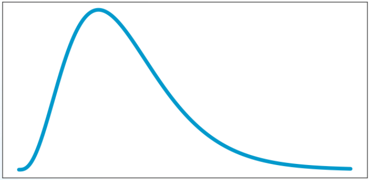
Статистика ФИВТ ПМИ

Прикладной поток

Лекция 9

Что численно характеризует симметричность распределения?



Перебираем моменты:

$a = \mathrm{E}X$ — отвечает за среднее значение

$\sigma^2 = \mathrm{D}X = \mathrm{E}(X - a)^2$ — отвечает за разброс значений

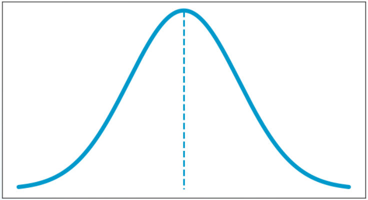
Идем дальше...

$$\kappa = \frac{1}{\sigma^3}\mathrm{E}(X - a)^3$$ — коэффициент асимметрии (skewness)

мера симметричности распределения

$$\gamma = \frac{1}{\sigma^4}\mathrm{E}(X - a)^4 - 3$$ — коэффициент эксцесса (kurtosis)

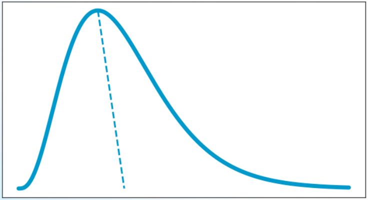
мера остроты пика распределения



$X \sim \mathcal{N}(0,1)$

$\kappa = 0$

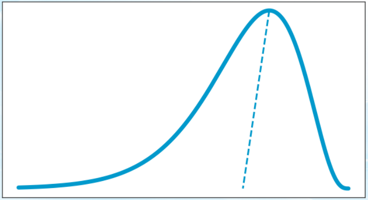
распределение симметрично



$X \sim \Gamma(1/2, 4)$

$\kappa = 1$

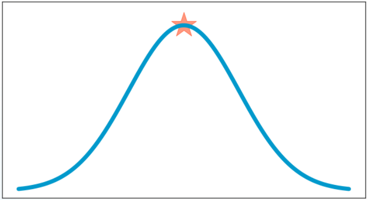
правый хвост тяжелее левого



$-X \sim \Gamma(1/2, 4)$

$\kappa = -1$

левый хвост тяжелее правого

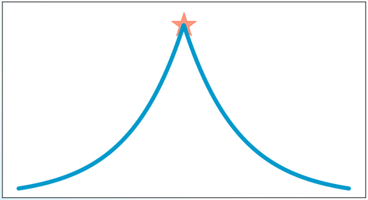


$X \sim \mathcal{N}(0,1)$

$\gamma = 0$

сглаженный пик

(тройка вычитается чтобы тут $\gamma = 0$)



$X \sim \mathit{Laplace}$

$\gamma = 3$

острый пик

$-X \sim U[0,1]$

$\gamma = -1.2$

ровный пик

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

Посчитаем оценку методом подстановки

$$\widehat{\kappa} = \frac{1}{\widehat{\sigma}^3} \int\limits_{\mathbb{R}} (x - \widehat{a})^3 d\widehat{F}_n(x) = \frac{1}{n\widehat{\sigma}^3} \sum_{i=1}^{n} (X_i - \widehat{a})^3.$$

Хотелось бы получить доверительный интервал для значения κ...

Пусть $\mathscr{P} = \{\mathsf{P}_\theta \mid \theta \in \Theta\}$ — параметрический случай.

Применяем многомерную ЦПТ:

$$\sqrt{n}\left(\begin{pmatrix} \overline{X} \\ \overline{X^2} \\ \overline{X^3} \end{pmatrix} - \begin{pmatrix} \mathsf{E}_\theta X_1 \\ \mathsf{E}_\theta X_1^2 \\ \mathsf{E}_\theta X_1^3 \end{pmatrix}\right) \xrightarrow{d_\theta} \mathcal{N}(0, \Sigma(\theta)), \qquad \text{где } \Sigma(\theta) = \ldots$$

Применяем дельта метод с функцией $\tau(x, y, z) = \frac{z - 3xy^2 + 2x^3}{y - x^2}$.

Далее берем производные, перемножаем матрицы... Что за жесть...

сколько задач на дельта метод можно решить

Посмотрим подробнее на жесть:

1. Посчитать первые 6 моментов, посчитать матрицу ковариаций;
2. Найти функцию для применения дельта-метода;
3. Взять производные;
4. Перемножить матрицы.

Три вопроса "А если ...":

1. А если лень?
2. А если это срочная бизнесовая задача,

   а не на сдачу семинаристам по статистике?
3. А если нет никакого параметрического семейства?

   То есть рассматривается непараметрический случай. Упс...

4.3. Бутстреп

Постановка задачи

$X_1, ..., X_n$ — выборка из неизв. распр. P;

$T(X_1, ..., X_n)$ — некоторая статистика;

$v = V(T(X_1, ..., X_n)) = G(\mathrm{P})$ — функционал, значение которого требуется оценить;

$\widehat{v} = G(\widehat{\mathrm{P}}_n)$ — оценка методом подстановки.

В примере выше:

$T(X_1, ..., X_n) = \widehat{\varkappa}$ — оценка коэфф. асимметрии

$v = \mathrm{D}\widehat{\varkappa}$ — дисперсия оценки коэфф. асимметрии

$\widehat{v} = \mathrm{D}_{\widehat{\mathrm{P}}_n}\widehat{\varkappa}$ — оценка дисперсии оценки коэфф. асимметрии

Пример: оценка дисперсии

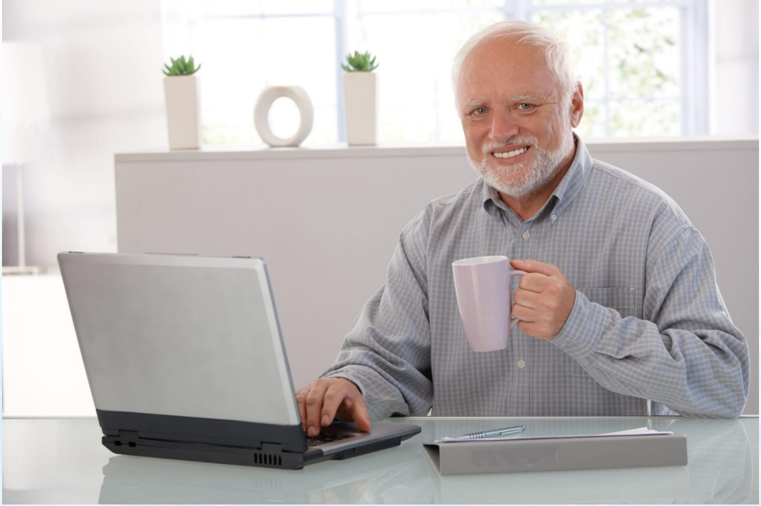
Дисперсия статистики

$$V(T(X_1,\dots,X_n)) = \mathrm{D}\,T(X_1,\dots,X_n) =$$
$$= \int\limits_{\mathscr{X}^n} T^2(x_1,\dots,x_n)dF(x_1)\dots dF(x_n) - \left(\int\limits_{\mathscr{X}^n} T(x_1,\dots,x_n)dF(x_1)\dots dF(x_n)\right)^2.$$

Оценка методом подстановки имеет вид

$$\widehat{v} = \mathrm{D}_{\widehat{\mathrm{P}}_n} T(X_1,\dots,X_n) =$$
$$= \int\limits_{\mathscr{X}^n} T^2(x_1,\dots,x_n)d\widehat{F}_n(x_1)\dots d\widehat{F}_n(x_n) - \left(\int\limits_{\mathscr{X}^n} T(x_1,\dots,x_n)d\widehat{F}_n(x_1)\dots d\widehat{F}_n(x_n)\right)^2 =$$
$$= \frac{1}{n^n}\sum_{i_1=1}^{n}\dots\sum_{i_n=1}^{n} T^2(X_{i_1},\dots,X_{i_n}) - \left(\frac{1}{n^n}\sum_{i_1=1}^{n}\dots\sum_{i_n=1}^{n} T(X_{i_1},\dots,X_{i_n})\right)^2,$$

Нужно совершить порядка n^n операций!!!



Решение проблемы

Монте-Карло!!!

Метод бутстрепа

Идея: приближенное вычисление $\widehat{v}$ методом Монте-Карло.

Этап 1. Генерация выборки из эмп. распределения $\widehat{\mathsf{P}}_n$.

Рассмотрим реализацию выборки $X_1(\omega) = x_1, ..., X_n(\omega) = x_n$.
Тогда реализацией $\widehat{\mathsf{P}}_n$ явл. распределение $U\{x_1, ..., x_n\}$.
(с учетом повторений)

Генерация случ. величины из $\widehat{\mathsf{P}}_n$:
выбор случайного элемента из мн-ва $\{X_1, ..., X_n\}$

Генерация выборки $X_1^*, ..., X_n^*$ из $\widehat{\mathsf{P}}_n$:
упоряд. выбор с возвращением n элементов из мн-ва $\{X_1, ..., X_n\}$.

Другой вид записи:

1. $i_1, ..., i_n \sim U\{1, ..., n\}$.
2. $X^* = (X_1^*, ..., X_n^*) = (X_{i_1}, ..., X_{i_n})$ — бутстрепная выборка.

Метод бутстрепа

Этап 2.

Процедуру генерации выборок повторить B раз:

$X_b^* = (X_{b1}^*, ..., X_{bn}^*)$, где $1 \leqslant b \leqslant B$.

Далее по каждой выборке посчитаем значение статистики T,

получив выборку значений $T_1^* = T(X_1^*), ..., T_B^* = T(X_B^*)$.

Этап 3.

Полученную выборку использовать для аппроксимации значения

оценки, которая называется *бутстрепной оценкой*.

Например, бутстрепная оценка дисперсии имеет вид

$$\widehat{v}_{boot} = \frac{1}{B}\sum_{b=1}^{B} {T_b^*}^2 - \left(\frac{1}{B}\sum_{b=1}^{B} T_b^*\right)^2,$$

Схема метода бутстрепа

$X = (X_1, \ldots, X_n)$ — выборка

$T(X_1, \ldots, X_n)$ — статистика

Задача: оценить распределение $T(X)$ или функционал $V(T(X))$.

$$
\left.
\begin{array}{lcl}
X_{11}^*, \ \ldots, \ X_{1n}^* & \longrightarrow & T(X_1^*) \\
 & & \ldots \\
X_{b1}^*, \ \ldots, \ X_{bn}^* & \longrightarrow & T(X_b^*) \\
 & & \ldots \\
X_{B1}^*, \ \ldots, \ X_{Bn}^* & \longrightarrow & T(X_B^*)
\end{array}
\right\} v_{boot} \text{ — бутстрепная оценка } v = V(T(X))
$$

Пример

$x = (5, 1, 3, 6, 4)$ — реализация выборки

$T(X_1, X_2, X_3, X_4, X_5) = \overline{X}$ — статистика

$T(5, 1, 3, 6, 4) = 3.8$ — реализация статистика

Задача: оценить дисперсию статистики, т.е. $v = V(T(X)) = \mathrm{D}\,T(X)$.

$$
\left.
\begin{array}{ccc}
5,\ 4,\ 3,\ 4,\ 6 & \longrightarrow & 4.4 \\
3,\ 1,\ 4,\ 6,\ 5 & \longrightarrow & 3.8 \\
6,\ 5,\ 6,\ 1,\ 6 & \longrightarrow & 4.8 \\
4,\ 1,\ 5,\ 6,\ 4 & \longrightarrow & 4.0 \\
1,\ 1,\ 4,\ 6,\ 5 & \longrightarrow & 3.4 \\
6,\ 4,\ 1,\ 5,\ 5 & \longrightarrow & 4.2 \\
6,\ 5,\ 6,\ 3,\ 6 & \longrightarrow & 5.2
\end{array}
\right\} \quad v_{boot} = 0.317
$$

Зоопарк: оценить дисперсию выборочного среднего

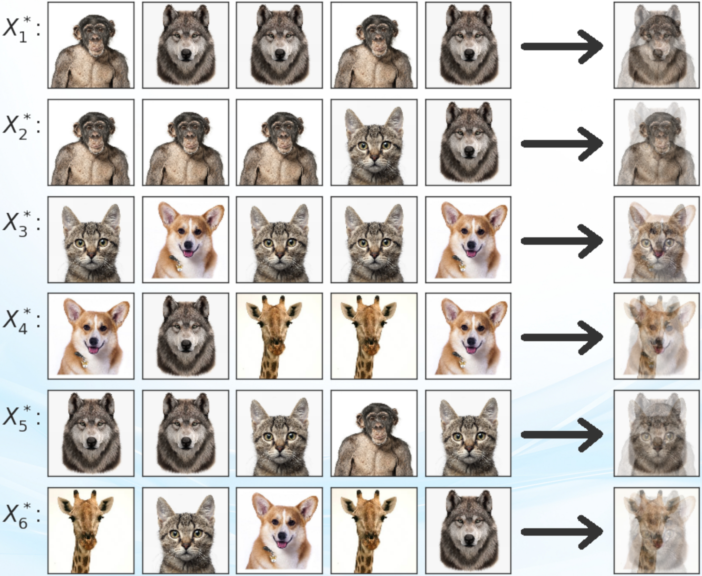
Выборка:

Задача:

для каждого пикселя и каждого цветового канала

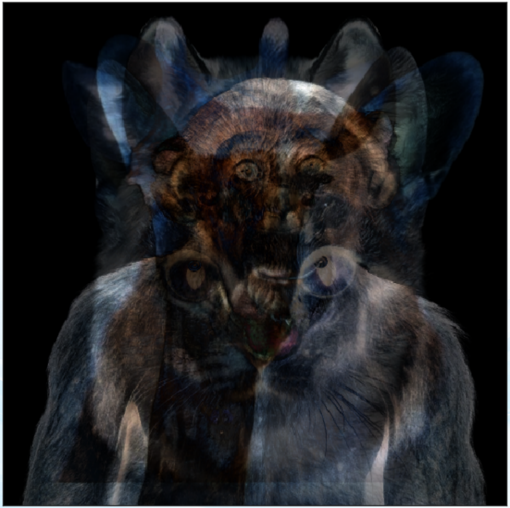
оценить дисперсию выборочного среднего.

Зоопарк: оценить дисперсию выборочного среднего



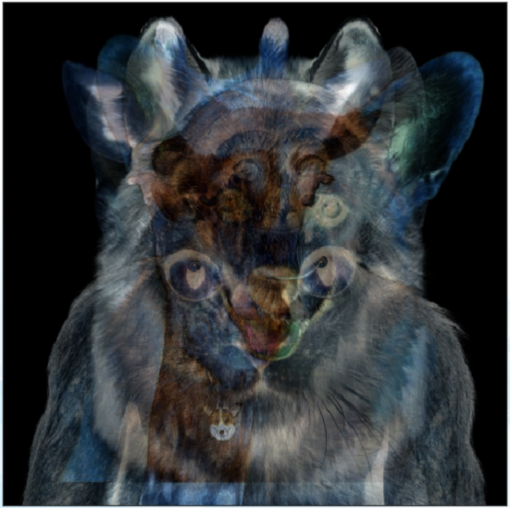
Зоопарк: оценить дисперсию выборочного среднего

Дисперсия по бутстрепной выборке средних:



Зоопарк: оценить дисперсию выборочного среднего

При большем количестве бутстрепных выборок:



Особенности

- Два этапа аппроксимации

$$v \underbrace{\approx}_{\text{метод подстановки}} \widehat{v} \underbrace{\approx}_{\text{Монте-Карло}} \widehat{v}_{boot}.$$

  Точность аппроксимации м. подстановки: $1/\sqrt{n}$

  Точность аппроксимации м. Монте-Карло: $1/\sqrt{B}$

- Число B стоит брать как можно больше.

- Размер бутстрепной выборки **всегда тот же**, что и у исходной.

  При генерации выборок иного размера распределение статистики T, вообще говоря, может быть другим.

  Например, дисперсия выборочного среднего зависит от размера выборки.

- Генерация бутстр. выборки проводится независимо с повторами.

  Иначе полученный набор даже не является выборкой.

Бутстрепные доверительные интервалы

1. **Нормальный интервал**

   Пусть $\widehat{\theta}$ — а.н.о. θ с ас. дисп. $\sigma^2(\theta)$.

   $\widehat{v}_{boot}$ — бутстрепная оценка дисперсии.

   Бутстрепный дов. интервал для параметра θ имеет вид

   $$\left(\widehat{\theta} - z_{(1+\alpha)/2}\sqrt{\widehat{v}_{boot}}, \quad \widehat{\theta} + z_{(1+\alpha)/2}\sqrt{\widehat{v}_{boot}}\right)$$

2. **Центральный интервал**

   $\theta = G(\mathrm{P})$ и $\widehat{\theta} = G(\widehat{\mathrm{P}}_n)$ — оценка методом подстановки.

   $\theta_1^*, ..., \theta_B^*$ — оценки по бутстрепным выборкам.

   Бутстрепный доверительный интервал имеет вид

   $$C^* = \left(2\widehat{\theta} - \theta^*_{(\lceil B(1+\alpha)/2\rceil)}, \quad 2\widehat{\theta} - \theta^*_{(\lfloor B(1-\alpha)/2\rfloor)}\right).$$

При достаточно слабых условиях на G: $\mathrm{P}\left(\theta \in C^*\right) \to \alpha$.

Бутстрепные доверительные интервалы

3. **Квантильный интервал**

   $\widehat{\theta}$ — некоторая оценка θ.

   $\theta_1^*, ..., \theta_B^*$ — оценки по бутстрепным выборкам.

   Бутстрепный доверительный интервал имеет вид

$$C^* = \left(\theta^*_{(\lfloor B(1-\alpha)/2 \rfloor)}, \quad \theta^*_{(\lceil B(1+\alpha)/2 \rceil)}\right).$$

Утв. Если существует монотонное преобразование φ, для которого $\varphi(\widehat{\theta}) \sim \mathcal{N}(\varphi(\theta), \sigma^2)$, то $\mathsf{P}\left(\theta \in C^*\right) = \alpha$.

На практике такое преобразование существует редко, но при этом часто может существовать приближенное преобразование.

Пример: построение дов. интервалов для θ

$x = (5, 1, 3, 6, 4)$ — реализация выборки

$\theta = \mathsf{E}X_1$ — параметр, $\widehat{\theta} = \overline{X}$ — оценка, $\widehat{\theta} = 3.8$ — реализация оценки

Реализации оценки параметра по бутстрепным выборкам ($B = 100$):

4.2, 4.2, 2.6, 3.2, 4.2, 3.8, 3.2, 3.6, 3.6, 3.4, 3.8, 4.4, 3.6, 3.2, 4.6, 4.2, 3.0, 3.2, 4.0, 3.0, 3.2, 3.0, 2.6, 3.0, 3.6, 3.4, 5.0, 4.8, 3.4, 2.6, 2.6, 3.6, 3.2, 4.2, 3.2, 3.4, 4.4, 4.2, 4.4, 3.4, 4.0, 2.4, 3.4, 3.8, 2.0, 3.0, 4.6, 3.2, 3.6, 3.6, 4.0, 3.8, 4.0, 3.4, 3.8, 3.8, 4.2, 3.2, 2.8, 4.0, 3.2, 3.4, 3.0, 4.0, 3.6, 3.4, 3.8, 3.2, 3.8, 2.6, 3.4, 5.0, 3.6, 3.0, 4.8, 4.2, 3.4, 5.2, 5.0, 3.4, 3.2, 3.6, 4.2, 3.4, 3.2, 3.8, 3.6, 3.8, 3.0, 2.8, 3.0, 4.0, 3.2, 3.6, 2.6, 3.2, 2.4, 3.6, 4.0, 4.2

1. **Нормальный интервал**

   $\widehat{\theta} = 3.8,\ v_{boot} = 0.394,\ z_{0.975} = 1.96$

   $(3.8 \pm 1.96 \cdot \sqrt{0.394}) = (2.57,\ 5.03)$

2. **Центральный интервал**

   $B(1+\alpha)/2 = 100 \cdot 0.975 = 97.5,\ B(1-\alpha)/2 = 100 \cdot 0.025 = 2.5$

   $\theta^*_{(\lceil 97.5 \rceil)} = 5, \quad \theta^*_{(\lfloor 2.5 \rfloor)} = 2.4$

   $(2 \cdot 3.8 - 5,\ 2 \cdot 3.8 - 2.4) = (2.6,\ 5.2)$

3. **Квантильный интервал**

   $(2.4,\ 5)$

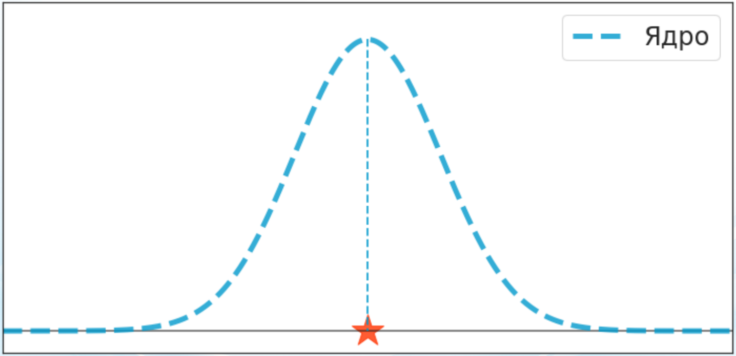
4.4. Ядерные оценки плотности

Kernel density estimation

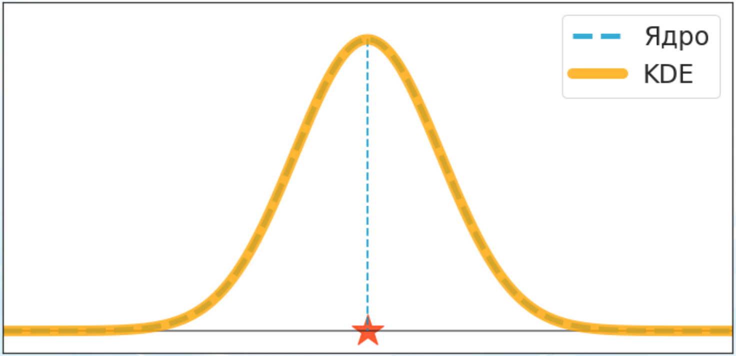
KDE

Ядерная оценка плотности: простые примеры

Ядерная оценка плотности: простые примеры

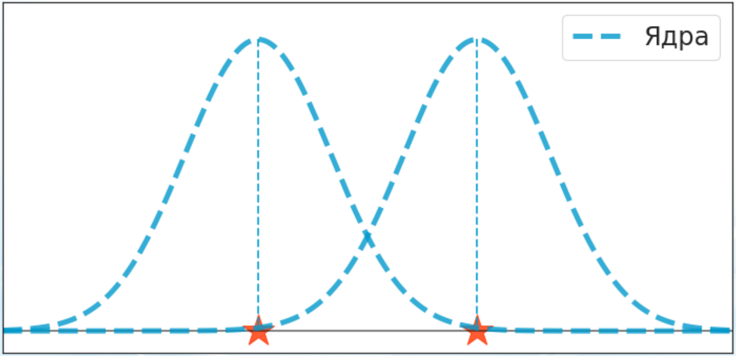


Ядерная оценка плотности: простые примеры

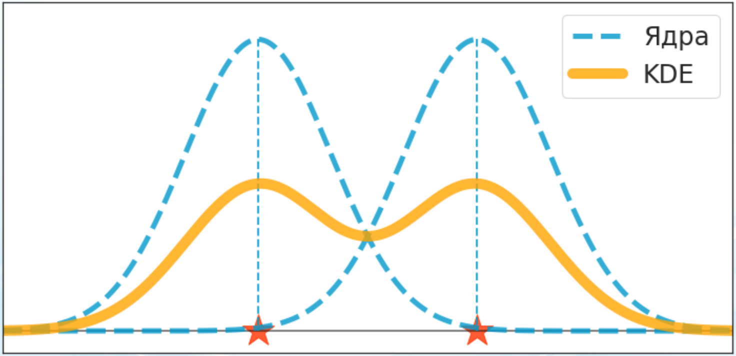


Ядерная оценка плотности: простые примеры

Ядерная оценка плотности: простые примеры

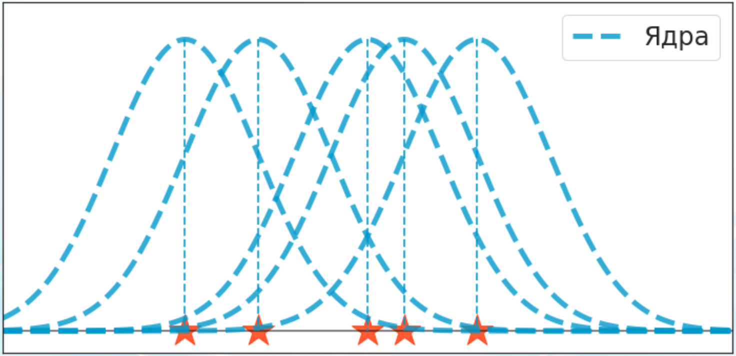


Ядерная оценка плотности: простые примеры

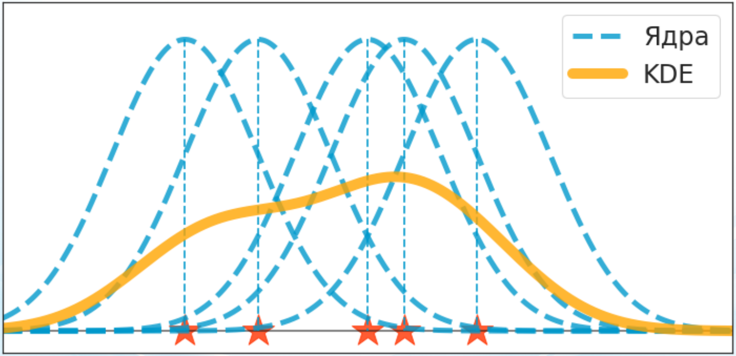


Ядерная оценка плотности: простые примеры

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

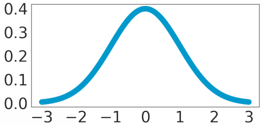
$$\widehat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} q\left(\frac{x - X_i}{h}\right)$$

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

Виды ядер

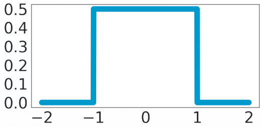
Гауссовское

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



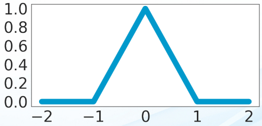
Прямоугольное

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



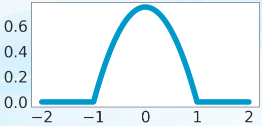
Треугольное

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



Епанечникова

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




ВСЁ!