

Распределения и описательные статистики

План

- Вспомним основные понятия теории вероятностей
- Поговорим про то, какими бывают распределения
- Научимся считать описательные статистики
- Разберёмся в выборках и их свойствах

Пакт

Заключим 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

$\text{Var}(X)$ – дисперсия

$\text{Cov}(X, Y), \rho(X, Y)$ – ковариация и корреляция

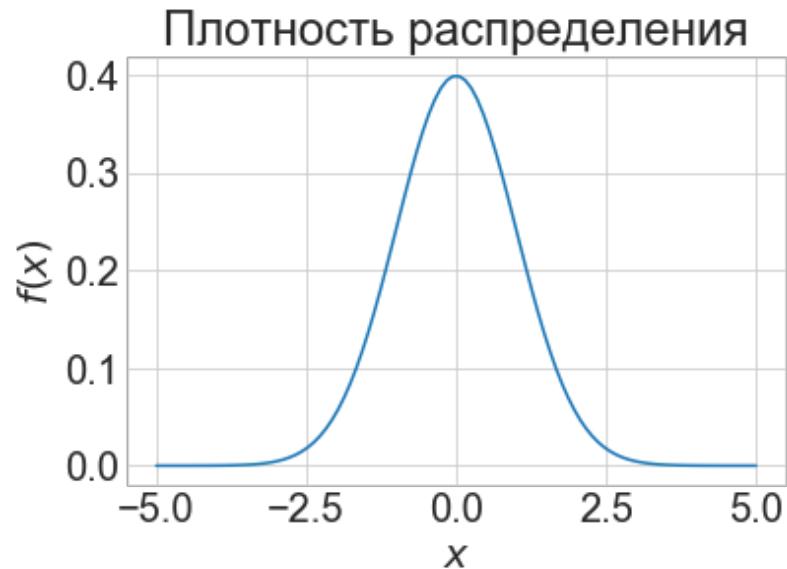
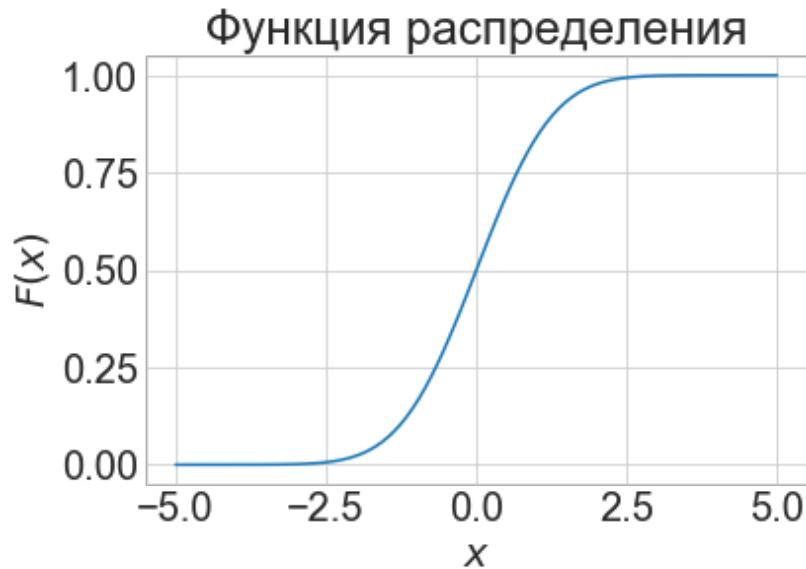
Как устроен мир



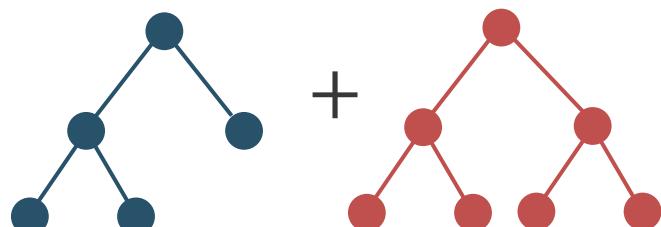
X

- 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

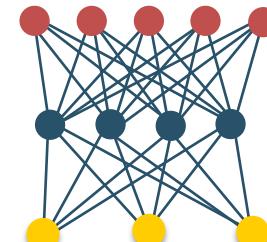


Модель – наше предположение о том, как процесс порождения данных устроен. За каждой моделью стоят какие-то предпосылки, описывающие наше незнание.



+

$$M = U V^T$$



Что мы будем делать?

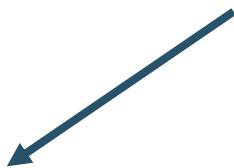
- Изучать выборки и их свойства
- Предполагать, какие процессы порождают данные, описывать своё незнание с помощью какой-то модели
- Разбираться, насколько наши предположения согласуются с выборками



Распределение случайной величины

Какими бывают случайные величины

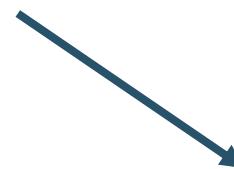
Случайные величины



Дискретные

Множество значений
конечно или счётно

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



Непрерывные

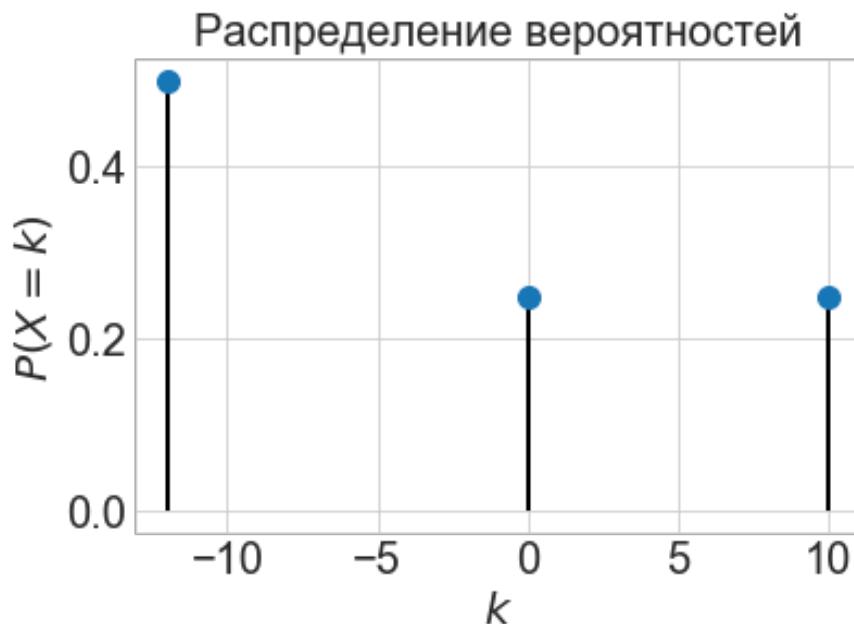
Принимают бесконечное,
континуальное число
значений

(рост, время ожидания
автобуса, вес)

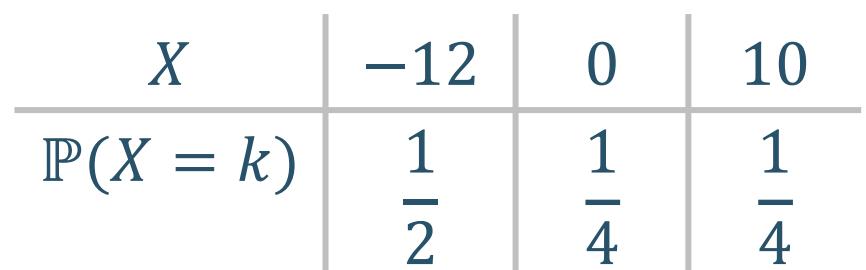
Дискретные случайные величины

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

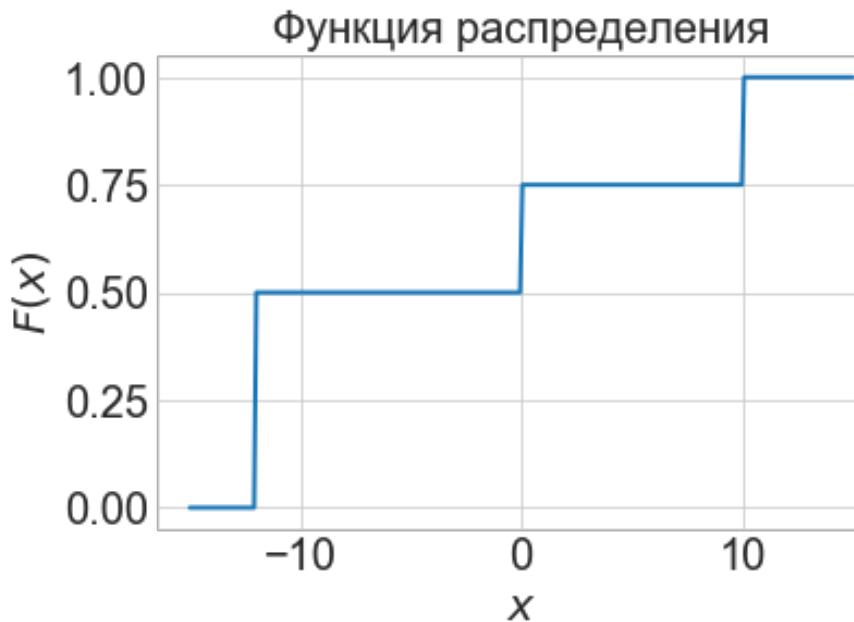


Дискретные случайные величины

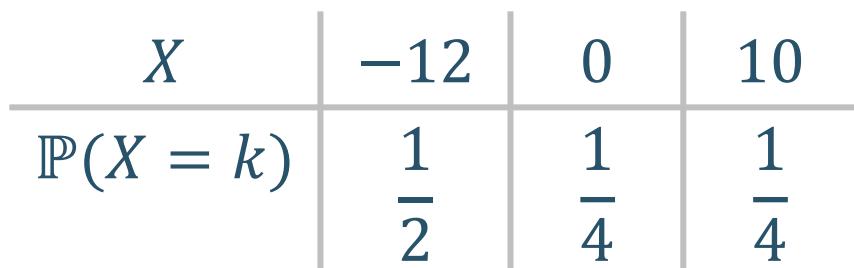
Функция распределения – функция, которая определяет вероятность события $X \leq x$, то есть

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

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

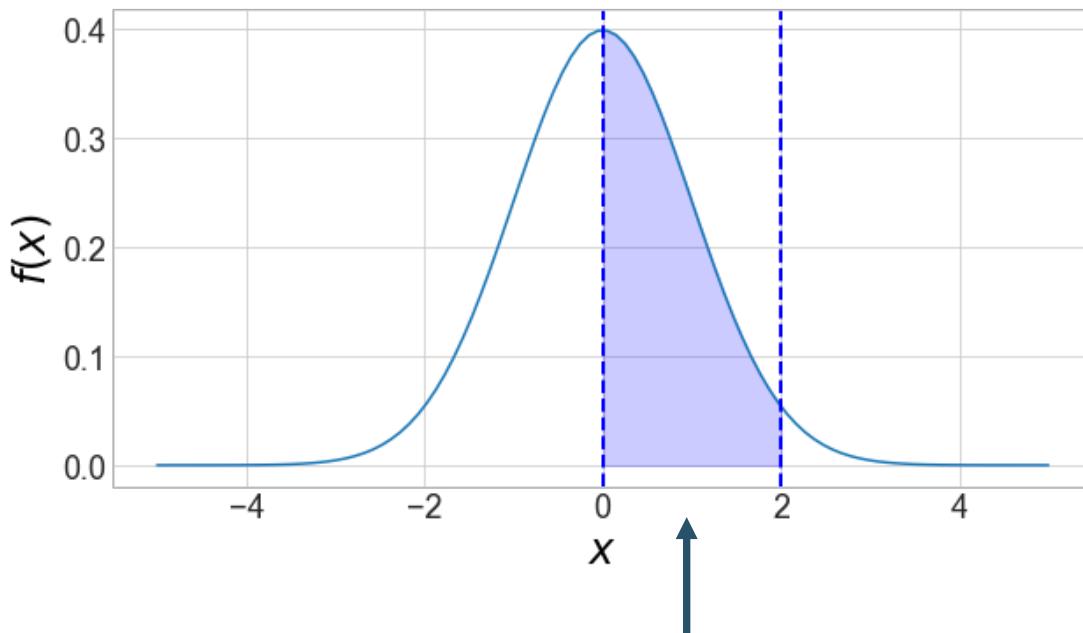


Пример: лотерея



Непрерывные случайные величины

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



Площадь равна вероятности попасть
на отрезок от нуля до двух

Пример:
нормальное
распределение

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

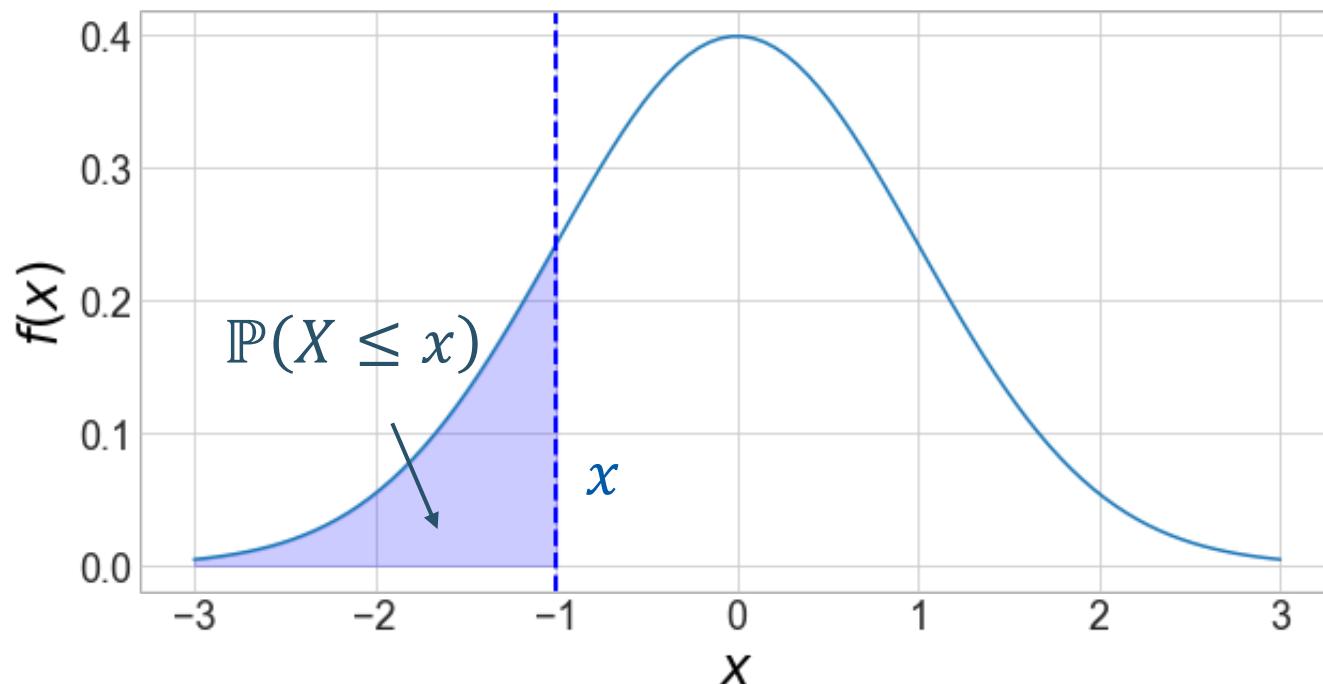
$$= \int_0^2 f(x) \, dx$$

Площадь под всей плотностью должна быть равна 1

Непрерывные случайные величины

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

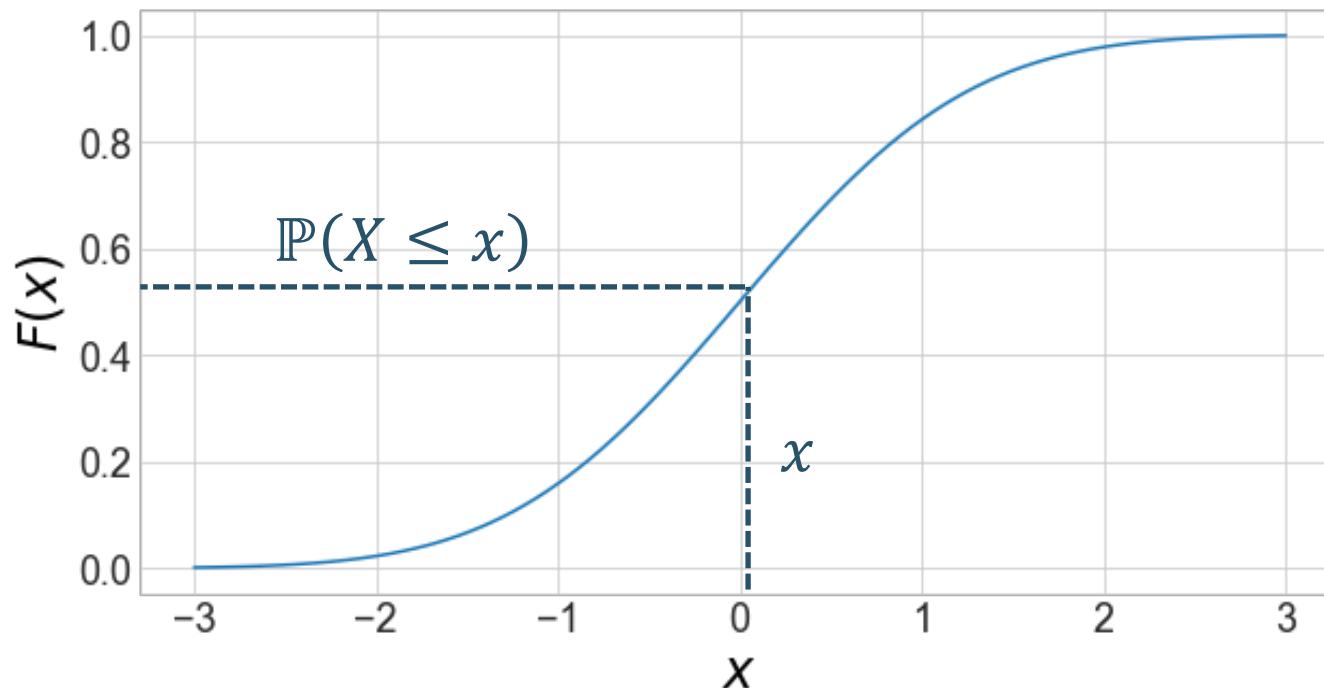
$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) \, dt, f(t) \text{ – плотность}$$



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

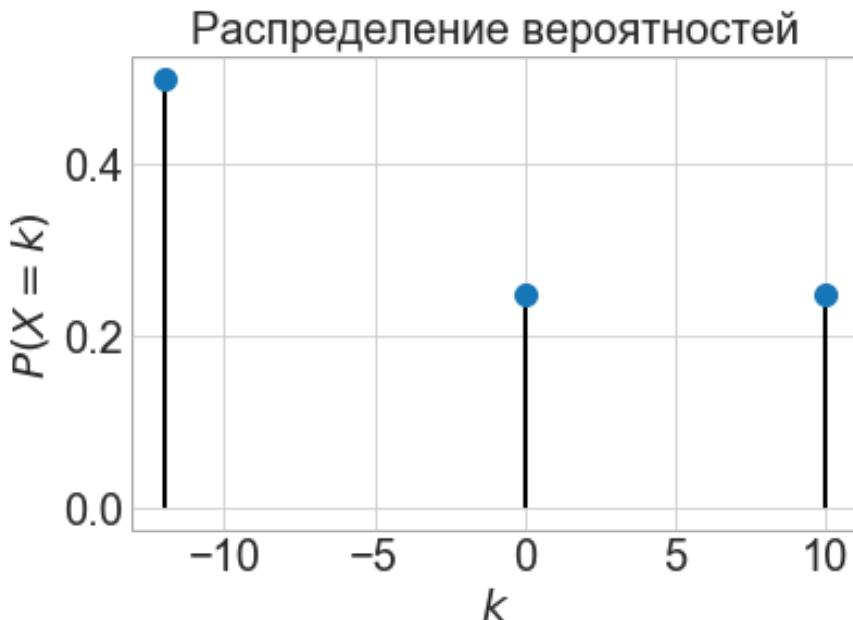
Характеристики случайных величин

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

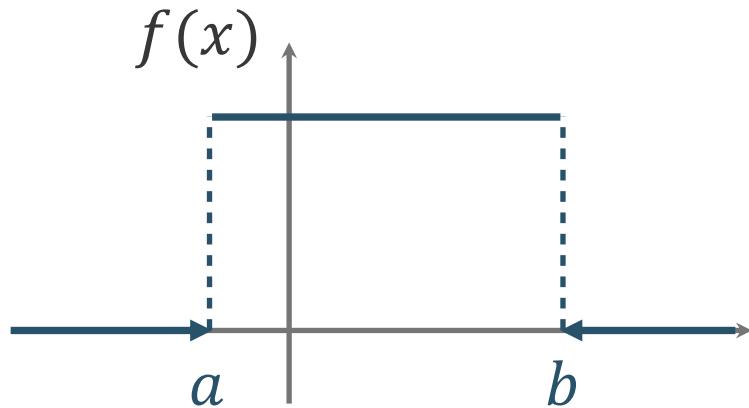
$$\mathbb{E}(X) = -12 \cdot 0.5 + 0 \cdot 0.25 + 10 \cdot 0.25 = -3.5 \text{ рубля}$$

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

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

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

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:

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

Дисперсия

Дисперсия – мера разброса случайной величины вокруг её среднего

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

Пример: лотерея

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

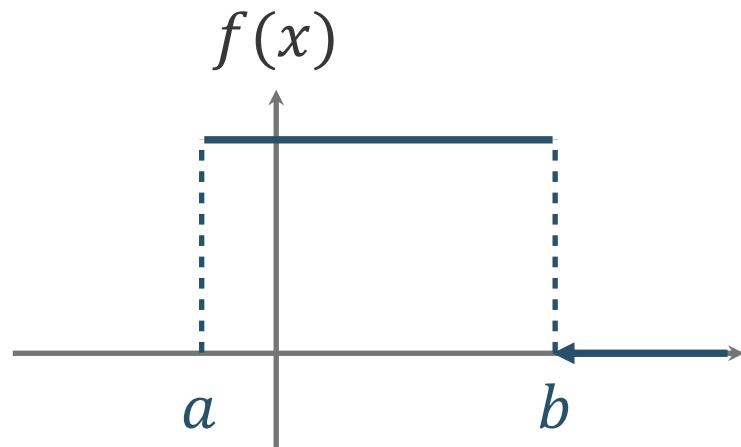
$$\mathbb{E}(X^2) = (-12)^2 \cdot 0.5 + 0^2 \cdot 0.25 + 10^2 \cdot 0.25 = 97 \text{ рублей}^2$$

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

Дисперсия

Дисперсия – мера разброса случайной величины вокруг её среднего

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



Пример: равномерное

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

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

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

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:

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

Свойства дисперсии

X, Y – случайные величины

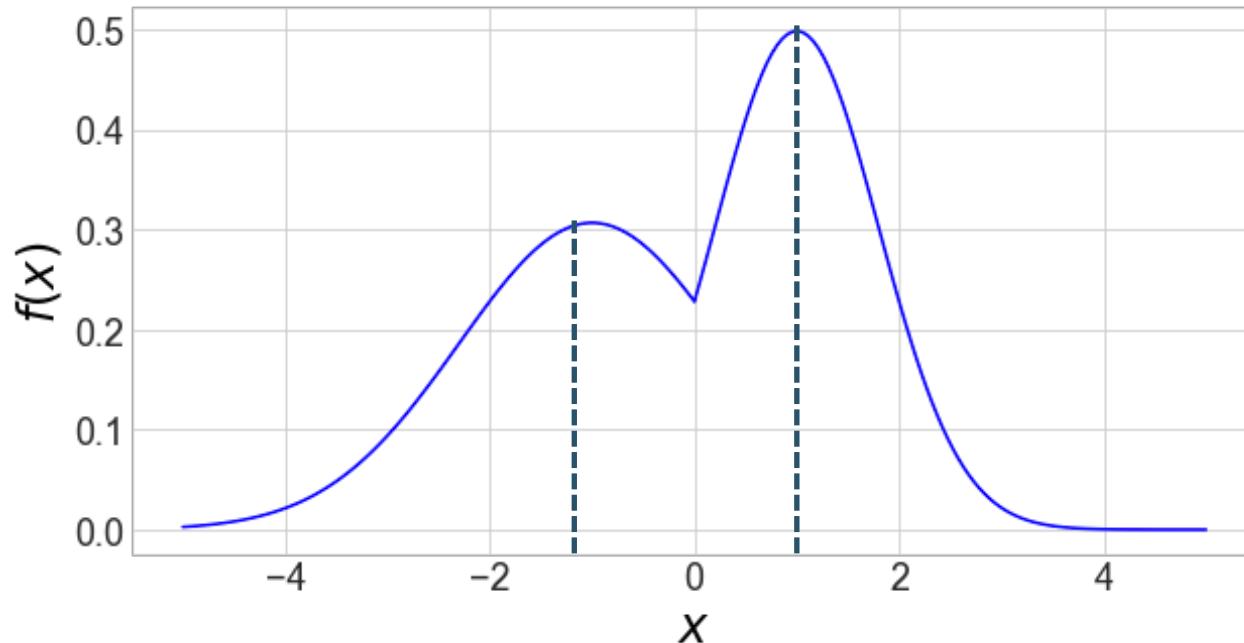
a – константа

1. $Var(a) = 0$
2. $Var(X + Y) = Var(X) + Var(Y)$, если независимы
3. $Var(a \cdot X) = a^2 \cdot Var(X)$
4. $Var(X - Y) = Var(X) + Var(Y)$, если независимы
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)

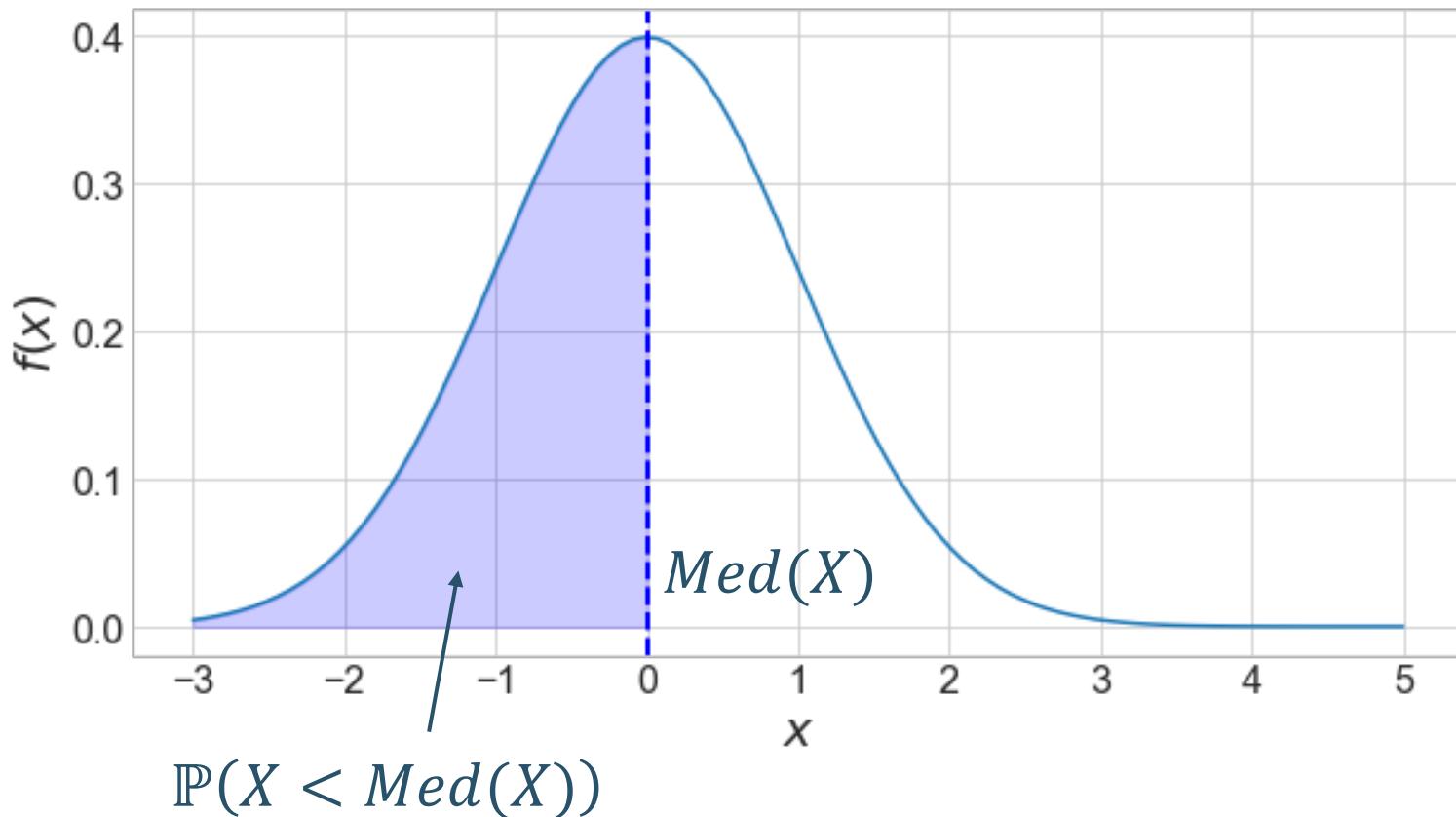
На практике встречаются мультимодальные распределения



Медиана

Медиана случайной величины – такое её значение, что

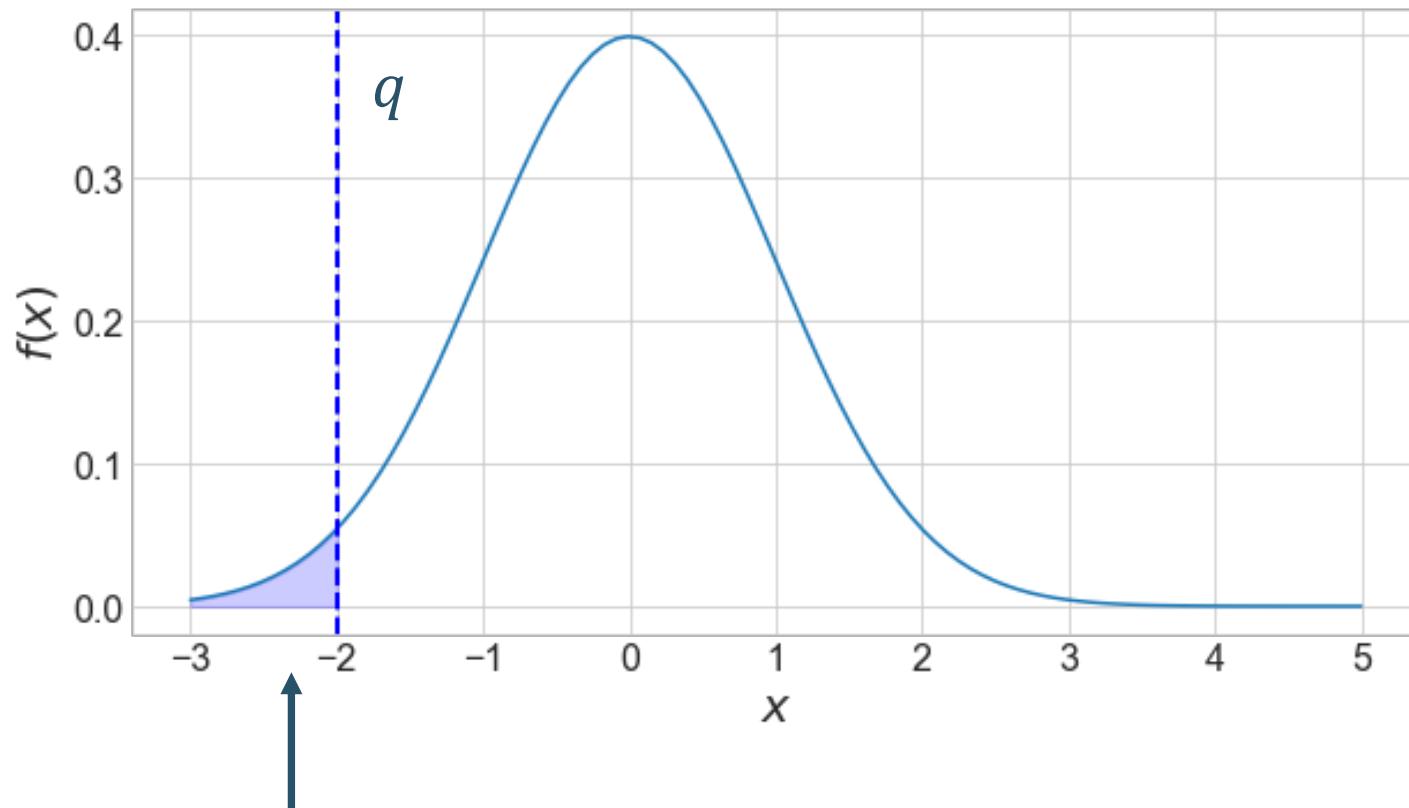
$$\mathbb{P}(X < \text{Med}(X)) = \mathbb{P}(X > \text{Med}(X)) = 0.5$$



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:

$$X \sim Bern(p)$$

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

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

Биномиальное распределение

- Число попаданий в баскетбольную корзину

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

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

Биномиальное распределение

$$Y_i \sim Bern(p)$$

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

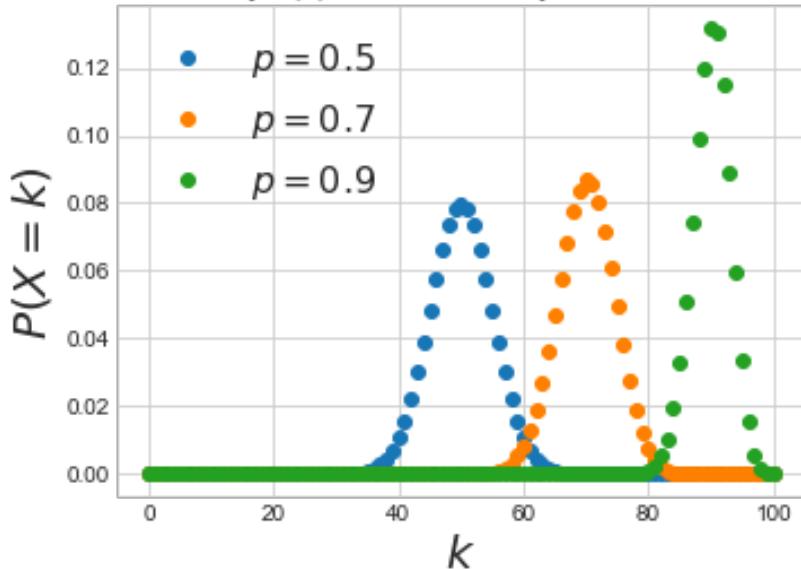
$$X = Y_1 + \dots + Y_n$$

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

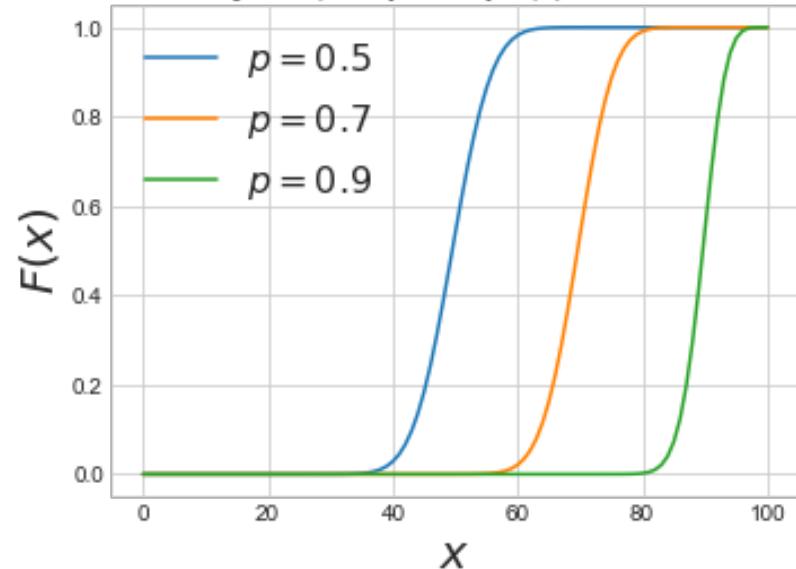
$$X \sim Bin(p, n)$$

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

Распределение вероятностей



Функция распределения



Геометрическое распределение

- Номер броска, когда произошло первое попадание в корзину

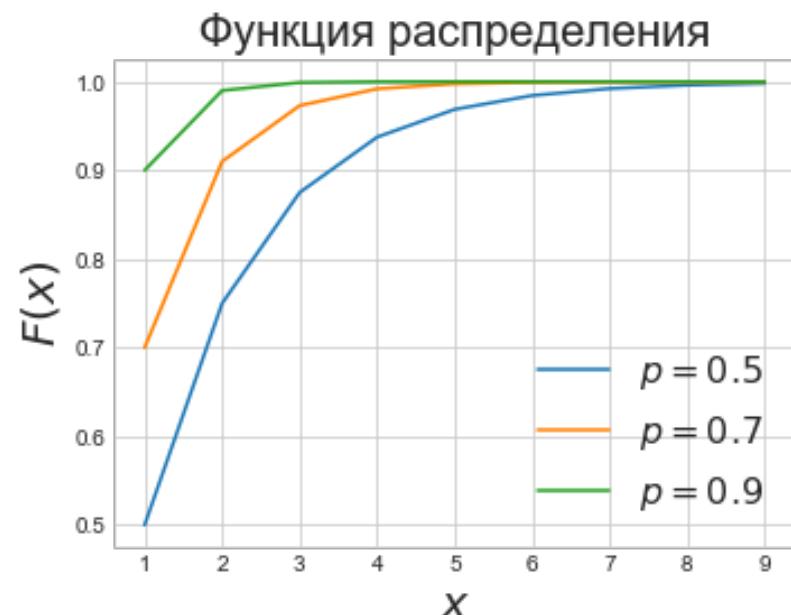
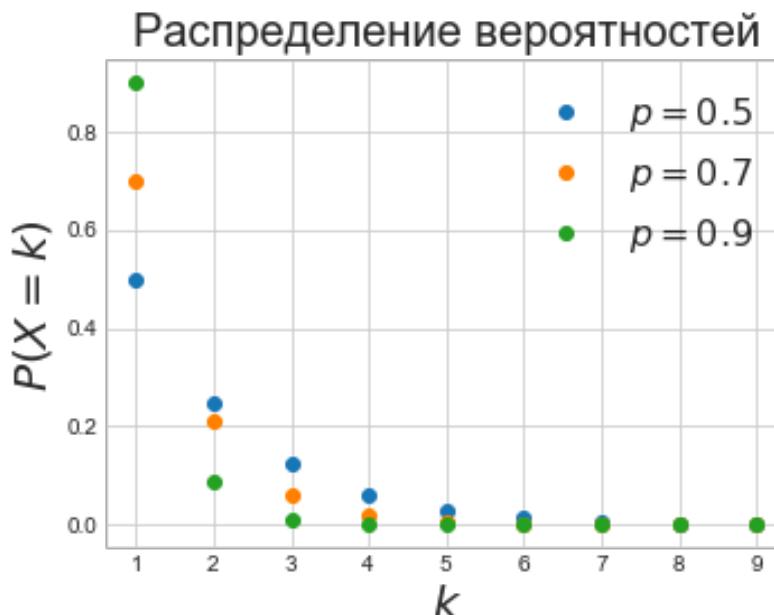
$$\mathbb{E}(X) = \frac{1}{p}$$

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

p – вероятность успеха

k принимает значения 1,2,3,...

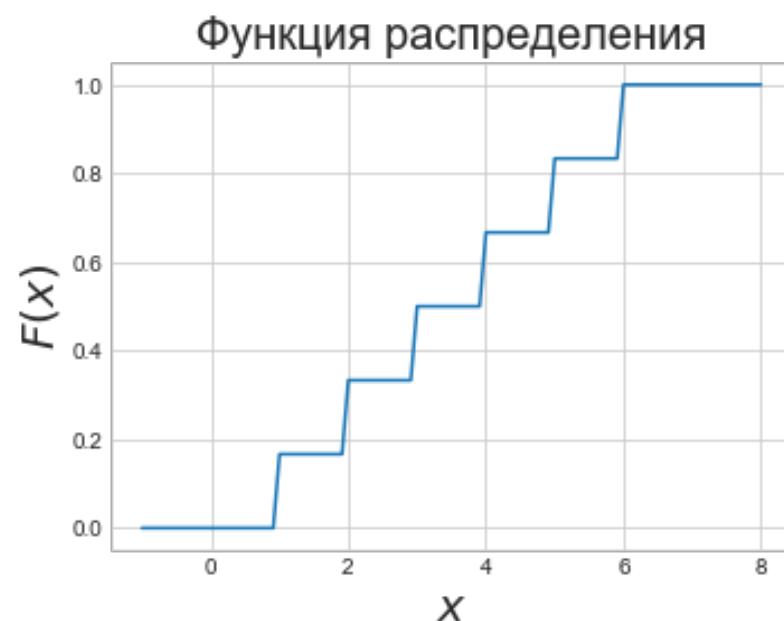
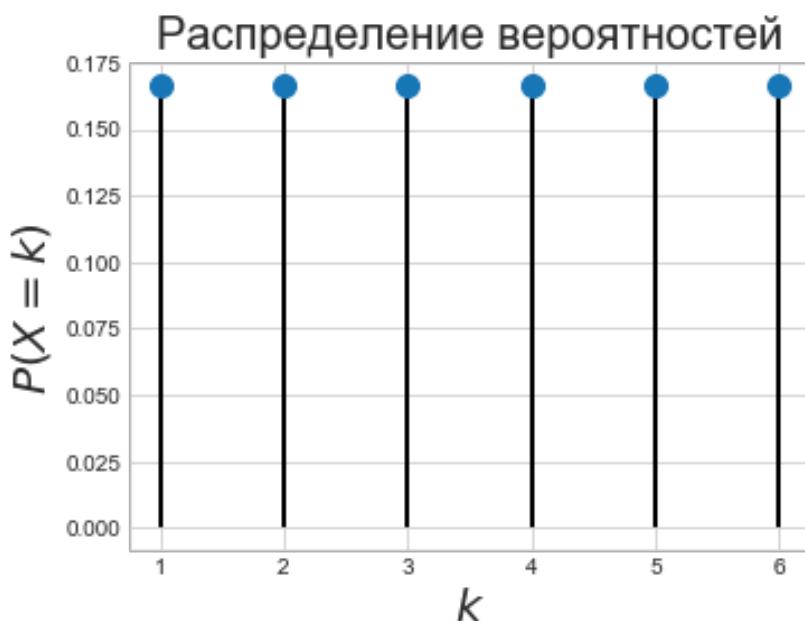
$$\mathbb{P}(X = k) = p \cdot (1 - p)^{k-1}$$



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

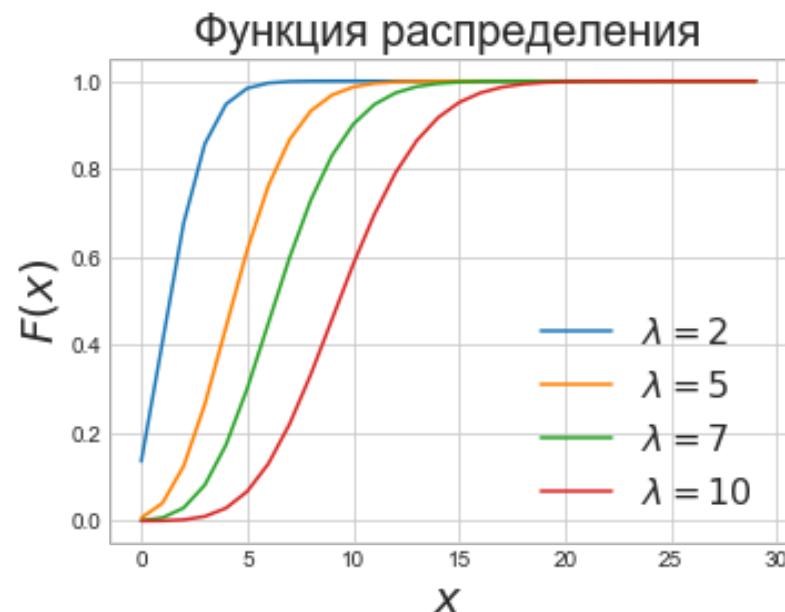
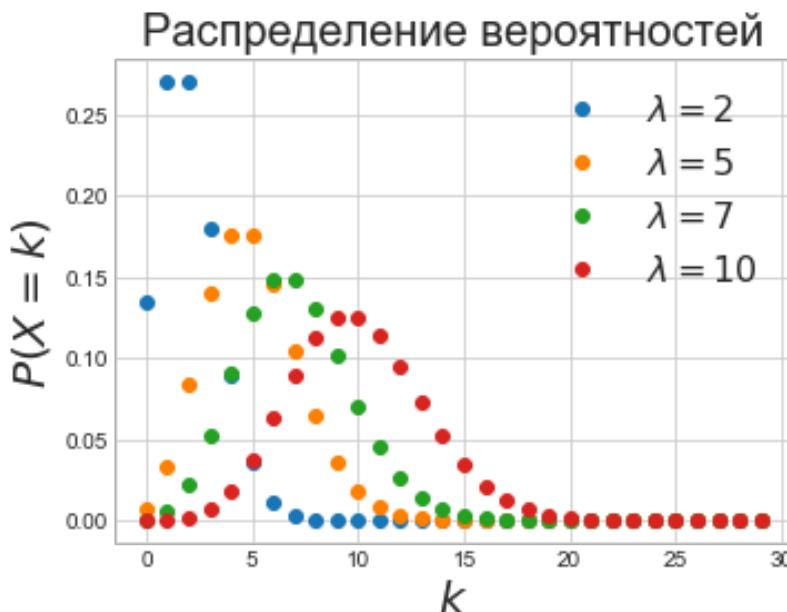
Распределение Пуассона

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

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

$$Var(X) = \lambda$$

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

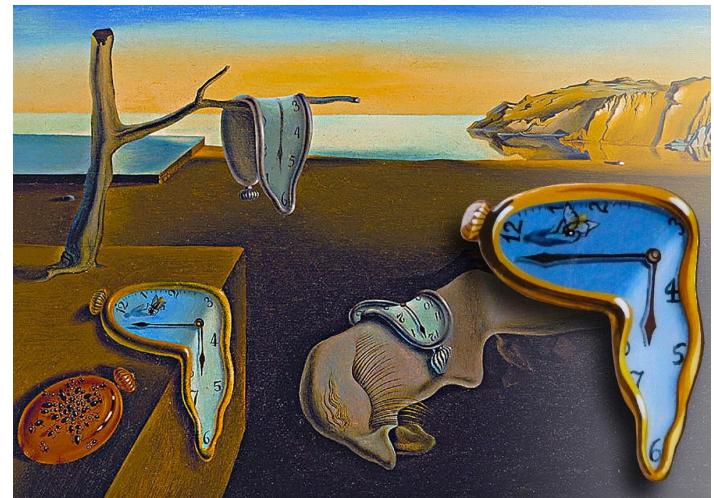


Время до ...

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

Экспоненциальная
случайная величина:
 $X \sim Exp(\lambda)$

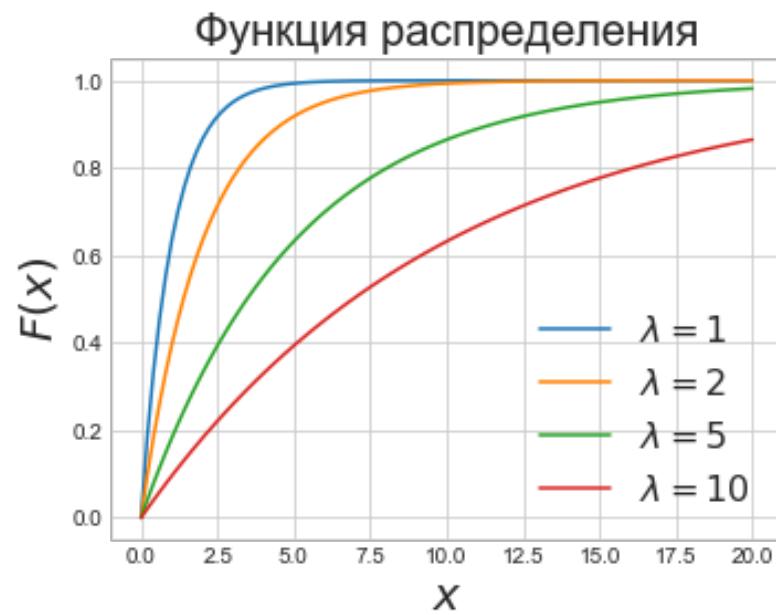
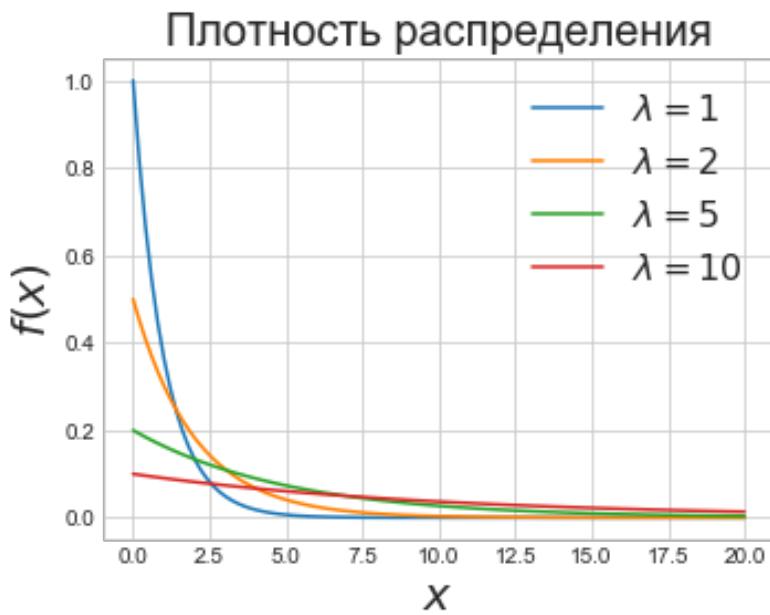
- Интервалы времени между событиями
- Модели времени жизни



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

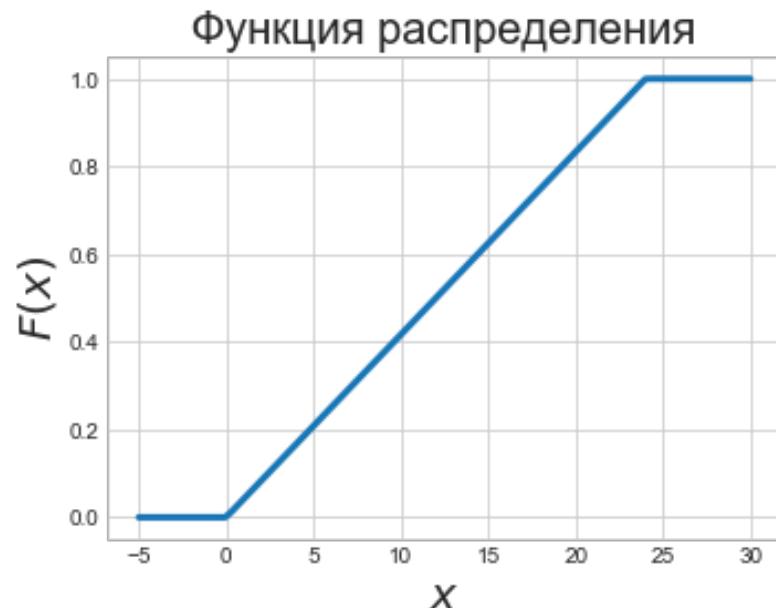
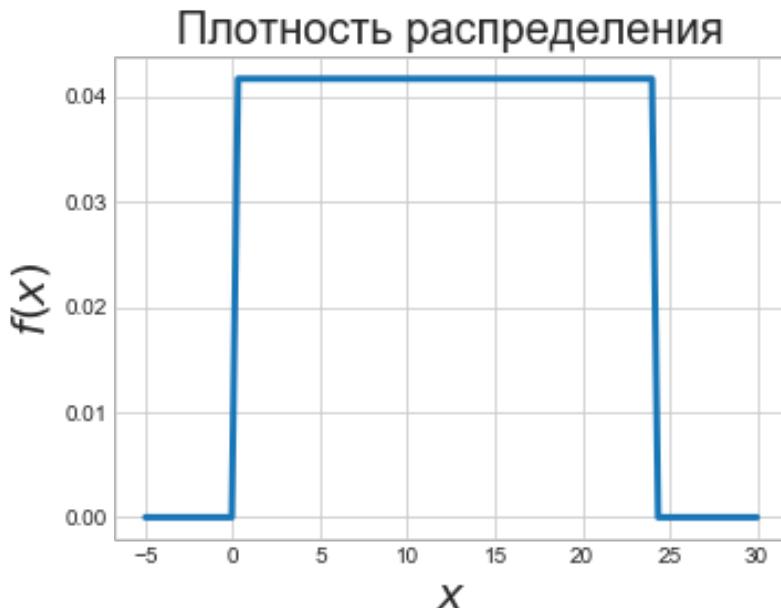
Равномерная случайная величина: $X \sim U[a; b]$

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

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

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

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



Нормальное распределение

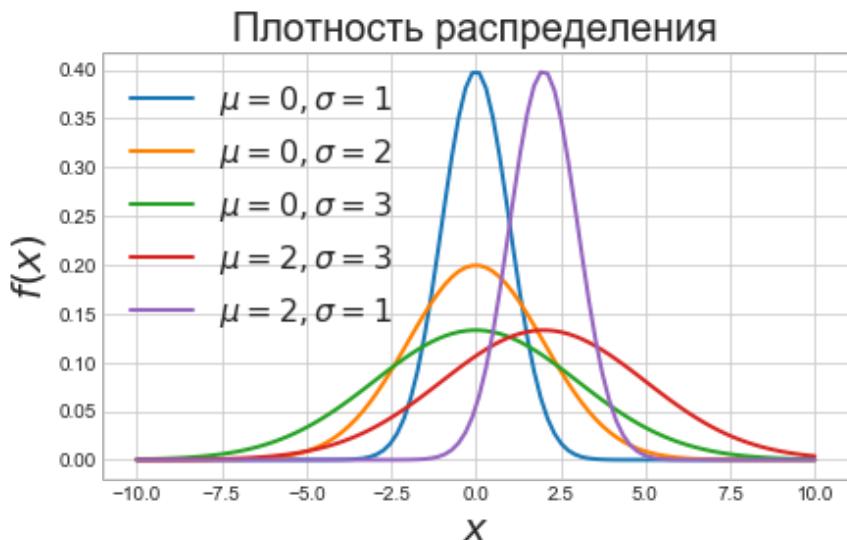
- Погрешность весов

Нормальная случайная величина:

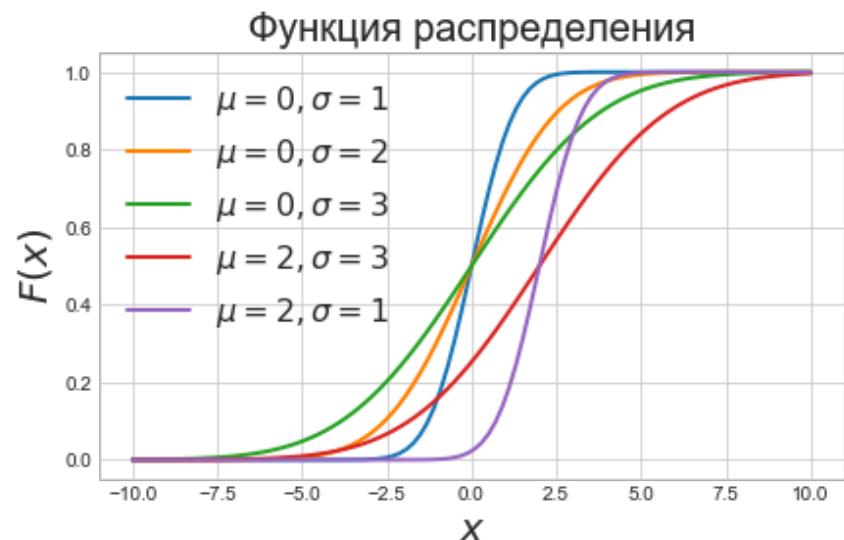
$$X \sim N(\mu, \sigma^2)$$

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

Функцию распределения нельзя найти в аналитическом виде, интеграл не берётся



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



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

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

Нормальное распределение и его свойства

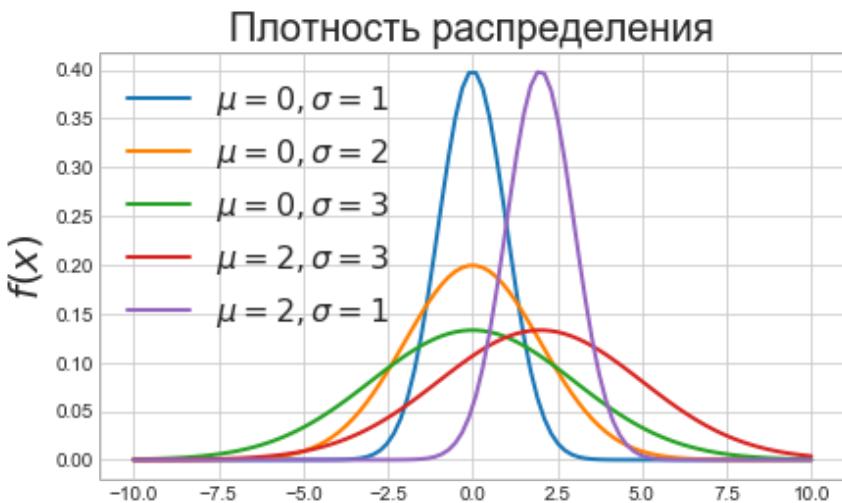
Нормальное распределение

- 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
- Давайте познакомиться с нормальным распределением поближе

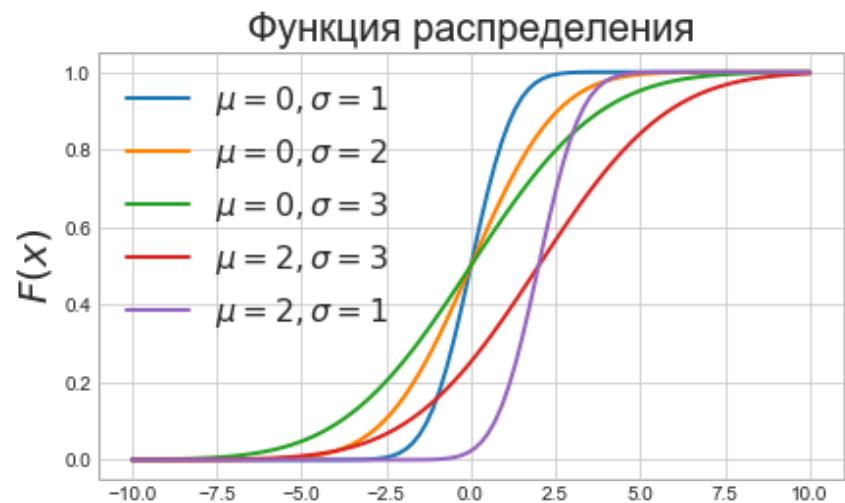
Нормальное распределение

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

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

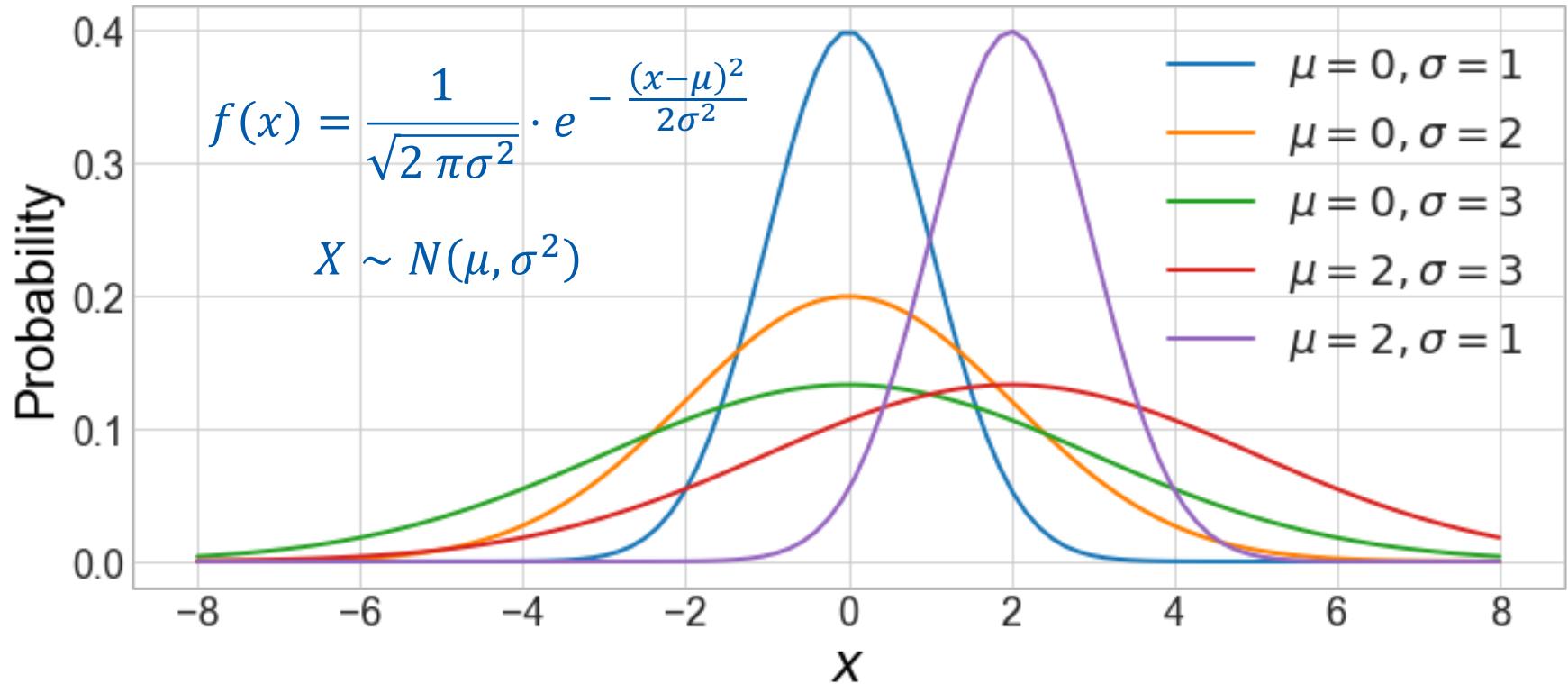


$$f(x) = \frac{1}{\sqrt{2 \pi \sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



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

Свойства нормального распределения



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

Как найти вероятность

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

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

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



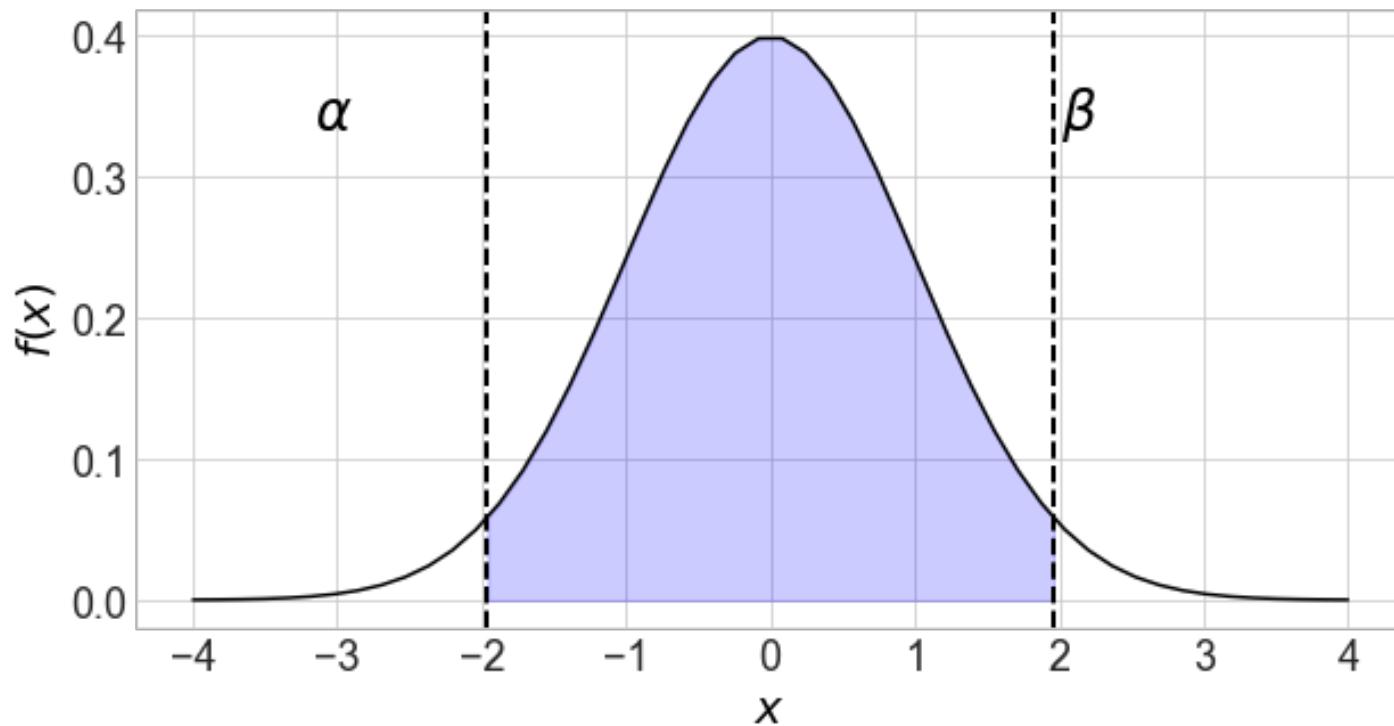
Обозначение
для функции
распределения $N(0,1)$

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:

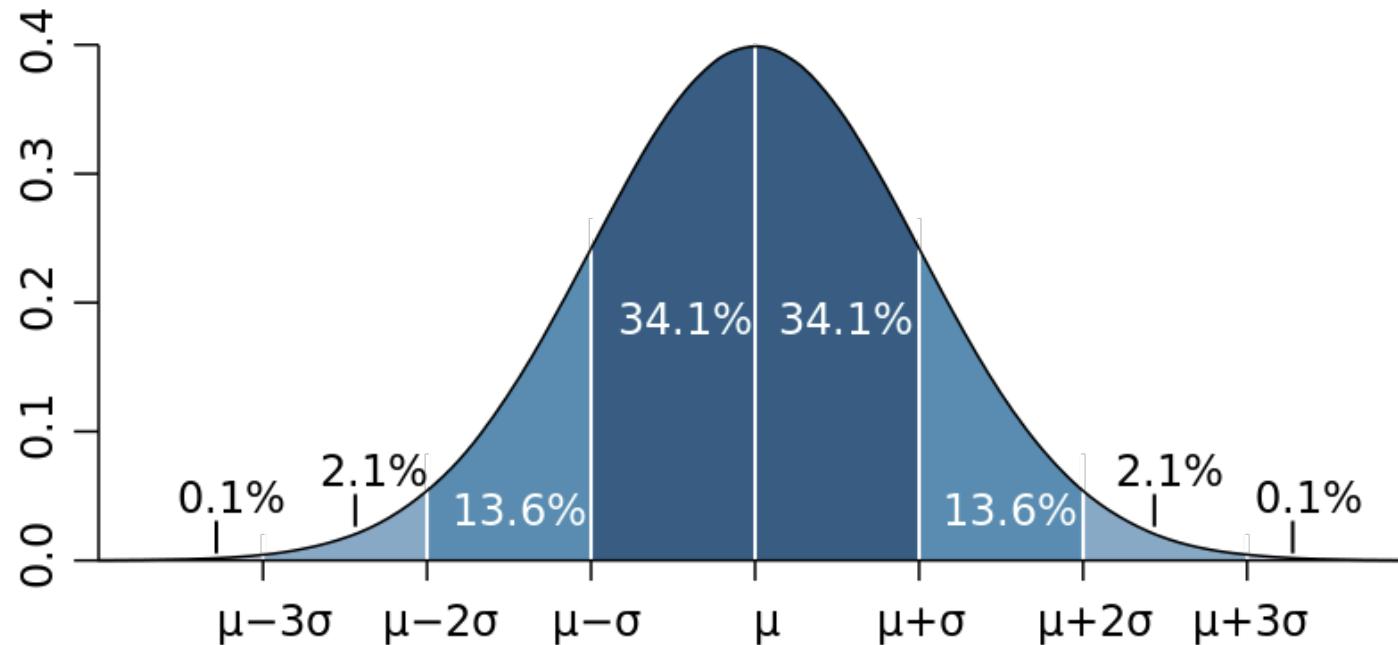
$$\mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$$

Правило двух сигм:

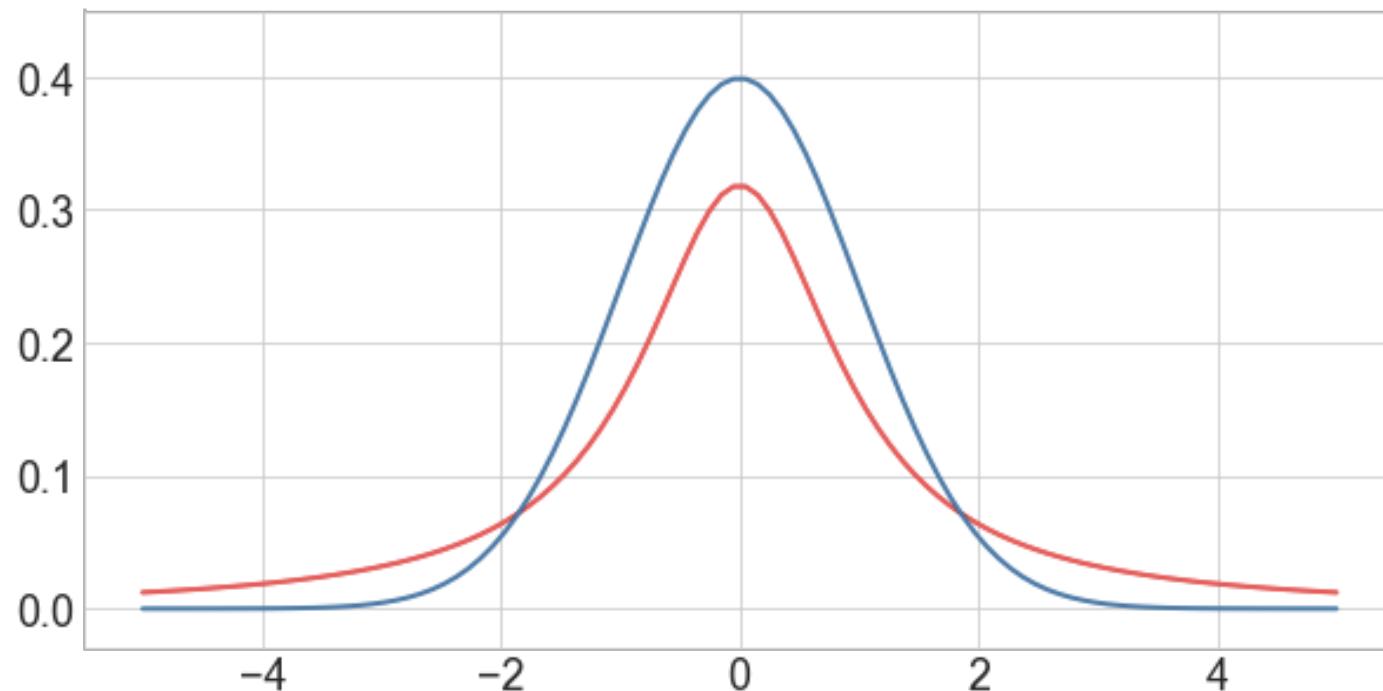
$$\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

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

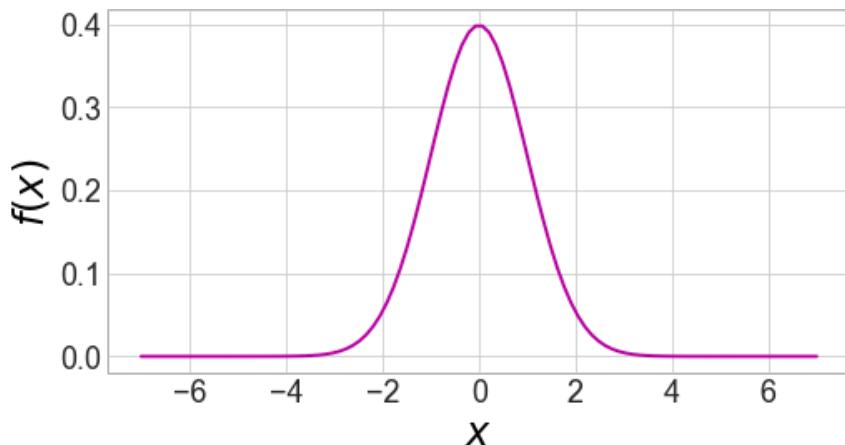
- Число 3 вычитается из куртосиса, чтобы эксцесс нормального распределения был равен нулю
- Если хвосты распределения легче, а пик острее, чем у нормального распределения, тогда $\beta_X > 0$
- Если хвосты распределения тяжелее, а пик более приплюснутый, тогда $\beta_X < 0$

Эксцесс и куртосис

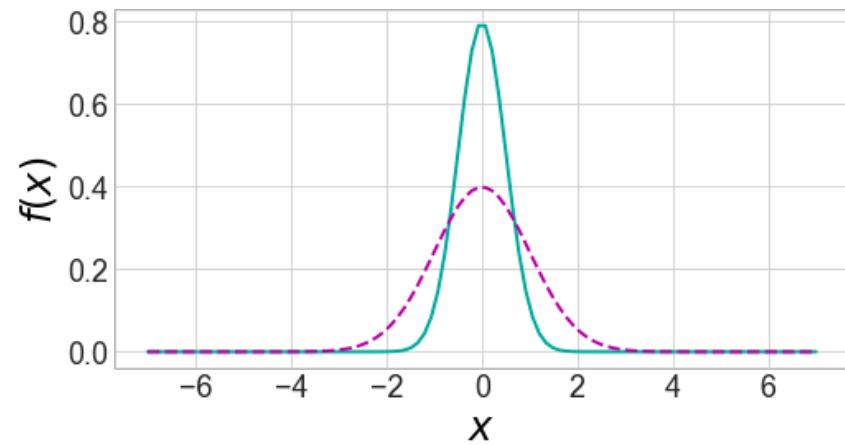
Эксцессом случайной величины X называют величину

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

Куртосис



Нормальное распределение
с нулевым эксцессом



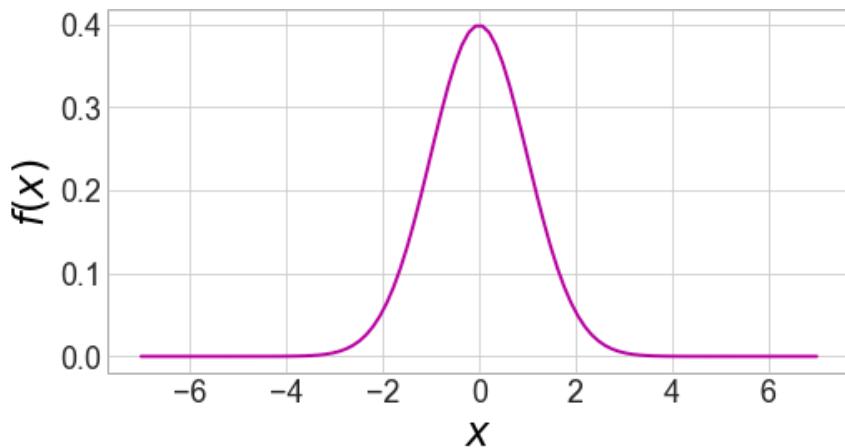
Положительный эксцесс

Эксцесс и куртосис

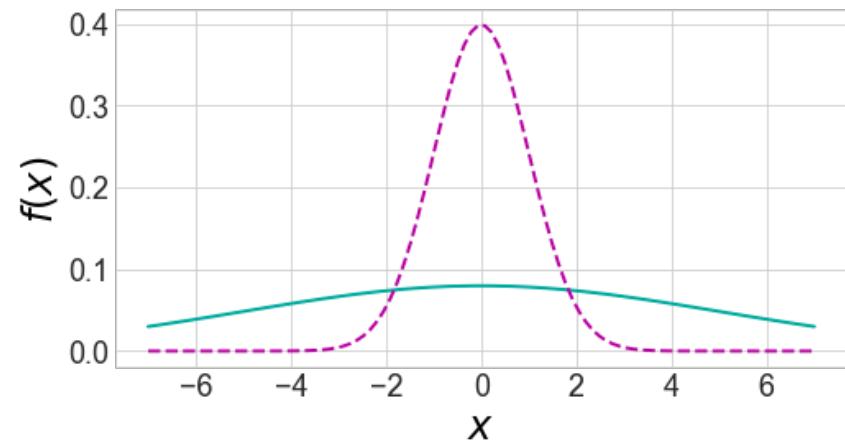
Эксцессом случайной величины X называют величину

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

Куртосис



Нормальное распределение
с нулевым эксцессом



Отрицательный эксцесс

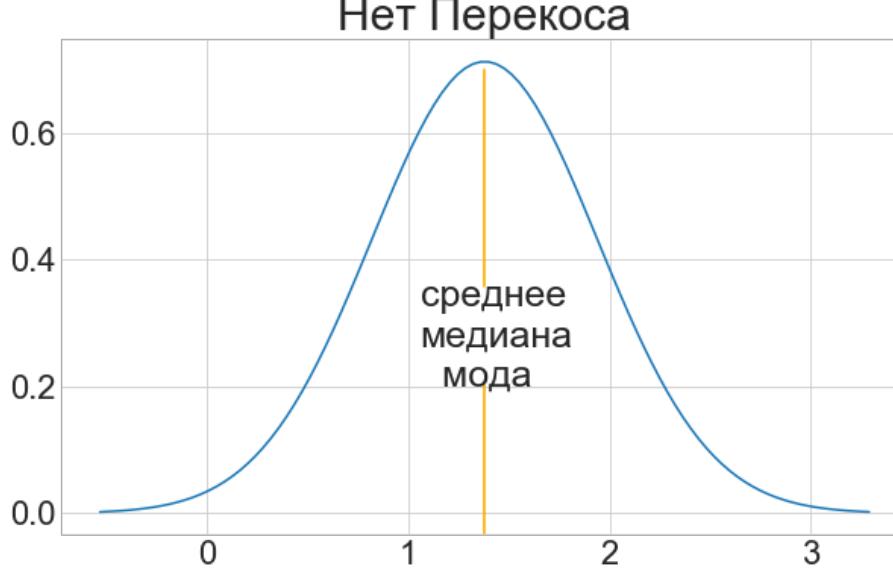
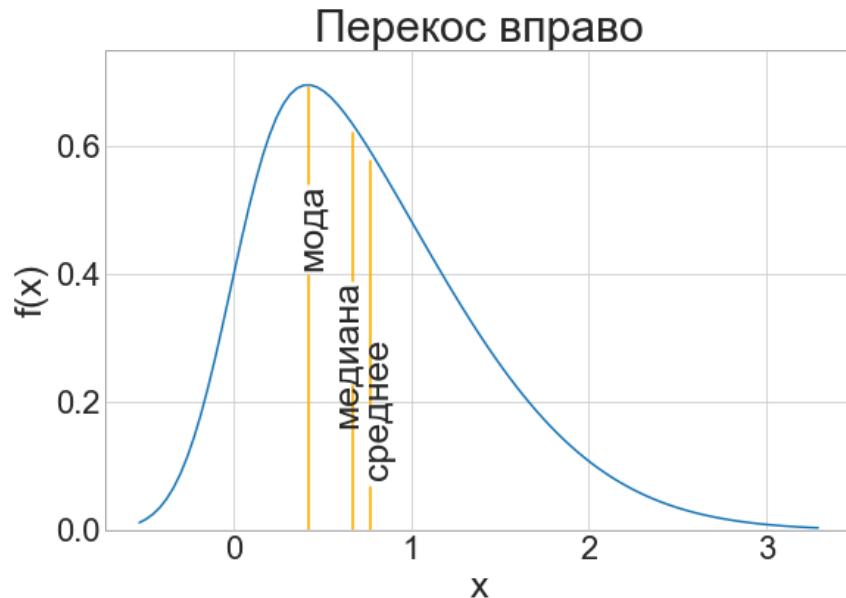
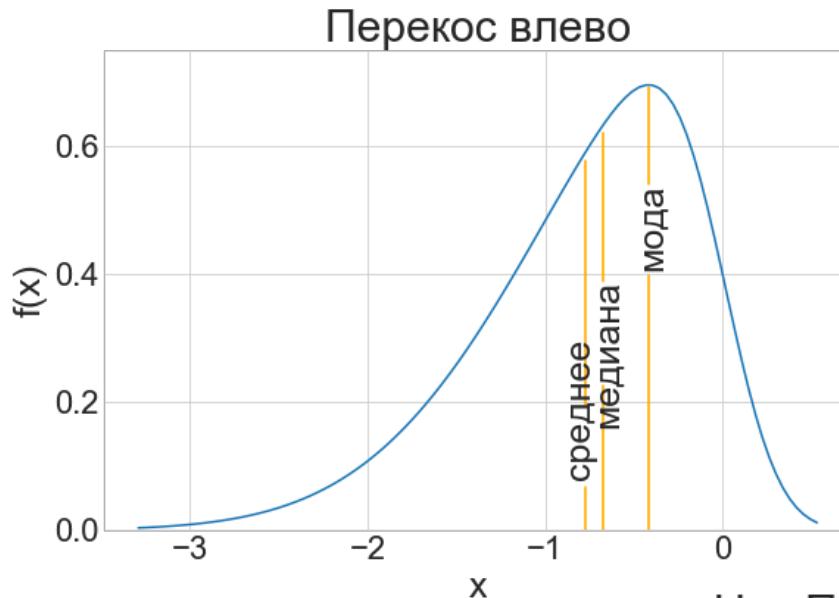
Коэффициент асимметрии (skewness)

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

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

Математическое
ожидание

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

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

Ковариационная
матрица

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

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

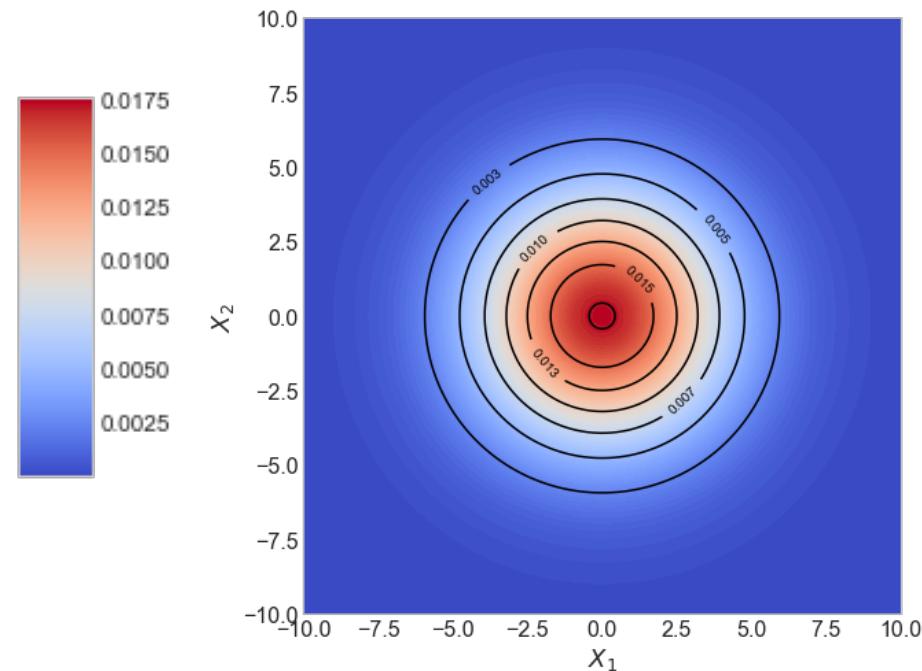
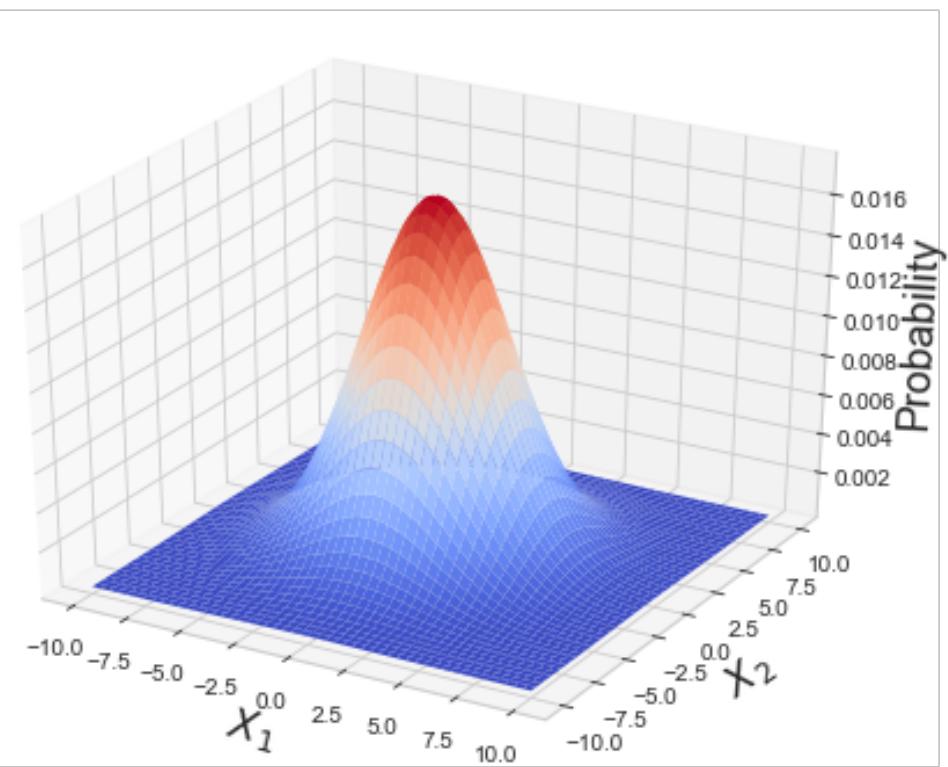
$$Cov(X_1, X_2) = \rho$$

Кратко пишут

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

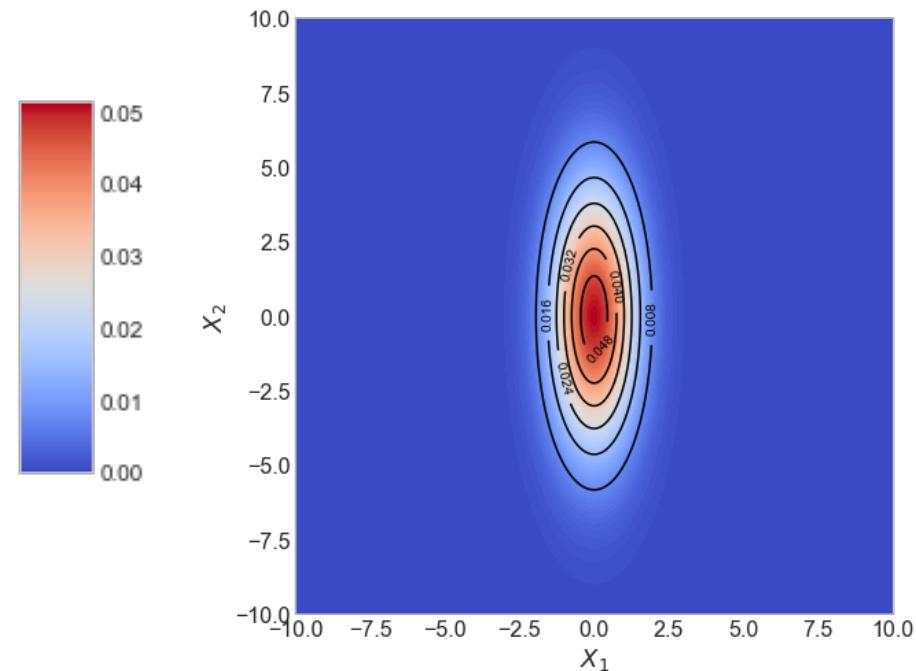
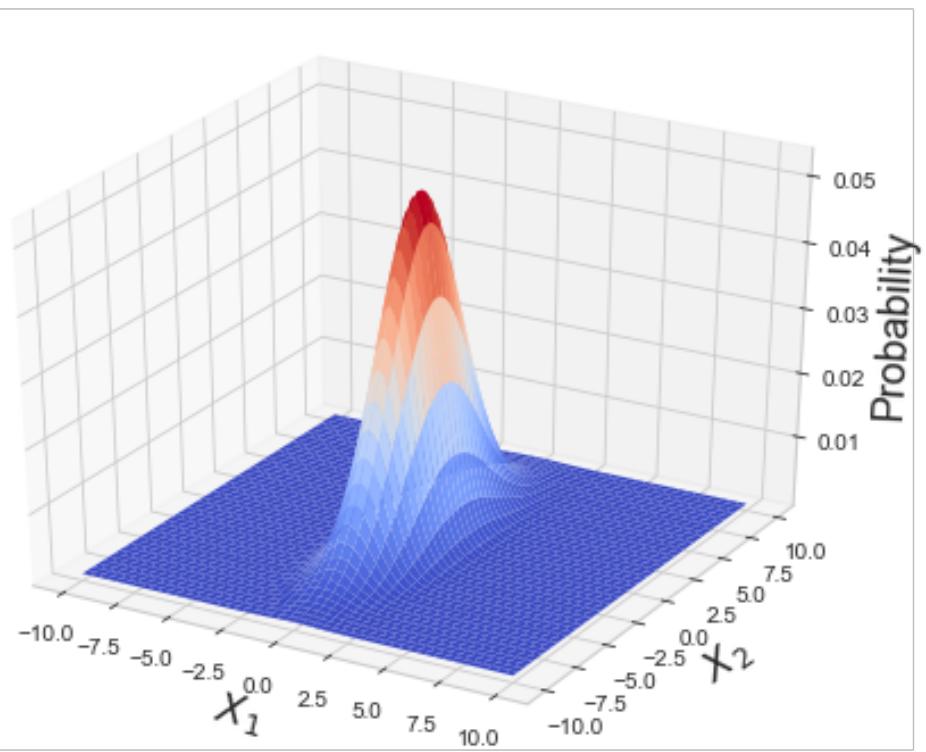
Многомерное нормальное

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



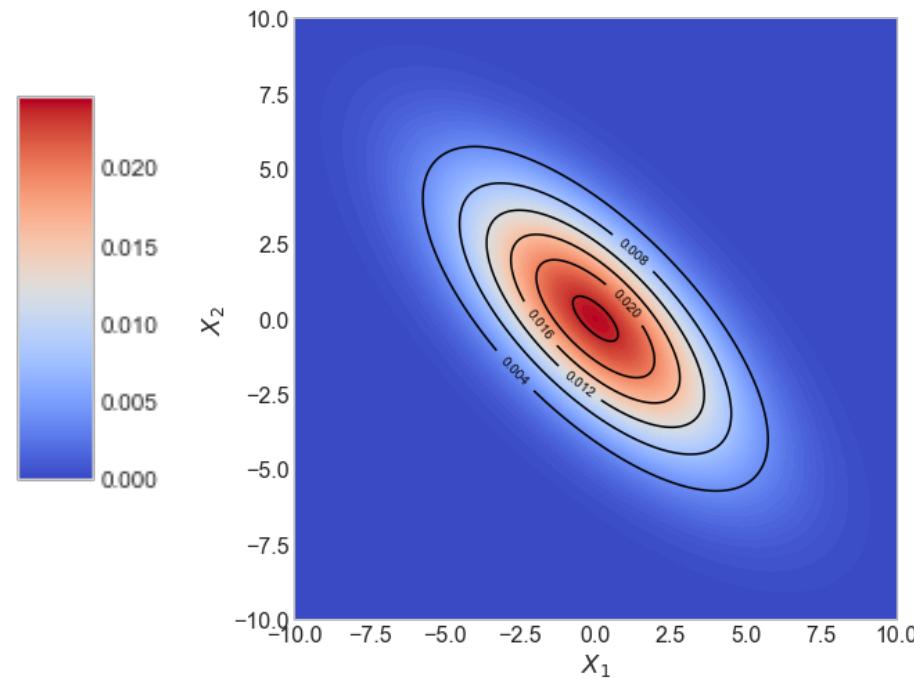
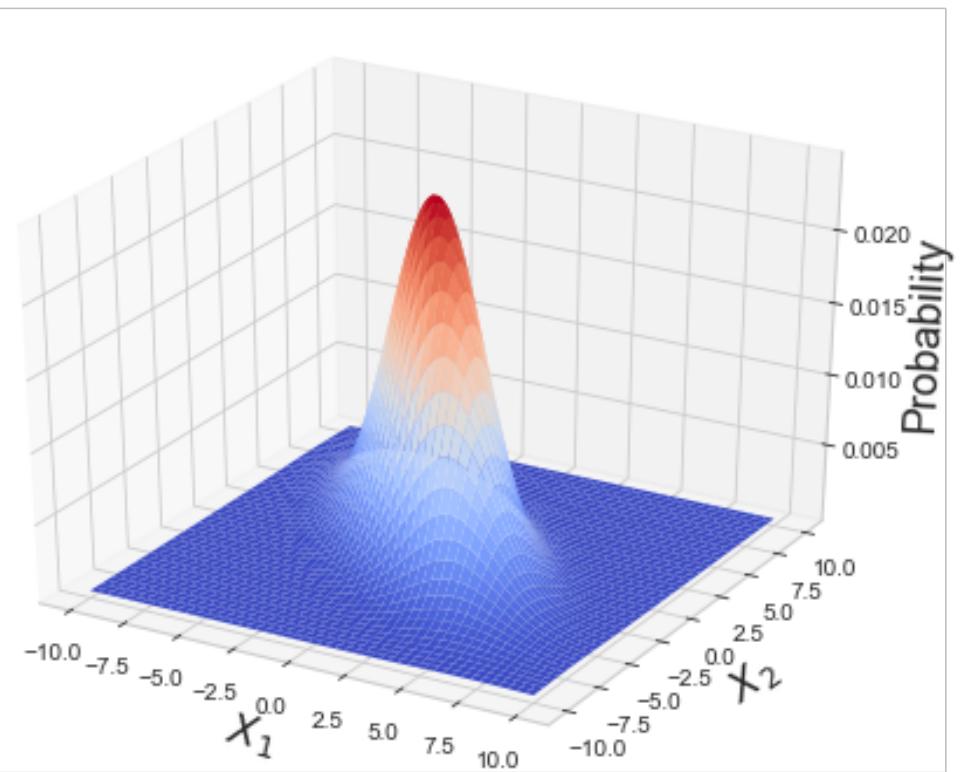
Многомерное нормальное

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



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

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

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

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

Резюме

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

Выборочные статистики

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.

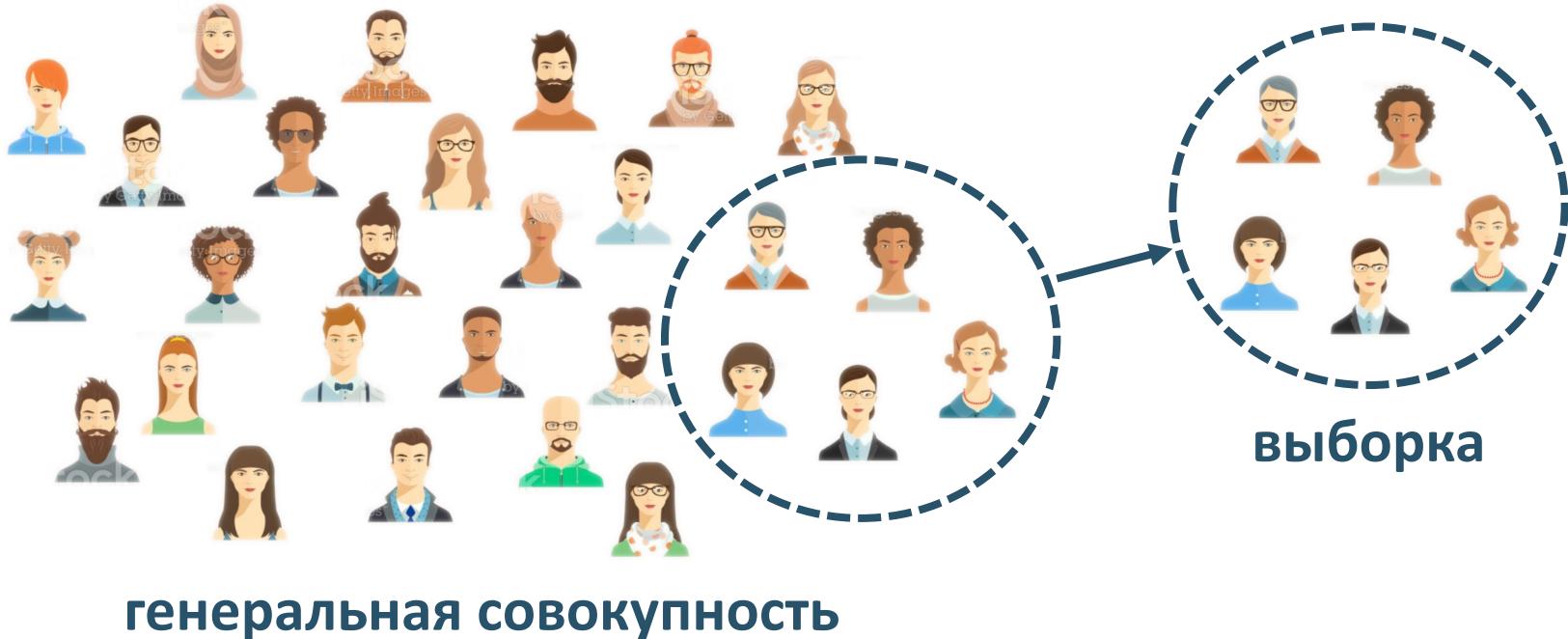
Генеральная совокупность и выборка

Генеральная совокупность – это все объекты, которые нас интересуют при исследовании

Выборка – это та часть генеральной совокупности, по которой мы собрали данные для исследования

Генеральная совокупность и выборка

- В городе живёт 1 млн. человек
- Провели опрос об уровне дохода (2.5 тыс. человек)
- Опубликовали средний доход по городу
- Опрашивать абсолютно всех людей в городе долго



Репрезентативность

- Выборки позволяют сделать выводы о всей генеральной совокупности
- Чтобы выводы были корректными, выборка должны быть **репрезентативной**
- **Репрезентативная выборка** – отражает свойства генеральной совокупности

Пример: Добрыня, Илья и Алёна исследуют рост людей.
Чья выборка репрезентативна?

- Добрыня опросил свою баскетбольную команду
- Илья опросил людей на остановке
- Алёна опросила всех своих подруг

Репрезентативность

- Выборки позволяют сделать выводы о всей генеральной совокупности
- Чтобы выводы были корректными, выборка должны быть **репрезентативной**
- **Репрезентативная выборка** – отражает свойства генеральной совокупности

Пример: Добрыня, Илья и Алёна исследуют рост людей.
Чья выборка репрезентативна?

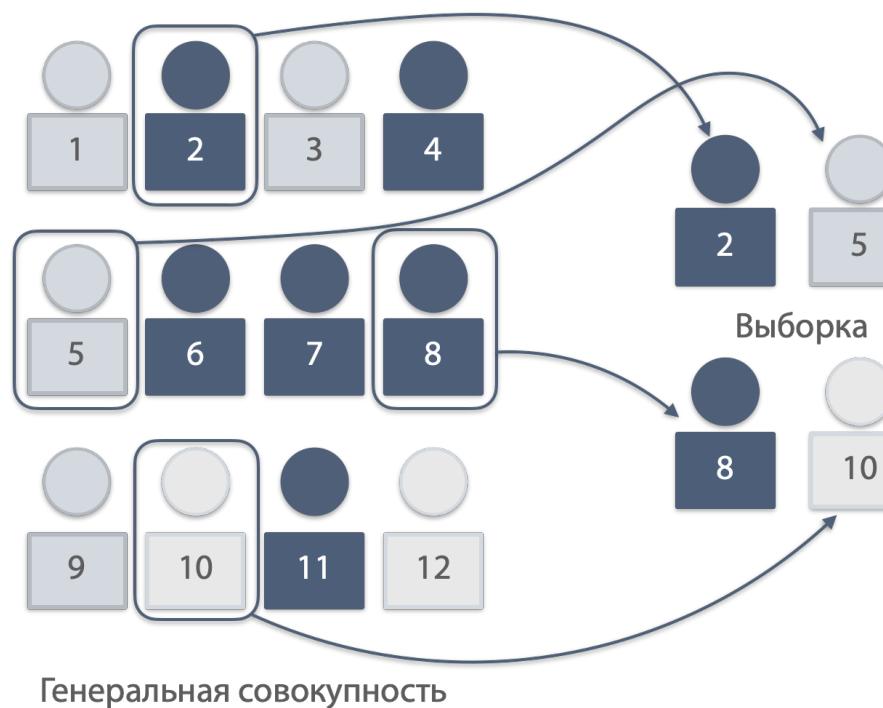
нет • Добрыня опросил свою баскетбольную команду

да • Илья опросил людей на остановке

нет • Алёна опросила всех своих подруг

Выборку бы, да репрезентативную бы

- Репрезентативность выборки определяет, насколько корректно делать выводы о всей генеральной совокупности, опираясь только на неё
- Один из способов достижения репрезентативности: случайный отбор наблюдений



Предпосылки

Выборка: X_1, X_2, \dots, X_n

Размер выборки

Одно наблюдение: X_i

- Каждое наблюдение можно рассматривать как случайную величину, которая имеет такое же распределение как и генеральная совокупность

Мы в дальнейшем будем всегда предполагать:

1. Наблюдения X_1, X_2, \dots, X_n независимы друг от друга
2. Наблюдения имеют одинаковое распределение (как у генеральной совокупности)

Краткая запись: $X_1, X_2, \dots, X_n \sim iid$

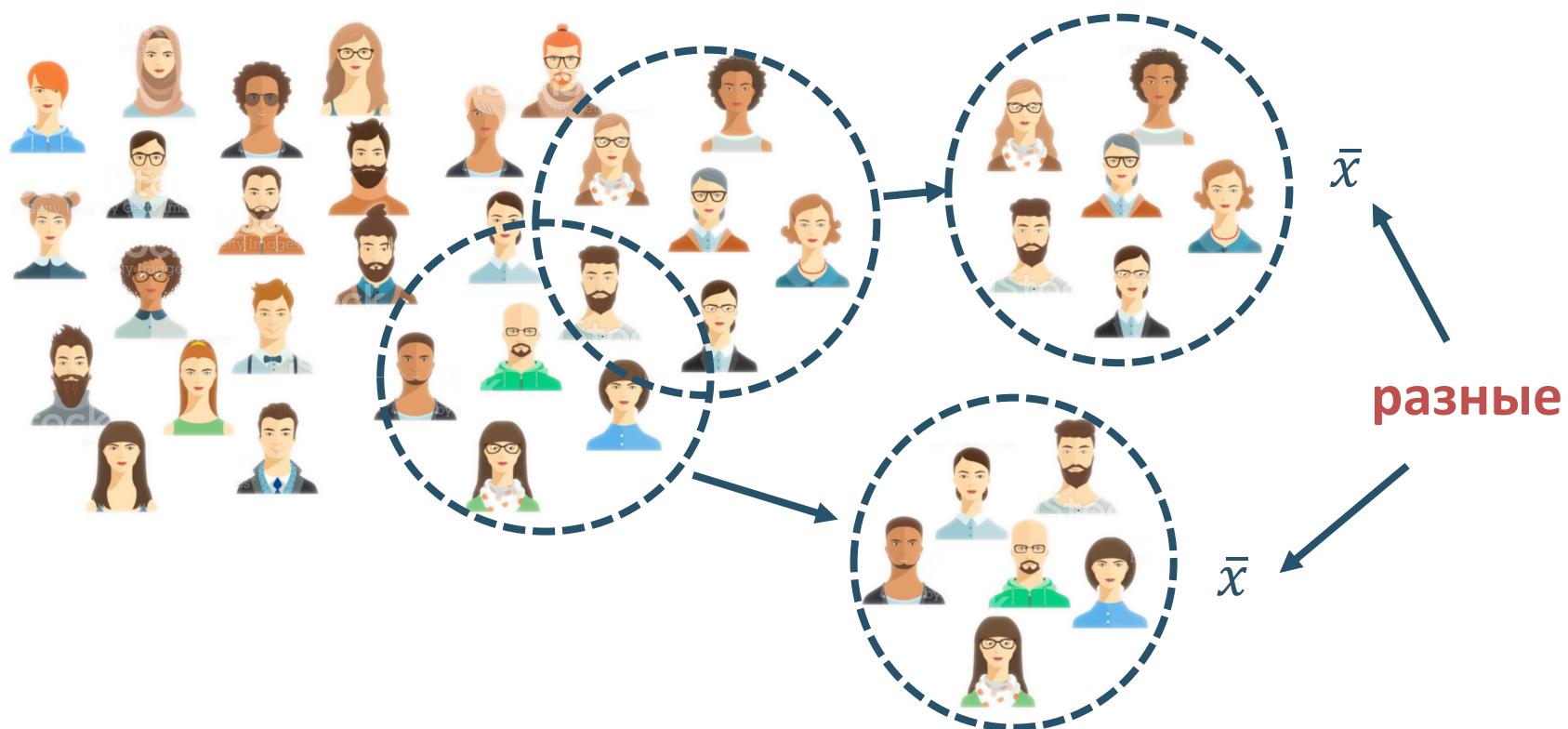
► *iid* расшифровывается как *identically independently distributed* (независимы и одинаково распределены)

Статистика

Выборка: $X_1, X_2, \dots, X_n \sim iid$

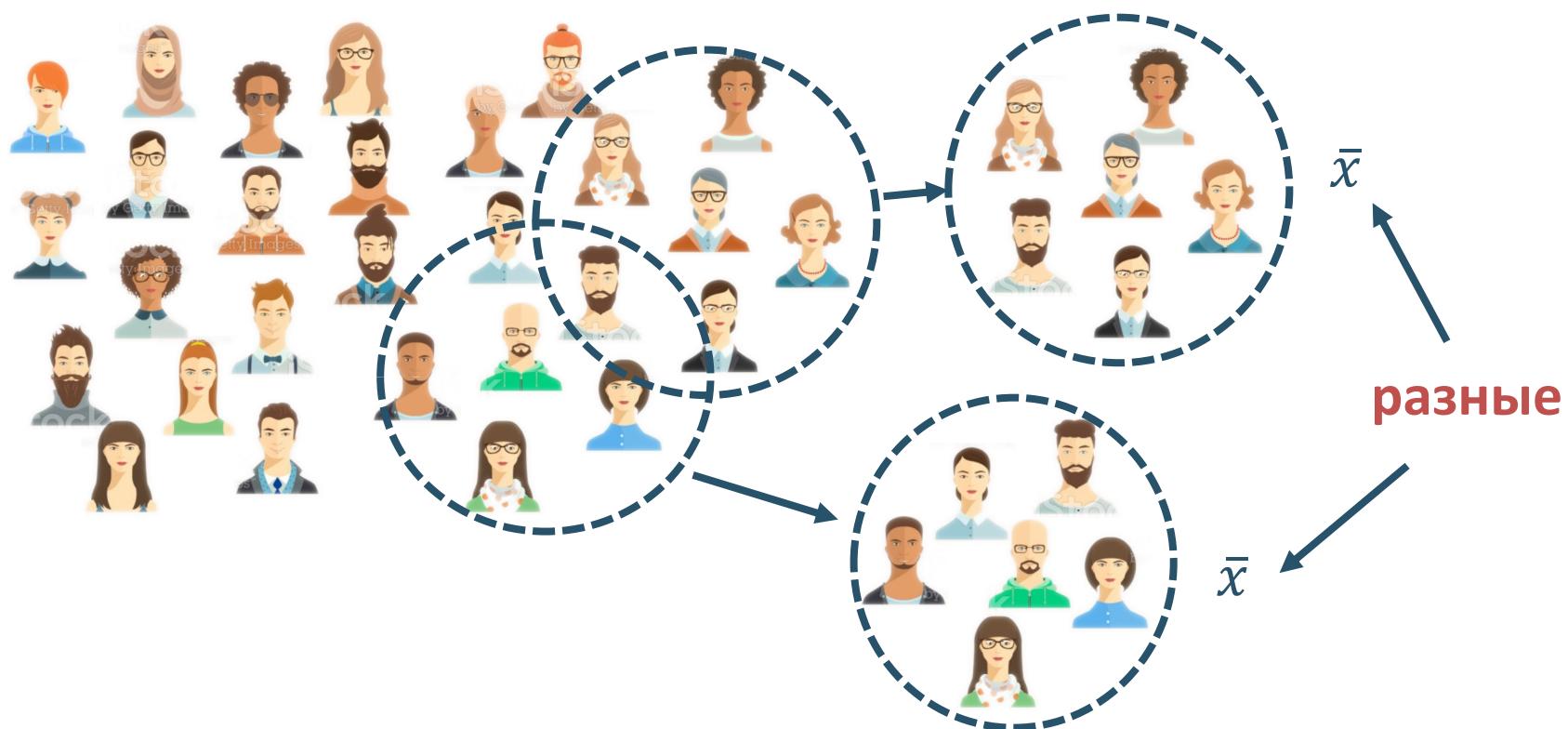
Статистика – любая функция от наблюдений

Примеры: среднее, медиана, максимум и т.п.



Статистика

Каждая статистика – случайная величина, так как она вычисляется на основе случайной выборки, т.е. на основе других случайных величин



Статистика

Каждая статистика – случайная величина, так как она вычисляется на основе случайной выборки, т.е. на основе других случайных величин

- ! Мы будем рассматривать любые статистики, посчитанные на основе выборки как случайные величины и изучать их свойства

Выборка

	Название	Сборы	Год
0	Мстители: Война бесконечности (2018)	2048359754	2018
1	Черная Пантера (2018)	1346913161	2018
2	Мир Юрского периода 2 (2018)	1309484461	2018
3	Суперсемейка 2 (2018)	1242805359	2018

- Строчка таблицы – наблюдение
- Столбец таблицы – переменная

Какими бывают переменные

Переменные



Категориальные

Принимают значения из какого-то ограниченного множества: пол, цвет машины, страна сборки и т.п.

Непрерывные

Могут принимать бесконечное число значений: возраст, вес, цены, кассовые сборы и т.п.

Какими бывают описательные статистики

Описательные статистики



Меры центральной тенденции

Отвечают на вопрос
“а на что похожи типичные
наблюдения из выборки”

Примеры: среднее, мода,
медиана



Меры разброса

Отвечают на вопрос
“а как сильно значения
в выборке могут отличаться от
типовых значений”

Примеры: дисперсия,
стандартное отклонение,
интерквантильный размах

Среднее

Выборочный аналог математического ожидания, рассчитывается по формуле:

$$\bar{x} = \frac{X_1 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

Пример: $x_1 = 1, x_2 = 5, x_3 = -4, x_4 = 3, x_5 = 0$

$$\bar{x} = \frac{1 + 5 + (-4) + 3 + 0}{5} = 1$$

Медиана

Чтобы найти медиану, данные нужно расположить в порядке возрастания. Медианой будет значение, которое оказалось в середине.

Пример 1: $x_1 = 1, x_2 = 5, x_3 = -4, x_4 = 3, x_5 = 0$

$$-4, 0, \textcolor{pink}{1}, 3, 5 \Rightarrow med = 1$$

Если число значений чётное, берётся среднее двух значений, которые «окружают» середину

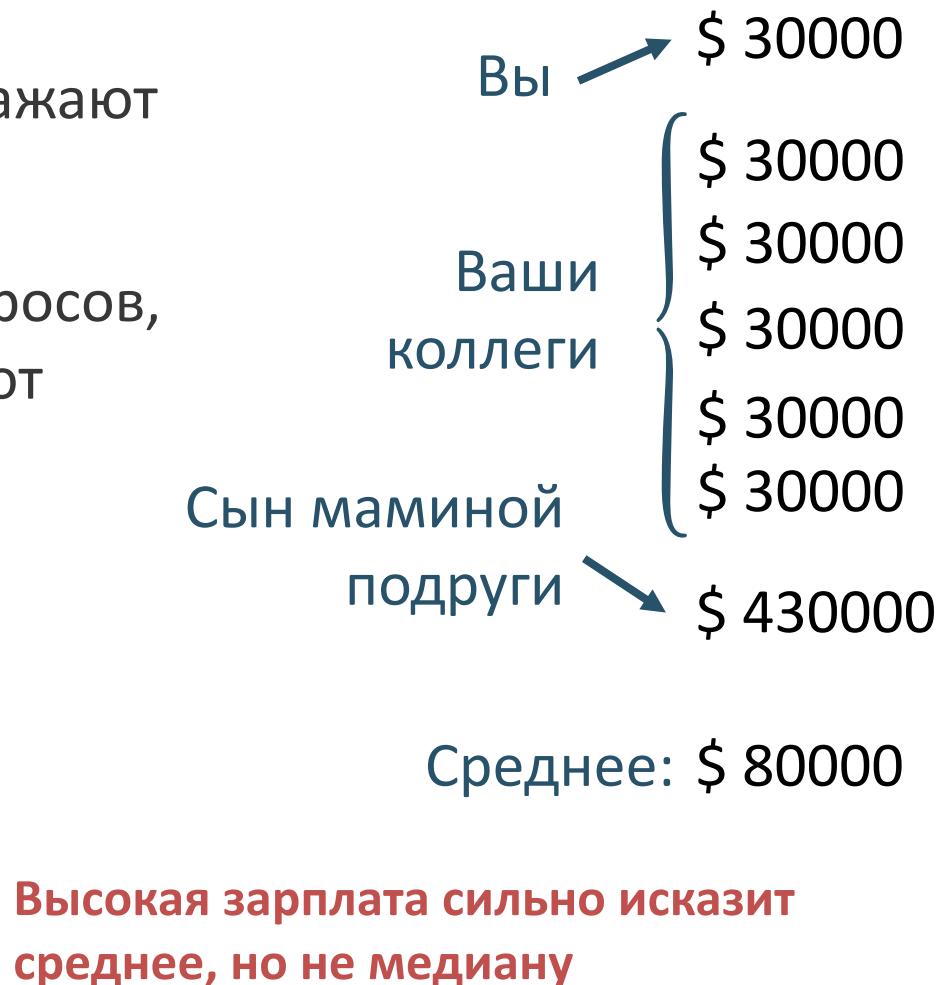
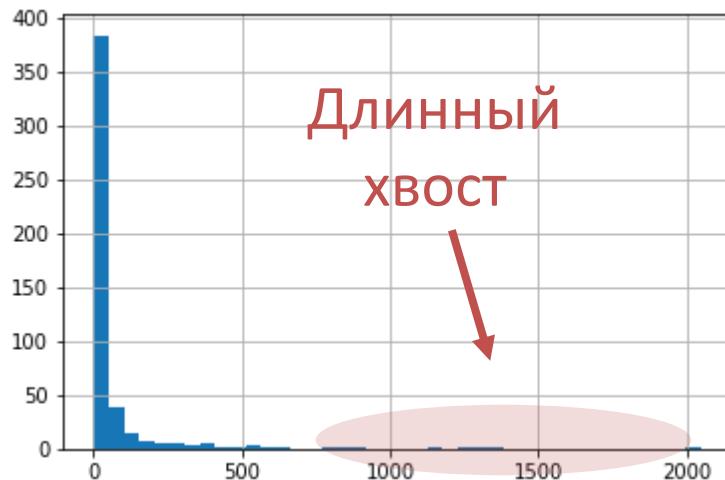
Пример 2: $x_1 = 1, x_2 = 5, x_3 = -4, x_4 = 3$

$$-4, \textcolor{pink}{1}, 3, 5 \Rightarrow med = \frac{1+3}{2} = 2$$

Среднее и медиана

Среднее чувствительно к выбросам в данных,
медиана не чувствительна

- Среднее и медиана отражают типичное значение
- Если в выборке нет выбросов, они примерно совпадают



Выборочная дисперсия

- Хочется понимать, насколько сильно элементы выборки отклоняются от своего типичного значения

Алёна: 18 лет,

Карина: 22 года

$$\bar{x} = \frac{18 + 22}{2} = 20$$

Отклонения от среднего:

$$x_1 - \bar{x} = 18 - 20 = -2$$

$$x_2 - \bar{x} = 22 - 20 = 2$$

Среднее отклонение:

$$\frac{-2 + 2}{2} = 0$$



Как быть?

Выборочная дисперсия

- Хочется понимать, насколько сильно элементы выборки отклоняются от своего типичного значения

Алёна: 18 лет,

Карина: 22 года

$$\bar{x} = \frac{18 + 22}{2} = 20$$

Отклонения от среднего:

$$x_1 - \bar{x} = 18 - 20 = -2$$
$$x_2 - \bar{x} = 22 - 20 = 2$$

Среднее отклонение:

$$\frac{-2 + 2}{2} = 0$$

Выход №1:

$$\frac{| -2 | + | 2 |}{2} = \frac{4}{2} = 2$$

Выборочная дисперсия

- ! Проблема меры, основанной на модуле в том, что она недифференцируема

Модуль неудобно использовать
при теоретических выкладках

Выборочная дисперсия

- Хочется понимать, насколько сильно элементы выборки отклоняются от своего типичного значения

Алёна: 18 лет,

Карина: 22 года

$$\bar{x} = \frac{18 + 22}{2} = 20$$

Отклонения от среднего:

$$x_1 - \bar{x} = 18 - 20 = -2$$
$$x_2 - \bar{x} = 22 - 20 = 2$$

Среднее отклонение:

$$\frac{-2 + 2}{2} = 0$$

Выход №2:

$$\frac{(-2)^2 + 2^2}{2} = \frac{4 + 4}{2} = 4$$

Выборочная дисперсия

- ! Для квадратичной функции всегда есть производная
 - Она обладает хорошими статистическими свойствами
 - Её удобно использовать для теоретических выкладок

Выборочная дисперсия

- Это мера разброса. Показывает, насколько сильно элементы выборки отклоняются от своего типичного значения

$$\hat{\sigma}^2 = \frac{(X_1 - \bar{x})^2 + \dots + (X_n - \bar{x})^2}{n} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{x})^2$$

Пример: $x_1 = 1, x_2 = 5, x_3 = -4, x_4 = 6 \quad \bar{x} = 2$

$$\hat{\sigma}^2 = \frac{(1 - 2)^2 + (5 - 2)^2 + (-4 - 2)^2 + (6 - 2)^2}{4}$$

$$\hat{\sigma}^2 = \frac{1 + 9 + 36 + 16}{4} = 15.5$$

Выборочная дисперсия

Удобнее искать дисперсию по более простой формуле:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n [X_i^2 - 2 \cdot X_i \cdot \bar{x} + \bar{x}^2] = \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \frac{2 \bar{x}}{n} \sum_{i=1}^n X_i + \frac{1}{n} \sum_{i=1}^n \bar{x}^2 \\ &= \overline{x^2} - 2 \bar{x}^2 + \bar{x}^2 = \overline{x^2} - \bar{x}^2\end{aligned}$$

Выборочная дисперсия

Удобнее искать дисперсию по более простой формуле:

$$\hat{\sigma}^2 = \bar{x^2} - \bar{x}^2$$

Пример: $x_1 = 1, x_2 = 5, x_3 = -4, x_4 = 6 \quad \bar{x} = 2$

$$\bar{x^2} = \frac{1^2 + 5^2 + (-4)^2 + 6^2}{4} = 19.5$$

$$\hat{\sigma}^2 = 19.5 - 4 = 15.5$$

Стандартное отклонение

- Дисперсия измеряется в квадратных величинах
- Чтобы вернуться назад к исходным величинам, можно взять из неё квадратных корень

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

лет = $\sqrt{\text{лет в квадрате}}$

Несмешённая выборочная дисперсия

- Обычно на практике используют другую формулу:

$$s^2 = \frac{(X_1 - \bar{x})^2 + \cdots + (X_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^n (X_i - \bar{x})^2$$

- Это несмешённая дисперсия
- Что это значит, мы подробно обсудим позднее

Перцентиль

Перцентиль порядка k – это такое число, что $k\%$ выборки меньше этого числа

- Перцентиль это выборочный аналог квантиля
- Проще всего вычислять его по упорядоченной выборке

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

- Квартили – перцентили с шагом в 0.25:

$$x_{(0.25 \cdot [n+1])} \quad x_{(0.5 \cdot [n+1])} \quad x_{(0.75 \cdot [n+1])}$$

медиана

Интерквантильный размах:

$$IQR = x_{(0.75 \cdot [n+1])} - x_{(0.25 \cdot [n+1])}$$

Перцентиль

Пример:

1, 5, 3, -1, -4, 3, 3, -10, 2, -1

упорядочили
выборку

-10, -4, -1, -1, 1, 2, 3, 3, 3, 5

$$\Rightarrow med = \frac{1+2}{2} = 1.5$$

позиция медианы:

$$0.5 \cdot (10 + 1) = 5.5$$

Перцентиль

Пример:

1, 5, 3, -1, -4, 3, 3, -10, 2, -1

-10, -4, -1, -1, 1, 2, 3, 3, 3, 5

упорядочили
выборку

$$\Rightarrow \frac{-4+(-1)}{2} = -2.5$$

позиция нижней квартили: $0.75 \cdot (10 + 1) = 8.25$

- Перцентили в спорных случаях можно считать по-разному, каждая библиотека предоставляет разные варианты

Перцентиль

Пример:

1, 5, 3, -1, -4, 3, 3, -10, 2, -1

-10, -4, -1, -1, 1, 2, 3, 3, 3, 5

упорядочили
выборку

$$\Rightarrow \frac{3 + 3}{2} = 3$$

позиция верхней квартили: $0.75 \cdot (10 + 1) = 8.25$

- Перцентили в спорных случаях можно считать по-разному, каждая библиотека предоставляет разные варианты

Резюме

- Репрезентативная выборка – отражает свойства генеральной совокупность
- Мы будем в дальнейшем предполагать, что все наблюдения, которые мы делаем, не зависят друг от друга

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

Гистограмма и эмпирическая функция распределения

Эмпирическая функция распределения

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

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

Эмпирическая функция распределения – функция, которая определяет для каждого x частоту события $X \leq x$, то есть

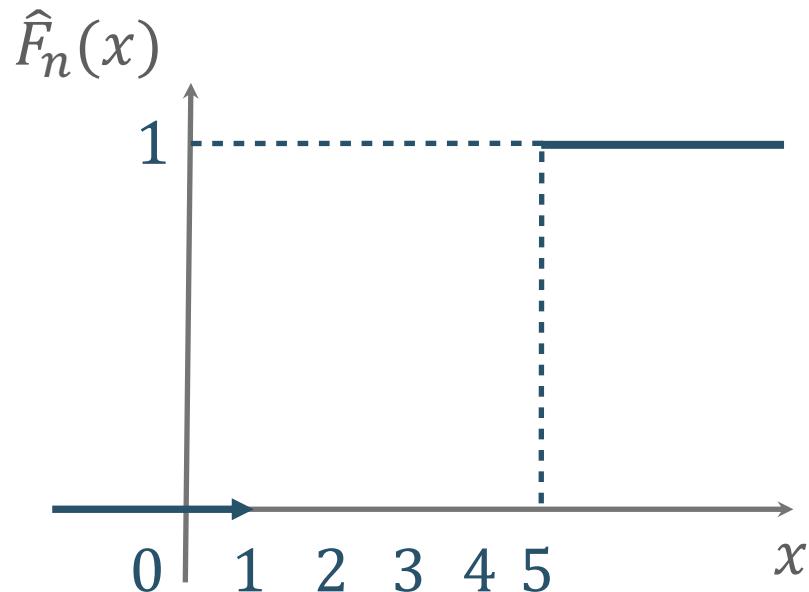
$$\hat{F}_n(x) = \widehat{\mathbb{P}}(X \leq x) = \frac{1}{n} \sum_{i=1}^n [X_i \leq x],$$

где $[]$ – индикаторная функция, то есть:

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

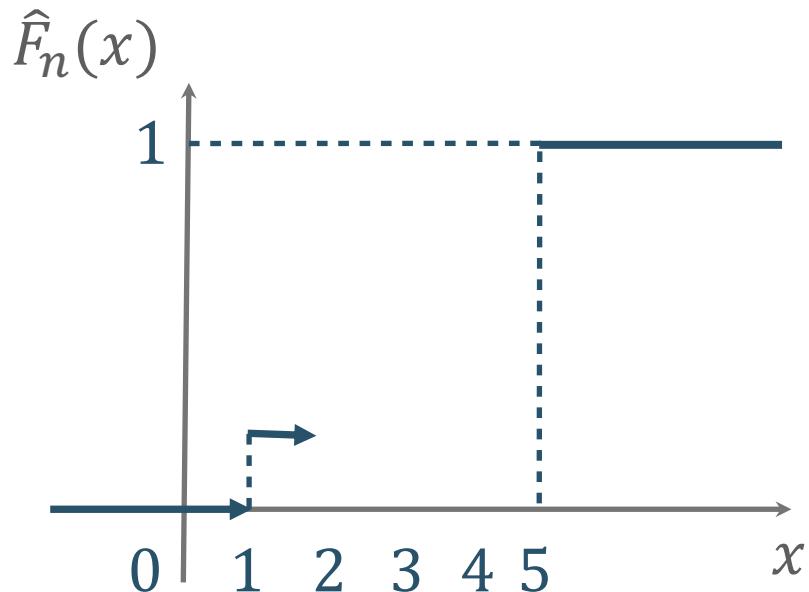
Эмпирическая функция распределения

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



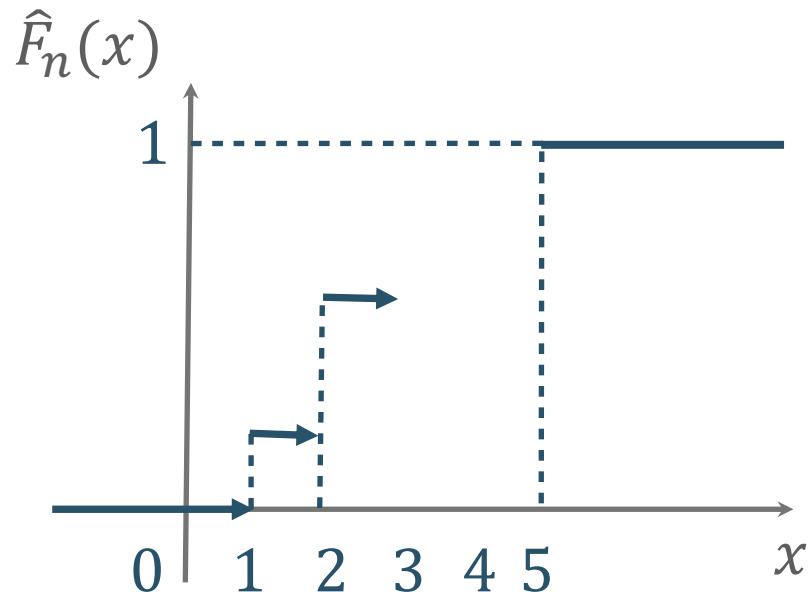
Эмпирическая функция распределения

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



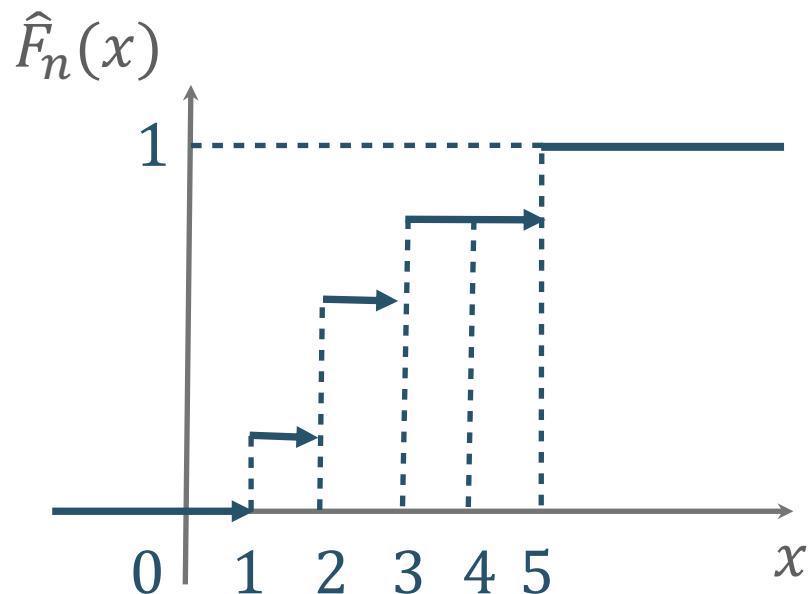
Эмпирическая функция распределения

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



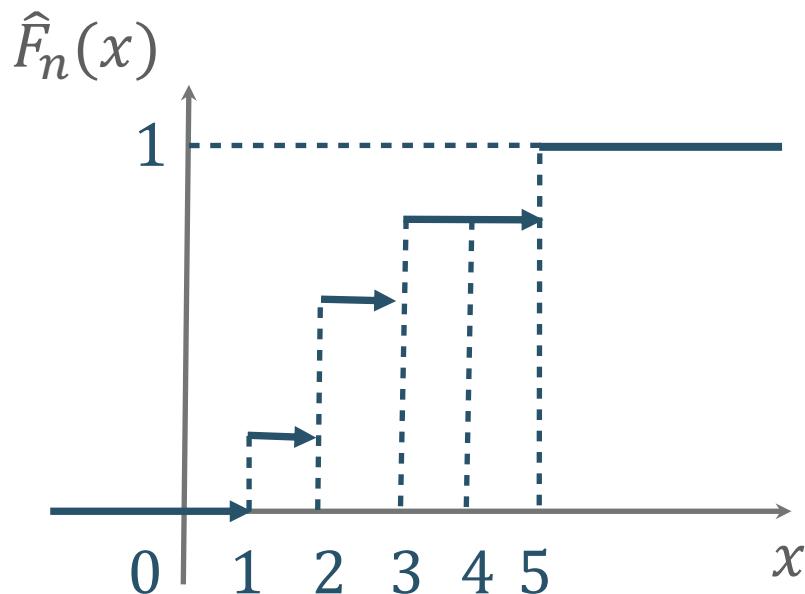
Эмпирическая функция распределения

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



Эмпирическая функция распределения

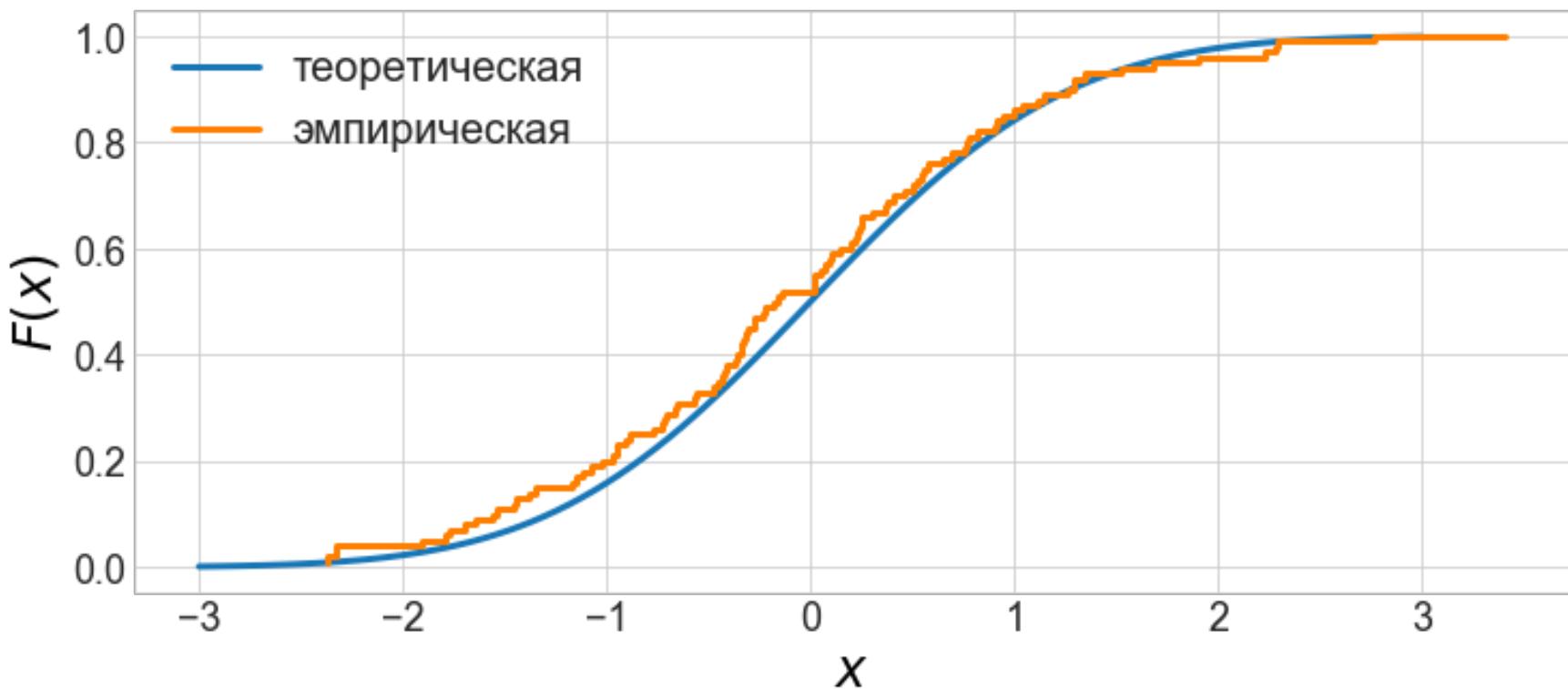
Пример: $x_1 = 2, x_2 = 5, x_3 = 2, x_4 = 3, x_5 = 1$



По аналогии строится
теоретическая функция
распределения для
дискретных случайных
величин

Эмпирическая функция распределения

Чем больше выборка, тем чаще ступеньки и тем больше эмпирическая функция распределения похожа на теоретическую.

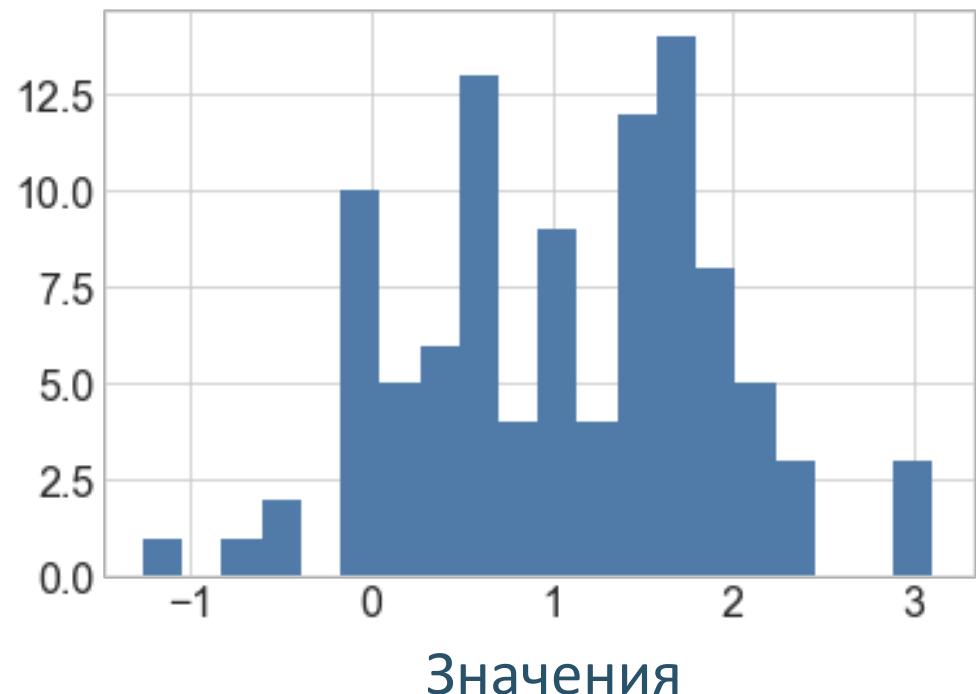


Гистограмма

Гистограмма – эмпирическая оценка **плотности** распределения.
По оси x откладывают значения,
по оси y частоты.

Область возможных значений обычно дробят на отрезки, **бины**.
Чем короче бины, тем детальнее рисуется гистограмма.

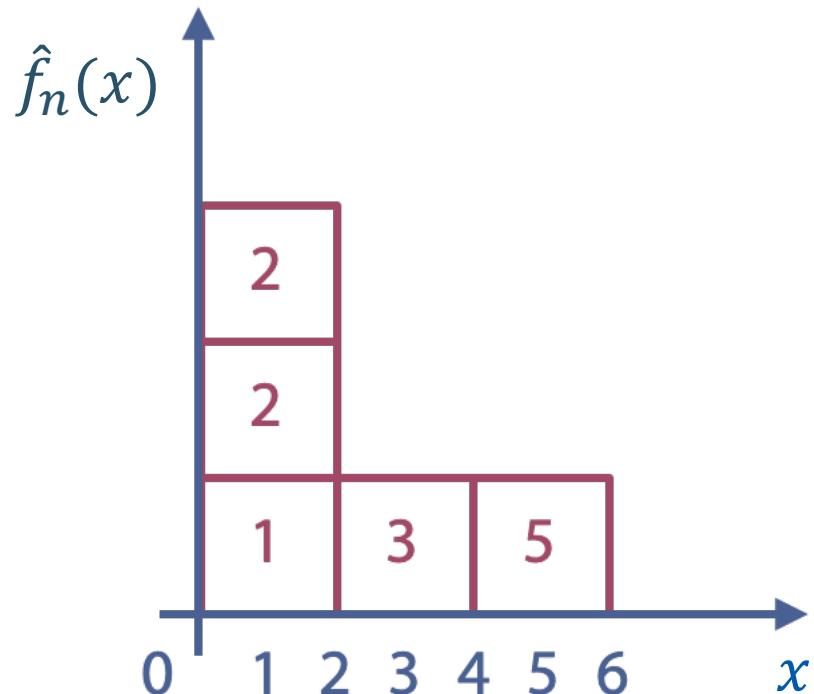
Сколько значений
попали в текущий
отрезок (бин)



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

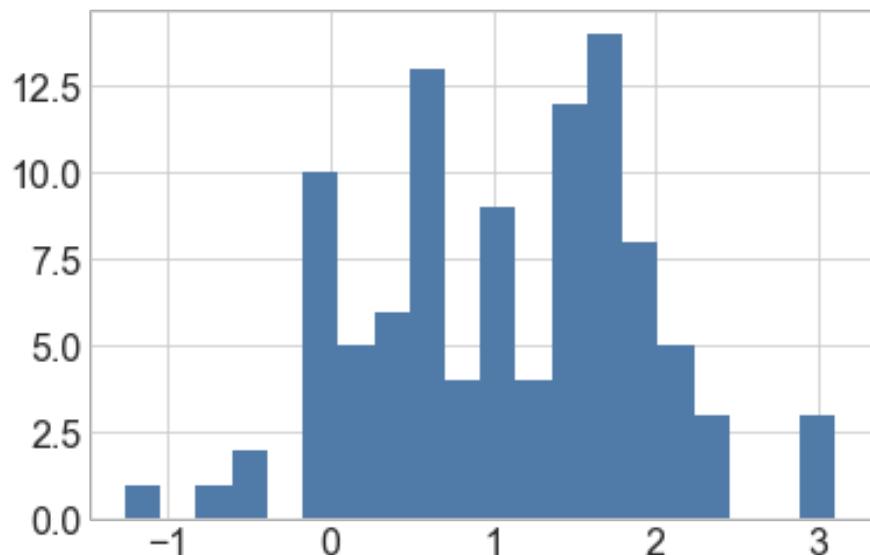
[ложь] = 0

Размер бина (длина отрезка):

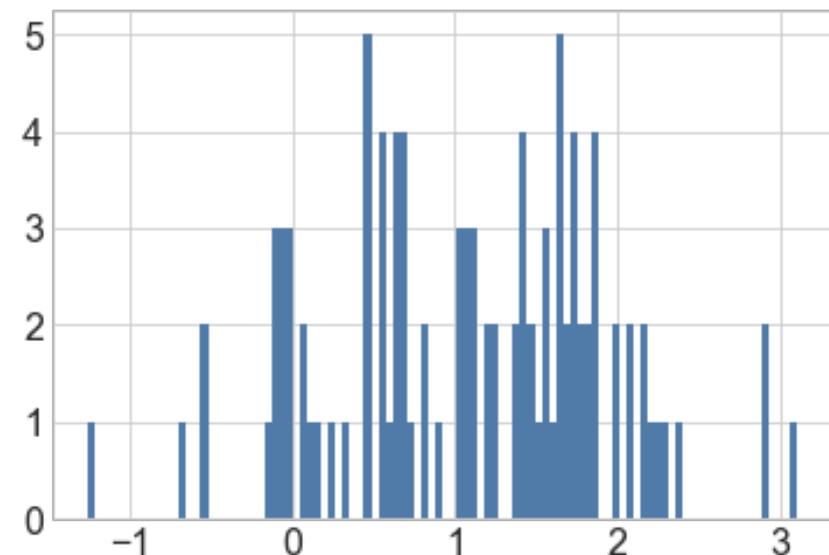
$$h = z_{k+1} - z_k$$

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



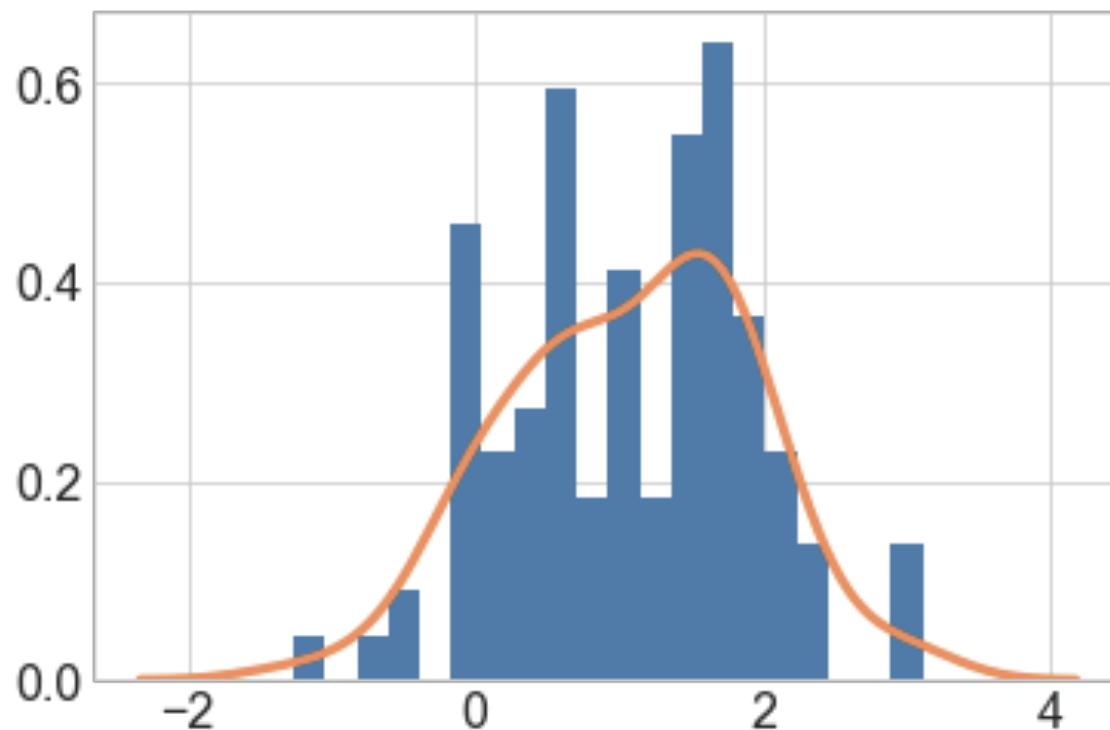
20 бинов



100 бинов

Ядерные оценки плотности

- Ядерные оценки плотности (kernel density estimation, KDE) 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

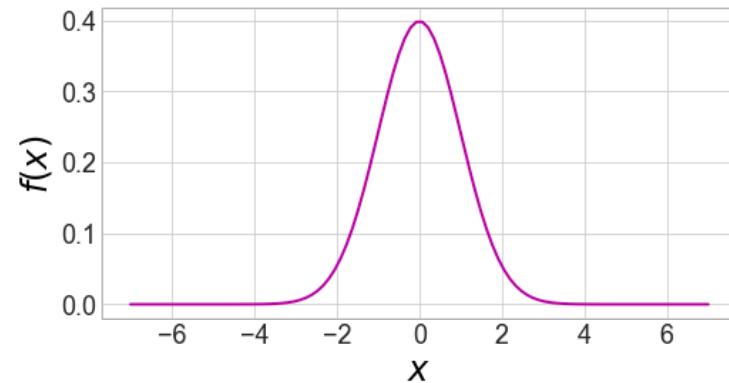
Ядерные оценки плотности

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

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

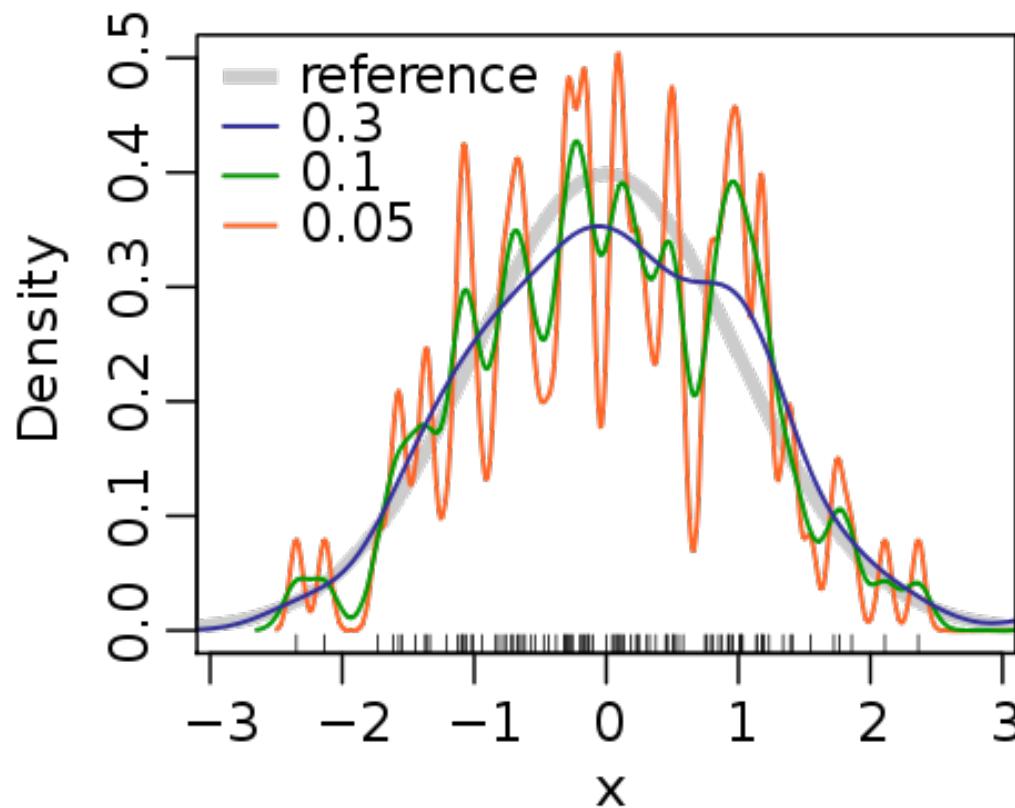
Ядерные функции бывают разными, чаще всего используют Гауссовское ядро:

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

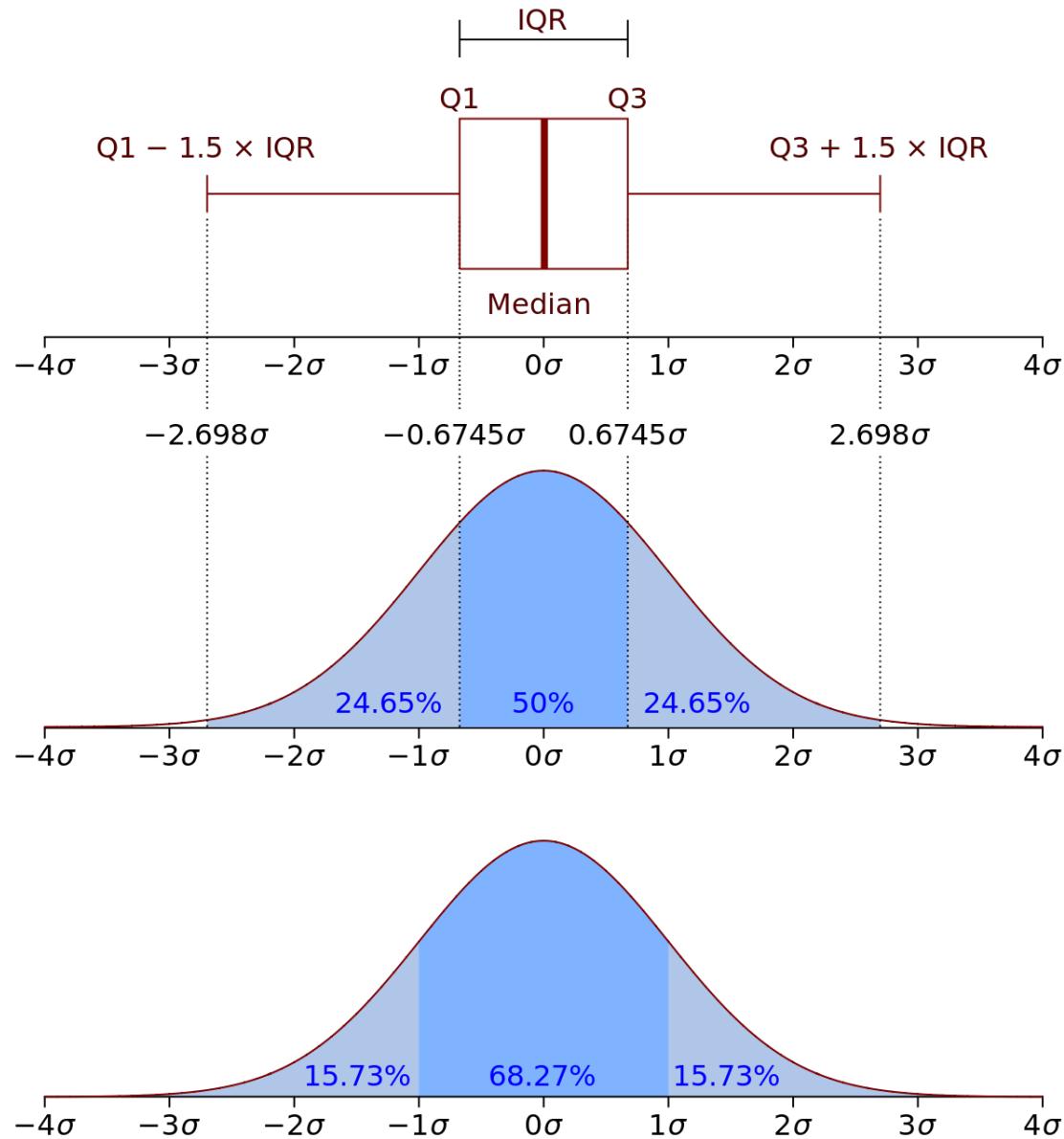


Пример

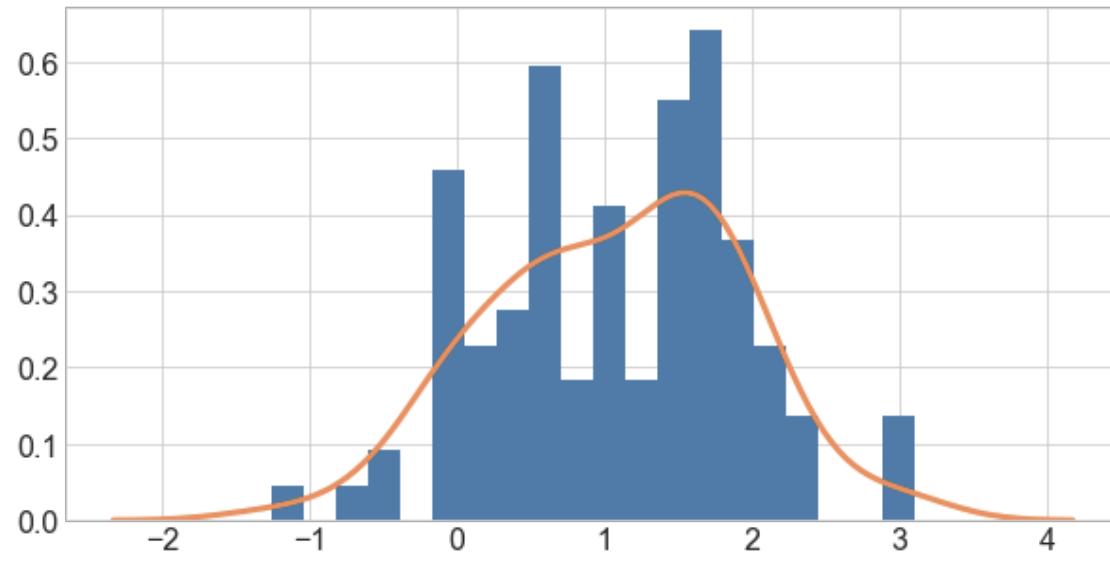
- Серые чёрточки на оси x – наблюдения
- Величина параметра h (ширина окна) влияет на то, насколько гладкой получается итоговая кривая



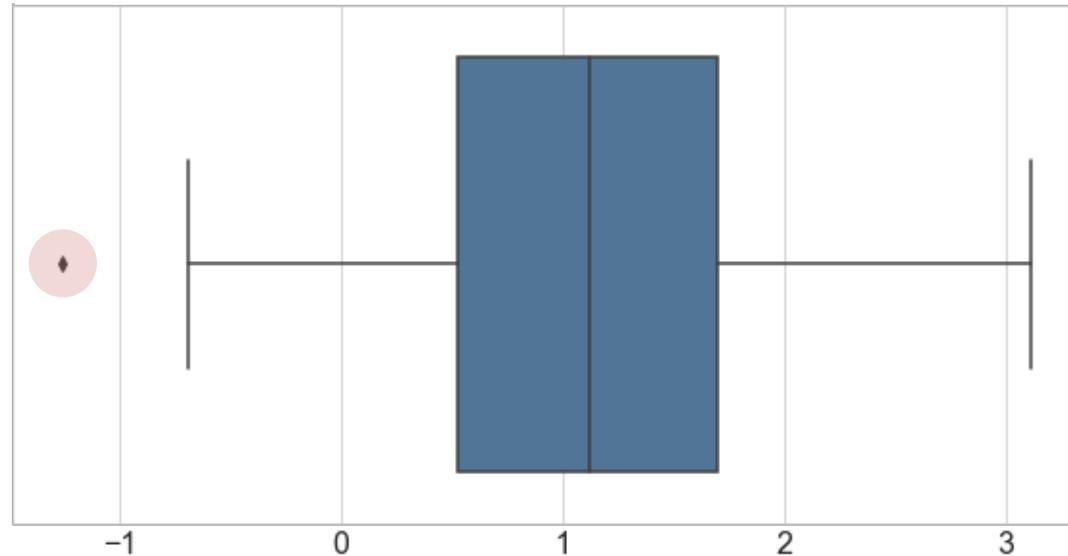
Ящик с усами (Boxplot)



Ящик с усами



Аномальное
значение



Ящик с усами (Boxplot)

- Позволяет обобщить данные
- Показывает наличие выбросов
- Даёт некоторое представление о симметрии данных
- Позволяет сравнить несколько переменных между собой

Резюме

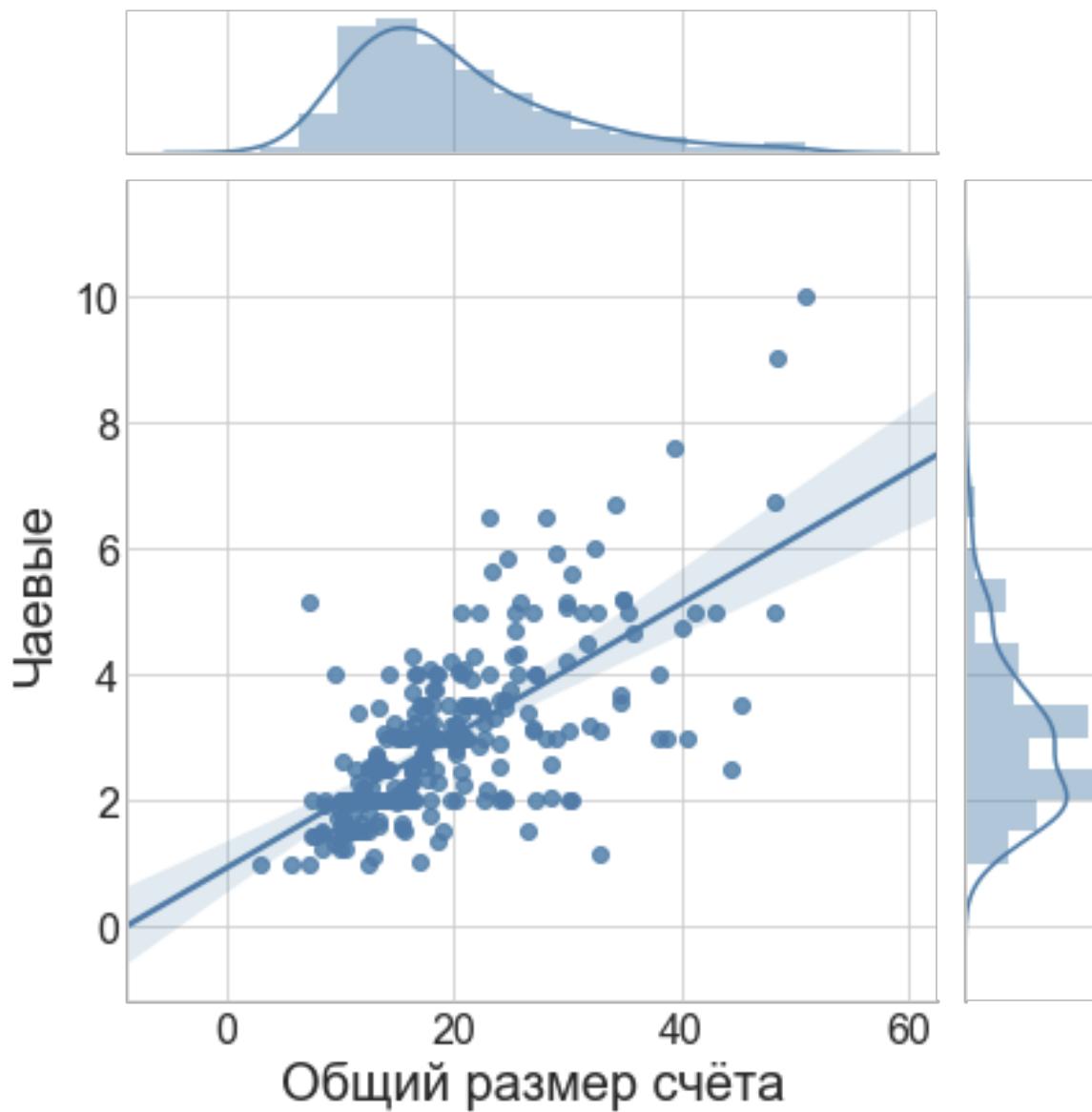
- Плотность распределения вероятностей и функцию распределения также можно оценить по выборке
- Для того, что получить непрерывную оценку плотности распределения используют ядерное сглаживание

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

- Ящик с усами позволяет визуализировать основные описательные статистики

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, b – константы

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

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

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

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

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:

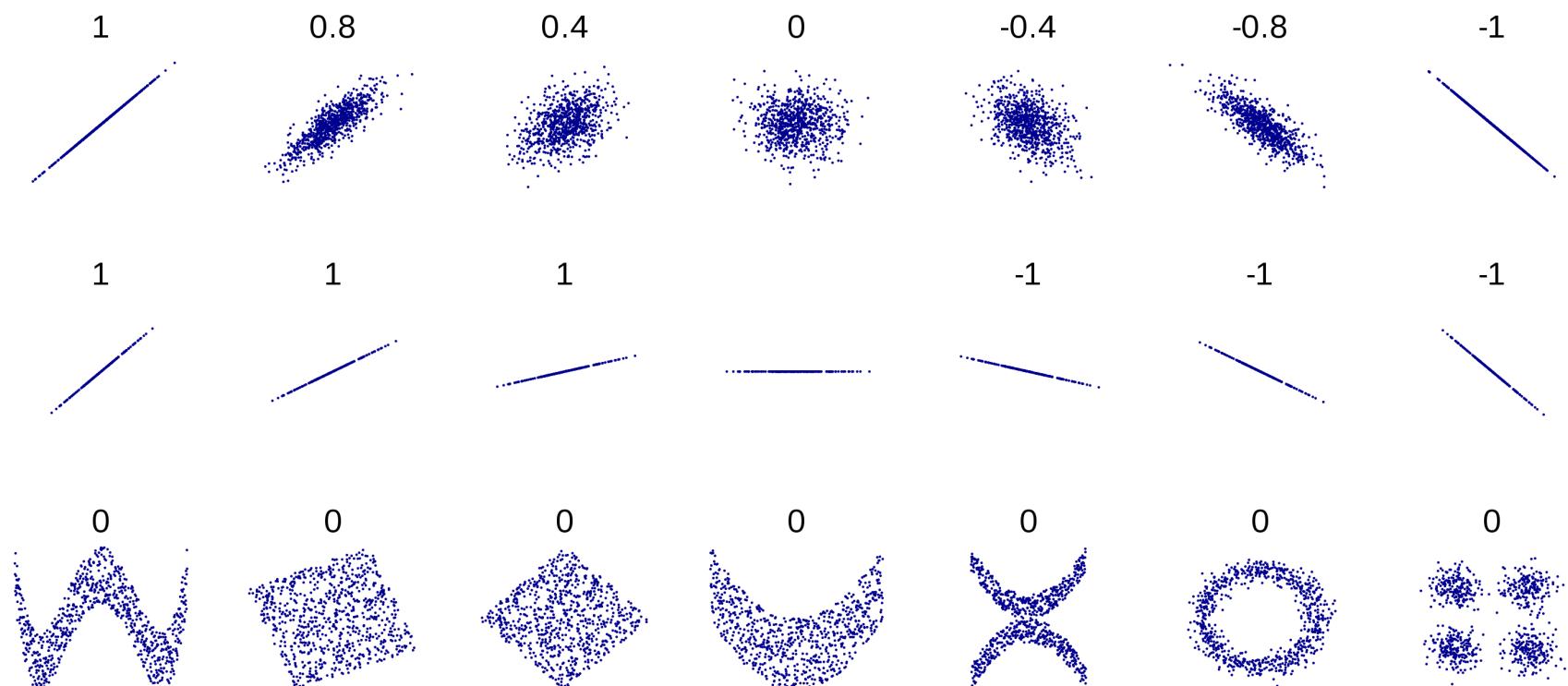
Выборочная ковариация:

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

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

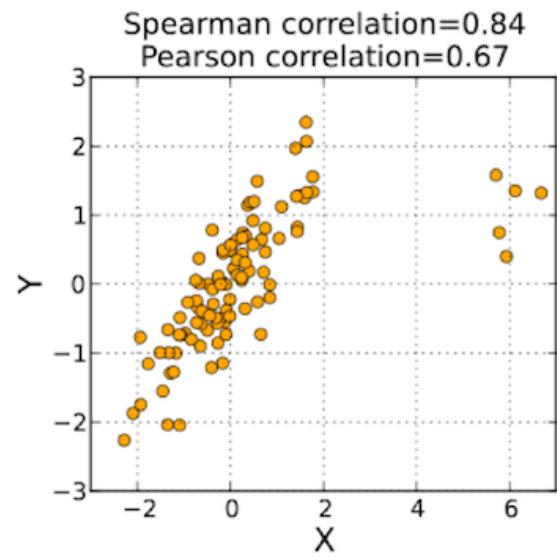
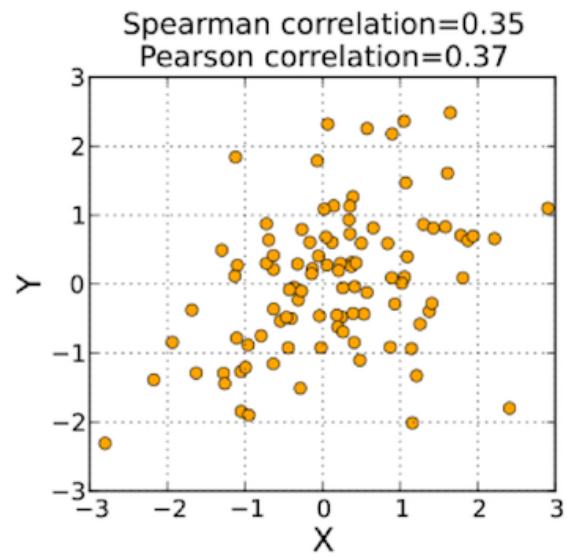
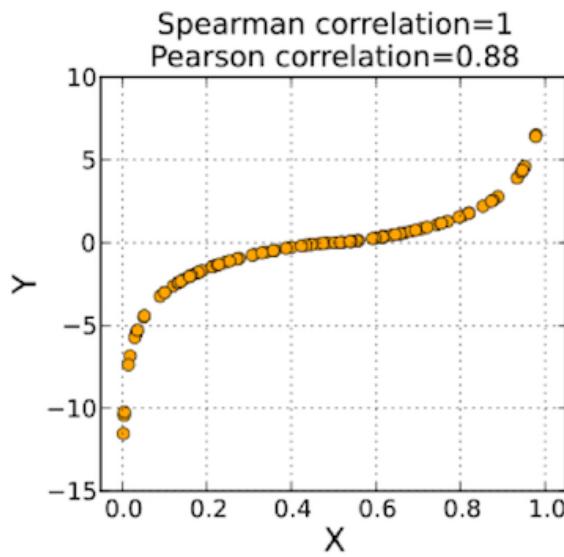
Корреляция Спирмена

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

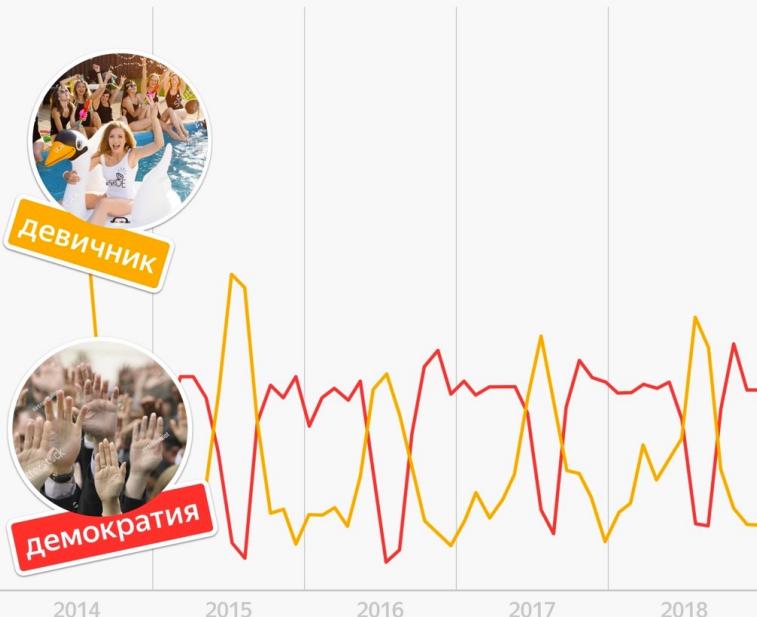


Корреляция Спирмэна пытается уловить
в данных монотонность

Корреляция не означает причинность

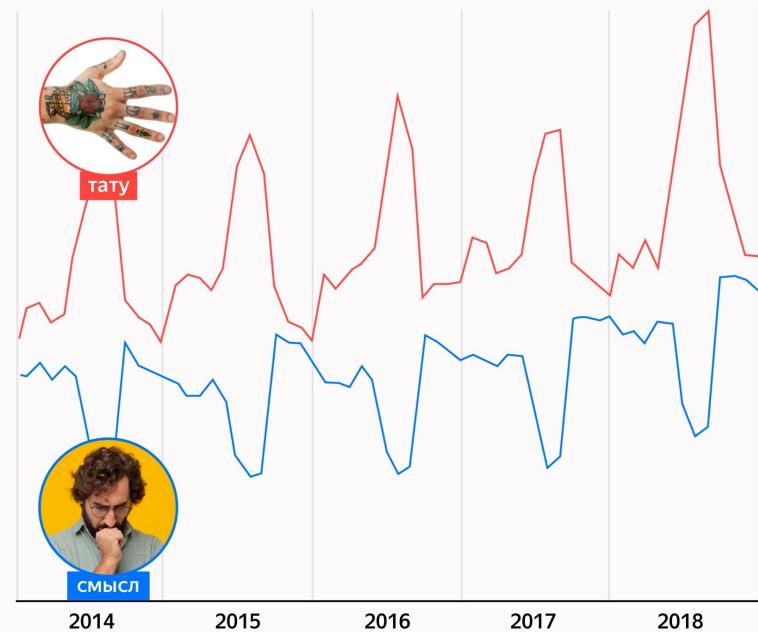
Парадоксы в Поиске — Яндекс

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



Парадоксы в Поиске — Яндекс

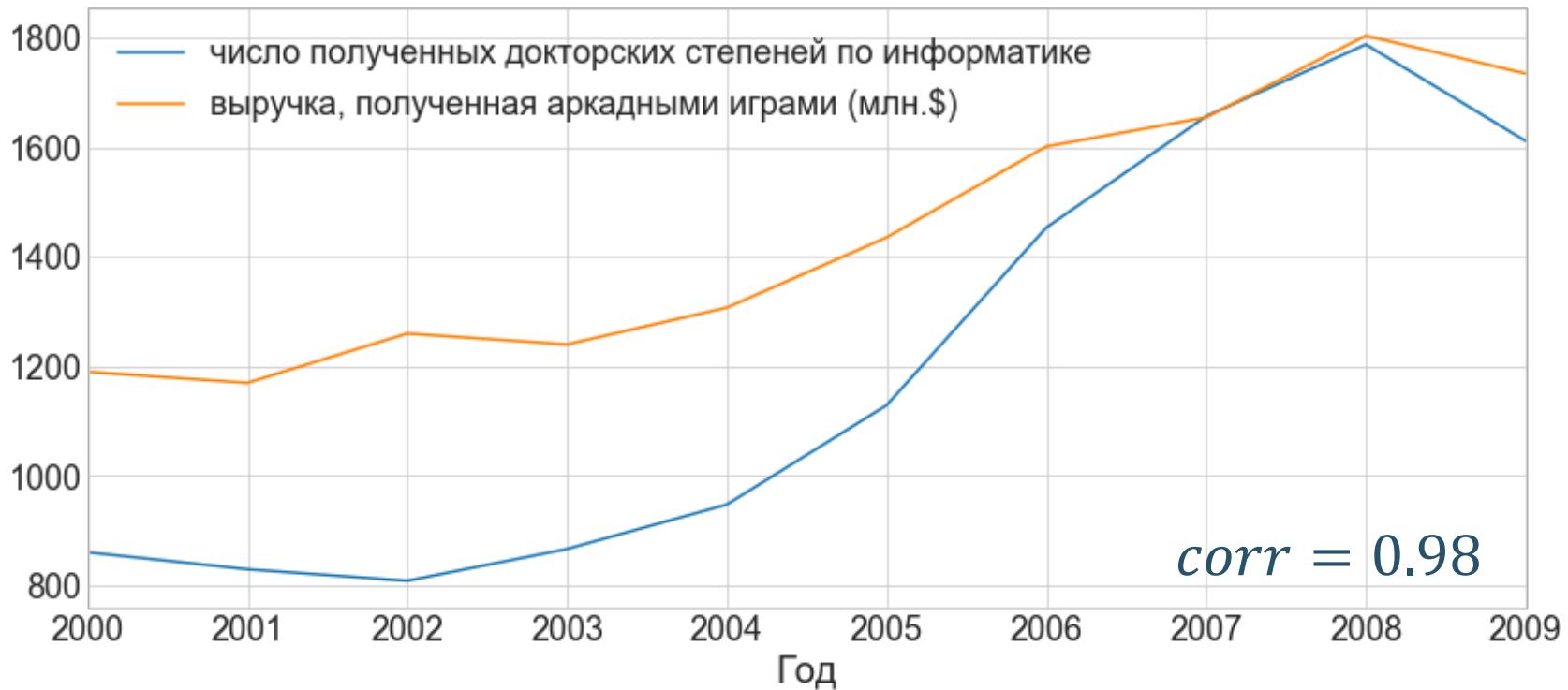
Когда в Поиске растёт интерес к **тату**, снижается доля запросов со словом **смысл**



► 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