Непараметрическое моделирование (одномерный случай)

Финансовая эконометрика

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

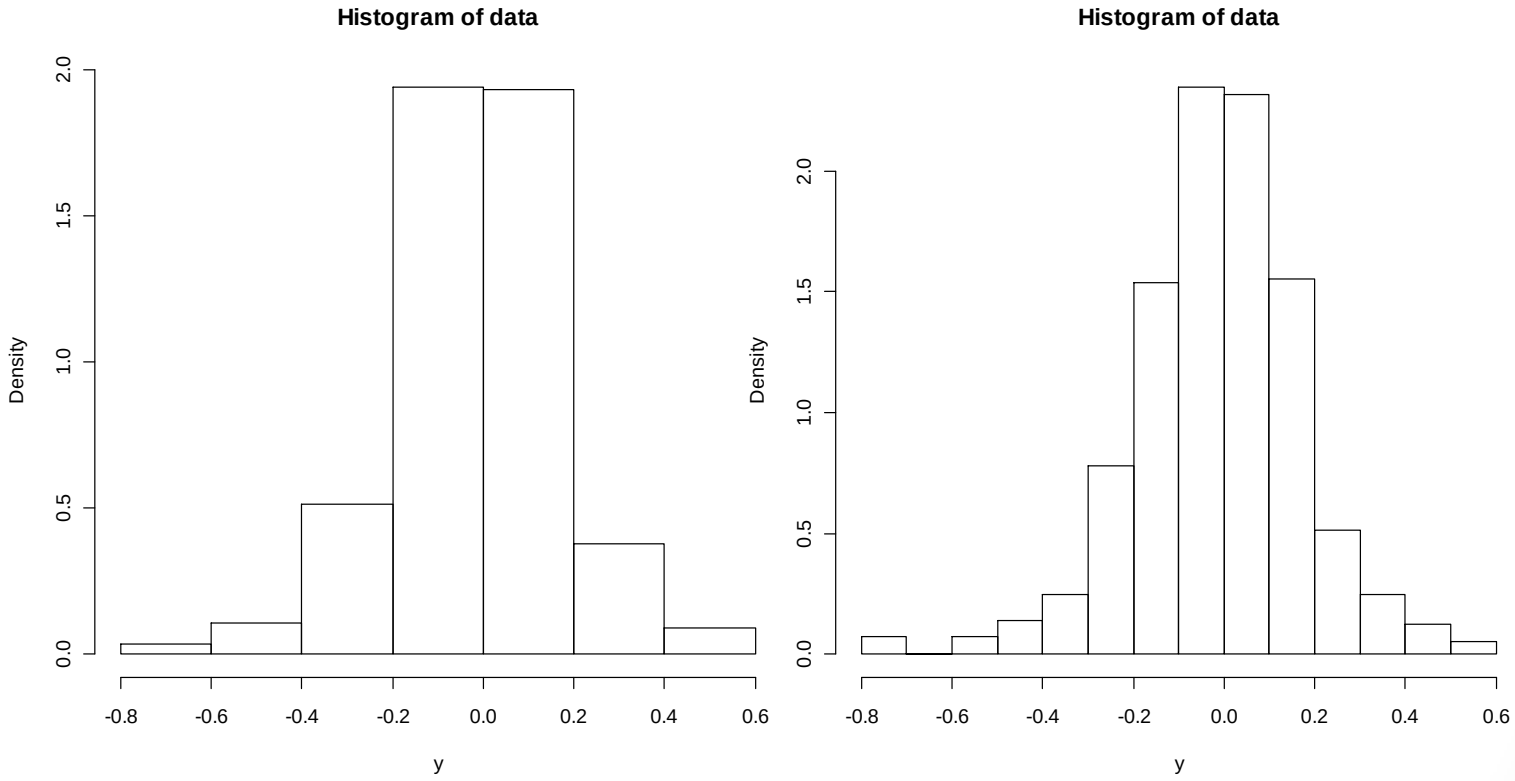
$$\hat{f}(y) = \frac{1}{n} \cdot \frac{\text{число наблюдений в интервале вокруг } y}{\text{длина интервала}} \quad (1)$$

Для построения гистограммы нужно определить:

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

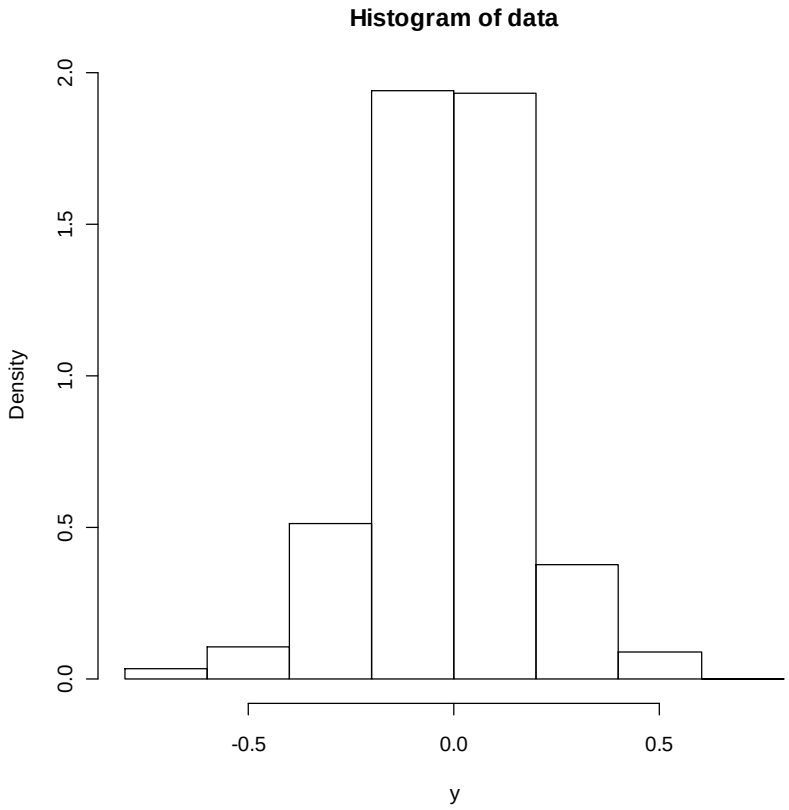
Параметры гистограммы

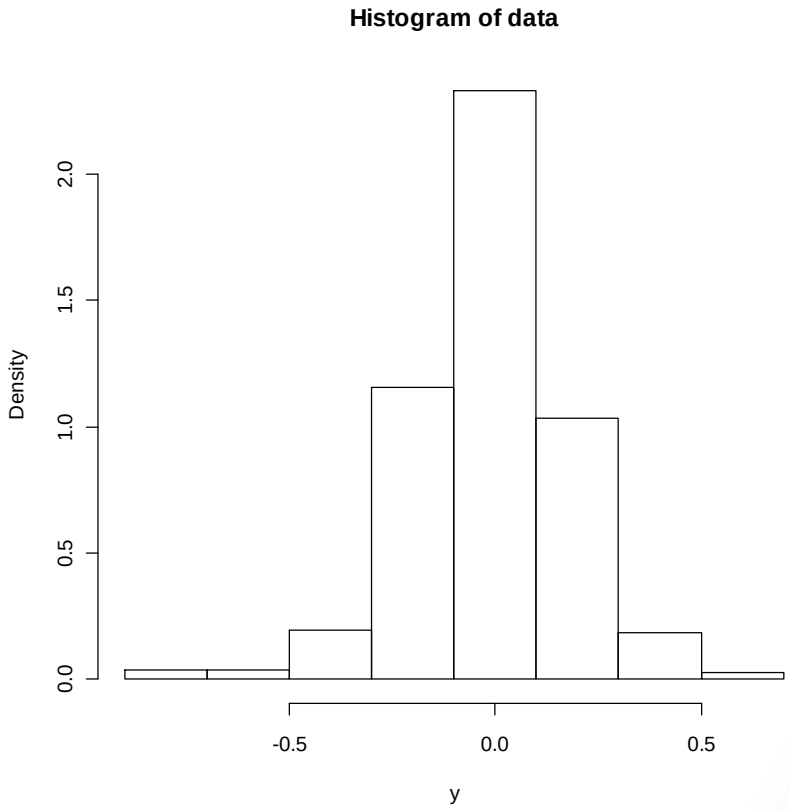
Длина интервалов влияет на детализацию гистограммы



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 нерешённой проблема области распределения

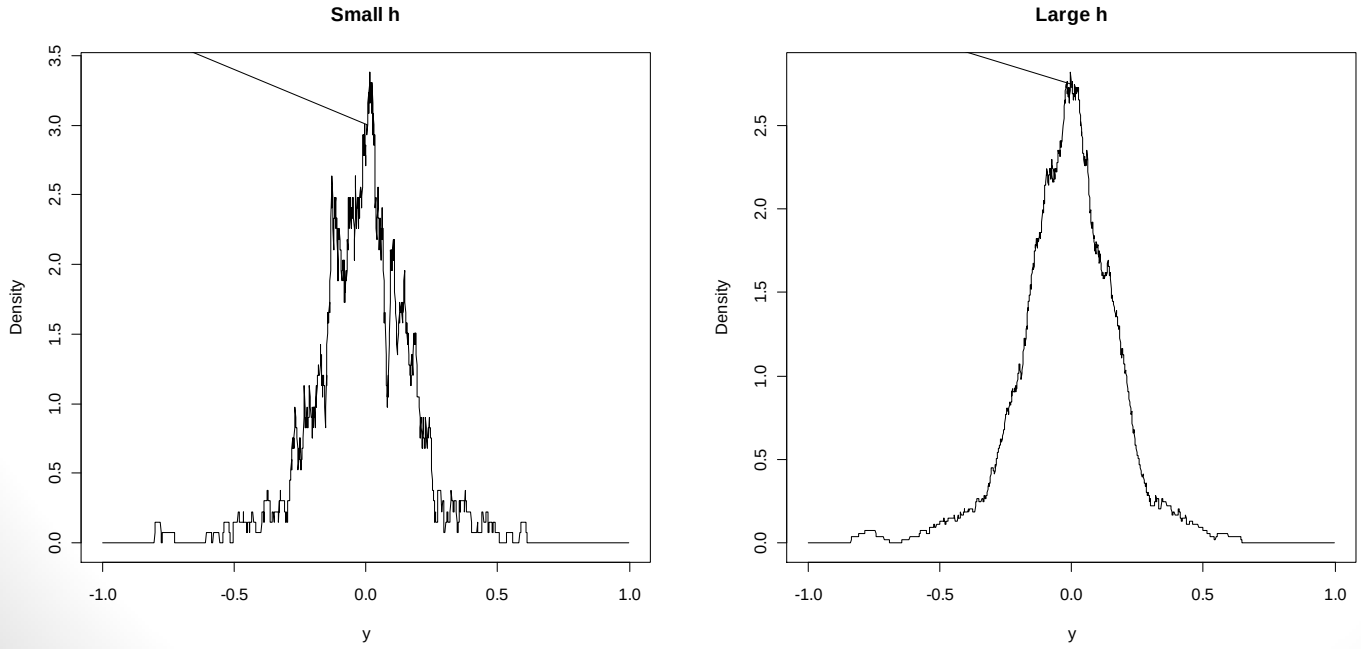
Простая непараметрическая оценка

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

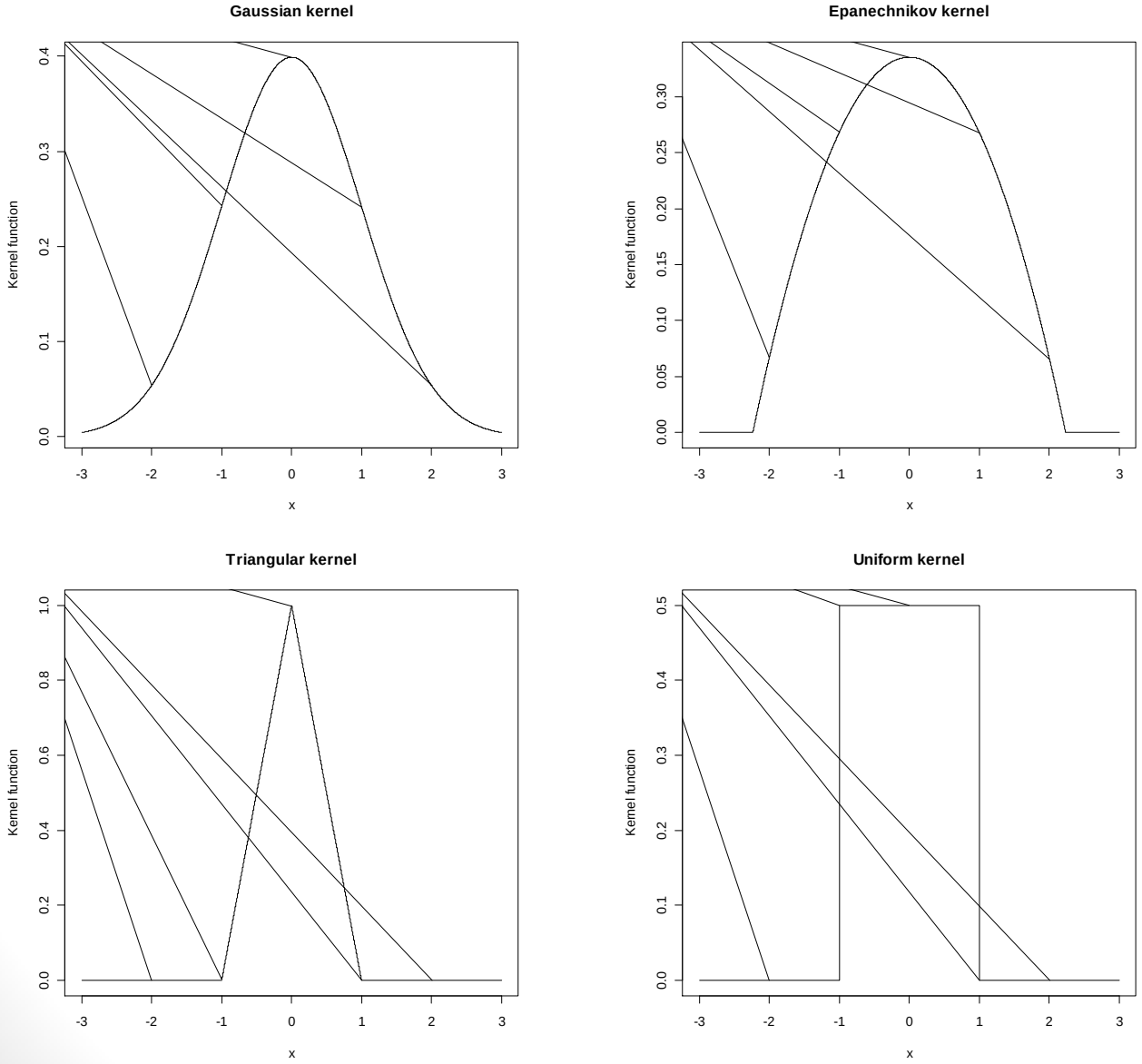
$K_E(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right) \cdot I\left(|x| < \sqrt{5}\right) \quad (7)$ — ядро Епанечникова

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

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

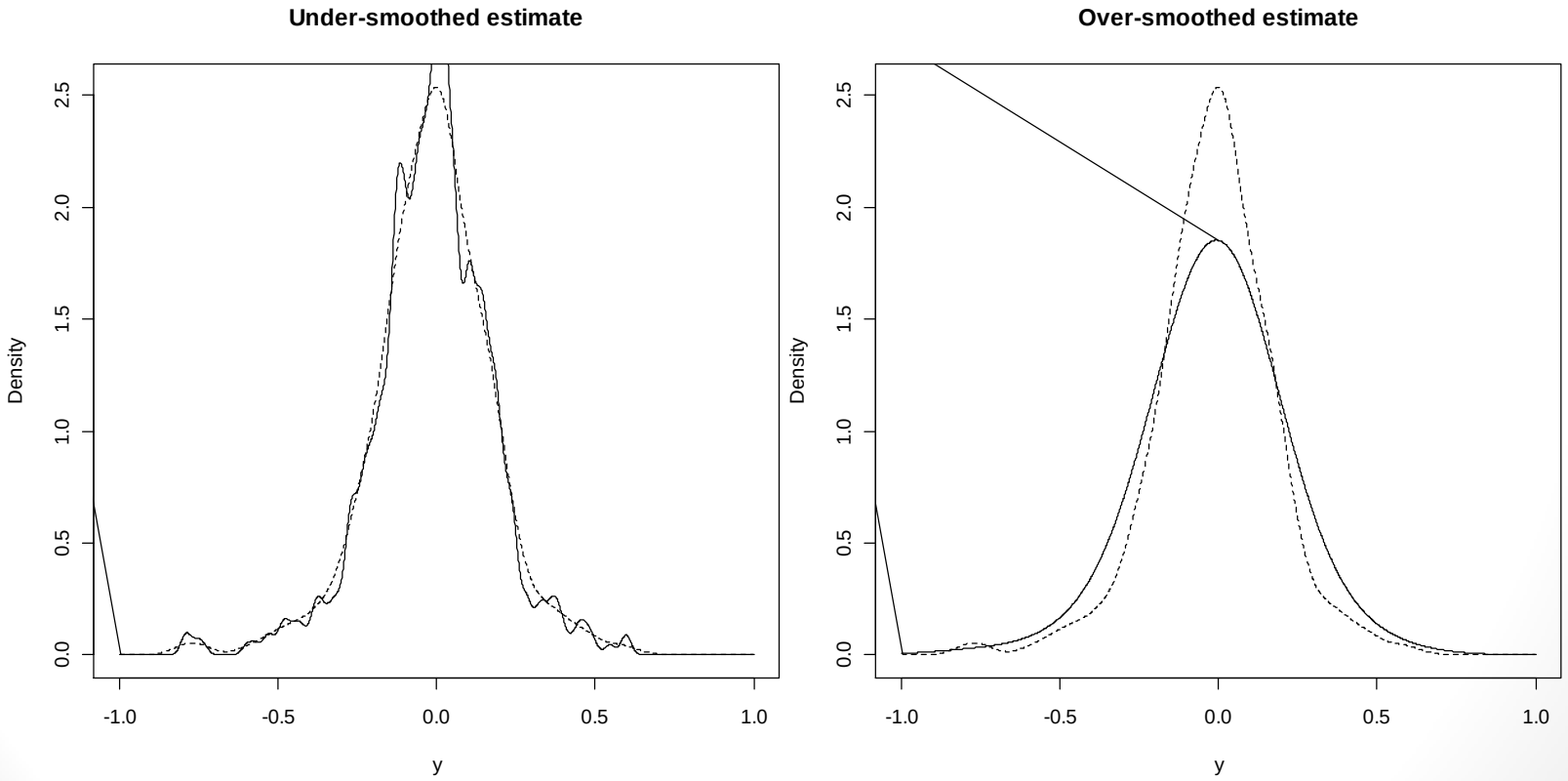
Вид этих функций представлен на следующем слайде

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)

Среднеквадратичная ошибка

Выбирать величину *h* следует так, чтобы оценка была как можно ближе к истинной плотности распределения, т.е. минимизировать разницу между $\hat{f}(y)$ и $f(y)$

Наиболее естественным кандидатом на эту разницу является среднеквадратичная ошибка (Mean Squared Error, MSE), рассчитываемая в конкретной точке *y*:

$$MSE(h,y) = E\left(\left(\hat{f}(y) - f(y)\right)^2\right) \quad (10)$$

Распишем выражение (10) подробнее:

$$MSE(h,y) = E\left(\left(\hat{f}(y) - f(y)\right)^2\right) = E\left(\hat{f}^2(y)\right) - 2f(y)E\left(\hat{f}(y)\right) + f^2(y)$$

$$= \left[E\left(\hat{f}^2(y)\right) - E^2\left(\hat{f}(y)\right)\right] + \left[E^2\left(\hat{f}(y)\right) - 2f(y)E\left(\hat{f}(y)\right) + f^2(y)\right] =$$

$$var\left(\hat{f}(y)\right) + \left(E\left(\hat{f}(y)\right) - f(y)\right)^2 \quad (11)$$

Дисперсия и смещение оценки

$$MSE(h, y) = var\left(\hat{f}(y)\right) + \left(E\left(\hat{f}(y)\right) - f(y)\right)^2 \quad (11)$$

Первое слагаемое выражения (11) соответствует дисперсии оценки, второе — квадрату её смещения

Если ширина интервала слишком большая, то оценка оказывается пересглаженной, и растёт смещение

Если значение *h* слишком маленькое, то это увеличивает дисперсию

Минимальное смещение достигается при максимальной дисперсии $(h = 0)$, а минимальная дисперсия — при максимальном смещении $(h \to +\infty)$

Нужно искать компромисс

Интегральная среднеквадратичная ошибка

Поскольку мы заинтересованы в минимизации отклонения между оценкой $\hat{f}(y)$ и плотностью $f(y)$ не только в конкретной точке *y*, рассмотрим интегральную среднеквадратичную ошибку (Mean Integrated Squared Error, MISE):

$$MISE(h) = E\left(\int_{-\infty}^{+\infty}\left(\hat{f}(y) - f(y)\right)^2 dy\right) \quad (12)^1$$

Мы можем переписать это так:

$$MISE(h) = \int E\left(\left(\hat{f}(y) - f(y)\right)^2\right)dy = \int MSE(h,y)dy \quad (13), —$$

или в следующем виде:

$$MISE(h) = \int var\left(\hat{f}(y)\right)dy + \int\left(E\left(\hat{f}(y)\right) - f(y)\right)^2 dy \quad (14)$$

[1] Далее вместо определённого интеграла по всей числовой оси будет использоваться неопределённый

Оптимальная ширина интервала

Минимизируя аппроксимацию к критерию *MISE*, обозначаемую *AMISE*, можно найти оптимальное значение параметра сглаживания:

$$h_{opt} = \left(\int x^2 K(x)dx\right)^{-\frac{2}{5}}\left(\int K^2(x)dx\right)^{\frac{1}{5}}\left(\int f''^2(y)\,dy\right)^{-\frac{1}{5}} n^{-\frac{1}{5}} \quad (15)$$

Замечания к формуле (15):

- h_{opt} стремится к нулю по мере роста объема выборки, но сравнительно медленно (по степенному закону);
- h_{opt} уменьшается, если $f(y)$ сильно варьируется, и возрастает, если функция плотности варьируется слабо;
- наиболее подходящее ядро $K(x)$ можно определить, исходя из значения критерия *MISE* (14)

Методы оценки оптимальной ширины интервала

В выражении для h_{opt} (15) остаётся неопределённость, связанная с незнанием истинной функции плотности $f(y)$

Мы рассмотрим два способа преодоления этой неопределённости:

1. Правило подстановки (Rule of Thumb);
2. Метод перекрёстной проверки (Cross-Validation)

Правило подстановки

Вместо $f(y)$ в выражение для оптимального интервала (15) подставляется какое-либо известное распределение

Если подставить нормальное распределение $N(\mu, \sigma^2)$ и использовать гауссовское ядро, то получим:

$$\hat{h}_{opt} \approx 1.059\sigma n^{-\frac{1}{5}} \quad (16)$$

В качестве оценки *σ* можно использовать выборочное стандартное отклонение, $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{j=1}^{n} y_j\right)}$, или межквартильное расстояние, $\frac{\hat{q}_3 - \hat{q}_1}{1.349}$, где $\hat{q}_i$ — выборочное значение *i*-го квартиля, 1.349 — межквартильное расстояние для стандартного нормального распределения

Модифицированное правило подстановки

Правило подстановки хорошо работает тогда, когда истинный закон распределения близок к подставляемому

Существует также модифицированное правило подстановки:

$$\hat{h}_{opt} = 0.9 \min\left(\hat{\sigma}, \frac{\hat{q}_3 - \hat{q}_1}{1.349}\right) n^{-\frac{1}{5}} \quad (17)$$

Модифицированное правило является более устойчивым к отклонениям истинного распределения от нормального закона

Метод перекрёстной проверки

Мы опишем вариацию метода, основанную на наименьших квадратах

Идея состоит в рассмотрении интегральной квадратической ошибки (Integrated Squared Error, ISE), аналогичной критерию *MISE* (12), но без математического ожидания:

$$ISE(h) = \int \left(\hat{f}(y) - f(y)\right)^2 dy =$$

$$\int \hat{f}^2(y)dy - 2\int \hat{f}(y)f(y)dy + \int f^2(y)dy \quad (18)$$

Последнее слагаемое не зависит от *h* не играет роли в оптимизации

Величина $\int \hat{f}(y)f(y)dy$ есть матожидание оценки, которое приближённо равно $E\left(\hat{f}(y)\right) \approx \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(y_i)$, где $\hat{f}_{-i}(y_i)$ — оценка плотности по всем наблюдениям, кроме y_i

Метод перекрёстной проверки

Таким образом, оптимизационная задача сводится к минимизации выражения

$$CV(h) = \int \hat{f}^2(y)dy - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(y_i) \quad (19)$$

Достоинства методов с фиксированной шириной интервала:

- простота вычислений;
- интуитивная понятность;
- оценки обладают известными статистическими свойствами

Недостатки:

- пересглаженный центр распределения;
- недосглаженные и тонкие хвосты

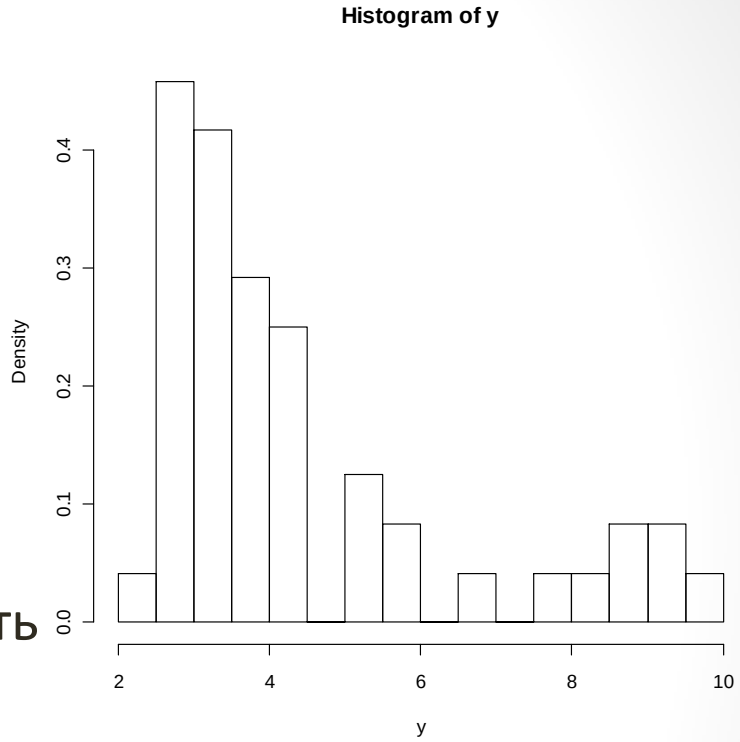
Пример 1. Острова

```
library(datasets)
y <- log(islands)
```

Построение гистограммы

```
hist(y,nclass=12,probability=TRUE)
```

- ***nclass*** определяет количество интервалов
- ***probability*** преобразует количество наблюдений в интервале в плотность распределения



С помощью дополнительного параметра ***breaks=c(y_1,…,y_k)*** задаётся разбиение на интервалы

Пример 1. Острова

Простая непараметрическая оценка плотности

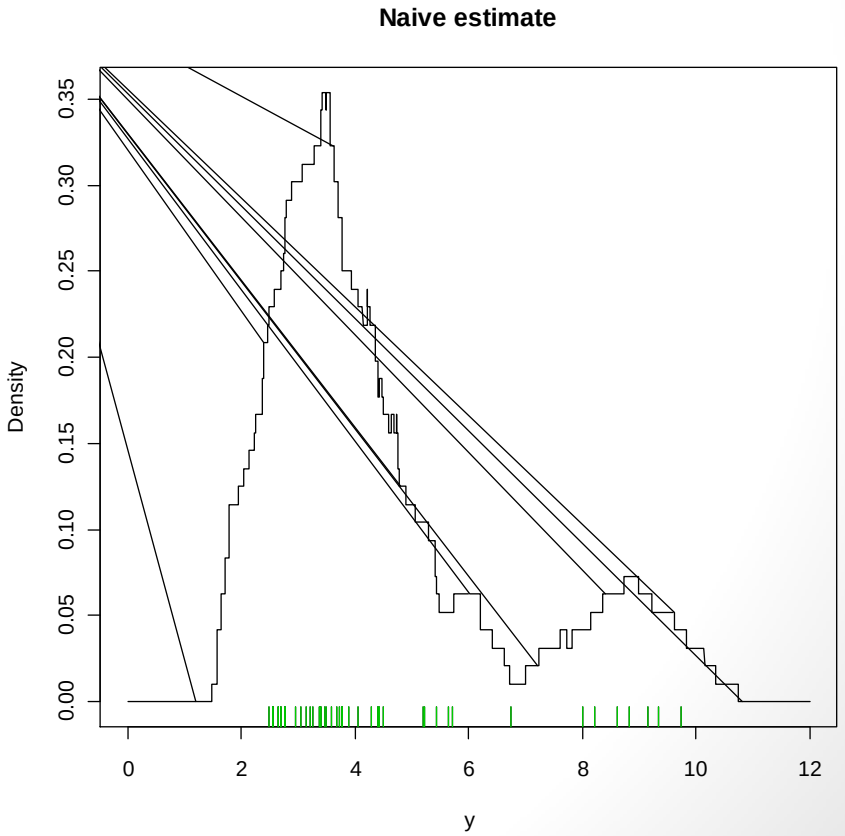
```
L <- 10^4; N <- length(y)
h <- 2   # ширина интервала

# в точках x будет оцениваться плотность
x <- seq(0,12,length=L)   # последовательность 0 – 12 длиной L
f.naive <- numeric()      # нулевой (пока) вектор оценок

# считаем количество элементов в интервалах x_i ± h/2
for (i in 1:L) f.naive[i] <- sum(1*((y>x[i]-h/2)&(y<x[i]+h/2)))
f.naive <- f.naive/(N*h)  # нормируем оценку
```

Пример 1. Острова

График простой оценки

```
plot(x,f.naive,type="l",main="Naive estimate",
xlab="y",ylab="Density")
rug(y,col=3)
```

- ***type*** определяет вид графика
  "l" — линии, "p" — точки, …
- ***main*** — заголовок
- ***xlab*** — подпись на оси x
- ***ylab*** — подпись на оси y



Пример 1. Острова

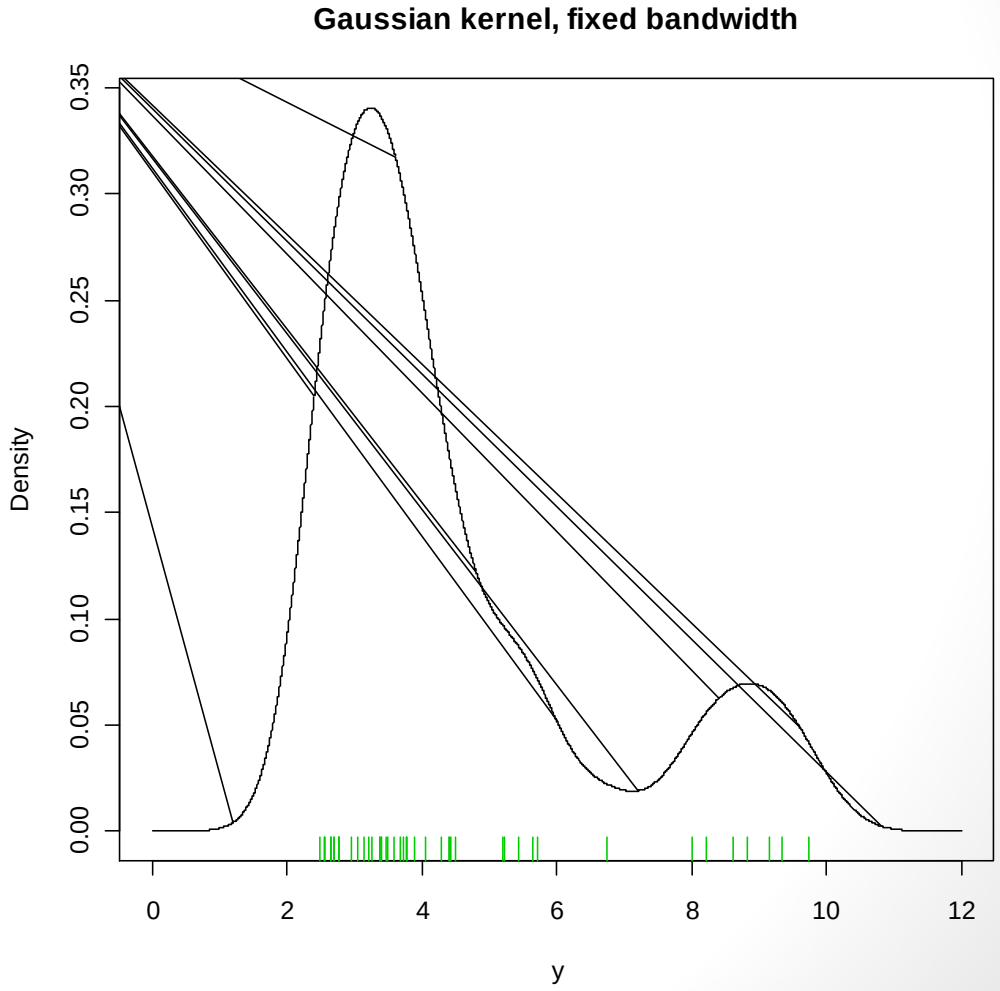
Ядерные оценки

```
library(np)
f.fix <- npudens(tdat=y,edat=x,
  ckertype="gaussian",bwtype="fixed")
```

- ***tdat*** — обучающая выборка
- ***edat*** — точки, в которых рассчитывается оценка
- ***ckertype*** — вид ядерной функции
  "gaussian", "epanechnikov", "uniform"
- ***bwtype*** определяет метод расчёта интервала *h*
  "fixed", "generalized_nn", "adaptive_nn"
- ***f$dens*** — искомые значения оценок

Пример 1. Острова

```
plot(x,f.fix$dens,type="l",
main="Gaussian kernel, fixed bandwidth",
xlab="y",ylab="Density")
```



Пример 1. Острова

Нахождение квантилей оценки распределения, $\hat{F}^{-1}(\alpha)$

оценка функции распределения

```
F.fix <- npudist(tdat=y,edat=x,ckertype="gaussian",bwtype="fixed")
```

поиск квантиля методом деления пополам

```
alpha <- 0.99
a <- 1; b <- L; ab <- trunc((a+b)/2)
while ((b-a)>2) {
  if  (F.fix$dist[ab]<=alpha)  a <- ab
  if  (F.fix$dist[ab]>=alpha)  b <- ab
  ab <- trunc((a+b)/2)
}
q.fix <- x[ab]
```

q.fix	10.00

Пример 1. Острова

Генератор случайных чисел

фиксированный интервал

```
M <- 10^6
y.fix.sim <- sample(x,prob=f.fix$dens,size=M,replace=TRUE)
q.fix <- sort(y.fix.sim)[alpha*M]
```

q.fix	10.01

Домашнее задание

- рассчитать оценки риска для биржевого индекса по всей совокупности наблюдений
- построить кривую VaR с помощью ядерной оценки плотности с фиксированным сглаживающим множителем h и проверить качество оценок риска
- написать функцию, которая будет возвращать квантиль заданного уровня по произвольной непараметрической модели, используя метод деления пополам или метод Монте-Карло на выбор пользователя

Исходные данные — котировки с сайтов finam.ru, finance.yahoo.com и др.