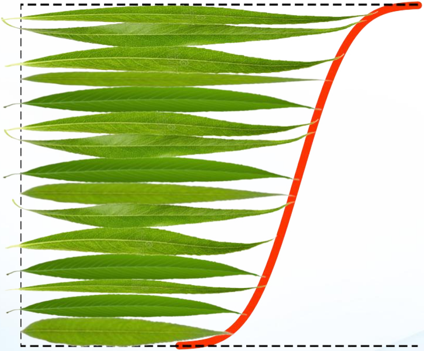
Распределения

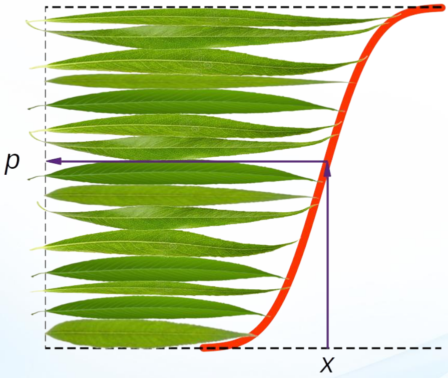
Соберем несколько листьев

Посмотрим на кончики



Приближенно получили функцию распределения нормального распределения.

Функция распределения



Функция распределения в точке x равна

доле листьев с длиной листа *не больше* x.

Виды распределений (основные)

Дискретные:

- Бернулли
- Биномиальное
- Равномерное
- Геометрическое

Абсолютно непрерывные:

- Нормальное
- Равномерное

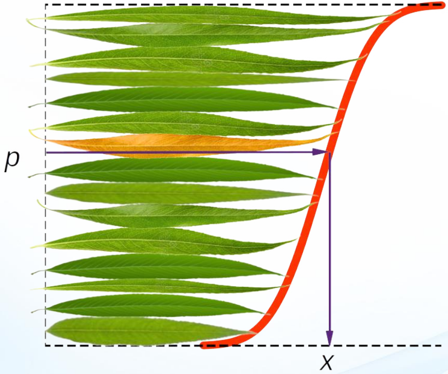
Что такое функция распределения

$F_\xi(x) = \mathsf{P}(\xi \leqslant x)$ — функция распределения случайной величины ξ.

Свойства из теории вероятностей:

1. Не убывает
2. Непрерывная справа, может иметь разрывы
3. $F(-\infty) = 0, F(+\infty) = 1$
4. Однозначно характеризует распределение.

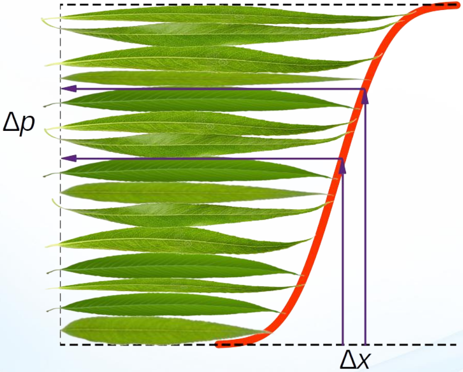
Возьмем значение p. Какой лист ему соответствует?



p-**квантиль** равна наименьшей длине листа, т.ч. есть не менее $p \cdot 100\%$ листьев с длиной листа не больше данного листа.

Формально: $u_p = \min\{x \mid F(x) \geqslant p\}$

Плотность



Плотность в точке x равна $\Delta p / \Delta x$,

т.е. доле листьев с длиной листа в окрестности x.

Что такое плотность

Свойства:

- лежит не ниже горизонтальной оси
- площадь под кривой равна 1
- неограничена сверху
- вероятности события $\{a \leq \xi \leq b\}$ соответствует площадь под кривой между точками a и b
- равна производной функции распределения

Формальные определения и свойства см. теорию вероятностей.

Дискретные распределения

Бернулли

Обозначение: $Bern(p)$

Параметры: $p \in (0, 1)$

Носитель: $\{0, 1\}$

Вероятность: $P(\{1\}) = p$

Математическое ожидание: p

Дисперсия: $p(1 - p)$

Интерпретация:

p — вероятность выпадения орла у монетки

Бернулли

- кто родится: мальчик или девочка?
- сдашь ты экзамен или нет?

Биномиальное

Обозначение: $Bin(n, p)$

Параметры: $n \in \mathbb{N}, p \in (0, 1)$

Носитель: $\{0, 1, \dots n\}$

Вероятность: $P(\{k\}) = C_n^k p^k (1-p)^{n-k}$

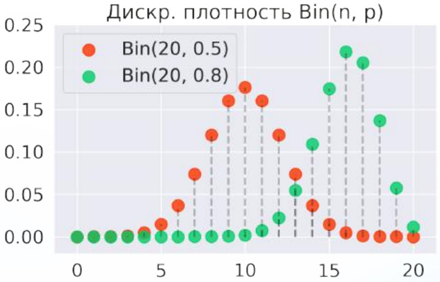
Математическое ожидание: np

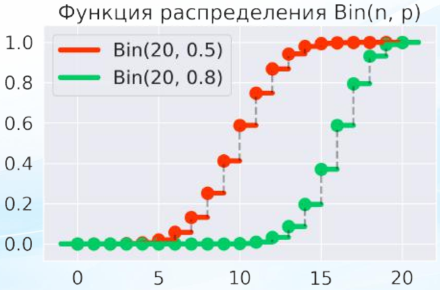
Дисперсия: $np(1-p)$

Интерпретация:

p — вероятность выпадения орла у монетки,

n — количество подбрасываний монетки





Биномиальное распределение

- кол-во людей, ответивших "да"в опросе
- кол-во дефектных продуктов на производстве
- кол-во выигранных матчей российской сборной

Равномерное

Обозначение: $U(1, 2 \dots N)$

Параметры: $N \in \mathbb{N}$

Носитель: $\{1, \dots N\}$

Вероятность: $P(\{k\}) = \frac{1}{N}$

Математическое ожидание: $\frac{N+1}{2}$

Дисперсия: $\frac{N^2-1}{12}$

Интерпретация:

N — количество шариков в мешке

Равномерное распределение

- бросок шестигранного кубика
- генерация случайной подвыборки для обзвона
- распределение встречаемости цифр в числе пи

Геометрическое

Обозначение: $Geom(p)$

Параметры: $p \in (0, 1]$

Носитель: $\mathbb{N}$

Вероятность: $P(\{k\}) = p(1-p)^{k-1}$

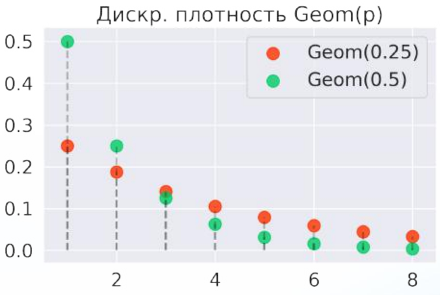
Математическое ожидание: $\frac{1}{p}$

Дисперсия: $\frac{1-p}{p^2}$

Интерпретация:

p — вероятность выпадения орла у монетки

Число $P(\{k\})$ интерпретируется как вероятность того, что в первый раз орел выпадет на k-ом подбрасывании монетки



Геометрическое распределение

- отток пользователей на k-й день использования продукта
- первое проявление плохого гена в k-ом поколении
- рождение двух девочек и затем мальчика

Абсолютно непрерывные распределения

Нормальное

Обозначение: $\mathcal{N}(a, \sigma^2)$

Параметры: $a \in \mathbb{R}, \sigma \in \mathbb{R}_+$

Носитель: $\mathbb{R}$

Плотность: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[\frac{-(x-a)^2}{2\sigma^2}\right]$

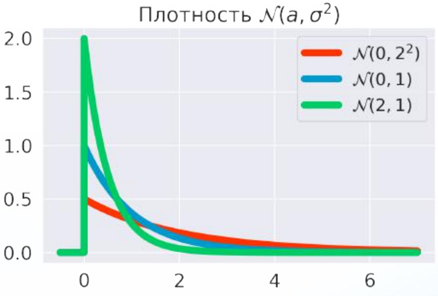
Математическое ожидание: a

Дисперсия: σ^2

Интерпретация:

a — среднее значение

σ — разброс значений



Нормальное распределение

- центральная предельная теорема
- моделирование погрешностей
- статистические методы
- броуновское движение

Равномерное

Обозначение: $U(a, b)$

Параметры: $a, b \in \mathbb{R}, a < b$

Носитель: $[a, b]$

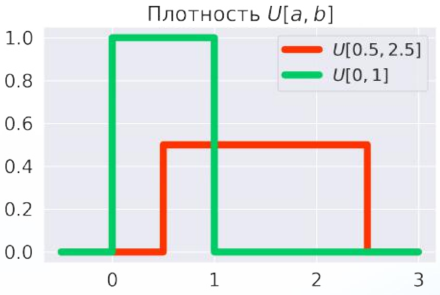
Плотность: $p(x) = \frac{1}{b-a} I(x \in [a, b])$

Математическое ожидание: $\frac{a+b}{2}$

Дисперсия: $\frac{(b-a)^2}{12}$

Интерпретация:

a и b — концы отрезка-носителя



Равномерное

- генерация случайной точки из отрезка
- генерация произвольных распределений
- байесовские методы

Генерация распределений

Генерация распределений

Задача: сгенерировать $\psi \sim Bern(p)$, имея $\xi \sim U(0,1)$

Решение: $\psi = I\{\xi \leqslant p\}$

Задача: сгенерировать $\psi \sim Bin(n,p)$, имея $\xi \sim U(0,1)$

Решение: $\psi = \sum_{i=1}^{n} \xi_i$,
где $\xi_i \sim \mathrm{Bern}(1/2)$ — независимые случайные величины.

Задача: сгенерировать $\xi \sim U(0,1)$, имея $\psi \sim Bern(1/2)$

Решение: запишем ξ в двоичной системе счисления: $\xi = 0,\xi_1\xi_2...\xi_n$,
где $\xi_i \sim \mathrm{Bern}(1/2)$ — независимые случайные величины.

Основные теоремы теории вероятностей

Закон больших чисел

Пусть $X_1, X_2, \ldots$ - последовательность независимых одинаково распределенных случайных величин, а $\overline{X_n} = \frac{X_1 + \ldots + X_n}{n}$ - арифметическое среднее первых n элементов

Слабый закон

$\overline{X_n} \xrightarrow{\mathbb{P}} \mu$, где μ - математическое ожидание X_1

Сильный закон

Если существует такая последовательность μ_n, что вероятность $\overline{X_n} - \mu_n \to 0$ равна 1 при $n \to \inf$, то:

$\overline{X_n} \to \mu$ почти наверное

Центральная предельная теорема

Пусть $X_1, X_2, \ldots$ - последовательность независимых одинаково распределенных случайных величин с математическим ожиданием μ и дисперсией σ^2, а $S_n = X_1 + \ldots + X_n$. Тогда:

$$\frac{S_n - \mu n}{\sigma \sqrt{n}} \to \mathcal{N}(0, 1)$$

Основная задача математической статистики

Введение

Теория вероятностей

Зная природу случайного явления,
посчитать характеристику этого явления.

Математическая статистика

По результатам экспериментальных данных
высказать суждение о том, какова была природа этого явления.

Классический пример

На курсе N студентов; из них M выбирает спецкурс по анализу данных.

Задача в теории вероятностей

$$\mathsf{P}(\text{среди случайных } n \text{ чел. ровно } m \text{ слушателей спецкурса}) - ?$$

Предполагается, что M известно.

Задача в математической статистике

Среди случайных n чел. есть m слушателей спецкурса.

Оценить M.

Предполагается, что M не известно.

Еще пример

$\xi \sim \mathcal{N}(a, \sigma^2)$ — случайная величина

Задача в теории вероятностей

Известно, что $a = 2.3, \sigma = 7.1$

$$\mathsf{P}(\xi \in [0,1]) - ?$$

$$\mathsf{E}\xi - ?$$

Задача в математической статистике

$x_1, ..., x_n$ — независимые реализации случайной величины ξ.

Оценить a и σ.

Вспоминаем оценки и погрешности в лабах!

Задача математической статистики

Пусть $x_1, ..., x_n$ — численные характеристики n-кратного повторения некоторого явления.

Будем их воспринимать как независимые реализации $\xi \sim \mathsf{P}$.

Задача: по значениям $x_1, ..., x_n$ высказать некоторое суждение о распределении P.

Решение: *статистический вывод* или *обучение*.

Основные понятия

Последовательность независимых одинаково распределенных случайных величин $X_1, \ldots X_n$ называется выборкой.

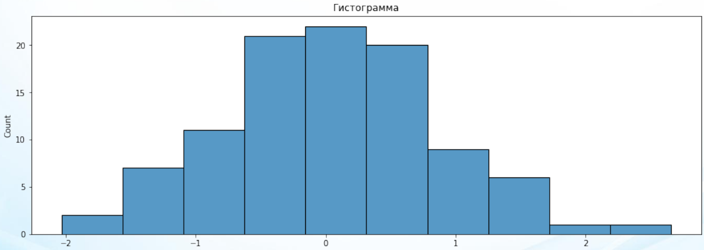
Их значения $x_1, \ldots, x_n$ как числа (на конкретном исходе) называются реализацией выборки.

Интуитивно: x_i - различные "измерения" какой-то величины.
Это имеющиеся у нас данные.

Давайте посмотрим, что вообще можно делать с данными!

Гистограмма

Пусть у нас есть реализация выборки $x_1, \ldots x_n \in \mathbb{R}$.
Идея: разделим всю числовую прямую на несколько "корзин" и посмотрим, сколько объектов (иксов) попало в каждую.

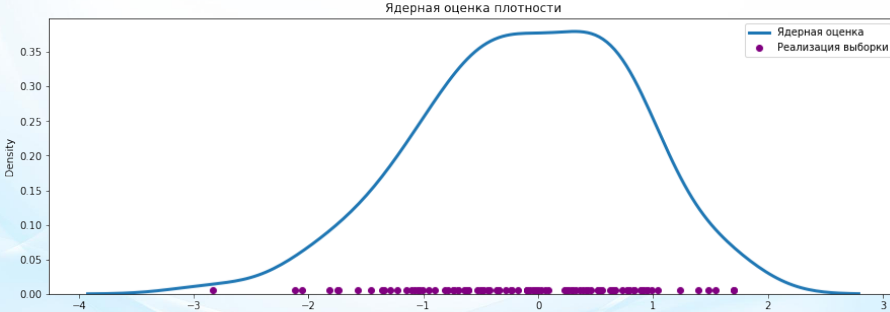


Можно построить график в виде столбиков, где высота столбика показывает, сколько объектов попало в соответствующую корзину.

Этот график по форме похож на график плотности распределения.

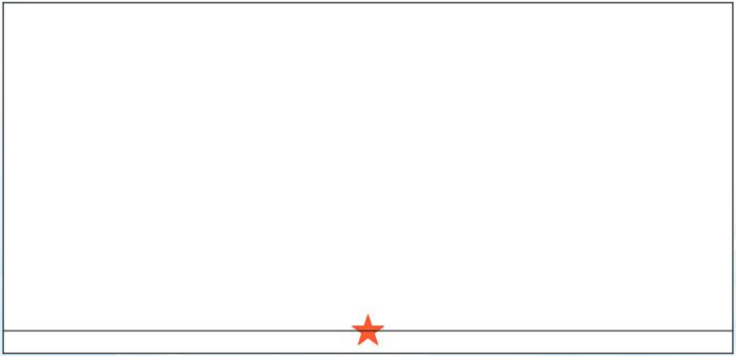
Ядерная оценка плотности

Идея: как-то оценить плотность распределения.

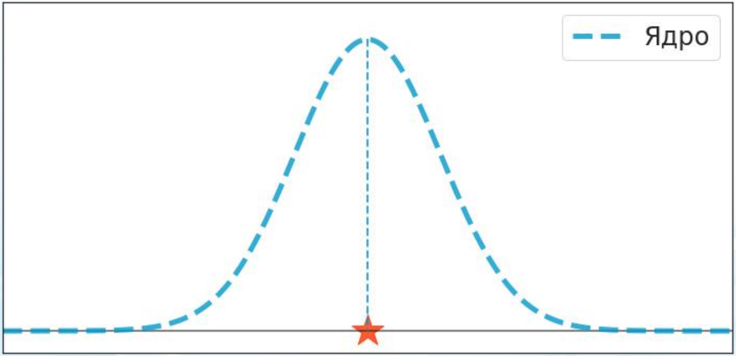


Как? Сейчас узнаем!

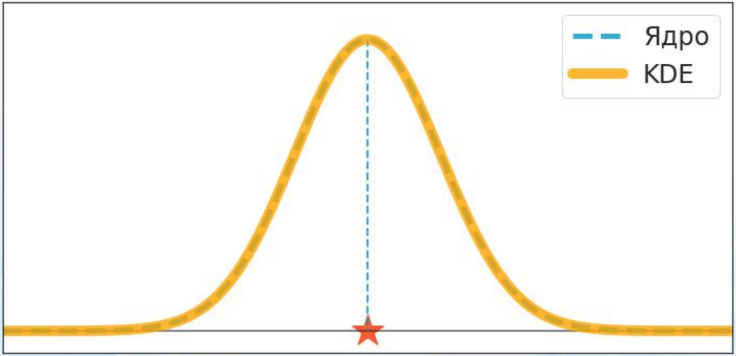
Ядерная оценка плотности: простые примеры



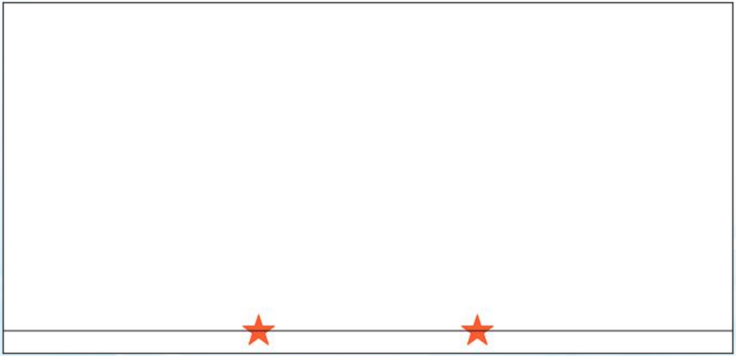
Ядерная оценка плотности: простые примеры



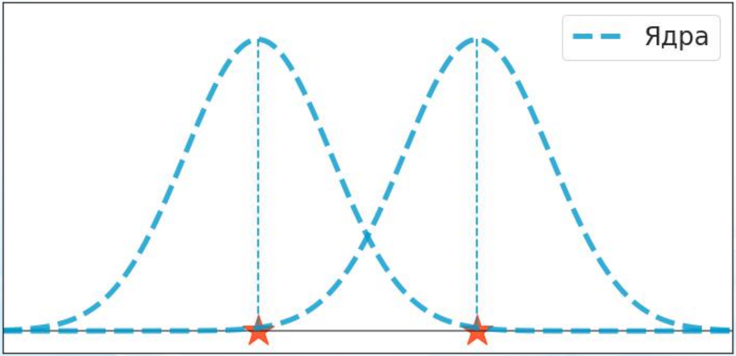
Ядерная оценка плотности: простые примеры



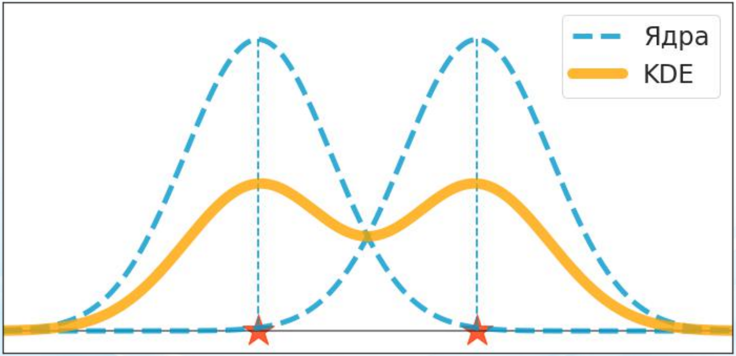
Ядерная оценка плотности: простые примеры



Ядерная оценка плотности: простые примеры

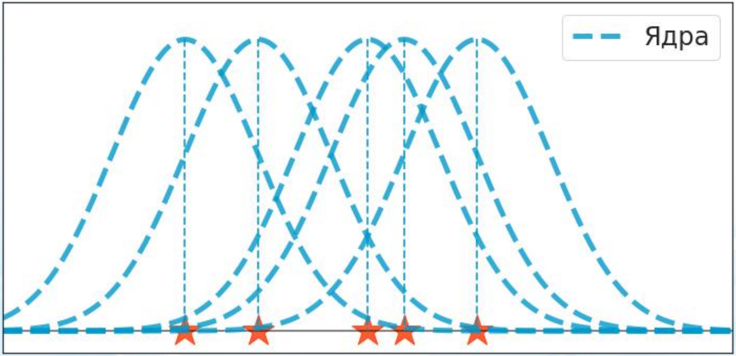


Ядерная оценка плотности: простые примеры

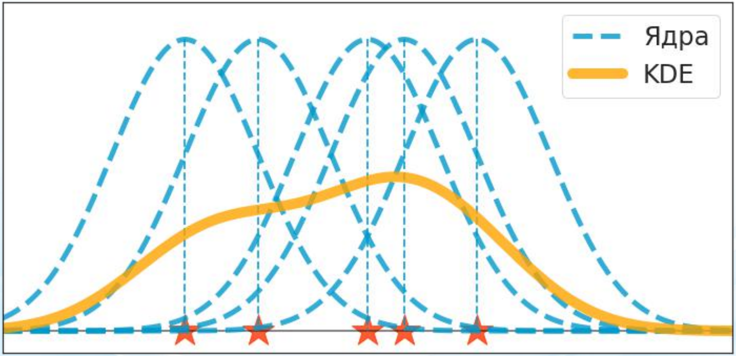


Ядерная оценка плотности: простые примеры

Ядерная оценка плотности: простые примеры



Ядерная оценка плотности: простые примеры



Определение

Пусть $X = (X_1, ..., X_n)$ — выборка из непрерывного распределения.

Выберем

- $q(x)$ — ядро = некоторая "базовая" симметричная плотность;
- $h > 0$ — ширина ядра, отвечающая за масштабирование.

Ядерная оценка плотности

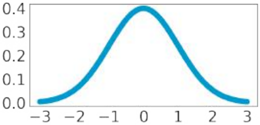
$$\widehat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} q\left(\frac{x - X_i}{h}\right)$$

Пояснение: в каждую точку выборки поставили отмасштабированное ядро и усреднили.

Виды ядер

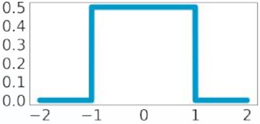
Гауссовское

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



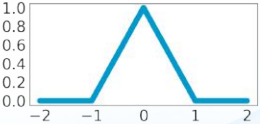
Прямоугольное

$$q(x) = \frac{1}{2} I\{|x| \leqslant 1\}$$



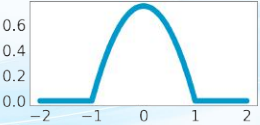
Треугольное

$$q(x) = (1 - |x|) I\{|x| \leqslant 1\}$$



Епанечникова

$$q(x) = \frac{3}{4}(1 - x^2) I\{|x| \leqslant 1\}$$



Давайте посмотрим на все это на практике!