

2. Основы теории вероятности.

Многомерное нормальное распределение.

Преобразование данных

План

- Вспомним основные понятия теории вероятностей
- Поговорим про то, какими бывают распределения
- Многомерное нормальное распределение
- Преобразование данных

Пакт

Заклучим соглашение!

X, Y, Z — случайные величины

x, y, z — какие-то конкретные значения

A, B, C — события

\mathbb{P} — вероятность

$\mathbb{E}(X)$ — математическое ожидание

$Var(X)$ — дисперсия

$Cov(X, Y), \rho(X, Y)$ — ковариация и корреляция

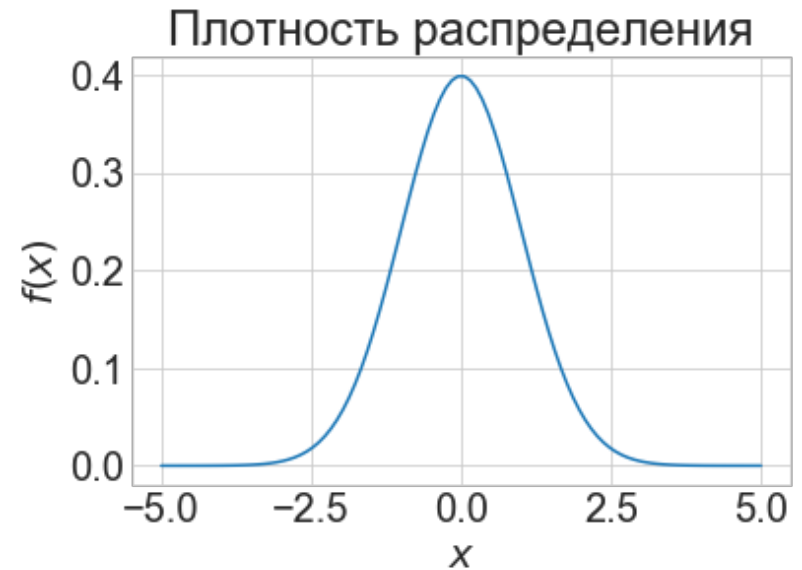
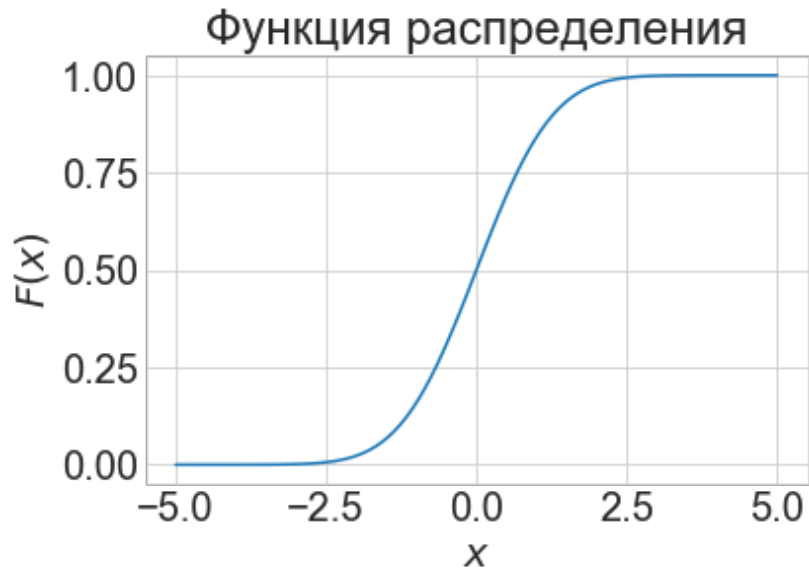
Как устроен мир



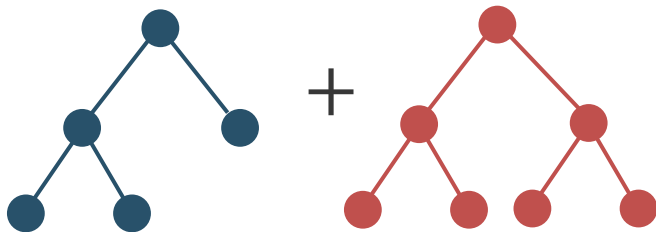
X

- Теория вероятностей изучает различные процессы порождения данных (некоторый сундук). В реальности мы не наблюдаем эти процессы.
- Однако эти процессы порождает **выборки**. Математическая статистика изучает их свойства и пытается восстановить структуру.

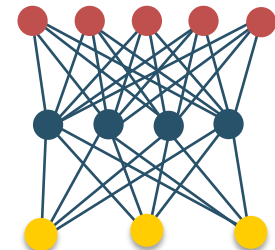
Устройство сундука



Модель – наше предположение о том, как процесс порождения данных устроен. За каждой моделью стоят какие-то предпосылки, описывающие наше незнание.



$$M = U V^T$$



Что мы будем делать?

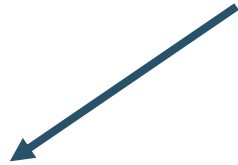
- Изучать выборки и их свойства
- Предполагать, какие процессы порождают данные, описывать своё незнание с помощью какой-то модели
- Разбираться, насколько наши предположения согласуются с выборками



Распределение случайной величины

Какими бывают случайные величины

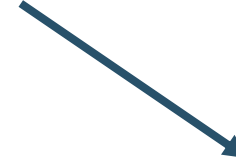
Случайные величины



Дискретные

Множество значений
конечно или счётно

(число звонков, число очков
на игральной кости, число
ошибок на страницу текста)



Непрерывные

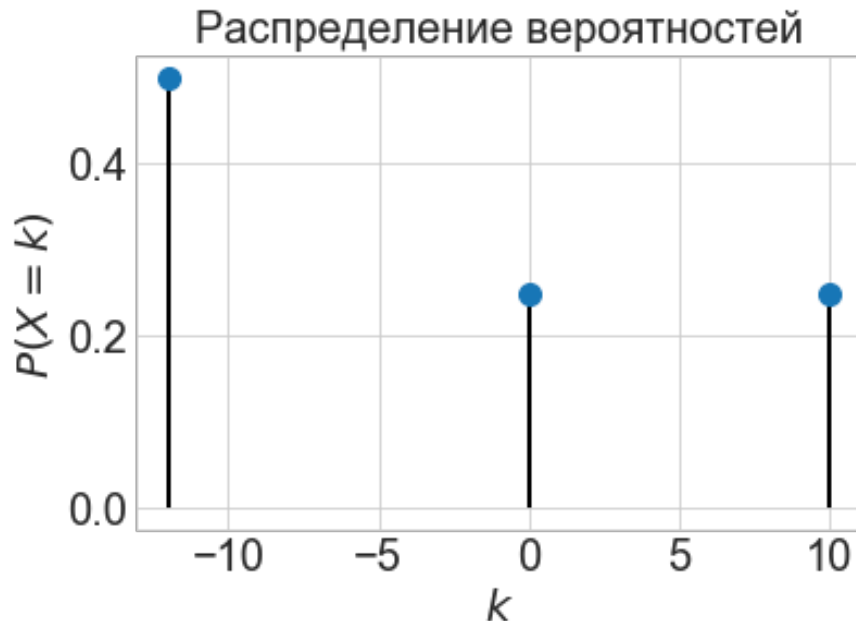
Принимают бесконечное,
континуальное число
значений

(рост, время ожидания
автобуса, вес)

Дискретные случайные величины

Распределение дискретной случайной величины – таблица, которая описывает, какие значения принимает случайная величина с какой вероятностью

Сумма вероятностей должна быть равна **1**, каждая вероятность лежит между **0** и **1**



Пример: лотерея

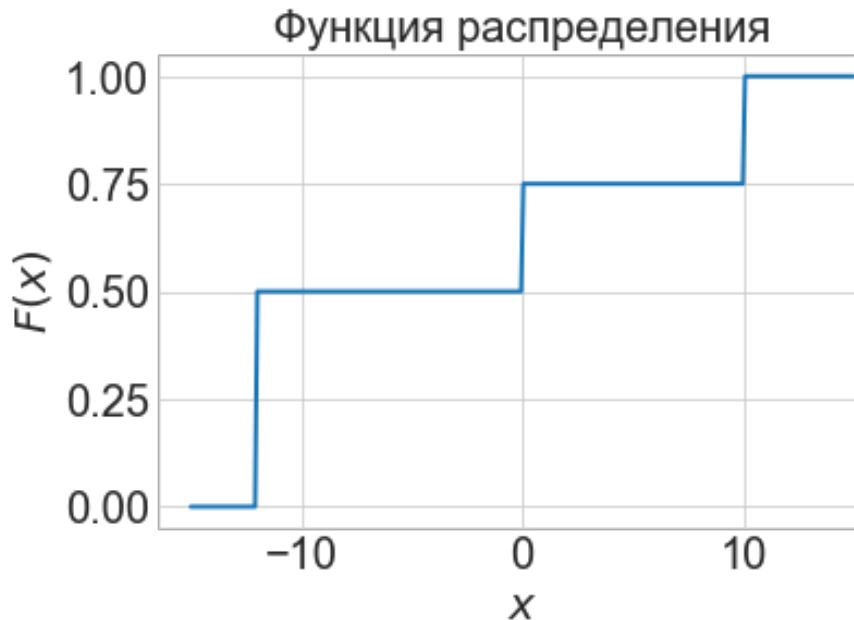
X	-12	0	10
$\mathbb{P}(X = k)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

Дискретные случайные величины

Функция распределения – функция, которая определяет вероятность события $X \leq x$, то есть

$$F(x) = \mathbb{P}(X \leq x) = \sum \mathbb{P}(X = k) \cdot [X \leq x],$$

$$[X \leq x] = \begin{cases} 1, & X \leq x \\ 0, & \text{иначе} \end{cases}$$

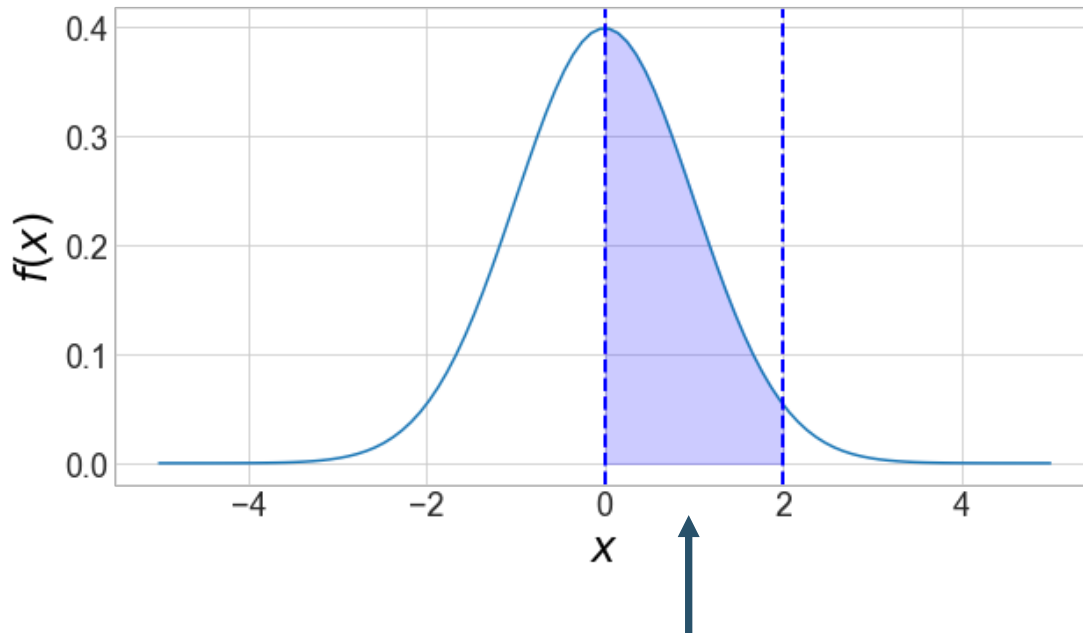


Пример: лотерея

X	-12	0	10
$\mathbb{P}(X = k)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

Непрерывные случайные величины

Распределение непрерывной случайной величины описывается **плотностью распределения вероятностей**.



Площадь равна вероятности попасть на отрезок от нуля до двух

Пример:
нормальное
распределение

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

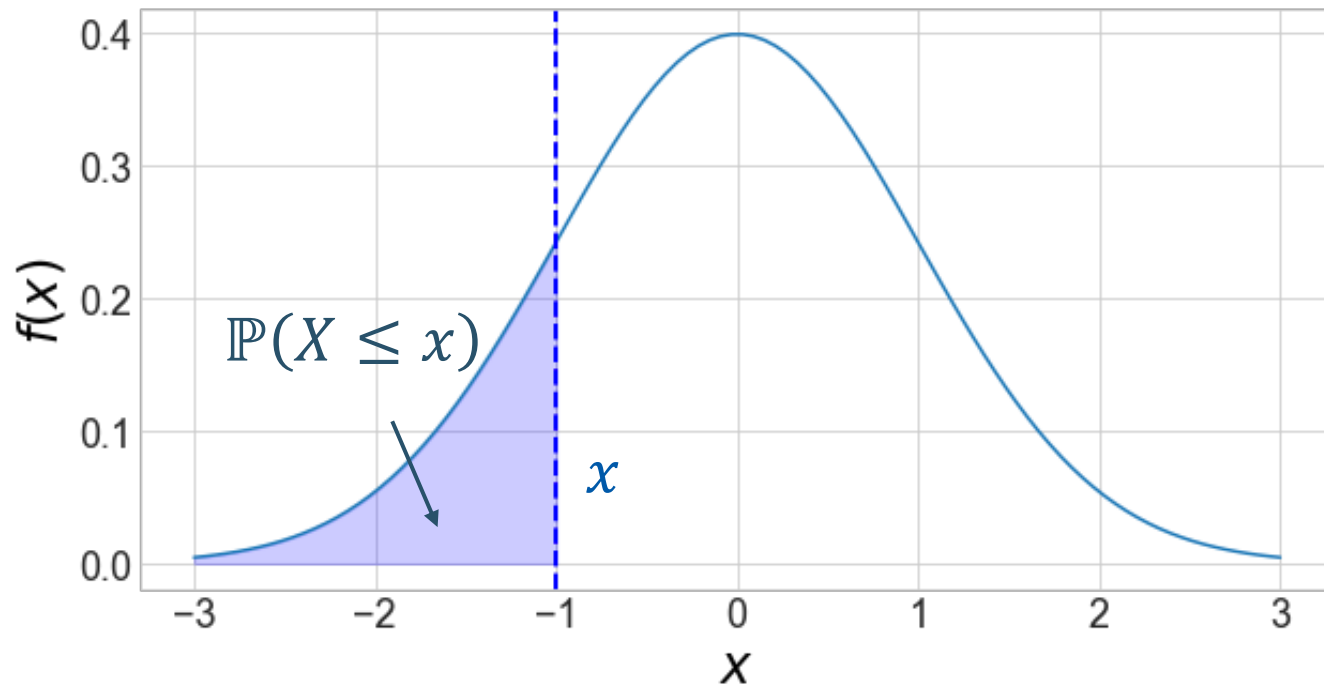
$$= \int_0^2 f(x) dx$$

Площадь под всей плотностью должна быть равна 1

Непрерывные случайные величины

Функция распределения – функция, которая определяет вероятность события $X \leq x$, то есть

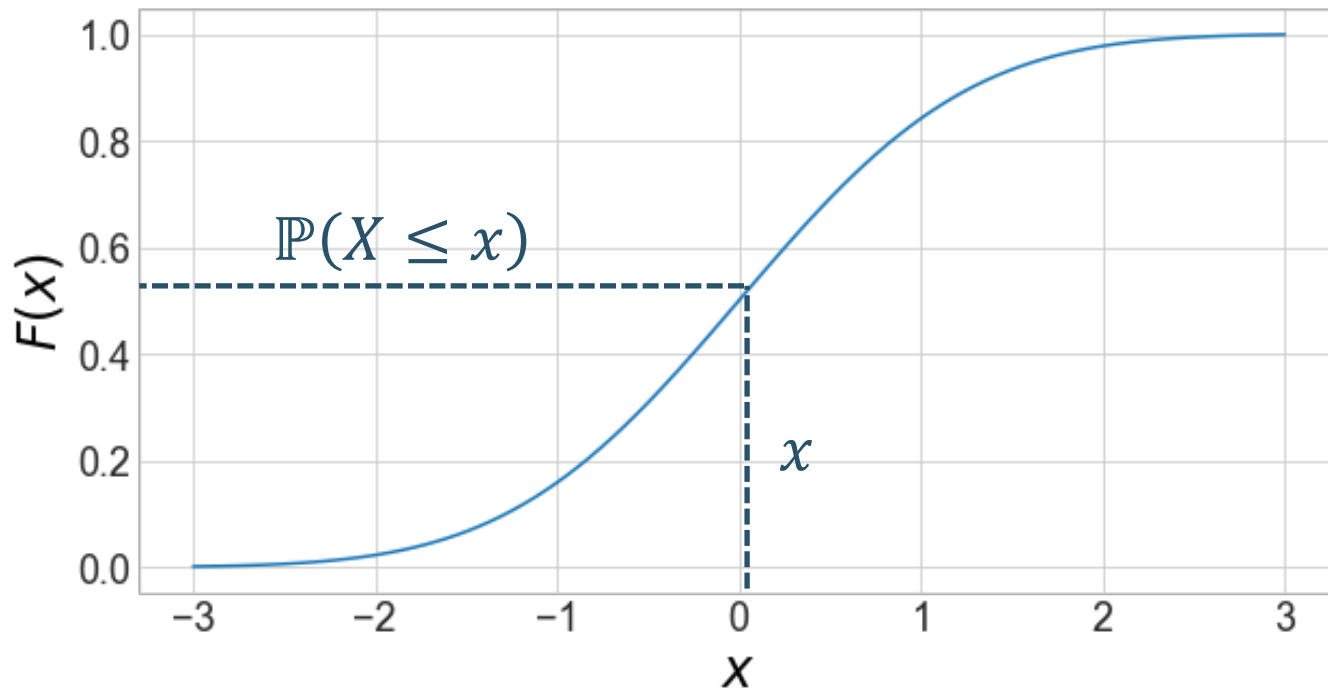
$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) dt, f(t) - \text{плотность}$$



Непрерывные случайные величины

Функция распределения – функция, которая определяет вероятность события $X \leq x$, то есть

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) dt, f(t) – \text{плотность}$$



Важные свойства

1. Плотность определена только для непрерывных случайных величин
2. $f(x) = F'(x)$
3. $\int_{-\infty}^{+\infty} f(t) dt = 1, \quad f(t) \geq 0 \quad \forall t$
4. $F(x)$ не убывает, лежит между 0 и 1
5. $\mathbb{P}(a \leq X \leq b) = \int_a^b f(t) dt = F(b) - F(a)$
6. Вероятность того, что непрерывная случайная величина попадёт в точку, равна нулю

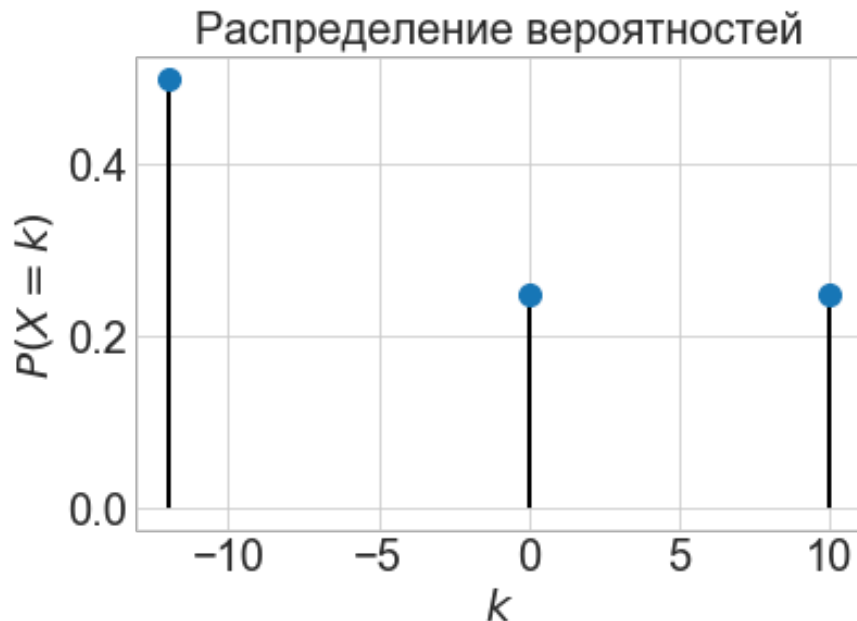
Характеристики случайных величин

Математическое ожидание

Математическое ожидание – среднее значение случайной величины

$$\mathbb{E}(X) = \sum_{k=1}^n k \cdot \mathbb{P}(X = k)$$

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} t \cdot f(t) dt$$



Пример: лотерея

X	-12	0	10
$\mathbb{P}(X = k)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

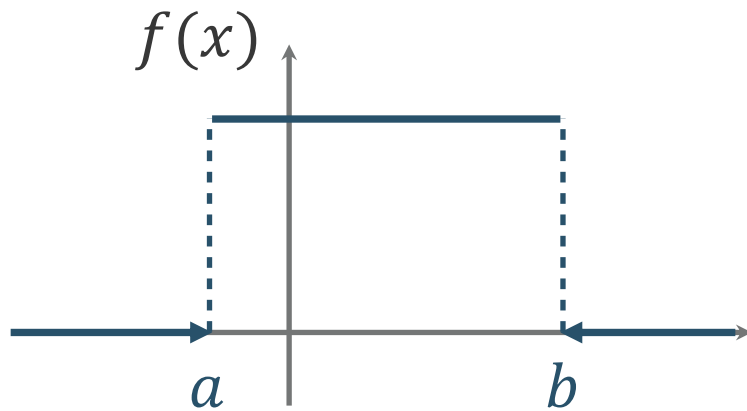
$$\mathbb{E}(X) = -12 \cdot 0.5 + 0 \cdot 0.25 + 10 \cdot 0.25 = -3.5 \text{ рубля}$$

Математическое ожидание

Математическое ожидание – среднее значение случайной величины

$$\mathbb{E}(X) = \sum_{k=1}^n k \cdot \mathbb{P}(X = k)$$

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} t \cdot f(t) dt$$



Пример: равномерное распределение

$$f(x) = \frac{1}{b-a}, x \in [a; b]$$

$$\mathbb{E}(X) = \int_a^b t \cdot \frac{1}{b-a} dt = \frac{1}{b-a} \cdot \frac{t^2}{2} \Big|_a^b = \frac{(b^2-a^2)}{2(b-a)} = \frac{a+b}{2}$$

Математическим ожиданием оказывается середина отрезка

Свойства математического ожидания

X, Y — случайные величины a — константа

1. $\mathbb{E}(a) = a$
2. $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
3. $\mathbb{E}(a \cdot X) = a \cdot \mathbb{E}(X)$
4. $\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$, если независимы
5. Математическое ожидание случайной величины — не случайно
6. $\mathbb{E}(X - \mathbb{E}(X)) = \mathbb{E}(X) - \mathbb{E}(\mathbb{E}(X)) = \mathbb{E}(X) - \mathbb{E}(X) = 0$

Дисперсия

Дисперсия – мера разброса случайной величины вокруг её среднего

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \sum_{k=1}^n (k - \mathbb{E}(X))^2 \cdot \mathbb{P}(X = k)$$

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \int_{-\infty}^{+\infty} (t - \mathbb{E}(X))^2 \cdot f(t) dt$$

Дисперсия

Дисперсия – мера разброса случайной величины вокруг её среднего

Более удобно искать дисперсию по формуле:

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}(X - \mathbb{E}(X))^2 \\ &= \mathbb{E}(X^2 - 2 \cdot X \cdot \mathbb{E}(X) + \mathbb{E}^2(X)) \\ &= \mathbb{E}(X^2) - 2 \cdot \mathbb{E}(X) \cdot \mathbb{E}(\mathbb{E}(X)) + \mathbb{E}^2(X) \\ &= \mathbb{E}(X^2) - 2 \cdot \mathbb{E}(X) \cdot \mathbb{E}(X) + \mathbb{E}^2(X) \\ &= \mathbb{E}(X^2) - \mathbb{E}^2(X) \end{aligned}$$

Дисперсия

Дисперсия – мера разброса случайной величины вокруг её среднего

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$$

Пример: лотерея

X	-12	0	10
$\mathbb{P}(X = k)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

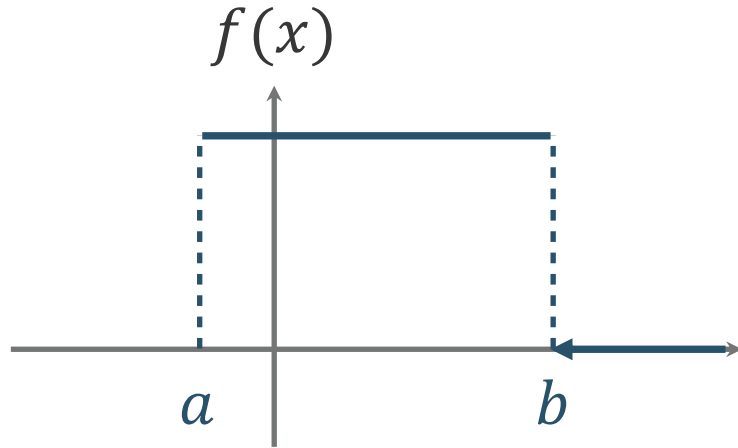
$$\mathbb{E}(X^2) = (-12)^2 \cdot 0.5 + 0^2 \cdot 0.25 + 10^2 \cdot 0.25 = 97 \text{ рублей}^2$$

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = 97 - 12.25 = 84.75 \text{ рублей}^2$$

Дисперсия

Дисперсия – мера разброса случайной величины вокруг её среднего

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$$



Пример: равномерное

$$f(x) = \frac{1}{b-a}, x \in [a; b]$$

$$\mathbb{E}(X^2) = \int_{-\infty}^{+\infty} t^2 \cdot \frac{1}{b-a} dt = \frac{1}{b-a} \cdot \frac{t^3}{3} \Big|_a^b = \frac{(b^3 - a^3)}{3(b-a)} = \frac{a^2 + ab + b^2}{2}$$

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) = \frac{a^2 + ab + b^2}{2} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

Среднеквадратическое отклонение

Дисперсия случайной величины имеет размерность, равную квадрату размерности самой величины

Чтобы вернуться к исходной размерности, из дисперсии часто извлекают корень и работают со среднеквадратическим отклонением:

$$\sigma(X) = \sqrt{\text{Var}(X)}$$

Свойства дисперсии

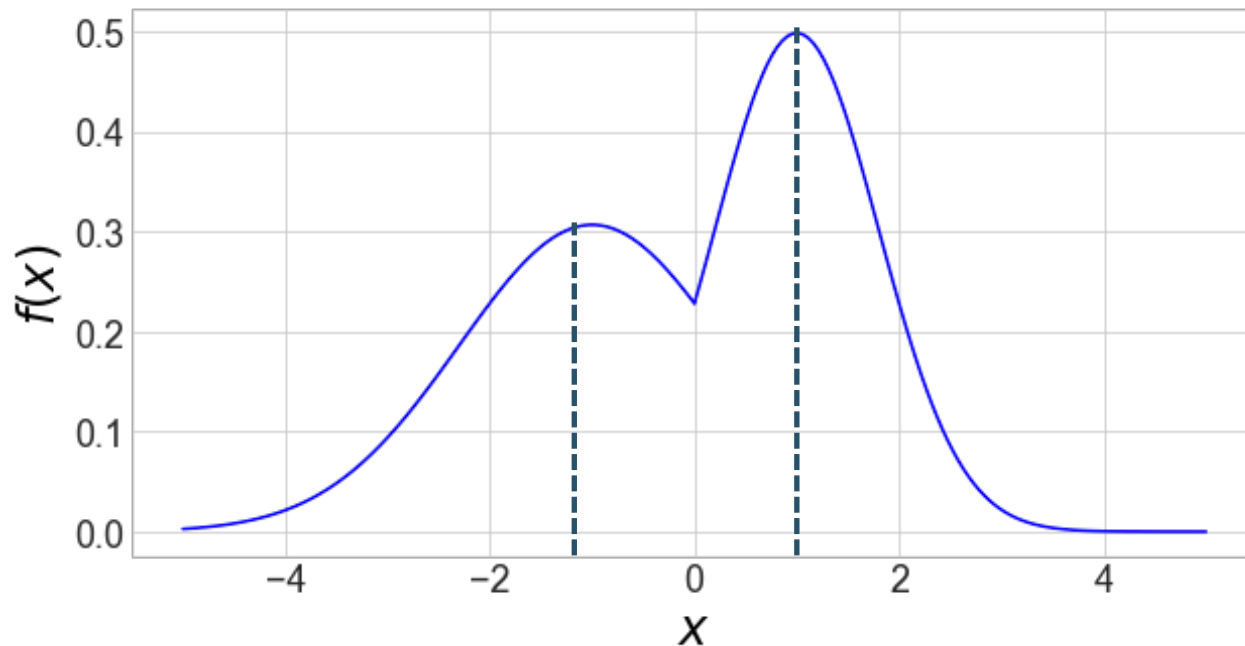
X, Y — случайные величины a — константа

1. $Var(a) = 0$
2. $Var(X + Y) = Var(X) + Var(Y)$, если независимы
3. $Var(X - Y) = Var(X) + Var(Y)$, если независимы
4. $Var(a \cdot X) = a^2 \cdot Var(X)$
5. Дисперсия случайной величины — не случайна

Мода

Мода случайной величины – значение, которому соответствует **наибольшая вероятность** (для дискретной случайной величины) и локальный **максимум плотности** распределения (для непрерывной случайной величины)

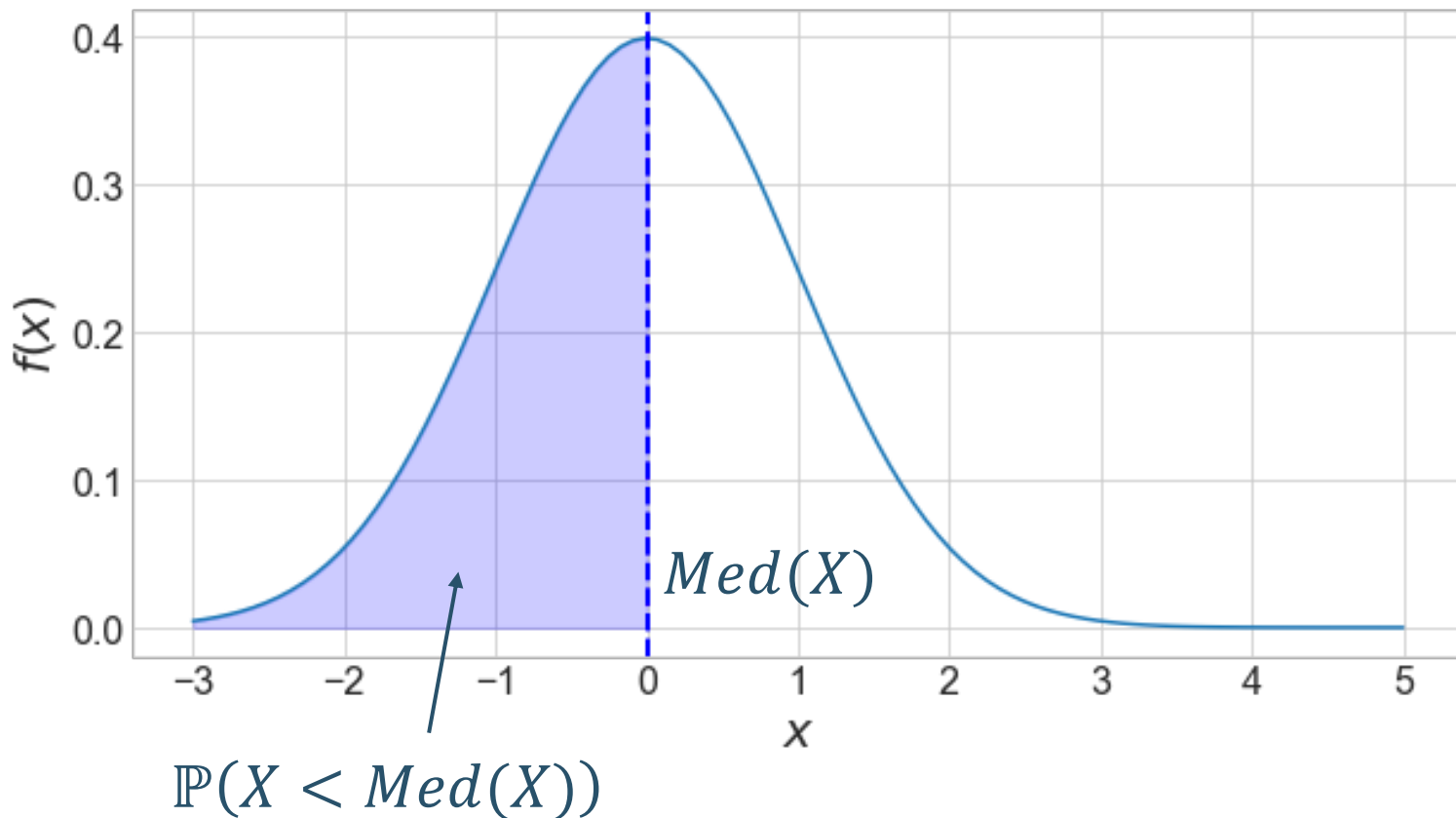
На практике встречаются мультимодальные распределения



Медиана

Медиана случайной величины – такое её значение, что

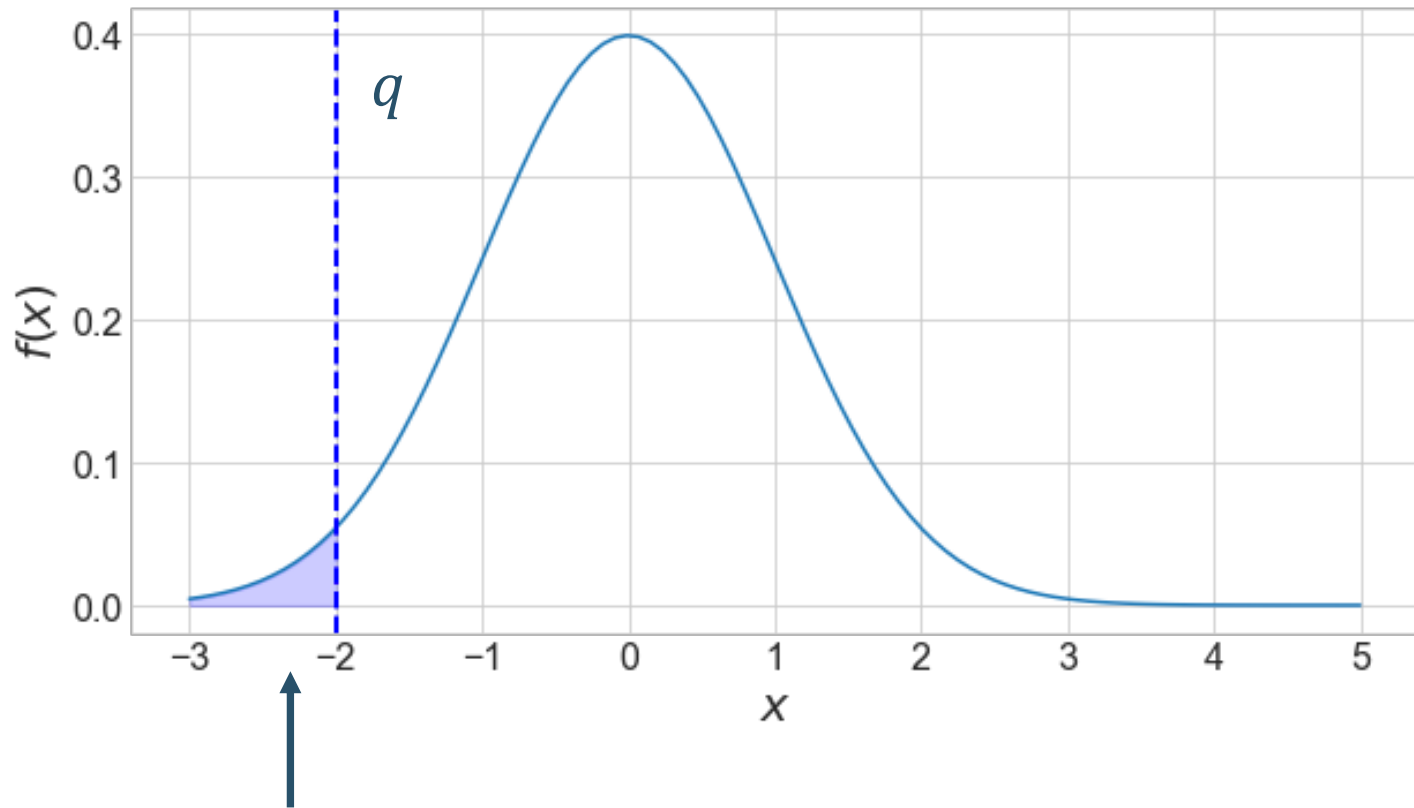
$$\mathbb{P}(X < \text{Med}(X)) = \mathbb{P}(X > \text{Med}(X)) = 0.5$$



Квантиль

Квантиль уровня γ – это такое число q , что

$$\mathbb{P}(X \leq q) = \gamma$$



Вероятность попасть в хвост равна γ

Резюме

- Мы вспомнили основные определения из теории вероятностей
- Мы поговорили про свойства математических ожиданий и дисперсий

Какими бывают случайные величины

Распределение Бернулли

- Пол родившегося ребёнка

	мальчик	девочка
X	0	1
$\mathbb{P}(X = k)$	$1 - p$	p

Распределение Бернулли:

$$X \sim \text{Bern}(p)$$

$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot (1 - p) = p$$

$$\text{Var}(X) = E(X^2) - E^2(X) = p - p^2 = p \cdot (1 - p)$$

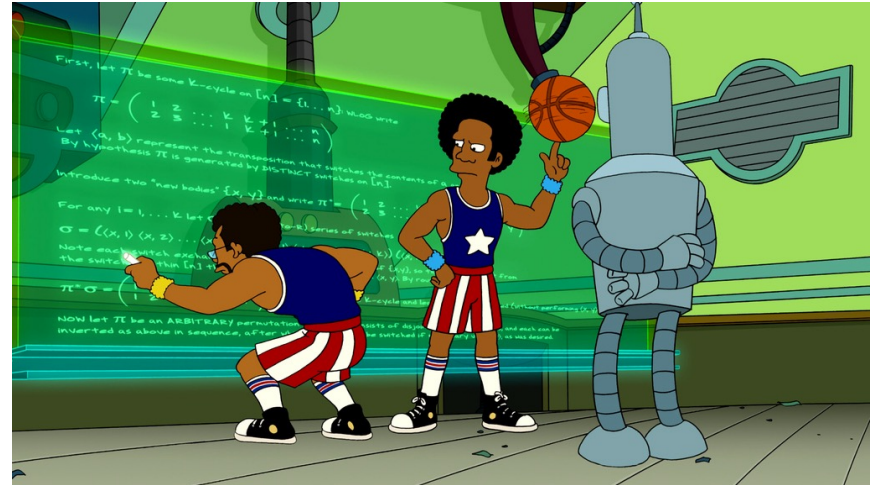
Биномиальное распределение

- Число попаданий в баскетбольную корзину

Биномиальная случайная величина: $X \sim \text{Bin}(p, n)$

n — число испытаний

p — вероятность успеха



Futurama s03 e14. Автор Мэтт Грейнинг. FOX Network.

$$\mathbb{P}(X = k) = C_n^k \cdot p^k (1 - p)^{n-k}$$

k принимает значения от 0 до n

Число сочетаний из n объектов по k :

$$C_n^k = \frac{n!}{(n - k)! \cdot k!}$$

n – количество объектов

k – количество сочетаний

Пример: В группе 20 человек, чтобы выполнить проект по python им нужно разбиться на команды по 3 человека. Посчитайте сколько различных команд можно составить из студентов группы?

$$\begin{aligned} C_{20}^3 &= \frac{20!}{(20 - 3)! \cdot 3!} = \frac{1 \cdot 2 \dots 20}{(1 \cdot 2 \dots 17) \cdot (1 \cdot 2 \cdot 3)} = \frac{18 \cdot 19 \cdot 20}{1 \cdot 2 \cdot 3} = \\ &= 6 \cdot 19 \cdot 10 = 1140 \end{aligned}$$

Биномиальное распределение

$$Y_i \sim \text{Bern}(p)$$

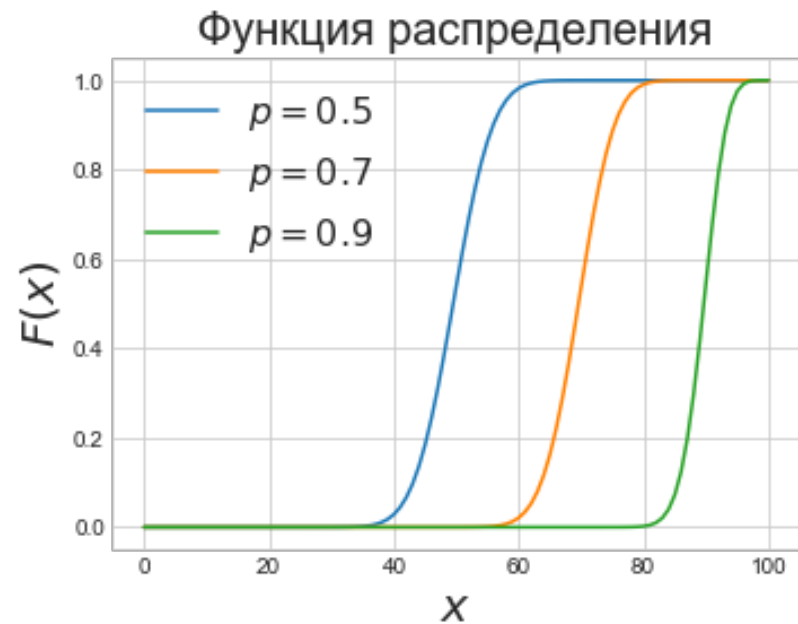
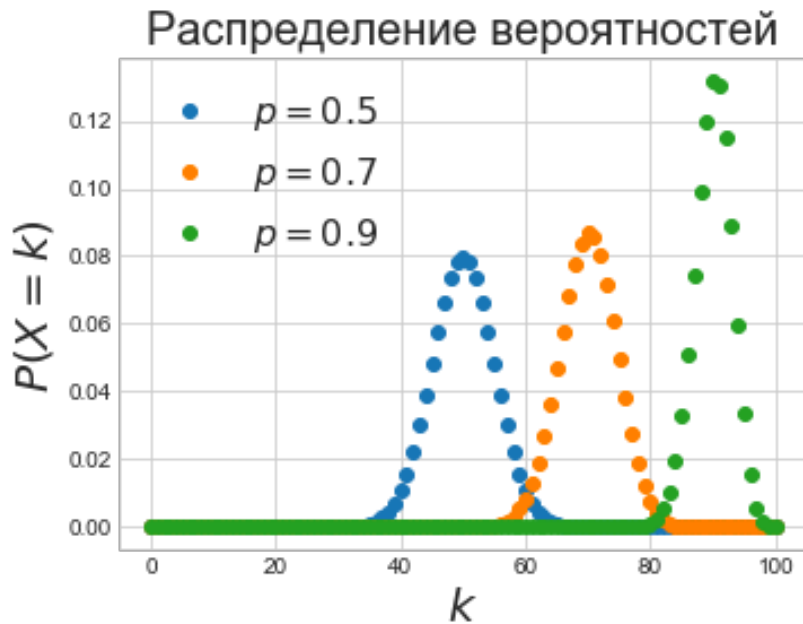
$$\mathbb{E}(X) = n \cdot p$$

$$X = Y_1 + \dots + Y_n$$

$$\text{Var}(X) = n \cdot p \cdot (1 - p)$$

$$X \sim \text{Bin}(p, n)$$

$$\mathbb{P}(X = k) = C_n^k \cdot p^k (1 - p)^{n-k}$$



Геометрическое распределение

- Номер броска, когда произошло первое попадание в корзину

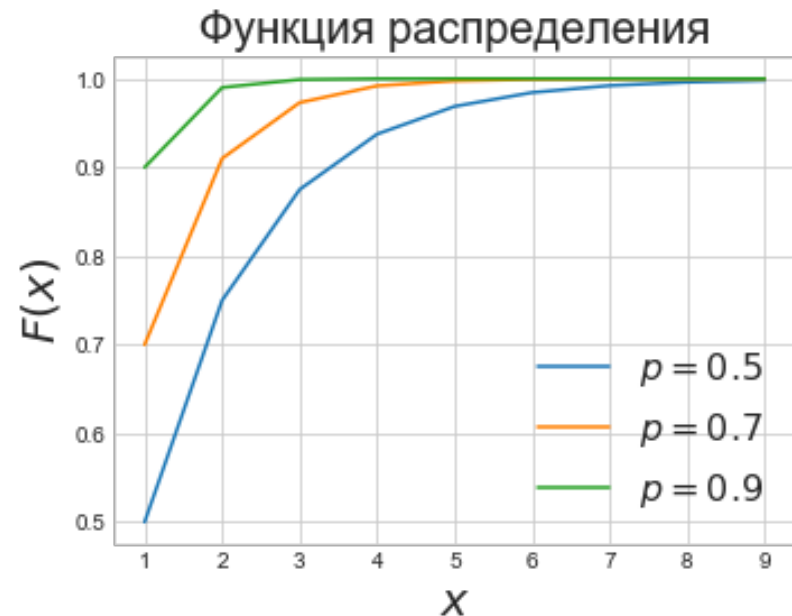
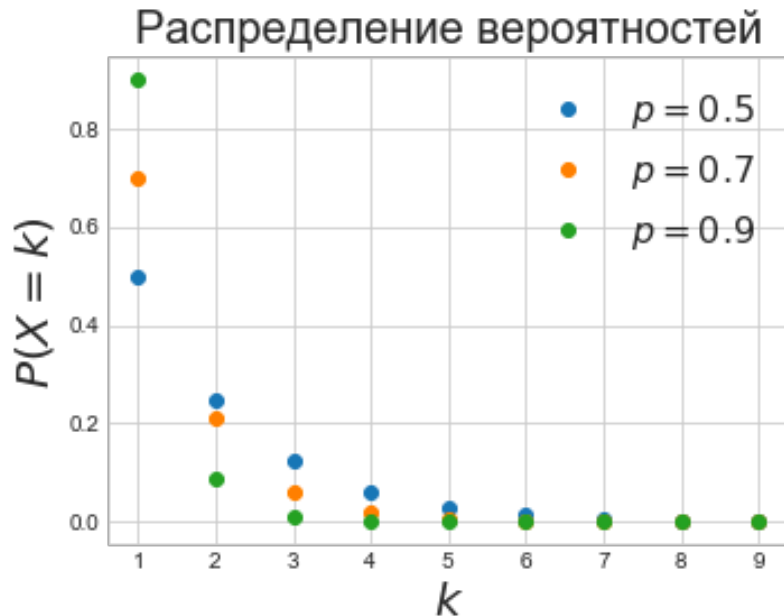
$$\mathbb{E}(X) = \frac{1}{p}$$

Геометрическая случайная величина: $X \sim \text{Geom}(p)$

$$\text{Var}(X) = \frac{1-p}{p^2}$$

p – вероятность успеха

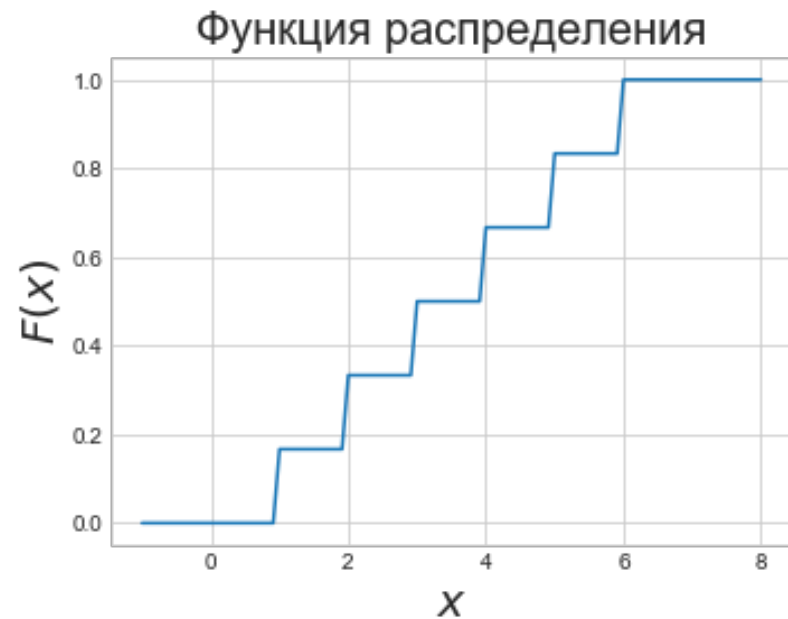
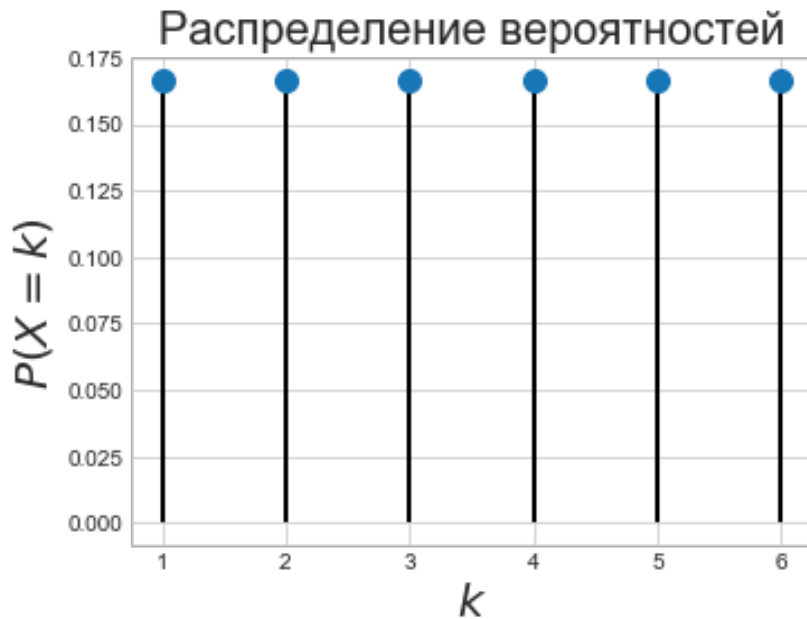
k принимает значения $1, 2, 3, \dots$ $\mathbb{P}(X = k) = p \cdot (1 - p)^{k-1}$



Произвольное дискретное распределение

- Подбрасывание игральной кости

X	1	2	3	4	5	6
$\mathbb{P}(X = k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$



Счётчики

- Число людей в очереди
- Число лайков под фото
- Число автобусов, проехавших за час мимо остановки

Пуассоновская случайная величина: $X \sim Poiss(\lambda)$

Распределение Пуассона хорошо описывает счётчики



♥ Нравится 15

Распределение Пуассона

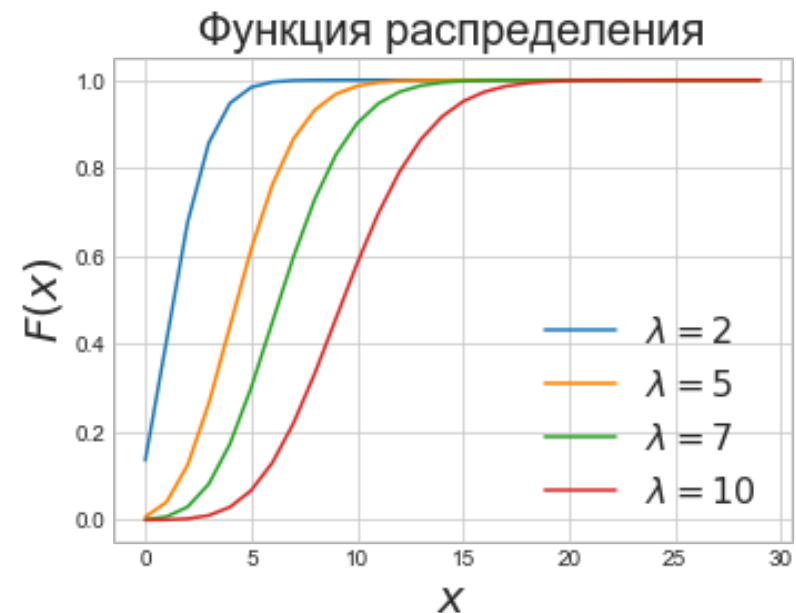
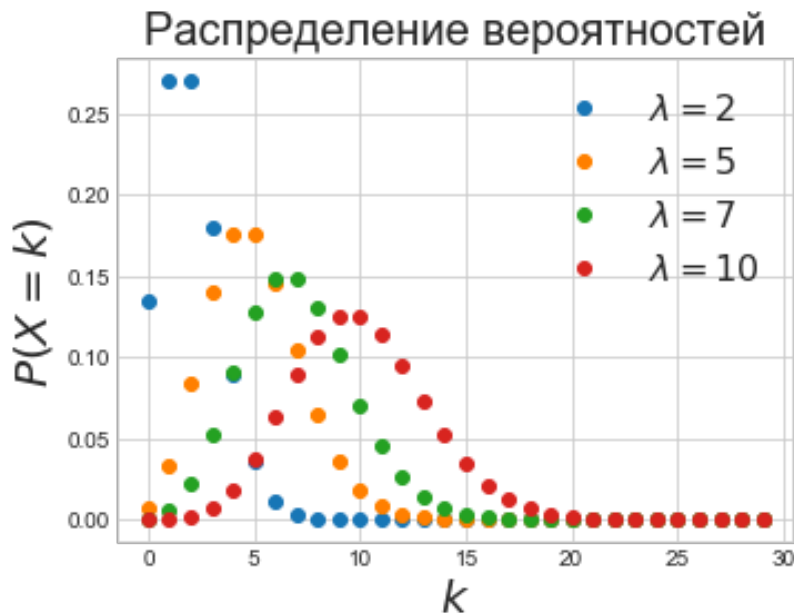
$$\mathbb{P}(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

$$X \sim \text{Poiss}(\lambda)$$

- Параметр λ интерпретируется как интенсивность потока событий
- k принимает значения $0, 1, 2, \dots$

$$\text{Var}(X) = \lambda$$

$$\mathbb{E}(X) = \lambda$$



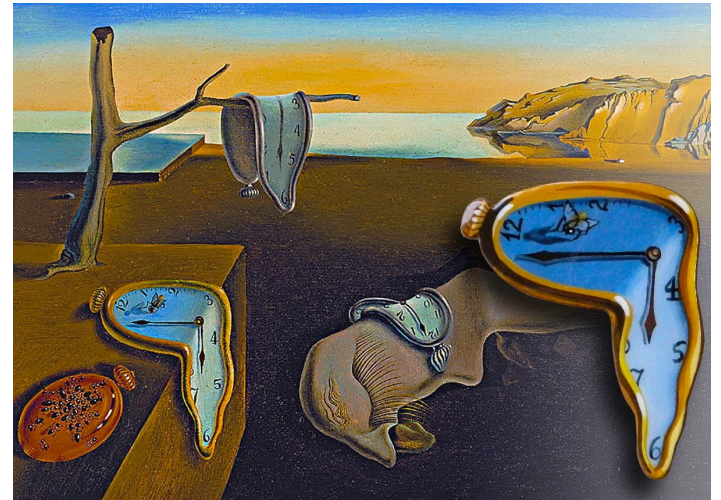
Время до ...

- Время ожидания трамвая
- Время до прихода нового человека в очередь
- Время до поломки механизма

**Экспоненциальная
случайная величина:**

$$X \sim \text{Exp}(\lambda)$$

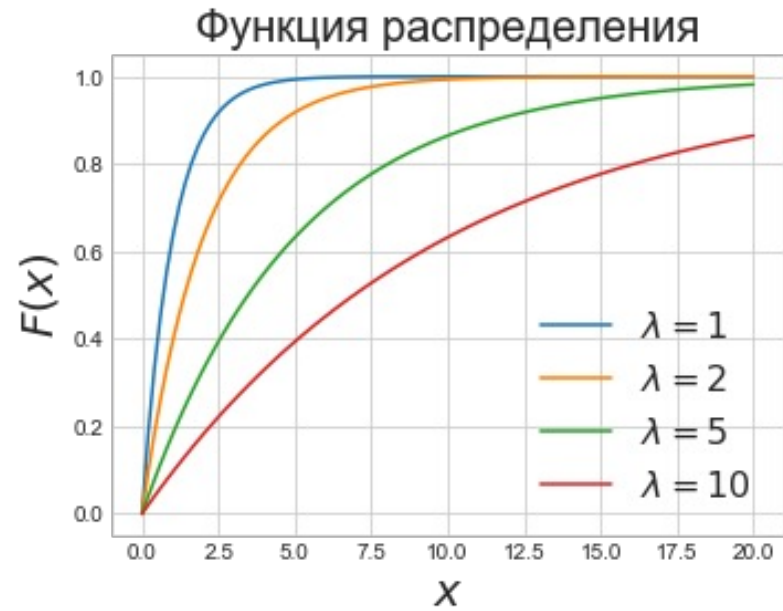
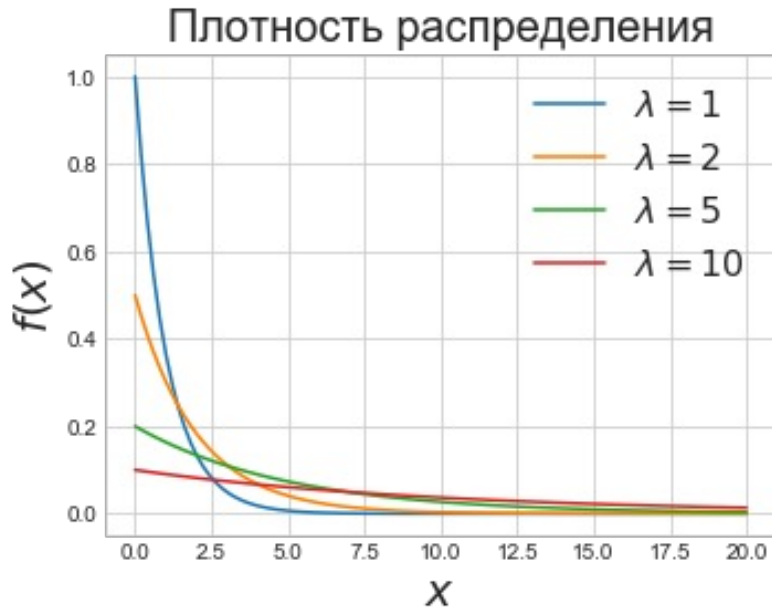
- Интервалы времени между событиями
- Модели времени жизни



Экспоненциальное распределение

$$f_X(x) = \lambda \cdot e^{-\lambda \cdot x}, x \geq 0$$

$$F_X(x) = 1 - e^{-\lambda \cdot x}, x \geq 0$$



У экспоненциального распределения нет памяти. Автобусы приходят на остановку случайно. Время, которое осталось ждать не зависит от того, сколько уже прошло времени.

$$\mathbb{E}(X) = \frac{1}{\lambda}$$

$$\text{Var}(X) = \frac{1}{\lambda^2}$$

Равномерное распределение

- Время рождения ребёнка

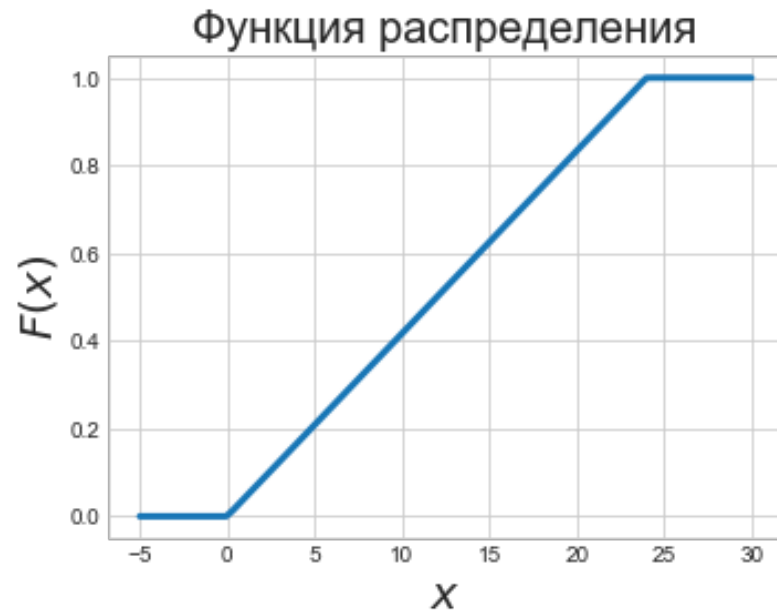
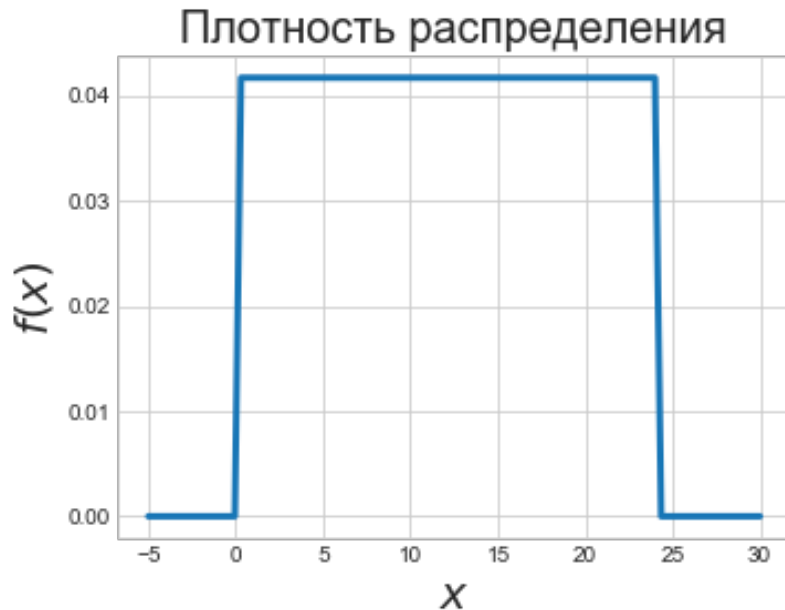
Равномерная случайная величина: $X \sim U[a; b]$

$$f_X(x) = \frac{1}{b - a}, x \in [a; b]$$

$$\mathbb{E}(X) = \frac{a + b}{2}$$

$$\text{Var}(X) = \frac{(b - a)^2}{12}$$

$$F_X(x) = \frac{x - a}{b - a}, x \in [a; b]$$



Нормальное распределение

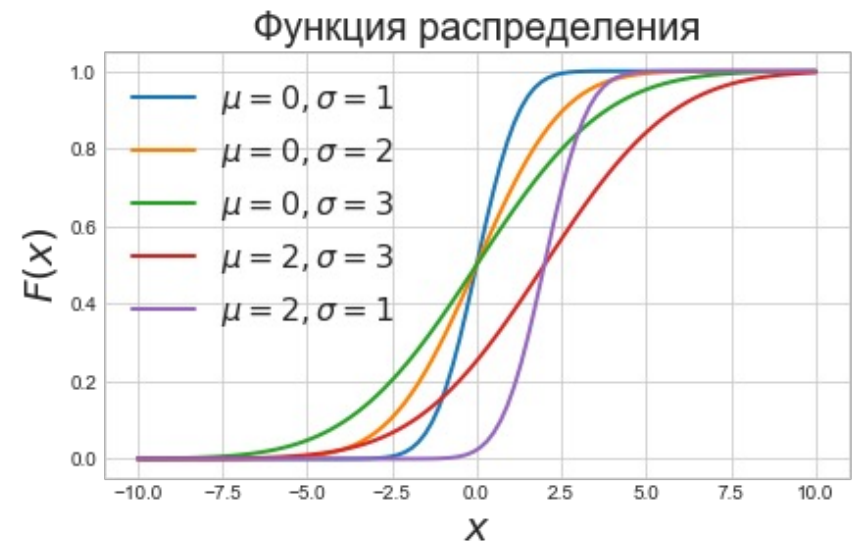
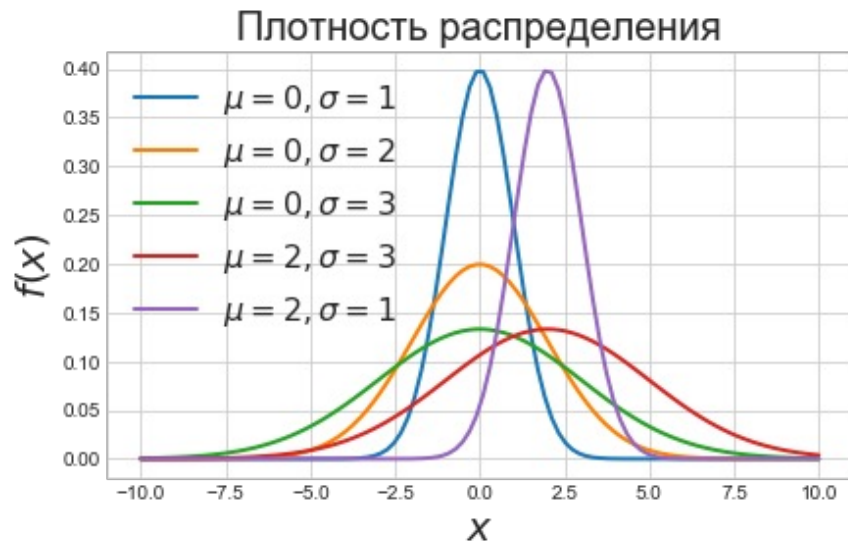
- Погрешность весов

Нормальная случайная величина:

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

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

Функцию распределения
нельзя найти в
аналитическом виде,
интеграл не берётся




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

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

Распределения бывают разными

Случайная величина	Распределение
Пол ребенка	$Bern(p)$
Попадания в корзину	$Binom(n, p)$
Число бросков до первого попадания	$Geom(p)$
Число людей в очереди	$Poiss(\lambda)$
Подбрасывание кости	Дискретное
Время между событиями	$Exp(\lambda)$
Время до поломки часов	$Exp(\lambda)$
Время рождения ребенка	$U[0; 24]$
Погрешность весов	$N(0, \sigma^2)$

Резюме

- Моделировать различные процессы можно с помощью различных законов распределения
- Наиболее подходящий закон выбирается с помощью здравого смысла
-  Мы перечислили лишь одни из вариантов моделирования. Эти распределения не истина в последней инстанции
- Все предпосылки, связанные с выбранным законом, должны проверяться по данным, в будущем мы научимся это делать

Резюме

Описательные статистики:

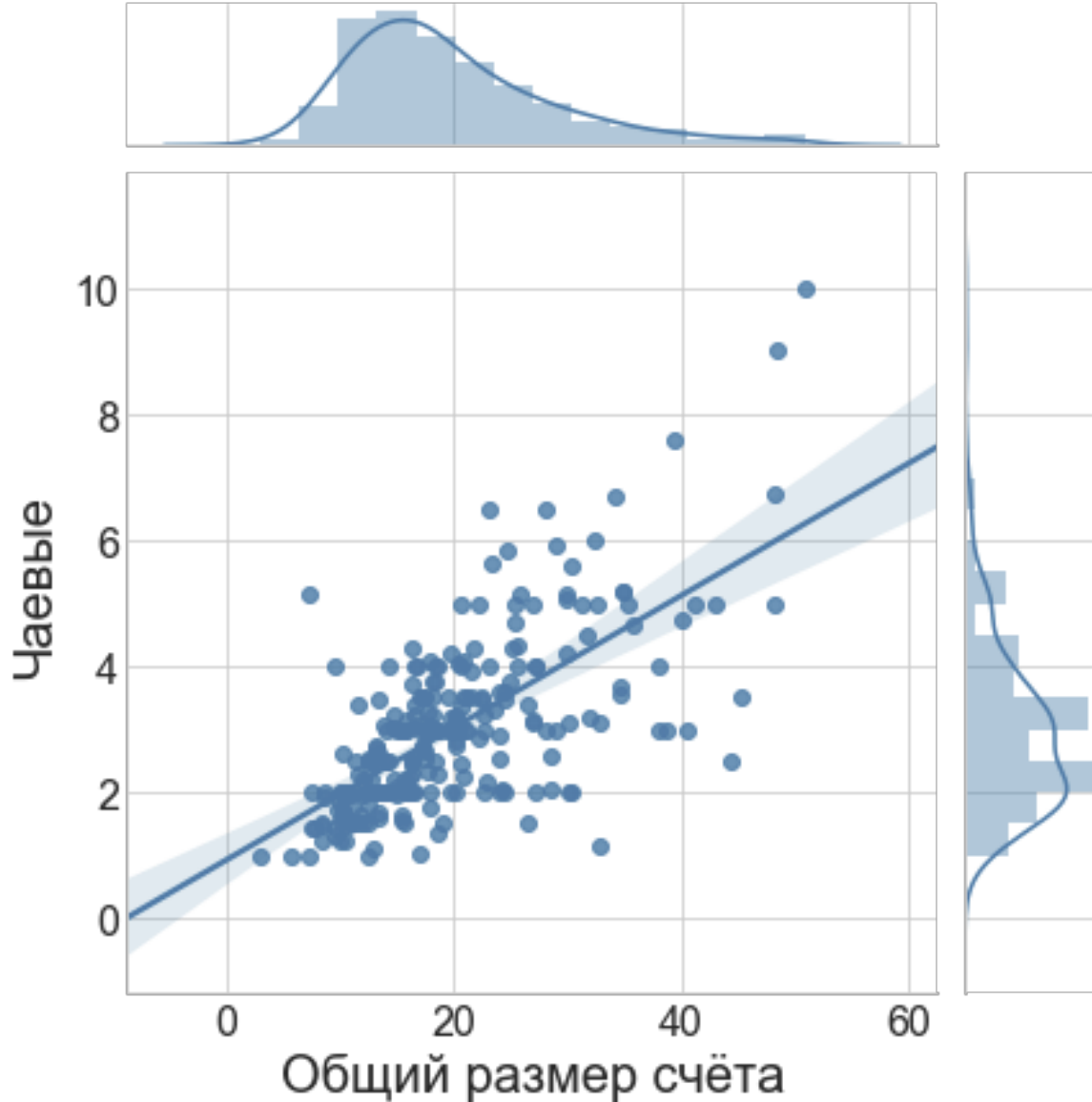
Теоретическая величина	Выборочный аналог
Математическое ожидание	Выборочное среднее
Дисперсия	Выборочная дисперсия
Квантиль	Перцентиль
Медиана	Выборочная медиана
Мода	Выборочная мода

Описание распределения:

Теоретическая величина	Выборочный аналог
Функция распределения	Эмпирическая функция распределения
Плотность распределения	Гистограмма

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

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

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

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

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, b — константы

7. Если случайные величины независимы,

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

8. Обратное неверно. Если ковариация равна нулю, случайные величины могут быть зависимы. Исключение — многомерное нормальное распределение.

9. Если X и Y зависимы, тогда

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

$$Var(aX + bY) = a^2 \cdot Var(X) + b^2 \cdot Var(Y) + 2 \cdot a \cdot b \cdot Cov(X, Y)$$

$$Var(aX - bY) = a^2 \cdot Var(X) + b^2 \cdot Var(Y) - 2 \cdot a \cdot b \cdot Cov(X, Y)$$

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:

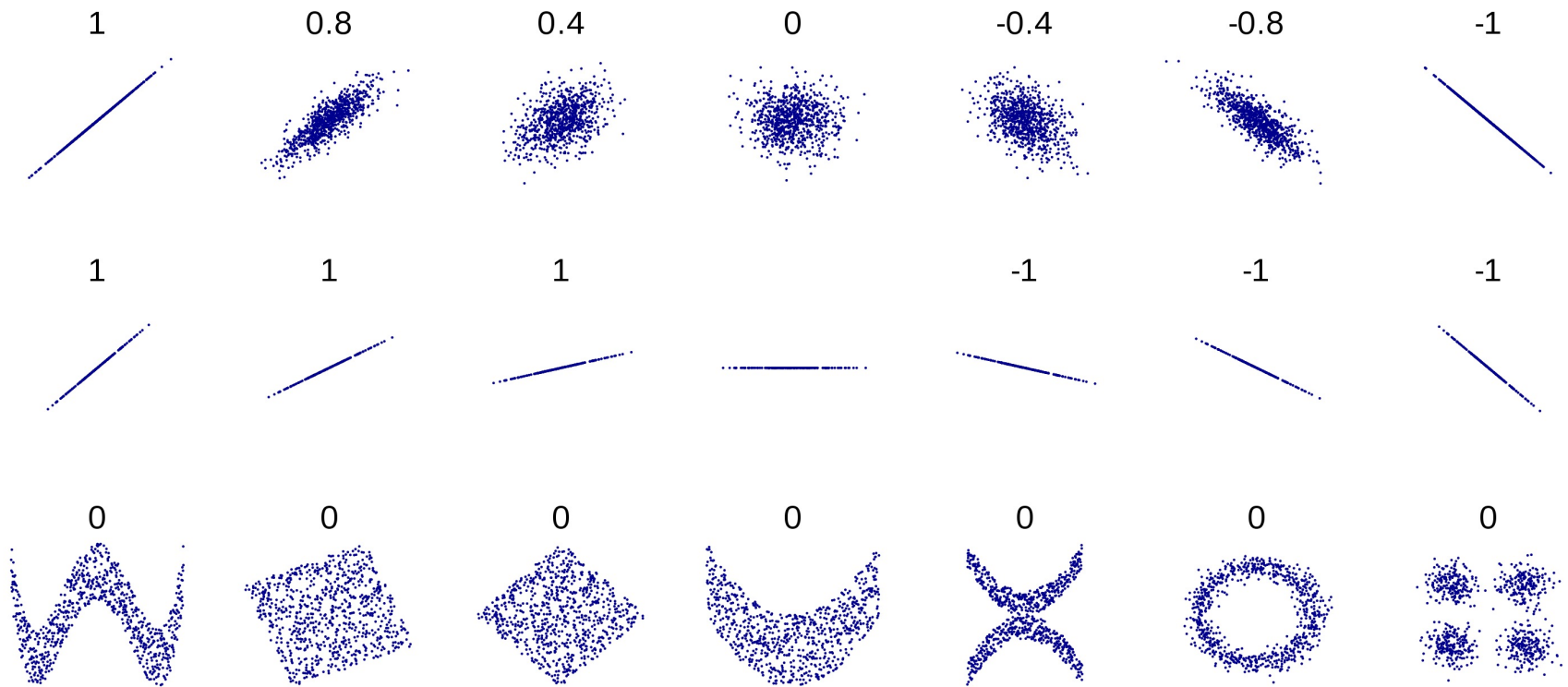
Выборочная ковариация:

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

Выборочная корреляция (корреляция Пирсона):

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

Корреляция Пирсона



Корреляция Пирсона



Корреляция Пирсона улавливает только линейную взаимосвязь и чувствительна к выбросам

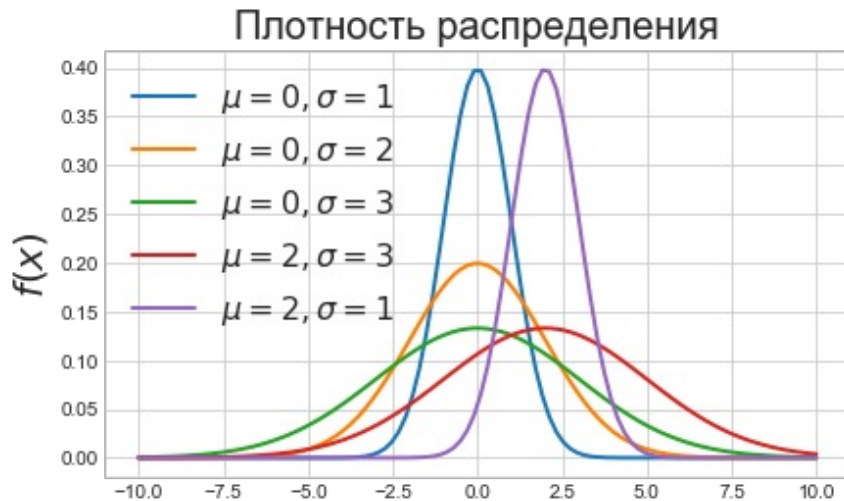
► Угадай корреляцию: <http://guessthecorrelation.com/>

Нормальное распределение

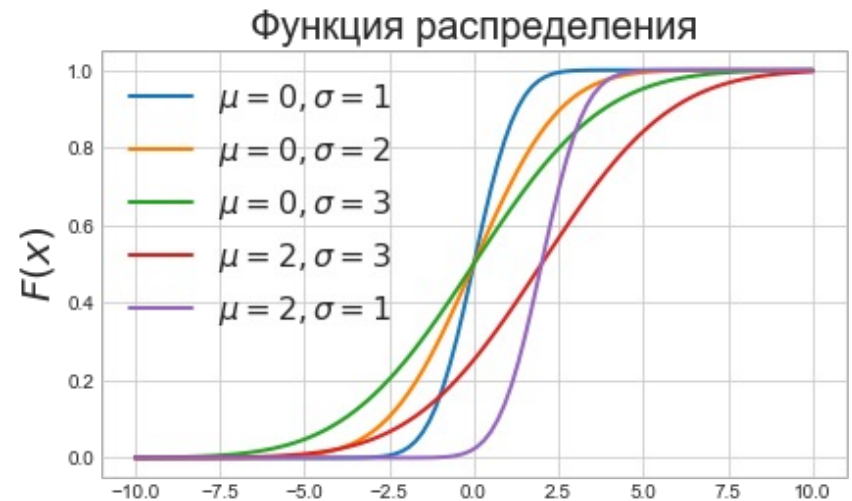
Нормальная случайная
величина: $X \sim N(\mu, \sigma^2)$

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

Функцию распределения
нельзя найти
в аналитическом виде,
интеграл не берётся



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




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

Многомерное нормальное

Многомерное нормальное


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



Математическое
ожидание

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

$$\mathbb{E}(X_2) = \mu_2$$



Ковариационная
матрица

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

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

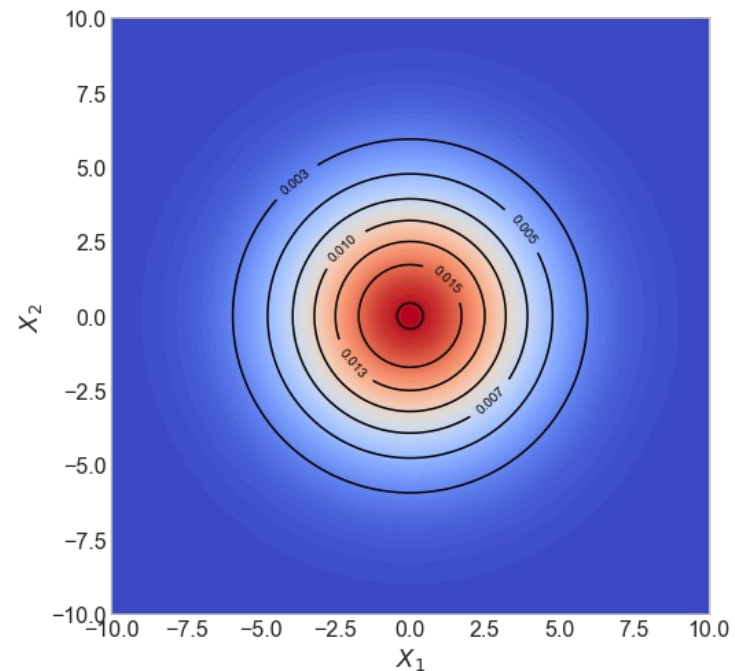
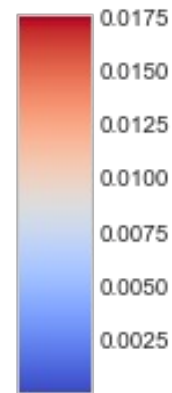
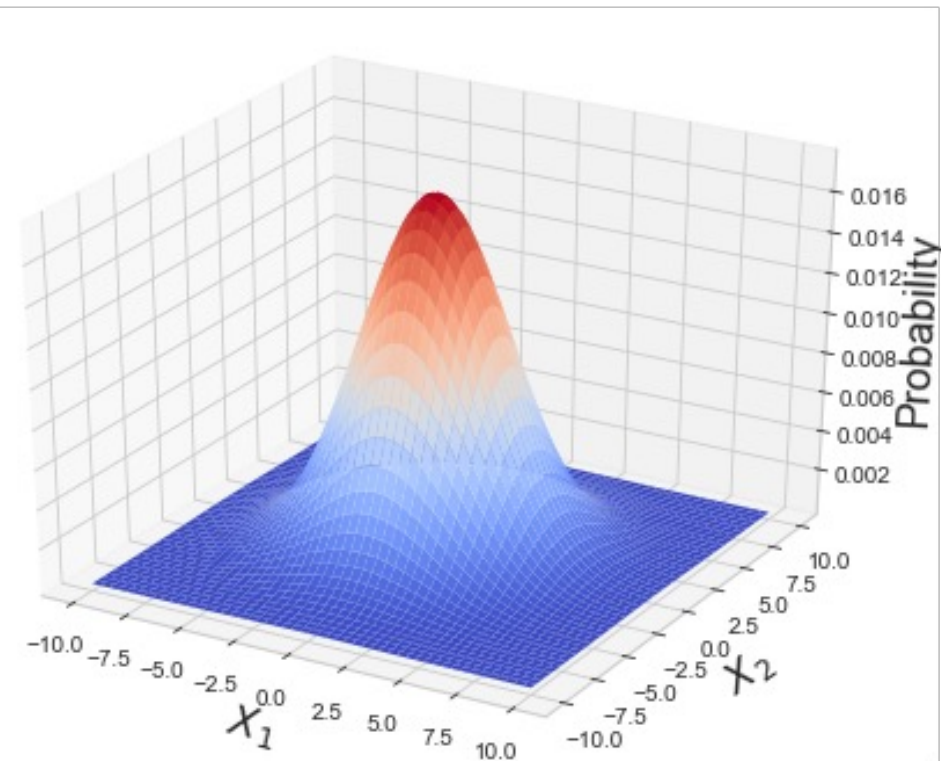
$$Cov(X_1, X_2) = \rho$$

Кратко пишут

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

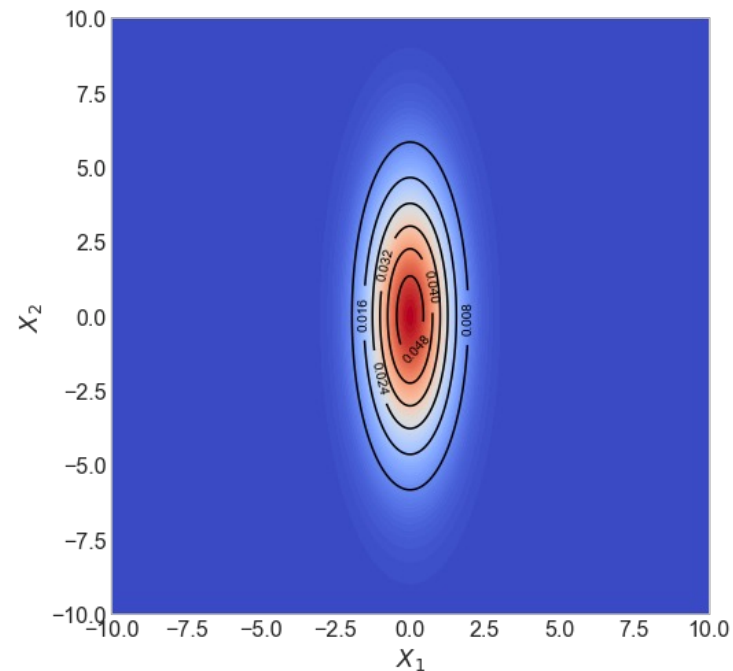
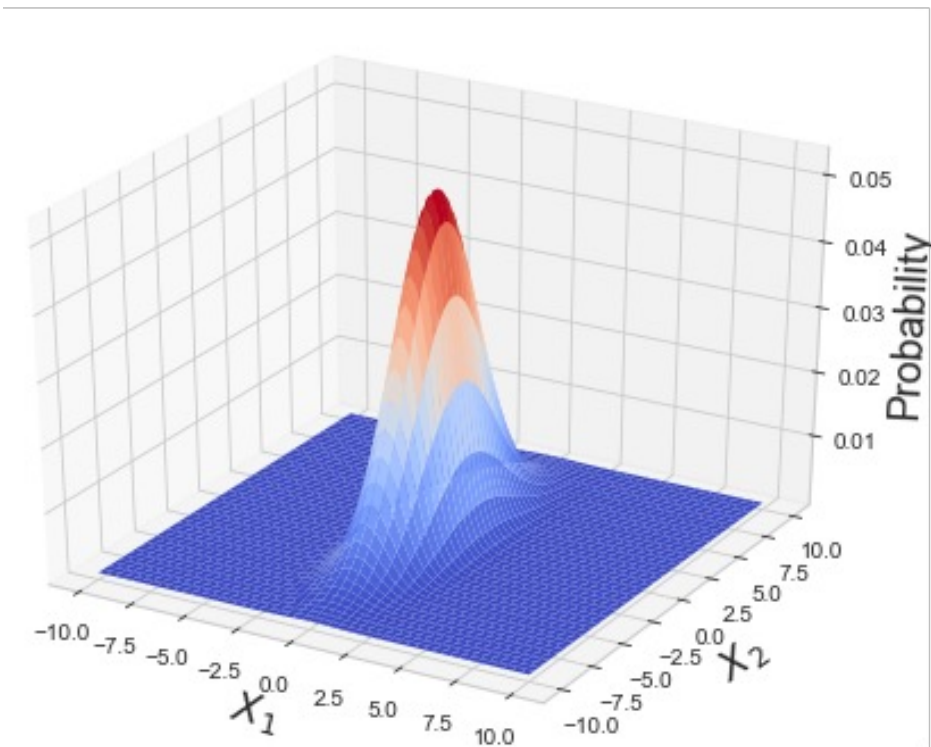
Многомерное нормальное

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



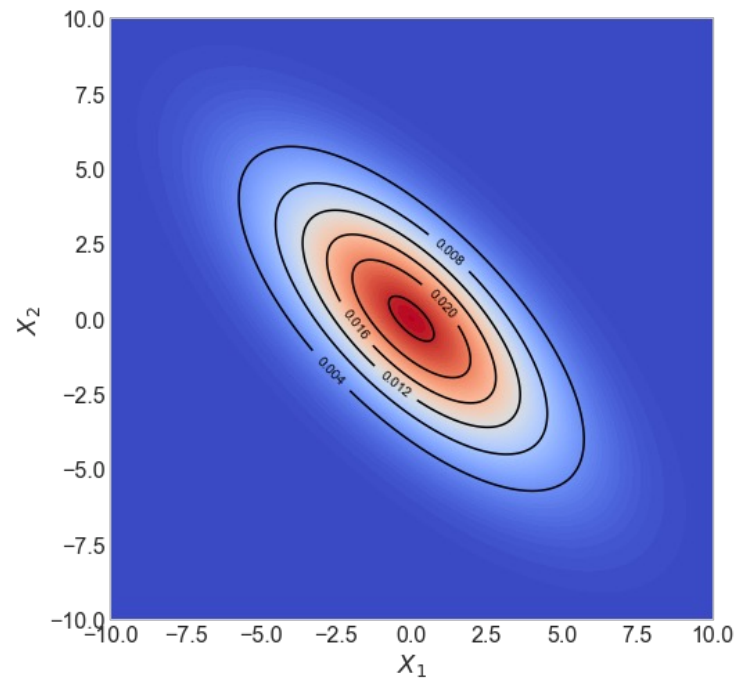
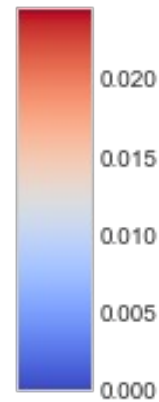
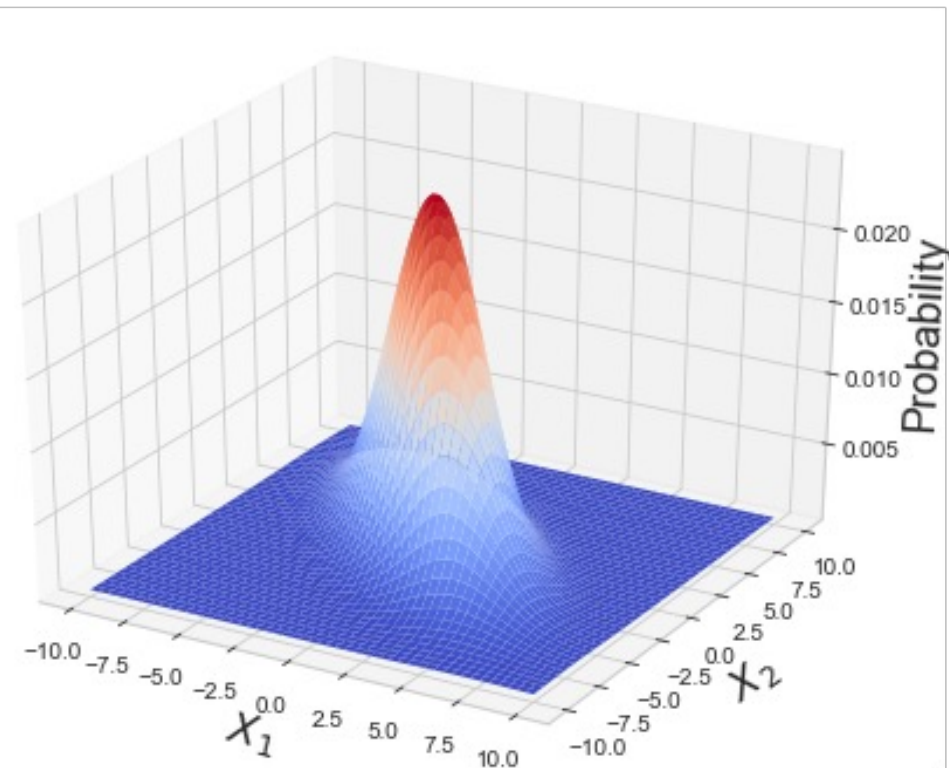
Многомерное нормальное

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



Многомерное нормальное

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



Многомерное нормальное

- По аналогии можно определить нормальное распределение для любой размерности

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

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

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

Резюме

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

**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, возникли пропуски
- Пропуски нужно как-то заполнить

Простые методы:

Замена средним значением / медианой / модой

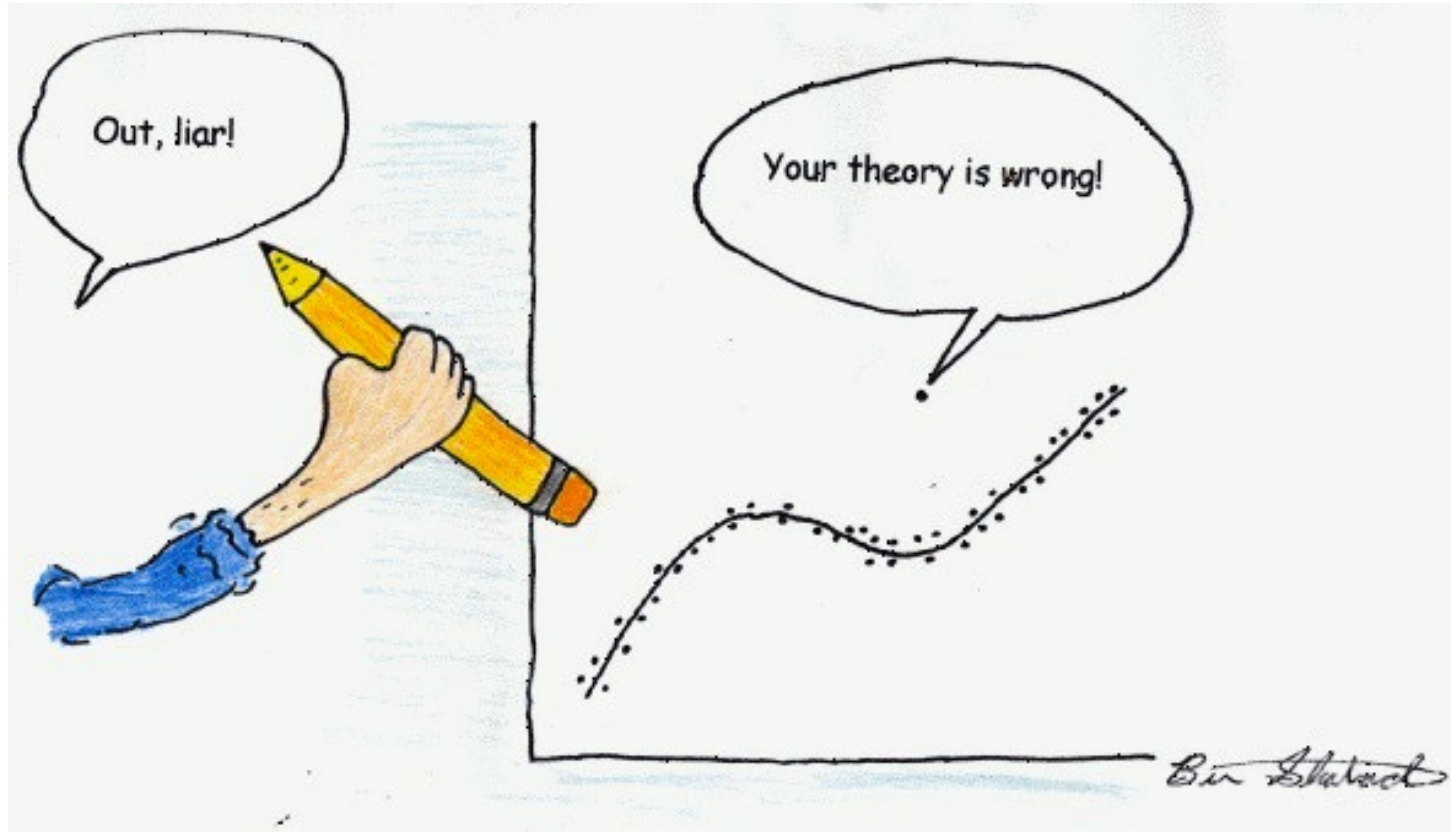
Сложные методы:

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



Среднее: \$ 80000

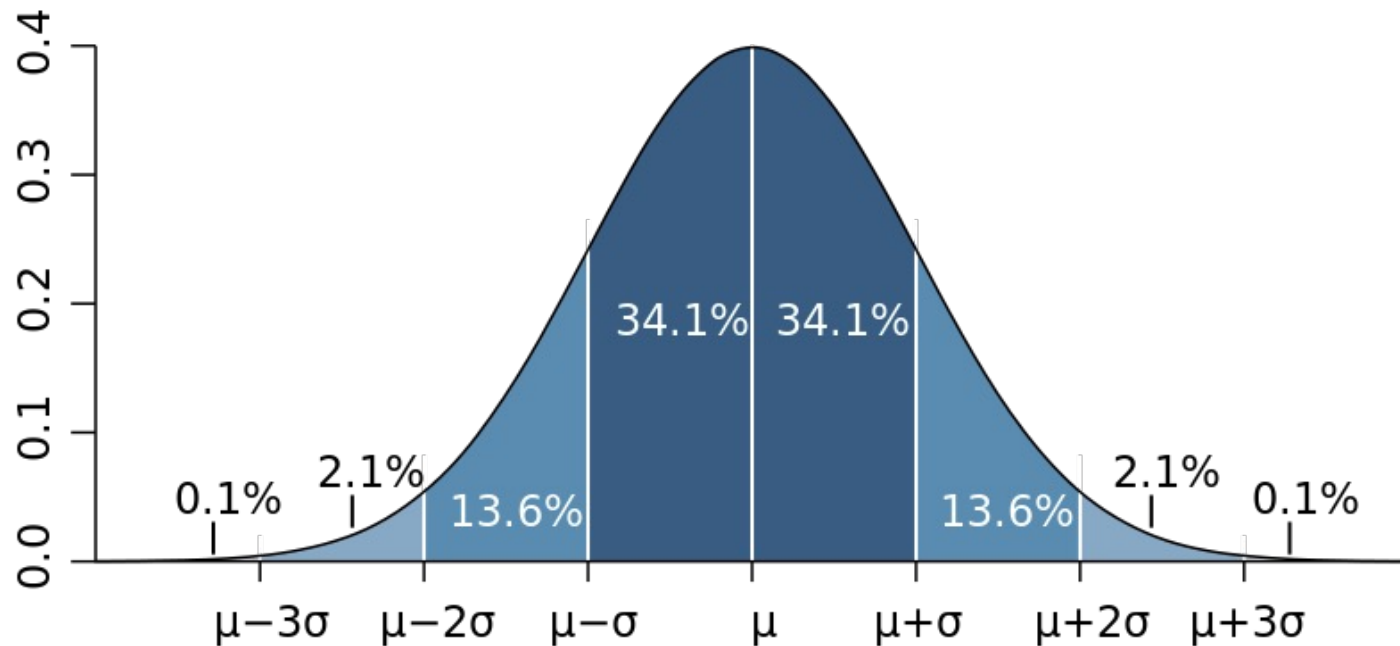
Медиана: \$ 30000

Высокая зарплата сильно исказит среднее, но не медиану

Поиск выбросов

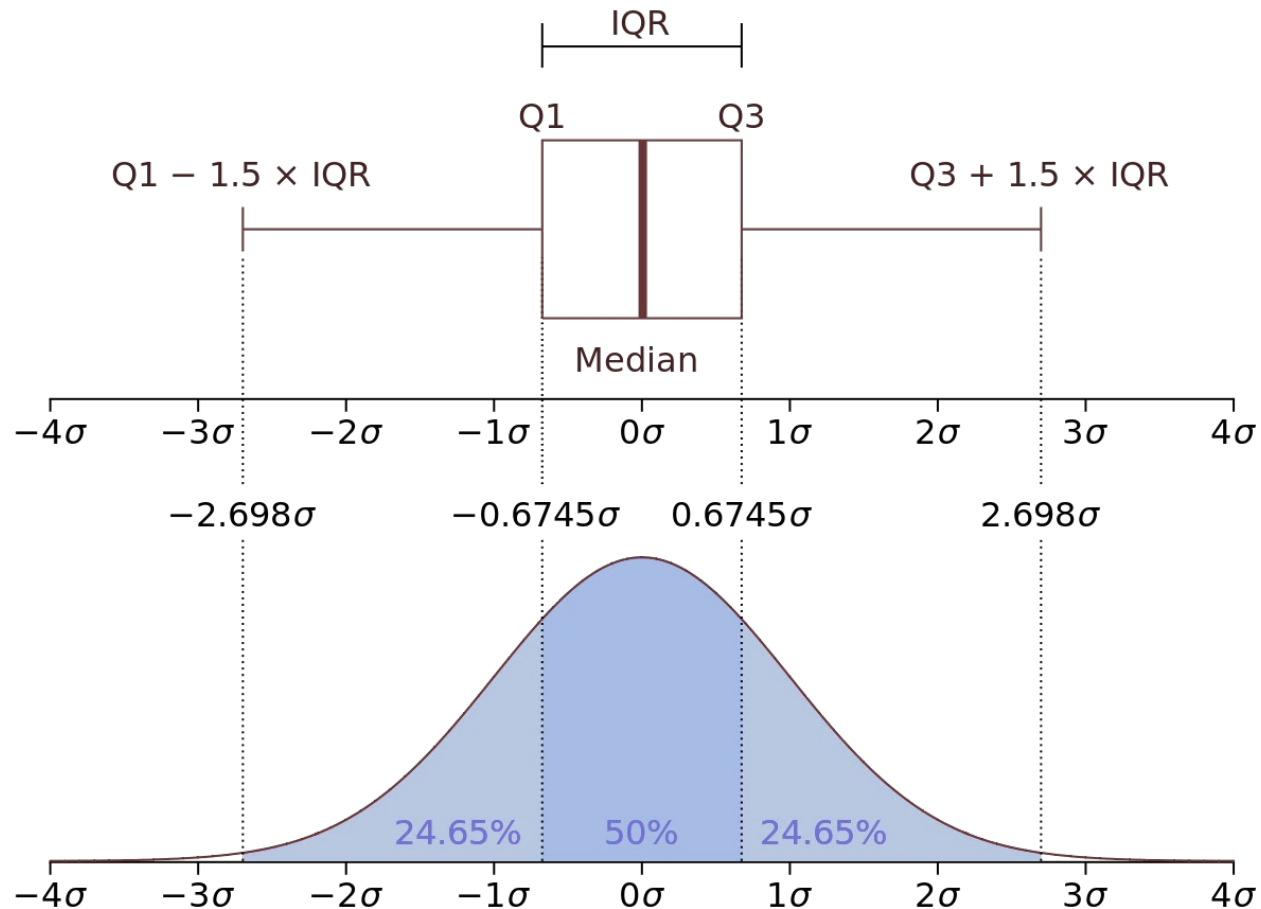
Правило трёх сигм: если данные распределены нормально и наблюдение оказалось за пределами интервала в три сигмы, это выброс

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



Поиск выбросов

Правило 1.5 интерквартильных размахов (IQR): если наблюдение оказались 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 не искажился бы, как среднее
- Можно придумывать статистики, основанные на медиане, которые будут нечувствительны к выбросам

Резюме

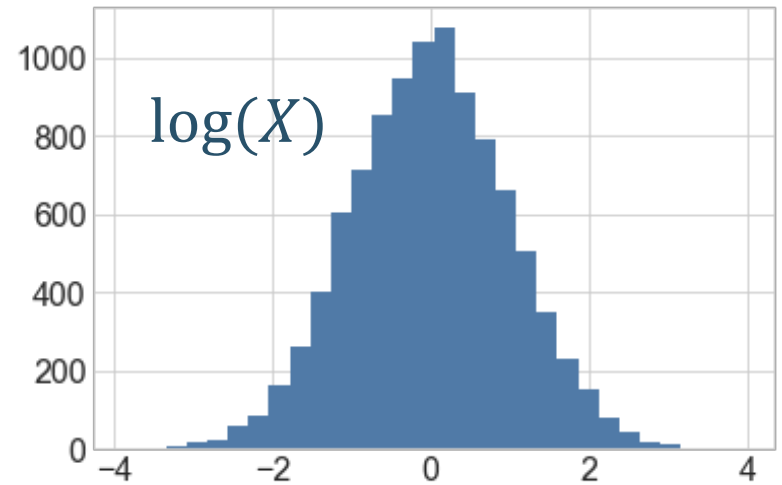
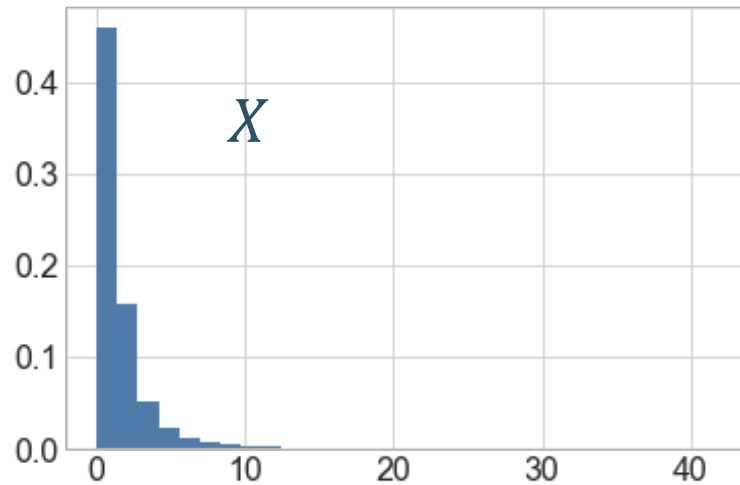
- Пропуски и выбросы – проблемы в данных, с которыми надо бороться
- Пропуски, если их мало, пытаются заполнять с помощью разных алгоритмов
- Выбросы либо сглаживают, либо выбрасывают из рассмотрения

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

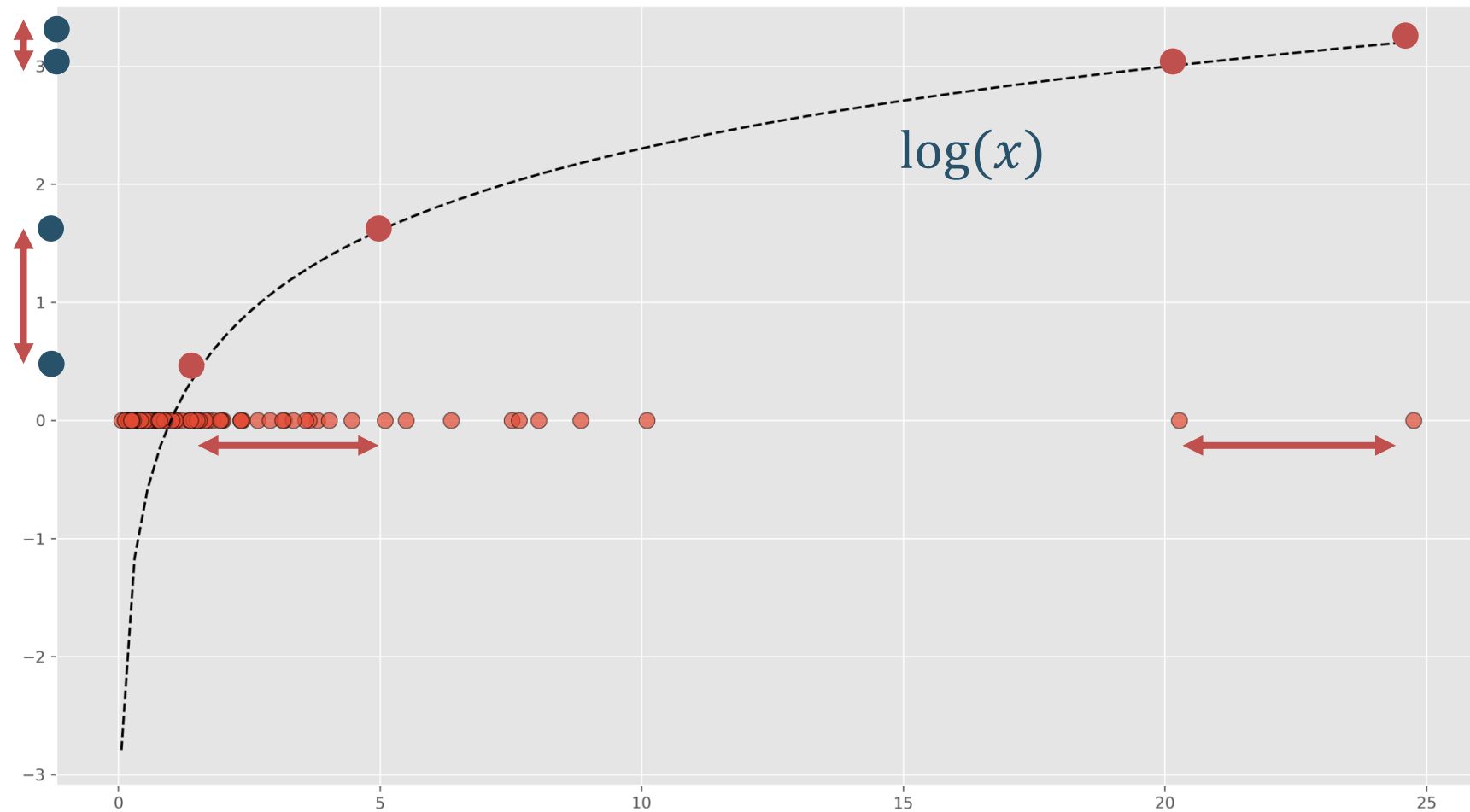


Логарифмирование значений позволяет
сгладить хвосты

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

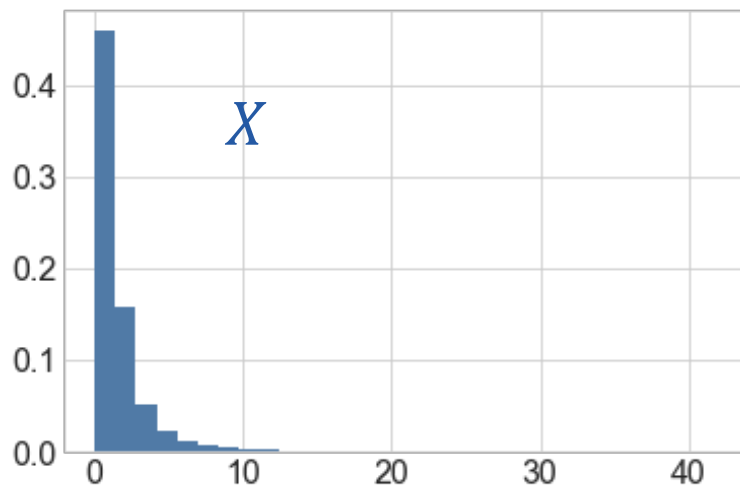


Логарифмирование

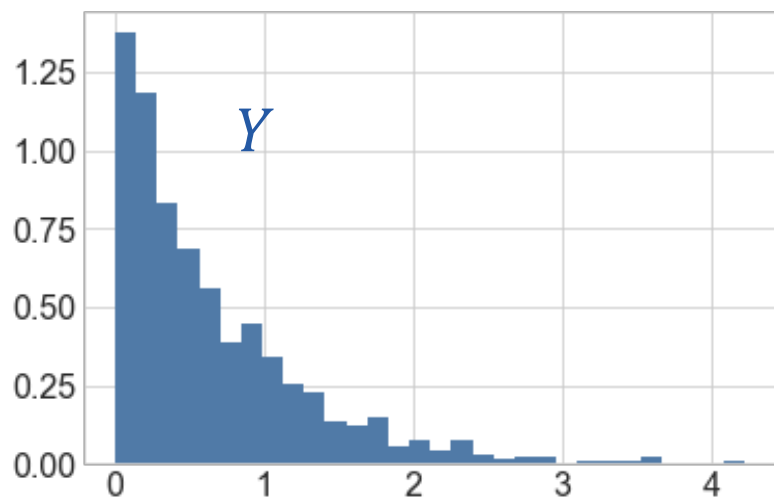
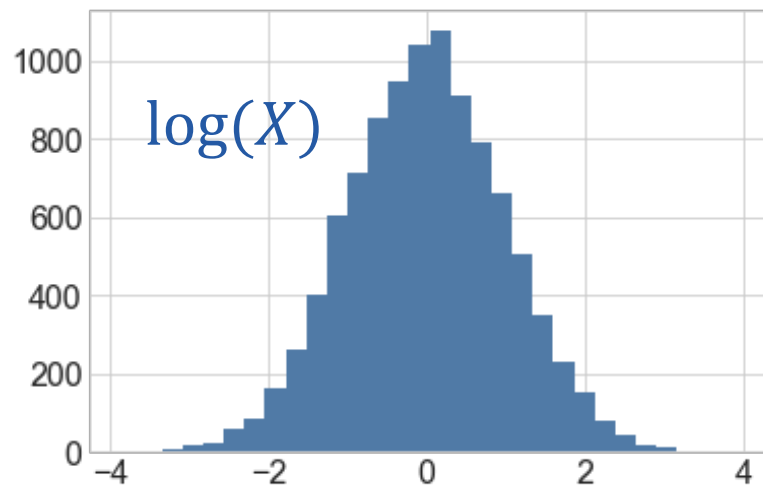
- Сгустки в начале оси абсцисс стали распределены более равномерно из-за того, что там логарифм растёт быстрее
- Расстояние между точками с большими значениями стало меньше, так как там логарифм растёт медленнее
- Чем правее мы движемся, тем медленнее растёт логарифм, скорость его роста это:

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

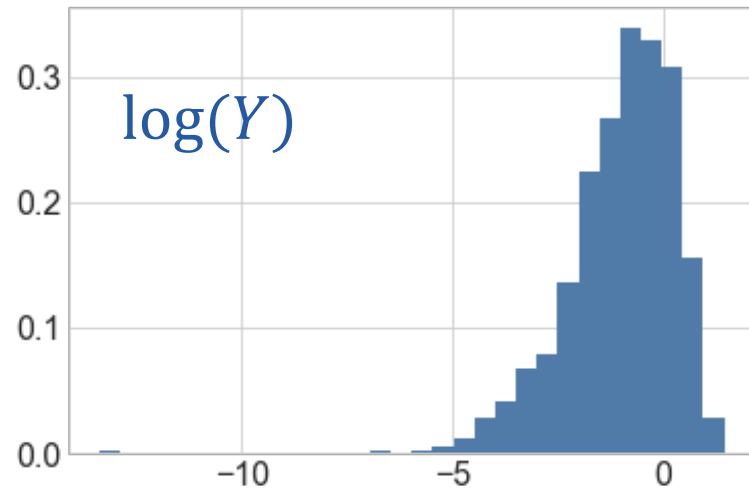
Повторим успех?



Логнормальное

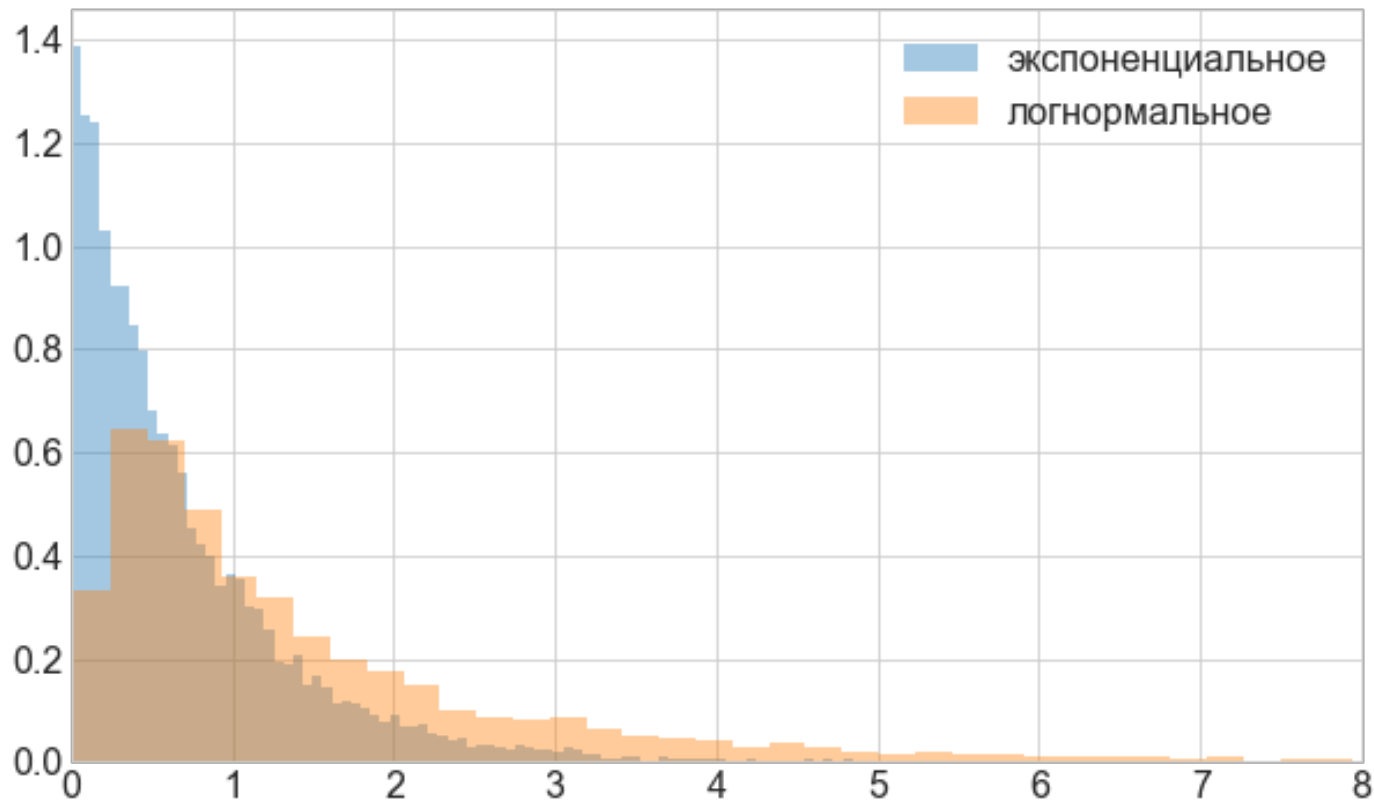


Экспоненциальное



Как повторить успех?

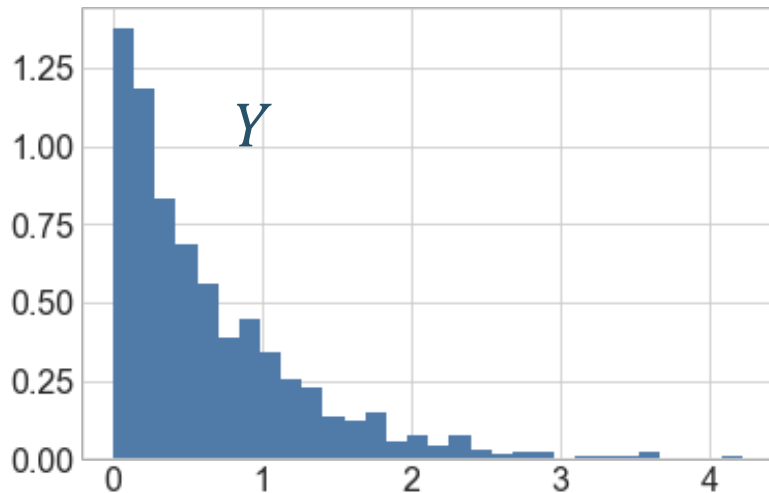
- Для экспоненциального распределения хвост легче, он не так резко убывает
- При его логарифмировании хвост появляется слева



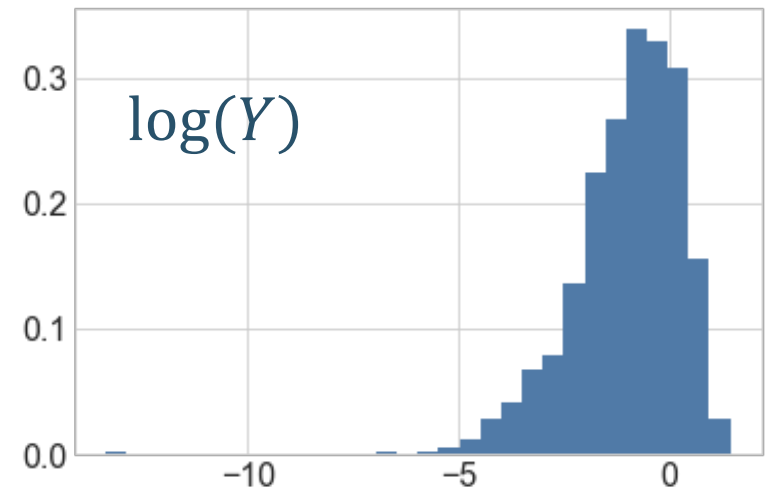
Как повторить успех?

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

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



Экспоненциальное



Как повторить успех?

- Найдём такое преобразование:

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

- Для удобства выберем конкретную константу и немного перепишем формулу:

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

Преобразование Бокса-Кокса

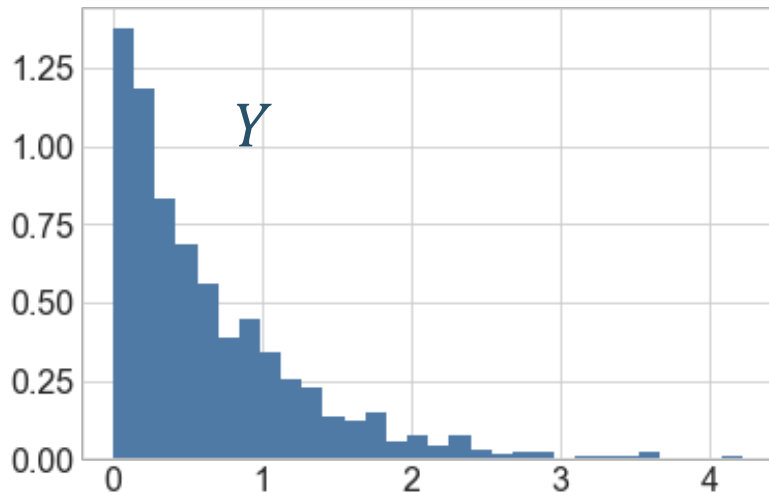
$$x_i^* = \begin{cases} \log(x), & p = 0 \\ \frac{x^p - 1}{p}, & 0 \leq p \leq 1 \end{cases}$$

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

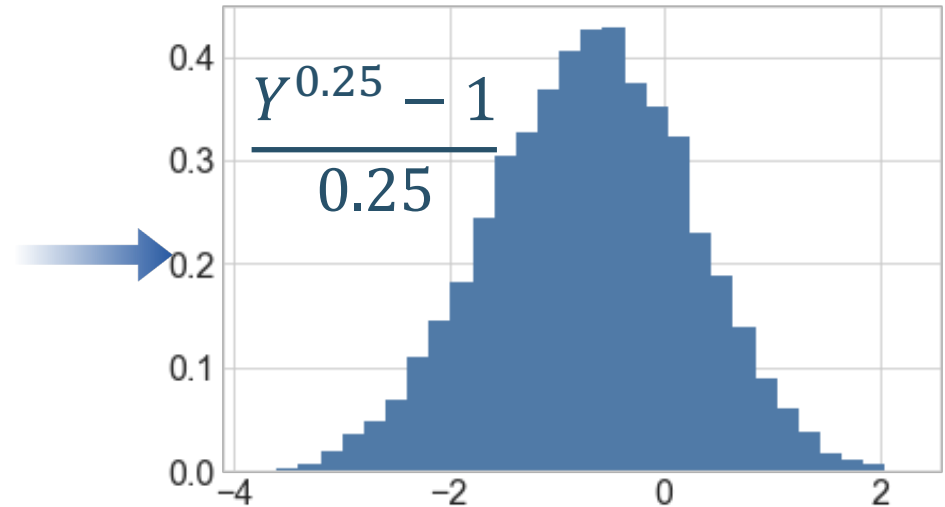
$$x_i^* = \begin{cases} \log(x + \alpha), & p = 0 \\ \frac{(x + \alpha)^p - 1}{p}, & 0 \leq p \leq 1 \end{cases}$$

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:

90 кг

1.9 м

Ярополк:

60 кг

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 одинаковому 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

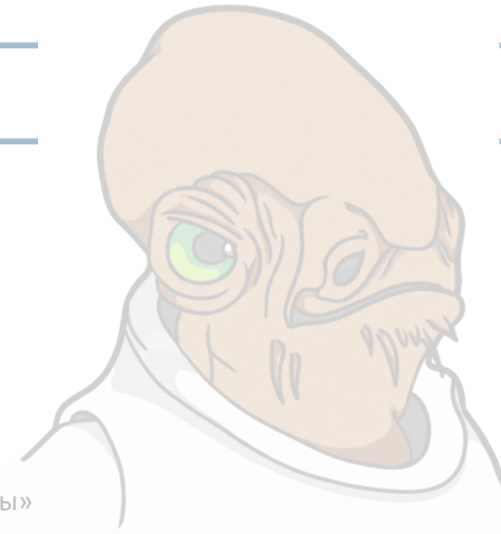
Бинарное кодирование (One Hot Encoding)

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

	х
0	Испания
1	Дача
2	Крым
3	Крым
4	Дача



	х_isp	х_da	х_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 (ONE)