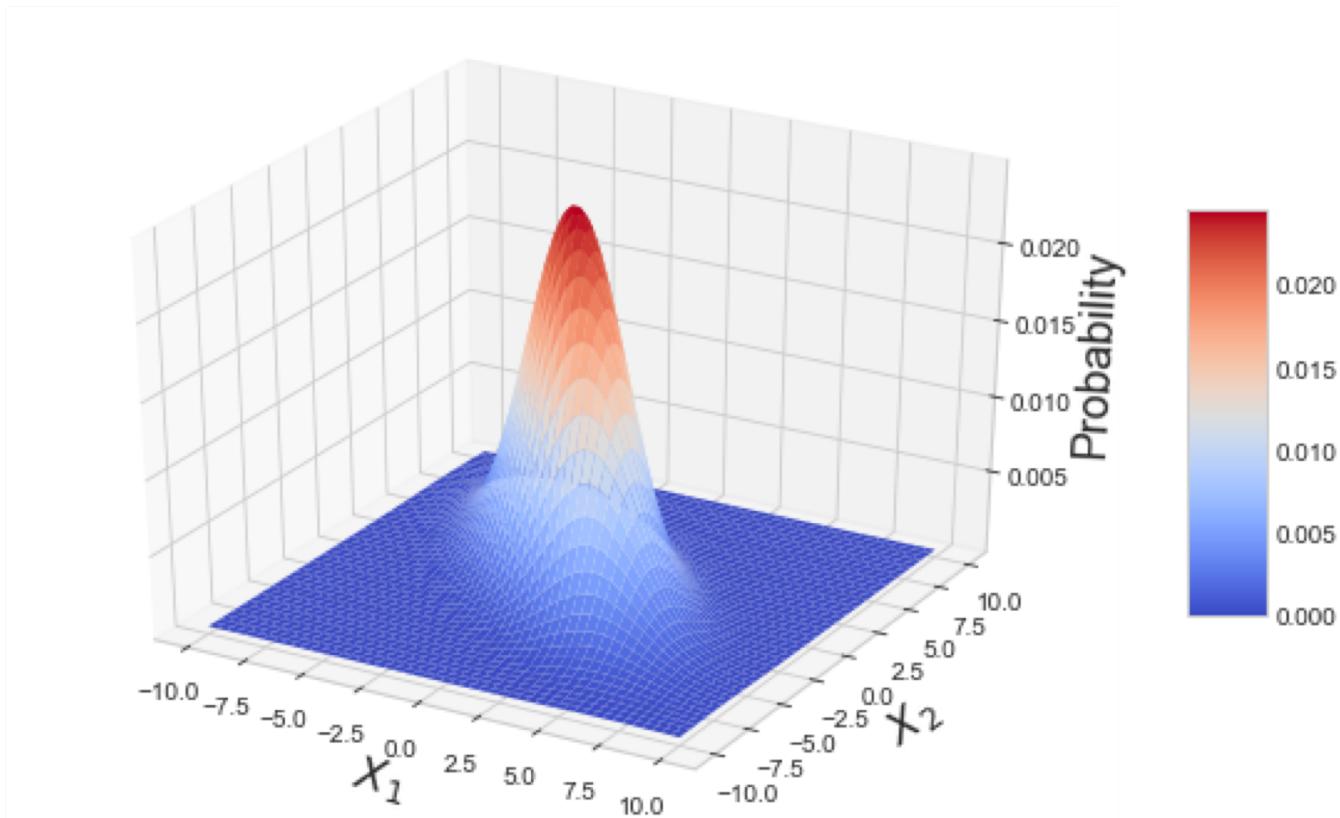
# Многомерное нормальное

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} \right]$$



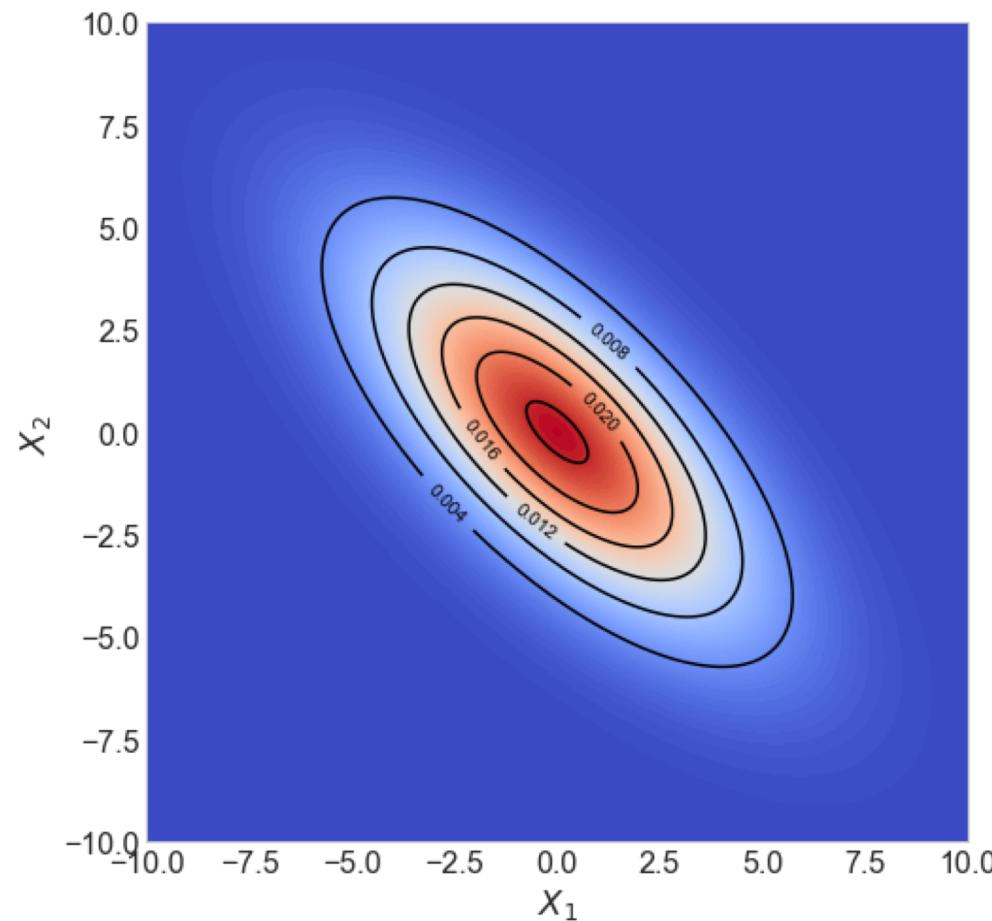
# Многомерное нормальное

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 9 & -6.3 \\ -6.3 & 9 \end{pmatrix} \right]$$



# Многомерное нормальное

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 9 & -6.3 \\ -6.3 & 9 \end{pmatrix} \right]$$



# Многомерное нормальное

- По аналогии можно определить нормальное распределение для любой размерности

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N(\mu, \Sigma)$$

$$\mu = \begin{pmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \mathbb{E}(X_3) \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Var(X_2) & Cov(X_2, X_3) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var(X_3) \end{pmatrix}$$

# Резюме

- Нормальное распределение довольно часто встречается на практике
- Важно научится хорошо с ним уметь работать

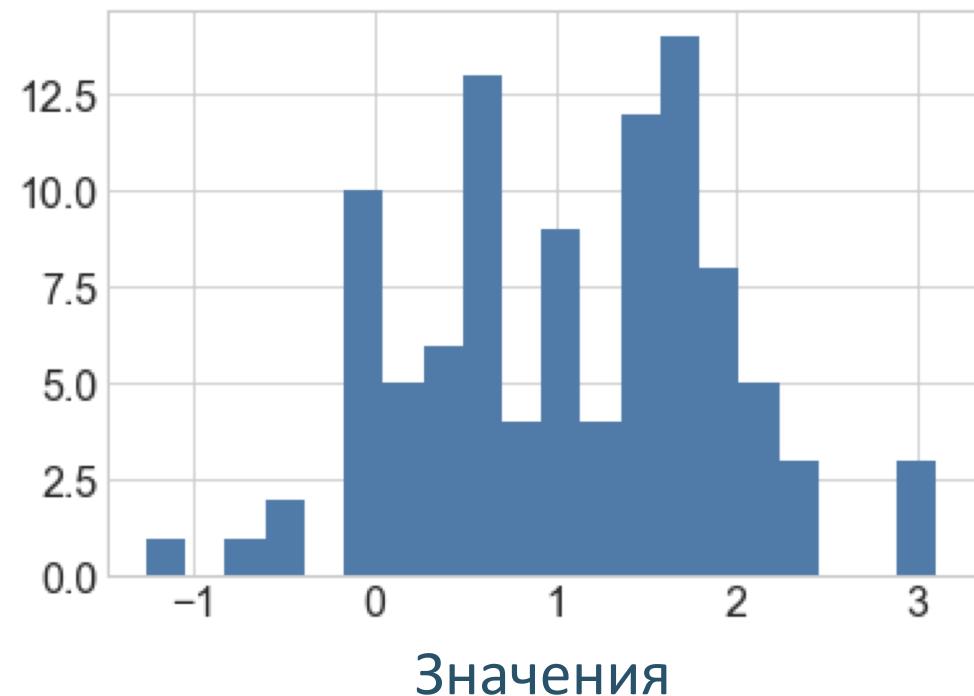
# **Ядерные оценки плотности**

# Гистограмма

Гистограмма – простейший непараметрический способ получить оценку плотности распределения

Непараметрический, так как не выдвигается никаких предположений о виде распределения

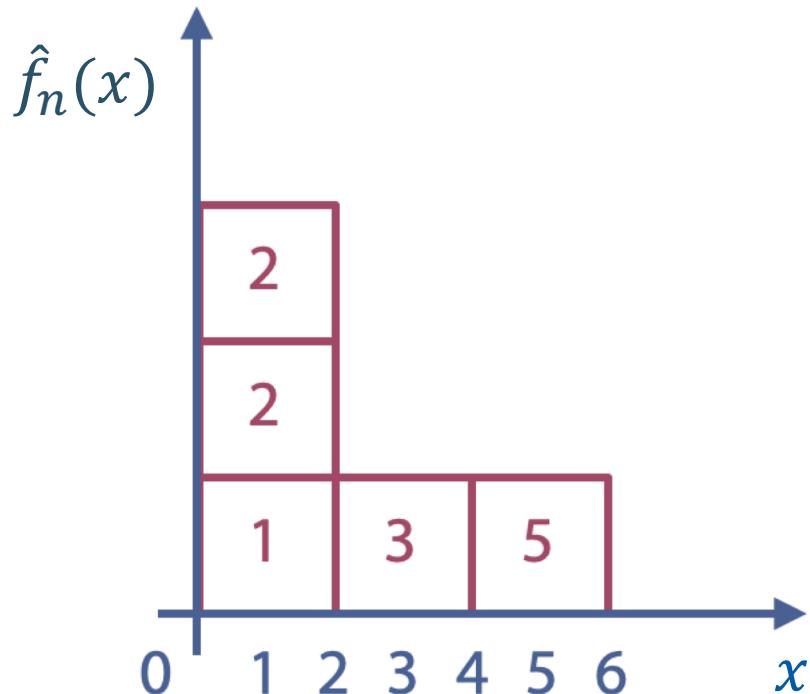
Сколько значений попали в текущий отрезок (бин)



# Гистограмма

Пример:  $x_1 = 2, x_2 = 5, x_3 = 2, x_4 = 3, x_5 = 1$

$$\hat{f}_n(x) = \frac{1}{n \cdot h} \cdot \sum [z_k < x_i \leq z_{k+1}],$$



Скобки – индикаторная  
функция:

[правда] = 1

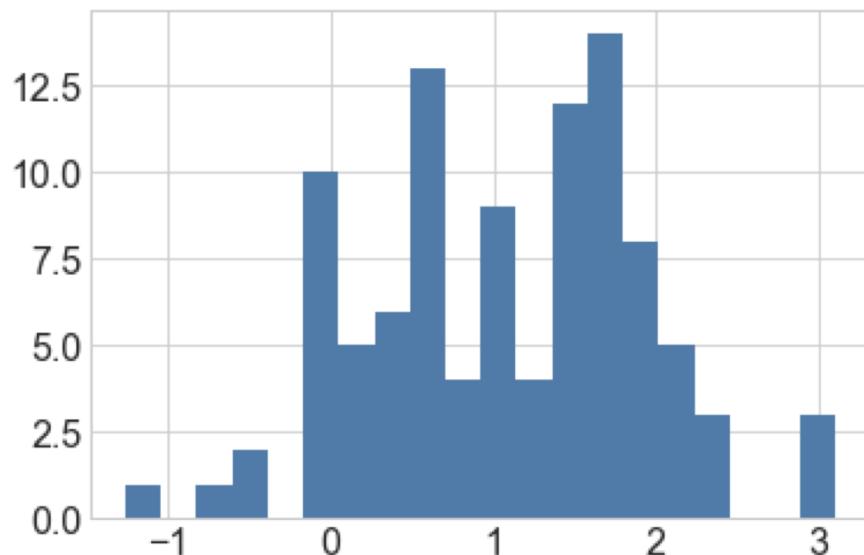
[ложь] = 0

Размер бина (длина отрезка):

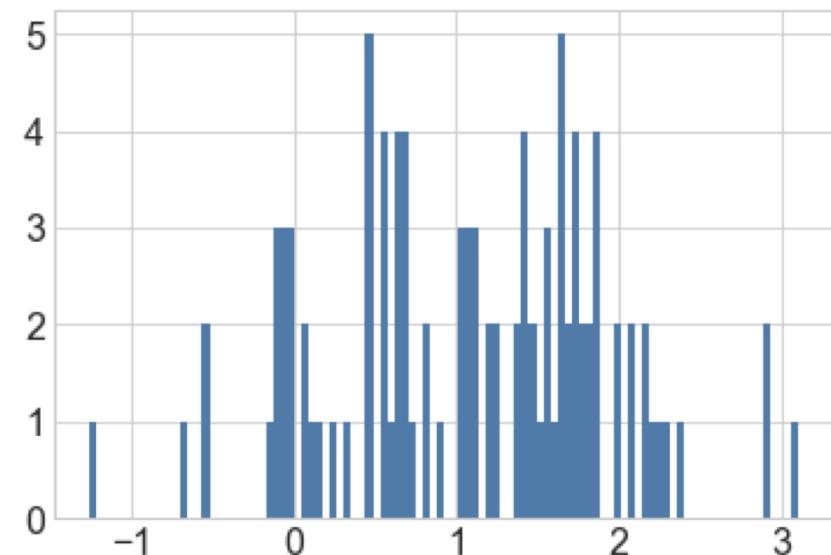
$$h = z_{k+1} - z_k$$

# Гистограмма

- Длина интервала  $h$  (бина) должна быть достаточно большой, чтобы в него попало существенное число наблюдений
- И при этом достаточно малой, чтобы не потерять важные детали распределения



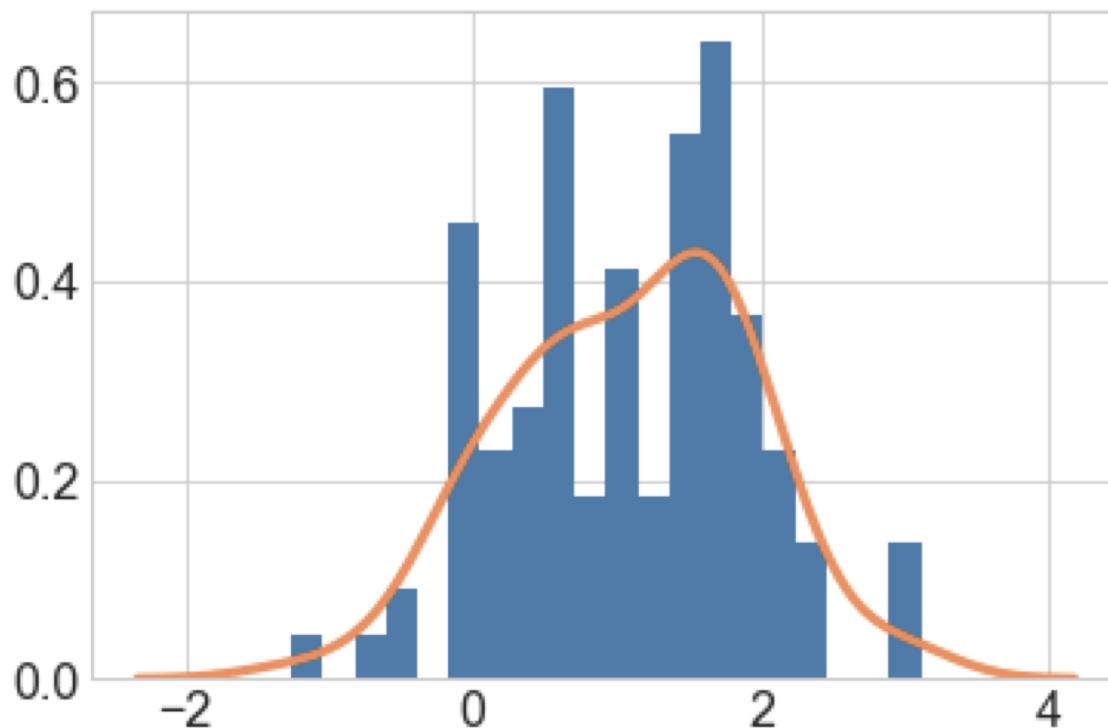
20 бинов



100 бинов

# Ядерные оценки плотности

- Ядерные оценки плотности позволяют получить график плотности в виде непрерывной кривой



# Ядерные оценки плотности

Гистограмма:

$$\hat{f}_n(x) = \frac{1}{n \cdot h} \cdot \sum [z_k < x_i \leq z_{k+1}]$$

↑  
↑  
границы  
фиксированы

Улучшение:

$$\hat{f}_n(x) = \frac{1}{n \cdot h} \cdot \sum \left[ x - \frac{h}{2} < x_i \leq x + \frac{h}{2} \right]$$

↑  
↑  
скользящие  
границы

$h$  – ширина окна

# Ядерные оценки плотности

$$\hat{f}_n(x) = \frac{1}{n \cdot h} \cdot \sum \left[ x - \frac{h}{2} < x_i \leq x + \frac{h}{2} \right]$$

- Перепишем оценку в более удобном виде:

$$\hat{f}_n(x) = \frac{1}{n \cdot h} \cdot \sum K\left(\frac{x - x_i}{h}\right) \quad K(z) = \left[-\frac{1}{2} < z \leq \frac{1}{2}\right]$$

- Такая функция придаёт каждому наблюдению вес либо 0 либо 1
- Чем дальше наблюдение от центра, тем меньше должен быть его вес, нужна другая функция

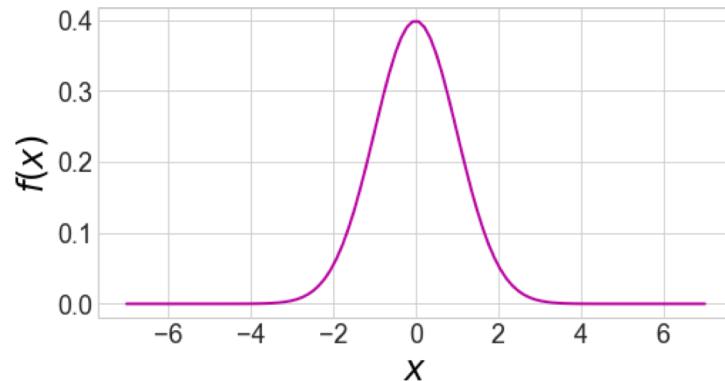
# Ядерные оценки плотности

Чтобы взвесить наблюдения, функцию  $K(z)$  (ядерную функцию) выбирают так, чтобы:

- Она была неотрицательной
- $\int K(z) dz = 1$  (сумма всех весов равна 1)

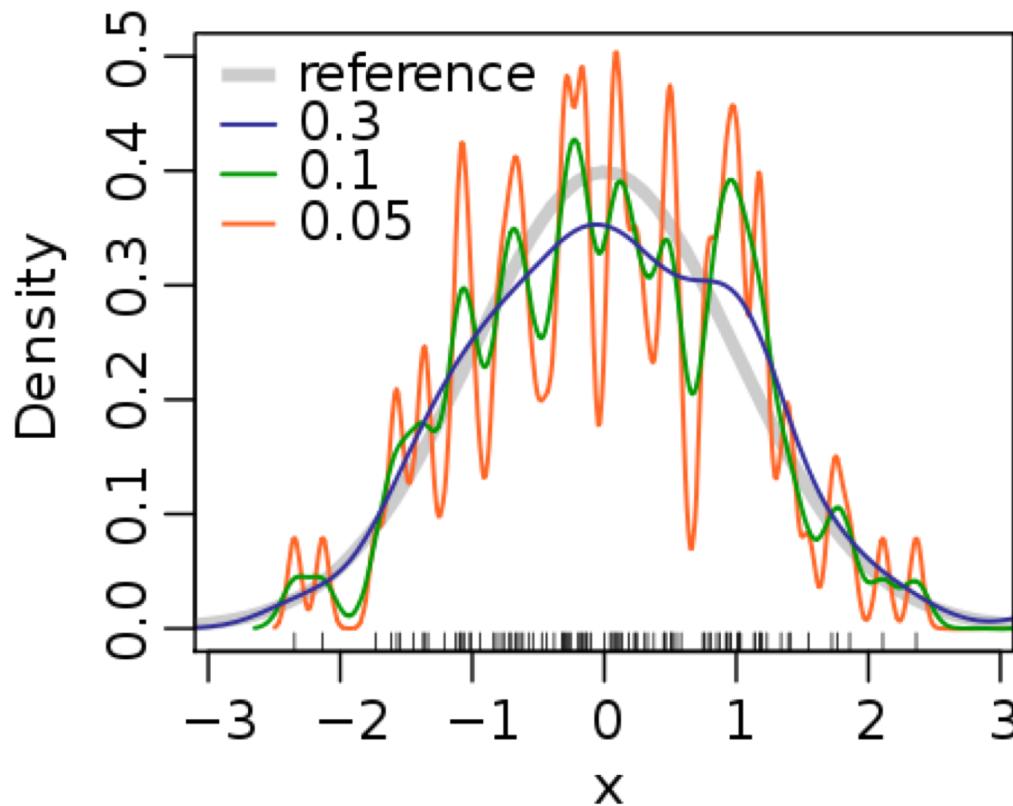
Ядерные функции бывают разными, чаще всего используют Гауссовское ядро:

$$K(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$$



# Пример

- Серые чёрточки на оси x – наблюдения
- Величина параметра  $h$  (ширина окна) влияет на то, насколько гладкой получается итоговая кривая



# Резюме

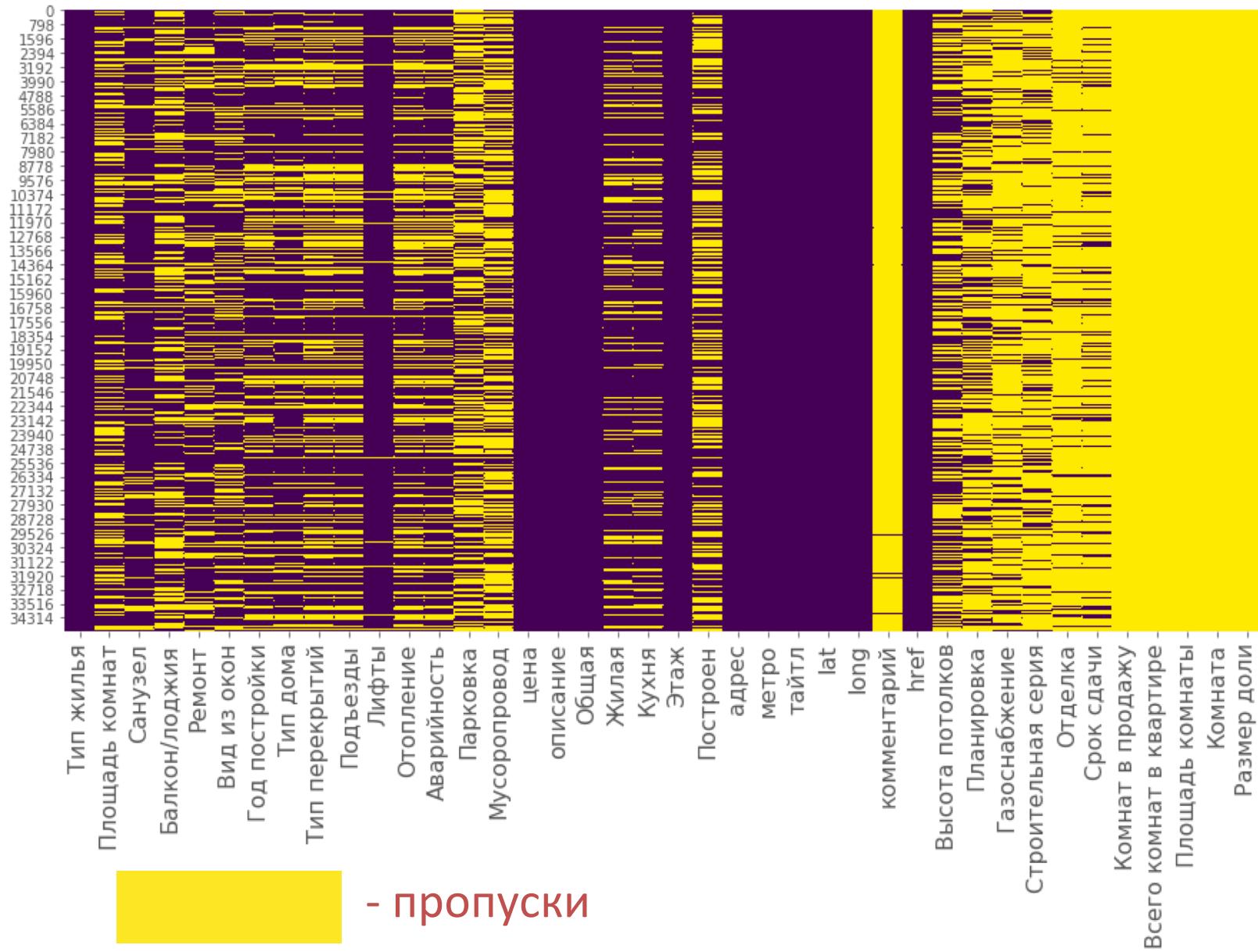
- Простейший способ построить непараметрическую оценку плотности распределения – гистограмма
- Для того, что получить непрерывную оценку плотности распределения используют ядерное сглаживание

# Проблемы с данными: пропуски и выбросы

# Пропуски в данных

- Большинство реальных данных имеют пропущенные значения
  - Ошибки при записи или измерении
  - Невозможность сбора данных

# Пропуски в данных



# Борьба с пропусками

- Удаление объектов с пропущенными значениями (строки)
- Удаление признаков с большим числом пропусков (столбцы)
- Такая стратегия может привести к проблемам:
  - У нас останется очень мало данных
  - В данных может возникнуть искажение (смещение)

**Пример:** среди пациентов масса измеряется только у тех, у кого высокое давление

# Борьба с пропусками

- Чтобы не возникало искажений, нужно понимать откуда, скорее всего, возникли пропуски
- Пропуски нужно как-то заполнить

# Борьба с пропусками

## Простые методы:

Замена средним значением / медианой / модой

## Сложные методы:

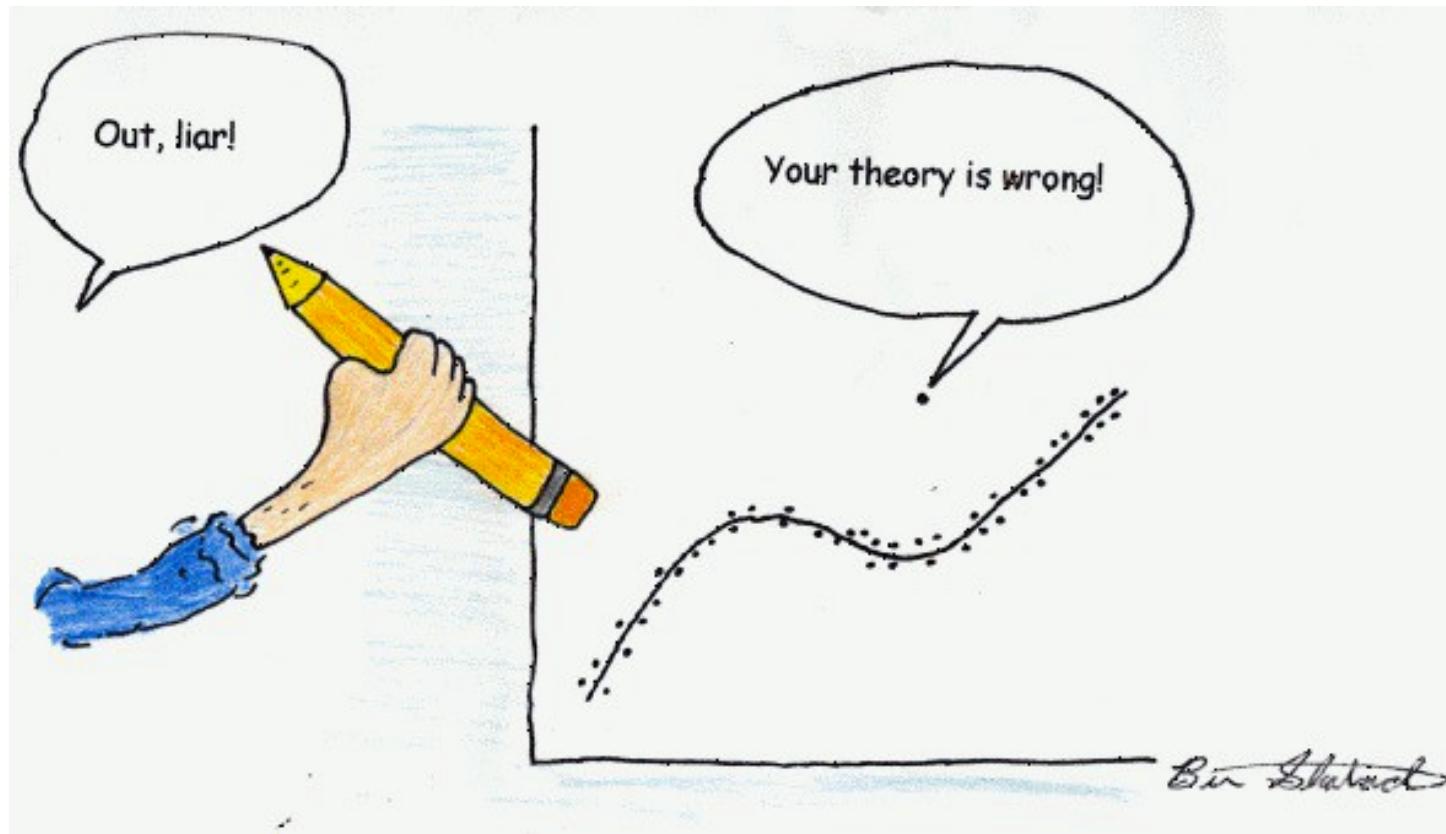
Основаны на машинном обучении, смотрят на другие примеры и пытаются предсказать что было пропущено



Мы чаще всего будем пользоваться простыми способами заполнения пропусков

# Выбросы (outliers)

Выброс – результат измерений, который сильно выделяется на общем фоне



# Проблемы из-за выбросов

Многие алгоритмы  
чувствительны к выбросам  
и переобучаются под них

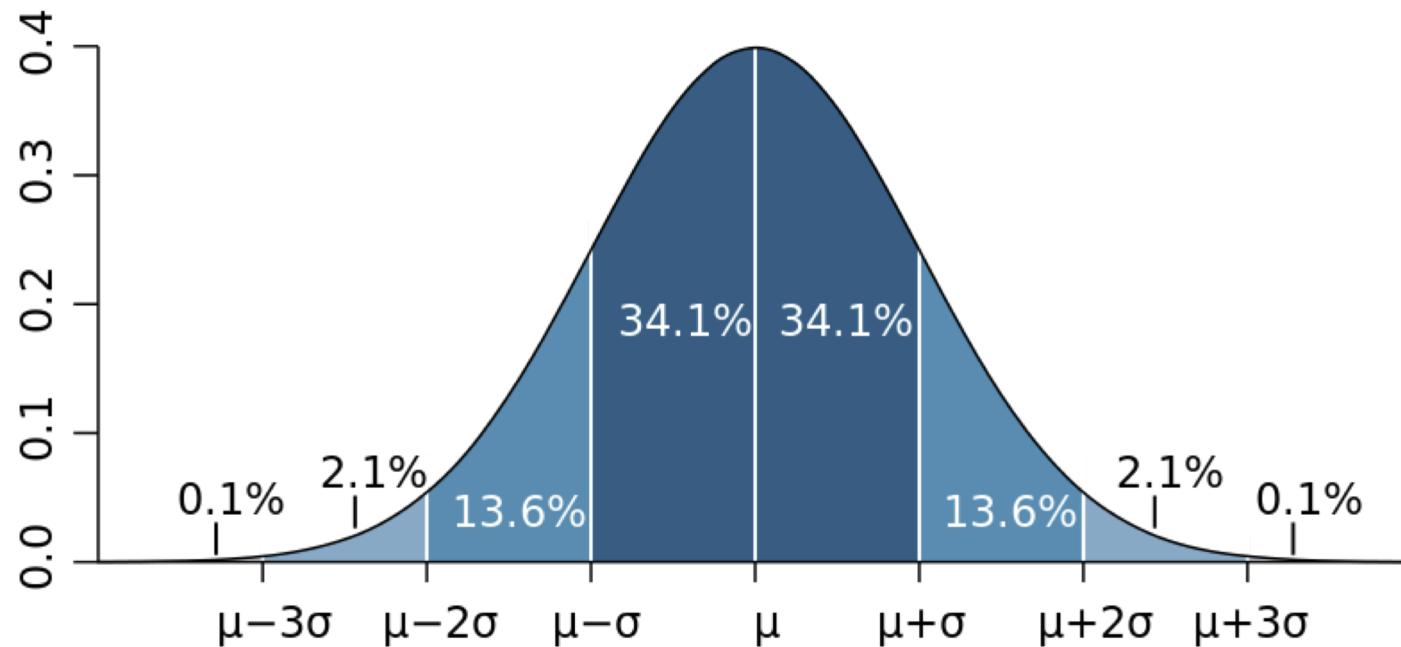
- **Пример:** среднее часто используют в качестве наивного прогноза
- С ним сравнивают насколько хорошо модель прогнозирует данные
- При наличии выброса итоговая статистика будет искажена



# Поиск выбросов

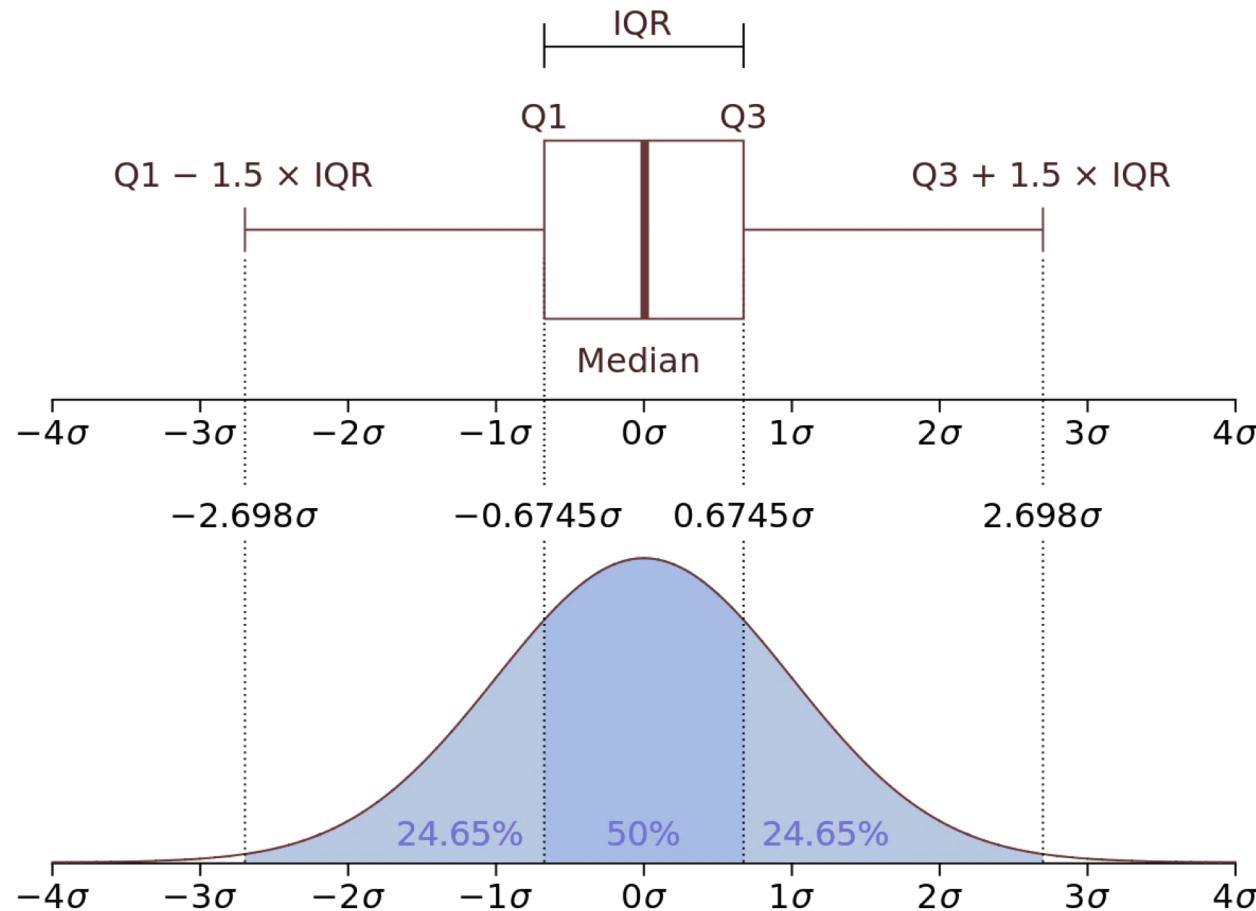
**Правило трёх сигм:** если данные распределены нормально и наблюдение оказалось за пределами интервала в три сигмы, это выброс

$$\mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$$



# Поиск выбросов

**Правило 1.5 интерквантильных размахов (IQR):** если наблюдение оказалось за пределами выделенного интервала, оно выброс. Иногда используют 3 IQR



# Поиск выбросов

- Также выбросы можно искать с помощью различных более сложных алгоритмов машинного обучения
- Многие алгоритмы устойчивы к выбросам
- **Пример:** если бы мы строили наивный прогноз на основе медианы, он был бы устойчивым к выбросам и неискаженным, как среднее
- Можно придумывать статистики, основанные на медиане, которые будут нечувствительны к выбросам

► <https://dyakonov.org/2017/04/19/поиск-аномалий-anomaly-detection/>

# Резюме

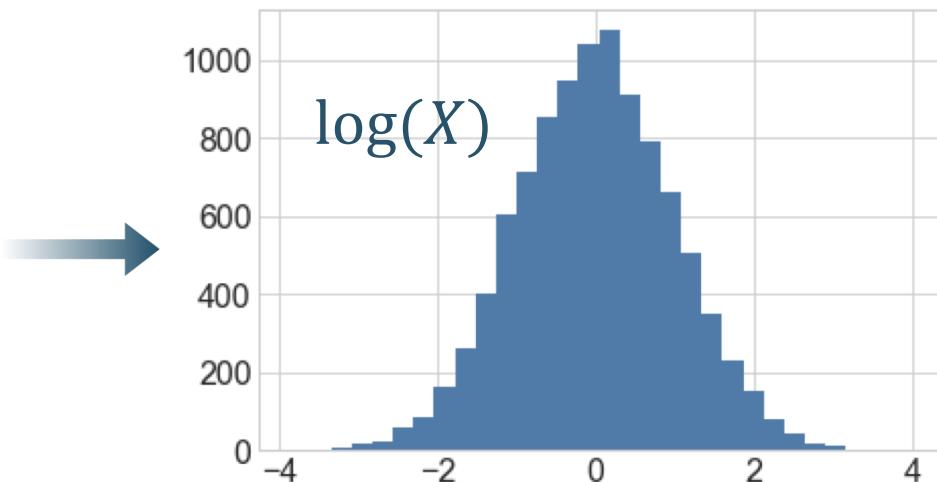
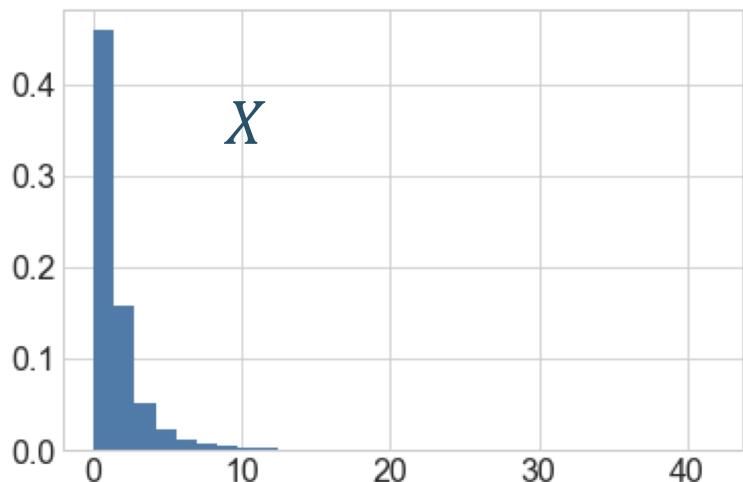
- Пропуски и выбросы – проблемы в данных, с которыми надо бороться
- Пропуски, если их мало, пытаются заполнять с помощью разных алгоритмов
- Выбросы либо сглаживают, либо выбрасывают из рассмотрения

# Преобразование Бокса-Кокса

# Длинные хвосты

- Выбросы связаны с шумом в данных
- Иногда данные имеют не очень удобное распределение (длинные хвосты)
- Из-за этого с ними сложно работать стандартными методами
- Чтобы с ними было удобнее работать, данные можно сгладить

# Логарифмирование

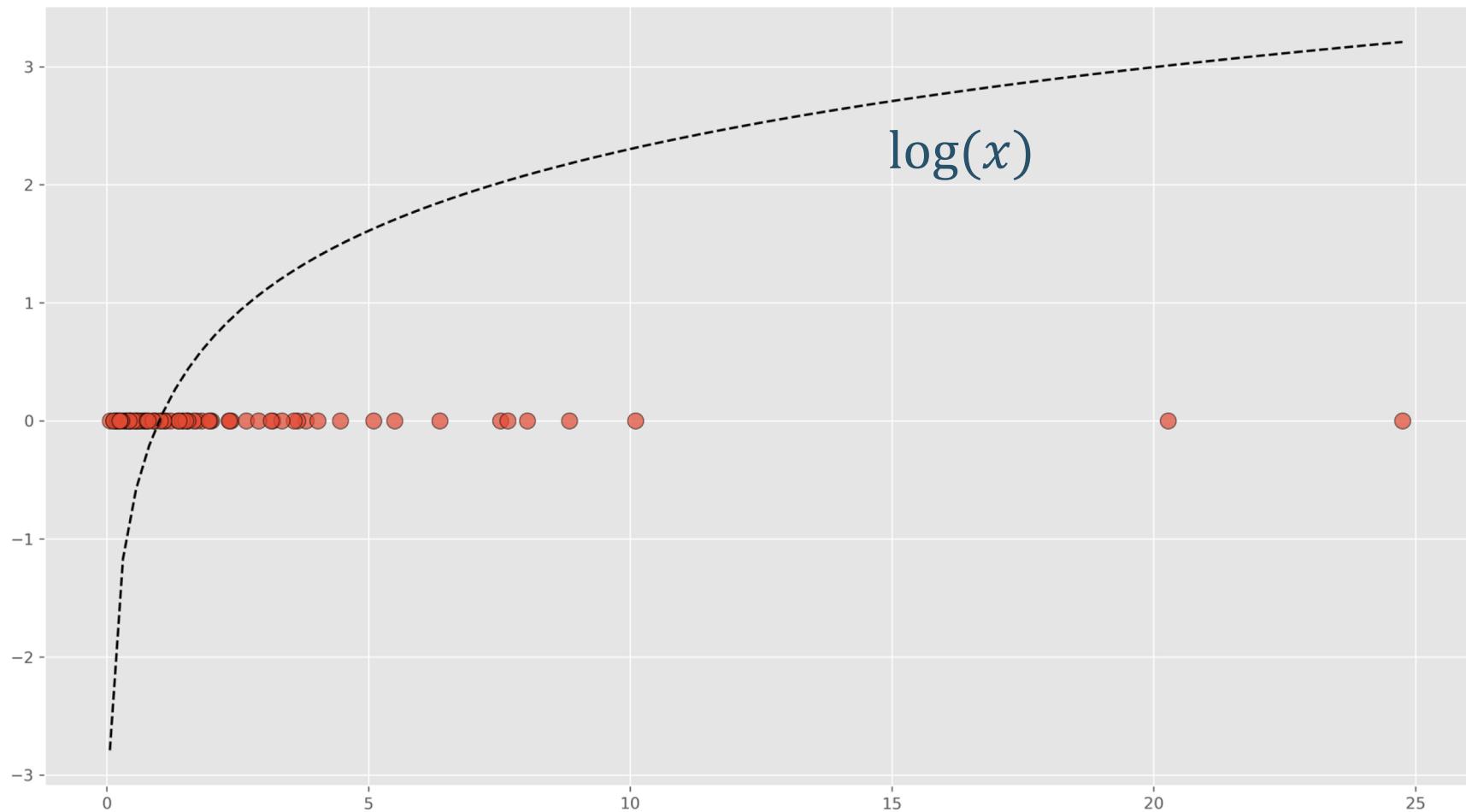


Логарифмирование значений позволяет  
сгладить хвосты

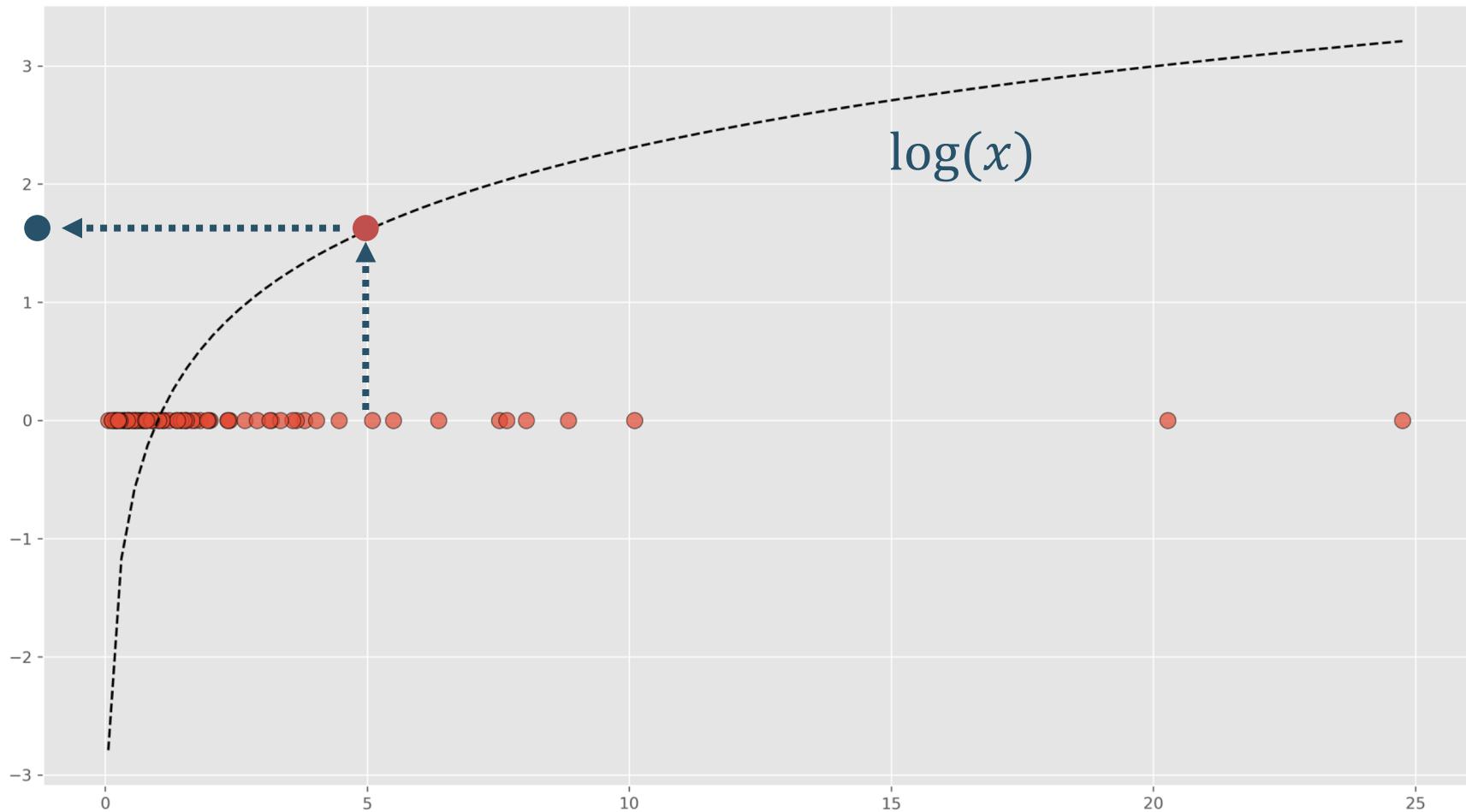
# Логарифмирование и нормальность

- Логарифмирование помогает сгладить длинные хвосты и получить куполообразное распределение
- Если в результате логарифмирования случайной величины получается нормальное распределение, такая случайная величина называется логнормальной
- Часто такой особенностью обладают цены

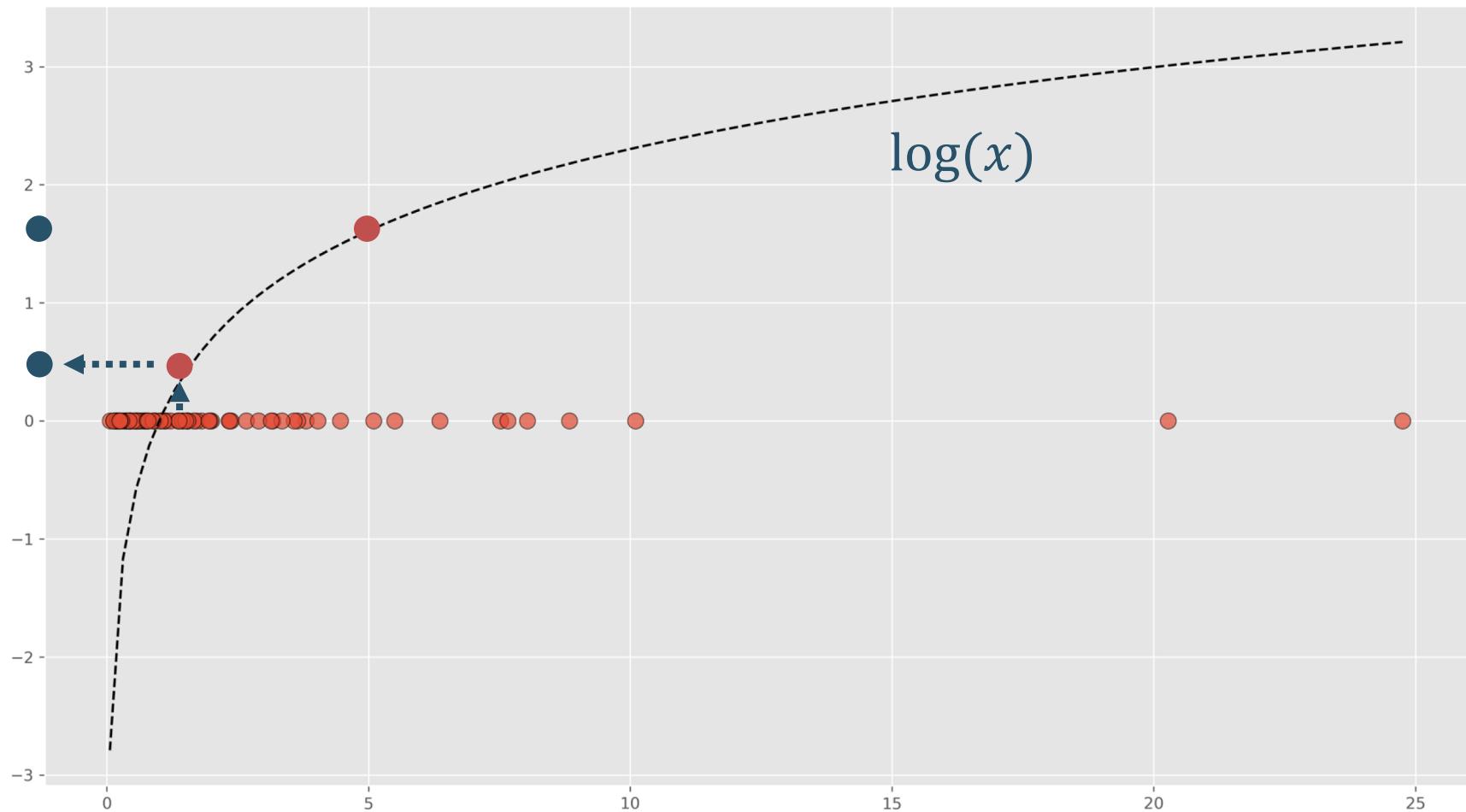
# Логарифмирование



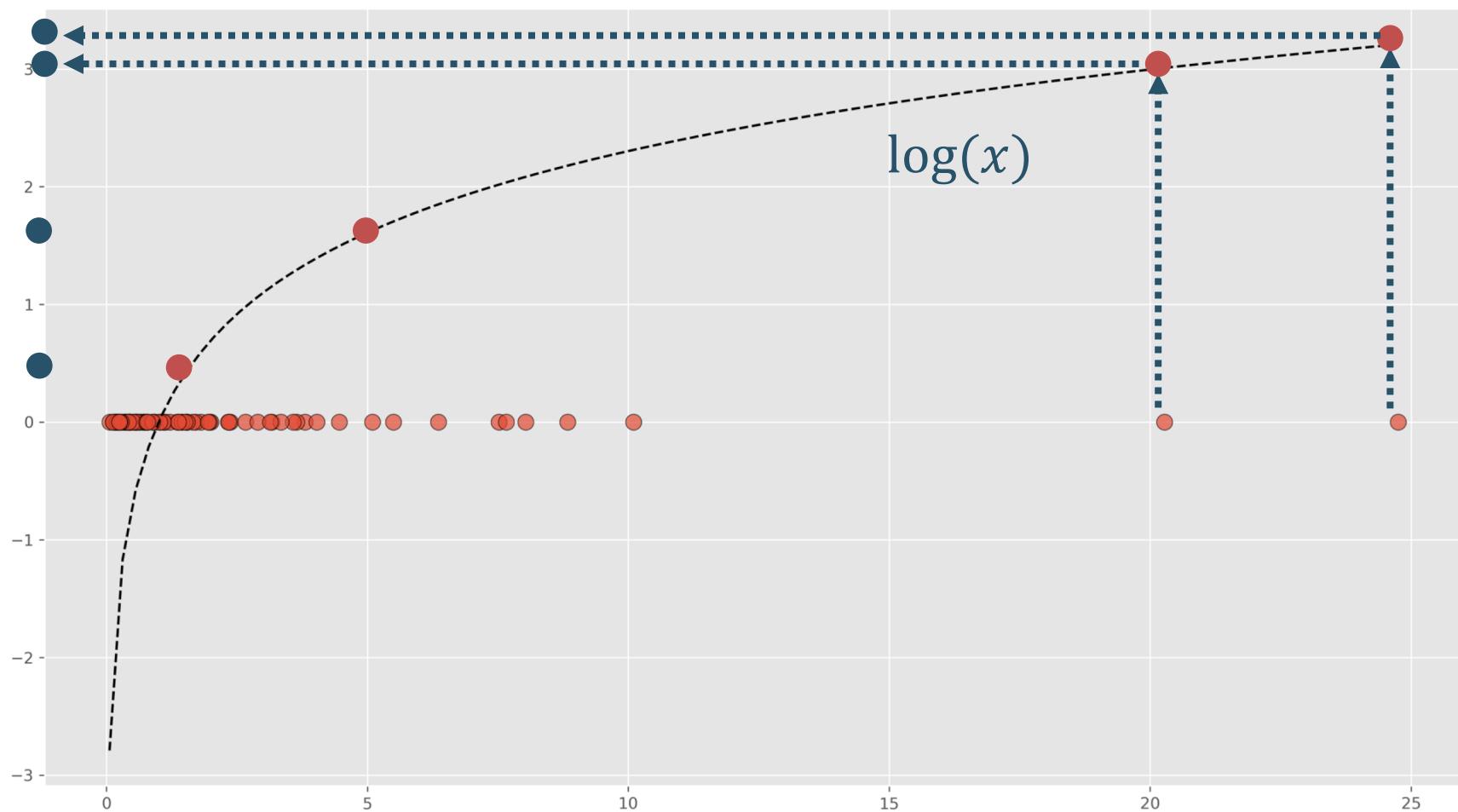
# Логарифмирование



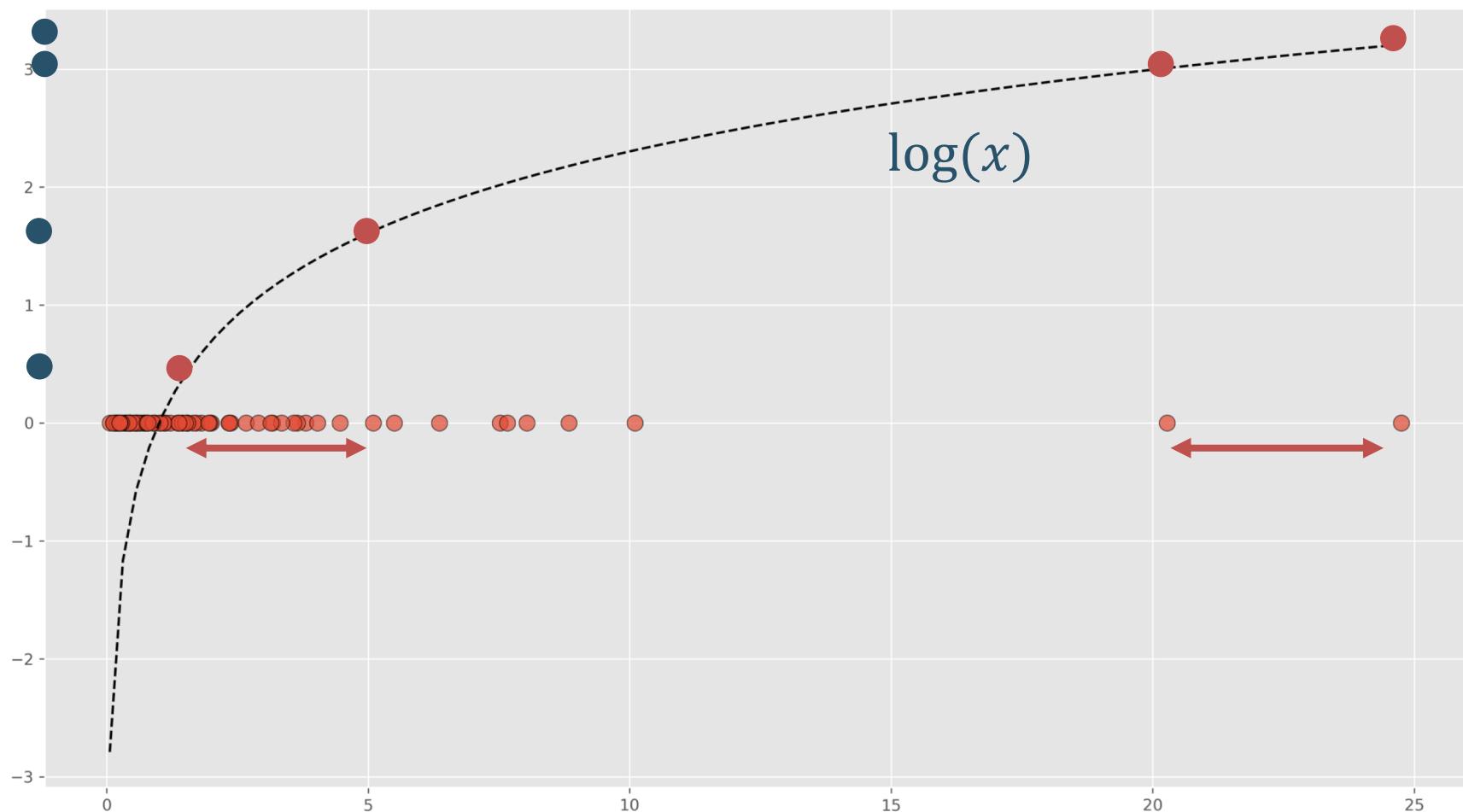
# Логарифмирование



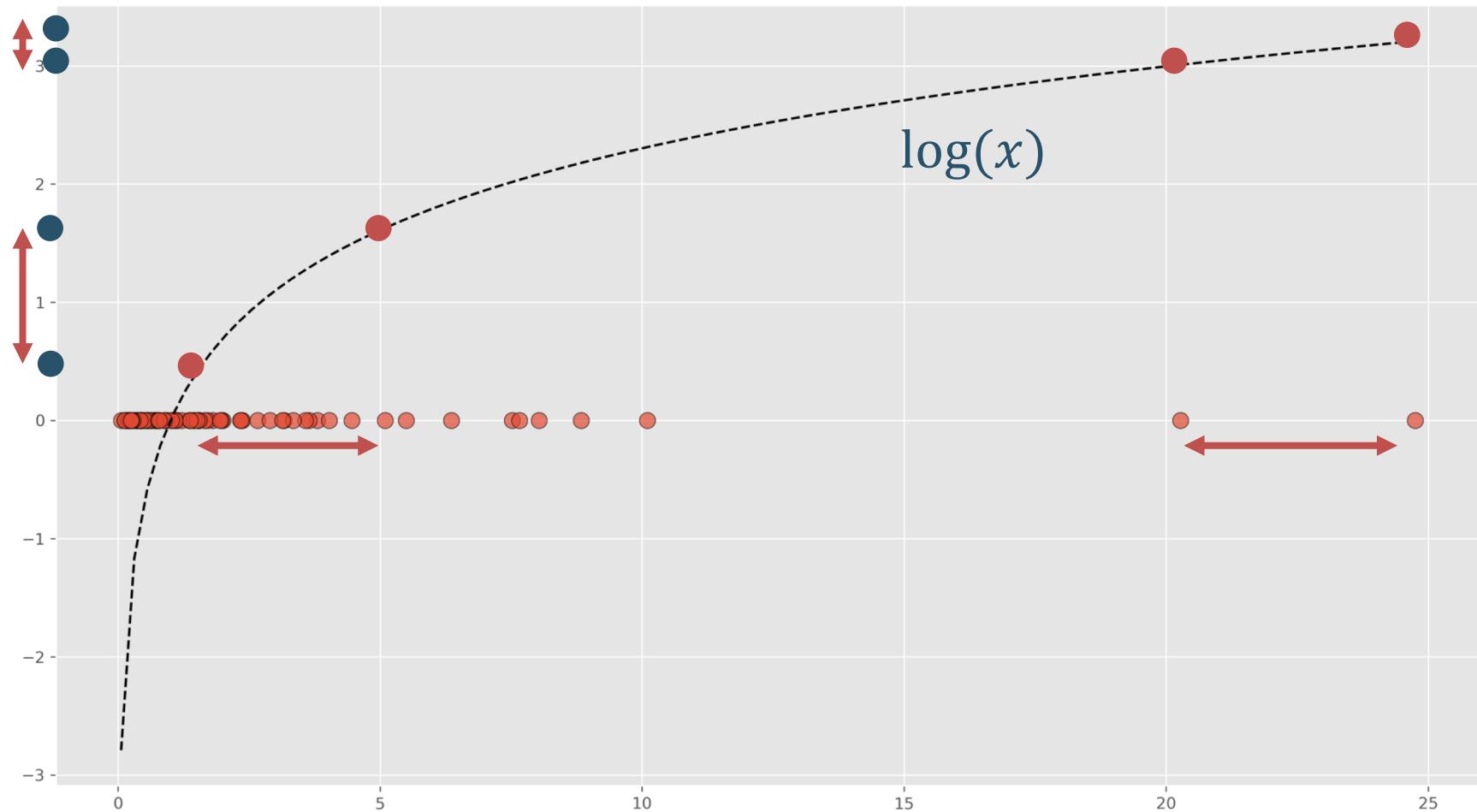
# Логарифмирование



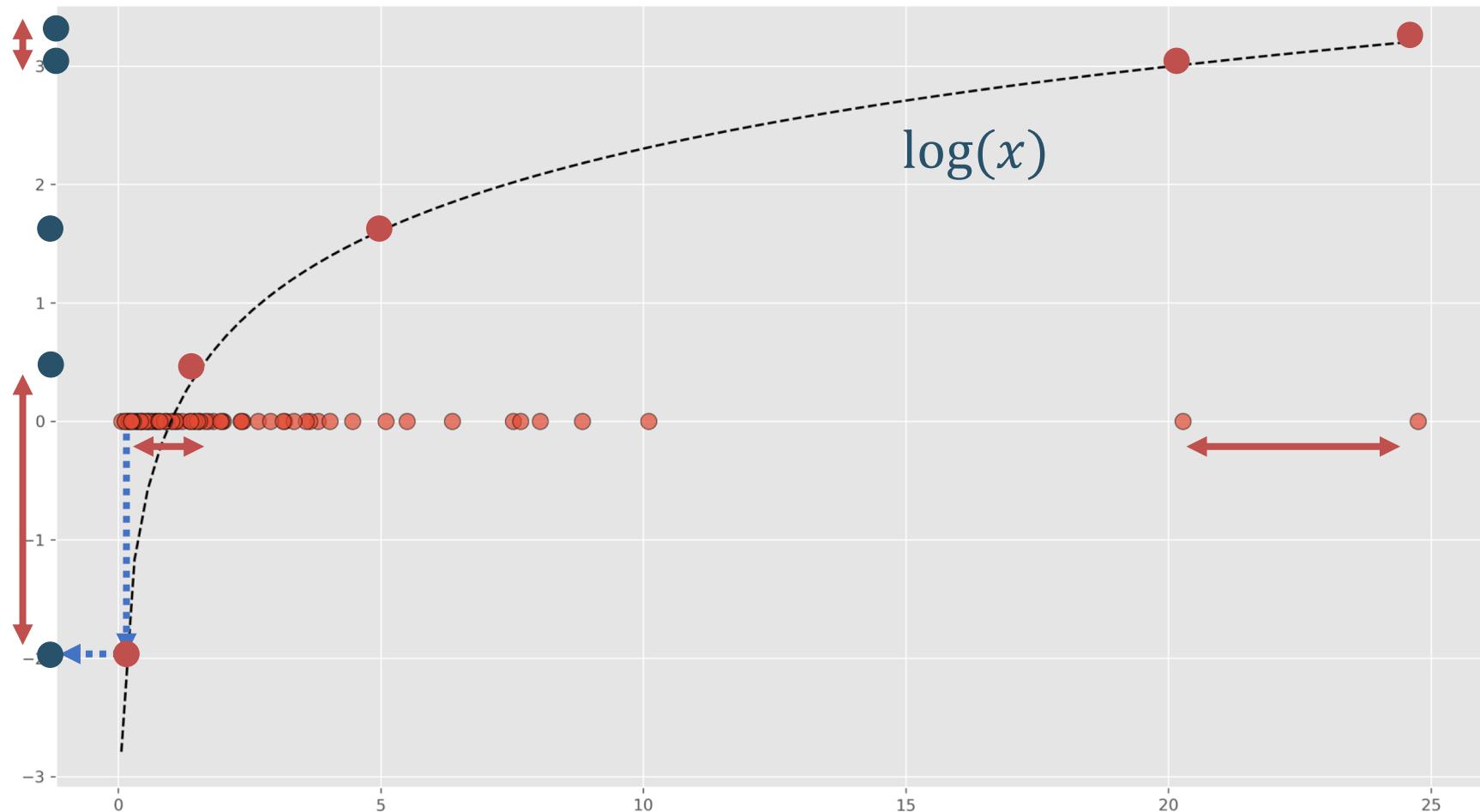
# Логарифмирование



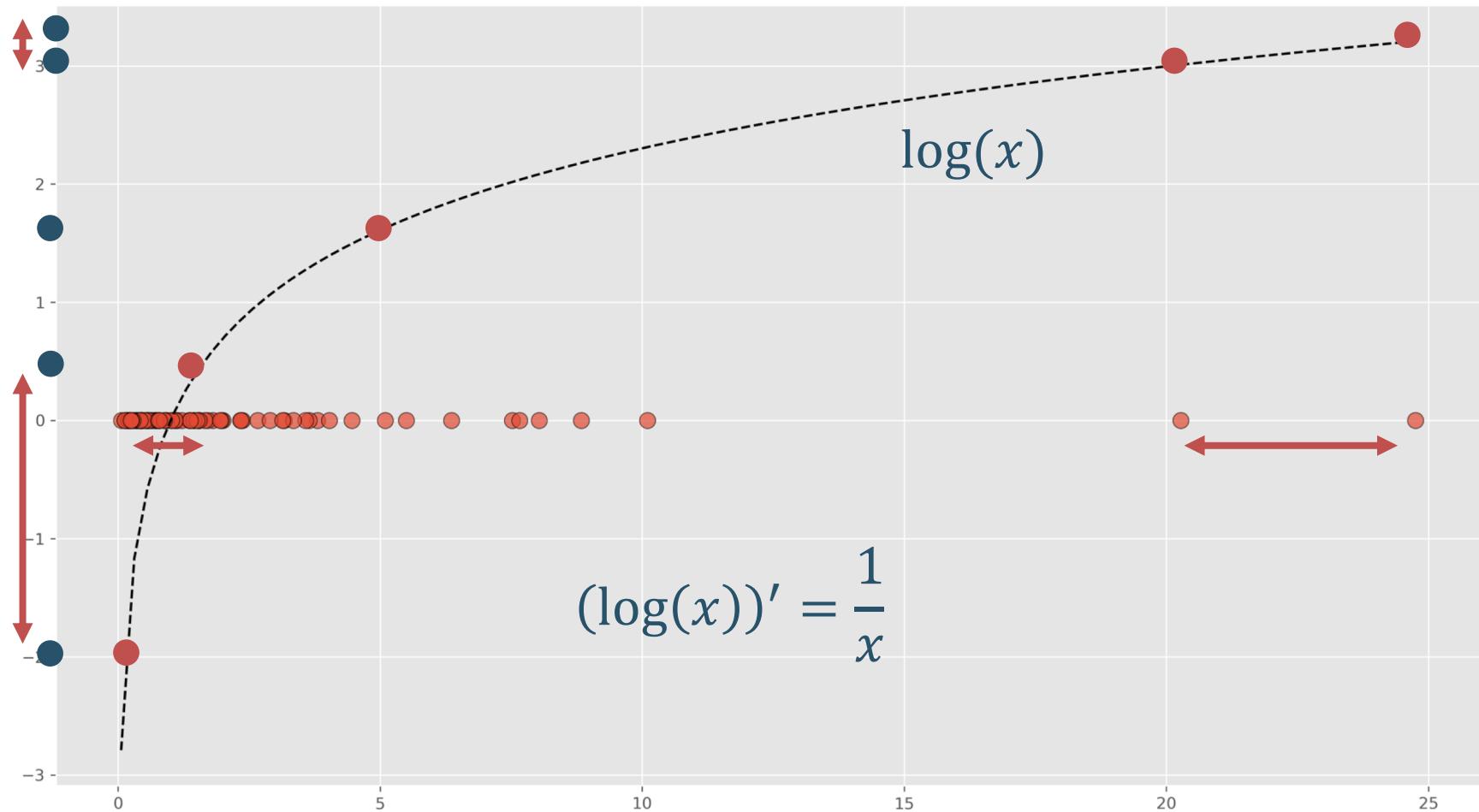
# Логарифмирование



# Логарифмирование



# Логарифмирование

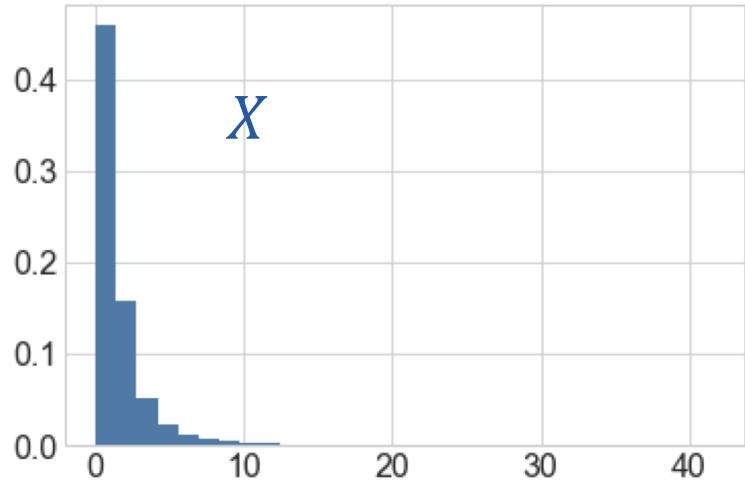


# Логарифмирование

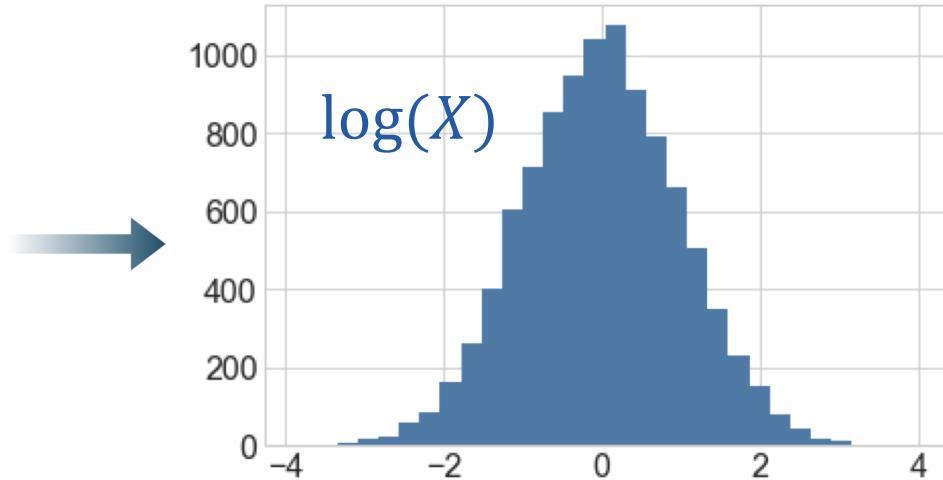
- Сгустки в начале оси абсцисс стали распределены более равномерно из-за того, что там логарифм растёт быстрее
- Расстояние между точками с большими значениями стало меньше, так как там логарифм растёт медленнее
- Чем правее мы движемся, тем медленнее растёт логарифм, скорость его роста это:

$$(\log(x))' = \frac{1}{x}$$

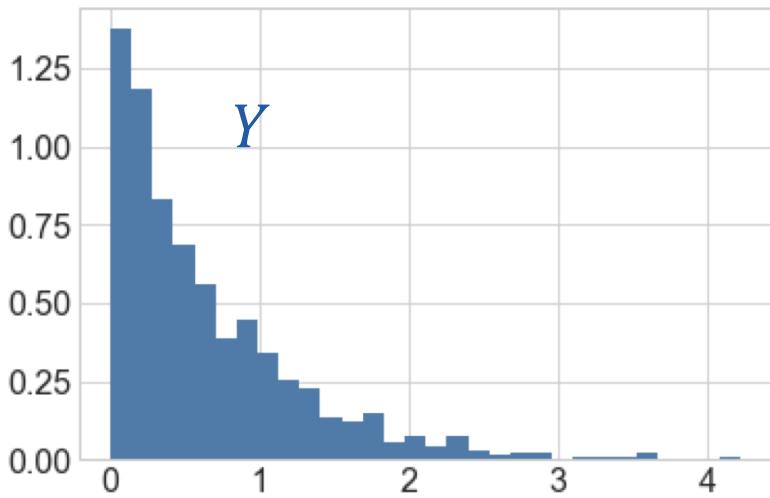
# Повторим успех?



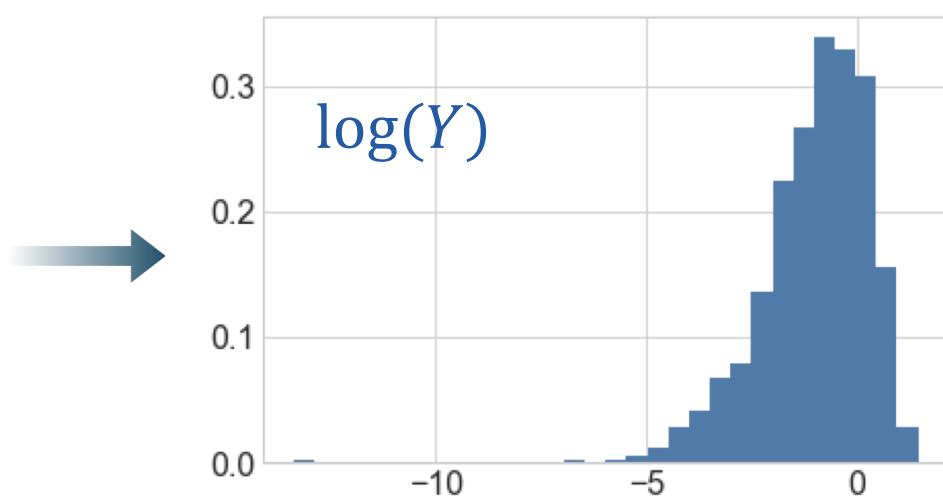
Логнормальное



$\log(X)$



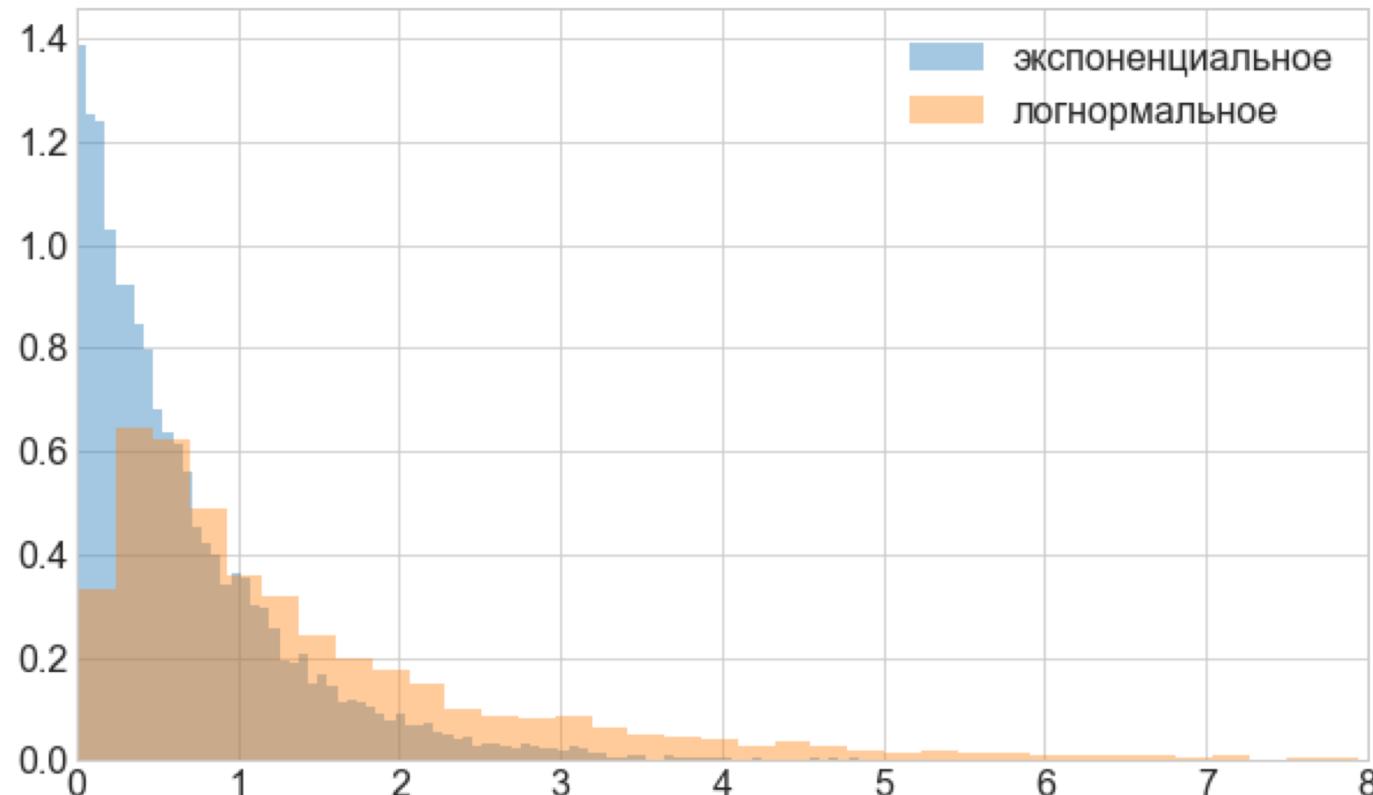
Экспоненциальное



$\log(Y)$

# Как повторить успех?

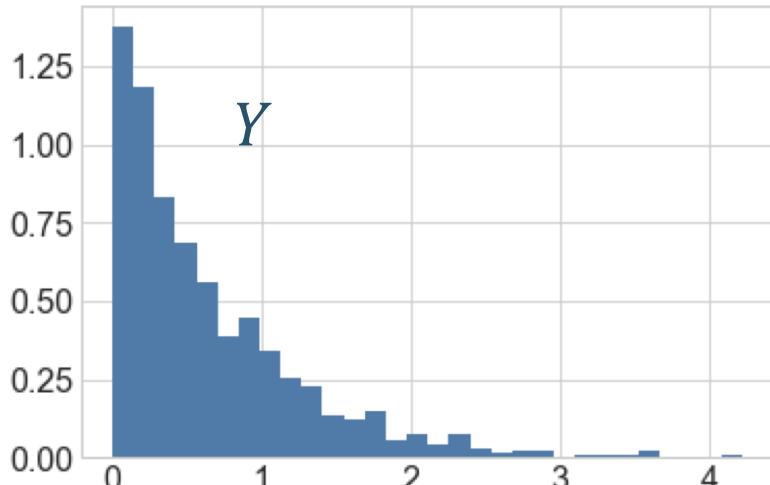
- Для экспоненциального распределения хвост легче, он не так резко убывает
- При его логарифмировании хвост появляется слева



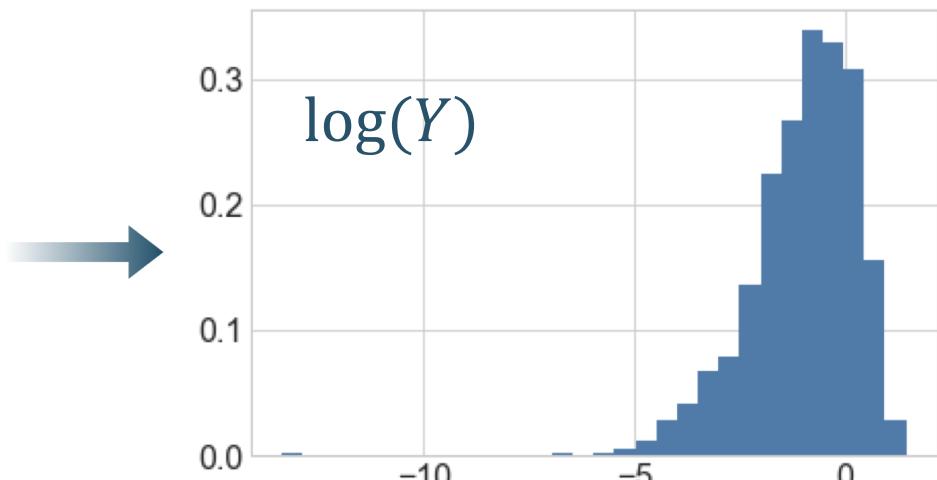
# Как повторить успех?

- Нужно выбрать преобразование с другой скоростью роста, точки слева разнеслись слишком далеко

$$\frac{1}{x} \longrightarrow \frac{1}{?}$$



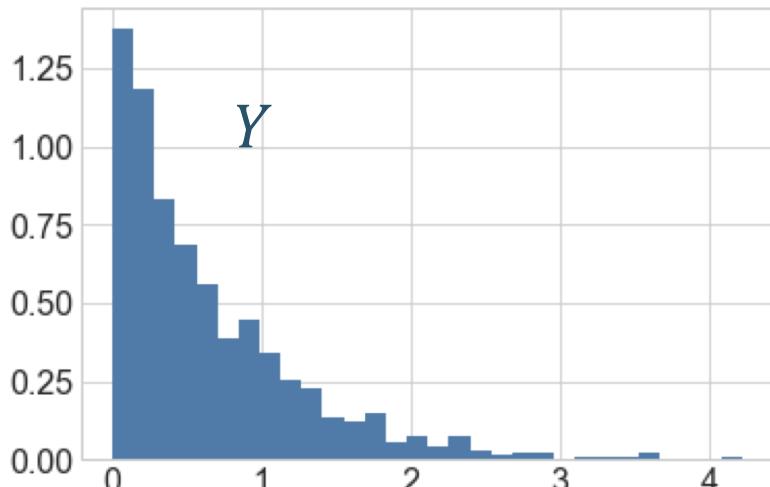
Экспоненциальное



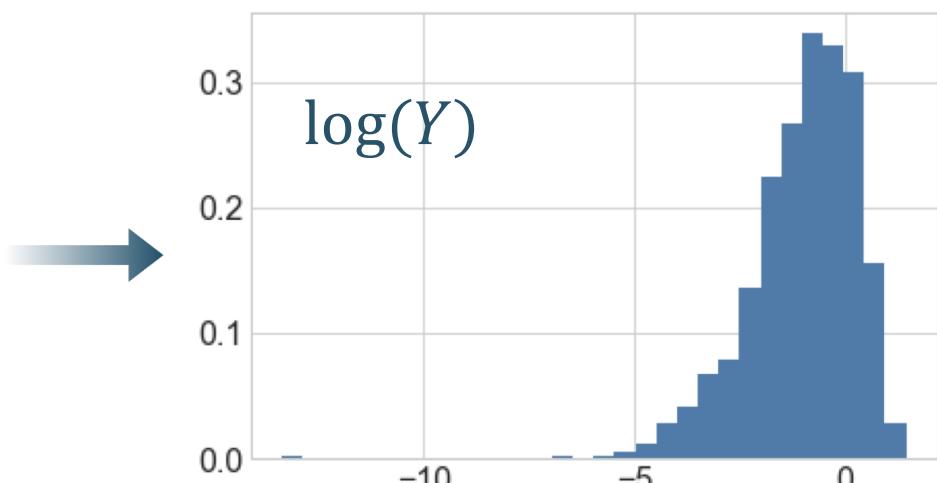
# Как повторить успех?

- Нужно выбрать преобразование с другой скоростью роста, точки слева разнеслись слишком далеко

$$\frac{1}{x} \longrightarrow \frac{1}{x^p} \quad 0 \leq p \leq 1$$



Экспоненциальное



# Как повторить успех?

- Найдём такое преобразование:

$$\int \frac{1}{x^p} dx = \frac{x^{1-p}}{1-p} + const$$

- Для удобства выберем конкретную константу и немножко перепишем формулу:

$$\int \frac{1}{x^{1-p}} dx = \frac{x^p}{p} - \frac{1}{p} = \frac{x^p - 1}{p}$$

# Преобразование Бокса-Кокса

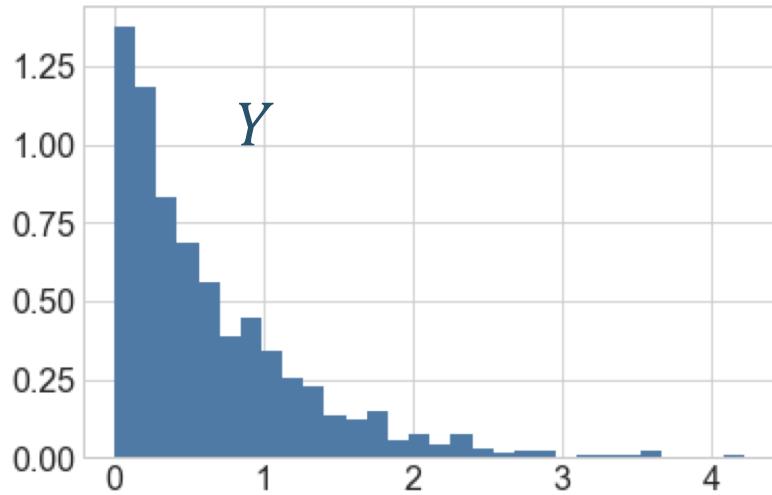
$$x_i^* = \begin{cases} \log(x), & p = 0 \\ \frac{x^p - 1}{p}, & 0 \leq p \leq 1 \end{cases}$$

- Параметр  $p$  можно выбрать, максимизируя корреляцию между квантилями нормального распределения и  $x^*$
- Если в выборке есть отрицательные значения, можно сдвинуть её в положительную область

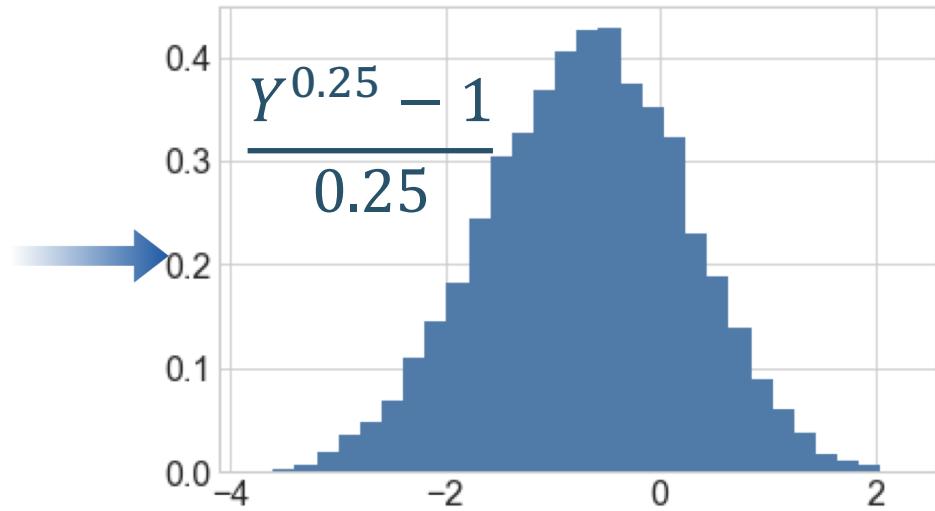
$$x_i^* = \begin{cases} \log(x + \alpha), & p = 0 \\ \frac{(x + \alpha)^p - 1}{p}, & 0 \leq p \leq 1 \end{cases}$$

# Как повторить успех?

- Если взять  $p = 0.25$ , получим трансформацию для экспоненциального распределения



Экспоненциальное



# Нормальность не панацея

- Если мы встретили на практике распределение, которое отличается от нормального, но в предпосылках того метода, который мы используем, есть нормальность, можно воспользоваться преобразованием Бокса-Кокса
- Нормальность не является необходимым свойством выборки, существует огромное количество методов, которые обходятся без неё

# Нормальность не панацея

- Проблема преобразования Бокса-Кокса в том, что чаще всего мы строим не интерпретируемую переменную

**Пример:** Время между поломками распределено экспоненциально, применив к нему преобразование Бокса-Кокса, мы получим не интерпретируемую переменную

# Проблемы с данными: масштабирование и категориальные переменные

# Похожесть

Добрыня:      Ярополк:

90 кг

60 кг

1.9 м

1.7 м

- В Анализе данных часто ищут похожие объекты на основе расстояния между ними
- Какое расстояние между Добрыней и Ярополком?

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

$$\sqrt{(1.9 - 1.7)^2 + (90 - 60)^2} = \sqrt{0.04 + 900}$$

Вес вносит в расстояние более весомый вклад из-за того, что он измерен в кг

# Похожесть

- Разный масштаб искажает подсчёт расстояний между объектами
- Позже мы узнаем, что разный масштаб портит сходимость многих алгоритмов машинного обучения
- **Решение:** отмасштабировать измеренные величины к одному диапазону, чтобы ни одно из измерений не выделялось
- Есть несколько способов масштабирования:
  - Нормализация
  - Масштабирование на отрезок  $[0; 1]$
  - Робастная нормализация (устойчивая к выбросам)

# Способы масштабирования

$i$  - номер наблюдения

**Нормализация (Standard Scaler):**

$$x_i^* = \frac{x_i - \bar{x}}{\hat{\sigma}}$$

**Масштабирование на отрезок [0; 1] (Minmax Scaler):**

$$x_i^* = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

**Устойчивая к выбросам нормализация (Robust Scaler):**

$$x_i^* = \frac{x_i - med(X)}{Q_3 - Q_1}$$

# Категориальные переменные

- Элементы неупорядоченного множества
- Город, цвет, марка машины, пол, тариф, ...

# Категориальные переменные

- Место, где провели отпуск: Крым, Дача, Испания
- Можно заменить Крым на 1, Дачу на 2, Испанию на 3

## Проблема 1:

- Мы ввели на объектах искусственный порядок, другой исследователь может ввести другой порядок и мы получим разные результаты

## Проблема 2:

- Разница между Крымом и Дачей равна 1, разница между Дачей и Испанией тоже равна 1, но эти переходы могут быть разными, непонятно как их оценить

# Бинарное кодирование (One Hot Encoding)

- Выход: закодировать каждое возможное значение как столбец из нулей и единиц (dummy-переменная)

	x
0	Испания
1	Дача
2	Крым
3	NaN
4	Дача

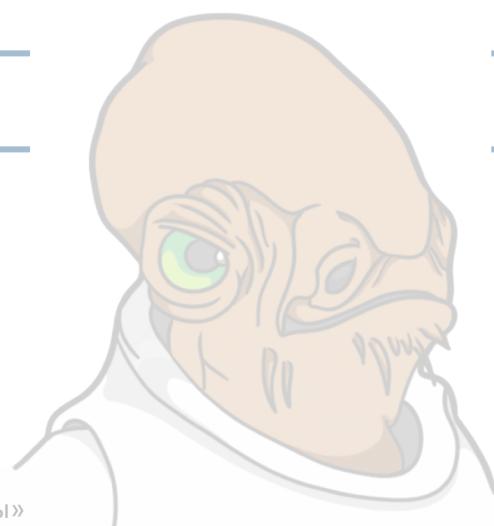


	x_isp	x_da	x_kr
0	1	0	0
1	0	1	0
2	0	0	1
3	0	0	0
4	0	1	0

# Бинарное кодирование (One Hot Encoding)

- **Dummy-ловушка** – ситуация, когда мы закодировали категориальную переменную набором столбцов, которые в сумме дают колонку из единиц

	x
0	Испания
1	Дача
2	Крым
3	Крым
4	Дача



	x_isp	x_da	x_kr
0	1	0	0
1	0	1	0
2	0	0	1
3	0	0	1
4	0	1	0

# Бинарное кодирование (One Hot Encoding)

- **Dummy-ловушка** – ситуация, когда мы закодировали категориальную переменную набором столбцов, которые в сумме дают колонку из единиц
- Ловушка состоит в том, что из-за нашей обработки в данных между столбцами возникает линейная зависимость, некоторые методы из-за этого некорректно работают

# Особенности бинарного кодирования

- Создаём много дополнительных колонок, с этим связано проклятье размерности (далее будем о нём говорить)
- Редкие категории нужно объединять в категорию “другое”
- Если признак начал принимать новое значение, мы его будем игнорировать
- Пропуски можно рассматривать как отдельную категорию, их можно не заполнять
- Можно попасть в dummy-ловушку

# Резюме

- Разный масштаб измерений может приводить к проблемам
- Категориальные признаки чаще всего нельзя использовать напрямую, один из способов работы с ними – бинарное кодирование (ОНЕ)