Введение в непараметрическое моделирование

ЦМФ

Гистограмма

Наиболее простая и широко используемая непараметрическая оценка плотности распределения

Пусть мы имеем выборку $(y_1, \dots, y_n)$. Разделим всю область определения случайной величины на несколько интервалов, тогда оценка плотности запишется в виде

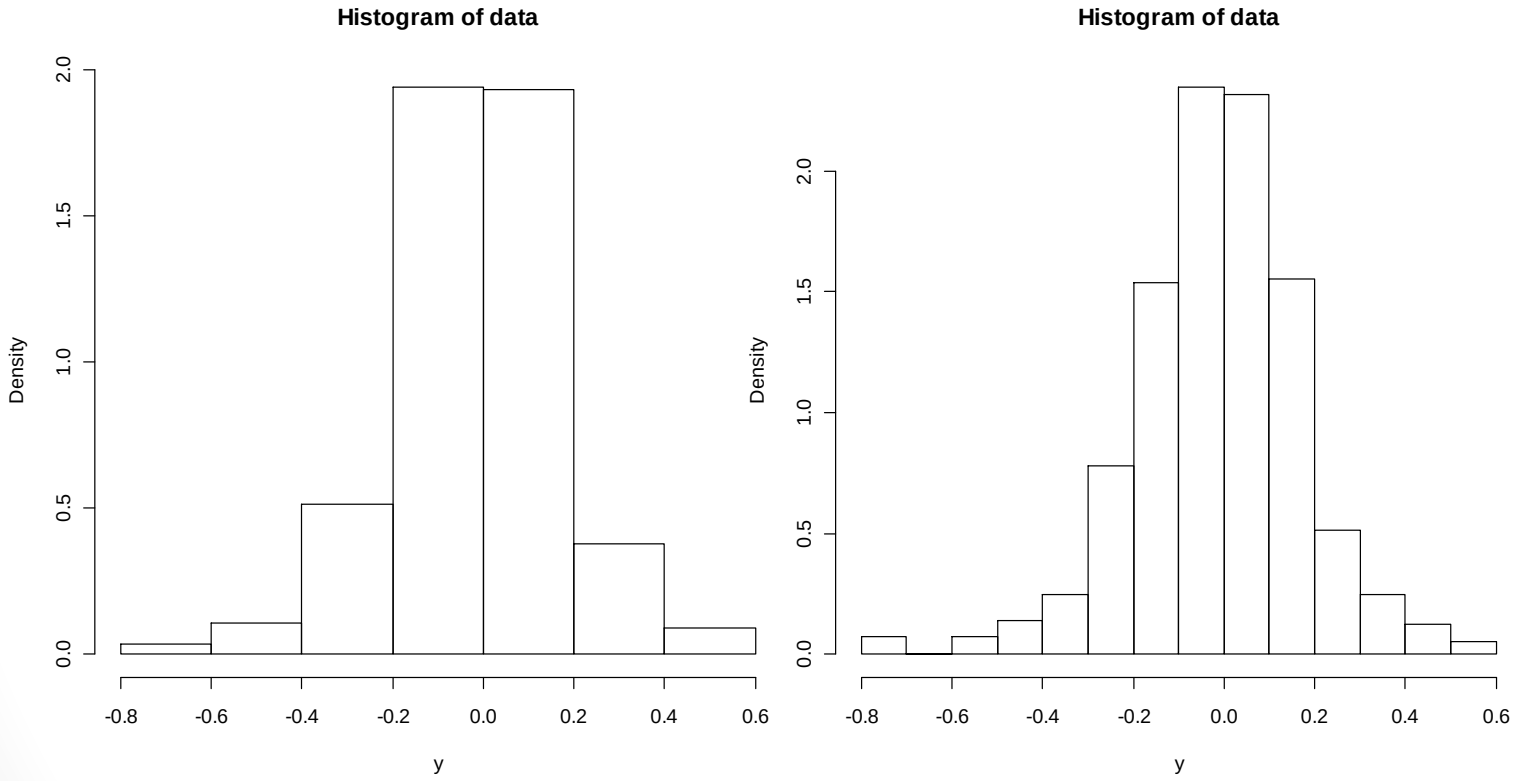
$$\hat{f}(y) = \frac{1}{n} \cdot \frac{\text{число наблюдений в интервале } y}{\text{длина интервала}} \quad (1)$$

Для построения гистограммы нужно определить:

1. Границы области определения;
2. Длину (количество) интервалов

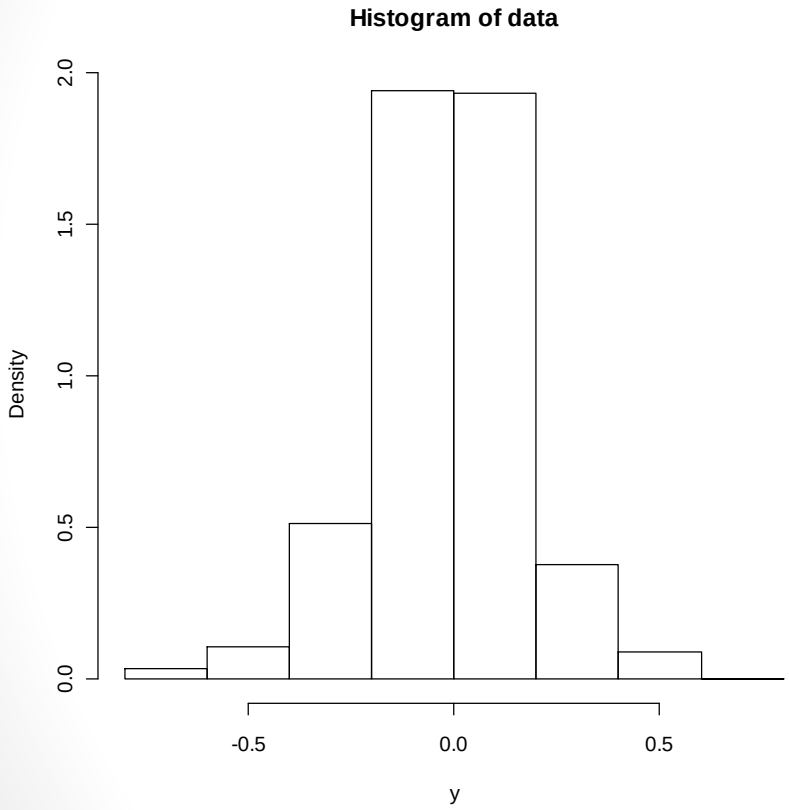
Параметры гистограммы

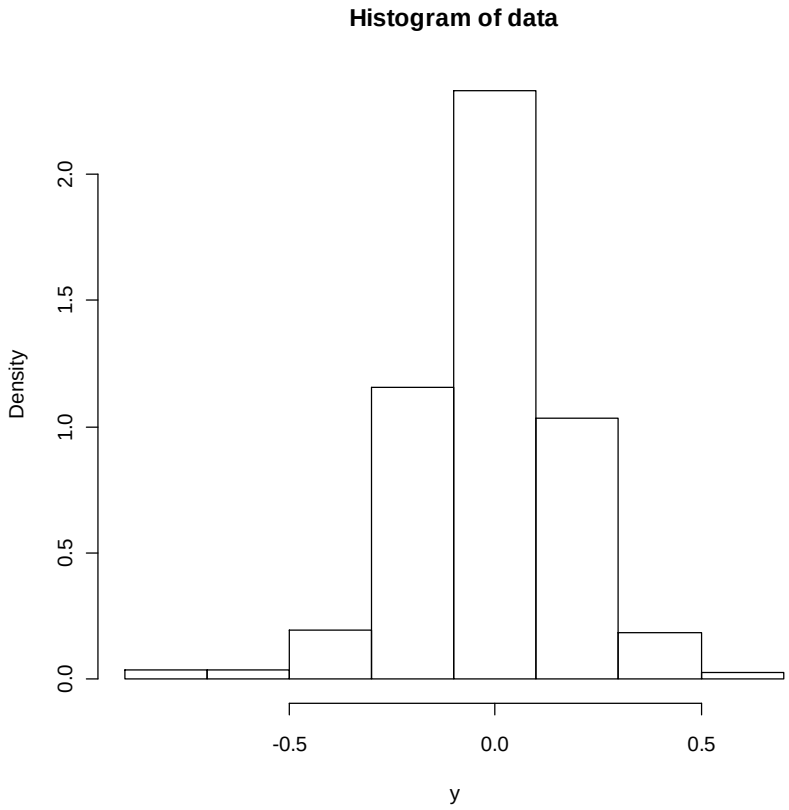
Длина интервалов влияет на детализацию гистограммы



Параметры гистограммы

Область определения может повлиять на форму





Оценка плотности распределения

Оценку (1) можно записать более формально. Пусть у нас есть m интервалов вида $(z_k; z_{k+1}),\ k \in \{1; \dots; m\},\ m < n.$ Все интервалы одинаковой длины $h = z_{k+1} - z_k$, тогда

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} I(z_k < y_i \leq z_{k+1}),\ z_k < y \leq z_{k+1} \quad (2)$$

Длина интервала *h* должна быть достаточно большой, чтобы в него попало существенное количество наблюдений, и достаточно малой, чтобы не потерять важные детали распределения

Остаётся нерешенной проблема области распределения

Выход — ядерные оценки

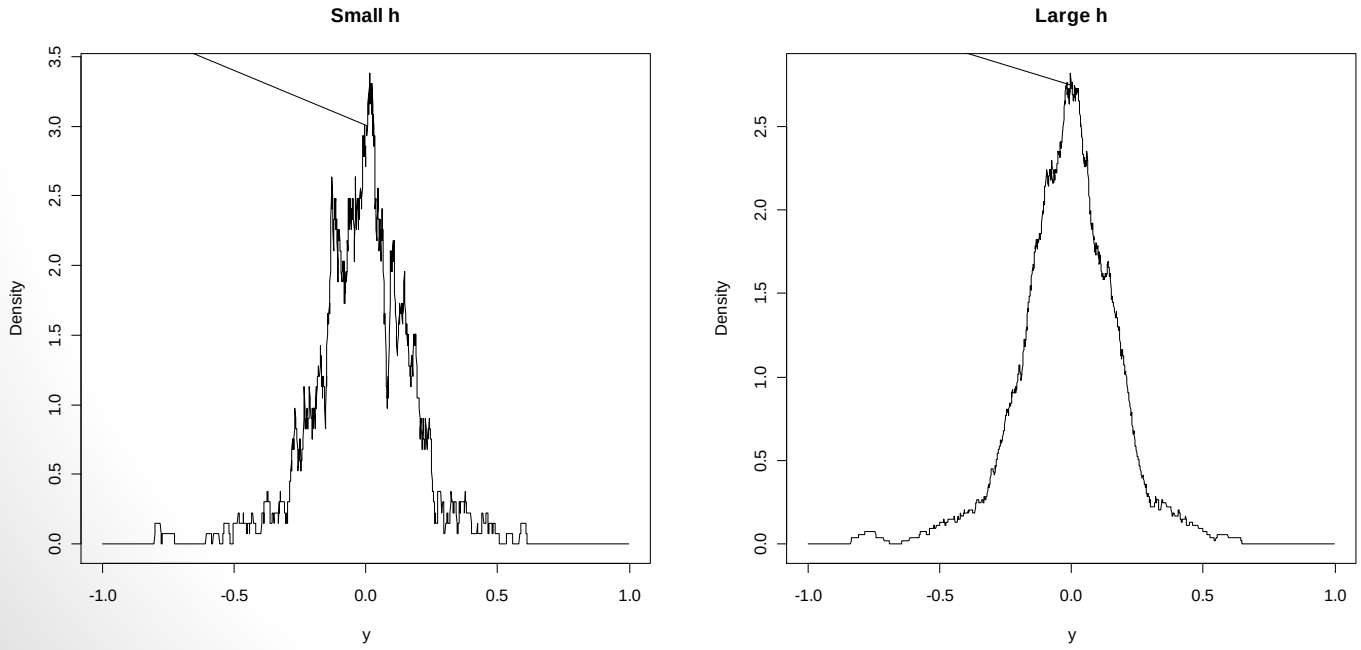
Простая непараметрическая оценка

Принцип построения простой (naïve) непараметрической оценки плотности в точке *y* состоит в подсчёте количества наблюдений, находящихся вблизи неё:

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} I\left(y - \frac{h}{2} < y_i < y + \frac{h}{2}\right) \quad (3),$$

где h — длина интервала

Большие значения *h* дают более гладкие оценки:



Ядерная оценка

Простая оценка нигде не дифференцируема. Чтобы понять это перепишем формулу (3) в следующем виде:

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} w\left(\frac{y-y_i}{h}\right), \text{ где } w(x) = I\left(|x| < \frac{1}{2}\right) \quad (4)$$

Проблема заключается в функции *w(x),* которая придаёт наблюдениям дискретные веса (0 или 1)

Проблема решается с помощью замены функции *w(x)* на ядерную функцию *K(x)* с плавно изменяющимися весами:

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y-y_i}{h}\right) \quad (5)$$

Для того, чтобы оценка $\hat{f}(y)$ была функцией плотности, ядро должно удовлетворять условию $\int_{-\infty}^{+\infty} K(x)dx = 1$

Любая функция плотности удовлетворяет этому условию

Ядерные функции

В качестве ядерных функций обычно используются симметричные одномодальные функции плотности

Наиболее часто используемые на практике ядра:

$K_G(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \quad (6)$ — гауссовское ядро

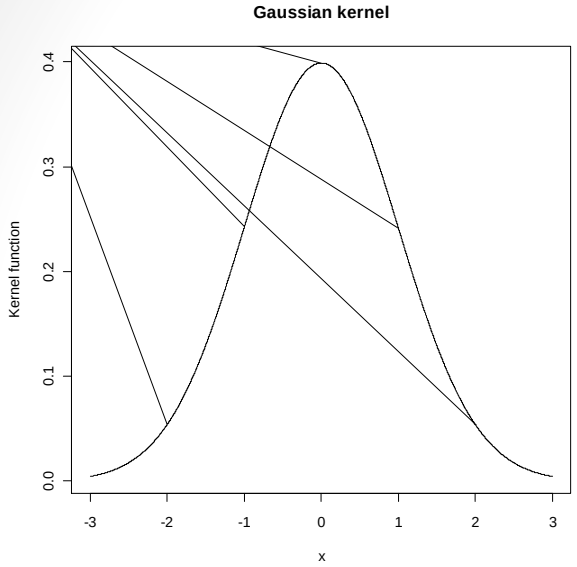
$K_E(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right) \cdot I\left(|x| < \sqrt{5}\right) \quad (7)$ — ядро Епанечникова

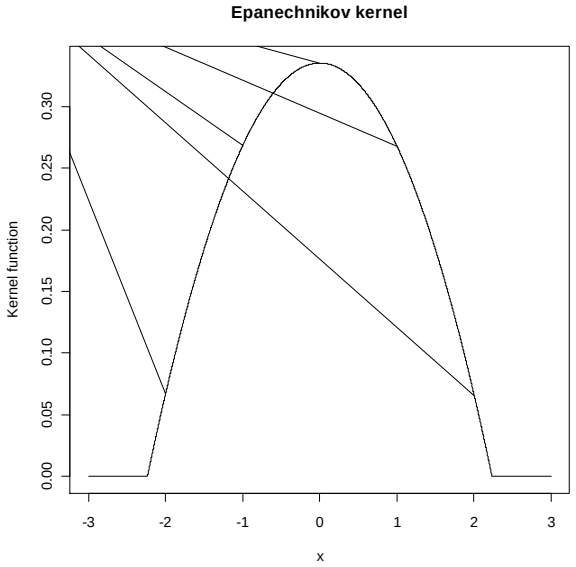
$K_T(x) = (1 - |x|) \cdot I(|x| < 1) \quad (8)$ — треугольное ядро

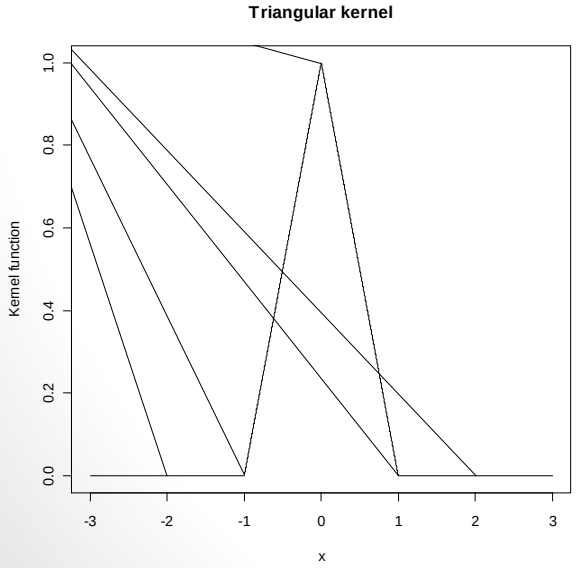
$K_U(x) = \frac{1}{2} I(|x| < 1) \quad (9)$ — прямоугольное (равномерное) ядро

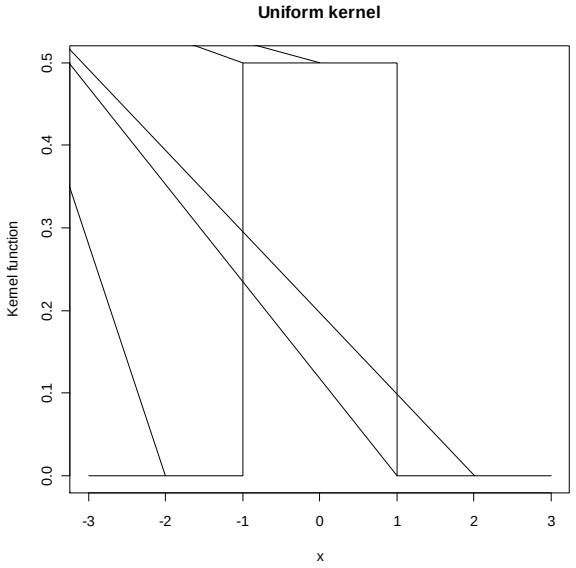
Вид этих функций представлен на следующем слайде

Ядерные функции



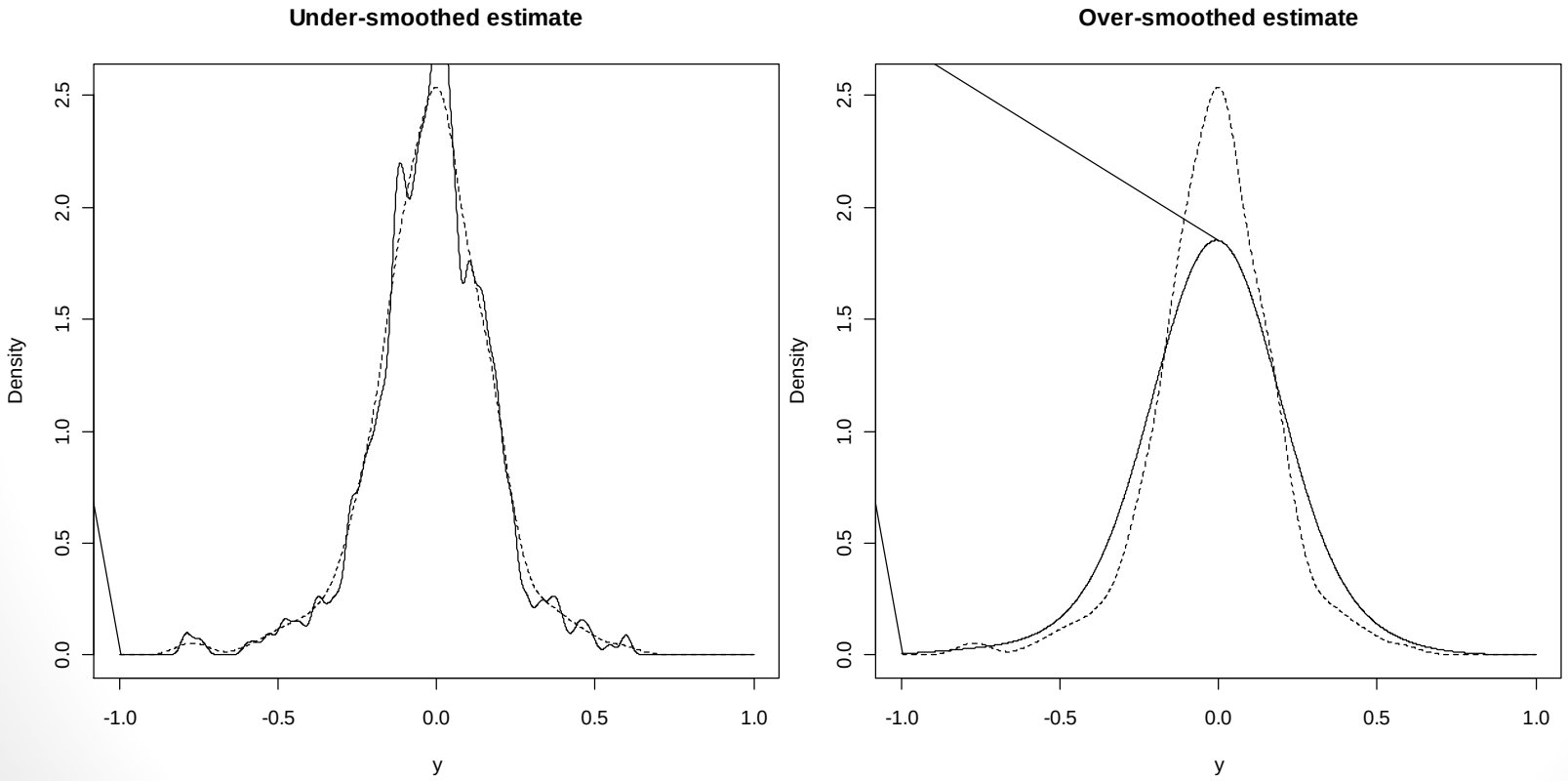






Влияние ширины интервала

Тогда как выбор ядра оказывает незначительное влияние на оценку плотности, выбор ширины интервала имеет решающее значение



Выбор ширины интервала

Существует два основных подхода к определению величины сглаживающего множителя (ширины интервала):

1. Фиксированная ширина интервала на всей выборке. В рамках этого подхода выделяют:
   - правило подстановки (rule of thumb);
   - метод перекрёстной проверки (cross-validation)
2. Ширина интервала меняется в зависимости от локальной концентрации наблюдений. Методы:
   - обобщённый метод ближайших соседей (generalized nearest neighbors);
   - адаптивный метод (adaptive nearest neighbors)