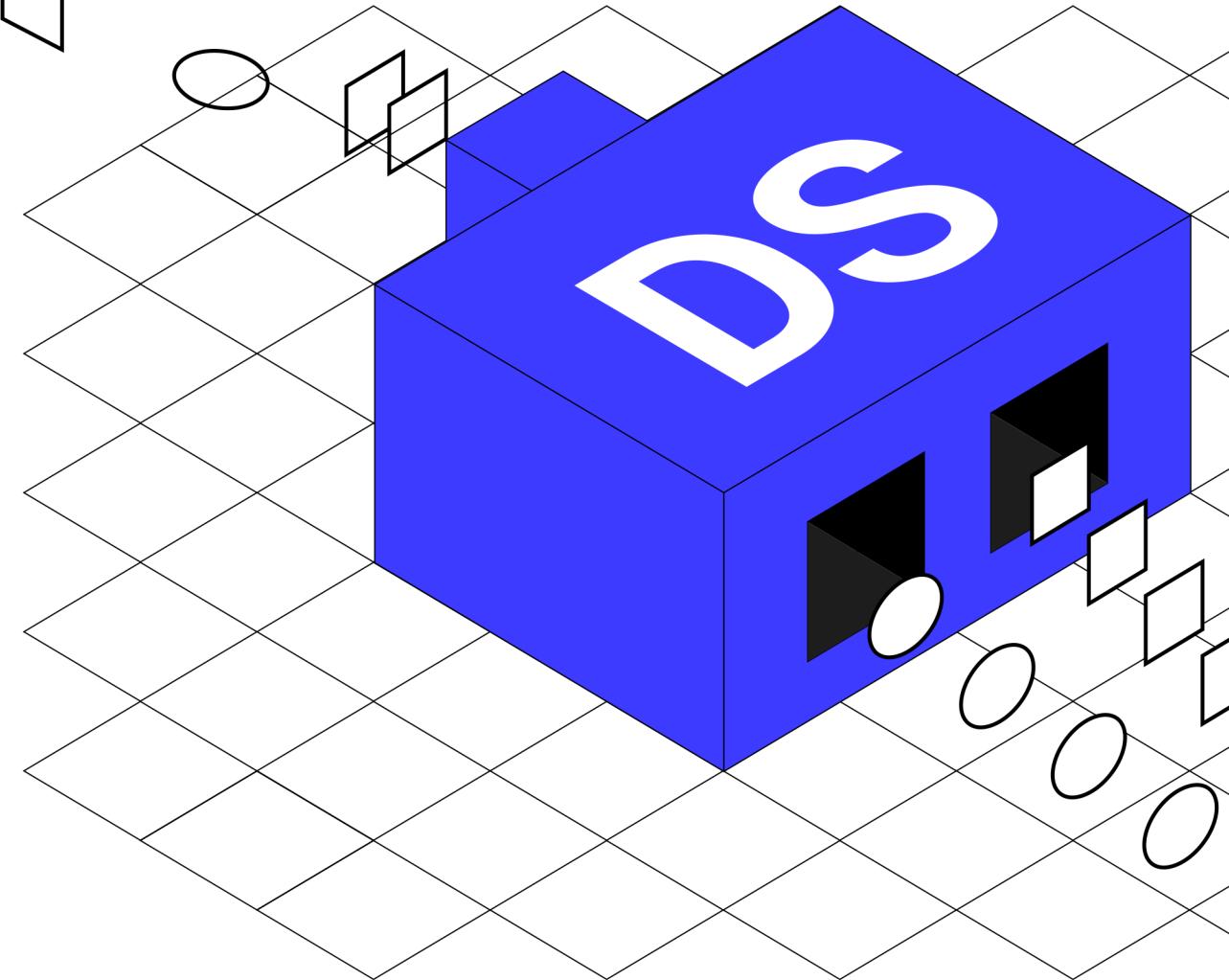


Skillbox

Data scientist с 0 до PRO



Базовая статистика для Data Scientist

О спикере

Лидия Храмова, team lead data scientist в QIWI

- Немного о моем проектном опыте и о том, как от Excel дойти до нейронок:
<https://youtu.be/tUBrP2o4Jpo?t=8479>



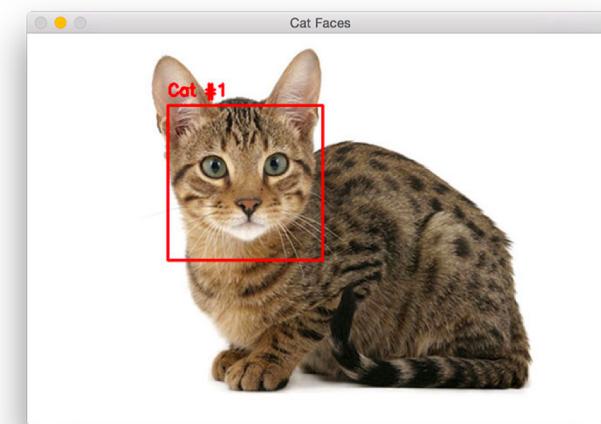
1. Случайные величины - интроверт

Для чего нужны знания математической статистики

- Чтобы делать верные выводы с учетом структуры и взаимосвязей данных
- Находить паттерны, аномалии легко и изящно
- База для понимания более сложных алгоритмов и моделей

Математическая статистика – наука, проверенная временем

- Термин «статистика» появился в 1794 году
- Первые статистические задачи были связаны с анализом ошибок геодезических и астрономических наблюдений
- Методы усложнялись и уже в 1940х годах появился прообраз современного распознавания изображений. Почти то самое распознавание капчи и котиков!



Вероятность и ее распределение

Вероятность

- Понятие появилось благодаря популярности игр, в особенности игры в кости
- В статистике броски кубика называются **случайным экспериментом**, а выпавшие на грани цифры – его **исходами**

Упрощенное определение вероятности

Классический – вероятностью случайного события А называется отношение числа n несовместных (то есть исключающих друг друга) равновероятных элементарных событий, составляющих событие А, к числу всех возможных элементарных событий N :

$$P(A) = \frac{n}{N}$$

Вероятность выпадения шестерки на кубике:

$$P(A) = \frac{1}{6}$$

Аксиоматическое определение вероятности

Аксиоматический подход – начинаем фиксировать реализации этой случайной величины в реальном мире и базировать наши решения уже на этой реальной выборке



Номер броска	Выпавшее число
1	6
2	3
...	...
N	4

Виды распределений



Дискретные - счетное число значений

- Температура тела человека
- Число бракованных изделий на линии
- Результаты футбольного матча

Известные законы

- Пуассон, Бернулли, Биномиальное



Непрерывные - бесконечное число значений на элементарном интервале

- Цена доллара
- Убытки при наступлении страхового случая

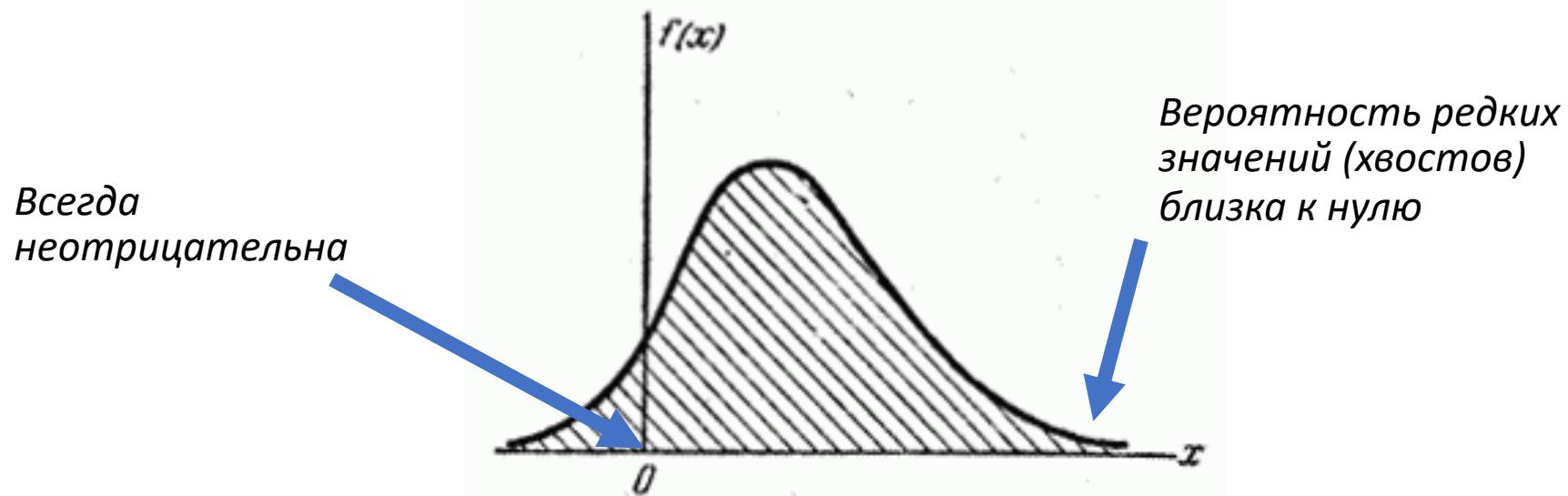
Известные законы

- Нормальное (Гауссово)
- Гамма

Плотность распределения

Плотность распределения случайной величины - это функция $f(x)$, характеризующая сравнительную вероятность попадания случайной величины на участок $x + \Delta x$

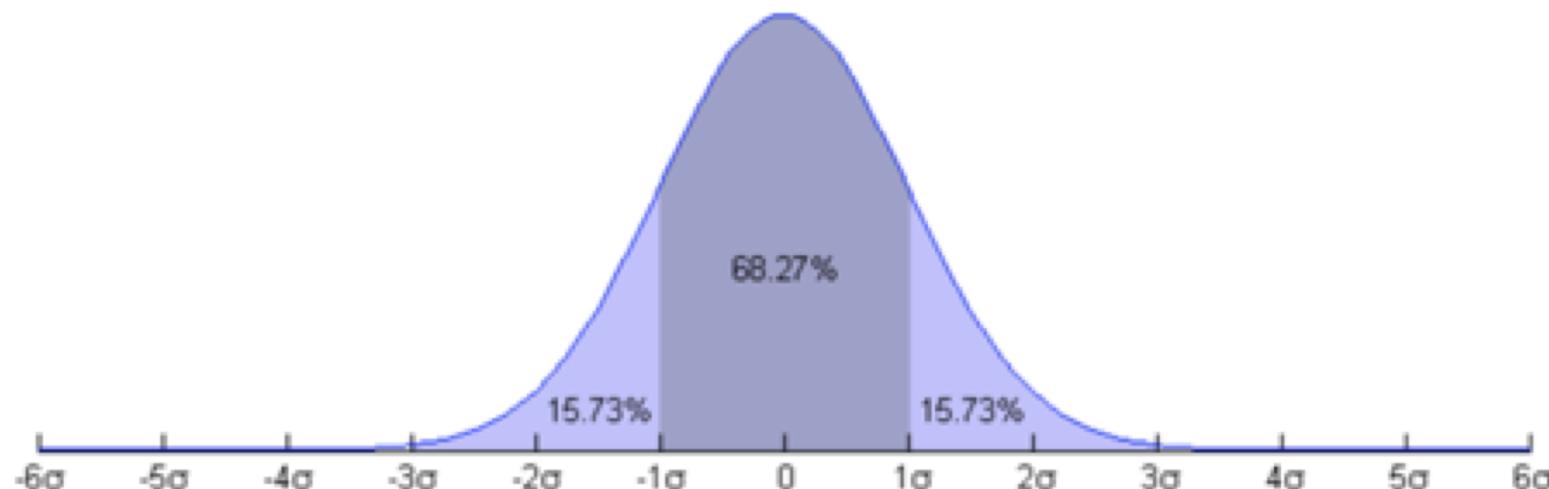
- Подходит как для дискретных, так и непрерывных случайных величин
- Может иметь различные параметры
- График плотности распределения называется **гистограммой или кривой распределения**



Плотность распределения Нормального закона

- Плотность распределения задается функцией Гаусса с параметрами μ и σ
- Гистограмма имеет вид симметричного колокола

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



Центральная предельная теорема (ЦПТ)

Сумма **независимых одинаково распределенных случайных величин** имеет распределение близкое к **нормальному распределению**

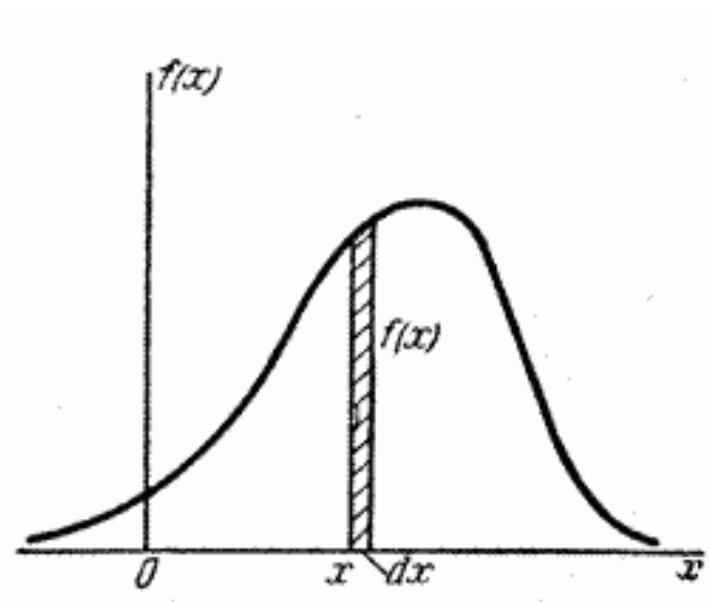
Нормальное распределение встречается очень часто)



Функция распределения

Функция распределения случайной величины - существует только для непрерывных случайных величин и представляет собой $f(x)$, характеризующую вероятность попадания x в элементарный интервал dx

- Подходит только для непрерывных случайных величин
- Графически выглядит так



Как измерить случайную величину

Точечные оценки случайной величины

- **Математическое ожидание** или **среднее** значение случайной величины. Часто обозначается \bar{x} , $m(x)$ или $EV(x)$ – от английского expected value (ожидаемое значение)
- **Дисперсия** – физически представляет собой меру **рассеяния** значений вокруг среднего случайной величины. Полезный показатель для анализа разброса в вашей выборке.

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

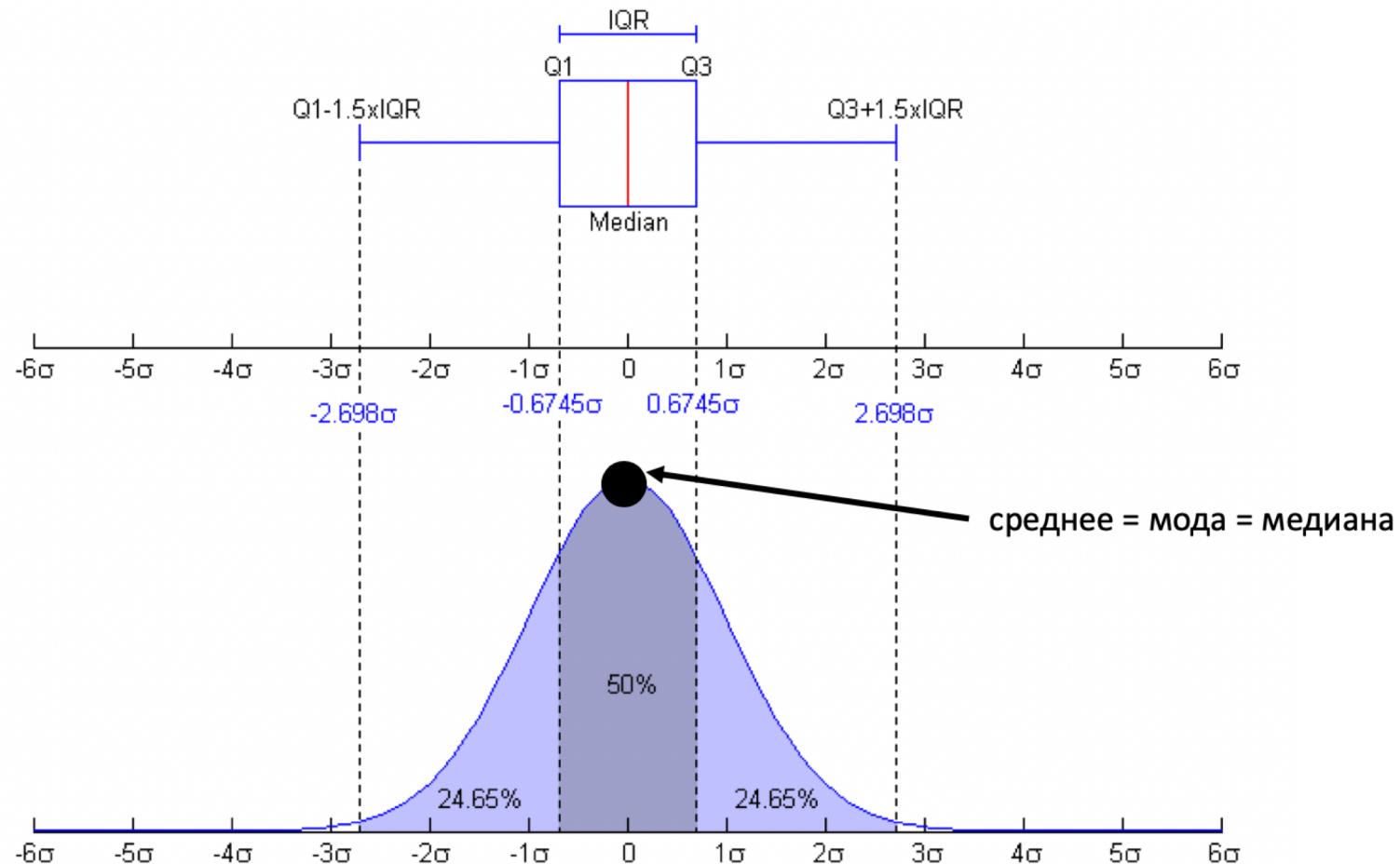
- **Среднеквадратическое отклонение, СКО, $S(x)$ и сигма** – еще одна мера рассеяния

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

Точечные оценки случайной величины

- **Мода или $Mo(x)$** – самое часто встречающееся значение случайной величины
- **Квантили или x_α** – семейство оценок, отвечающих на вопрос – какое значение x_α случайной величины не превысит с заданной вероятностью α .
- **Квартили** - квантили с вероятностями 0.25, 0.5, 0.75, разделяющие всю совокупность возможных исходов на четыре равных промежутка. 0.5 квартиль в статистике еще называют медианой.
- **Интерквартильный размах или IQR** – аналог дисперсии, полезный для оценки разброса случайной величины. Разница между 0.75 и 0.25 квантилем - $x_{0,75} - x_{0,25}$

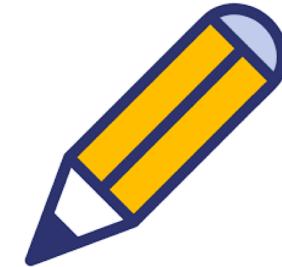
Точечные оценки Гауссова закона



Практика 1.1 Анализ распределения случайной величины

Для практики понадобится

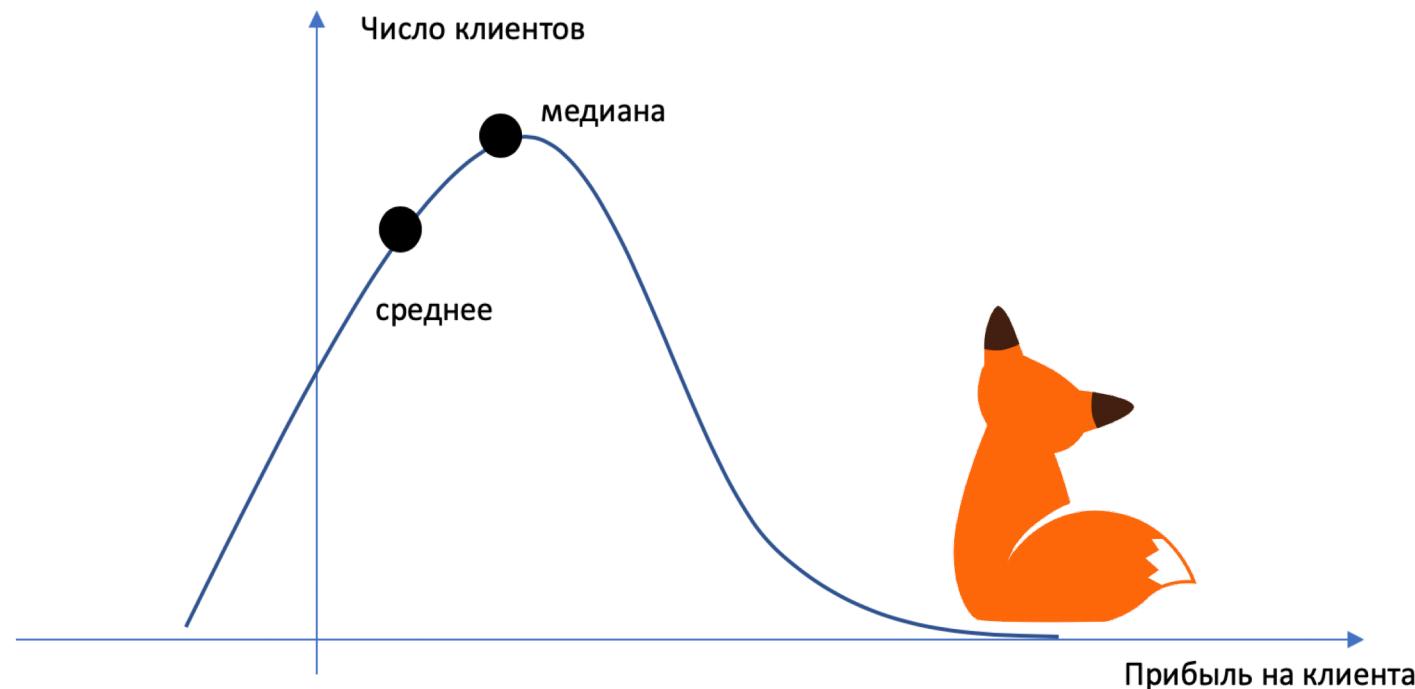
- Сборка Anaconda for python 3.5 и выше
- Все примеры будут продемонстрированы на сборке с python 3.6



Проклятие
ассиметричности,
бимодальность или
где чаще всего
ошибаются аналитики

Асимметричность в данных

- Средняя оценка оказалась нерелевантной – посчитав только среднее мы не приняли во внимание, что у нашего распределения есть так называемые «хвосты» - наличие выбросов слева или справа



Что делать?

- Использовать весь набор точечных оценок (квантили, показатели разброса) для оценки спектра распределения
- Не забывать про гистограмму
- Использовать **специальные коэффициенты формы распределения**

Коэффициенты асимметрии и эксцесса

- Коэффициент асимметрии (*skewness*) – характеризует смещенность распределения и вычисляется по формуле

$$y_1 = \frac{m[(x - m(x))^3]}{S(x)^3}$$

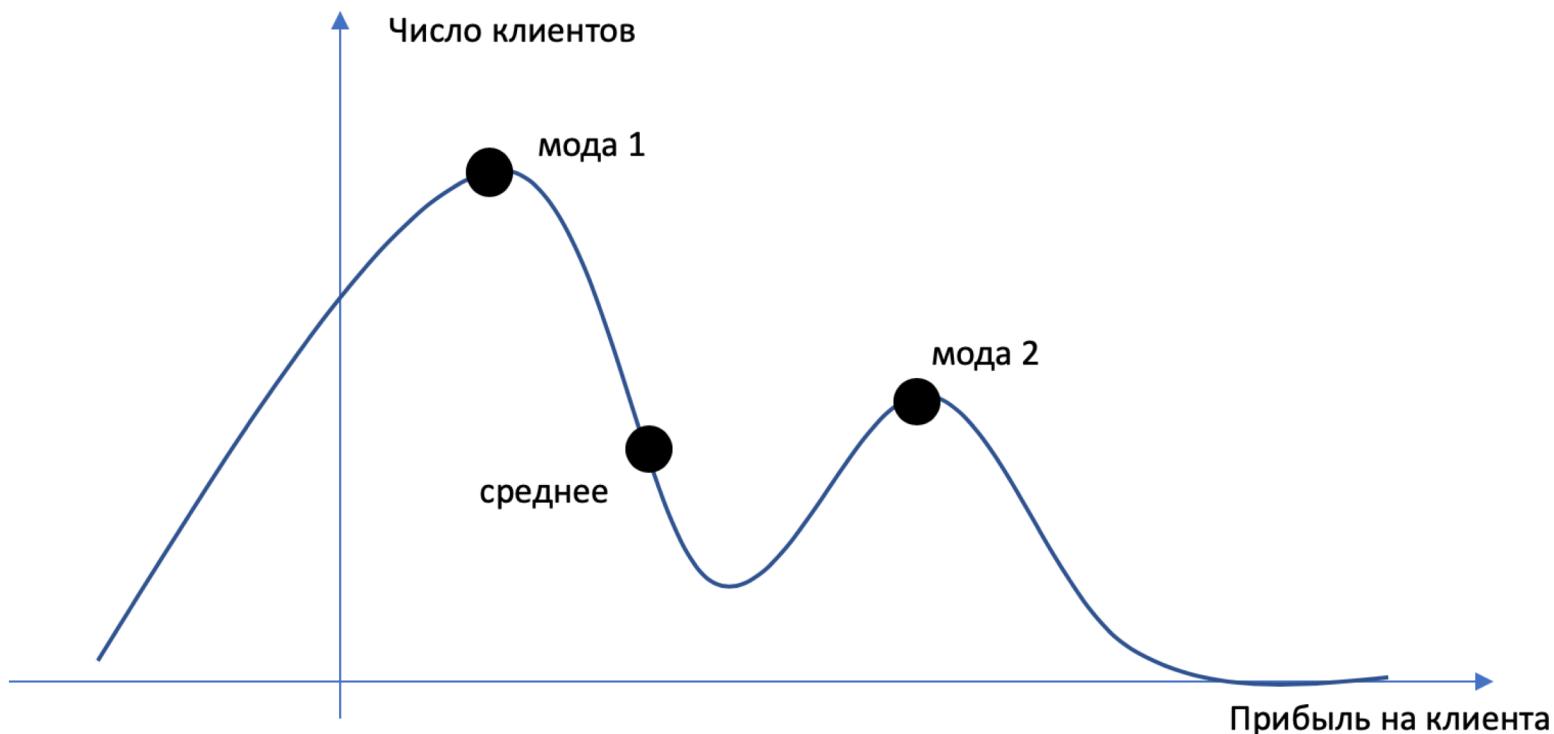
Положителен, если правый хвост распределения длиннее левого, отрицателен – если левый длинные правого. **У нормального распределения равен 0.**

- Коэффициент эксцесса (*kurtosis*) характеризует остроконечность распределения. Положительный при островом пике у среднего значения и отрицателен при гладком. **У нормального распределения эксцесс равен 0.**

$$y_2 = \frac{m[(x - m(x))^4]}{S(x)^4}$$

Еще страшнее - мультимодальность

- В выборке существуют несколько модальных значений
- Мультимодальность говорит о наличии нескольких кластеров внутри изучаемой метрики



Практика 1.2 Анализ распределения ассиметричной случайной величины

2. Корреляционный анализ

Линейная корреляция

- Корреляционно-регрессионный анализ (КРА) помогает находить взаимосвязи между признаками
- **Линейная корреляция** - это статистическая взаимосвязь двух и более случайных величин, при которой изменений значений одной сопутствует изменений значений второй.

Линейная связь бывает **прямой** и **обратной**:

- Прямая связь - изменение числа покупок мороженого в зависимости от температуры воздуха. Чем жарче, тем больше мороженого продается
- Обратная связь – потребление сахара и здоровье зубов – чем больше сладостей, тем вреднее для эмали

Оценка для количественных данных

Линейный коэффициент корреляции Пирсона

$$r(x, y) = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Принимает значения от -1 до 1, где -1 – максимальная обратная связь, а 1 – максимальная прямая связь

Шкала Чеддока для количественных данных

Количественная мера тесноты связи	Качественная характеристика силы связи
0,1 - 0,3	Слабая
0,3 - 0,5	Умеренная
0,5 - 0,7	Заметная
0,7 - 0,9	Высокая
0,9 - 0,99	Весьма высокая

Шкала используется для модуля коэффициента Пирсона

Оценка для ранговых данных

Линейный коэффициент корреляции Кендалла

Значения количественного показателя x выставляют в порядке возрастания и присваивают им ранги, значение рангового показателя y - сортируют и рассчитывают следующий коэффициент:

$$\tau = \frac{2S}{n(n - 1)}$$

Где $S = P - Q$, P - суммарное число наблюдений, следующих за текущими наблюдениями с большим значением рангов Y , Q - суммарное число наблюдений, следующих за текущими наблюдениями с меньшим значением рангов Y

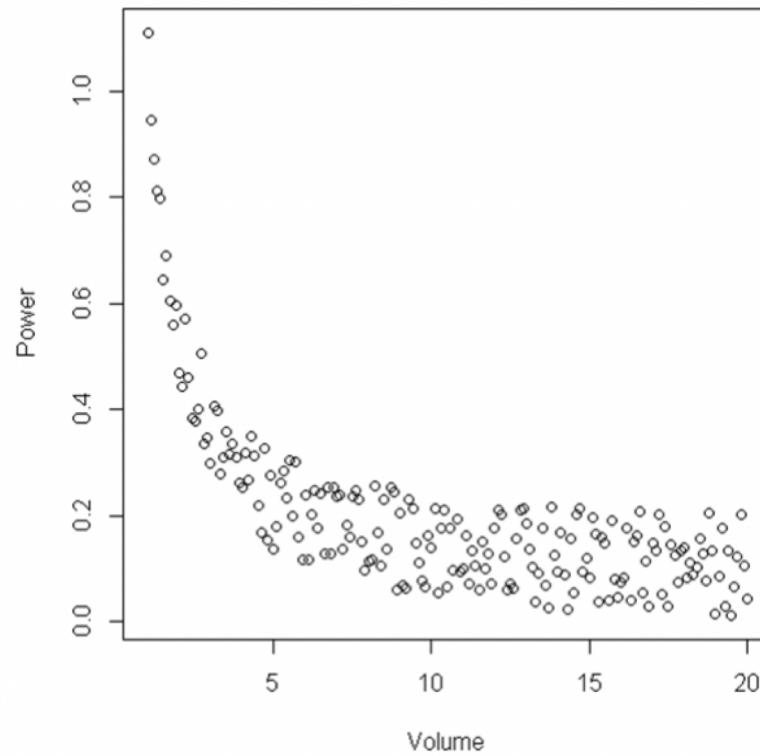
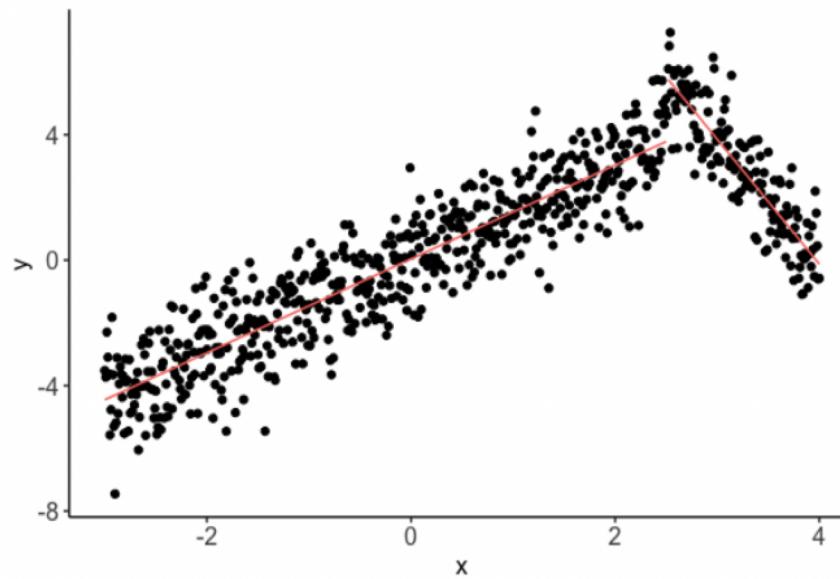
Матрица парных корреляций

Полезна для анализа попарных корреляций в большом датасете размерности $N * N$, где N - число фич

-	y	x_1	x_2	x_3
y	1	r_{yx1}	r_{yx2}	r_{yx3}
x_1	r_{x1y}	1	r_{x1x2}	r_{x1x3}
x_2	r_{x2y}	r_{x2x1}	1	r_{x2x3}
x_3	r_{x3y}	r_{x3x1}	r_{x3x2}	1

Мифы про связи – не все линейно

- Для базового поиска нелинейных связей поможет **диаграмма рассеяния** (scatter plot)

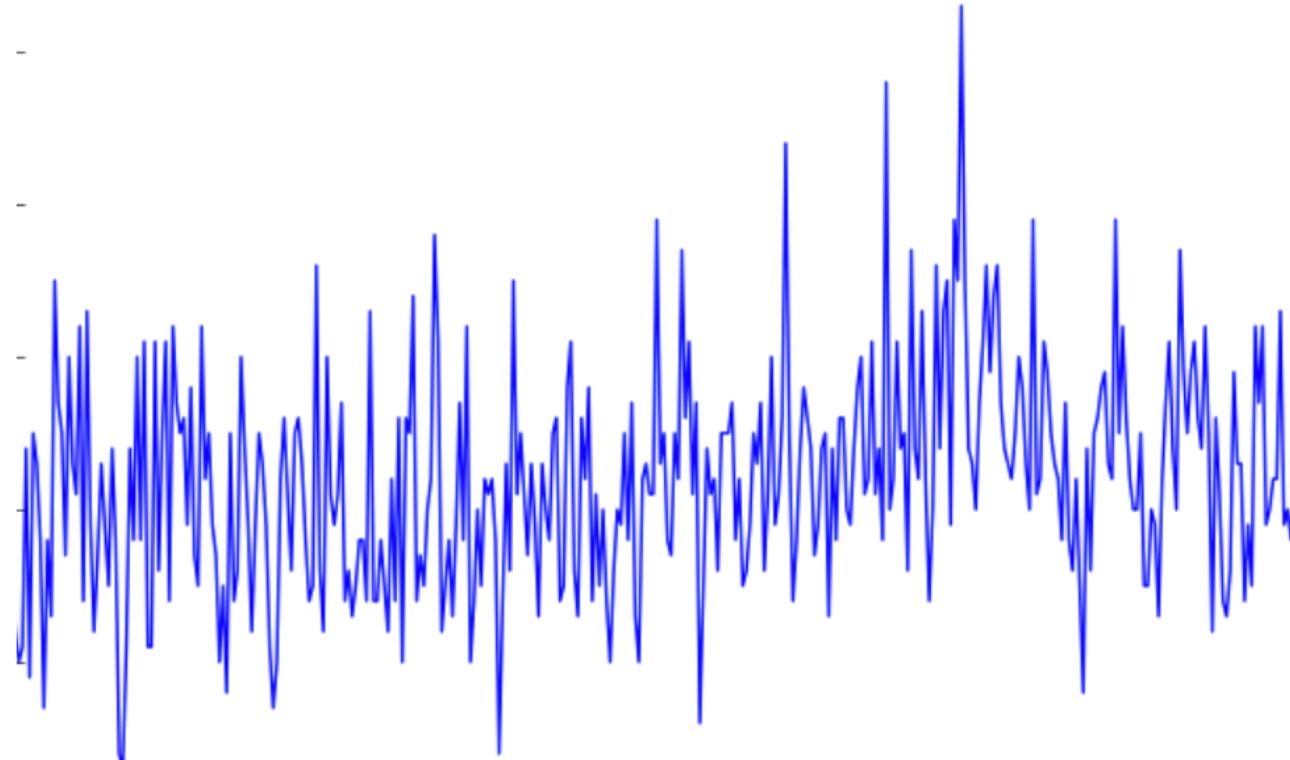


Практика 2 Пойск взаимосвязей в данных

3. Доверительные интервалы

Точных оценок недостаточно

- Что делать, если нужно оценить очень нестабильный процесс?



Понятие доверительного интервала

Доверительный интервал – это особый, **интервальный** вид оценки параметров случайной величины, который позволяет оценить их с заданной **доверительной вероятностью**

Доверительный интервал уровня $\alpha = 95\%$ для параметра означает, что при проведении большого числа независимых экспериментов в 95 % случаев истинное значение параметра будет лежать внутри этого интервала.

Как найти доверительный интервал

Для нормально распределенных выборок существуют два варианта интервалов:

- Для математического ожидания μ с известной дисперсией – вычисляется по формуле

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

где z_α - квантиль нормального распределения уровня α .

- При неизвестной дисперсии

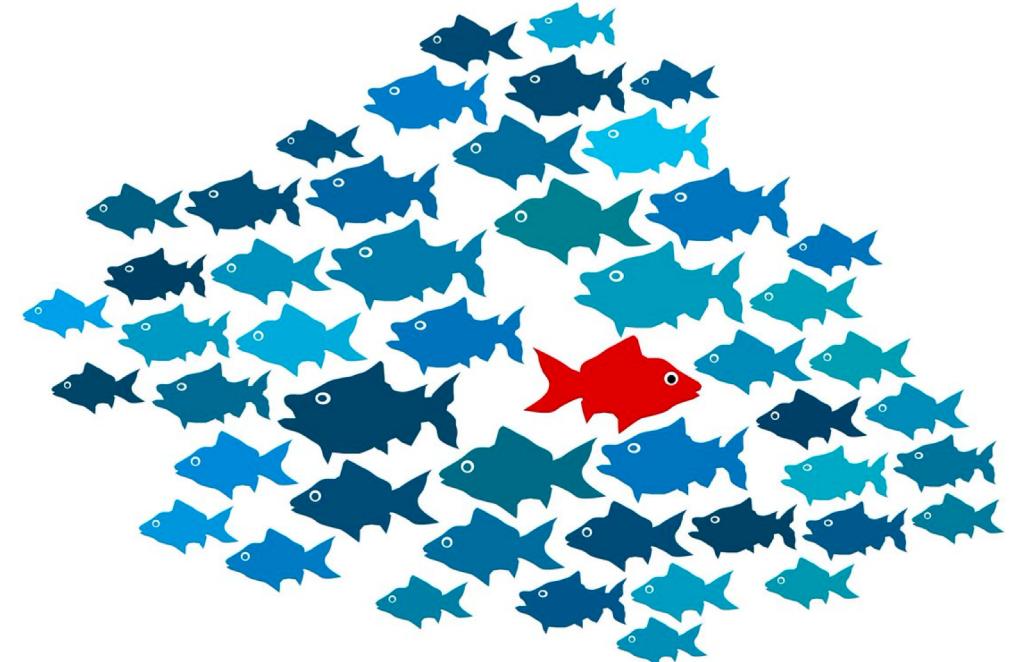
$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Практика 3. Построение доверительных оценок

4. Аномалии в данных

Что такое аномалии и что с ними делать?

- Аномалией называют наблюдение, отличное от основной совокупности
- Аномалии не всегда нужно удалять из данных, но всегда стоит **фильтровать и исследовать** отдельно

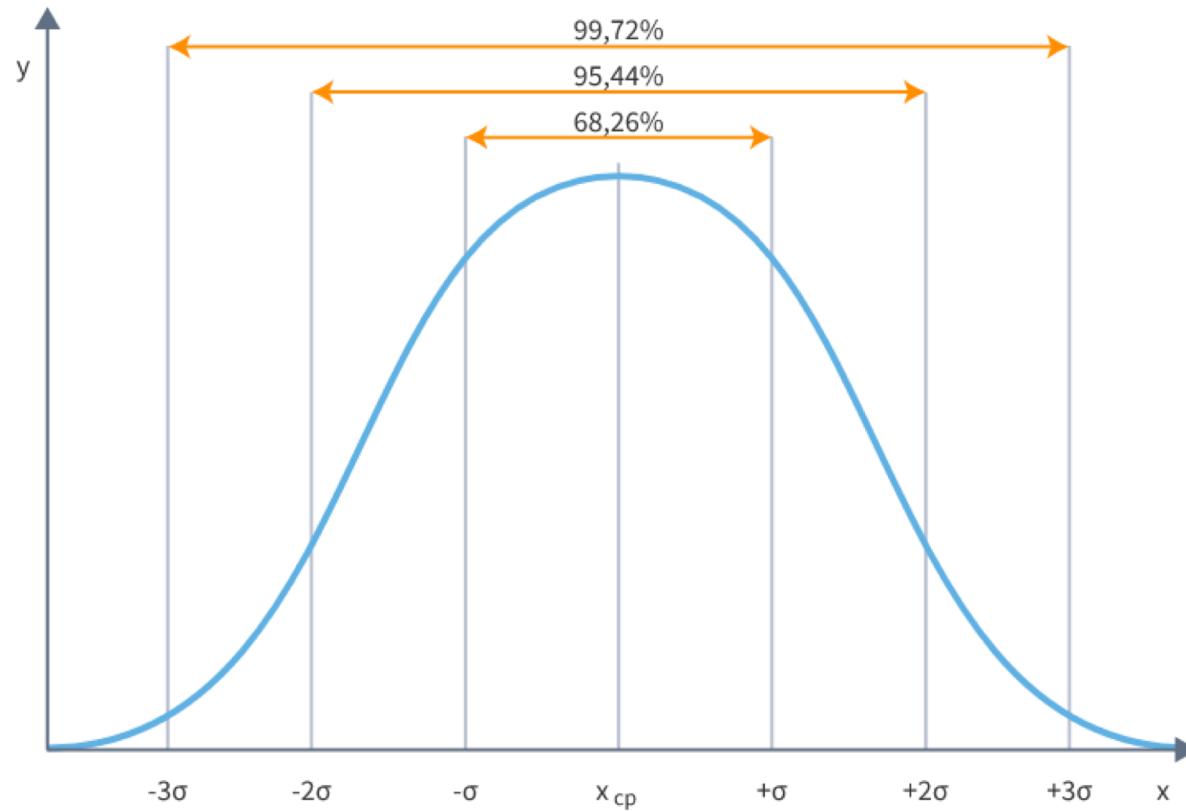


Инструменты поиска аномалий

1. **Для одномерного пространства** – основаны на свойствах распределения - IQR, квантили, доверительные интервалы, правило трех сигм
2. **Для многомерного пространства** – основаны на алгоритмах машинного обучения, например Isolation forest - <https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf>
One Class Support Vector Machine - <https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf>

Правило трех сигм

Для нормально распределенных данных вероятность выхода за коридор $m(x) \pm 3\sigma$ практически нулевая



Практика 4. Поиск аномалий в данных

Вместо заключения

1. Помните базовые принципы
2. Экспериментируйте с данными
3. Комбинируйте подходы

