

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

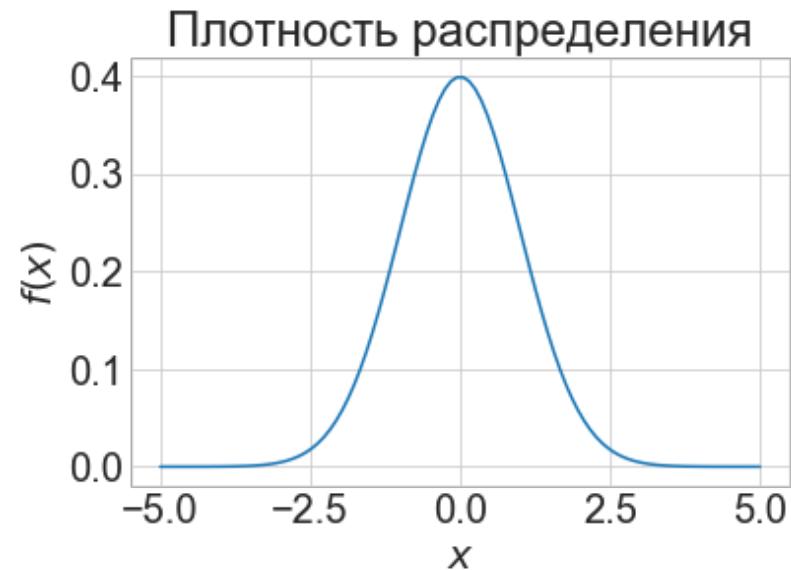
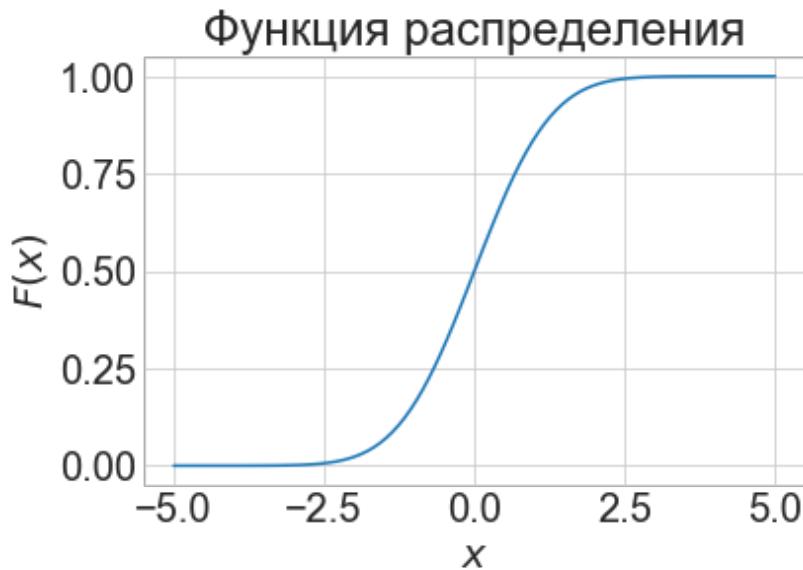
Как устроен мир



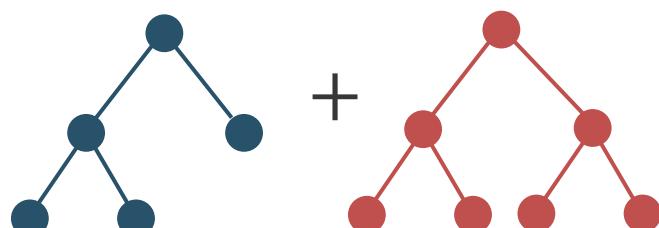
X

- 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.

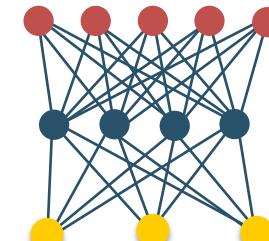
Устройство сундука



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?

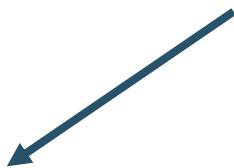
- Изучать выборки и их свойства
- Предполагать, какие процессы порождают данные, описывать своё незнание с помощью какой-то модели
- Разбираться, насколько наши предположения согласуются с выборками



Распределение случайной величины

Какими бывают случайные величины

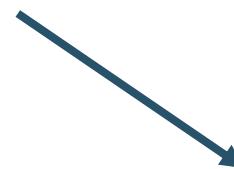
Случайные величины



Дискретные

Множество значений
конечно или счётно

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



Непрерывные

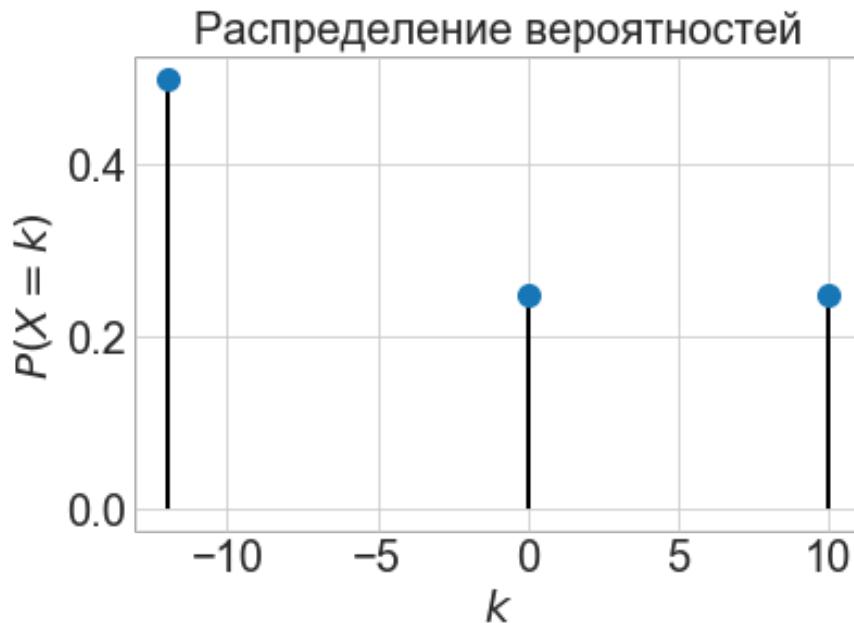
Принимают бесконечное,
континуальное число
значений

(рост, время ожидания
автобуса, вес)

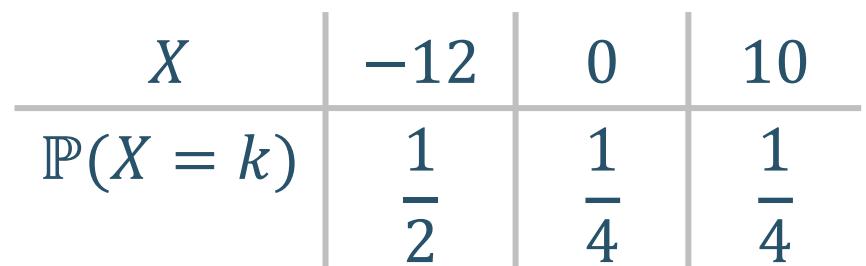
Дискретные случайные величины

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

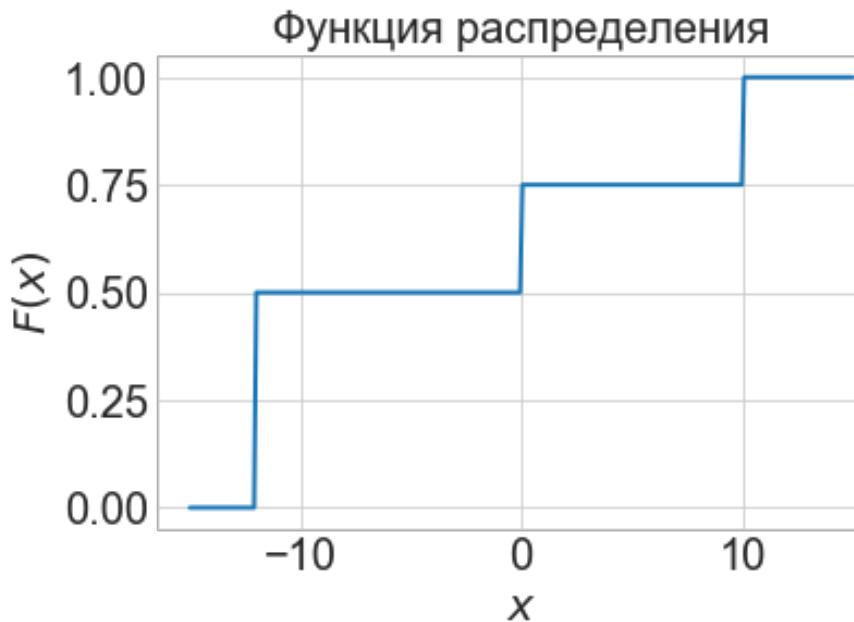


Дискретные случайные величины

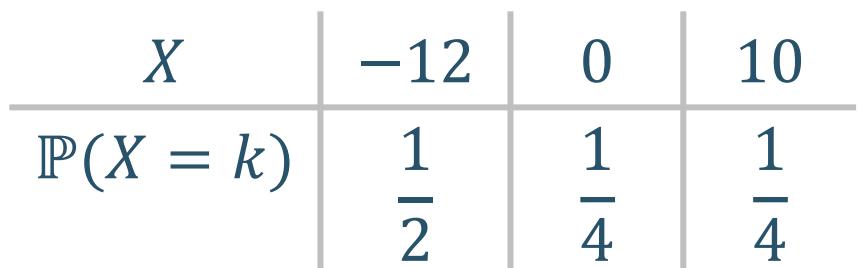
Функция распределения – функция, которая определяет вероятность события $X \leq x$, то есть

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

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

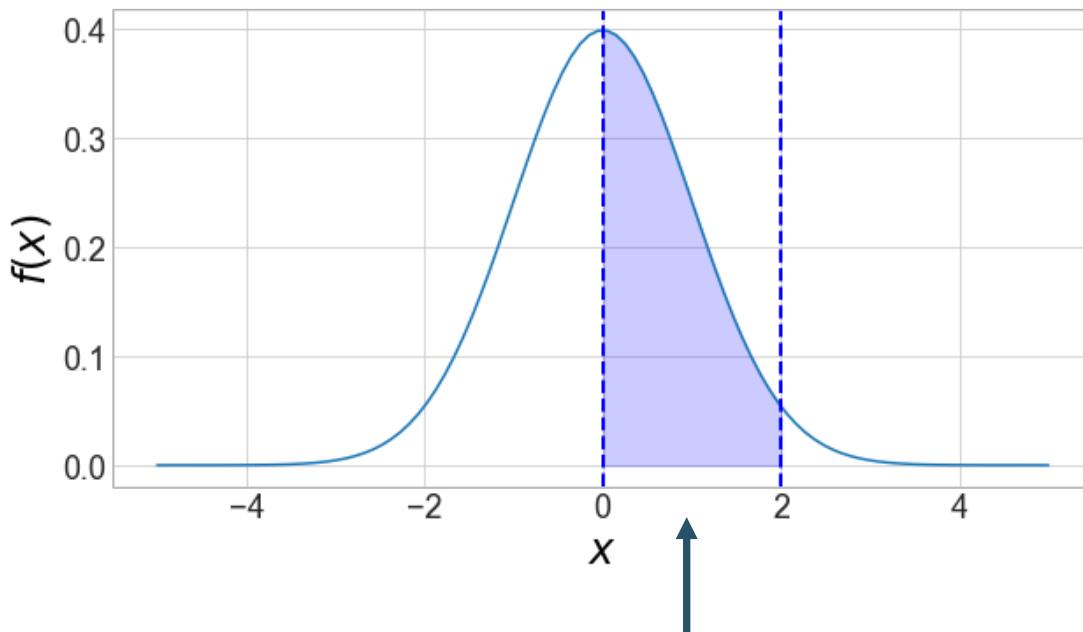


Пример: лотерея



Непрерывные случайные величины

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



Площадь равна вероятности попасть
на отрезок от нуля до двух

Пример:
нормальное
распределение

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

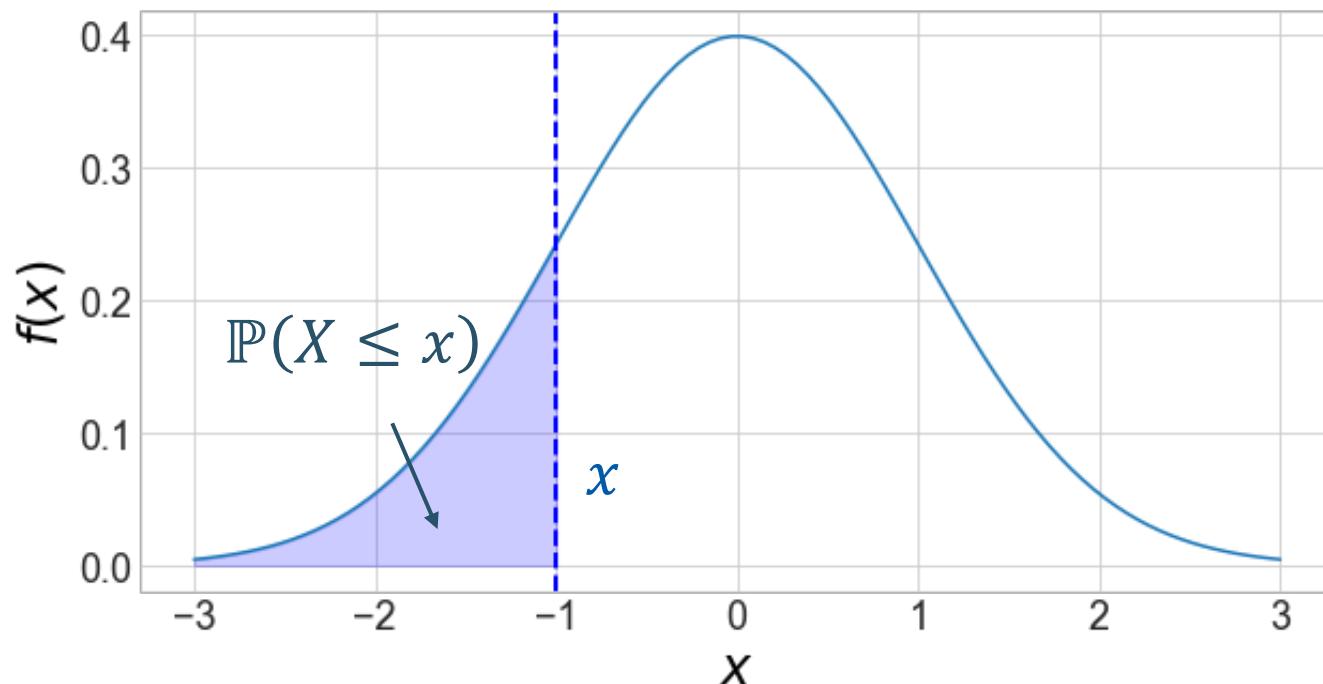
$$= \int_0^2 f(x) \, dx$$

Площадь под всей плотностью должна быть равна 1

Непрерывные случайные величины

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

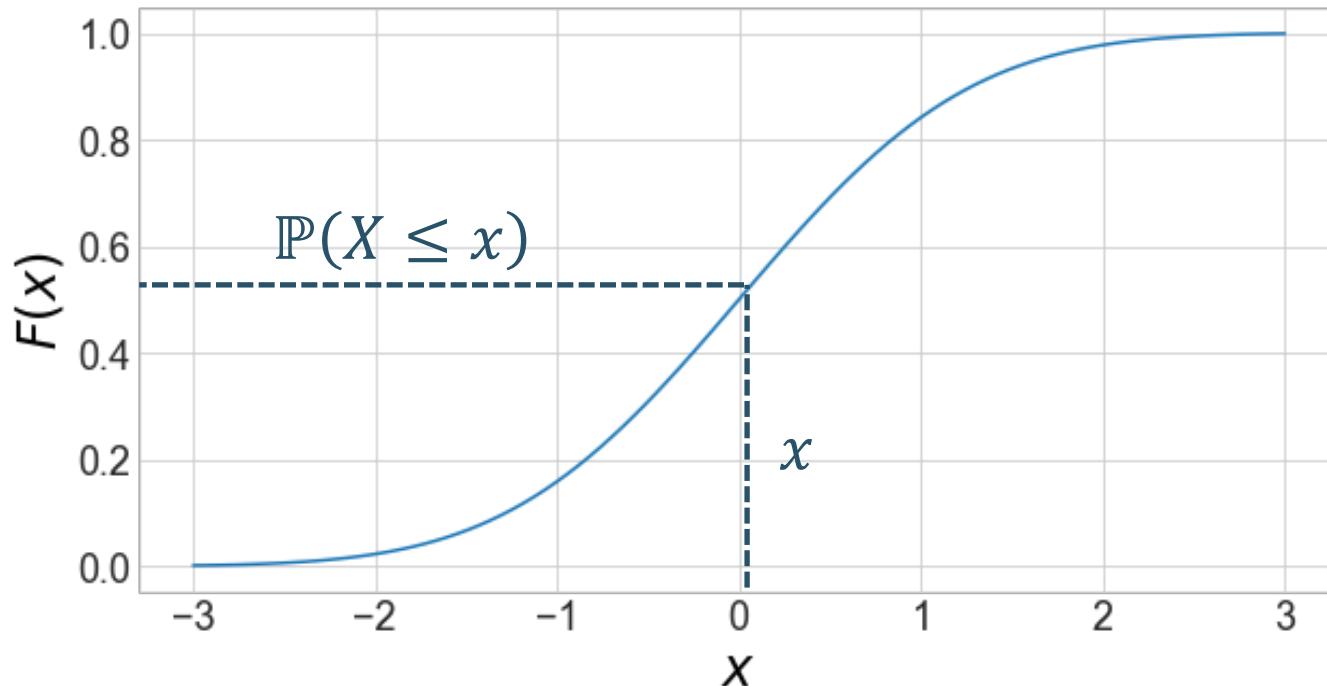
$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t) \, dt, f(t) \text{ – плотность}$$



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

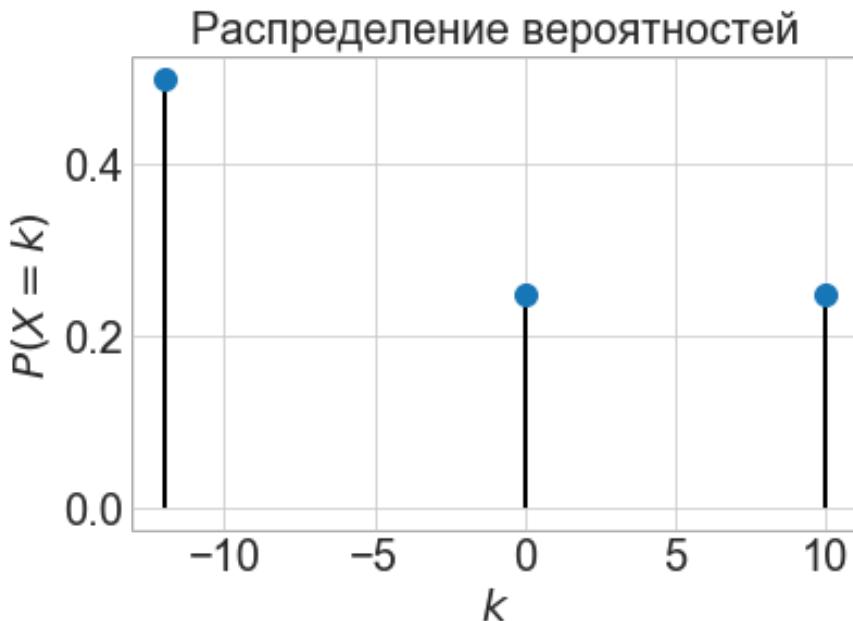
Характеристики случайных величин

Математическое ожидание

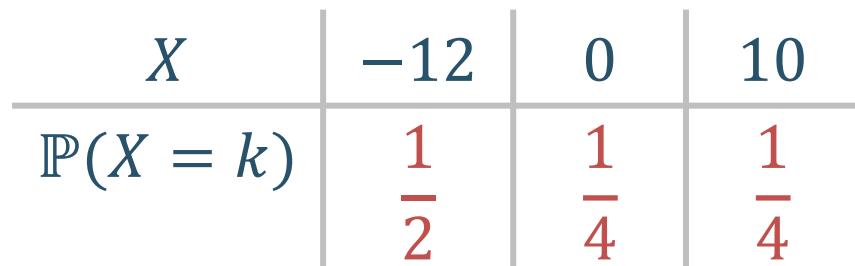
Математическое ожидание – среднее значение случайной величины

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

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



Пример: лотерея



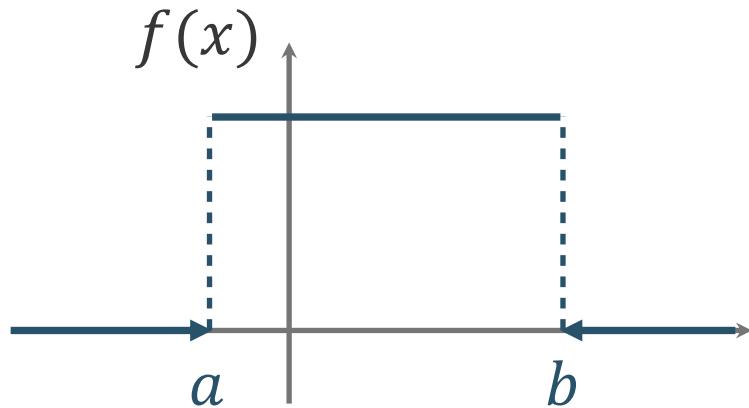
$$\mathbb{E}(X) = -12 \cdot 0.5 + 0 \cdot 0.25 + 10 \cdot 0.25 = -3.5 \text{ рубля}$$

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

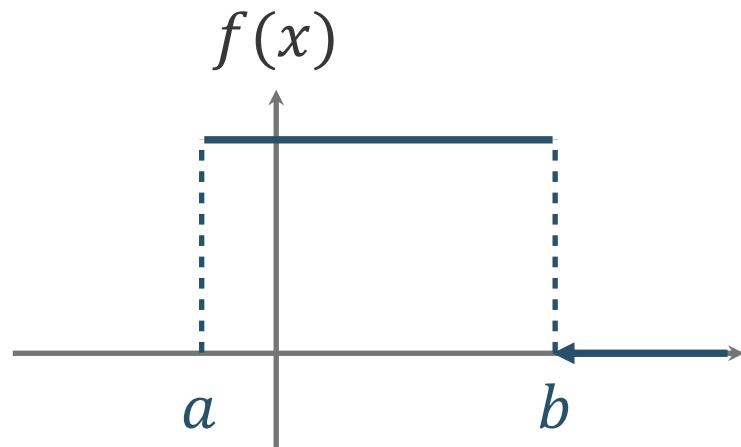
$$\mathbb{E}(X^2) = (-12)^2 \cdot 0.5 + 0^2 \cdot 0.25 + 10^2 \cdot 0.25 = 97 \text{ рублей}^2$$

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

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

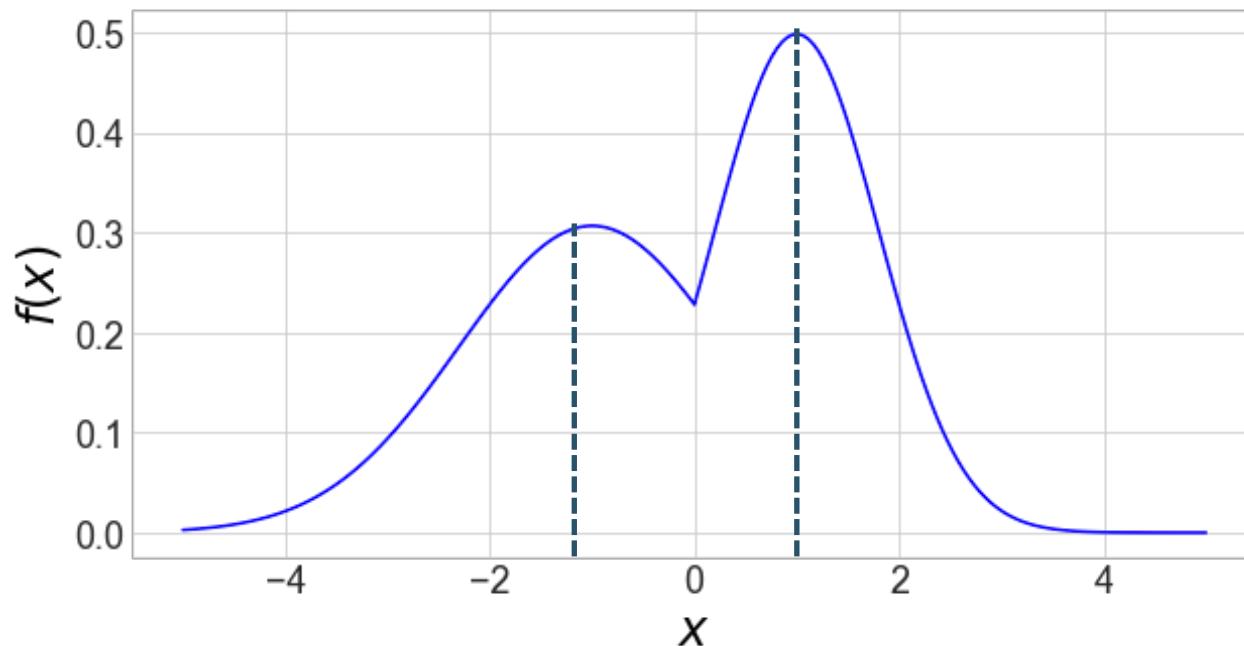
X, Y – случайные величины a – константа

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)

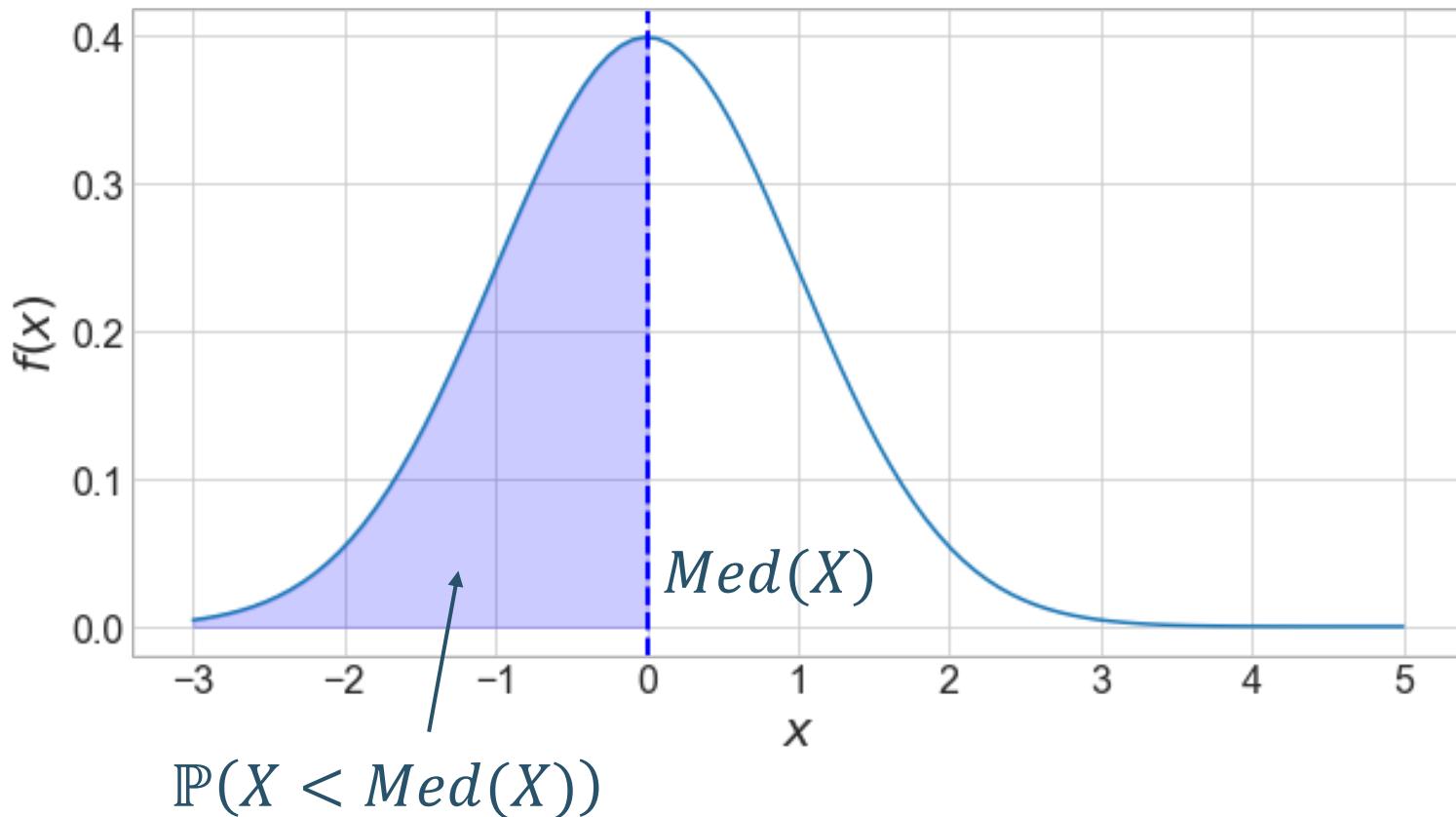
На практике встречаются мультимодальные распределения



Медиана

Медиана случайной величины – такое её значение, что

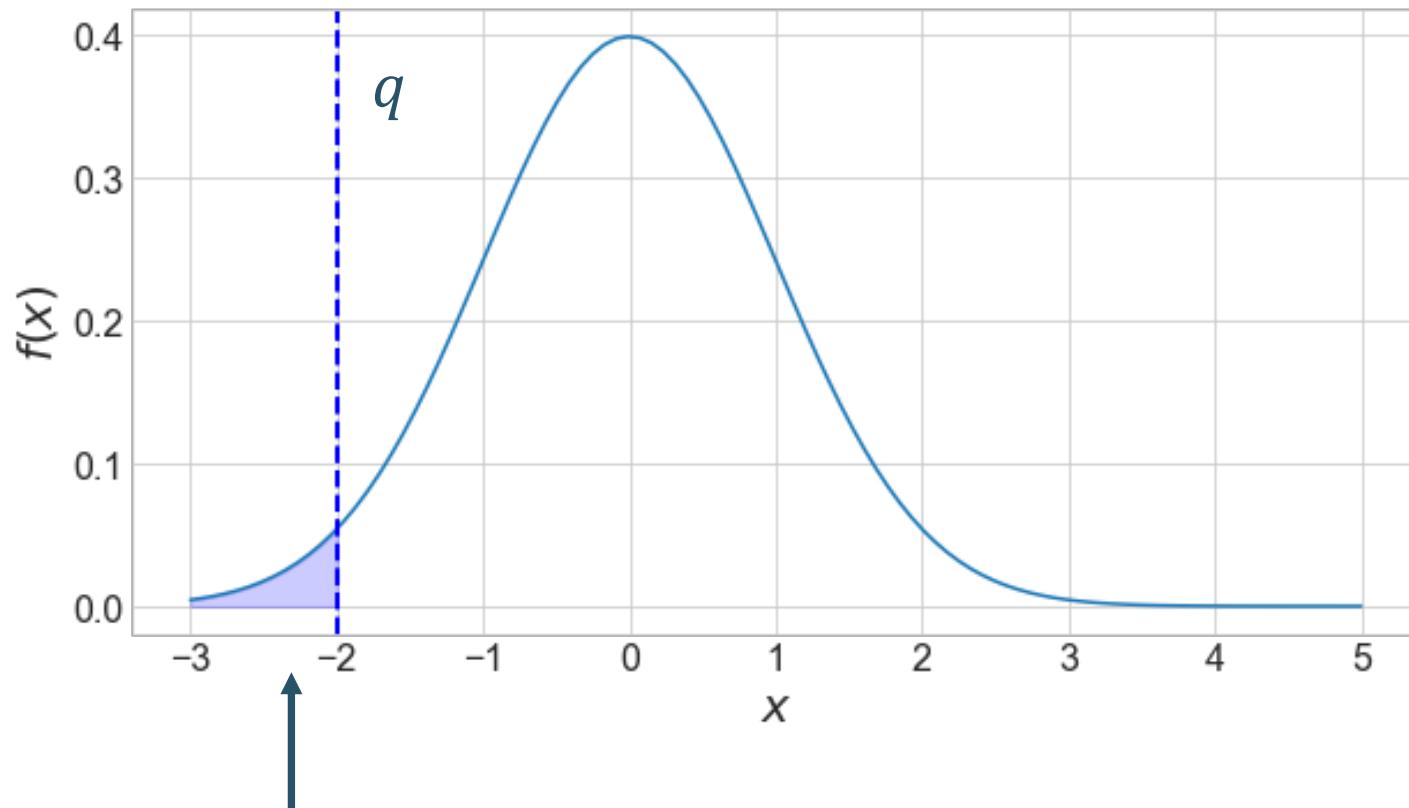
$$\mathbb{P}(X < \text{Med}(X)) = \mathbb{P}(X > \text{Med}(X)) = 0.5$$



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

$$Y_i \sim Bern(p)$$

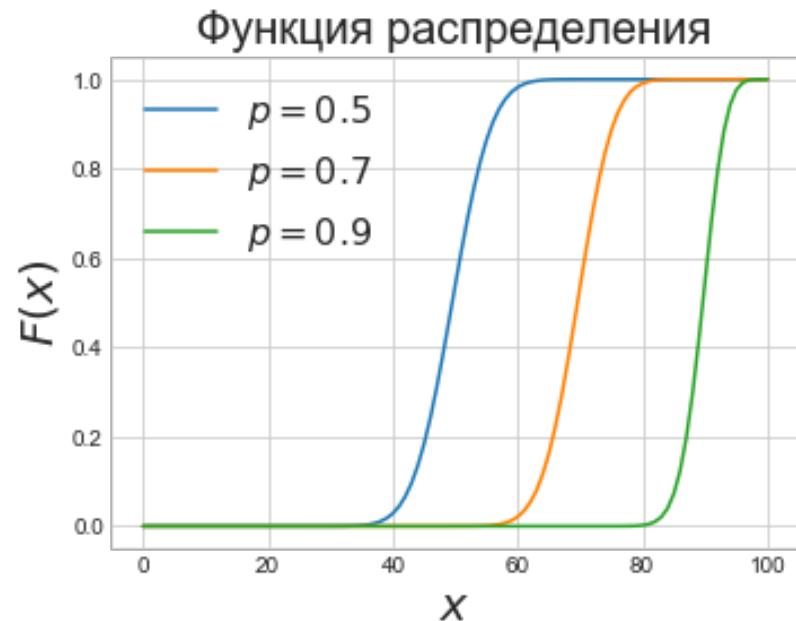
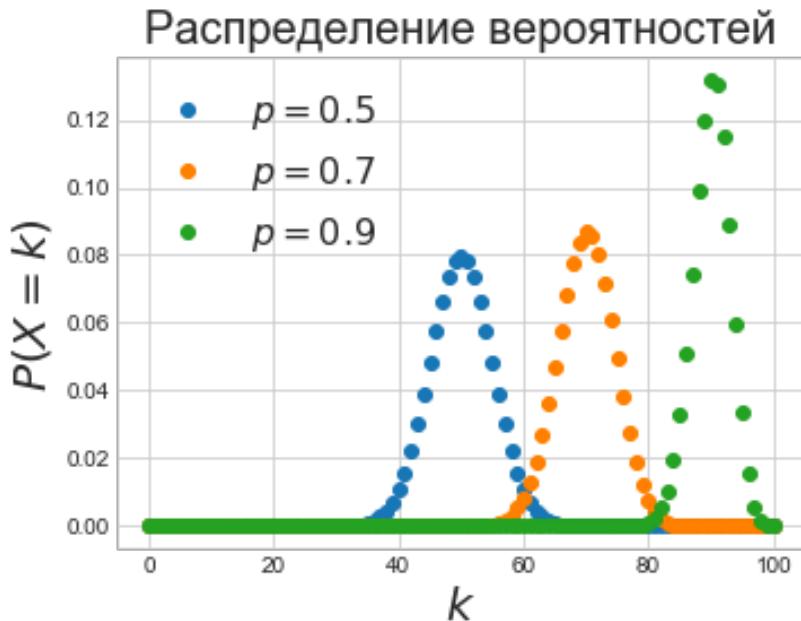
$$\mathbb{E}(X) = n \cdot p$$

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

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

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

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



Геометрическое распределение

- Номер броска, когда произошло первое попадание в корзину

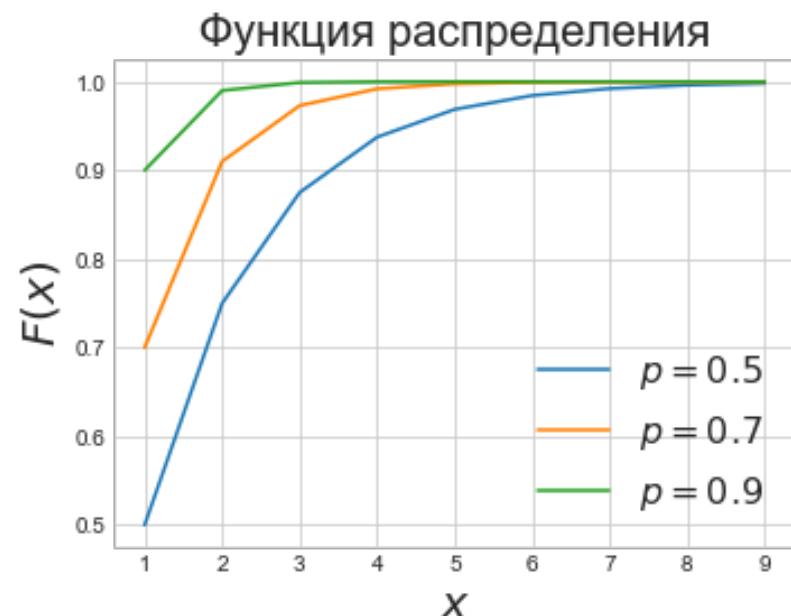
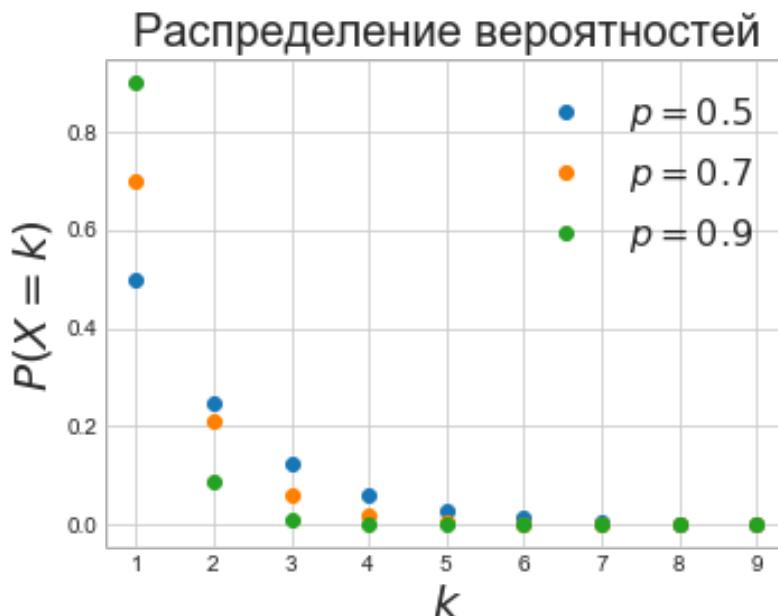
$$\mathbb{E}(X) = \frac{1}{p}$$

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

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

p – вероятность успеха

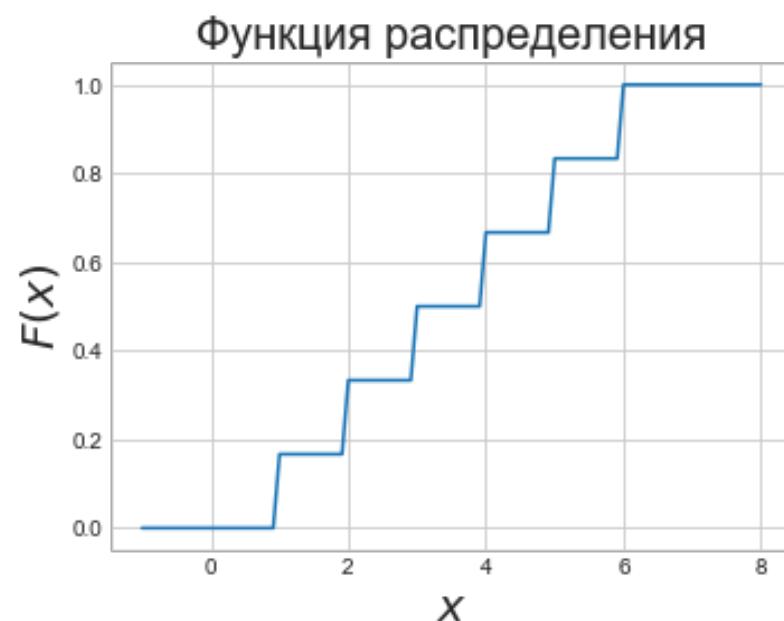
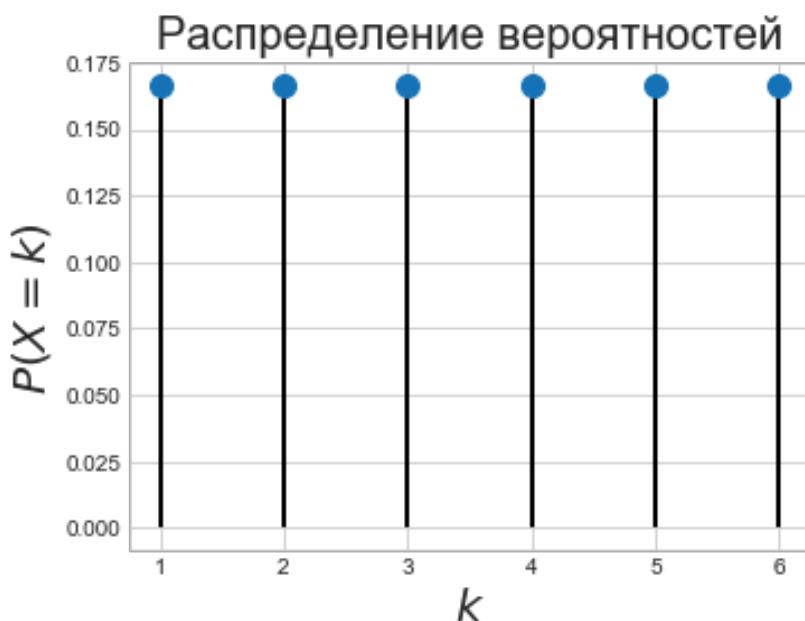
k принимает значения $1, 2, 3, \dots$ $\mathbb{P}(X = k) = p \cdot (1 - p)^{k-1}$



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

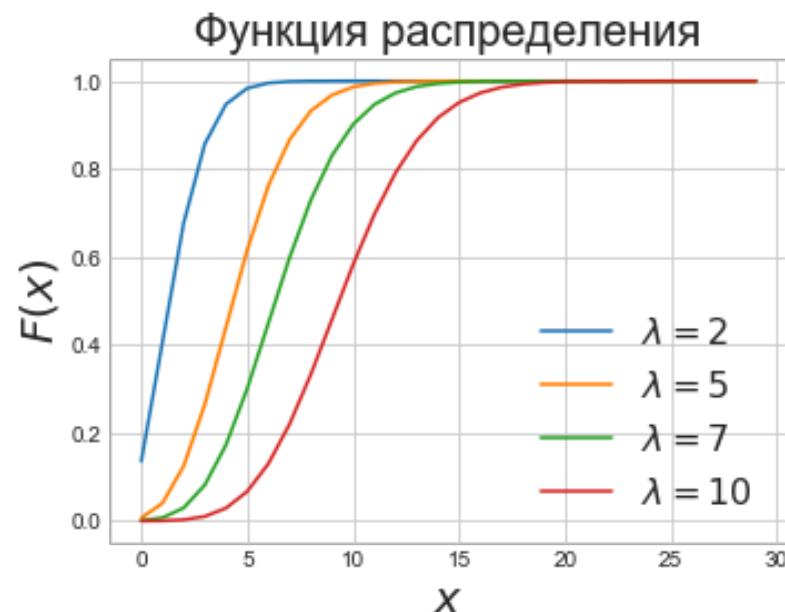
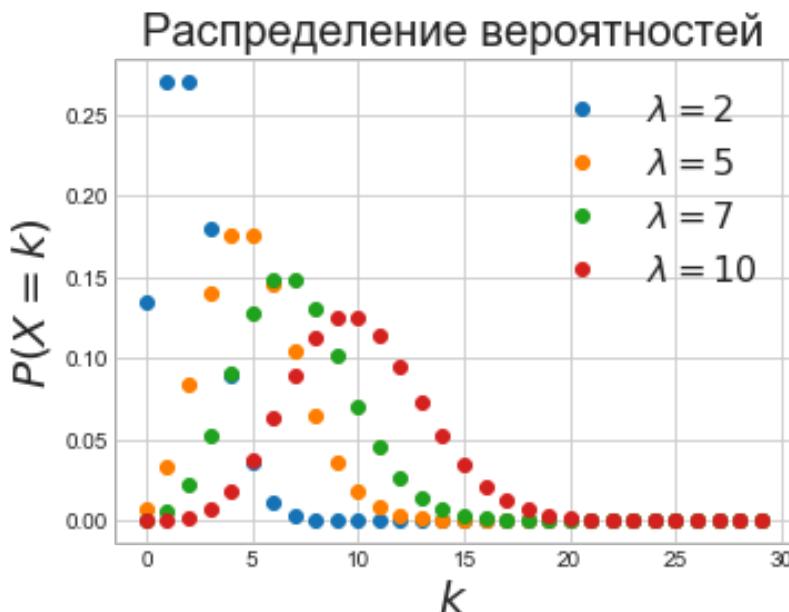
Распределение Пуассона

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

- 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:
 $X \sim Exp(\lambda)$

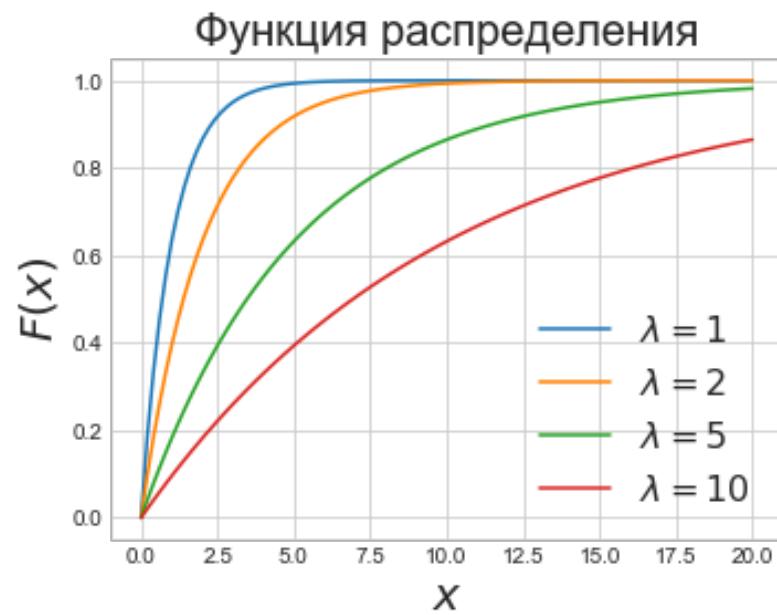
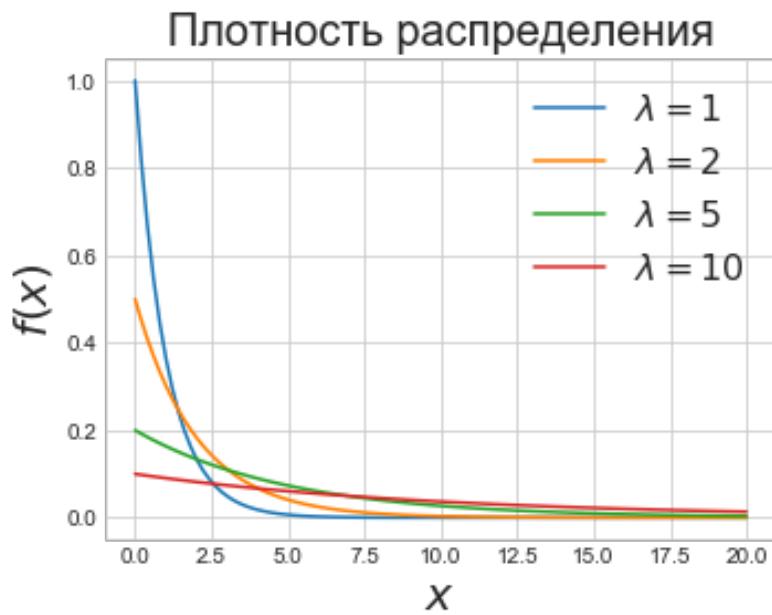
- Интервалы времени между событиями
- Модели времени жизни



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

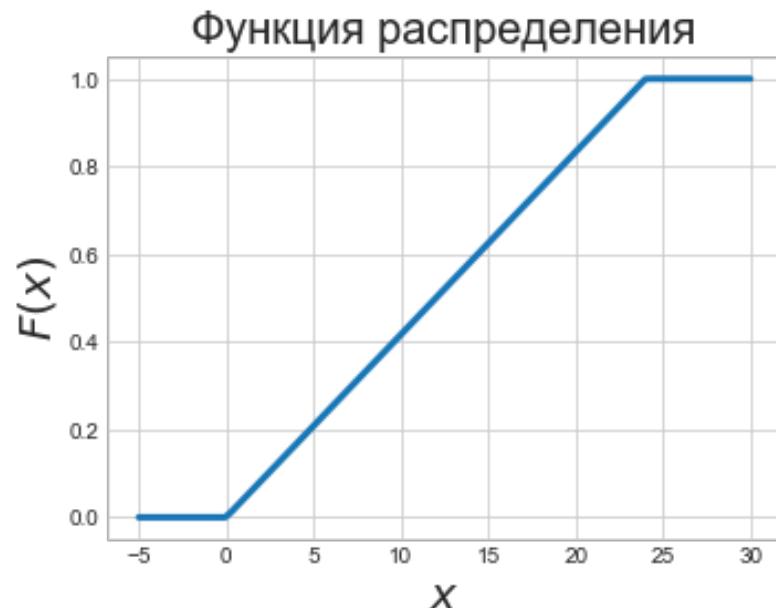
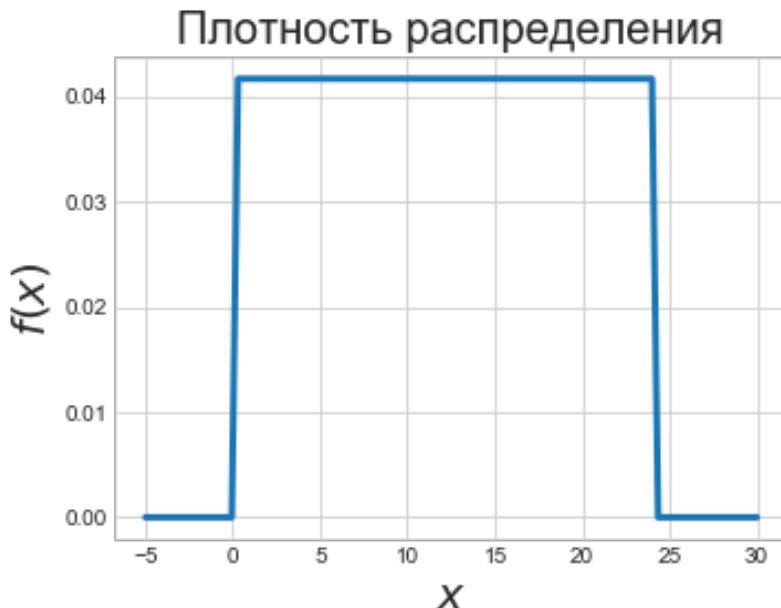
Равномерная случайная величина: $X \sim U[a; b]$

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

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

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

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



Нормальное распределение

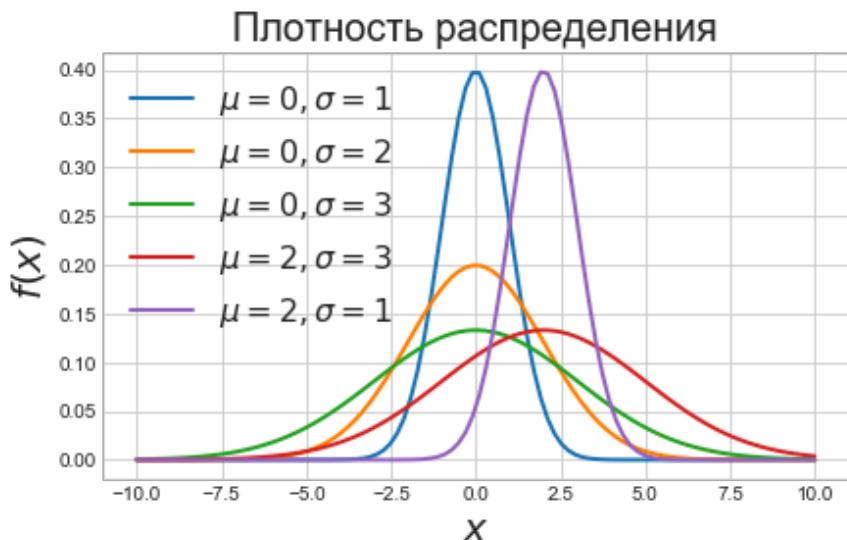
- Погрешность весов

Нормальная случайная величина:

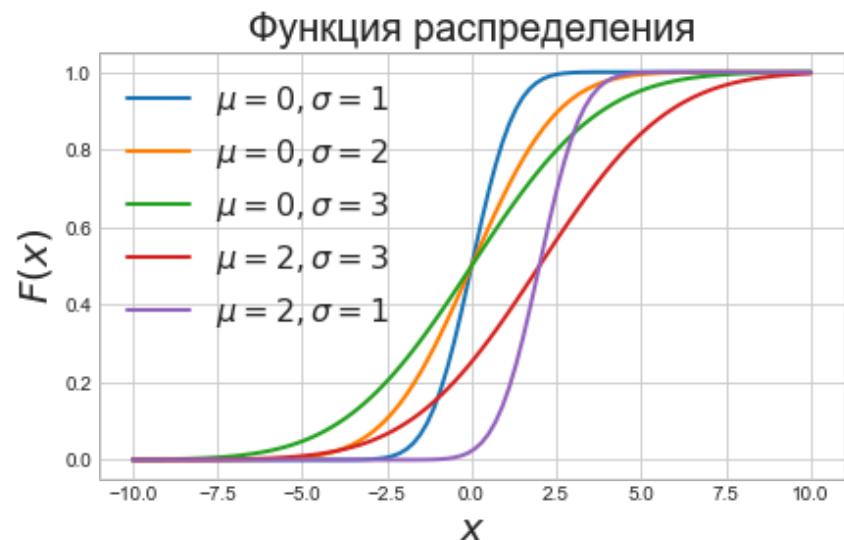
$$X \sim N(\mu, \sigma^2)$$

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

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