ТВиМС 1

Байрамкулов Аслан

Руководитель группы моделирования профиля
МТС Big Data

План:

- Частотный и байесовский взгляд на статистику
- Случайные величины, их основные свойства. Функция и плотность распределения
- Независимые случайные величины
- Условные вероятности
- Формула полной вероятности. Формула Байеса
- Примеры разных распределений
- Гистограммы

Разные взгляды на статистику

Взгляд на статистику

Томас Байес

Рональд Фишер

Байесовский взгляд

- Лаплас развил байесовские идеи. Сторонник детерминизма.
- Точное предсказание вселенной в случае возможности измерения положения каждого атома. (Но издержки огромны)
- Возникающая неопределенность – результат огромного разрыва между совершенством природы и несовершенством человеческого познания.
- Таким образом, случайность – следствие нашей ограниченности
- Вероятность – способ измерения случайности (субъективно)

Частотный взгляд

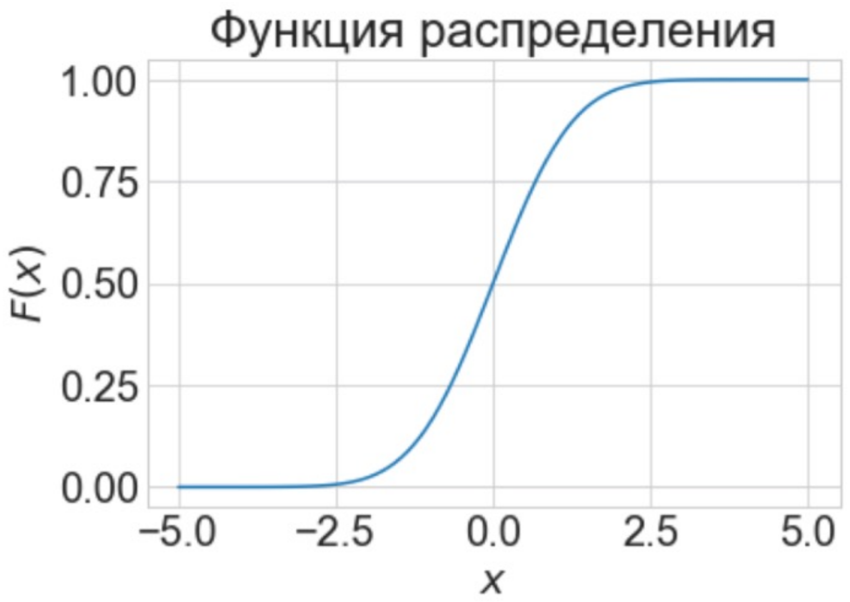
- Вероятность не субъективна. Она должна быть объективной мерой оценки возникновения события.
- Оцениваем вероятность только повторяющихся событий (которые происходят > 1 раза)

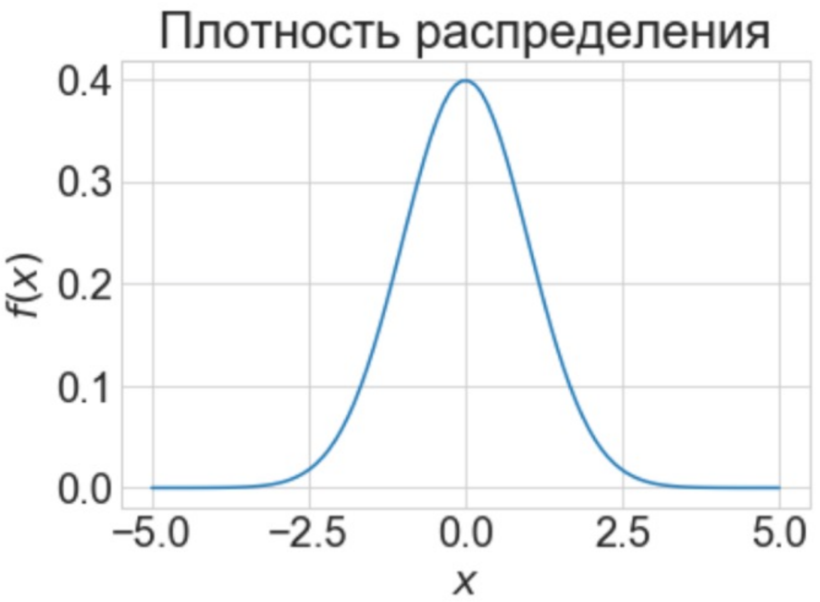
ТВ vs МС

- Существуют различные процессы порождения данных. Механизмы порождения изучаются **теорией вероятностей** (**прямой ход**)

- Получающиеся данные – объект исследования **математической статистики**. Она пытается изучить процесс порождения на основе экспериментальных выборок (**обратный ход**)

Механизм порождения





- Модель – наше предположение о том, как устроен процесс порождения данных. Каждая модель подкреплена предпосылками, описывающими наше незнание.

Детализация МС

- Эксперимент порождает данные на основе неизвестного механизма
- На основе экспериментальных данных мы пытаемся восстановить структуру неизвестного механизма
- Восстановление происходит в рамках выбранной нами модели

- Изучение данных и их свойств
- Формализация своих гипотез и предположений в виде моделей
- Состыковка наших предположений и имеющихся данных

Случайные величины

Случайная величина

- **Случайная величина** X – произвольная измеримая функция, заданная на пространстве элементарных событий **Ω** и принимающая значения в R
- Это означает, что каждому элементарному событию w мы будем ставить в соответствие некоторое число X(w).

Примеры случайных величин:

- Число солнечных дней N в году
- Число выпавших очков Q при бросании игральной кости
- Время пробуждения
- Величина конверсии P интернет-магазина

Случайная величина

Случайные величины

Дискретные:

- множество значений конечно или счётно
- примеры: число звонков, ошибок в тексте…

Непрерывные:

- Бесконечное число значений
- Примеры: рост, вес, время ожидания автобуса

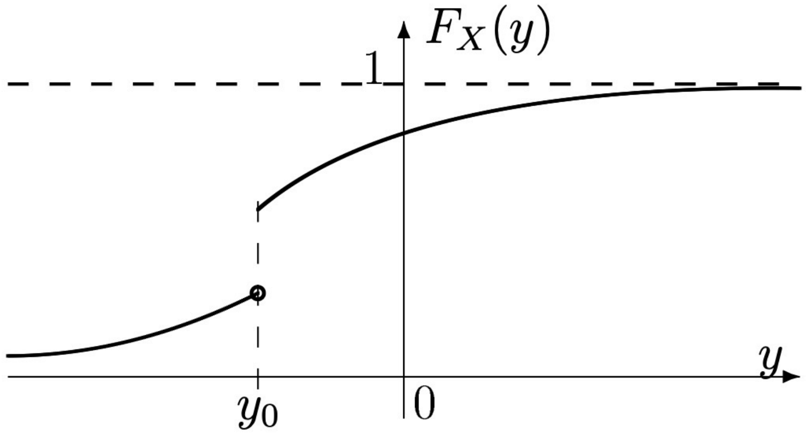
Функция распределения

- Функцией распределения случайной величины X называется:

$$F_X(y) = \mathbf{P}(\omega : \; X(\omega) < y) = \mathbf{P}(X < y), \quad -\infty < y < \infty.$$

Основные свойства:

- $0 \leq F!y \leq 1$ для $\forall y$
- Функция распределения монотонно не убывает
- Существуют пределы $\lim F!\ y = 0$ и $\lim F!\ y = 1$
- $\forall y\ F!\ y - 0 = F!\ y$, - функция распределения непрерывна слева



Функция распределения

- Для любых чисел a < b получим вероятность P попадания в полуинтервал:

$$\mathbf{P}(a \leq X < b) = \mathbf{P}(X < b) - \mathbf{P}(X < a) = F_X(b) - F_X(a)$$

Виды функций распределений:

- Дискретные
- Непрерывные
- Сингулярные*
- Смешанные*

Условные вероятности. Формула Байеса

Независимые случайные величины

Определение. Случайные величины $X_1, X_2, \ldots, X_n$ называются *независимыми*, если для любых борелевских множеств $B_1 \in \mathcal{B}(\mathbb{R}), \ldots, B_n \in \mathcal{B}(\mathbb{R})$ выполняется соотношение

$$\mathbf{P}(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) = \mathbf{P}(X_1 \in B_1)\mathbf{P}(X_2 \in B_2)\ldots\mathbf{P}(X_n \in B_n) \qquad (1)$$

Из этого определения вытекает, к примеру, попарная независимость случайных величин: если положить $B_3 = B_4 = \ldots = B_n = \mathbb{R}$, то будем иметь

$$\mathbf{P}(X_1 \in B_1, X_2 \in B_2) = \mathbf{P}(X_1 \in B_1)\mathbf{P}(X_2 \in B_2).$$

Условные вероятности

- **Условной вероятностью** (или вероятностью события A при условии, что произошло событие B) называется:

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(AB)}{\mathbf{P}(B)}.$$

$$A = \{6\}, \quad B = \{2, 4, 6\}, \quad AB = \{6\},$$

$$\mathbf{P}(A|B) = \frac{1/6}{1/2} = \frac{1}{3},$$

Формула полной вероятности

Пусть нас интересует вероятность некоторого события A и предположим, что наряду с A есть некий набор вспомогательных событий H1, ..., Hn, которые принято называть гипотезами и которые удовлетворяют следующим двум требованиям:

$$1)\ H_i H_j = \varnothing \quad (i \neq j);$$

$$2)\ A \subset \bigcup_{i=1}^{n} H_i.$$

Тогда справедлива формула полной вероятности:

$$\mathbf{P}(A) = \sum_{i=1}^{n} \mathbf{P}(A|H_i)\mathbf{P}(H_i).$$

Формула Байеса

Вероятности P(H1), P(H2), P(H3), . . . , P(Hn) называются **априорными**. После получения дополнительной информации в ходе проведения случайного эксперимента вероятности меняются! То есть требуется вычисление новых вероятностей P(Hi|A), при условии того, что произошло некоторое событие A. Вероятности P(Hi|A) называют **апостериорными**, то есть полученными в результате опыта

Если $H_1, H_2, H_3, \ldots, H_n$ полная система событий, а вероятность события A не равна нулю, то:

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \ldots + P(A|H_n)P(H_n)}$$

Формула Байеса

В салоне связи было проведено исследование продаж розовых телефонов. Выяснилось, что посетители женщины этот телефон покупают в 55% случаях, мужчины – в 5% случаях и дети – в 15% случаях. Среди посетителей салона 50% женщин, 40% мужчин и 10% детей. Найти вероятность того, что случайный покупатель приобретет этот товар.

Решение. *Рассмотрим события* A={*куплен розовый телефон*}, H_1={*посетителем была женщина*}, H_2={*посетителем был мужчина*} *и* H_3={*посетителем был ребенок*}. *По условию даны вероятности* $P(H_1) = 0.5$, $P(H_2) = 0.4$, $P(H_3) = 0.1$, $P(A|H_1) = 0.55$, $P(A|H_2) = 0.05$, $P(A|H_3) = 0.15$. *По формуле полной вероятности находим*

$$P(A) = P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + P(A|H_3)P(H_3) =$$
$$0.55 \cdot 0.5 + 0.55 \cdot 0.5 + 0.4 \cdot 0.05 + 0.1 \cdot 0.15 = 0.31.$$

Сводка

- Моделировать внутренности механизма порождения данных можно с помощью различных законов распределения
- Наиболее подходящий закон выбирается с помощью здравого смысла
- Все предпосылки, связанные с выбранным законом, должны проверяться по данным

Распределения

Дискретные распределения

- Случайная величина X называется **дискретной**, если существует конечная или счетная последовательность чисел yk такая, что:

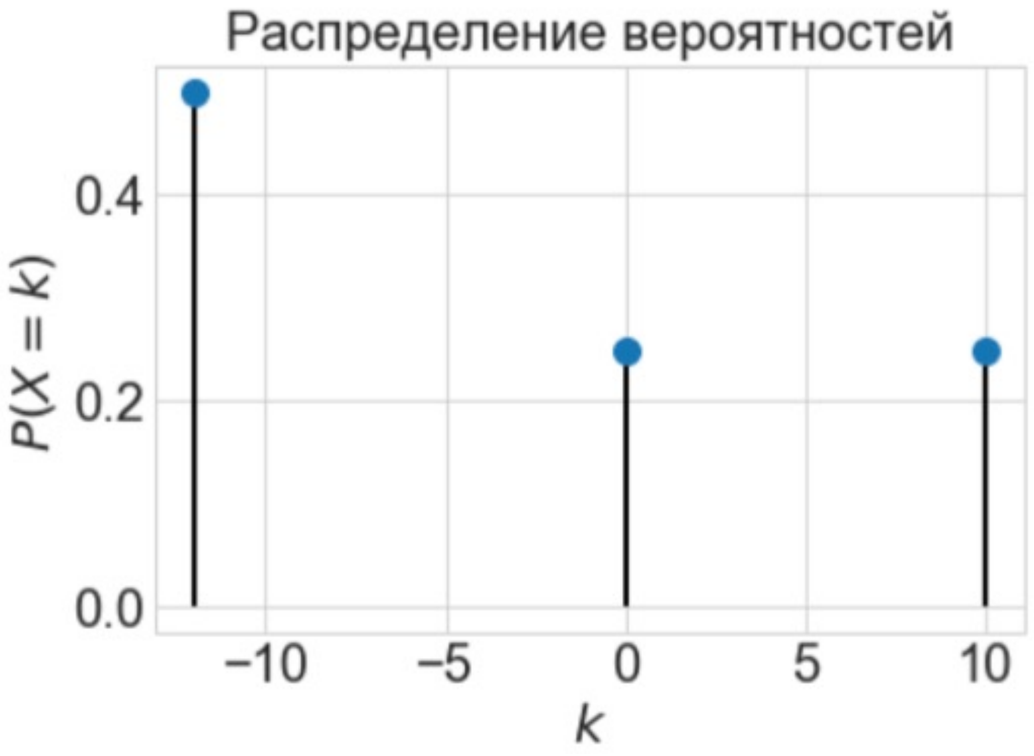
$$\sum_{k=1}^{\infty} \mathbf{P}(X = y_k) = 1.$$

- Дискретную величину можно охарактеризовать таблицей, обозначив вероятность каждого конкретного значения yk:

$$p_k = \mathbf{P}(X = y_k),\ k = 1,\, 2,\, \ldots,$$

Значения	y_1	y_2	y_3	$\ldots$
Вероятности	p_1	p_2	p_3	$\ldots$

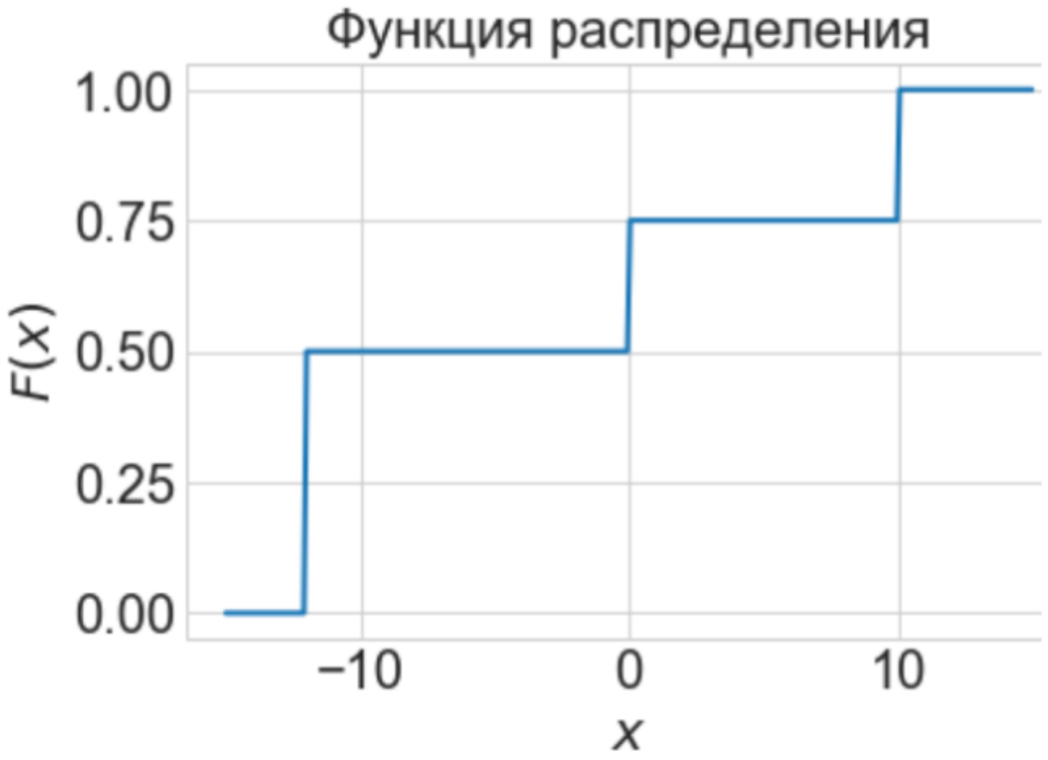
Дискретные распределения



Пример: лотерея

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

Дискретные распределения



Пример: лотерея

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

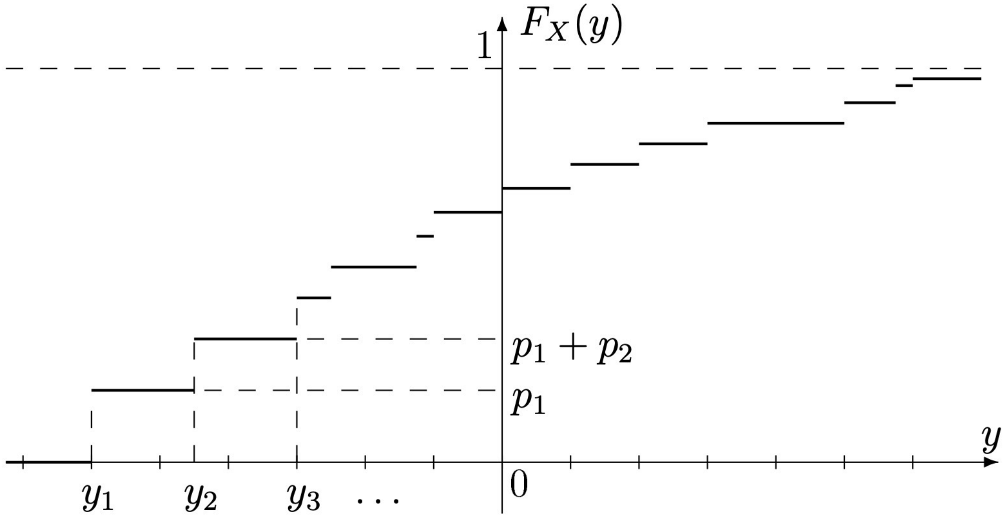
Дискретные распределения

- Вероятность попадания значений случайной величины в некоторый интервал можно легко найти суммированием элементов таблицы:

Значения	y_1	y_2	y_3	$\ldots$
Вероятности	p_1	p_2	p_3	$\ldots$

$$\mathbf{P}(a < X < b) = \sum_{k:\ a<y_k<b} p_k.$$

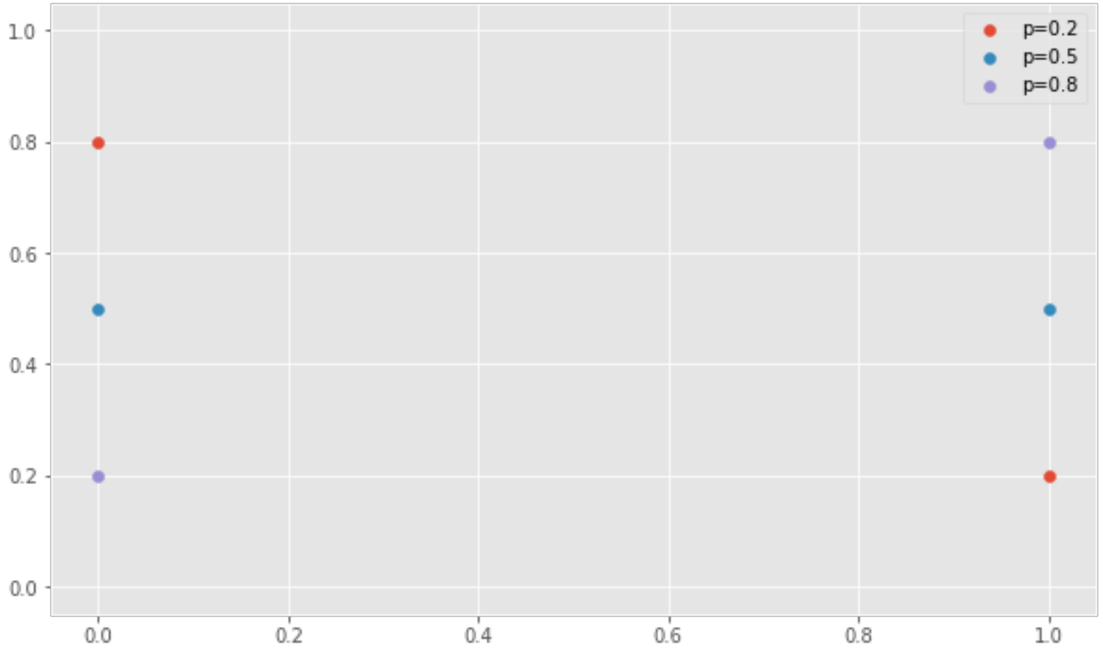
- Функция распределения для упорядоченной по возрастанию некоторой дискретной случайной величины yk будет выглядеть ступенчато:



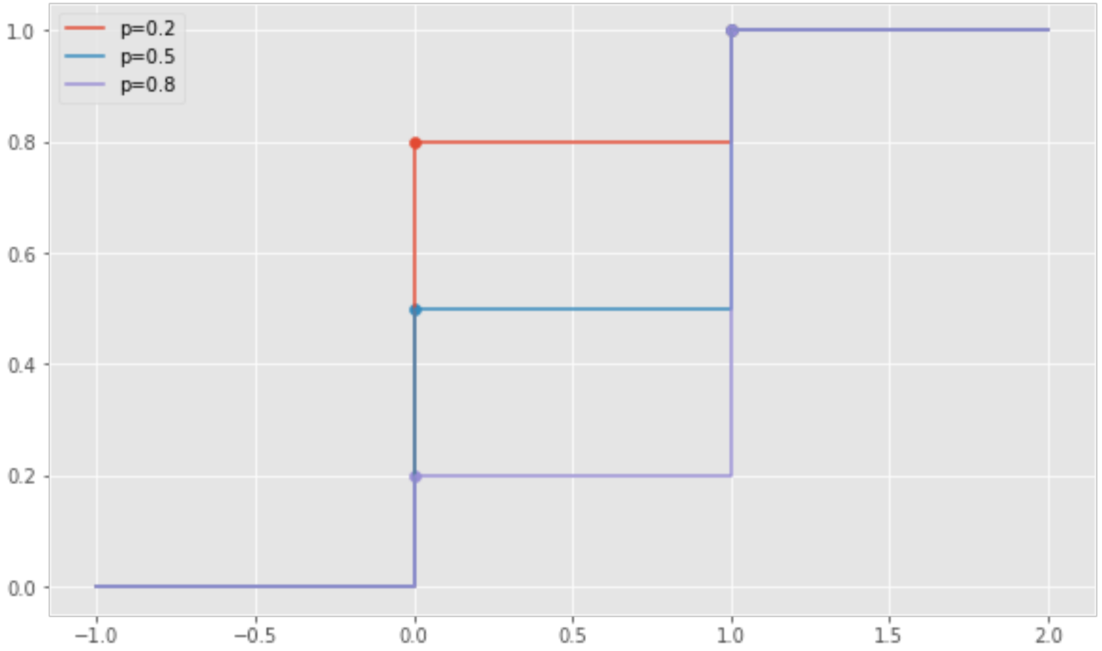
Распределение Бернулли

2. *Распределение Бернулли* B_p: $X \in B_p$, если $\mathbf{P}(X = 1) = p$, $\mathbf{P}(X = 0) = 1 - p$, $0 < p < 1$.

Распределение Бернулли. Функция вероятности



Распределение Бернулли. Функция распределения



Распределение Бернулли

- Пол родившегося ребёнка

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

Распределение Бернулли:

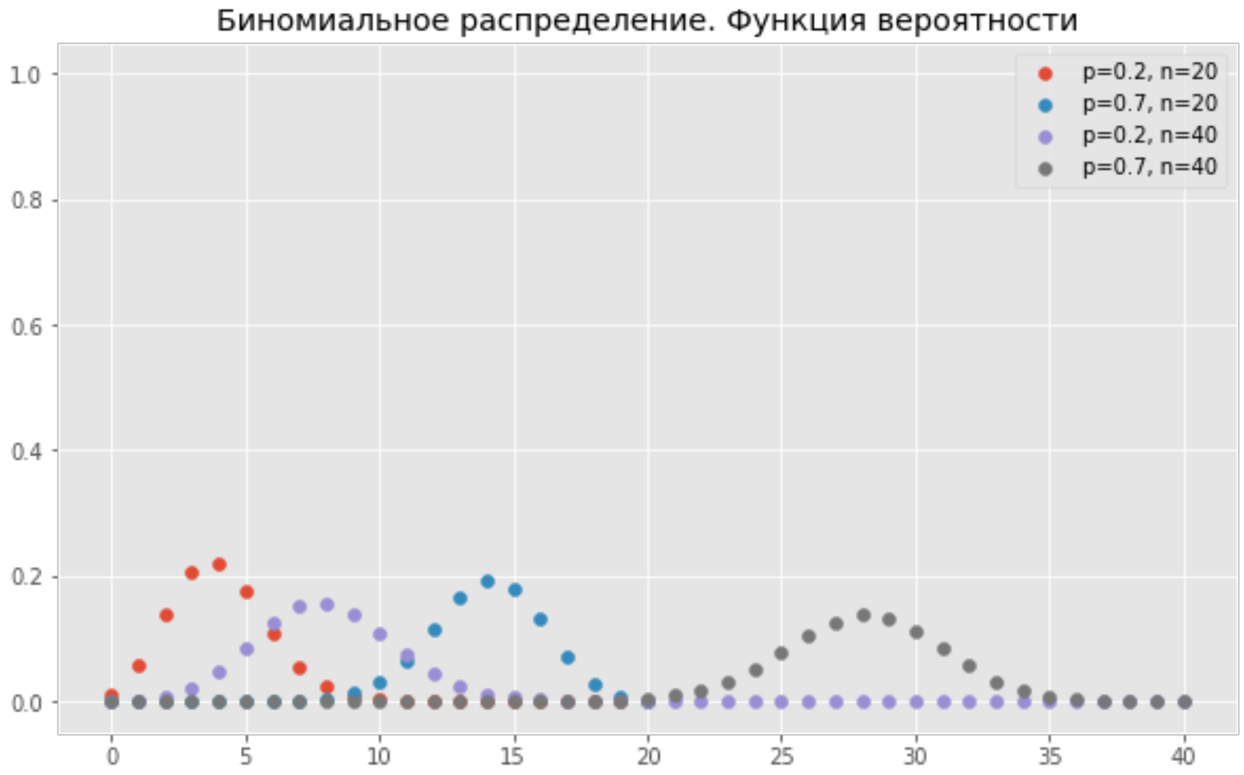
$$X \sim Bern(p)$$

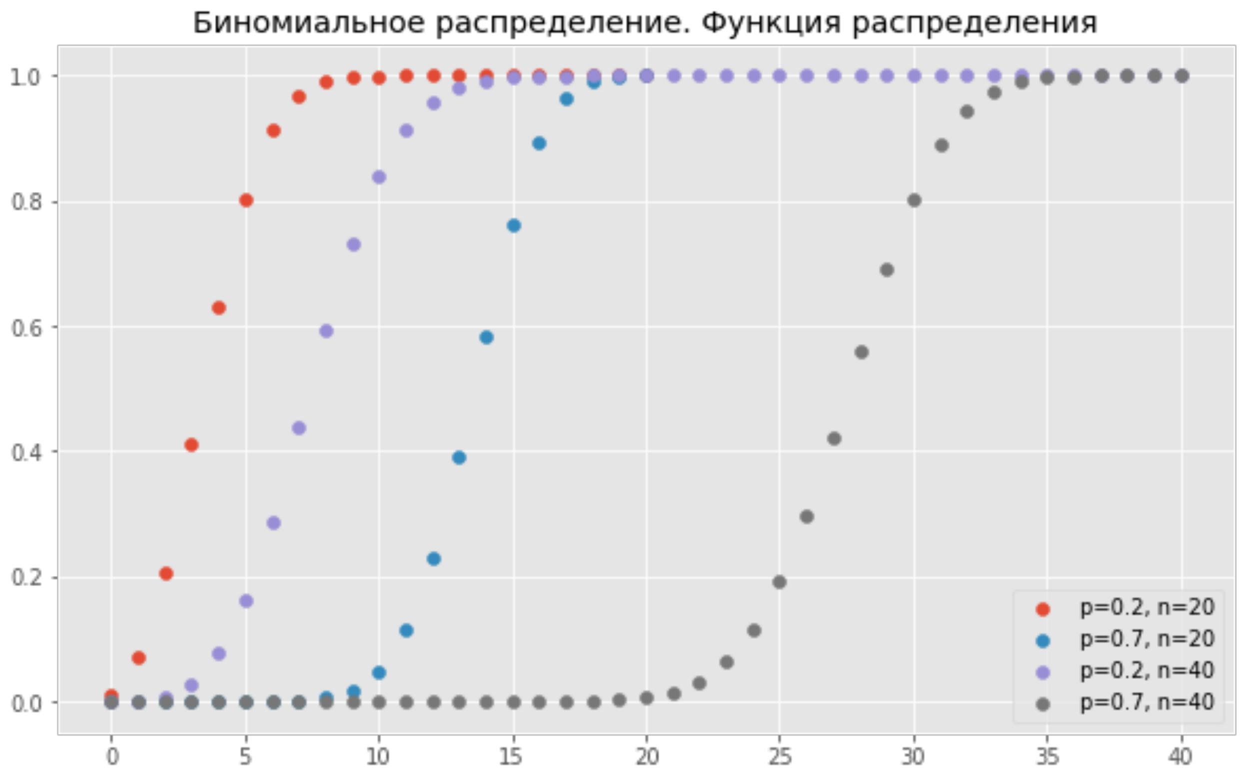
$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot (1 - p) = p$$

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

Биномиальное распределение

3. *Биномиальное распределение* $B_{n,p}$: $X \in B_{n,p}$, если $\mathbf{P}(X = k) = C_n^k p^k (1-p)^{n-k}$, $k = 0, 1, \ldots, n$ (в частности, $B_{1,p} = B_p$).





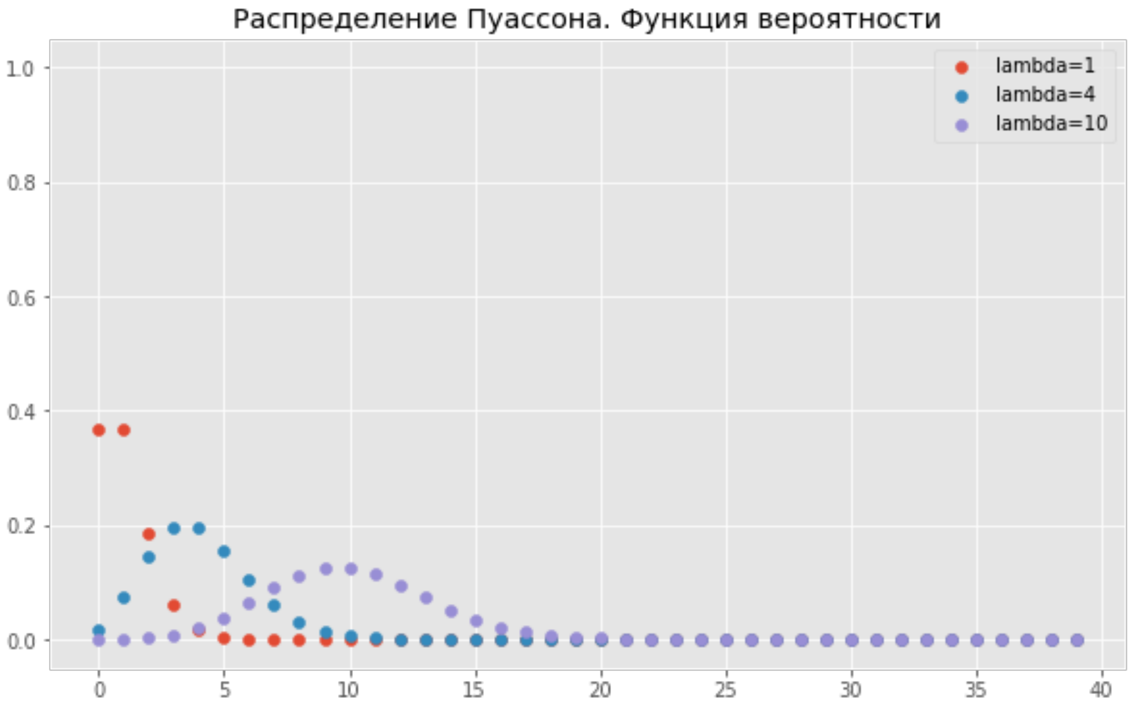
Распределение Пуассона

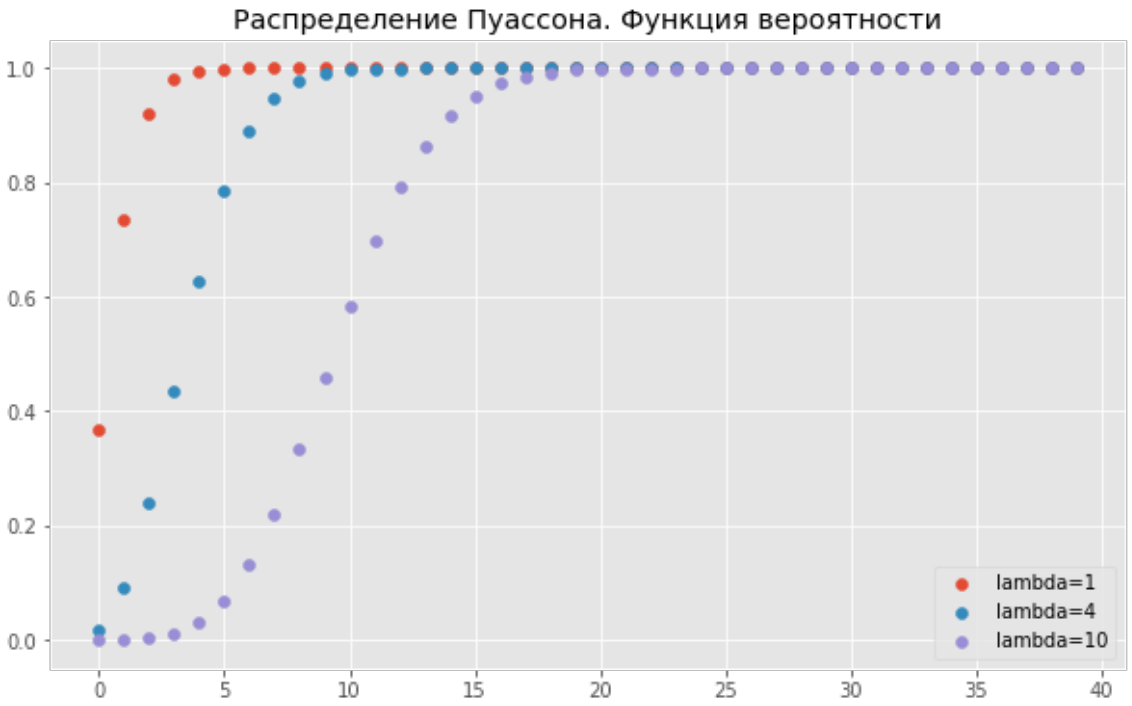
- Число людей в очереди
- Число лайков под фото
- Число автобусов, проехавших за час мимо остановки

Распределение Пуассона хорошо описывает счётчики

Распределение Пуассона

4. *Распределение Пуассона* Π_λ: $X \in \Pi_\lambda$, если $\mathbf{P}(X = k) = \dfrac{\lambda^k}{k!} e^{-\lambda}$, $k = 0, 1, 2, \ldots;$ $\lambda > 0$.





Непрерывные распределения

- Функция распределения случайной величины FX(y) называется абсолютно непрерывной, если для любого значения y:

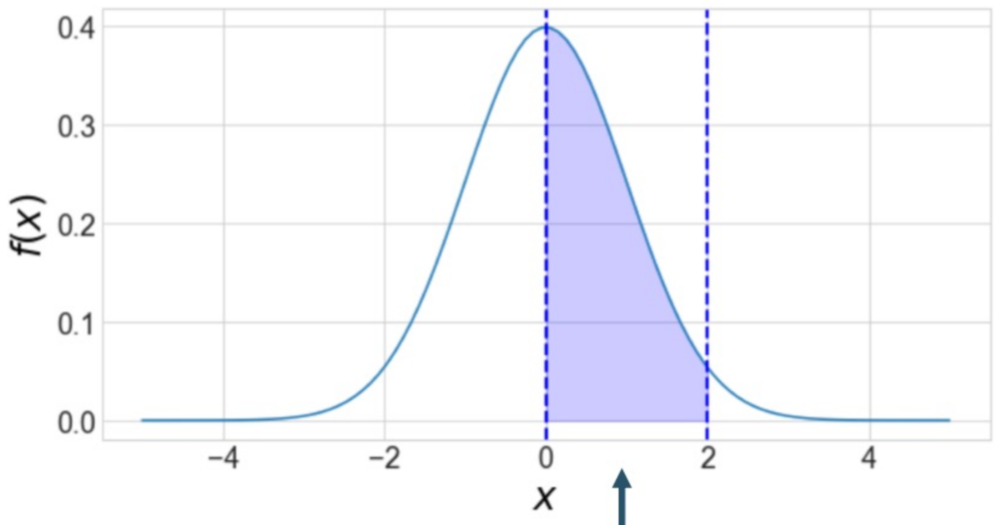
$$F_X(y) = \int\limits_{-\infty}^{y} f(t)\,dt;$$

- Стоящая под знаком интеграла функция f(t) называется плотностью распределения. Для всех точек, где производная функции распределения существует (а она существует почти везде), можем выразить плотность, как:

$$f_X(t) = \frac{dF_X(t)}{dt}$$

Плотность распределения

- Распределение непрерывной случайно величины описывается плотностью распределения вероятностей.



Пример: нормальное распределение

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

Площадь равна вероятности попасть на отрезок от нуля до двух

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

- Площадь под всей плотностью должны быть равна 1.

Плотность распределения

1. Плотность определена только для непрерывных случайных величин

2. $f(x) = F'(x)$

3. $\int_{-\infty}^{+\infty} f(t)\ dt = 1, \quad f(t) \geq 0\ \ \forall\, t$

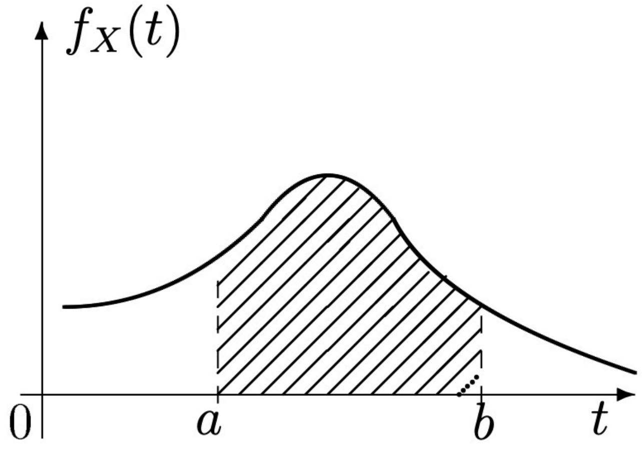
4. $F(x)$ не убывает, лежит между 0 и 1

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

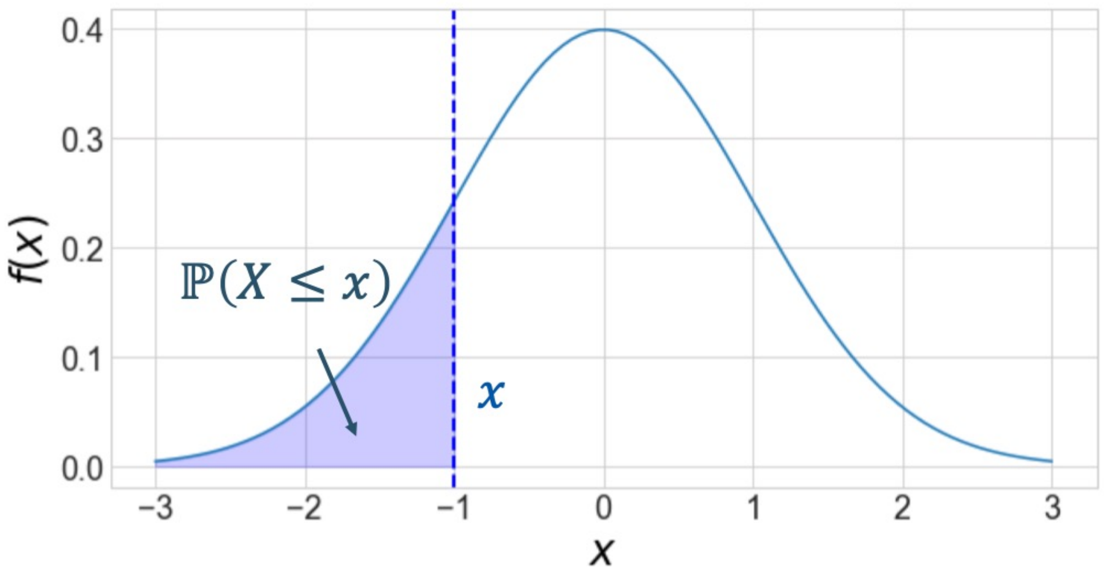
$$f_X(t) = \frac{dF_X(t)}{dt}$$

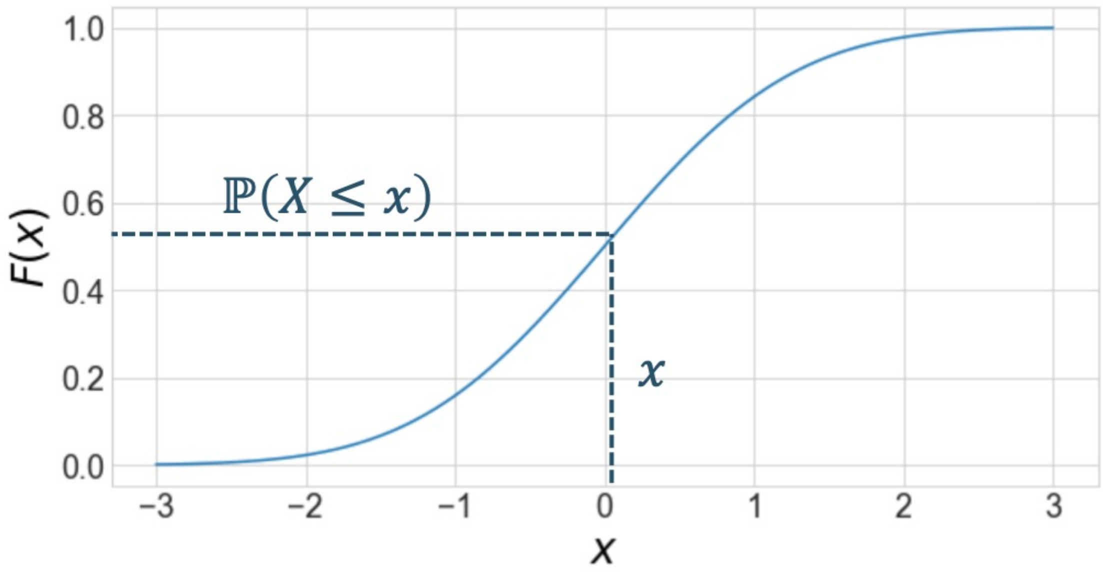
$$F_X(y) = \int\limits_{-\infty}^{y} f(t)\,dt;$$



Плотность распределения

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



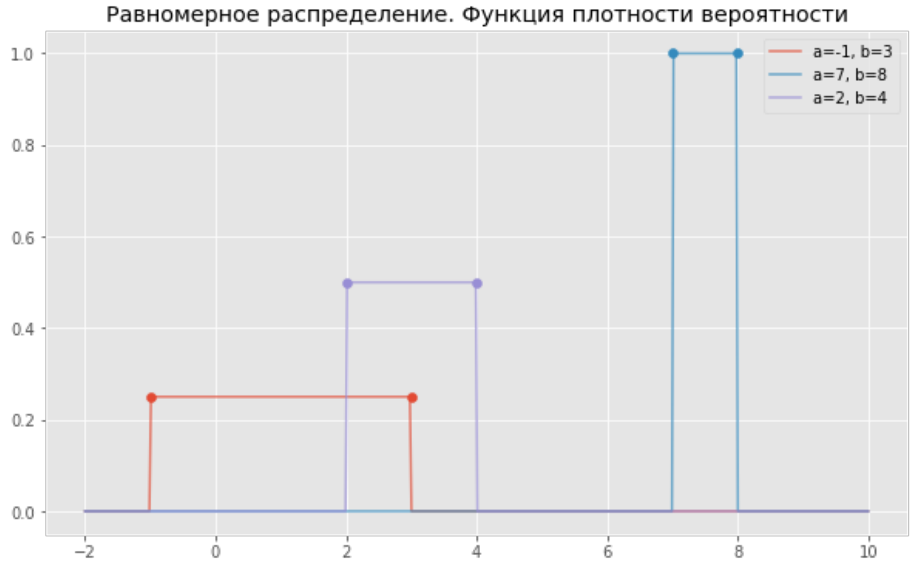


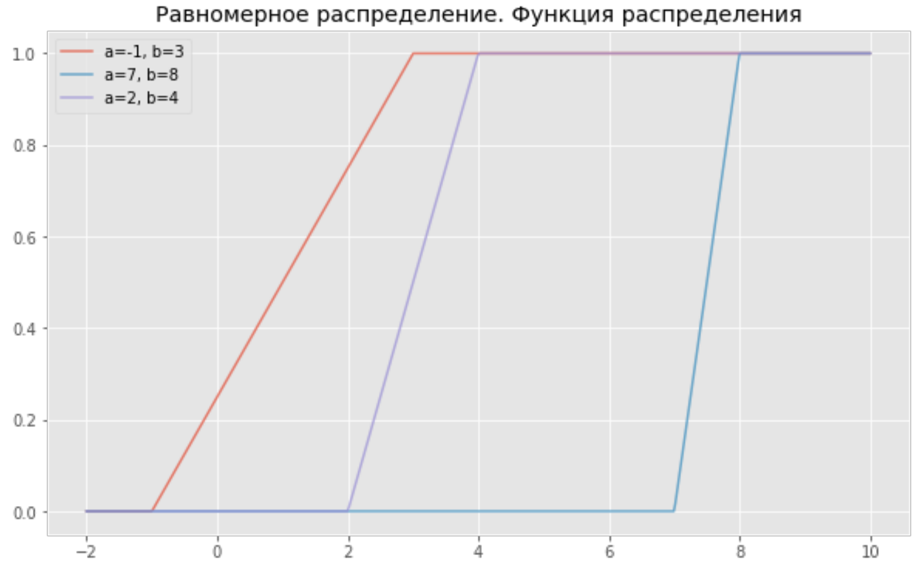
Равномерное распределение

Равномерное распределение на отрезке $[a,\, b]$.

$$u_{a,b}(t) = \begin{cases} \dfrac{1}{b-a}, & t \in [a,\, b], \\ 0, & \text{иначе.} \end{cases}$$

$$U_{a,b}(y) = \begin{cases} 0, & y \le a, \\ \dfrac{y-a}{b-a}, & y \in [a,\, b] \\ 1, & y > b. \end{cases}$$





Нормальное распределение

Нормальное (гауссовское) распределение Φ_{α,σ^2}.

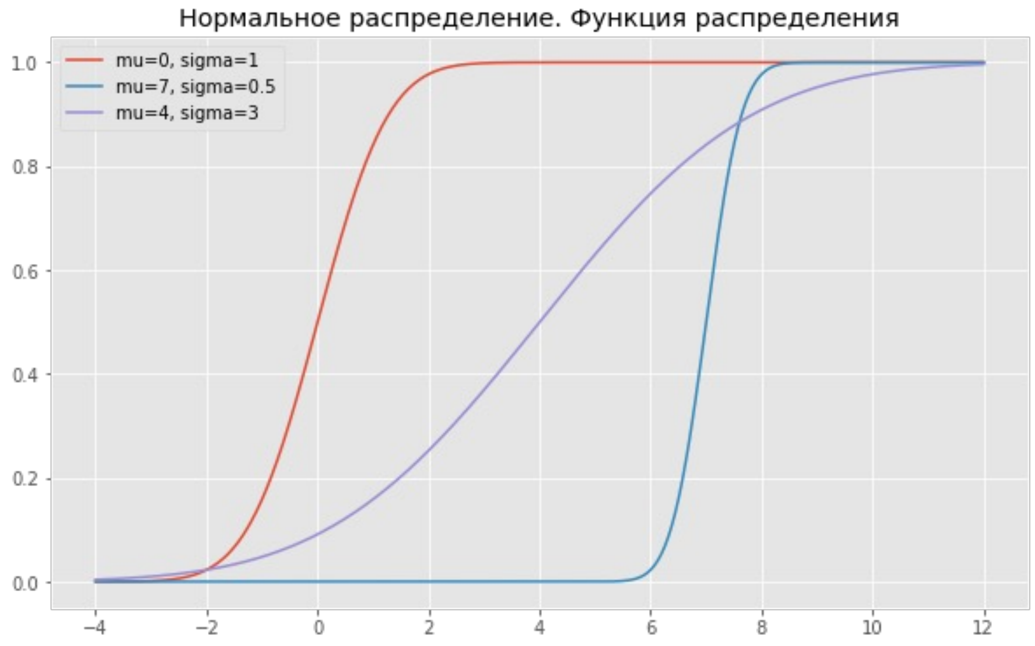
$$\varphi_{\alpha,\sigma^2}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\alpha)^2/2\sigma^2}, \quad -\infty < t < \infty.$$

Параметр α отвечает за сдвиг, а параметр σ2 за размах и максимальное значение функции плотности.

Нормальное распределение

Нормальное (гауссовское) распределение Φ_{α,σ^2}.

$$\Phi_{\alpha,\sigma^2}(y) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{-\infty}^{y} e^{-\frac{(t-\alpha)^2}{2\sigma^2}} \, dt.$$

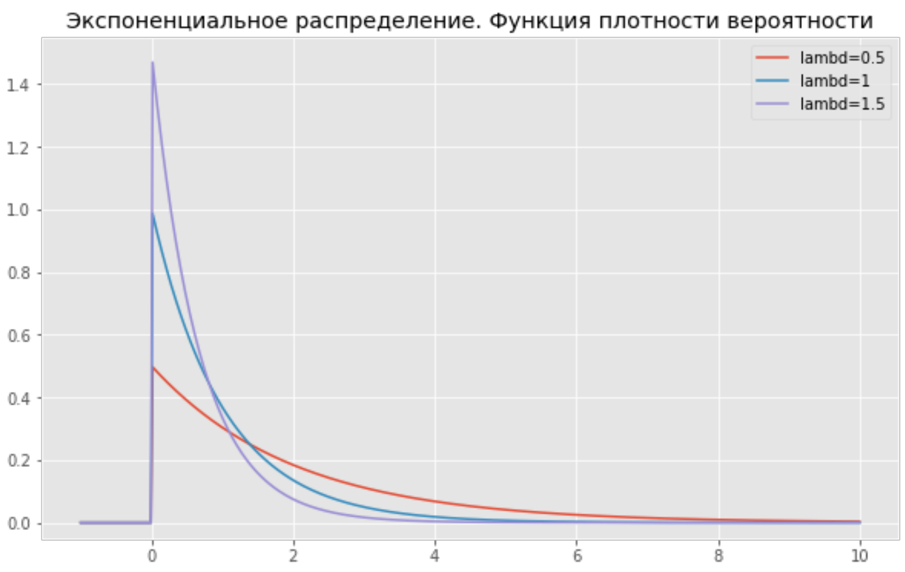


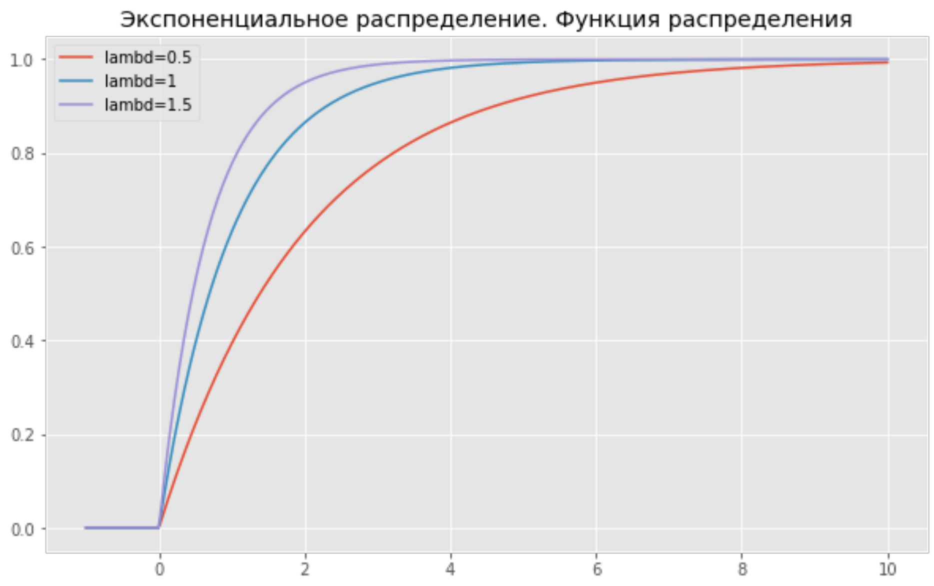
Экспоненциальное распределение

3. *Показательное (экспоненциальное) распределение* E_α.

$$e_\alpha(t) = \begin{cases} \alpha\, e^{-\alpha t}, & t > 0, \\ 0, & t \le 0. \end{cases}$$

$$E_\alpha(y) = \begin{cases} 0, & y \le 0, \\ 1 - e^{-\alpha y}, & y > 0. \end{cases}$$

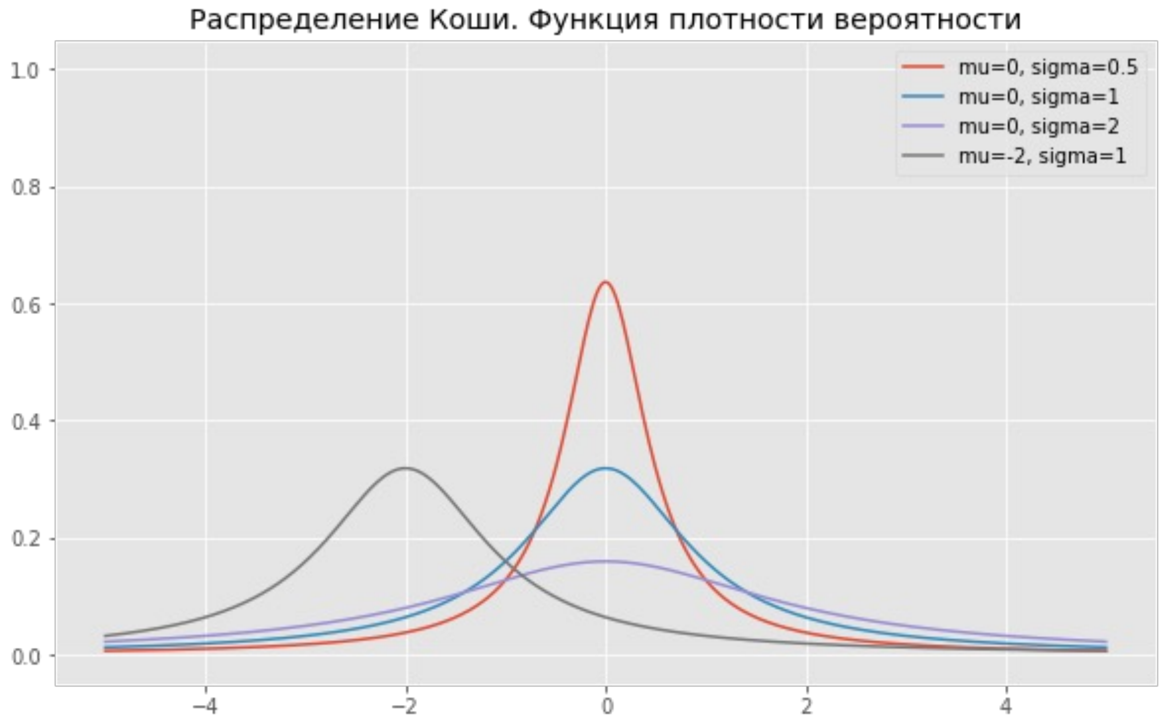


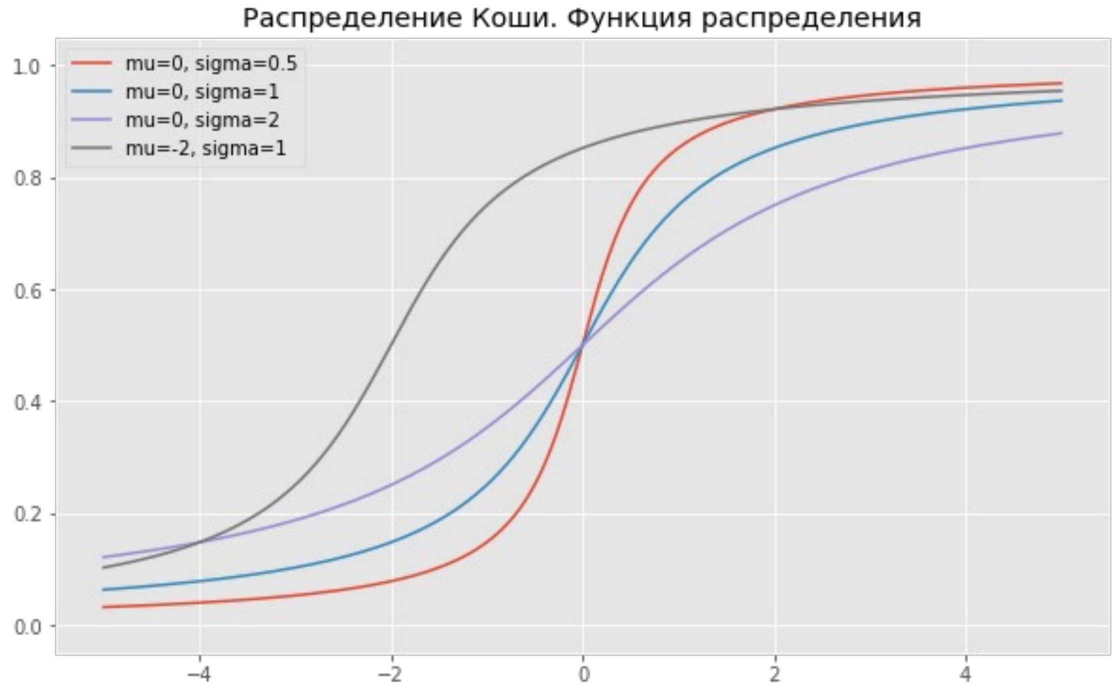


Распределение Коши

$$k(t) = \frac{1}{\pi}\frac{1}{1+t^2}, \qquad -\infty < t < \infty.$$

$$K(y) = \frac{1}{\pi}\int\limits_{-\infty}^{y}\frac{1}{1+t^2}dt = \frac{1}{2} + \frac{1}{\pi}\operatorname{arctg} y.$$





Кейсы распределений

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	$N(0, \sigma^2)$

Гистограмма и эмпирическая функция распределения

Эмпирическая функция распределения

- Функция распределения – функция, которая определяет вероятность события X ≤ x, то есть:

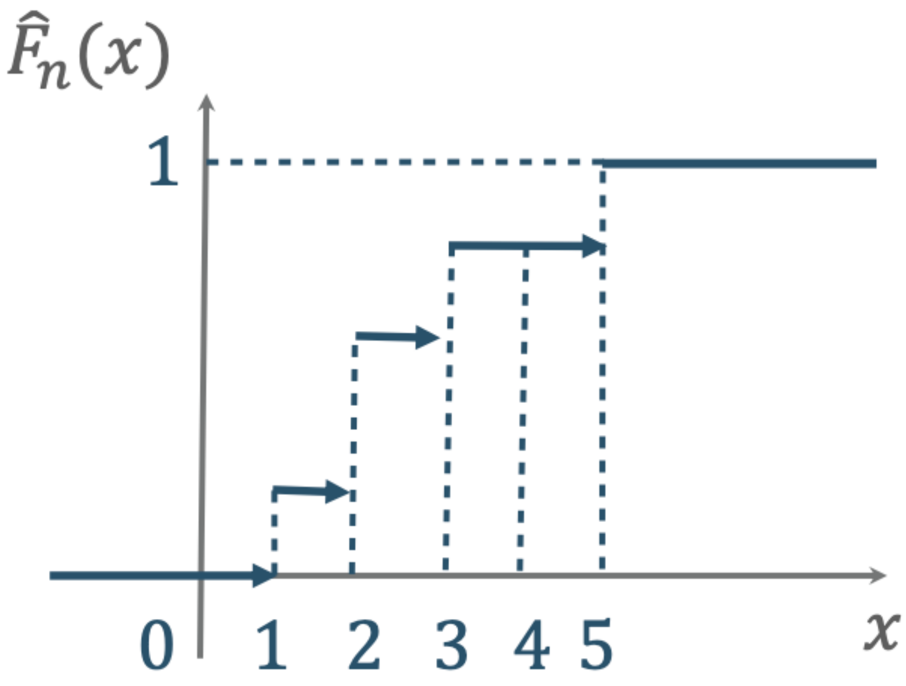
$$F(x) = \mathbb{P}(X \leq x)$$

- Эмпирическая функция распределения – функция, которая определяет для каждого x частоту события X ≤ x то есть:

$$\hat{F}_n(x) = \hat{\mathbb{P}}(X \leq x) = \frac{1}{n}\sum_{i=1}^{n}[X_i \leq x], \qquad [X_i \leq x] = \begin{cases} 1, X_i \leq x \\ 0, \text{иначе} \end{cases}$$

Эмпирическая функция распределения

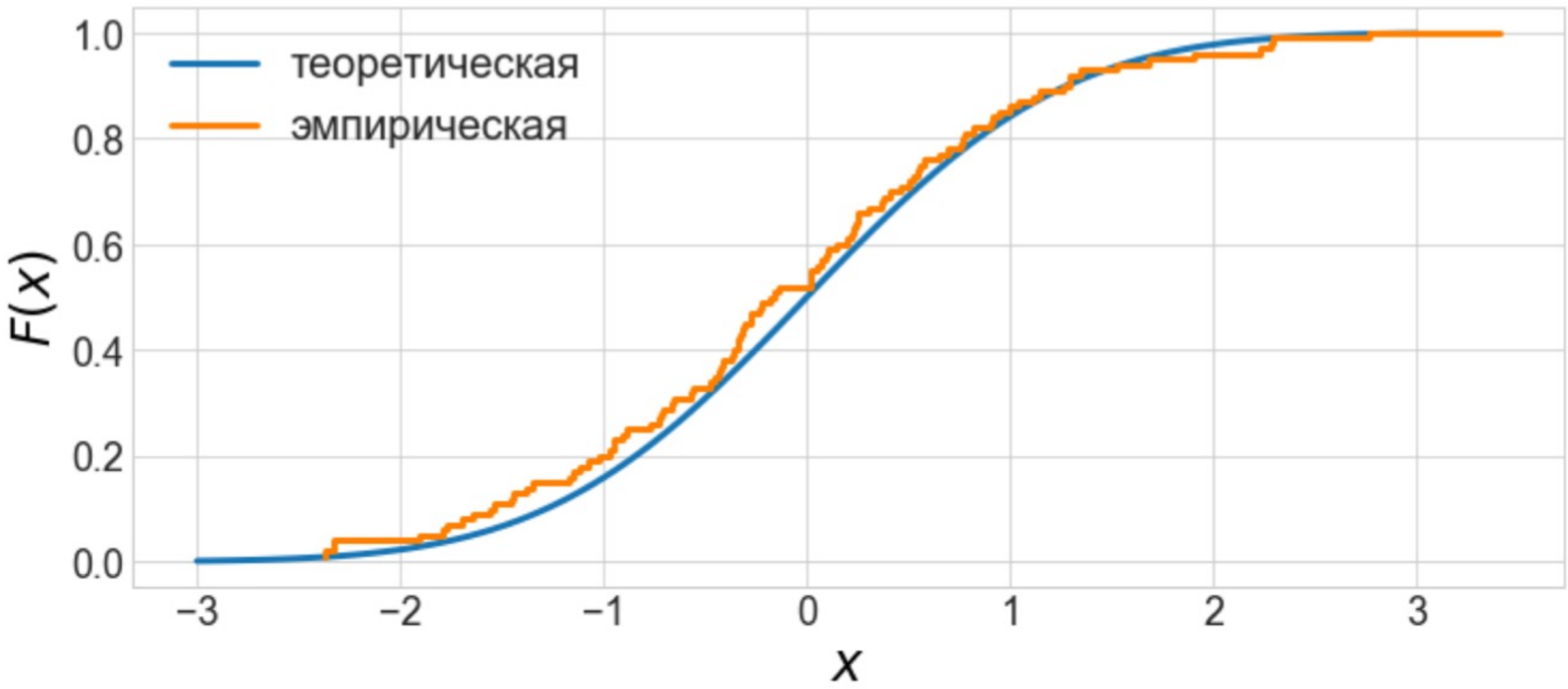
$$x_1 = 2, x_2 = 5, x_3 = 2, x_4 = 3, x_5 = 1$$



- 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.



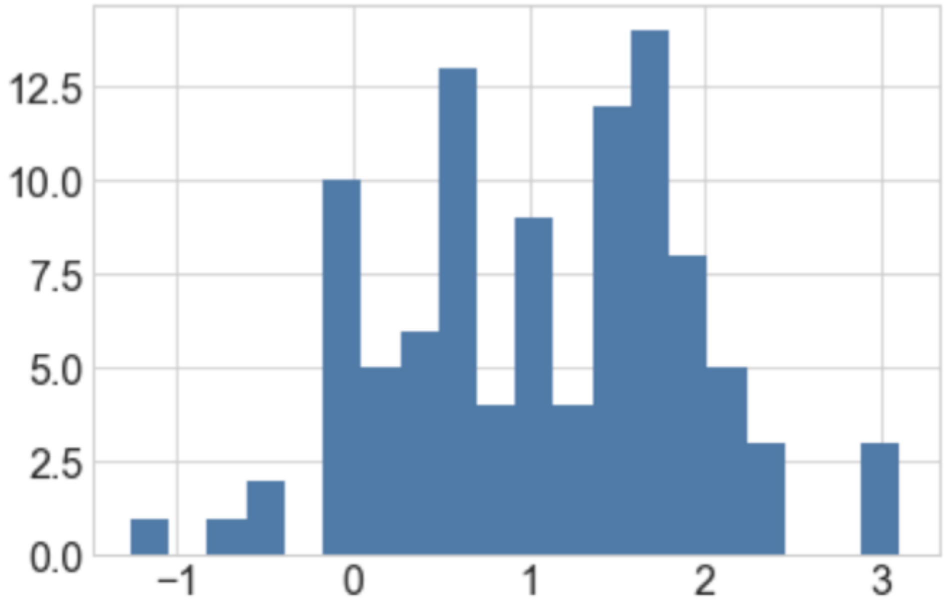
Гистограмма

- 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.

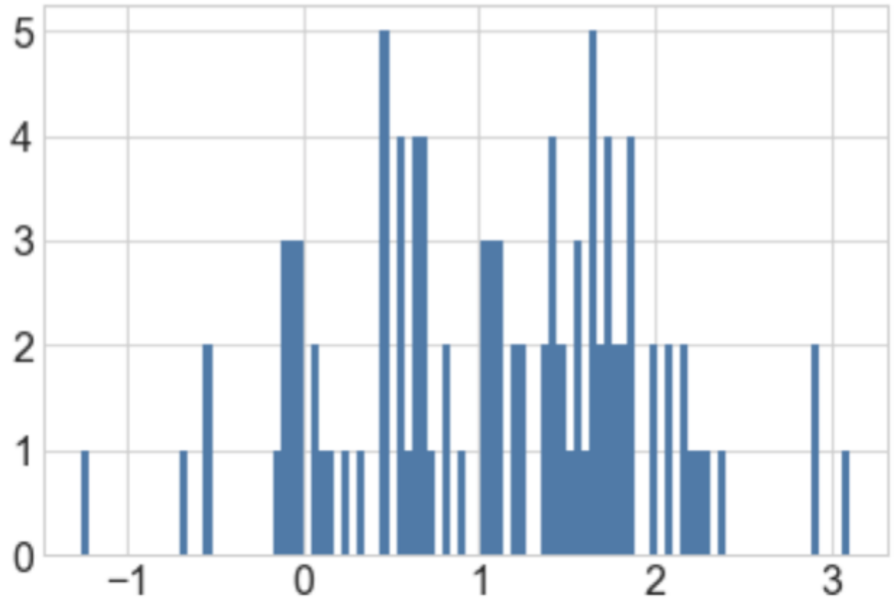
Сколько значений попали в текущий отрезок (бин)

Гистограмма

- Чем короче бины, тем чувствительнее гистограмма к шуму
- Выборка объёма 100 из нормального распределения
- По гистограмме можно попытаться оценить плотность распределения случайной величины



20 бинов



100 бинов

Что мы узнали

- Разобрали матричную терминологию: определитель, ранг, обратная матрица
- Рассмотрели работу со СЛАУ
- Обсудили вопросы обусловленности СЛАУ и матриц
- Определили понятия собственных векторов и значений
- Детально рассмотрели механизм разложения матриц и рассмотрели наиболее актуальное в анализе данных – SVD разложение

Полезные ссылки

- https://github.com/FUlyankin
- http://tvims.nsu.ru/chernova/tv/tv_nsu07.pdf
- https://tvims.nsu.ru/chernova/ms/ms_nsu14.pdf
- **Наглядная математическая статистика. Лагутин Михаил Борисович**