Статистика

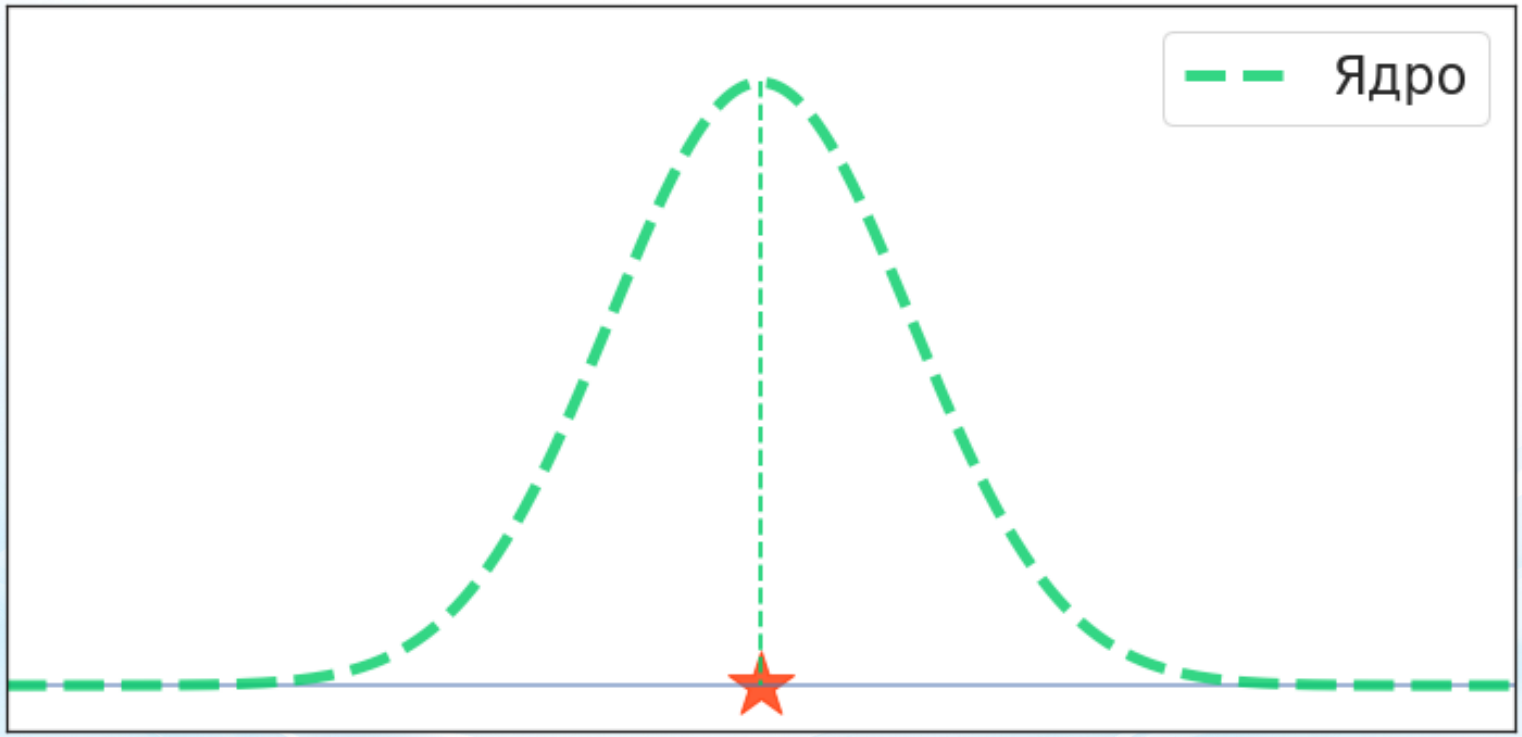
DS-поток

Лекция 10

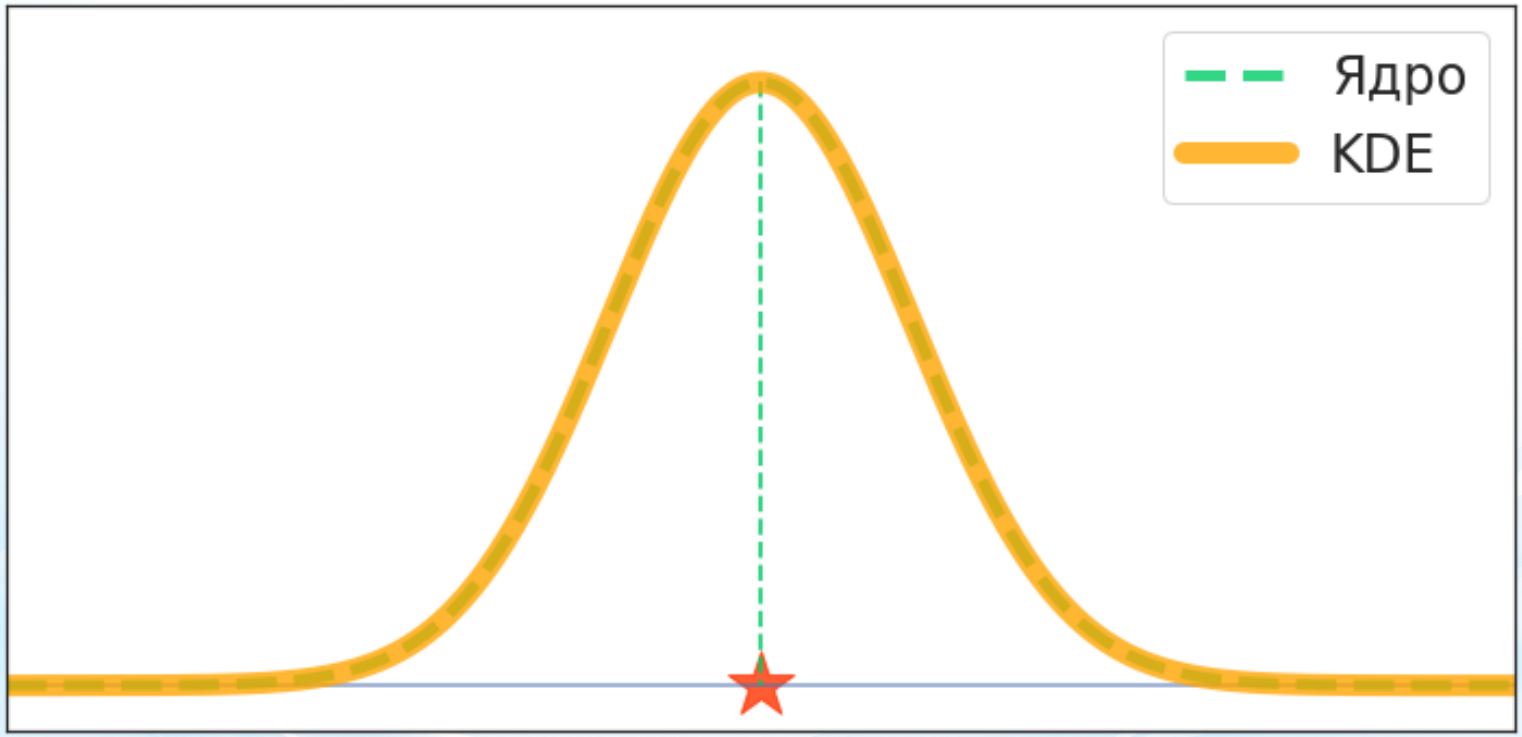
5.5 Ядерные оценки плотности

Ядерная оценка плотности: простые примеры

Ядерная оценка плотности: простые примеры

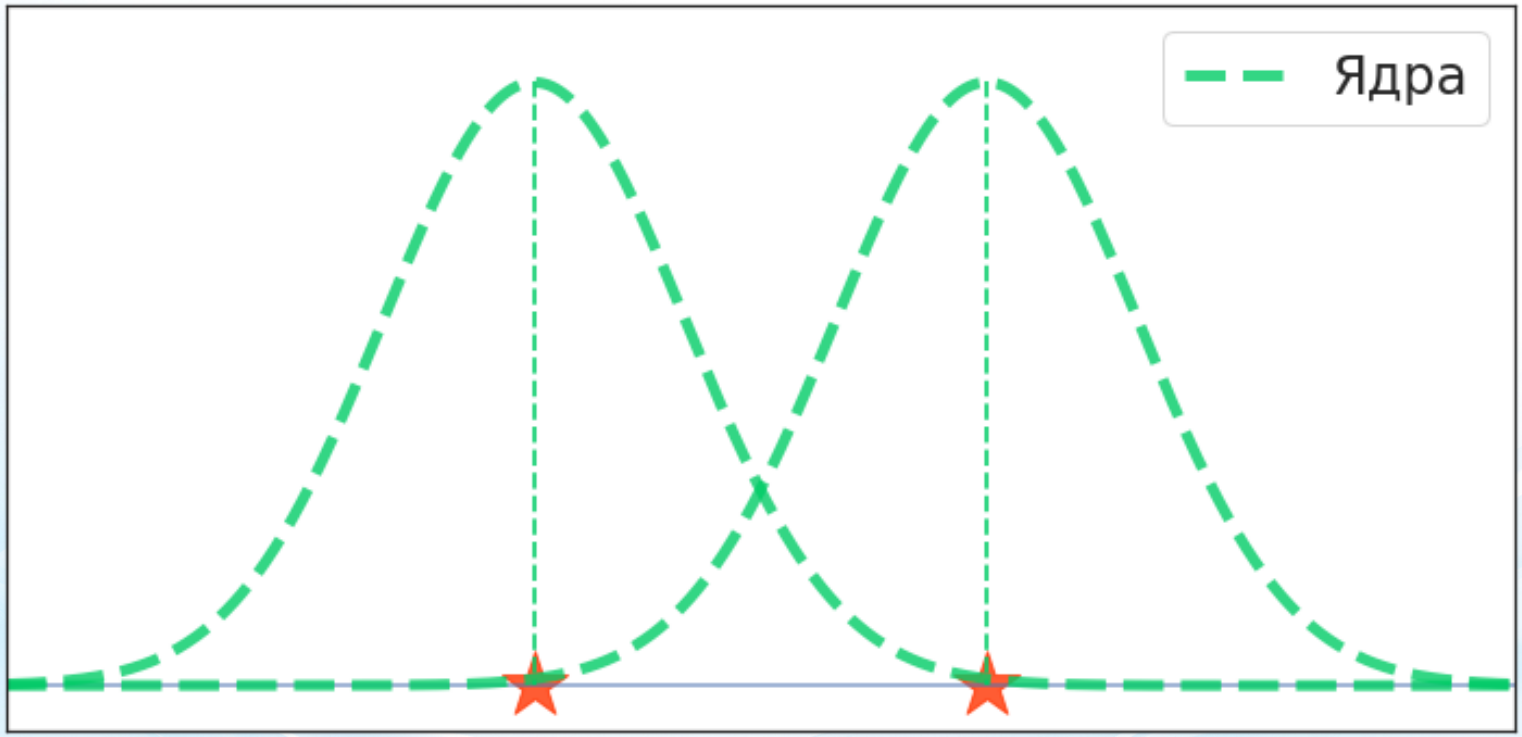


Ядерная оценка плотности: простые примеры

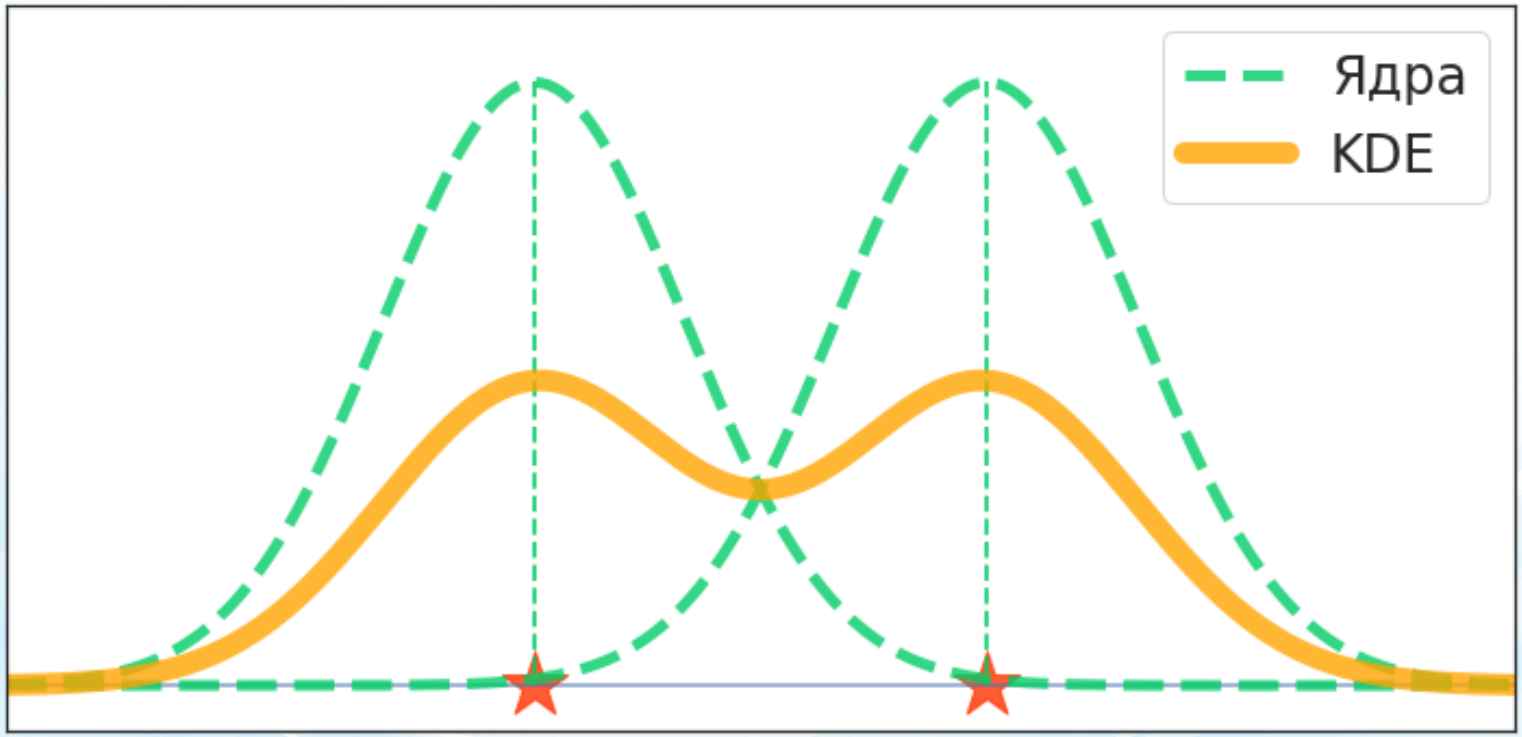


Ядерная оценка плотности: простые примеры

Ядерная оценка плотности: простые примеры

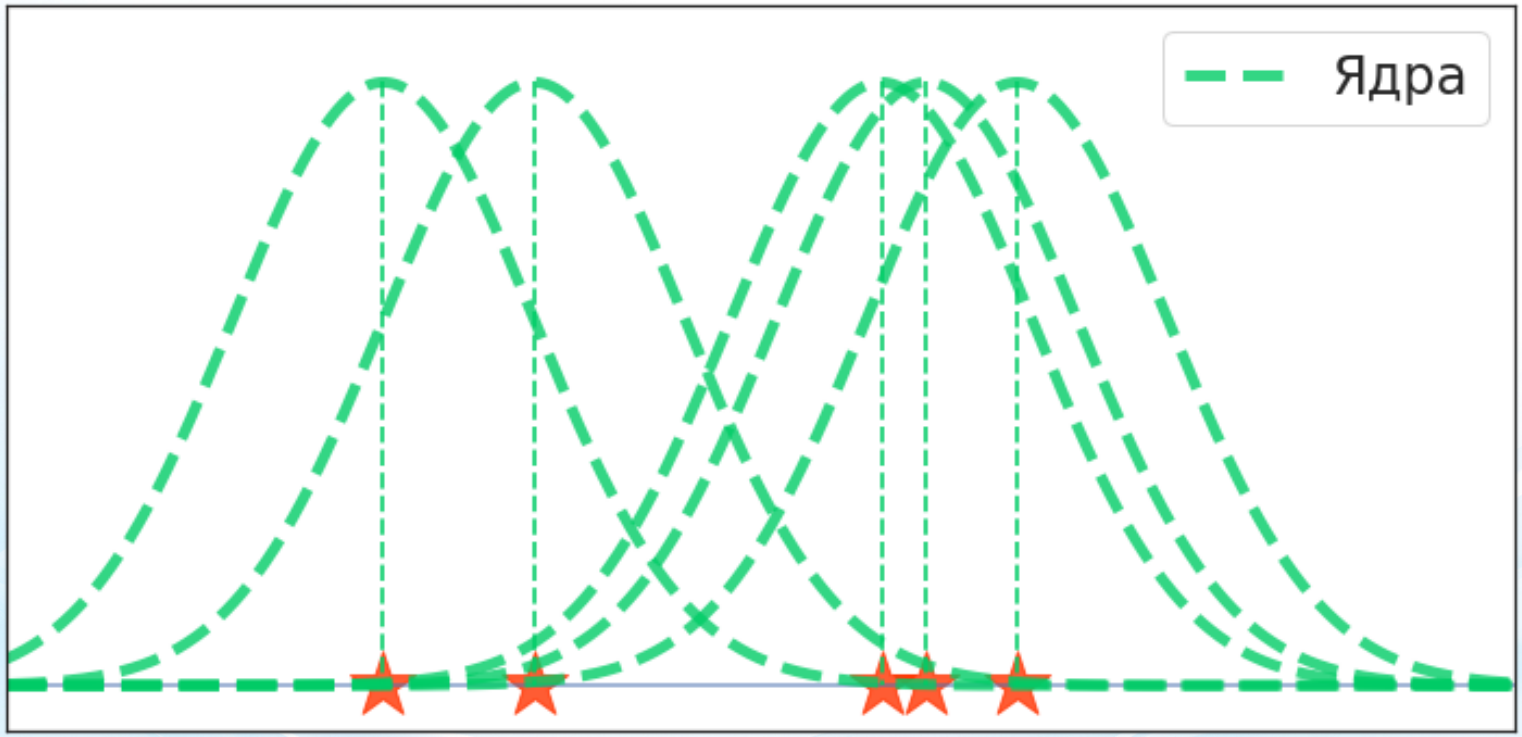


Ядерная оценка плотности: простые примеры

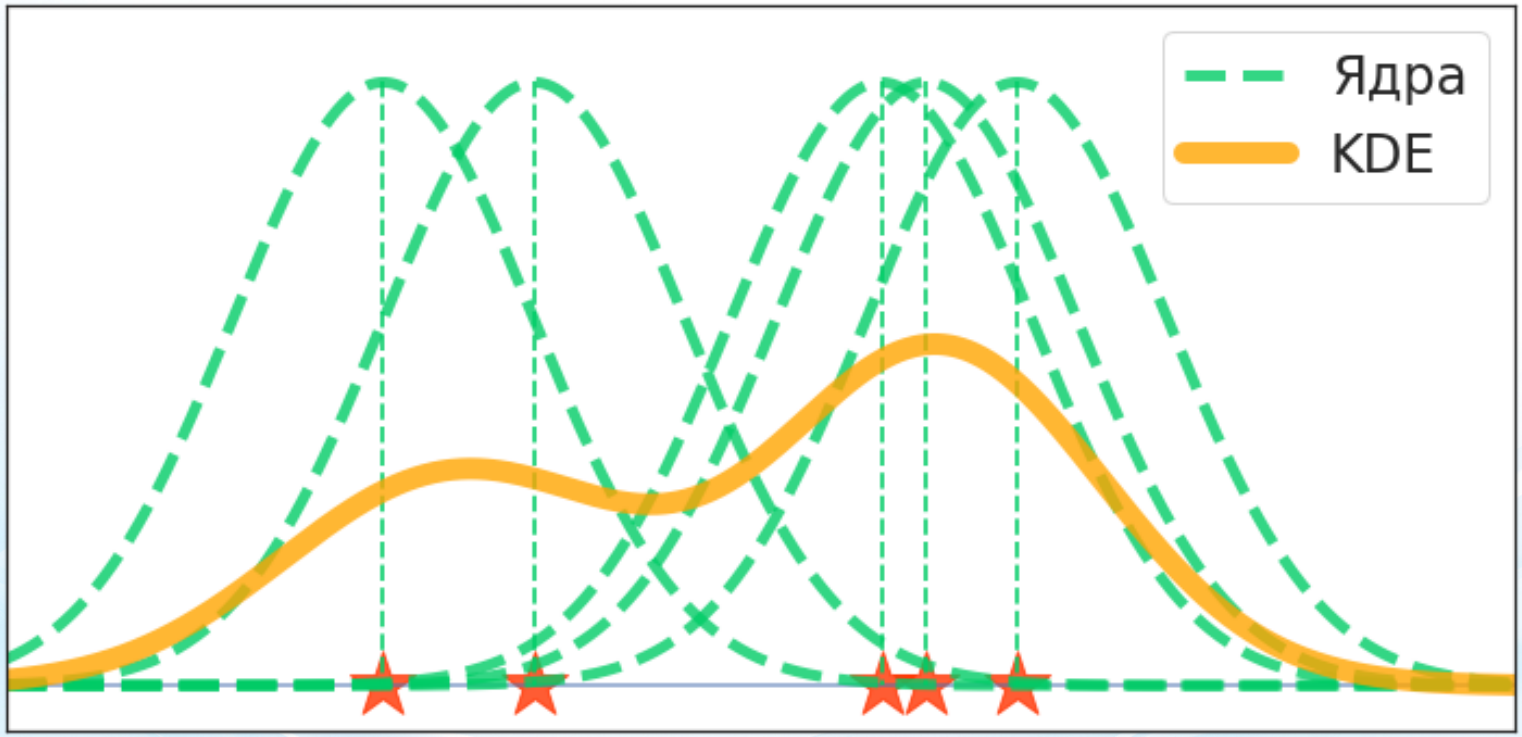


Ядерная оценка плотности: простые примеры

Ядерная оценка плотности: простые примеры



Ядерная оценка плотности: простые примеры



Определение

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

Рассматриваем $\mathscr{P}$ — все абс.-непрерывные распределения на $\mathbb{R}^d$.

Ядерная оценка плотности

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

где h и q — гиперпараметры:

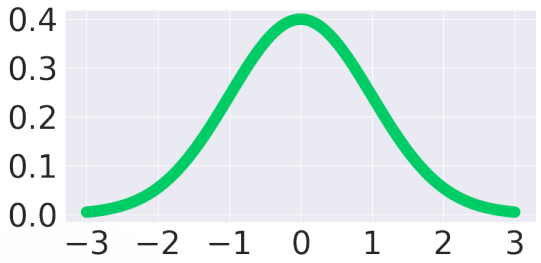
- $q(x)$ — ядро = некоторая "базовая" симметричная плотность;
- $h > 0$ — ширина ядра, отвечающая за масштабирование.

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

Виды ядер

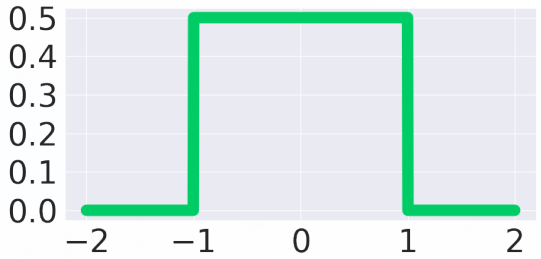
Гауссовское

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



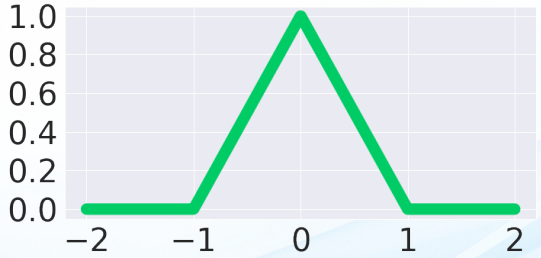
Прямоугольное

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



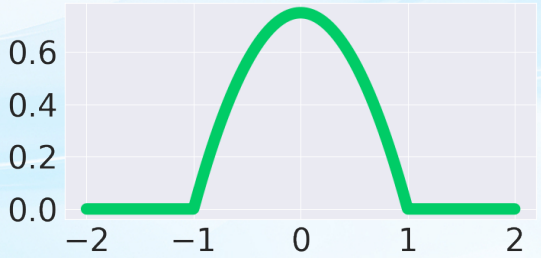
Треугольное

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



Епанечникова

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



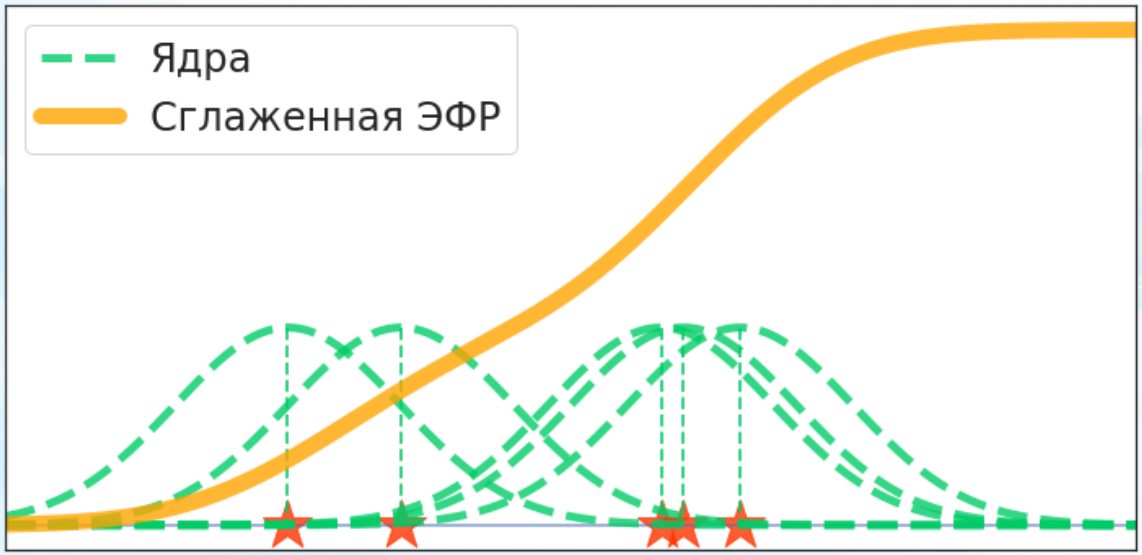
Связь с ЭФР

Ядерной оценке плотности $\widetilde{p}_h(x)$ соответствует **сглаженное эмпирическое распределение**

$$\widetilde{\mathrm{P}}_h(B) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Q}\left(\frac{B - X_i}{h}\right),$$

$$\frac{B - X_i}{h} = \left\{ \frac{x - X_i}{h} \,\middle|\, x \in B \right\},$$

где Q — распределение, соотв. плотности $q(x)$.

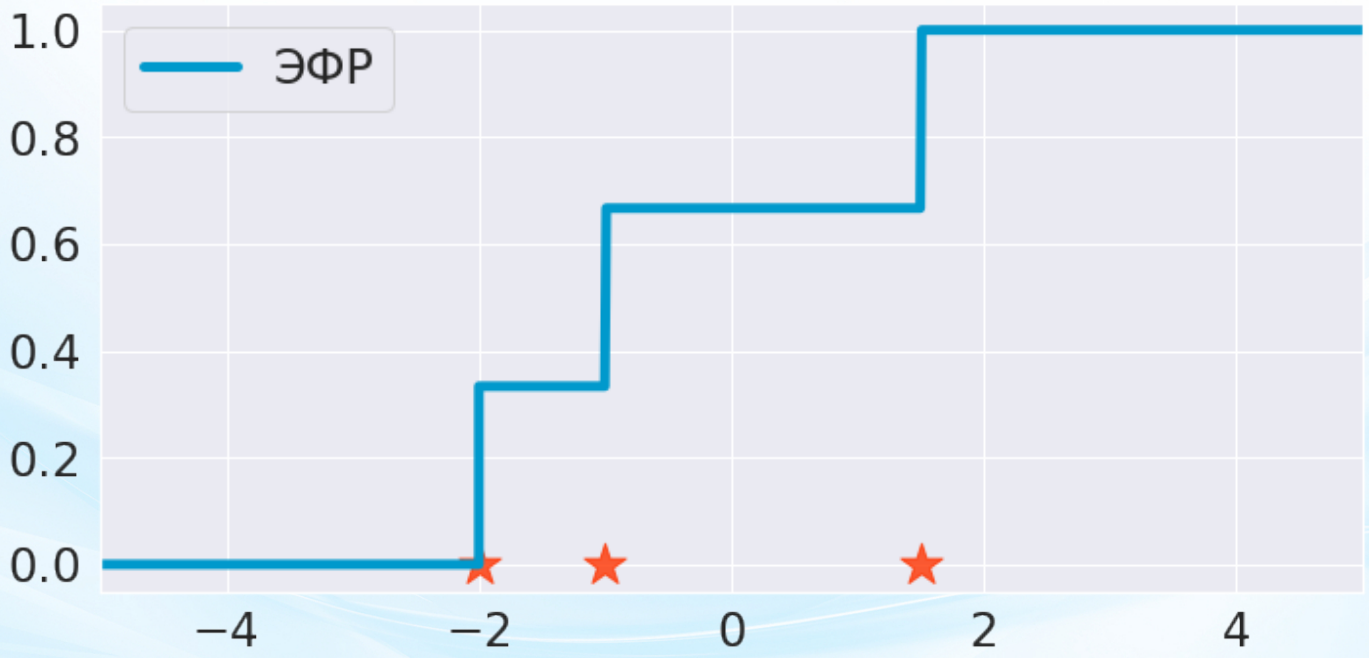


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

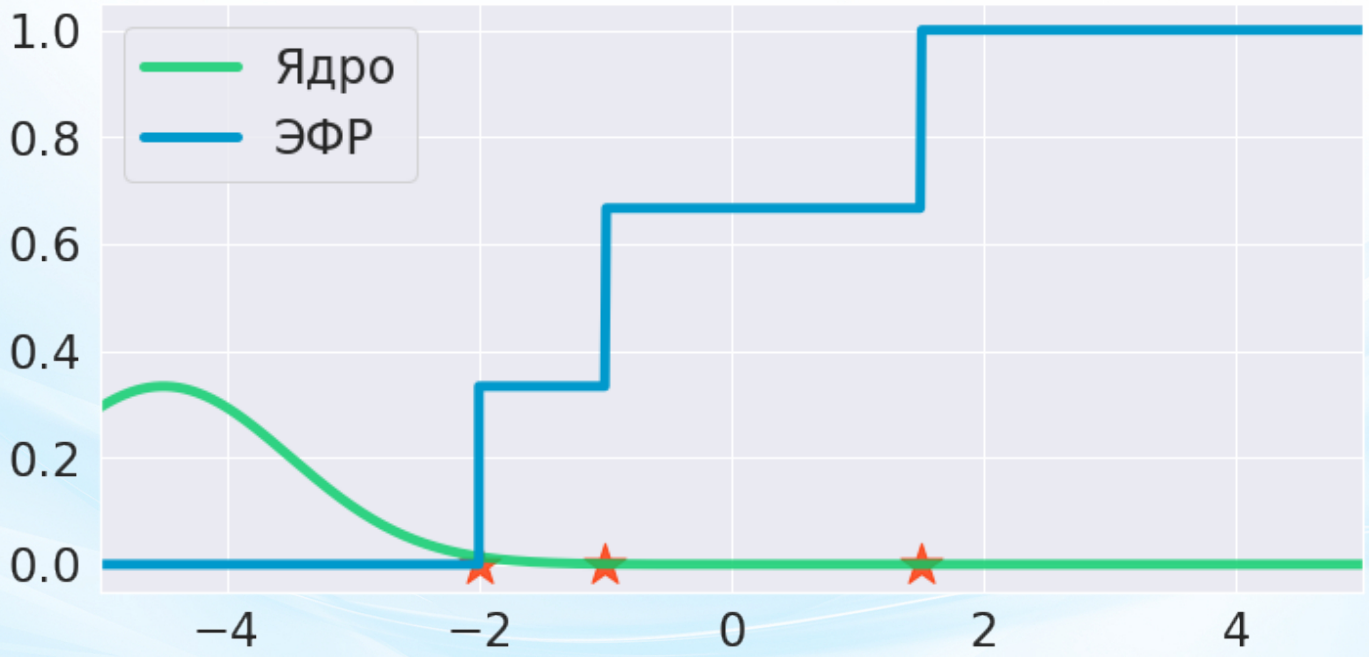


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

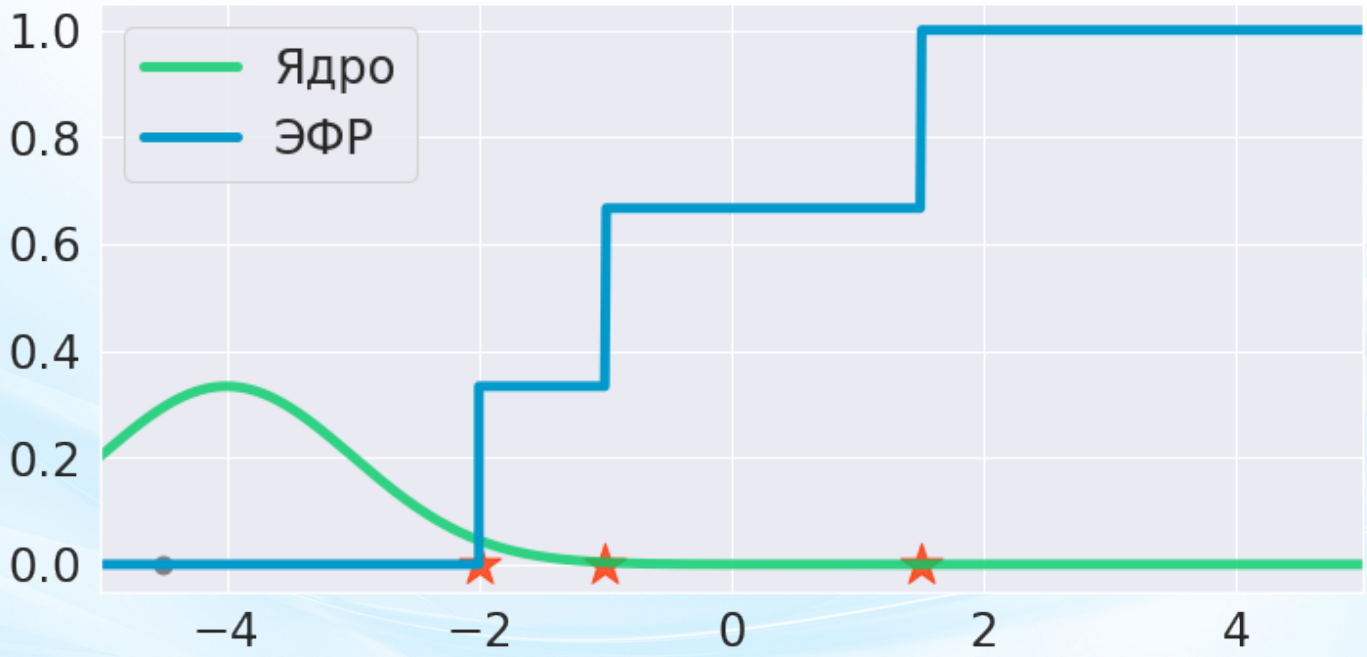


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

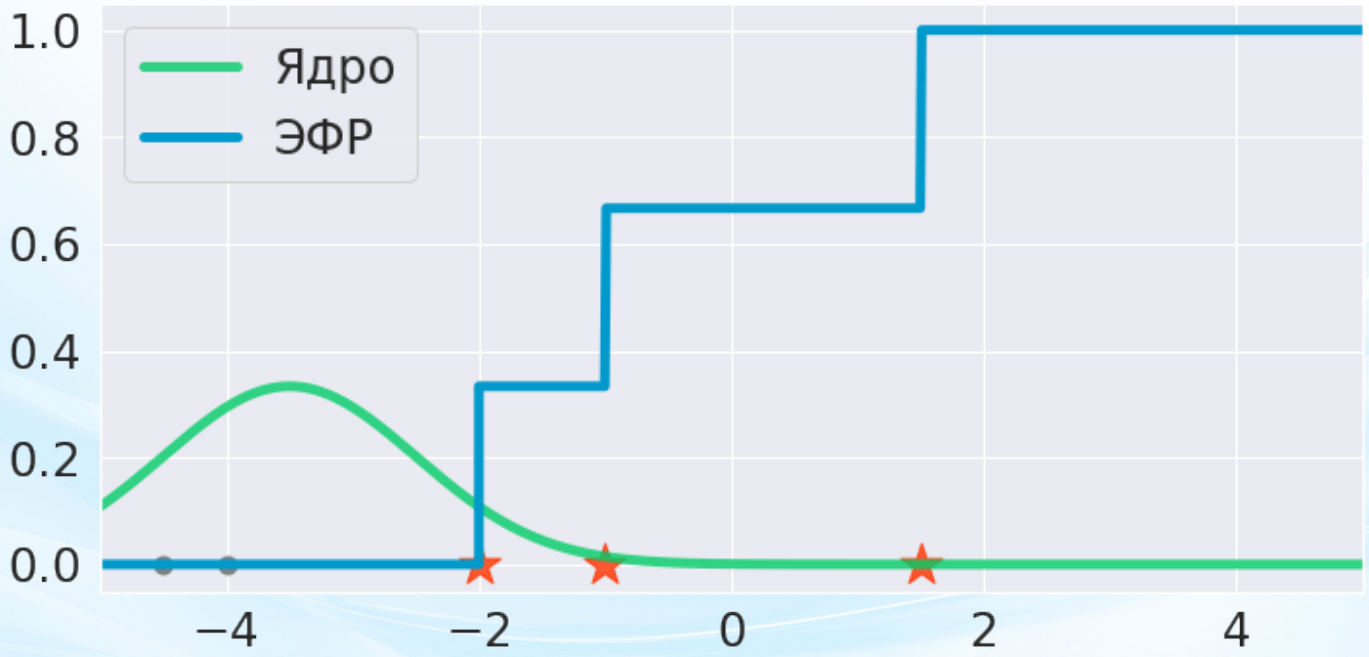


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

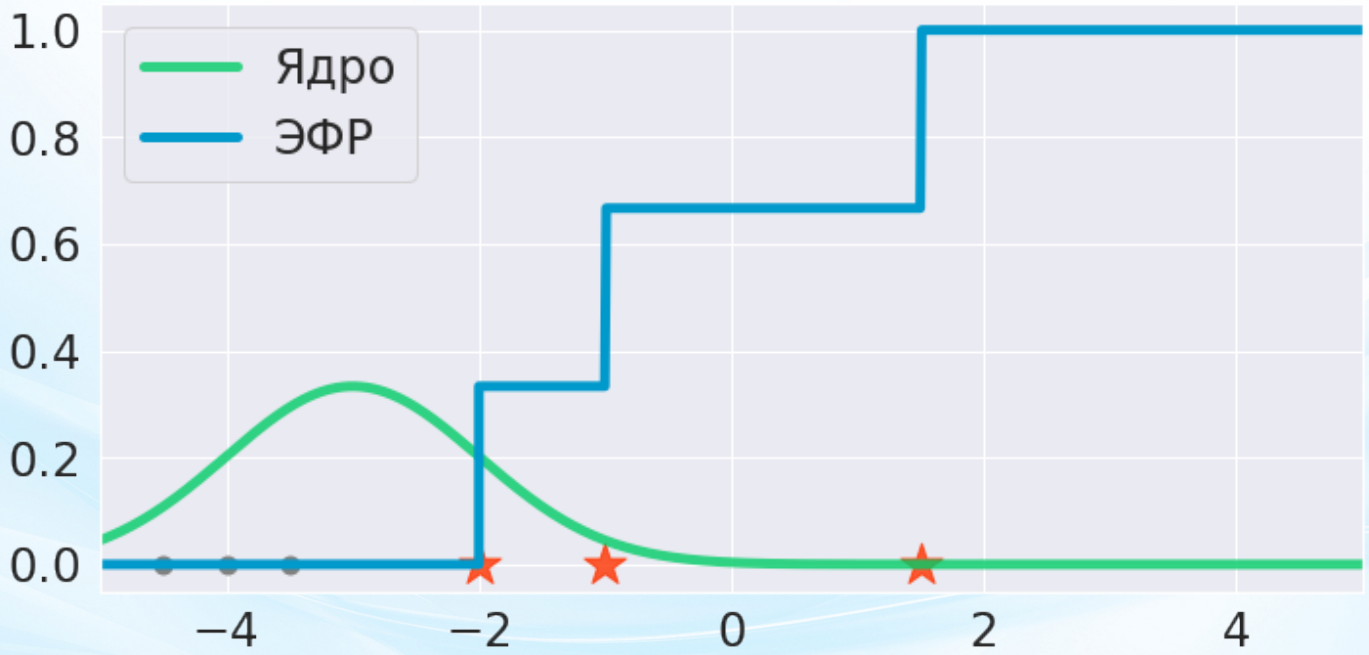


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

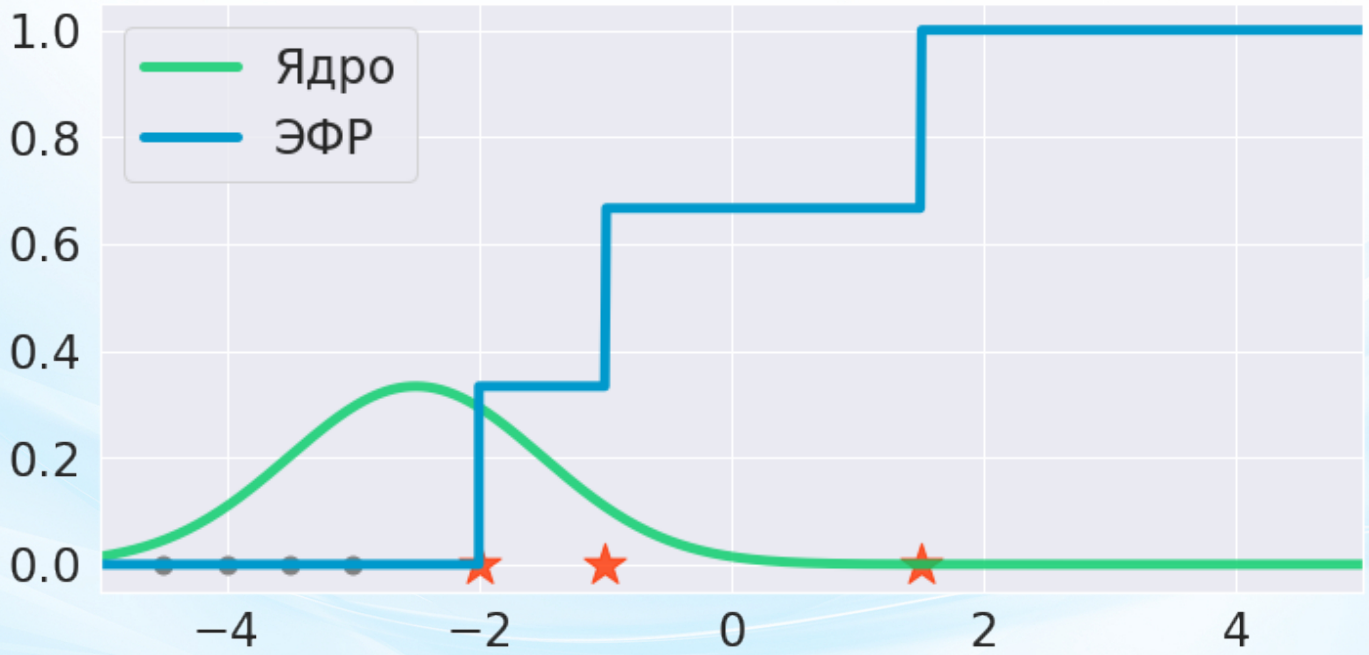


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

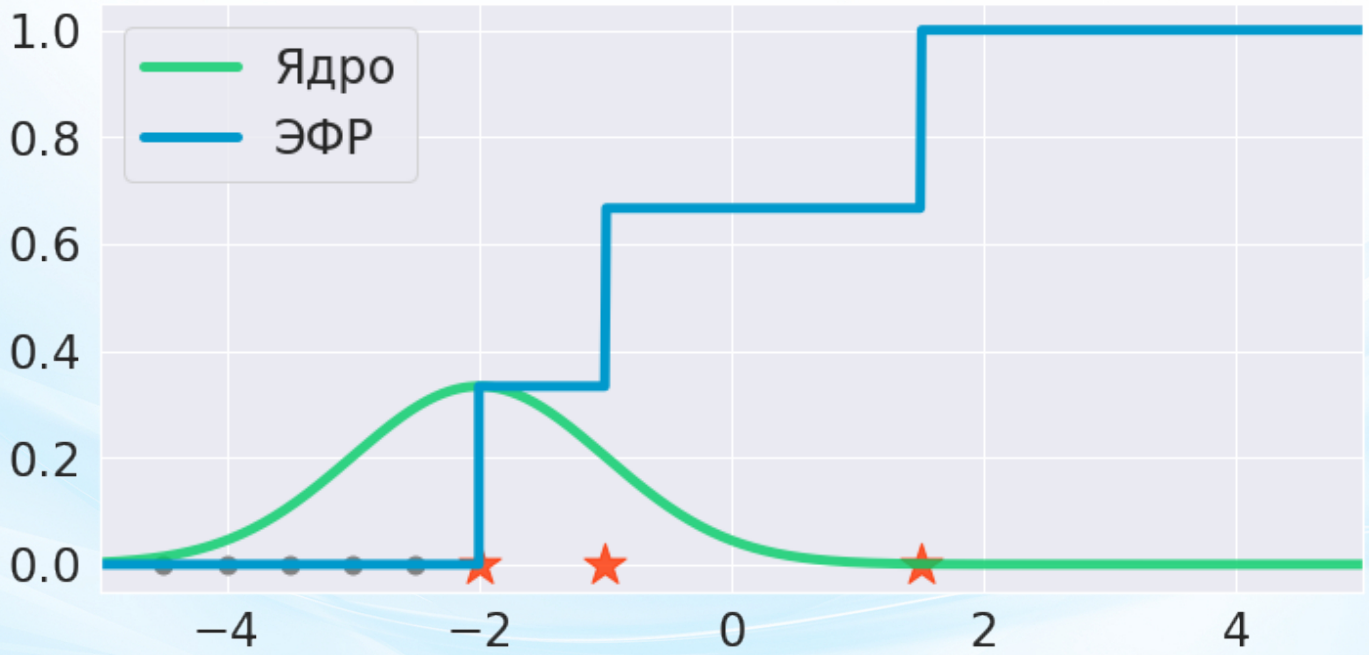


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

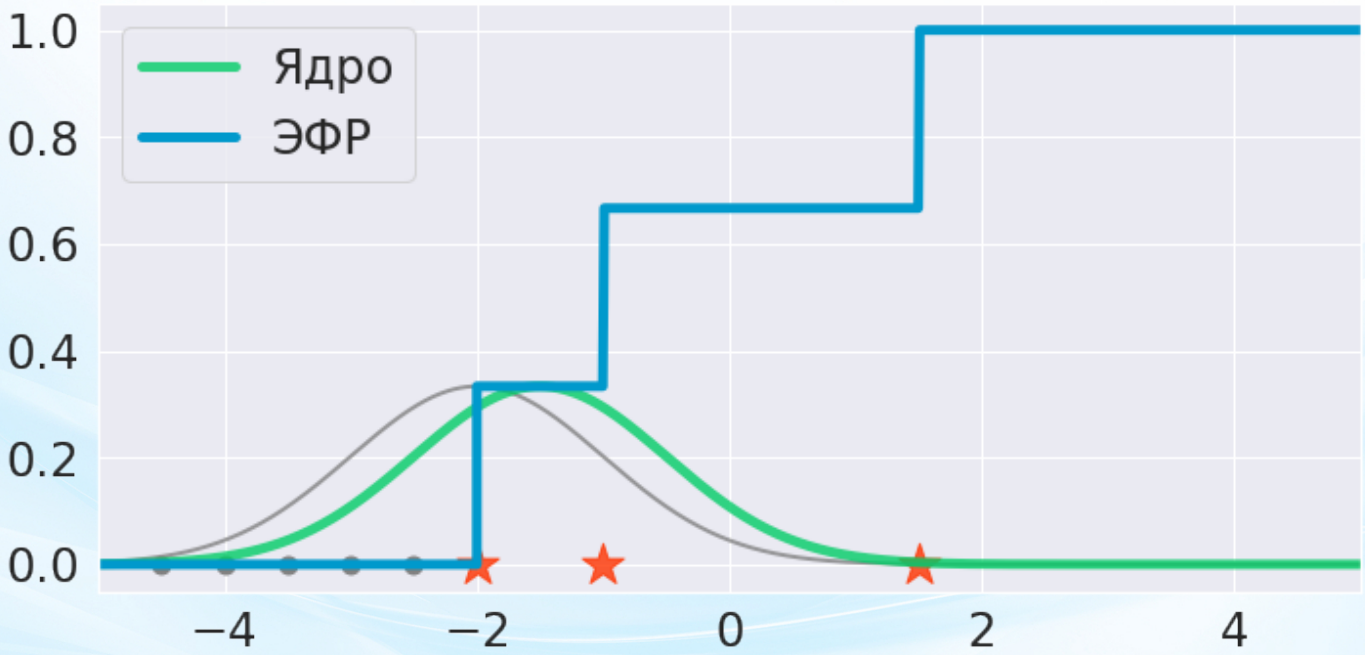


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

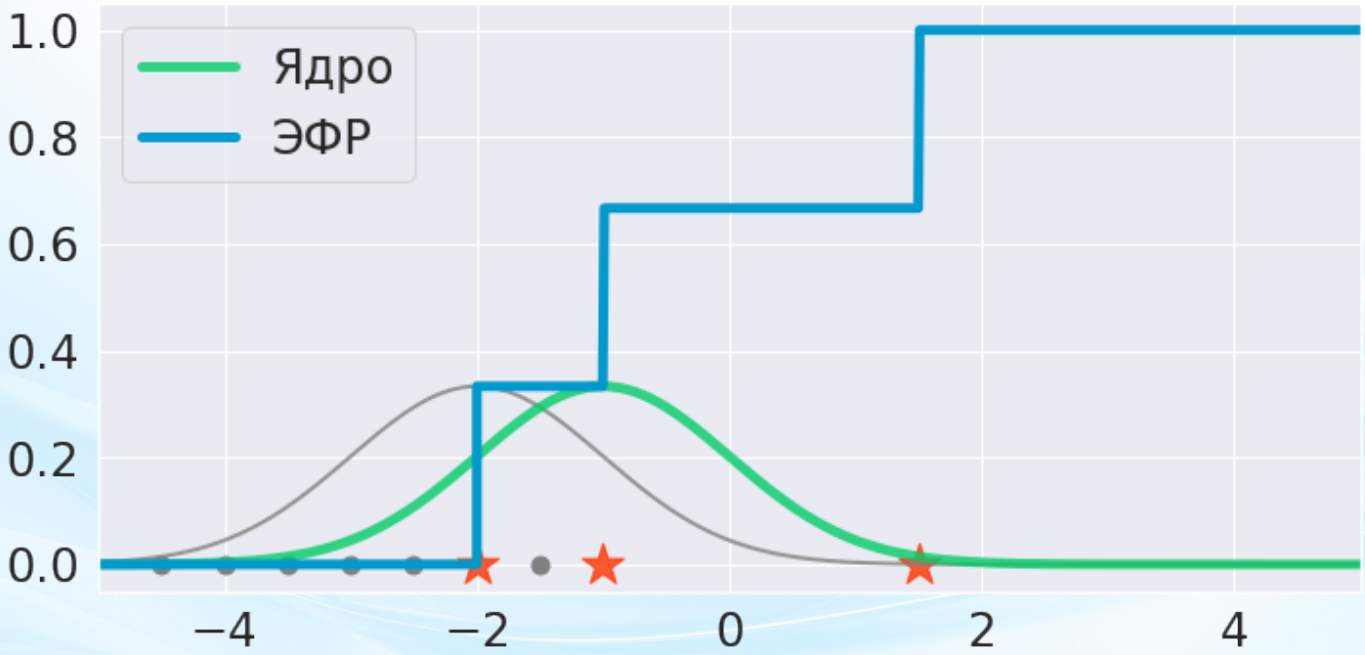


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

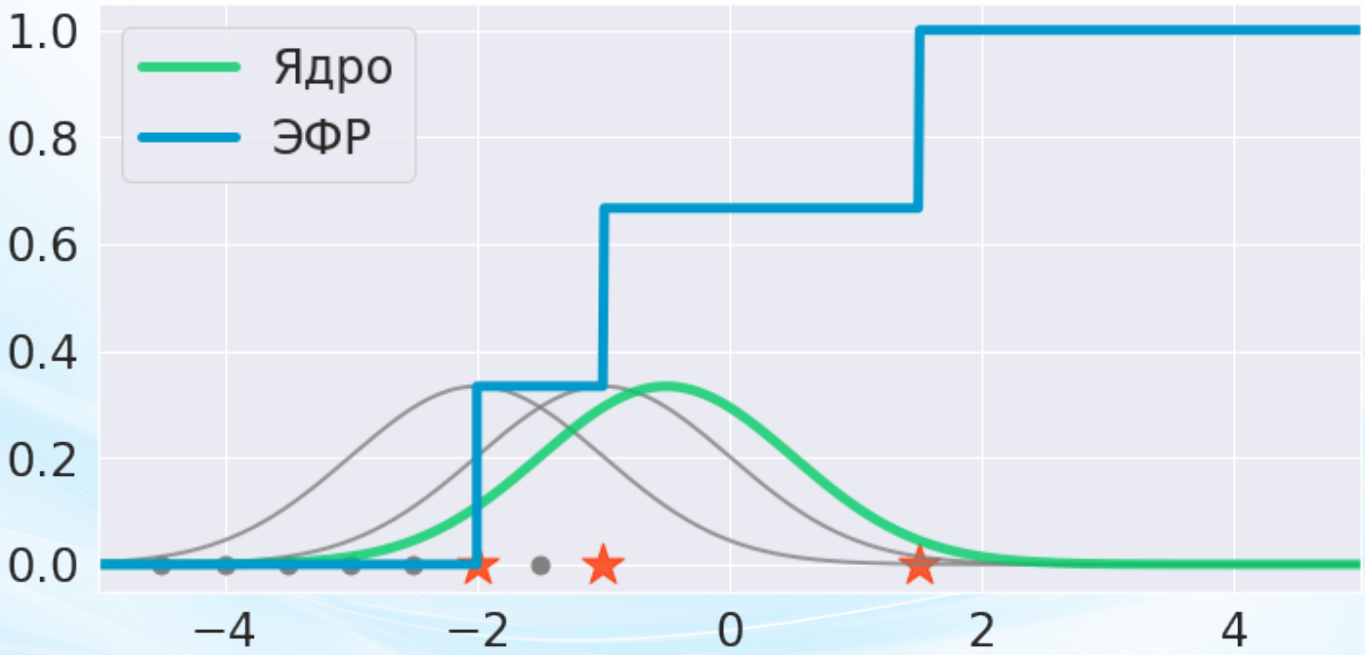


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

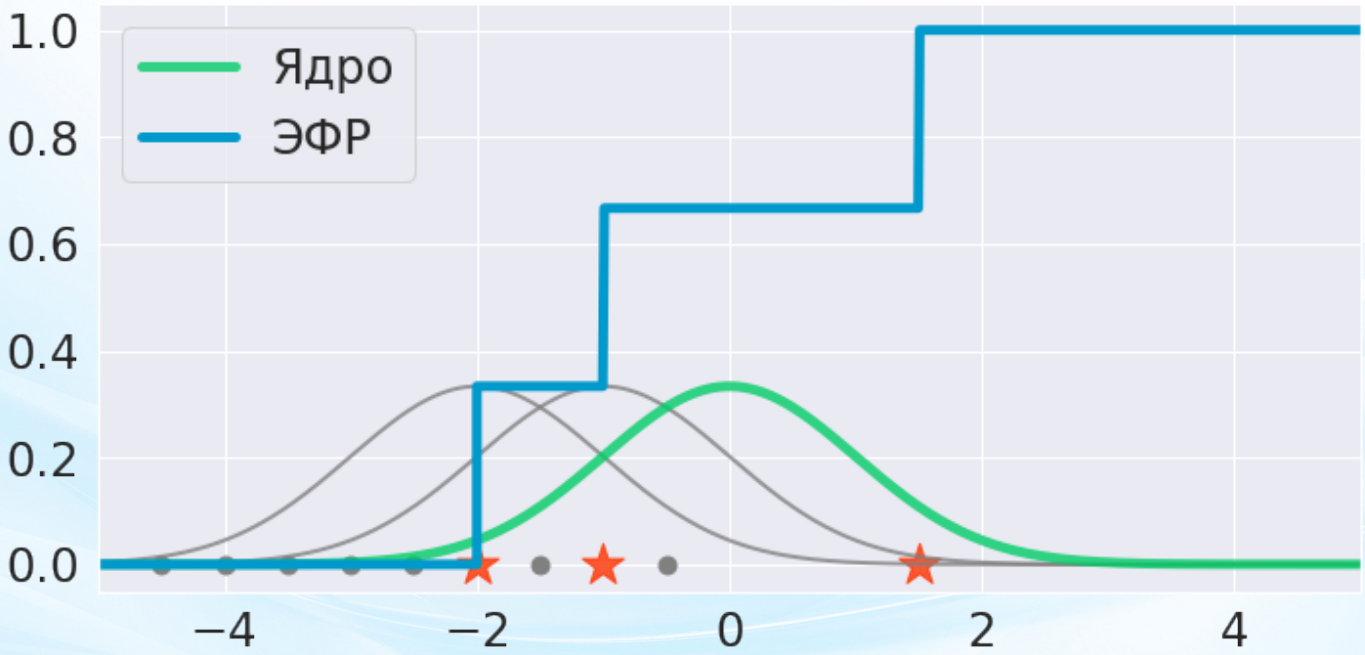


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

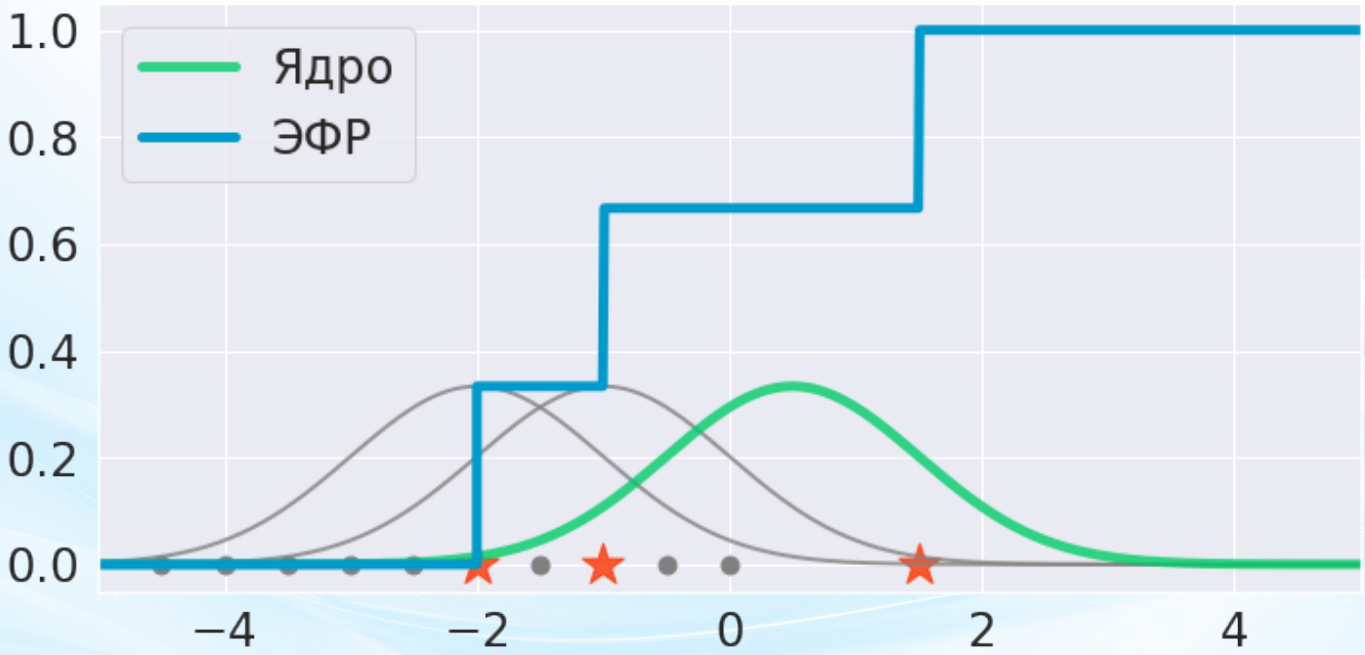


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

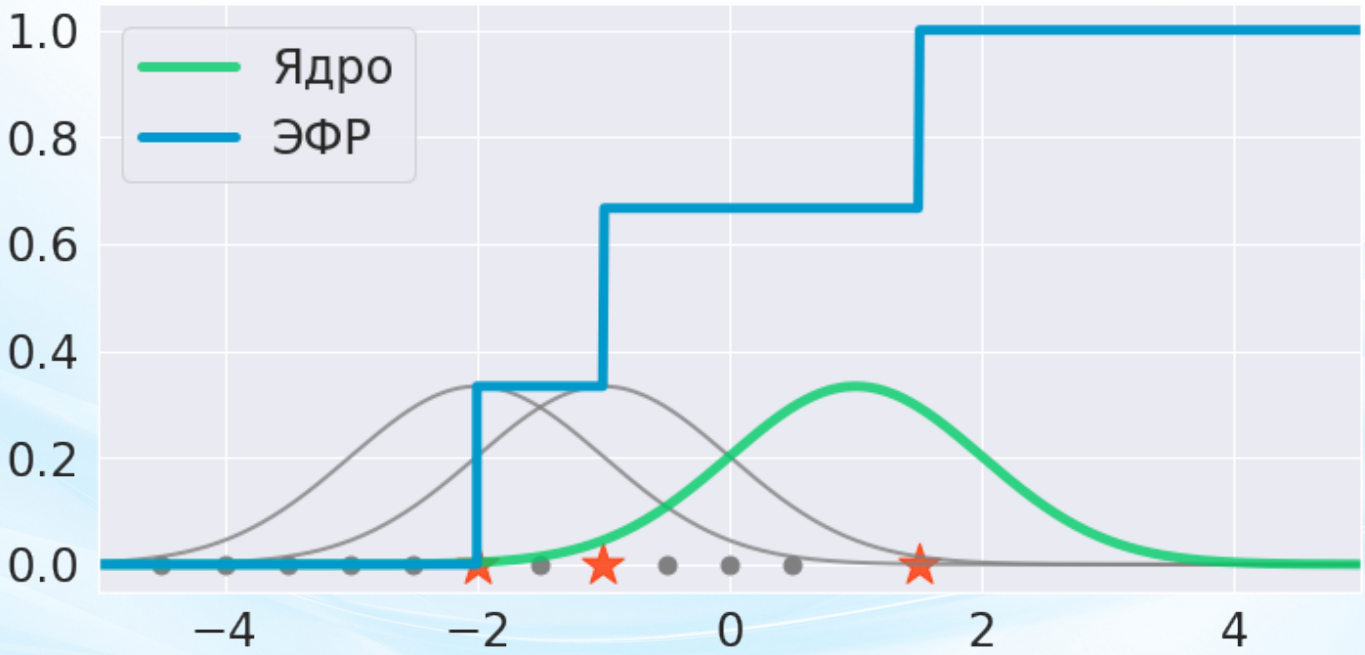


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

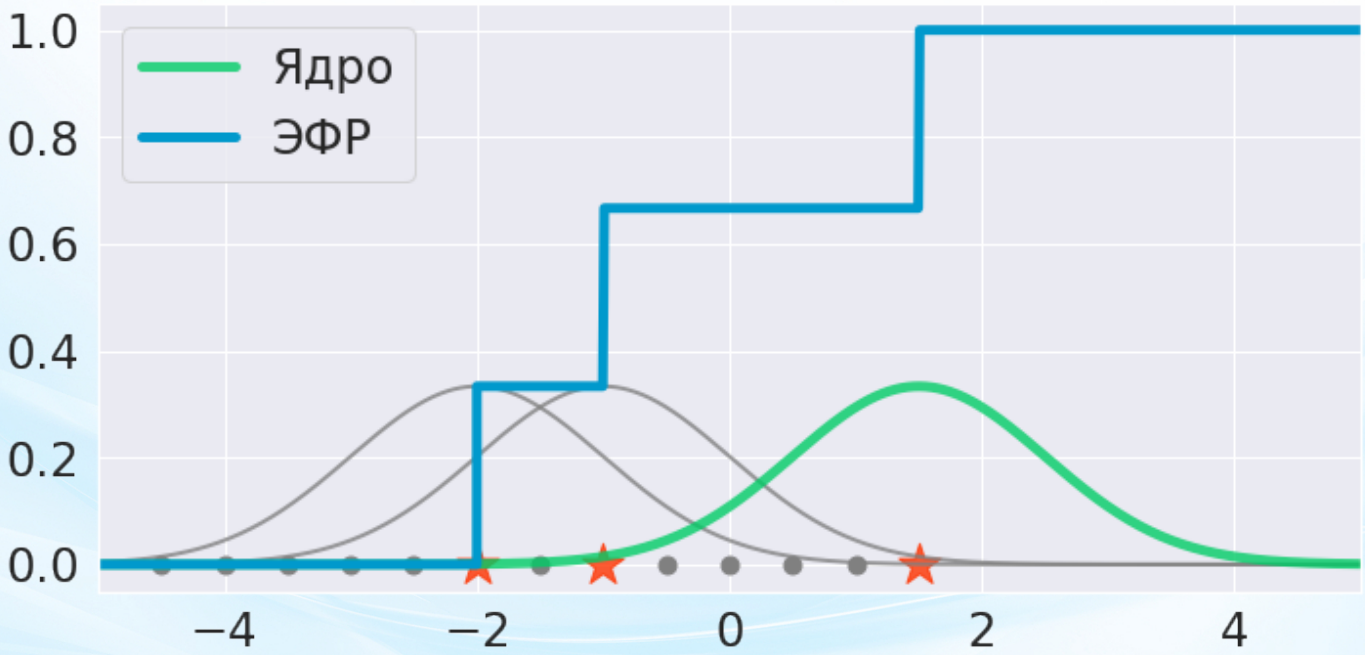


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

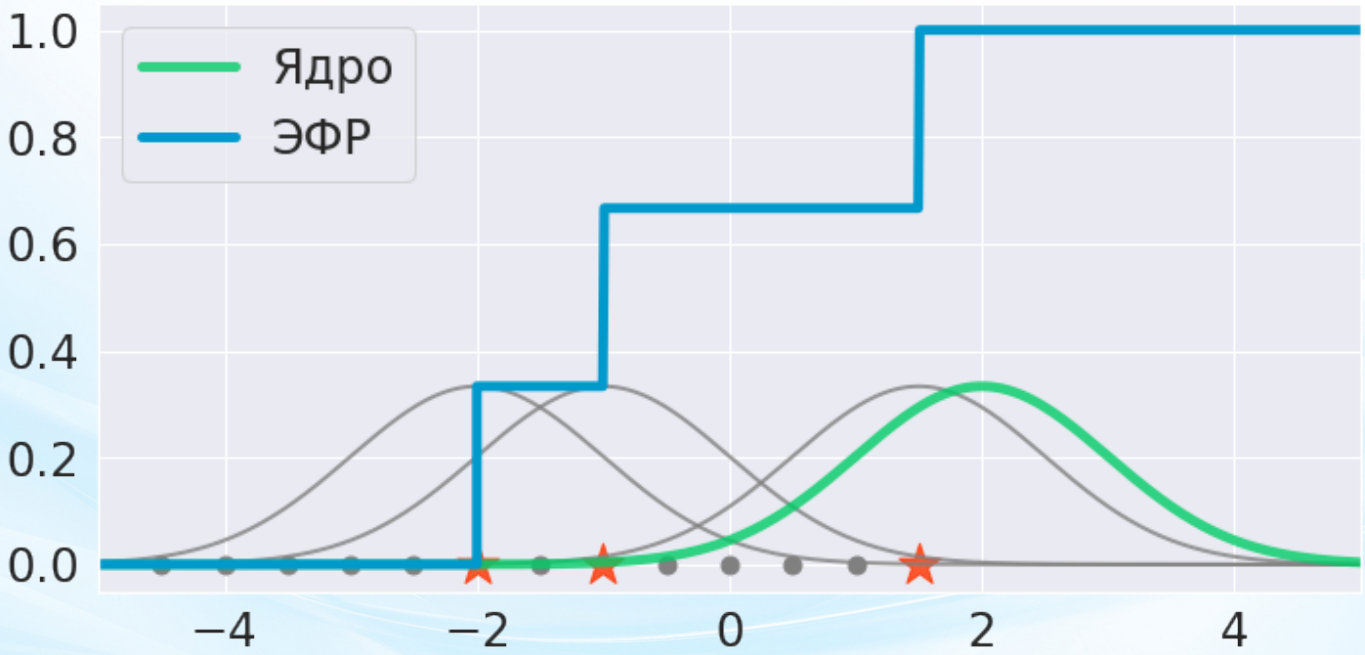


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

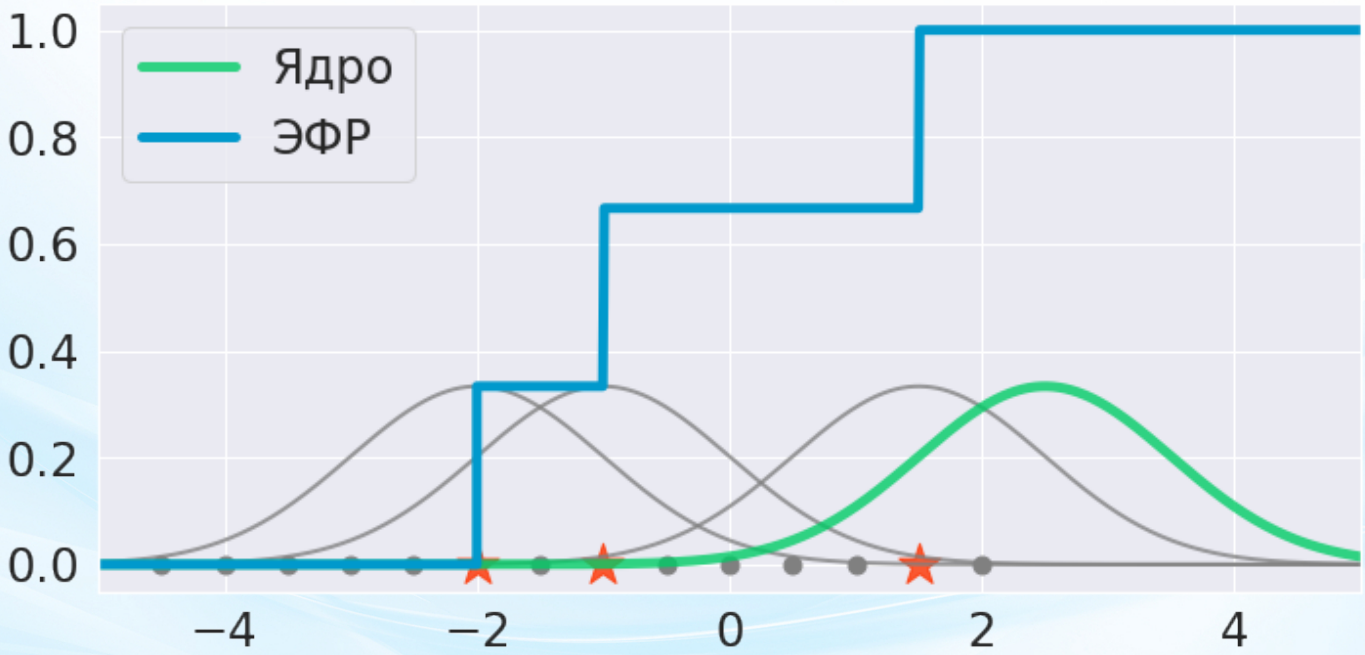


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

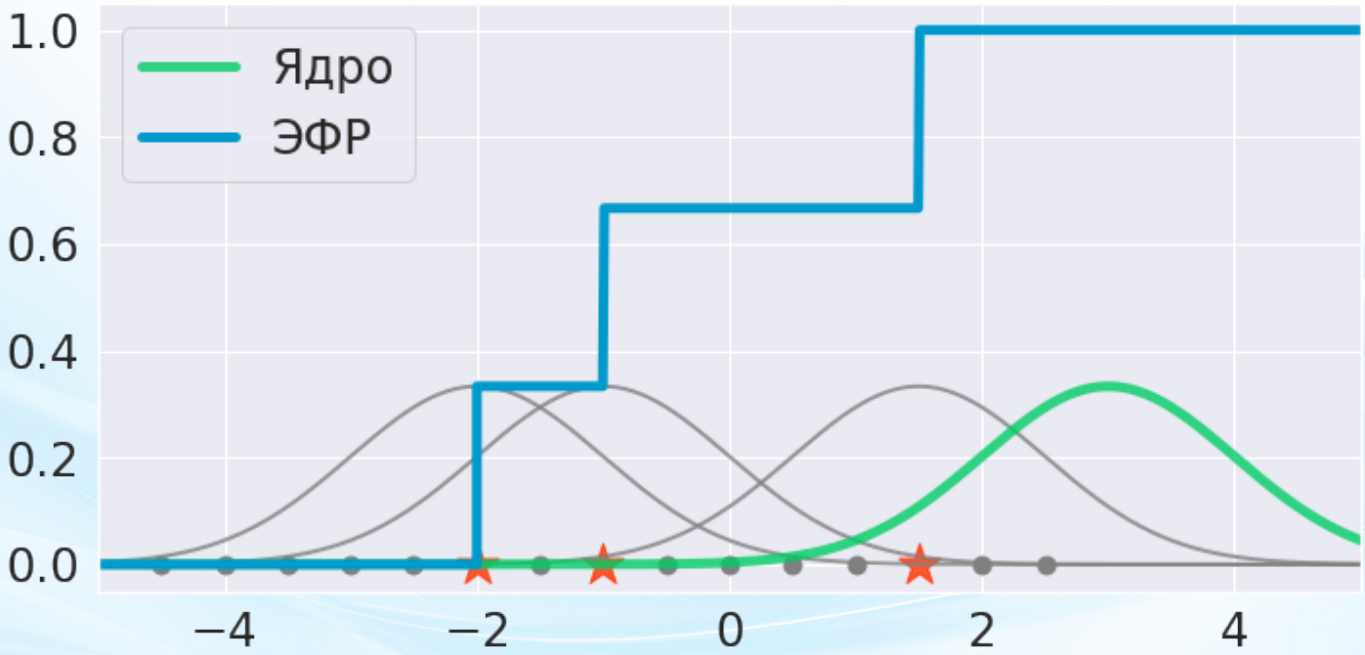


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

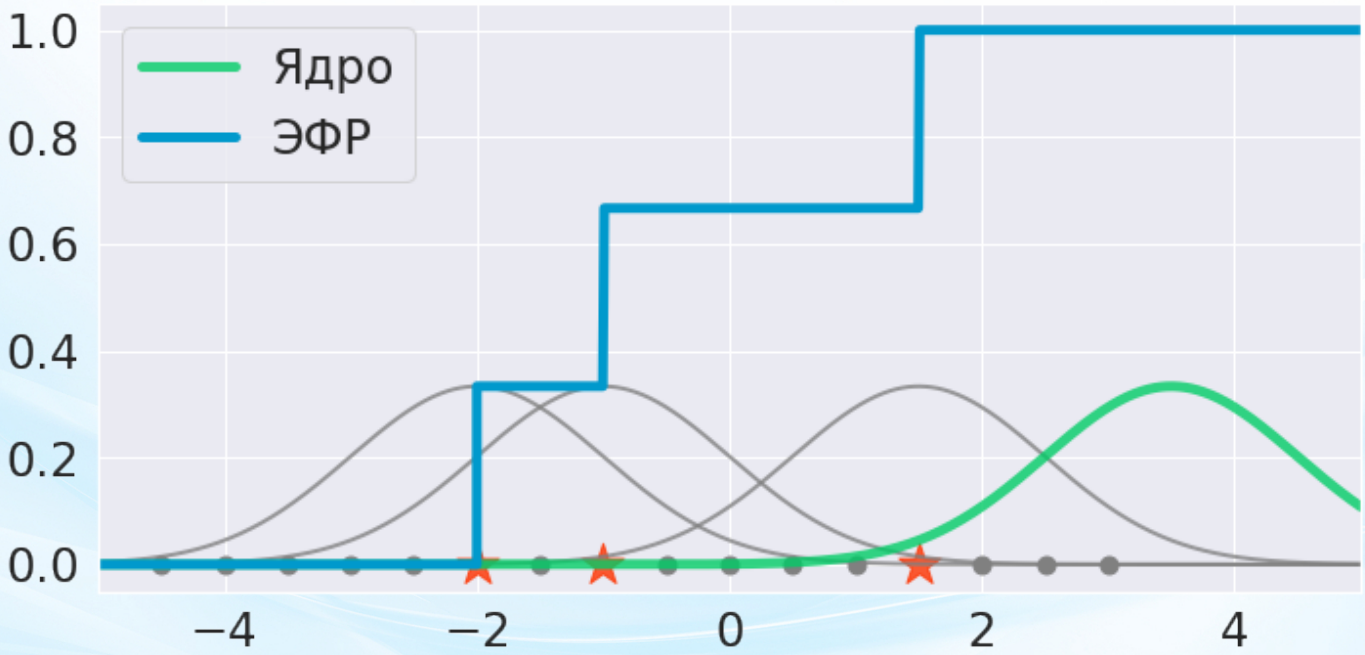


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

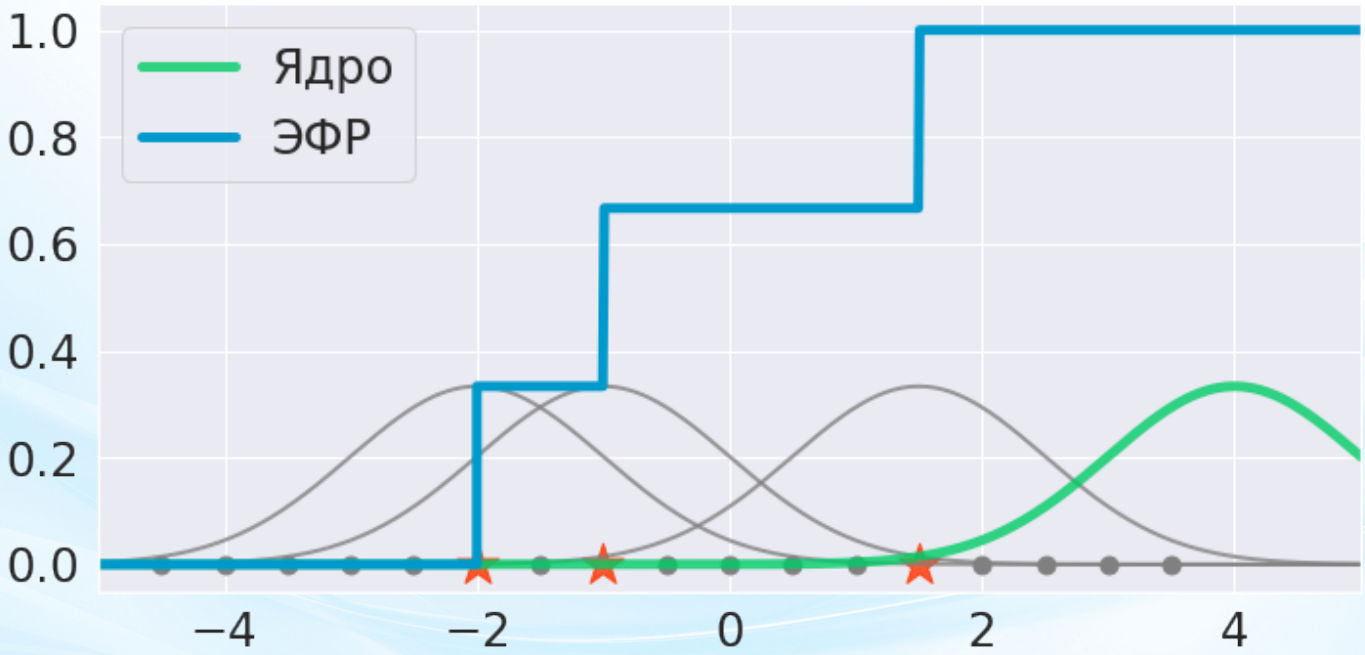


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

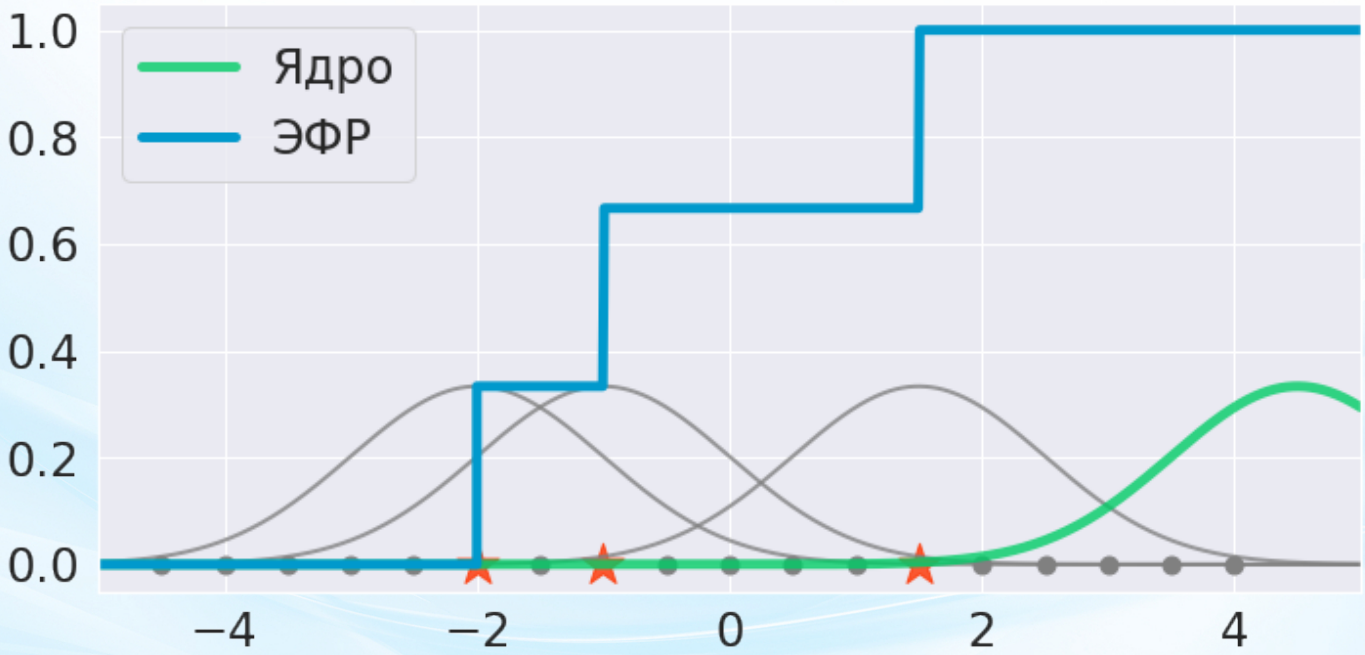


Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.



Связь с ЭФР

Утверждение

$\widetilde{\mathrm{P}}_h = \widehat{\mathrm{P}}_n * \mathrm{Q}(\cdot/h)$ — свертка ЭФР $\widehat{\mathrm{P}}_n$ и $\mathrm{Q}(\cdot/h)$,

то есть распределения Q с масштабом h.

Сходимость для случая $\mathscr{X} = \mathbb{R}$

Аналог теоремы Гливенко-Кантелли

$$\sup_{x \in \mathbb{R}} \left| \widetilde{F}_h(x) - F(x) \right| \xrightarrow{\text{п.н.}} 0,$$

где $\widetilde{F}_h(x)$ — функция распределения, соотв. плотности $\widetilde{p}_h(x)$.

Теорема об асимптотике

Пусть

1. $\alpha = \int\limits_{\mathbb{R}} q^2(y)dy < \infty$;
2. плотность $q(x)$ непрерывна и ограничена;
3. $h_n \to 0, nh_n \to \infty$ при $n \to \infty$.

Тогда ядерная оценка плотности представима в виде

$$\widetilde{p}_h(x) = p_h(x) + \frac{\xi_n(x)}{\sqrt{nh_n}},$$

где $p_h(x) = \mathsf{E}\widetilde{p}_h(x) \longrightarrow p(x)$,
$\xi_n(x) \xrightarrow{d} \mathcal{N}(0, \alpha p(x))$

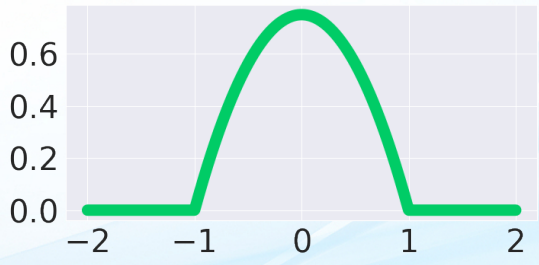
Каковы оптимальные h и q?

Рассмотрим минимизацию среднеквадратической ошибки:

$$\int_{\mathbb{R}} \mathsf{E}\left(\widetilde{p}_h(x) - p(x)\right)^2 dx \to \min_h$$

Тогда оптимальные параметры

- $h_n^* \sim n^{-1/5}$
- $q^*(x)$ — ядро Епаничнекова
- Скорость сходимости $\sim n^{-2/5}$



На практике полученными формулами пользоваться проблематично

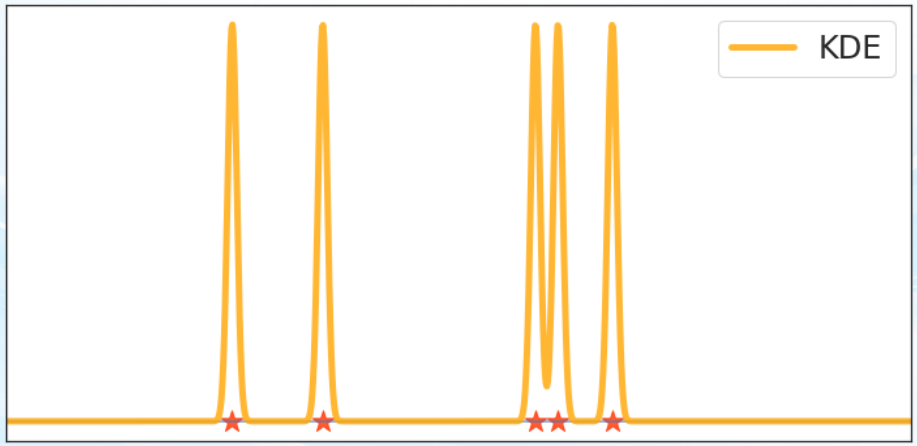
Подбор ширины ядра по выборке

Задача: подобрать оптимальную h для конкретной выборки.

Рассмотрим h как параметр и будем искать оценку макс. правд.

$$\prod_{i=1}^{n} \widetilde{p}_h(X_i) \longrightarrow \max_h$$

Этой задаче соответствует $h = 0$, что соответствует эмпирическому распределению. Это не то, что хочется.



Подбор ширины ядра по выборке

Leave-one-out оценка

Рассмотрим ядерную оценку плотности, исключив элемент X_i.

$$\widetilde{p}_h^{-i}(x) = \frac{1}{(n-1)h} \sum_{\substack{k=1 \\ k \neq i}}^{n} q\left(\frac{x - X_k}{h}\right),$$

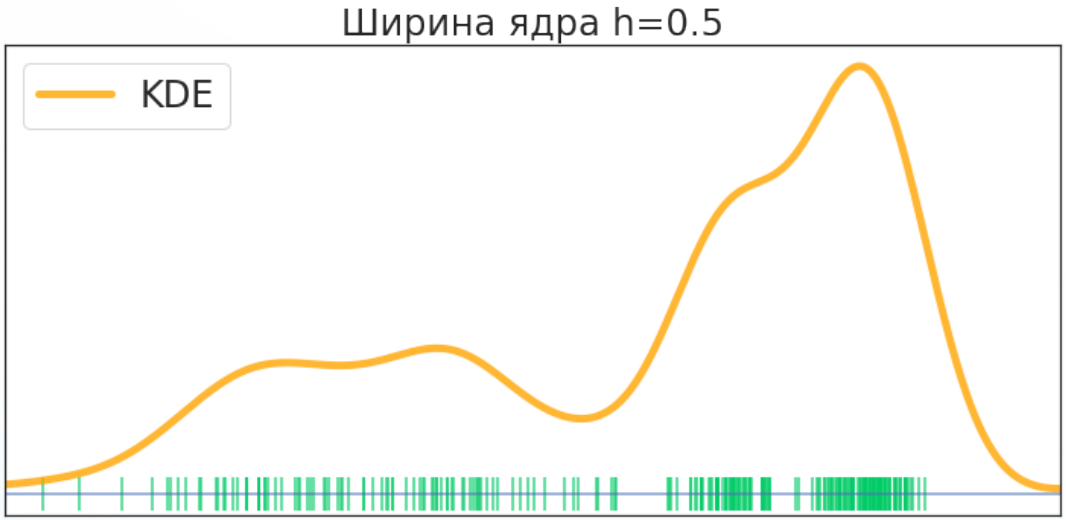
ширина ядра выбирается из максимизации функционала

$$F(h) = \log \prod_{i=1}^{n} \widetilde{p}_h^{-i}(X_i) = \sum_{i=1}^{n} \log \sum_{\substack{k=1 \\ k \neq i}}^{n} q\left(\frac{X_i - X_k}{h}\right) - n \log(n-1)h.$$

Поскольку h — одномерная величина,

максимум можно найти по сетке.

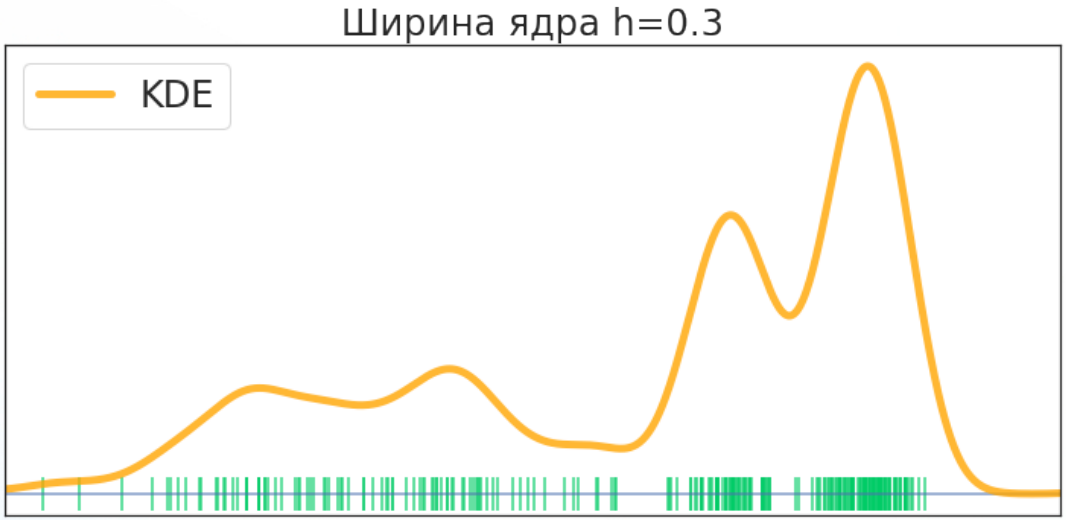
Подбор ширины ядра по выборке



Если в точки расположены где-то достаточно плотно, а где-то — сильно разреженно, то имеет смысл брать разную ширину ядра для разных точек.

Например, $h(x) = \|x - X_{(k)}\|$, где $X_{(k)}$ — k-й ближайший сосед для x.

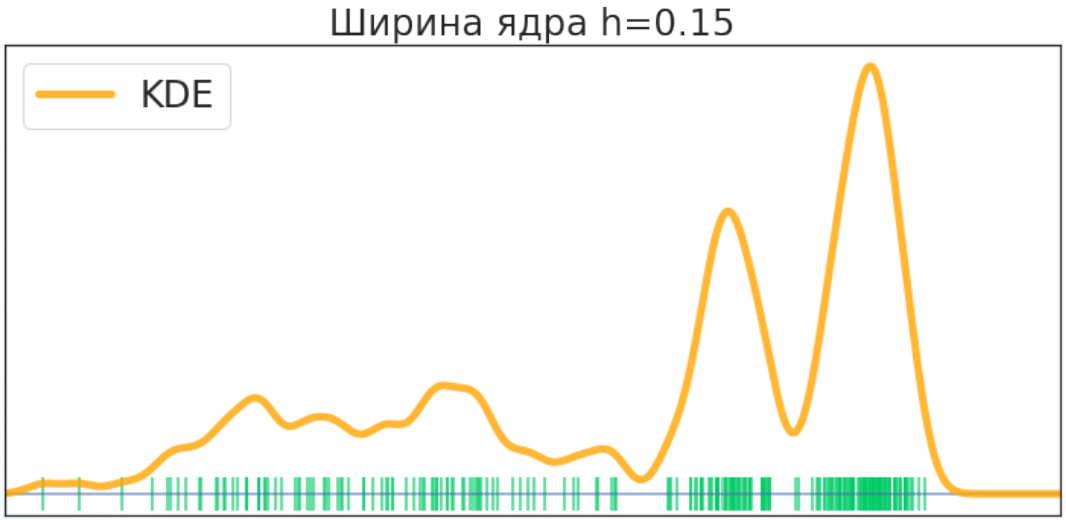
Подбор ширины ядра по выборке



Если в точки расположены где-то достаточно плотно, а где-то — сильно разреженно, то имеет смысл брать разную ширину ядра для разных точек.

Например, $h(x) = \|x - X_{(k)}\|$, где $X_{(k)}$ — k-й ближайший сосед для x.

Подбор ширины ядра по выборке



Если в точки расположены где-то достаточно плотно, а где-то — сильно разреженно, то имеет смысл брать разную ширину ядра для разных точек.

Например, $h(x) = \left\|x - X_{(k)}\right\|$, где $X_{(k)}$ — k-й ближайший сосед для x.

5.6 Обучение на основе ближайшего соседа

kNN

Приближенный поиск соседей

Метод локального усреднения

Метод ближайших соседей (kNN)

Пусть $\mathscr{X}$ — метрическое пространство.

$x_1, ..., x_n \in \mathscr{X}$ — обучающая выборка.

$Y_1, ..., Y_n$ — соответствующая целевая переменная.

Предположение:
свойства объекта меняются не сильно в его окрестности.

Тогда давайте смотреть на свойства k ближайших соседей.

Примеры.
Пусть $x \in \mathscr{X}$ — исследуемый объект.
$x_{(1)}, ..., x_{(k)}$ — k его соседей в порядке удаления от x.

1. *Классификация.*
   Предсказание — наиболее часто встречаемый класс.

2. *Регрессия.*
   Предсказание — усреднение отклика по соседям.

Взвешенный метод ближайших соседей

Пусть $x \in \mathscr{X}$ — исследуемый объект.

$x_{(1)}, ..., x_{(k)}$ — k его соседей в порядке удаления от x.

$Y_1, ..., Y_k$ — соответствующий отклик.

$w_1, ..., w_k$ — вклад k-го соседа, определяемый пользователем.

Способы определения веса:

- $w_j = 1 - j/k$ — зависящий от номера соседа;
- $w_j = \left\|x - x_{(j)}\right\|^{-1}$ — зависящий от расстояния до соседа.

$$\widehat{y}(x) = \arg\max_y \sum_{j=1}^{k} w_j I\{Y_j = k\} \text{ — классификация}$$

$$\widehat{y}(x) = \frac{\sum_{j=1}^{k} w_j Y_j}{\sum_{j=1}^{k} w_j} \text{ — регрессия}$$

Свойства

1. k — гиперпараметр модели;
2. Не редко на практике показывает хорошие результаты.
3. **Дорогое применение:**
   для каждого x результат вычисляется за $O(n \ln n)$.

Способы решения проблемы:

- Хранение данных в виде дерева: K-D Tree, Ball Tree.
- Приближенный поиск: Locality-sensitive hashing —

  вероятностный метод понижения размерности многомерных данных.
  Подбирает хеш-функций так, чтобы похожие объекты
  с высокой степенью вероятности попадали в одну корзину.

5.6 Обучение на основе ближайшего соседа

kNN

Приближенный поиск соседей

Метод локального усреднения

Locality-Sensitive Hashing (LSH)

Идея: если две точки расположены близко, то после некоторой проекции они тоже будут близко.

Пусть $h(x)$ — дискретная проекция. Будем искать такую проекцию, что для некоторых чисел $R_1 < R_2$ и $p_1 > p_2$ выполнено

1. если $\|x - y\| \leqslant R_1$, то $\mathrm{P}(h(x) = h(y)) \geqslant p_1$;
2. если $\|x - y\| \geqslant R_2$, то $\mathrm{P}(h(x) = h(y)) \leqslant p_2$;

Смысл условия:

если точки близки друг к другу, то с большой вероятностью они окажутся в одной корзине. Иначе — с малой вероятность.

Хотим выбрать такую h, что $p_1 \gg p_2$

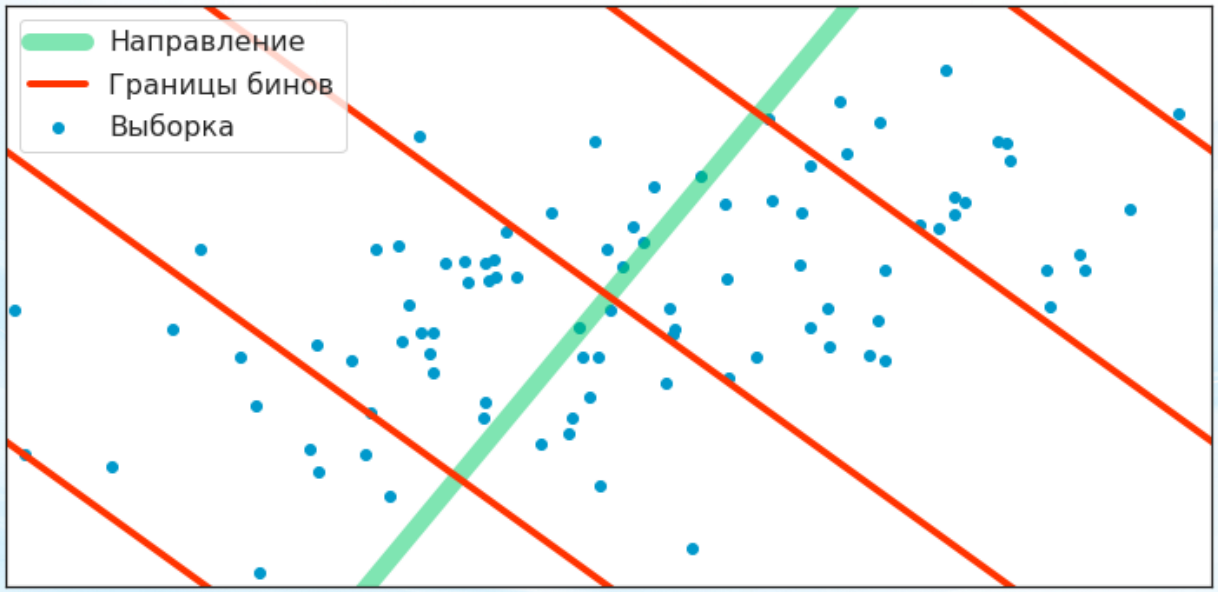
Выбор проекций

Рассмотрим функцию $h(x) = \big(h_1(x), ..., h_K(x)\big)$,

порожденную независимыми случайными проекциями

$$h_k(x) = \left\lfloor \frac{a_k^T x - b_k}{w} \right\rfloor,$$

где $a_k \sim \mathcal{N}(0, I_d)$, $b_k \sim U(0, w)$, w — ширина корзин.



Выбор проекций

Имеем:

Все пространство разделено на непересекающиеся пространства.

Поиск соседей:

В каждом подпространстве выполним

полный поиск ближайших соседей.

Влияние ширины корзин w**:**

при увеличении w увеличивается количество точек,

которые попадают в корзину. Соответственно, повышается точность

поиска, но и увеличивается требуемое время на поиск.

Комбинирование проекций

Пусть $\|x - y\| \leqslant R_1$ и пусть $\mathrm{P}(h_k(x) = h_k(y)) = \rho$.

Тогда из независимости $\mathrm{P}(h(x) = h(y)) = \rho^k$ — малое число.

Повторим процедуру определения бакетов L раз,

используя независимый набор случайных процекций $h^1(x), ..., h^L(x)$,

где $h^\ell(x) = \left(h_1^\ell(x), ..., h_K^\ell(x)\right)$.

Поиск соседей для x**:** ищем среди всех точек y,

для которых хотя бы для одного ℓ выполнено $h^\ell(x) = h^\ell(y)$.

Тогда $\mathrm{P}(\exists \ell : \ h^\ell(x) = h^\ell(y)) = 1 - \left(1 - \rho^k\right)^L$ — большое число.

Число L обычно определяется из допустимой вероятности ошибки:

если $\mathrm{P}(\forall \ell : \ h^\ell(x) \neq h^\ell(y)) \leqslant \delta$ то берем $L = \frac{\log \delta}{\log(1-\rho^k)}$.

5.6 Обучение на основе ближайшего соседа

kNN

Приближенный поиск соседей

Метод локального усреднения

Метод локального усреднения

$$\widehat{y}(x) = \sum_{i=1}^{n} w_i(x)\, Y_i \Bigg/ \sum_{i=1}^{n} w_i(x),$$

где $w_i(x) \geqslant 0$ убывает при удалении x от X_i.

Варианты:

1. $w_i(x) = I\{|x - X_i| \leqslant c\}$ — усреднение по окрестности x;
2. $w_i(x) = (c - |x - X_i|)^k \, I\{|x - X_i| \leqslant c\}$

   — взвешенное усреднение по окрестности x;
3. Усреднение по k ближайшим соседям;
4. Ядерная оценка (см. далее).

Ядерная оценка Надарая-Ватсона

$$w_i(x) = \frac{1}{h} q\left(\frac{x - X_i}{h}\right),$$

где q — ядро (симметричная плотность),

$h > 0$ — ширина ядра.

Вероятностная интерпретация

Пусть $X_1, ..., X_n$ случайны и $Y_i = f(X_i, \varepsilon_i)$ — отклик. Тогда

$\frac{1}{n}\sum_{i=1}^{n} w_i(x) = \frac{1}{nh}\sum_{i=1}^{n} q\left(\frac{x-X_i}{h}\right)$ — ядерная оценка плотности X;

$\frac{1}{n}\sum_{i=1}^{n} w_i(x) Y_i = \frac{1}{nh}\sum_{i=1}^{n} Y_i \cdot q\left(\frac{x-X_i}{h}\right)$ — "ядерное мат. ожид." $\mathsf{E}Y$;

$\widehat{y}(x) = \sum_{i=1}^{n} w_i(x) Y_i \Big/ \sum_{i=1}^{n} w_i(x)$ — "ядерное УМО" $\mathsf{E}(Y|X)$.

Ядерная оценка: теорема о сходимости

Пусть

1. $\int\limits_{\mathbb{R}} |q(y)|\,dy < \infty$;
2. $yq(y) \to 0$ при $|y| \to \infty$;
3. $\mathsf{E}Y^2 < \infty$;
4. $h_n \to 0, nh_n \to \infty$ при $n \to \infty$.

Тогда $\widehat{y}(x) \xrightarrow{\mathsf{P}} y(x)$ в точках непрерывности функции $f(x)$, плотности $p_X(x)$ и условной дисперсии $\sigma^2(x) = \mathsf{D}(Y|X = x)$, если при этом $p(x) > 0$.

Наилучшая скорость сходимость квадратичного риска достигается при $h \sim n^{-1/5}$.

Ядерная оценка: выбор ширины ядра

Функционал вида leave one out:

$$F(h) = \sum_{i=1}^{n} \left(Y_i - \widehat{y}_{-i}(x_i)\right)^2,$$

где $\widehat{y}_{-i}(x)$ — ядерная оценка, построенная по выборке, из которой было исключено i-е наблюдение.

Утверждение

$$F(h) = \sum_{i=1}^{n} \left(Y_i - \widehat{y}(x_i)\right)^2 \Bigg/ \left(1 - \frac{q(0)}{\sum_{k=1}^{n} q\left(\frac{x_i - x_k}{h}\right)}\right)$$

Выбор h: $F(h) \to \min\limits_{h}$

Ядерная оценка: доверительная лента

Предположения:

$\mu(x)$ — ожидаемый отклик;

$Y_i = \mu(x_i) + \varepsilon_i$ — наблюдаемый отклик, $\mathrm{E}\varepsilon_i = 0, \mathrm{D}\varepsilon_i = \sigma^2$;

q — гауссовское ядро.

Доверительная лента уровня доверия α:

$$\Big(\widehat{y}(x) - z_{3h}\delta(x),\ \ \widehat{y}(x) + z_{3h}\delta(x)\Big)$$

$$\delta(x) = \widehat{\sigma}\sqrt{\sum_{i=1}^{n} w^2(x_i)}\Big/\sum_{i=1}^{n} w(x_i), \qquad p = \frac{1+\alpha^{1/3h}}{2},$$

$$\widehat{\sigma}^2 = \frac{1}{2(n-1)}\sum_{i=1}^{n-1}(Y_{i+1} - Y_i)^2$$

Локальная линейная регрессионная модель

Модель $f(x) = x^T\theta(x)$, т.е. для каждого x свои коэффициенты.

Для каждого предсказания применяется взвешенный МНК:

$$\sum_{i=1}^{n} w_i(x)\left(Y_i - X_i^T\theta(x)\right)^2 \longrightarrow \min_{\theta(x)},$$

где $w_i(x) = \frac{1}{h(x)}q\left(\frac{x-X_i}{h(x)}\right)$,

$h(x) = \left\|x - X_{(k)}\right\|$, где $X_{(k)}$ — k-й ближайший сосед для x.

Взвешенный МНК:

$$\widehat{\theta}(x) = \left(X^T W(x) X\right)^{-1} X^T W(x) Y,$$

$$W(x) = diag\left(w_1^2(x), ..., w_n^2(x)\right).$$

5.7 Анализ зависимостей

Анализ зависимостей

Даны **парные** выборки:

$X = (X_1, ..., X_n)$

$Y = (Y_1, ..., Y_n)$

Задачи:

- Зависимы ли выборки?

  H_0: выборки независимы vs. H_1: выборки зависимы

- Количественная оценка степени

  неслучайности их совместного изменения.

5.7 Анализ зависимостей

Коэффициенты корреляции

Таблицы сопряженности 2×2

Таблицы сопряженности, общий случай

Коэффициент корреляции

Пусть ξ, η — случайные величины.

$$corr(\xi, \eta) = \frac{cov(\xi, \eta)}{\sqrt{D\xi D\eta}} \text{ — коэффициент корреляции}$$

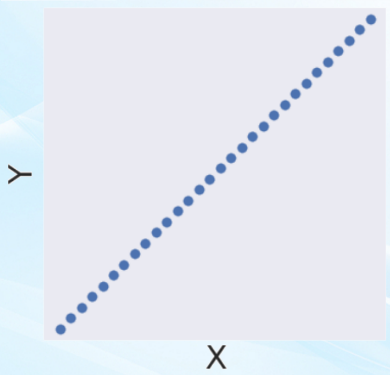
Свойства:

- $|corr(\xi, \eta)| \leqslant 1$;
- $|corr(\xi, \eta)| = 1 \Leftrightarrow \xi$ и η линейно зависимы п.н.;
- ξ и η независимы $\rightarrow corr(\xi, \eta) = 0$. Обратное не верно;
- Является мерой линейной зависимости.

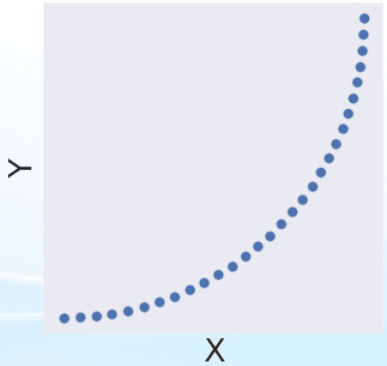
Коэффициент корреляции Пирсона

Метод подстановки: подставим в $corr(X_1, Y_1)$ эмпир. распр.

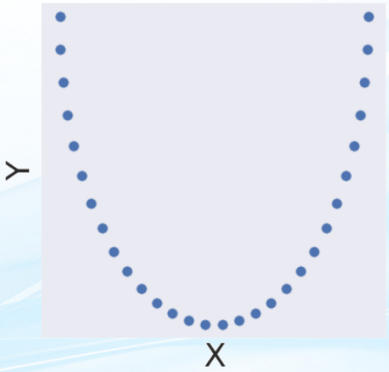
$$\widehat{\rho} = \frac{cov_{\mathsf{P}^*}(X_1, Y_1)}{\sqrt{\mathsf{D}_{\mathsf{P}^*} X_1 \mathsf{D}_{\mathsf{P}^*} Y_1}} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \sum_{i=1}^{n} (Y_i - \overline{Y})^2}}$$



$\widehat{\rho} = 1$
$pvalue = 0$



$\widehat{\rho} = 0.91$
$pvalue = 2 \cdot 10^{-12}$



$\widehat{\rho} = 0$
$pvalue = 1$

Коэффициент корреляции Пирсона

Свойства:

- $|\widehat{\rho}| \leqslant 1$;
- $|\widehat{\rho}| = 1 \Leftrightarrow$ точки лежат на одной прямой;
- Работает только для нормальных выборок

  для линейной зависимости;
- Не устойчив к выбросам.
- H_0: выборки независимы

  Если H_0 верна и выборки нормальные, то

$$T(X, Y) = \frac{\widehat{\rho}\sqrt{n-2}}{\sqrt{1-\widehat{\rho}^2}} \sim T_{n-2}.$$

Критерий $\{|T(X, Y)| > t_{n-2,1-\alpha/2}\}$.

Коэффициент корреляции Спирмена

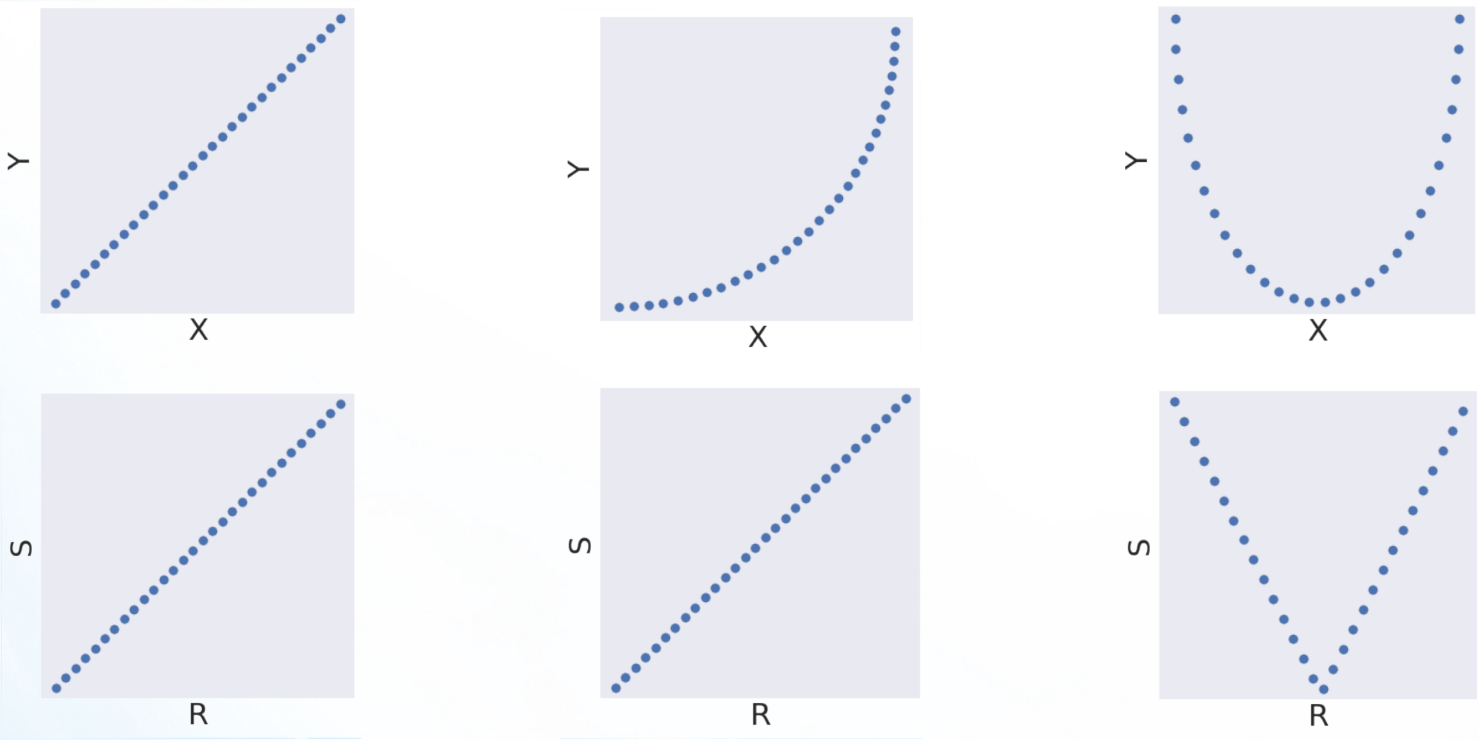
Пусть R_i — ранг наблюдения X_i в выборке X, то есть $X_{(R_i)} = X_i$.

Пусть S_i — ранг наблюдения Y_i в выборке Y, то есть $Y_{(S_i)} = Y_i$.

X_i	7.3	2.2	0.3	6.2	1.6	6.2	9.6
R_i	6	3	1	4.5	2	4.5	7

К.к. Спирмена = к.к. Пирсона по выборкам $(R_1, ..., R_n)$ и $(S_1, ..., S_n)$.

$$\rho_S = \frac{\sum_{i=1}^{n} \left(R_i - \overline{R}\right)\left(S_i - \overline{S}\right)}{\sqrt{\sum_{i=1}^{n} \left(R_i - \overline{R}\right)^2 \sum_{i=1}^{n} \left(S_i - \overline{S}\right)^2}} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (R_i - S_i)^2$$



$\rho_S = 1$
$pvalue = 0$

$\rho_S = 1$
$pvalue = 0$

$\rho_S = 0$
$pvalue = 1$

Свойства:

- $|\rho_S| \leqslant 1$, причем $|\rho_S| = 1 \Leftrightarrow$ точки лежат на монотонной кривой;
- Если H_0 верна, то $E\rho_S = 0, D\rho_S = \frac{1}{n-1}$;
- Если H_0 верна, то $\rho_S / \sqrt{D\rho_S} \xrightarrow{d_0} \mathcal{N}(0,1)$.

  Критерий $\{|\rho_S / \sqrt{D\rho_S}| > z_{1-\alpha/2}\}$;
- Устойчив к выбросам.

Коэффициент корреляции Кендалла

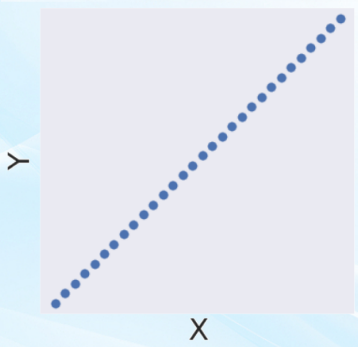
Пары (X_i, Y_i) и (X_j, Y_j) согласованы, если

$$\operatorname{sign}(X_i - X_j)\operatorname{sign}(Y_i - Y_j) = 1.$$

Пусть S — число согласованных пар,

R — число несогласованных.

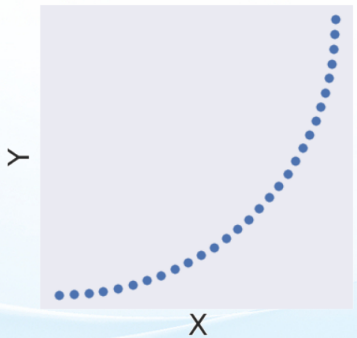
$$\tau = \frac{S - R}{S + R} = 1 - \frac{4}{n(n-1)}R$$



$S = \frac{n(n-1)}{2}, R = 0$

$\tau = 1$

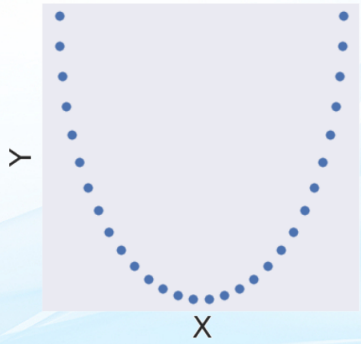
$pvalue = 0$



$S = \frac{n(n-1)}{2}, R = 0$

$\tau = 1$

$pvalue = 0$



$S = \frac{n(n-1)}{4}, R = \frac{n(n-1)}{4}$

$\tau = 0$

$pvalue = 1$

Коэффициент корреляции Кендалла

Свойства:

- $|\tau| \leqslant 1$;
- $|\tau| = 1 \Leftrightarrow$ точки лежат на монотонной кривой;
- Если H_0 верна, то $E\tau = 0, D\tau = \frac{2(2n+5)}{9n(n-1)}$;
- Если H_0 верна, то $\tau/\sqrt{D\tau} \xrightarrow{d_0} \mathcal{N}(0,1)$.

  Критерий $\{|\tau/\sqrt{D\tau}| > z_{1-\alpha/2}\}$;
- Если H_0 верна, то $corr(\rho_S, \tau) = \frac{2n+2}{\sqrt{4n^2+10n}}$;
- Менее чувствителен к большим различиям между рангами, чем ρ_S;
- Точнее оценивается по выборкам малых размеров.

Еще раз формулы

Пирсон:

$$\widehat{\rho} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \sum_{i=1}^{n} (Y_i - \overline{Y})^2}}$$

Спирмен: R и S — ранги наблюдений в выборках X и Y

$$\rho_S = \frac{\sum_{i=1}^{n} (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n} (R_i - \overline{R})^2 \sum_{i=1}^{n} (S_i - \overline{S})^2}} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (R_i - S_i)^2$$

Кендалл: S — число соглас. пар, а R — число несоглас.

$$\tau = \frac{S - R}{S + R} = 1 - \frac{4}{n(n-1)} R$$

5.7 Анализ зависимостей

Коэффициенты корреляции

Таблицы сопряженности 2×2

Таблицы сопряженности, общий случай

Осенний семестр (2018)

Результаты решения теор. задачи:

Семинар	I	II	III	IV
Справились	0	5	3	2
Не справились	8	2	4	5

Факты:

1. Случайное разбиение на группы;
2. Задача на алгоритмы и методы оптимизации

   $\Longrightarrow$ не должна зависеть от семинариста по статистике;
3. Дедлайн перед семинаром;
4. На первом семинаре задача была разобрана.

Хотим воспользоваться методом проверки статистических гипотез.

Какие взять H_0 и H_1?

Презумпция невиновности: не виновны пока нет доказательств.

H_0: решаемость задачи не зависит от семинара

H_1: решаемость задачи зависит от семинара

Упростим данные

Разбиралась ли задача до семинара?	Нет	Да
Справились	0	10
Не справились	8	11

Математическая формулировка

Даны **парные** выборки

$X = (X_1, ..., X_n) \sim Bern(p_1)$

$Y = (Y_1, ..., Y_n) \sim Bern(p_2)$

H_0: выборки X и Y независимы

H_1: выборки X и Y зависимы

	$Y_i = 0$	$Y_i = 1$	$\sum$
$X_i = 0$	a	b	$a + b$
$X_i = 1$	c	d	$c + d$
$\sum$	$a + c$	$b + d$	n

Вероятность таблицы с фиксированными суммами задается гипергеометрическим распределением:

$$P(table) = \frac{C_{a+b}^{a} C_{c+d}^{c}}{C_n^{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}.$$

p-value = сумма вероятностей по всем возможным вариантам таблицы с такими же суммами по строкам и столбцам, имеющим вероятность не больше, чем у полученной таблицы.

Точный тест Фишера

Особенности:

1. Критерий является точным (неасимптотическим);
2. Вычислительно затратный $\Rightarrow$ используется для малых выборок;
3. Что в сложных случаях? Увидим далее!

Пример про задачу 7 из ДЗ-12

Разбиралась ли задача до семинара?	Нет	Да
Справились	0	10
Не справились	8	11

```
scipy.stats.fisher_exact([[0, 8], [10, 11]])
```

вернет p-value $= 0.0265$.

Вывод: гипотеза о независимости отвергается.

Численные характеристики взаимосвязи

Даны **парные** выборки

$X = (X_1, ..., X_n) \sim Bern(p_1)$

$Y = (Y_1, ..., Y_n) \sim Bern(p_2)$

	$Y_i = 0$	$Y_i = 1$	$\sum$
$X_i = 0$	a	b	$a + b$
$X_i = 1$	c	d	$c + d$
$\sum$	$a + c$	$b + d$	n

$$Q = \frac{ad - bc}{ad + bc}$$ — **коэффициент ассоциации**

$$V = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$ — **коэффициент контингенции**

В обоих случаях:

$0 \Longrightarrow$ полное отсутствие взаимосвязи

$\pm 1 \Longrightarrow$ полная связь

Определение числа наблюдений (при $a + b = c + d$)

Задаем:

α — ур. значимости

β — мощность

$$\left.\begin{array}{l} p_1 = a/b \\ p_2 = c/d \end{array}\right\} \text{значимый эффект}$$

	$Y_i = 0$	$Y_i = 1$	$\sum$
$X_i = 0$	a	b	$a + b$
$X_i = 1$	c	d	$c + d$
$\sum$	$a + c$	$b + d$	n

Тогда необходимое число наблюдений в каждой строке равно

$$K \Big/ (\arcsin\sqrt{p_1} - \arcsin\sqrt{p_2})^2$$

K	$\beta = 0.8$	$\beta = 0.9$	$\beta = 0.99$
$\alpha = 0.05$	12885	17250	30161
$\alpha = 0.01$	16474	21369	35537
$\alpha = 0.001$	19172	24426	43945

5.7 Анализ зависимостей

Коэффициенты корреляции

Таблицы сопряженности 2×2

Таблицы сопряженности, общий случай

Категориальные признаки

Даны **парные** выборки

$X = (X_1, ..., X_n)$, причем $X_i \in \{1, ..., k_1\}$

$Y = (Y_1, ..., Y_n)$, причем $Y_i \in \{1, ..., k_2\}$

Таблица сопряженности:

	1	...	j	...	k_2	$\sum$
1	n_{11}	...	n_{1j}	...	n_{1k_2}	$n_{1\bullet}$
...	...		...		...	...
i	n_{i1}	...	n_{ij}	...	n_{ik_2}	$n_{i\bullet}$
...	...		...		...	...
k_1	$n_{k_1 1}$	...	$n_{k_1 j}$	...	$n_{k_1 k_2}$	$n_{k_1 \bullet}$
$\sum$	$n_{\bullet 1}$	...	$n_{\bullet j}$	...	$n_{\bullet k_2}$	n

Элементы таблицы:

$n_{ij} = \#\{s \mid X_s = i, Y_s = j\}$

$n_{i\bullet} = \#\{s \mid X_s = i\}$

$n_{\bullet j} = \#\{s \mid Y_s = j\}$

Вероятностные модели

Случай 1: X и Y случайны.

$\pi_{ij} = \mathrm{P}(X_1 = i, Y_1 = j) \implies \{\pi_{ij}\}_{ij}$ — совместное распределение;

$\pi_{i\bullet} = \mathrm{P}(X_1 = i) \implies \mathrm{P} = \{\pi_{i\bullet}\}_i$ — распределение X;

$\pi_{\bullet j} = \mathrm{P}(Y_1 = j) \implies \mathrm{Q} = \{\pi_{\bullet j}\}_j$ — распределение Y;

Определение: X и Y **независимы**, если $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \quad \forall i, j$.

Случай 2: X неслучаен, Y случаен.

$\implies$ суммы по строкам $n_{i\bullet}$ фиксированы.

$\pi_{j|i} = \mathrm{P}_i(Y_1 = j)$ — вероятность события $Y_1 = j$ если $X_1 = i$;

$\mathrm{P}_i = \{\pi_{j|i}\}_i$ — распределение Y если $X_1 = i$, т.е. X — параметр.

Определение: X и Y **независимы**, если $\mathrm{P}_1 = \ldots = \mathrm{P}_{k_1}$.

Критерий хи-квадрат (обе вер. модели)

H_0: выборки X и Y независимы

$$\chi^2(X,Y) = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}}$$

Если H_0 верна, то $\chi^2(X,Y) \xrightarrow{d} \chi^2_{(k_1-1)(k_2-1)}$

$\Longrightarrow$ критерий $\left\{\chi^2(X,Y) > \chi^2_{(k_1-1)(k_2-1),1-\alpha}\right\}$.

Условия применимости:

1. $n \geqslant 40$;
2. $\frac{n_{i\bullet}n_{\bullet j}}{n} < 5$

   не более чем в 5% ячеек.

Коэффициент корреляции Крамера

$$\varphi_C(X,Y) = \sqrt{\frac{\chi^2(X,Y)}{n(\min(k_1,k_2)-1)}}$$

$0 \Longrightarrow$ полное отсутствие взаимосвязи;

$1 \Longrightarrow$ совпадение переменных.

Пример

	Вернул кредит	Не вернул кредит
Android	850	870
iOS	380	410

H_0: зависимости возвращаемости кредита от типа ОС нет;

H_1: зависимость есть.

Критерий хи-квадрат: $\chi^2(X, Y) = 0.325,\ pvalue = 0.569,$

Численные характеристики: $\varphi_C(X, Y) = 0.008,\ Q = 0.026,\ V = 0.012$


ВСЁ!