Гистограмма ядерная оценка плотности поиск наилучшей проекции

Аббакумов Вадим Леонардович

План лекции

- Гистограмма
- Непараметрические оценки плотности – улучшенная гистограмма
- Ящики с усами
- Непараметрические оценки плотности при поиске наилучшей проекции (Projection Pursuit, Independent Component Analysis).

1. Гистограмма

- Алгоритм построения гистограммы

Данные

$$x_1, x_2, x_3, \ldots, x_n$$

$$x_{(1)} \le x_{(2)} \le x_{(3)} \le \ldots \le x_{(n)}$$

$$a \le x_{(1)} \le x_{(2)} \le x_{(3)} \le \ldots \le x_{(n)} \le b$$

Интервалы

- Разбиваем интервал [*a*,*b*] на *k* непересекающихся интервалов.

- Интервалы обозначаем

$$\Delta_i \qquad i = 1, 2, \ldots, k$$

- Обычно одинаковой длины.
- Лучше разной длины, но не умеем.

- Как выбирать интервалы и их число – открытый вопрос…

Столбцы

- Обозначения: в интервал Δ_i попало n_i наблюдений
- $$\sum_{i=1}^{k} n_i = n$$

Высота столбца

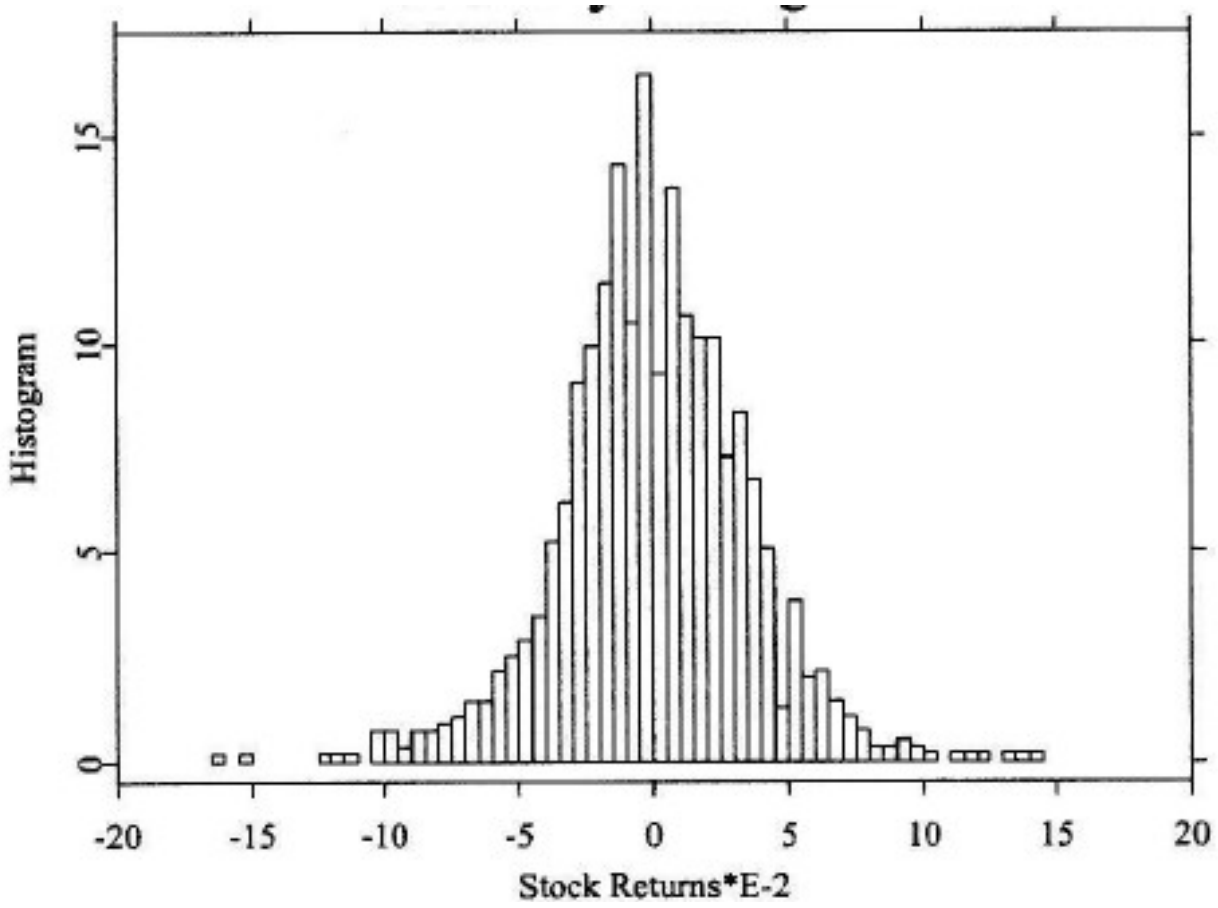
- «Наивная» формула

$$h_i = c \cdot n_i$$

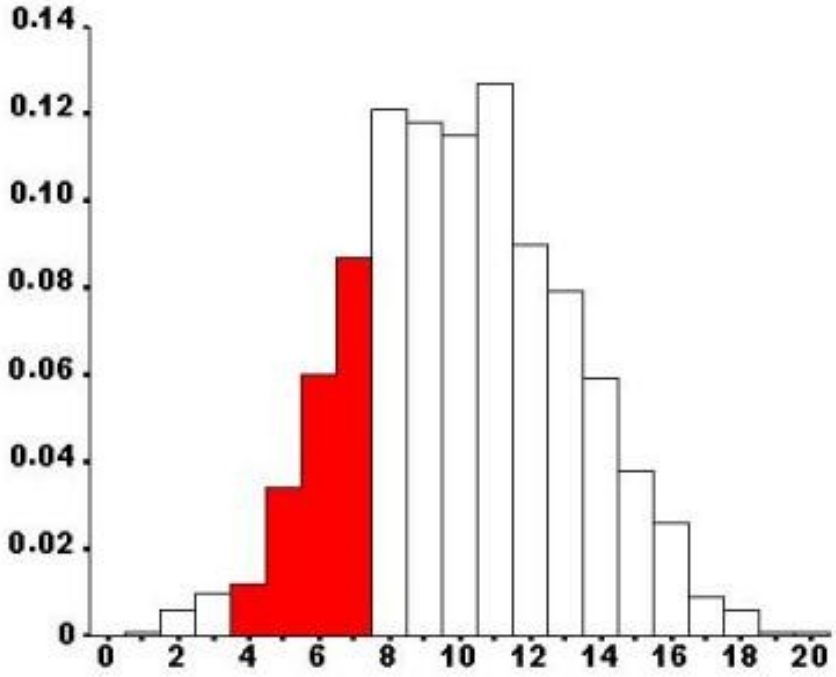
- «Научная» формула

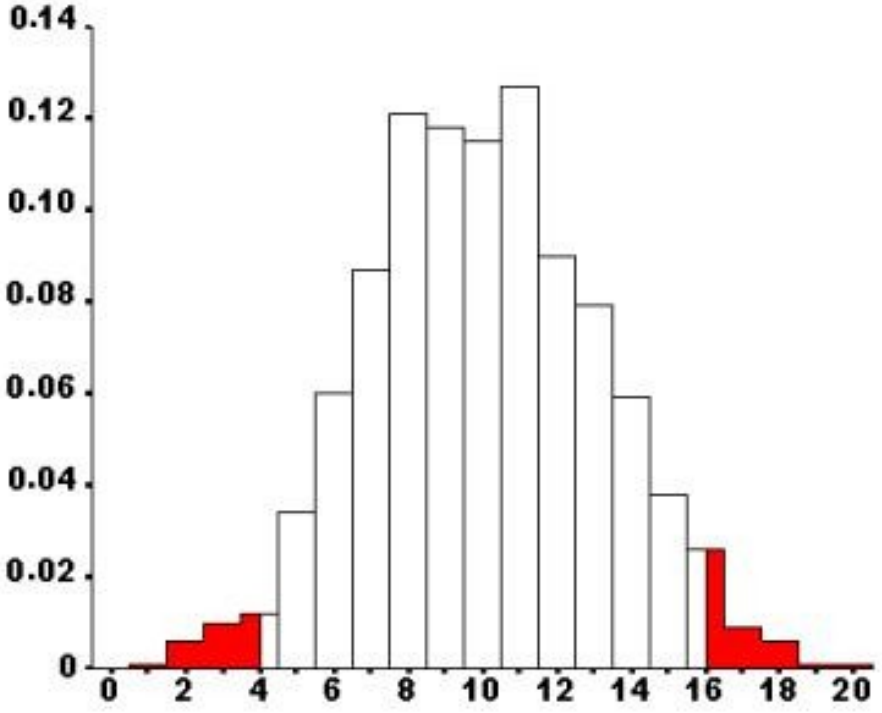
$$h_i = \frac{n_i}{n \cdot \left| \Delta_i \right|}$$

График

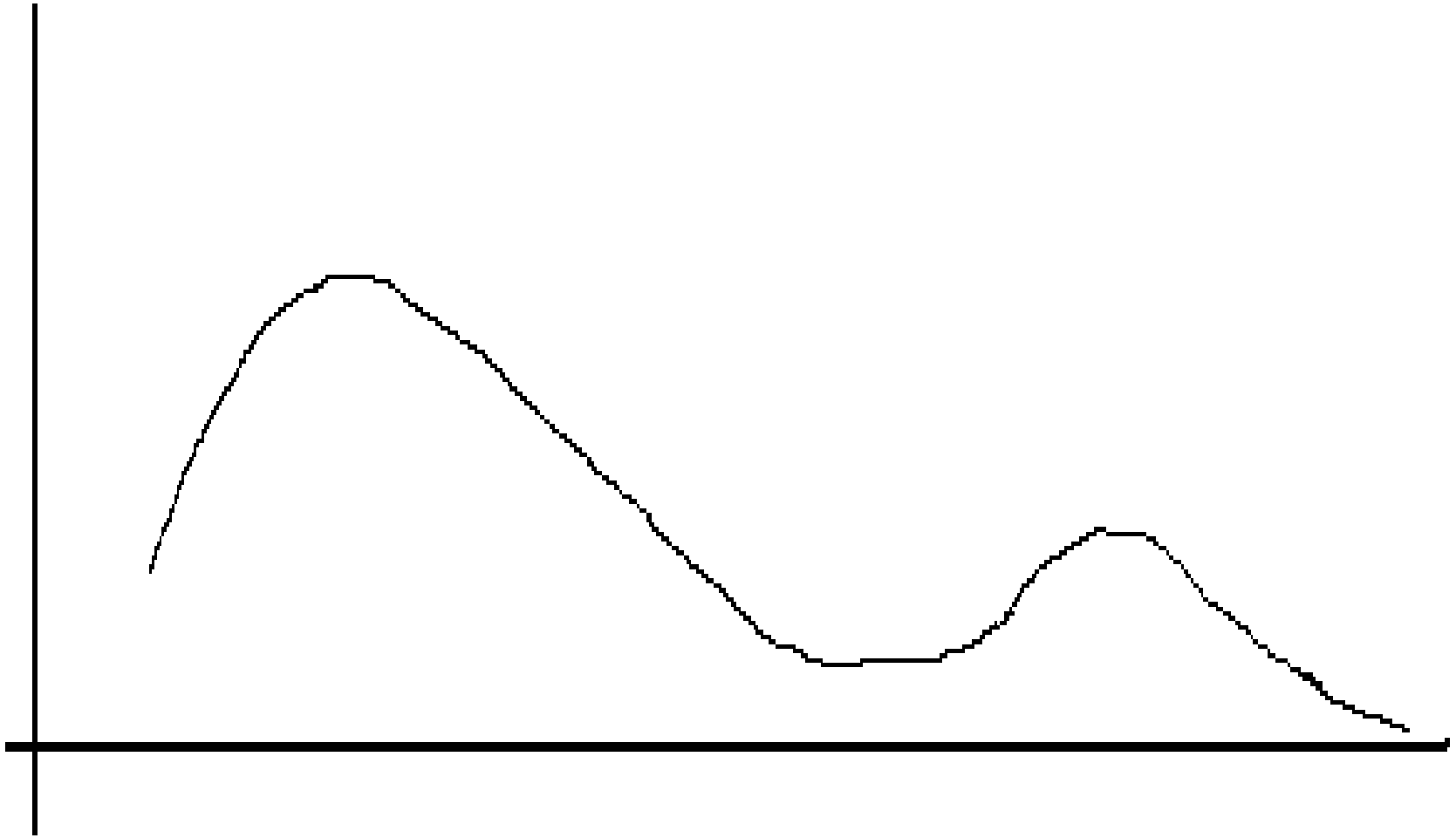


Анализ гистограммы – сравнение площадей

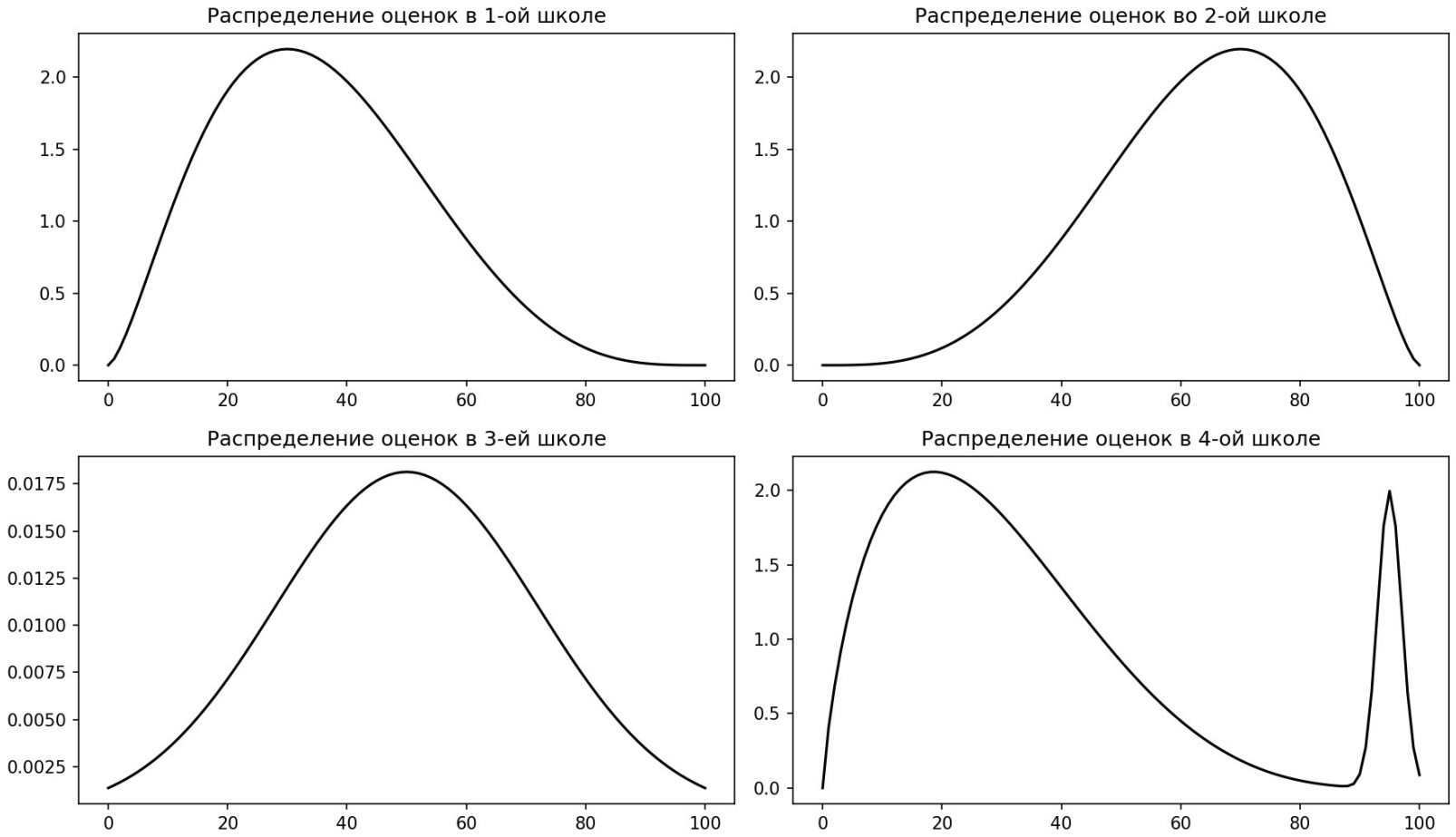




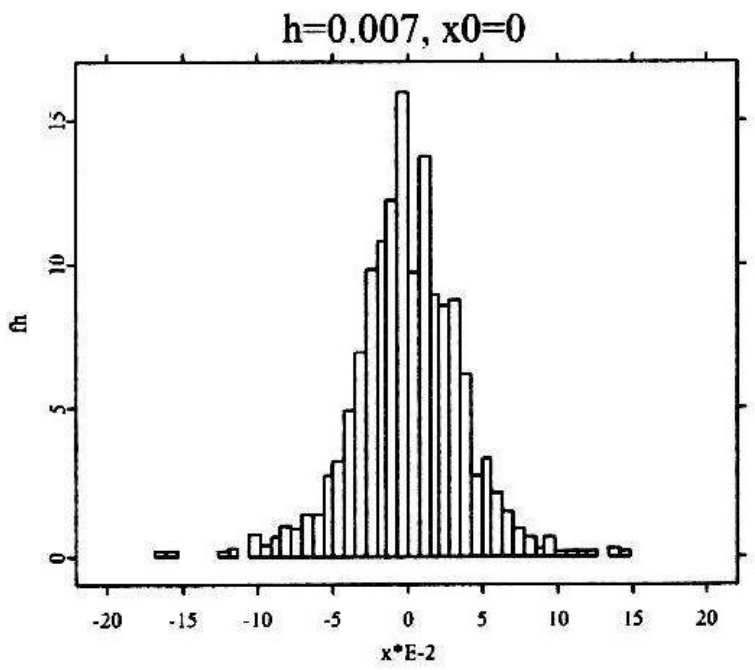
Форд мустанг

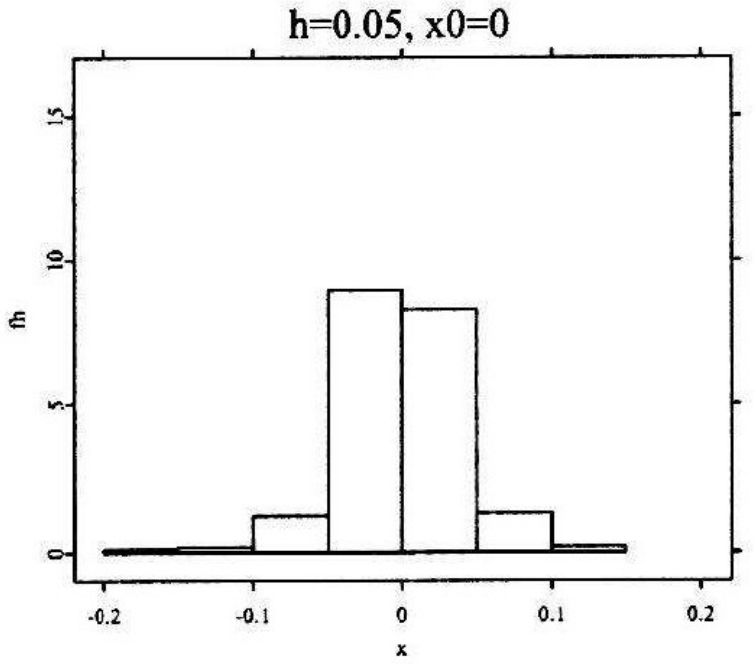


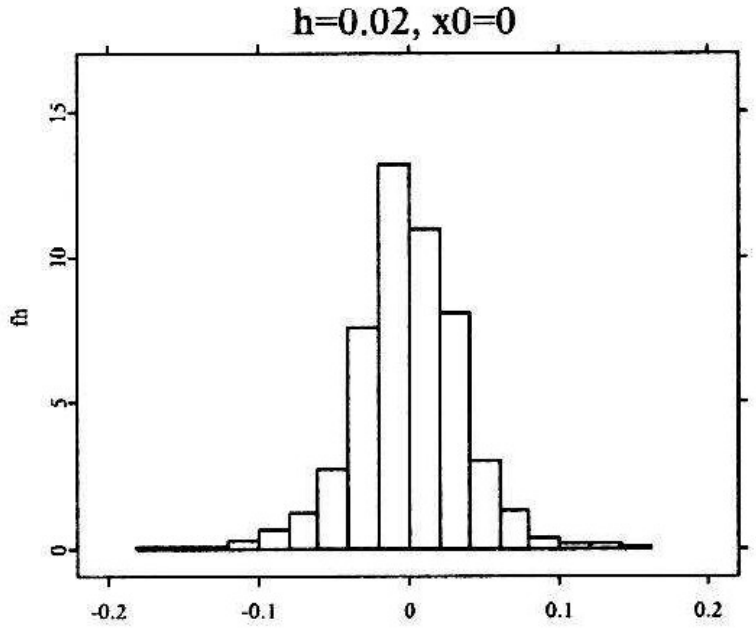
Выбор среди четырех школ

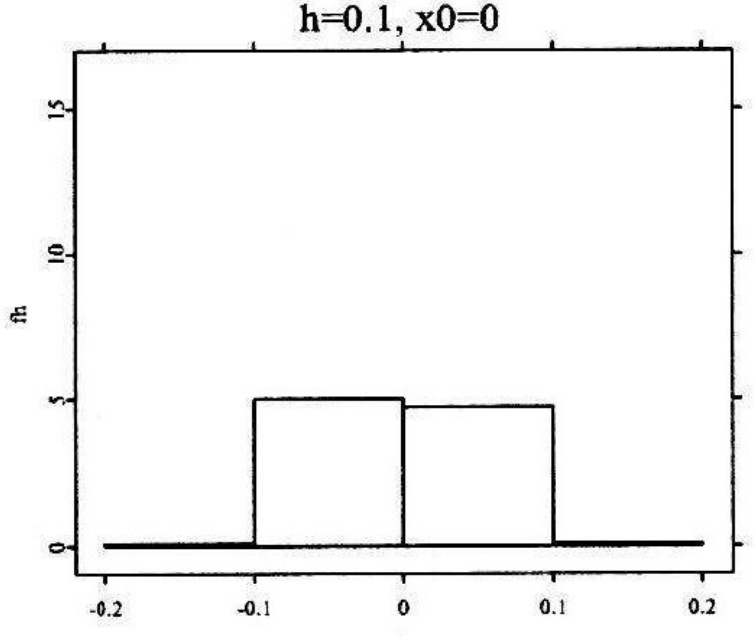


- Вид гистограммы зависит от ширины интервала

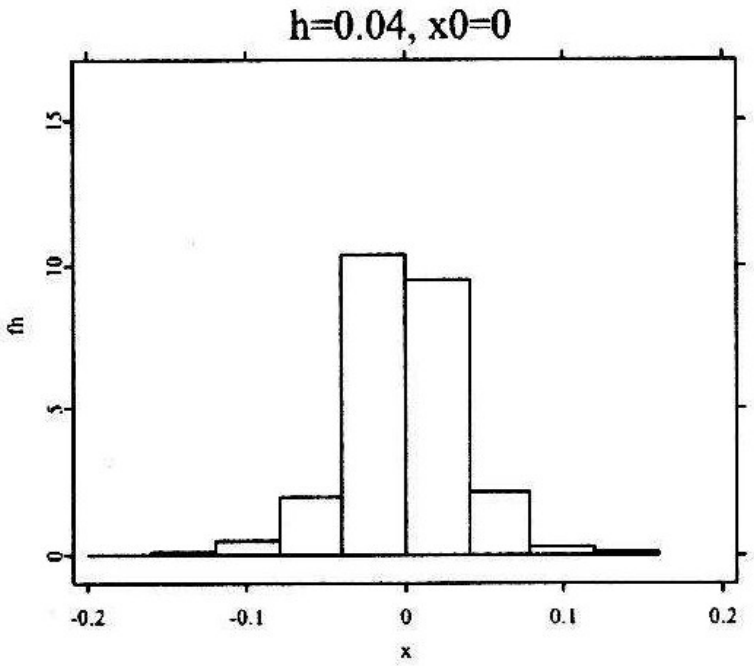

h=0.007, x0=0
fh
0
5
10
15
-20
-15
-10
-5
0
5
10
15
20
x*E-2

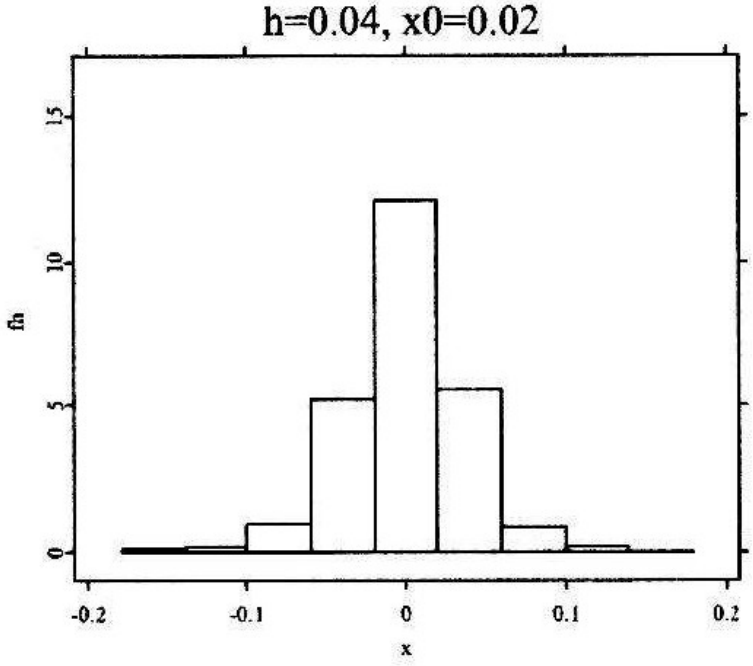

h=0.05, x0=0
fh
0
5
10
15
-0.2
-0.1
0
0.1
0.2
x

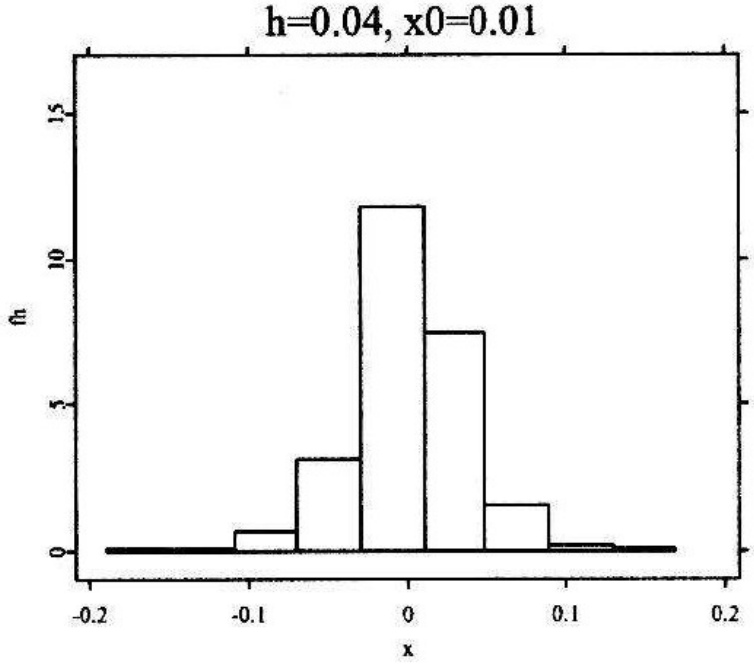

h=0.02, x0=0
fh
0
5
10
15
-0.2
-0.1
0
0.1
0.2

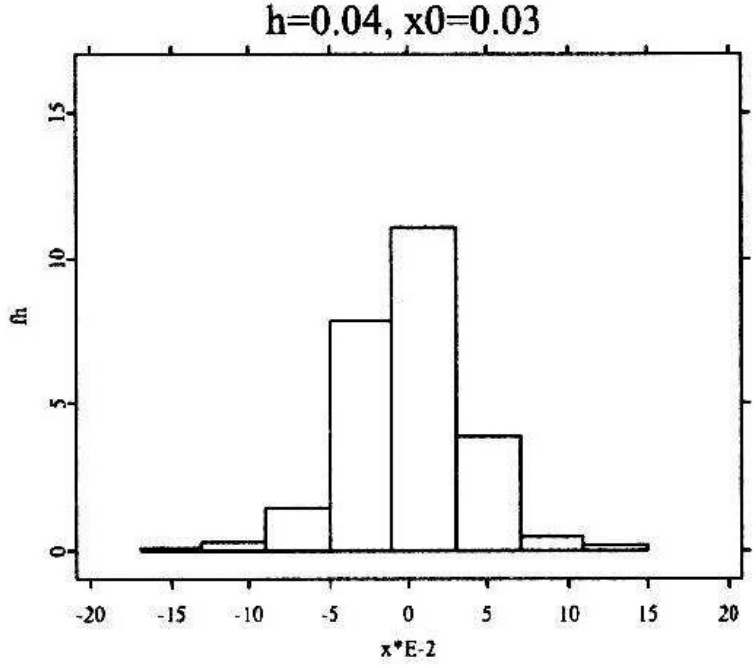

h=0.1, x0=0
fh
0
5
10
15
-0.2
-0.1
0
0.1
0.2

- Вид гистограммы зависит от выбора начальной и конечной точек a и b


h=0.04, x0=0
fh
0
5
10
15
-0.2
-0.1
0
0.1
0.2
x


h=0.04, x0=0.02
fh
0
5
10
15
-0.2
-0.1
0
0.1
0.2
x


h=0.04, x0=0.01
fh
0
5
10
15
-0.2
-0.1
0
0.1
0.2
x


h=0.04, x0=0.03
fh
0
5
10
15
-20
-15
-10
-5
0
5
10
15
20
x*E-2

Рекламная пауза

- Плотность распределения позволяет сосчитать вероятность

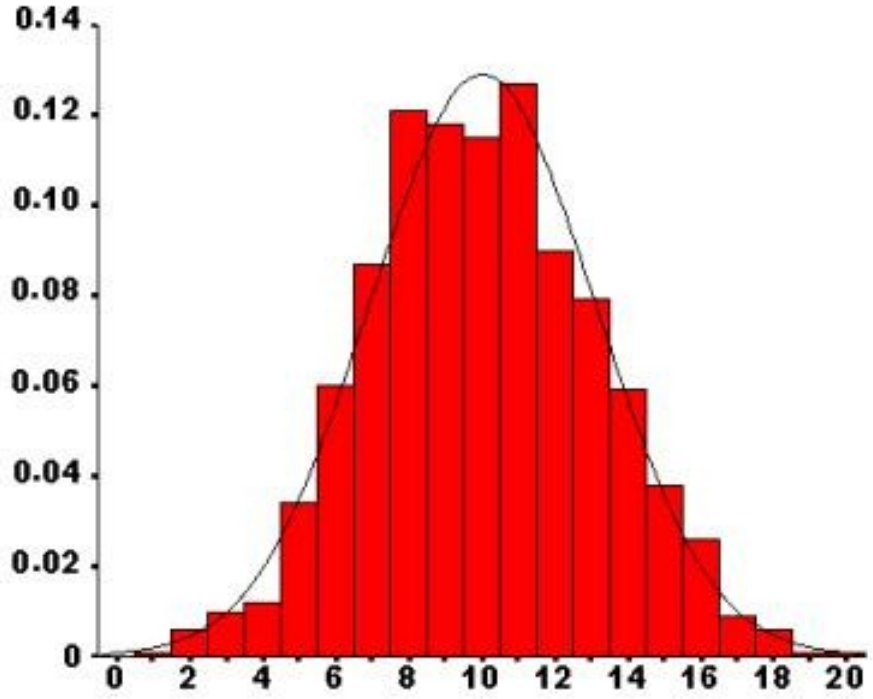
$$P\{ X \in [a,b)\} = \int_a^b f(t)\,dt$$

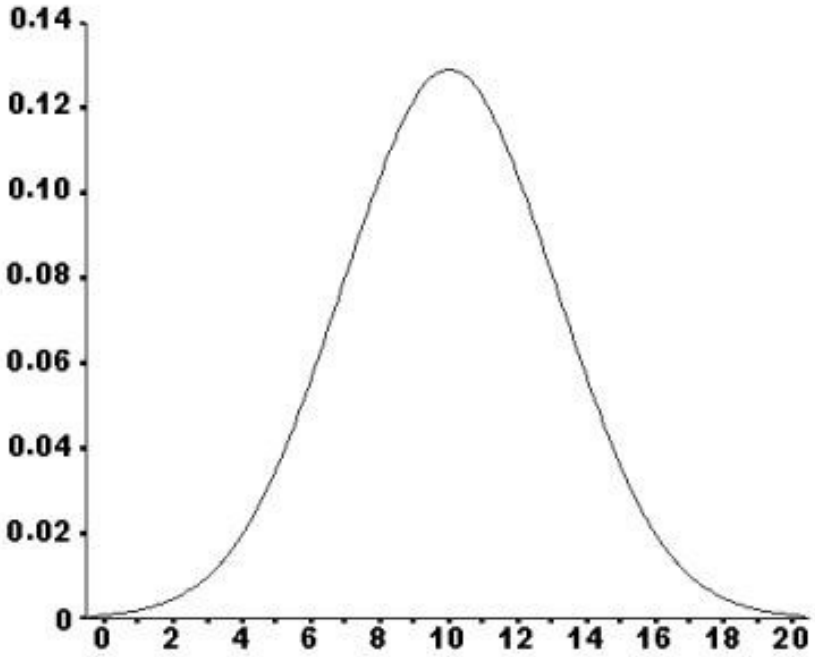
- Обычно плотность неизвестна
- Имеются «только» данные

- гистограмма является приближением к плотности распределения наблюдаемой величины

- Одновременно сама гистограмма является плотностью распределения, которым приближаем.

Гистограмма – оценка плотности





Гистограмма – оценка плотности. Обозначения

- Обозначим оцениваемую плотность распределения $f(t)$
- Обозначим гистограмму $f_n(t)$
- Пусть число интервалов k зависит от числа наблюдений n: $k = k_n$
- Обозначим $|\Delta| = \max\limits_{i=1,2\ldots,k_n} |\Delta_i|$

Гистограмма – оценка плотности. Предположения

- Пусть

$$\lim_{n\to\infty}|\Delta| = 0$$

$$\lim_{n\to\infty} k_n = \infty$$

$$\lim_{n\to\infty} \frac{k_n}{n} = 0$$

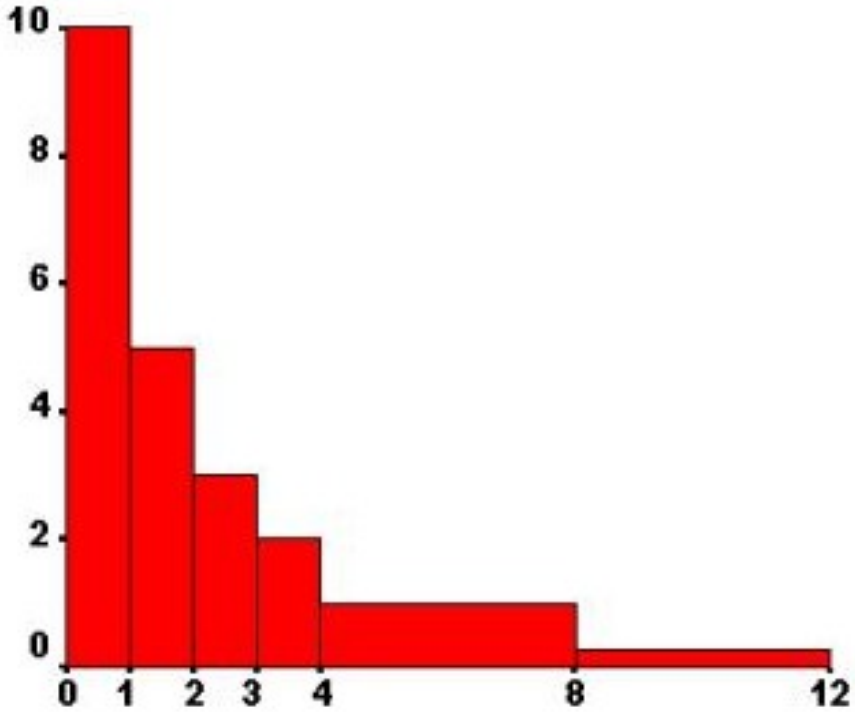
Сходимость по вероятности

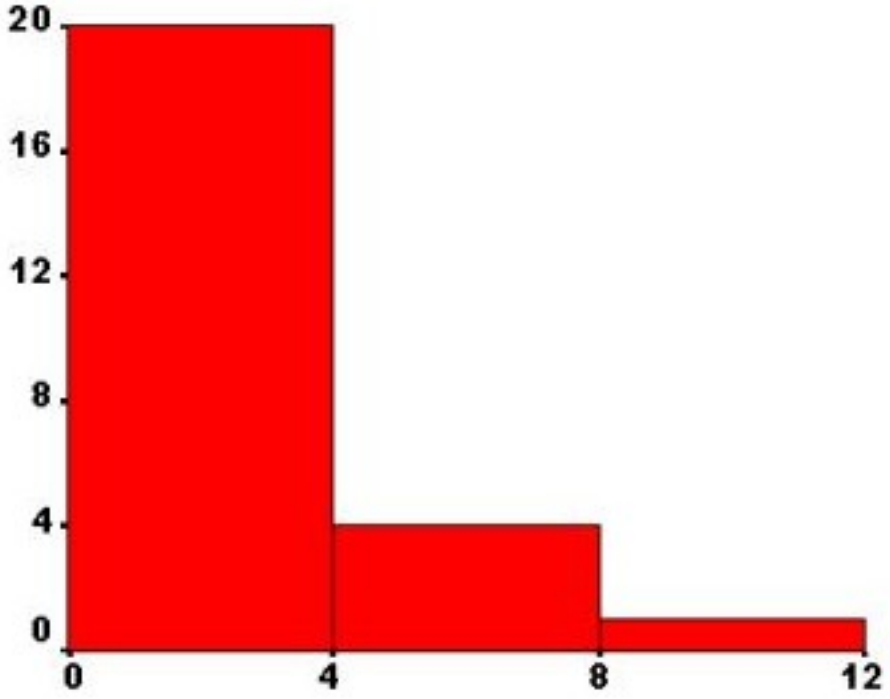
$$f_n(t) \xrightarrow[n\to\infty]{P} f(t)$$

Рекомендуется строить несколько гистограмм

- с разными значениями параметров
- Если время позволяет

Иногда интервалы надо брать разными





Достоинства гистограммы

- дает наглядное представление о распределении
- выбросы, мультимодальность
- позволяет сравнивать распределения
- просто вычисляется
- является оценкой плотности

Недостатки гистограммы

- выбор границ a и b произволен.
- выбор числа интервалов k произволен
- не непрерывна
- равна нулю вне интервала $[a, b]$
- неудобно на одном графике строить несколько гистограмм

Отступление. Формула Sturges’a

- Формула Sturges’a:
- *Ширина интервала вычисляется по формуле*
- $h = R / k$
  - где R – размах выборки,
  - $k = 1 + log_2 n$.

- Формула предложена в работе
- Sturges, H. (1926)

  The choice of a class-interval.

  J. Amer. Statist. Assoc., 21, 65–66.

- и вывод формулы содержит ошибки…

- Ошибки описаны в статье
- Rob J. Hyndman

  The problem with Sturges’ rule for constructing histograms

  1995

- тем временем применение формулы Sturges’a стало повсеместным

Альтернатива 1: формула Scott'a

- $h = 3.5s / n^{1/3}$.
- где s – выборочное стандартное отклонение
- Scott, D.W. (1979)

  On optimal and data-based histograms.

  Biometrika, 66, 605–610.

Альтернатива 2: формула Freedman и Diaconis

- $h = 2IQ / n^{1/3}$
- *где IQ – межквартильный размах*

- Freedman, D. and Diaconis, P. (1981)

  On the histogram as a density estimator: L2 theory.

  Zeit. Wahr. ver. Geb., 57, 453–476.

Куда все смотрели 100 лет?

- Если размер выборки меньше 200 наблюдений, все формулы дают близкие результаты.
- Если больше 200 наблюдений, то формула Sturges’a занижает число интервалов.
- Формула Freedman и Diaconis’a устойчивее к выбросам.

2. Обобщения гистограммы

- Непараметрические оценки плотности
  - Ядерные оценки плотности

Ядерные оценки плотности

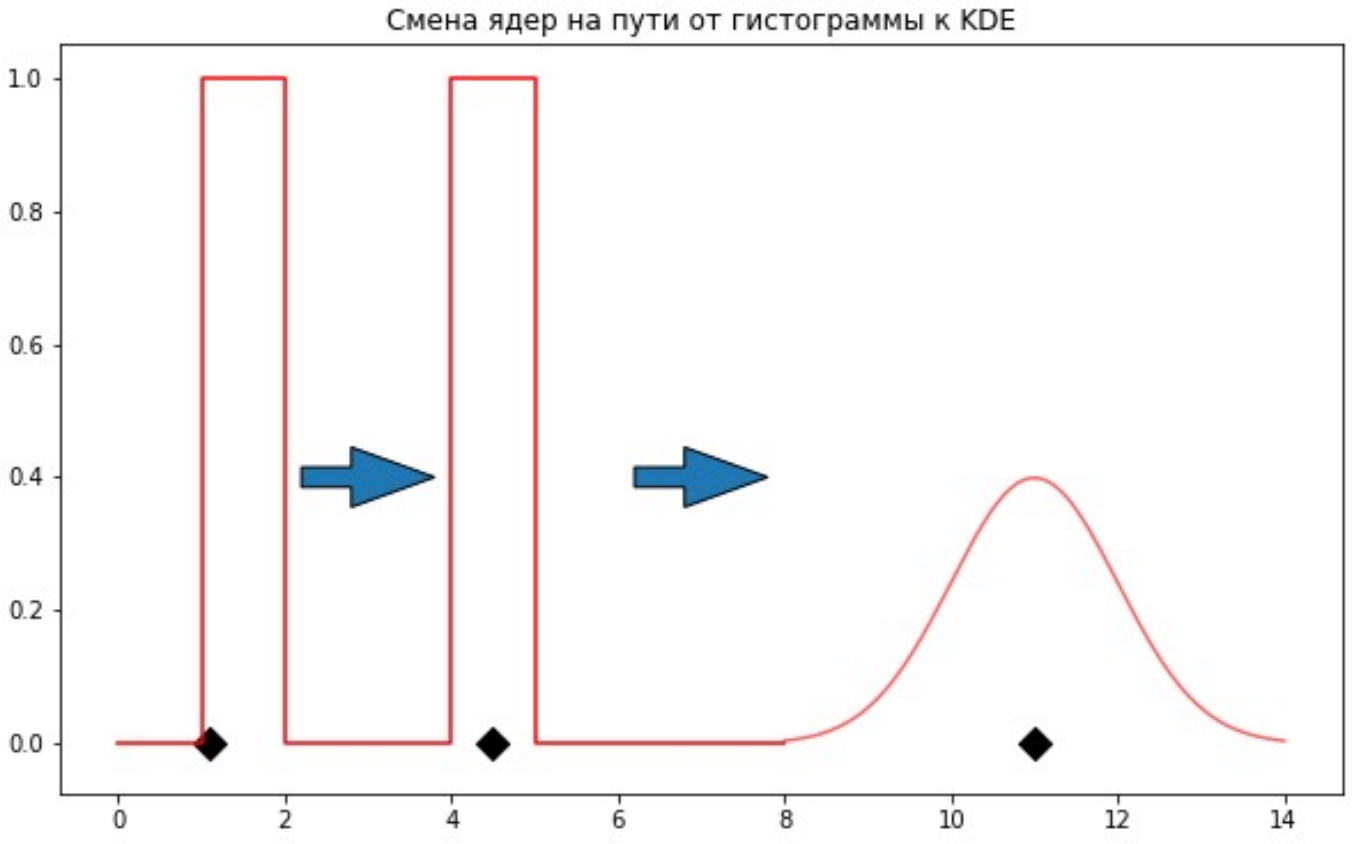
- Над каждой точкой – функция.
- Функции усредняем

- Площадь подграфика среднего равна 1.

Новый взгляд на гистограмму

- Гистограмма – функция
- Нормированная гистограмма – плотность распределения
- Гистограмма –среднее арифметическое плотностей распределения
- Каждому наблюдению соответствует своя плотность равномерного распределения

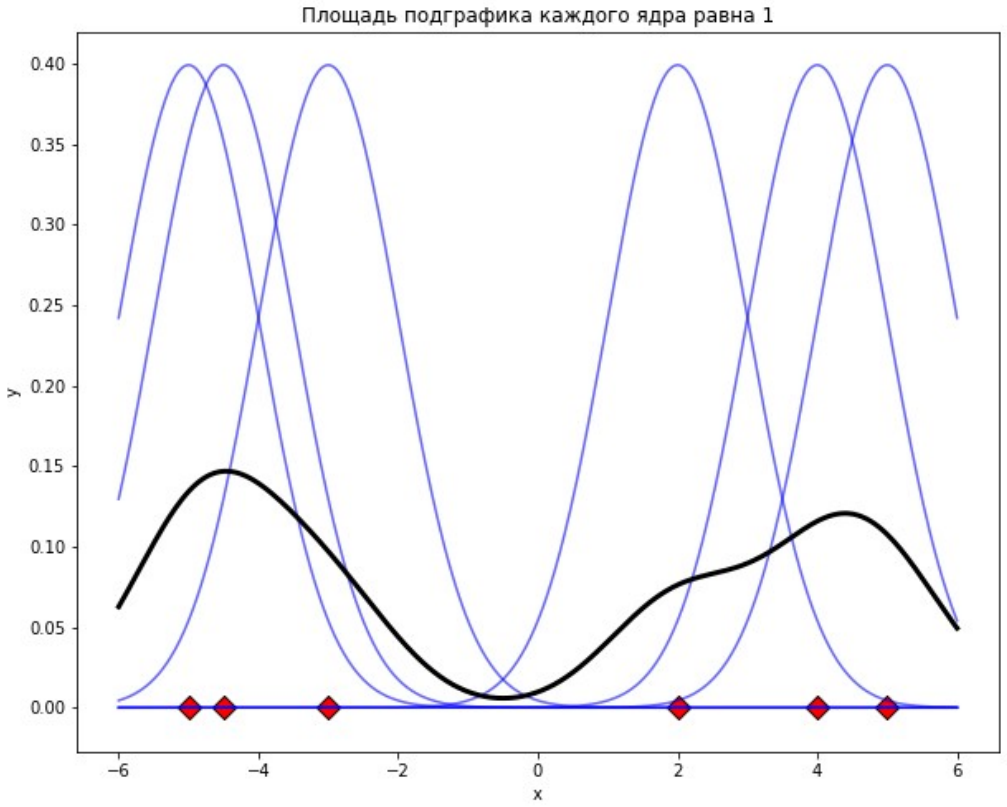
Улучшаем вид ядер



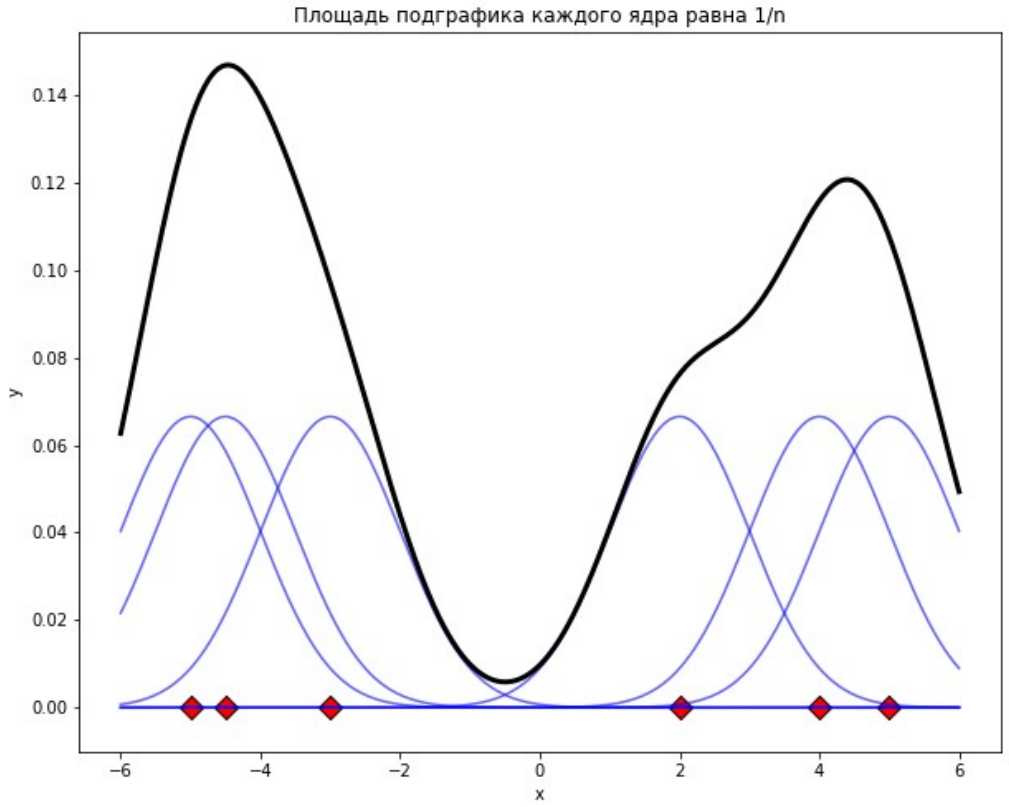
Ядерная оценка плотности

- Чтобы получить ядерную оценку плотности, считаем среднее значение этих функций

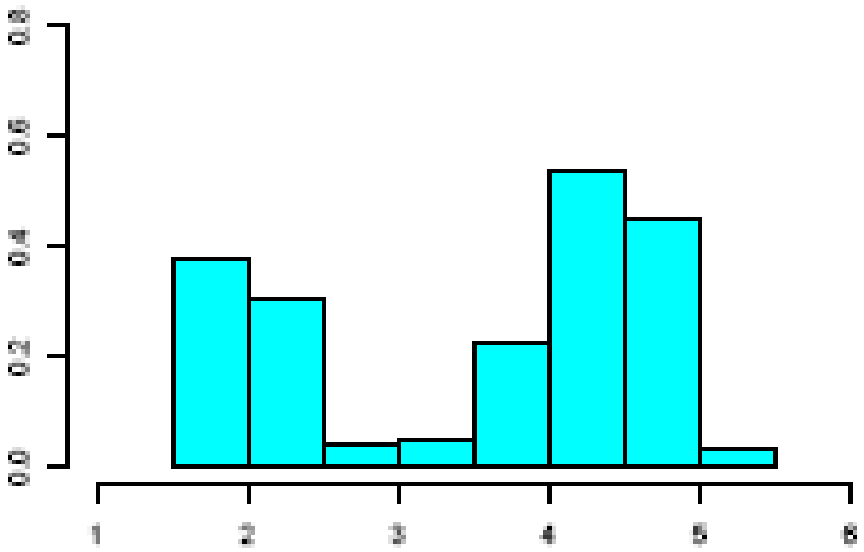
Правильный вариант: усреднение

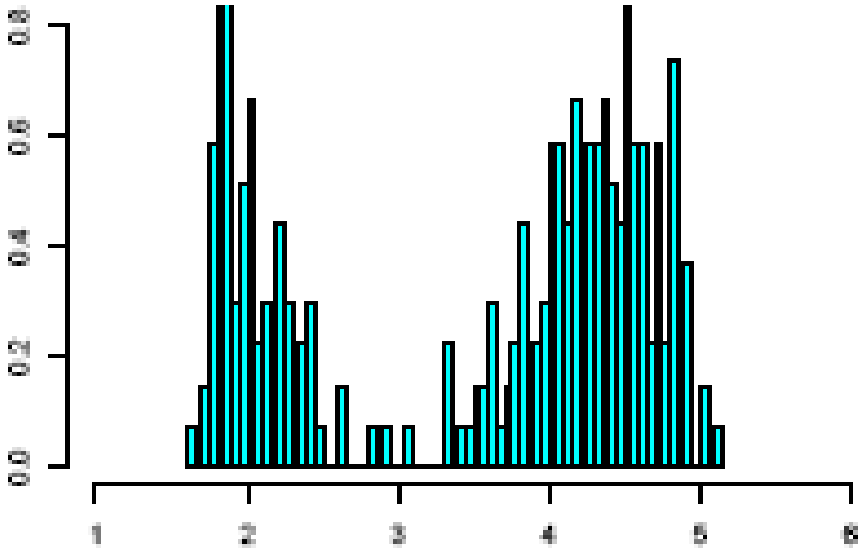


Наглядный вариант: сумма

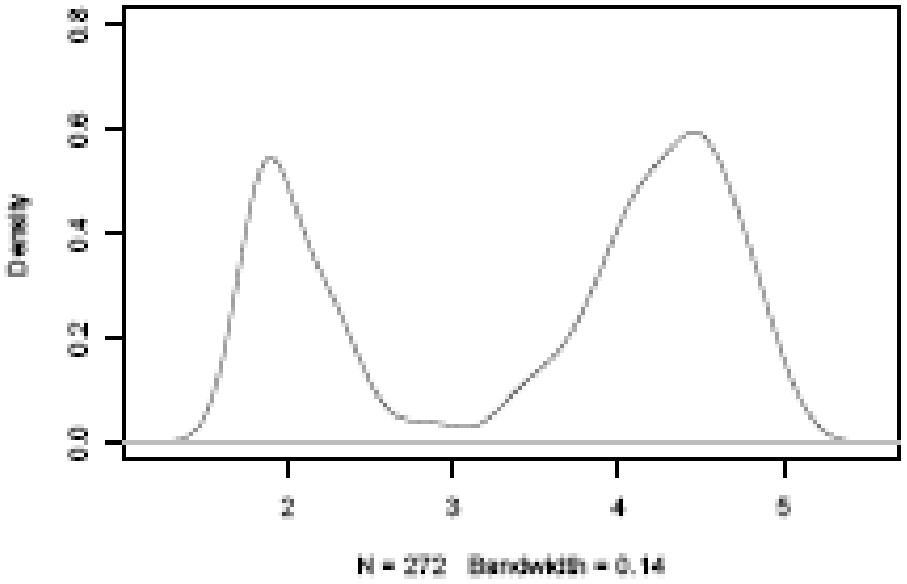


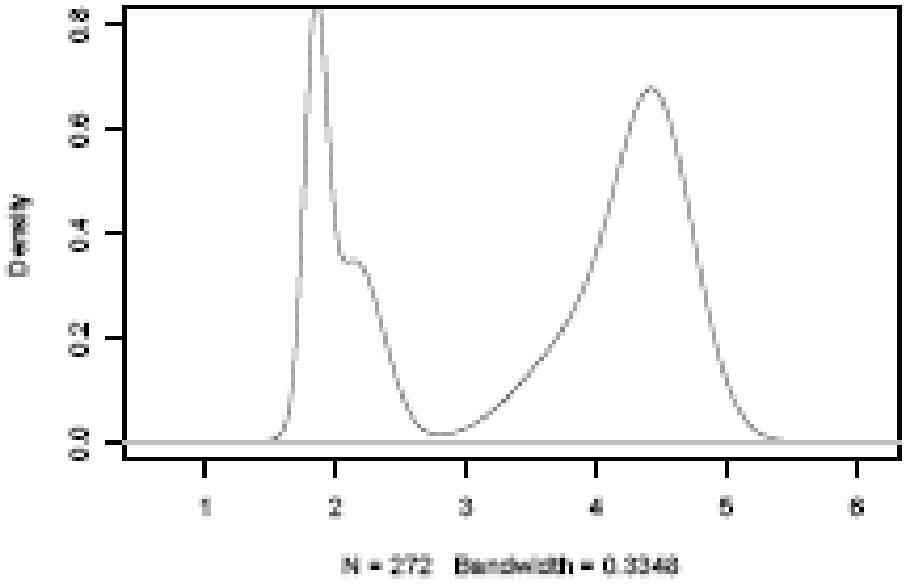
Сравним гистограмму и





ядерную оценку плотности





- Обобщения гистограмм называют
- «непараметрические оценки плотности»
- Ядерные оценки плоттности

Реализация ядерных оценок плотности в Питоне

- Убогая формула для вычисления ширины окна (предполагается нормальность)
- Мало вариантов ядра (несущественно)
- Нет модификаций для случая интервалов

- Яркий пример:
- seaborn.kdeplot и scipy.stats.gaussian_kde

Первые попытки процедур построения KDE

- В R в 2011 году было 20+ пакетов (Deng, Wickham Density estimation in R)
- **KDEpy**
- Все остальные предполагают нормальность оцениваемого распределения

В питоновском сообществе

Kernel Density Estimate

называют

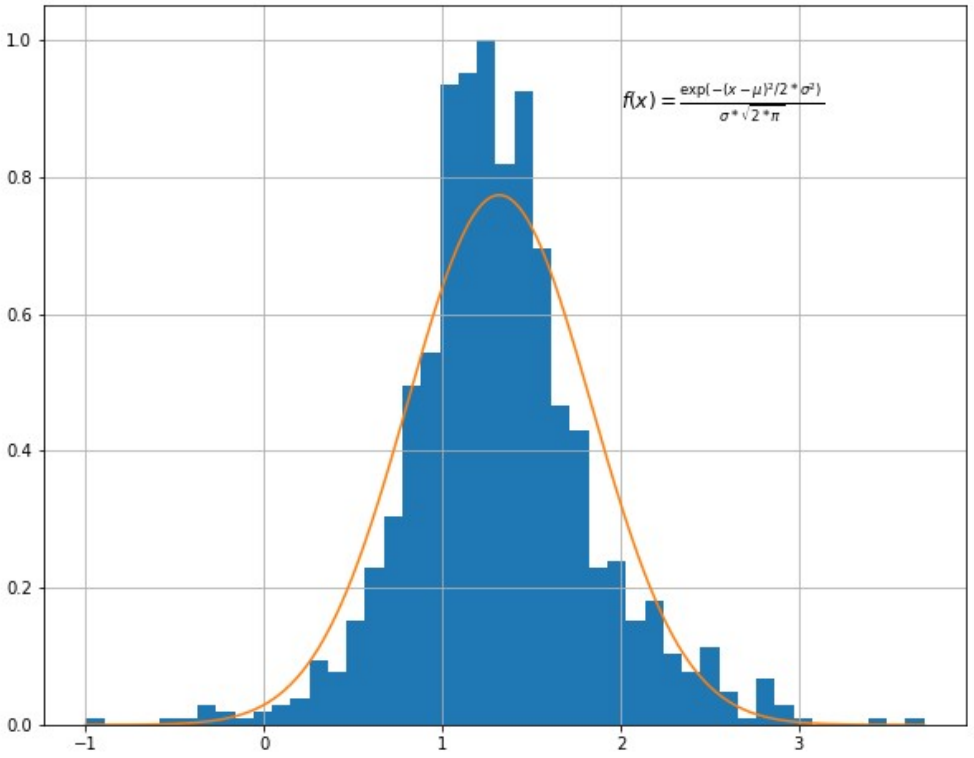
Density Plots

Многие уверены, что KDE разгадывает истинную плотность распределения переменной

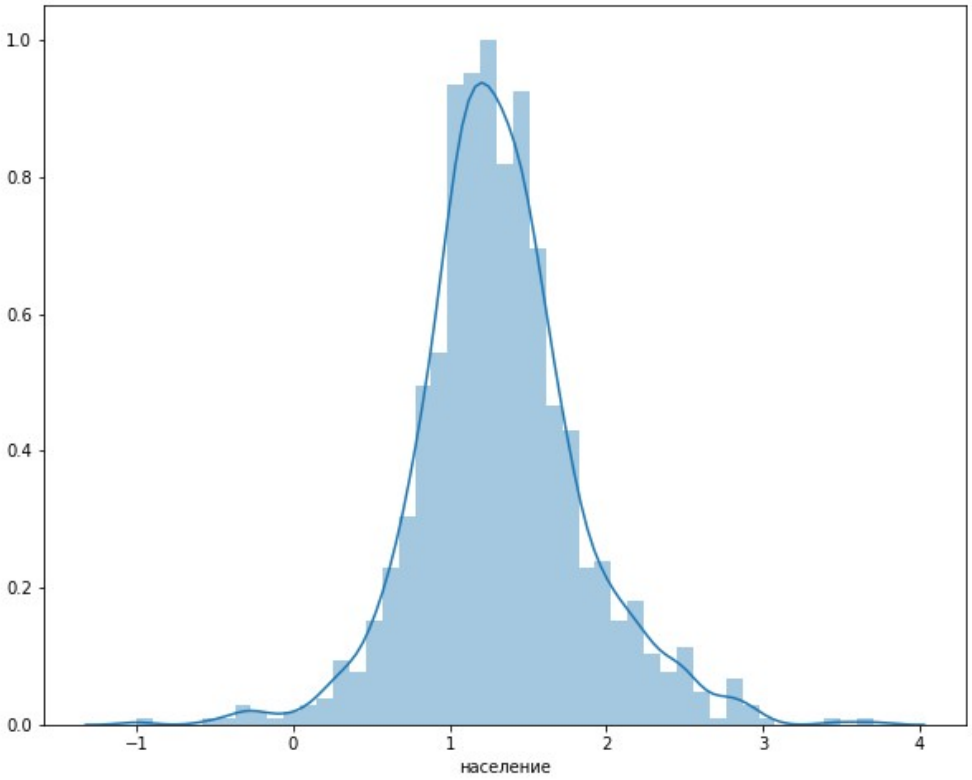
Поэтому два графика

Гистограмма и нормальная плотность

Оцен

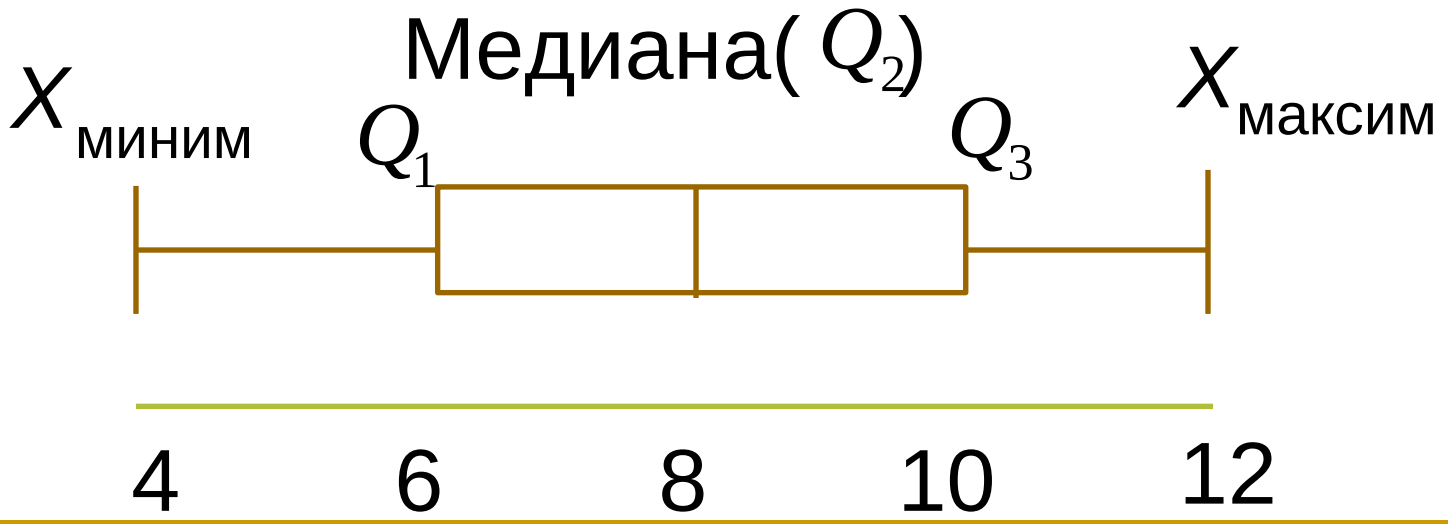


Гистограмма и KDE
Одно и то же, два раза



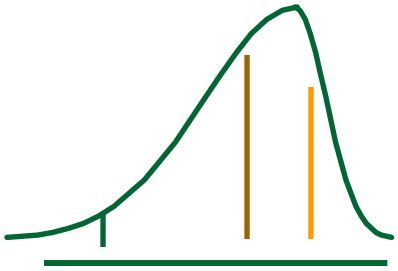
5-числовая сводка

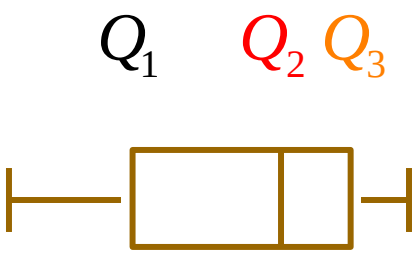
- минимальное,Q1,медиана, Q3, максимальное
- Ящик с усами - график 5-числовой сводки



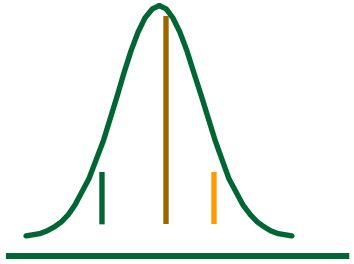
Форма распределения и ящиковая диаграмма

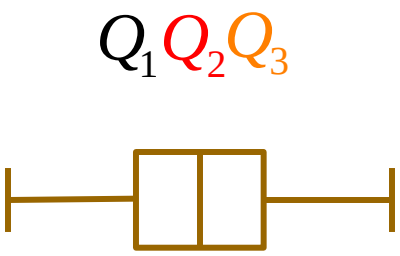
Скос влево (отрицательный) среднее < медианы



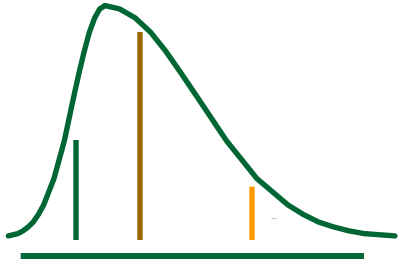


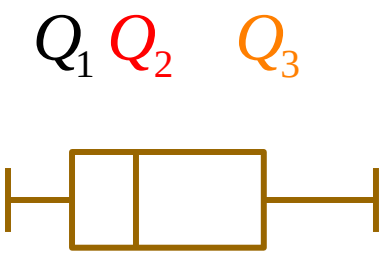
Симметричное (ноль)

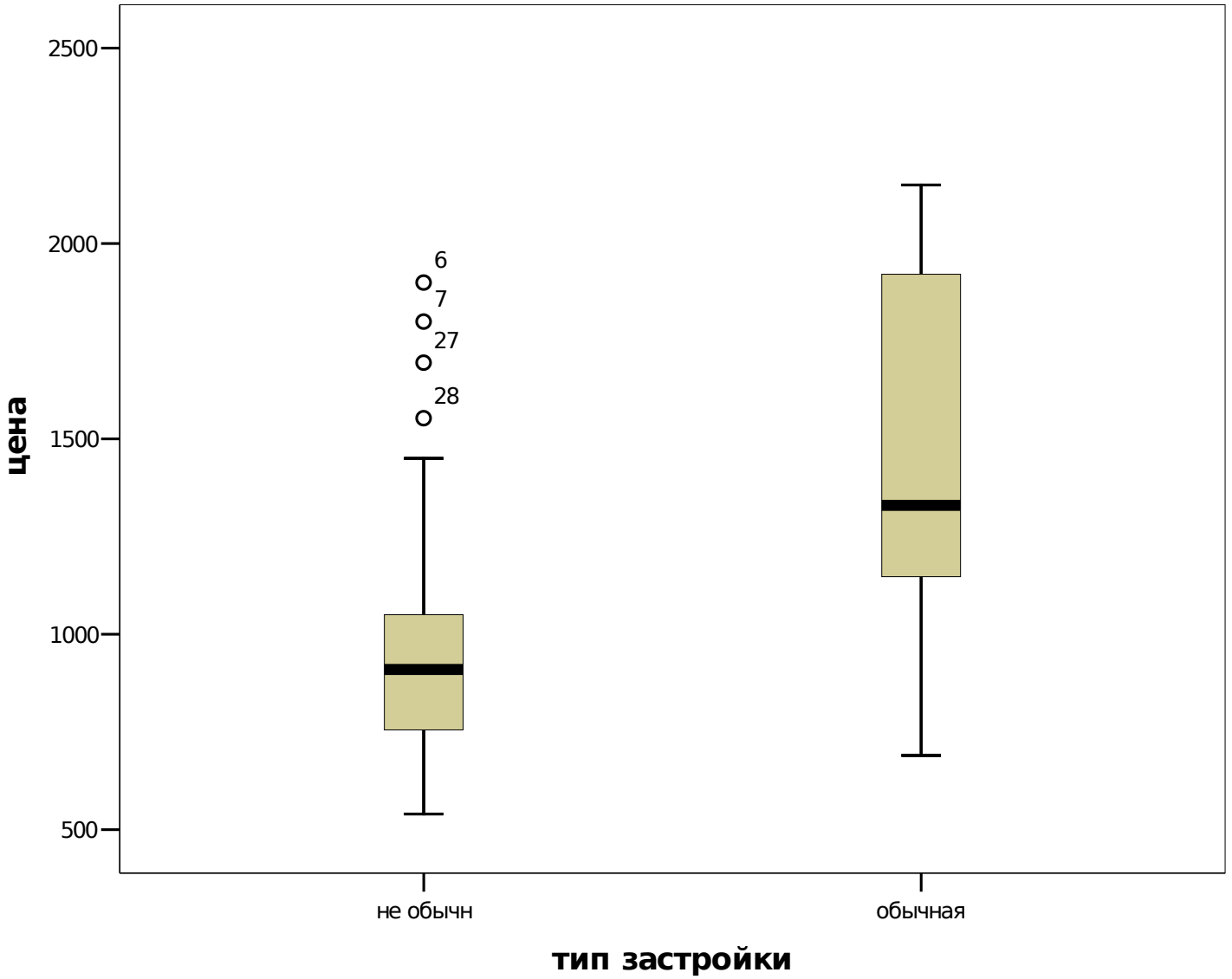


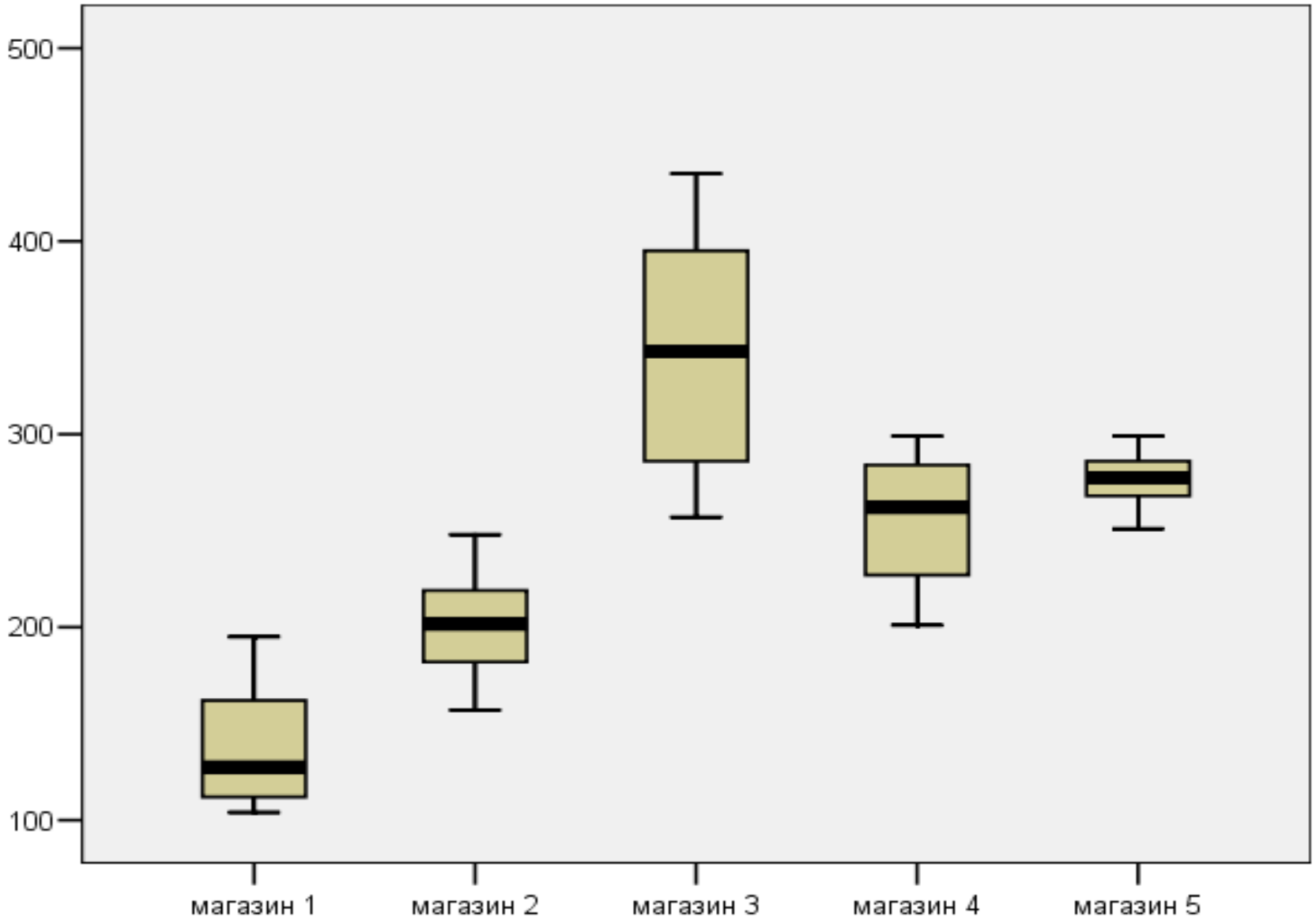


Скос вправо (положительный) среднее > медианы






2500
2000
1500
1000
500
цена
6
7
27
28
не обычн
обычная
тип застройки

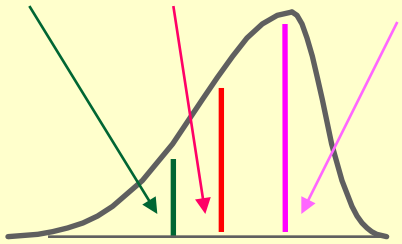

500
400
300
200
100
магазин 1
магазин 2
магазин 3
магазин 4
магазин 5

Форма распределения - терминология

- Для оценки формы используется гистограмма или ящики с усами
- Форма симметричная или скошенная
- взаимное расположение среднего и медианы
- Но есть еще мультимодальные распределения

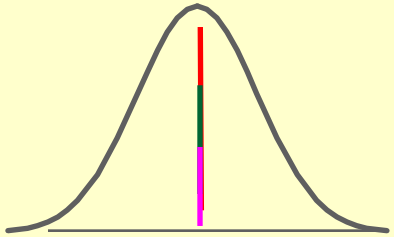
Left-Skewed

среднее < медиана(< мода?)



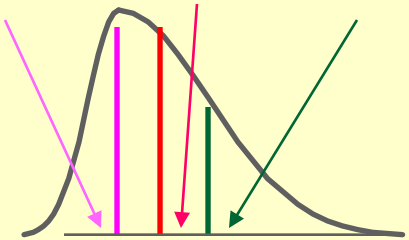
симметричная

среднее = медиана(=мода?)



Right-Skewed

(мода<?) медиана < среднее



Поиск наилучшей проекции (Projection Pursuit, Independent Component Analysis).

- Многомерные данные
- Проклятье размерности
- Гистограммы неприменимы
- Поиск наилучшей проекции

Многомерные данные

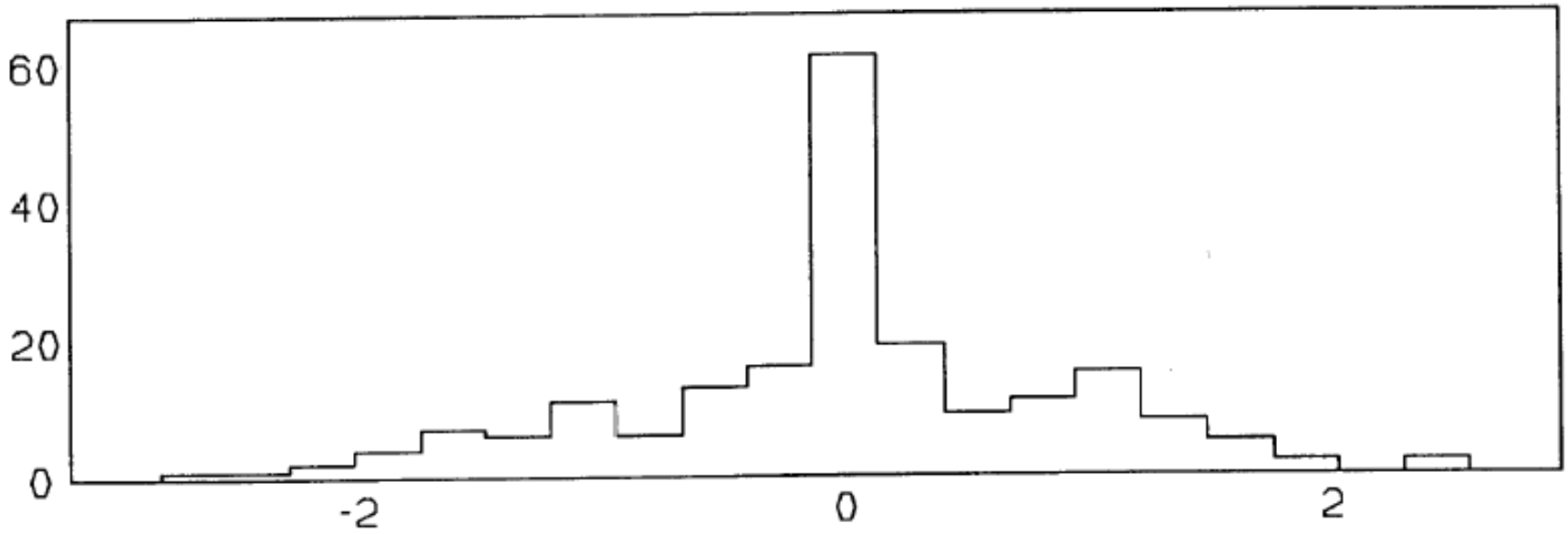
- Анализ главных компонент
- Факторный анализ

- Зачем нужно что-то еще?
- Не ориентированы на поиск кластеров в данных

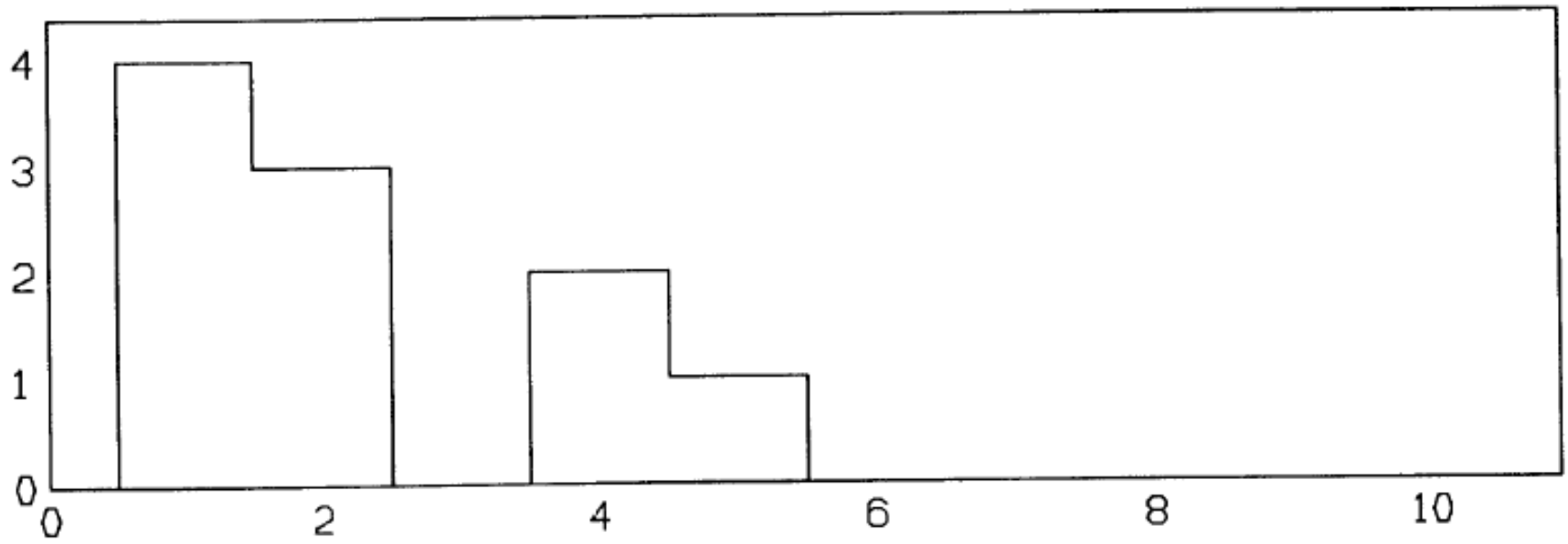
Многомерные данные

- Поиск наилучшей проекции

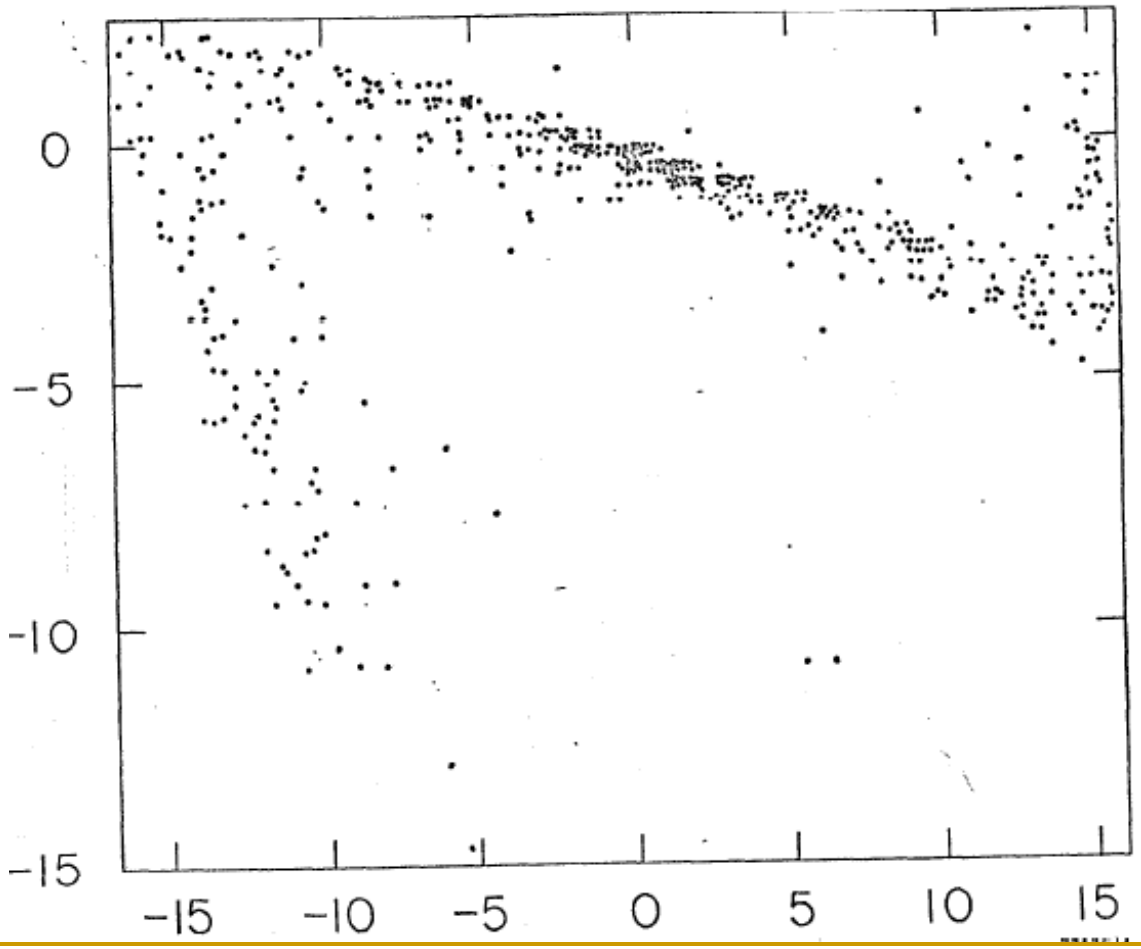
Пример 1
Проекция на произвольную прямую



Пример 1
Информативная проекция



Пример 2



Кварталы города Бостон 1

- 506 наблюдений для каждого квартала города Бостон и его пригородов.
- Кварталы определялись так же, как и при переписи населения США.
- Фиксировалось 13 независимых наблюдений и одна зависимая.

Кварталы города Бостон 2

- X1 = уровень преступности
- X2 = разреженность населения (обратно к плотности)
- X3 = доля предприятий не торгового характера среди всех предприятий квартала
- X4 = район у реки (1 = да, 0 = нет)
- X5 = загазованность
- X6 = среднее число комнат в квартире/доме среди домов этого квартала
- X7 = доля старых домов (построенных до 1940)

Кварталы города Бостон 3

- X8 = расстояние до промышленных/деловых районов города (взвешенное)
- X9 = расстояние до радиальных автодорог
- X10 = величина налога на недвижимость в районе в $10,000
- X11 = число учеников не одного учителя (в школах района)
- X12 = 1000(B - 0.63)^2, где B - доля афроамериканцев
- X13 = процент населения с низким соц. статусом
- X14 = медиана цен жилых домов в 1000 долларов

Кварталы города Бостон 4

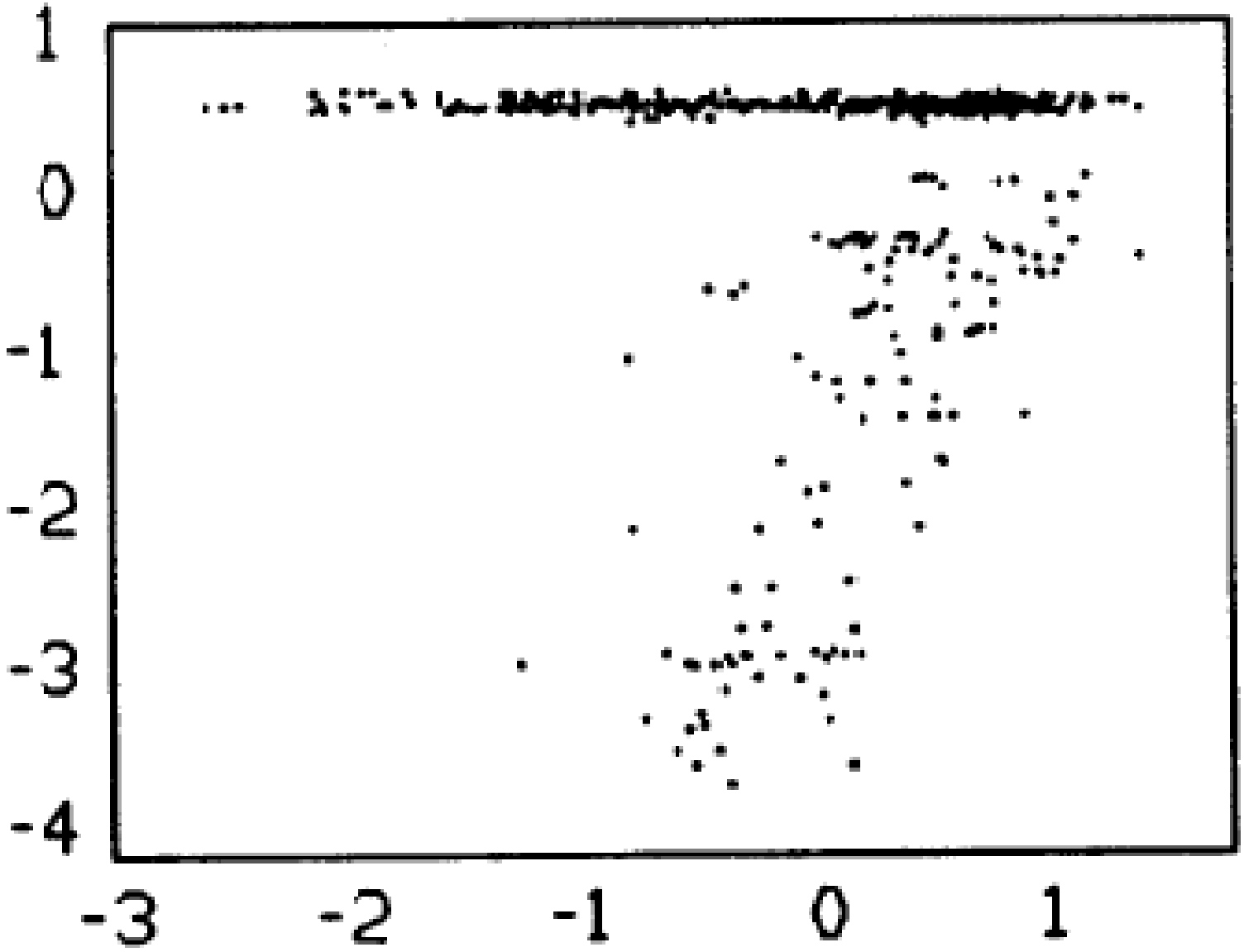
- Преобразуем переменные

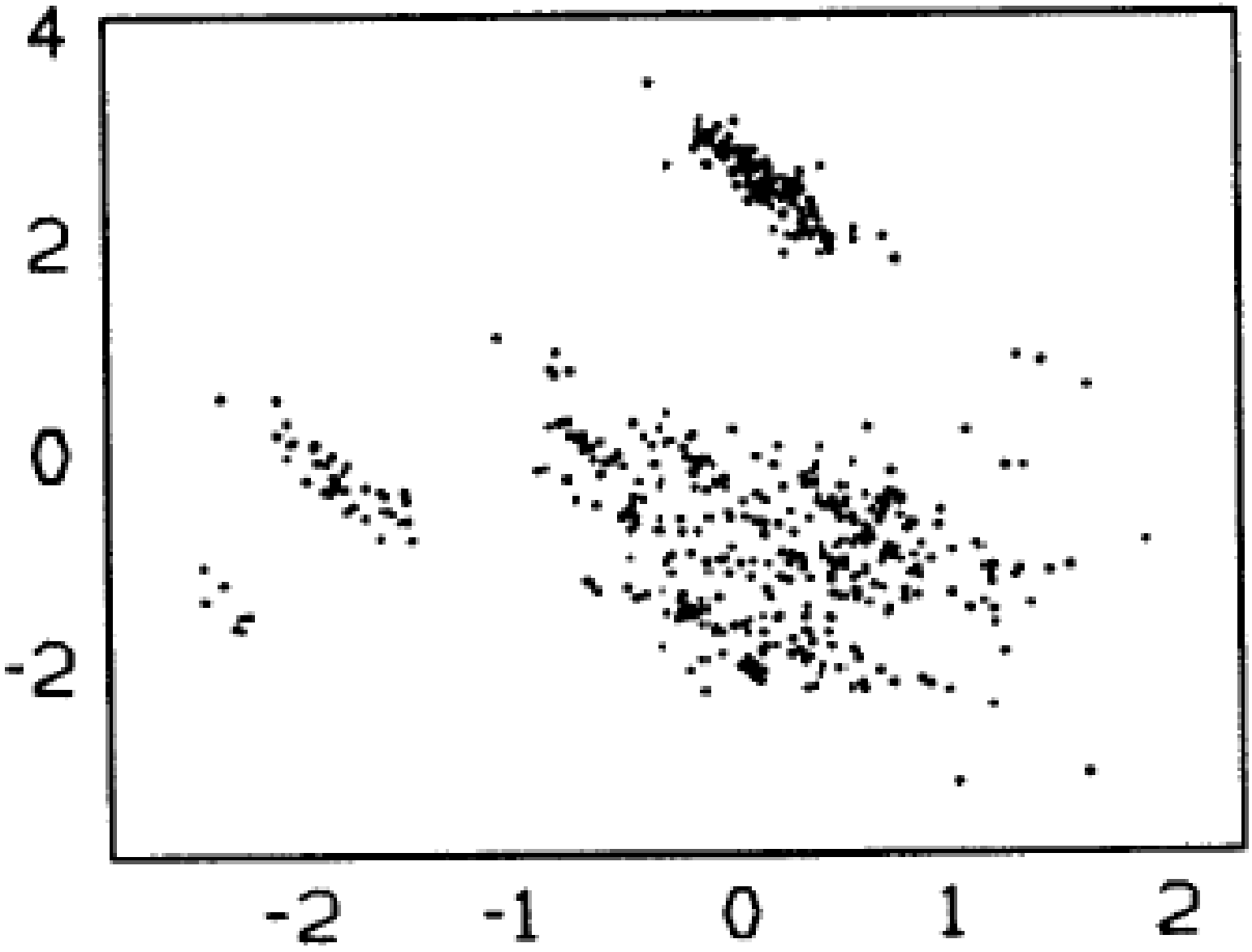
Кварталы города Бостон 5

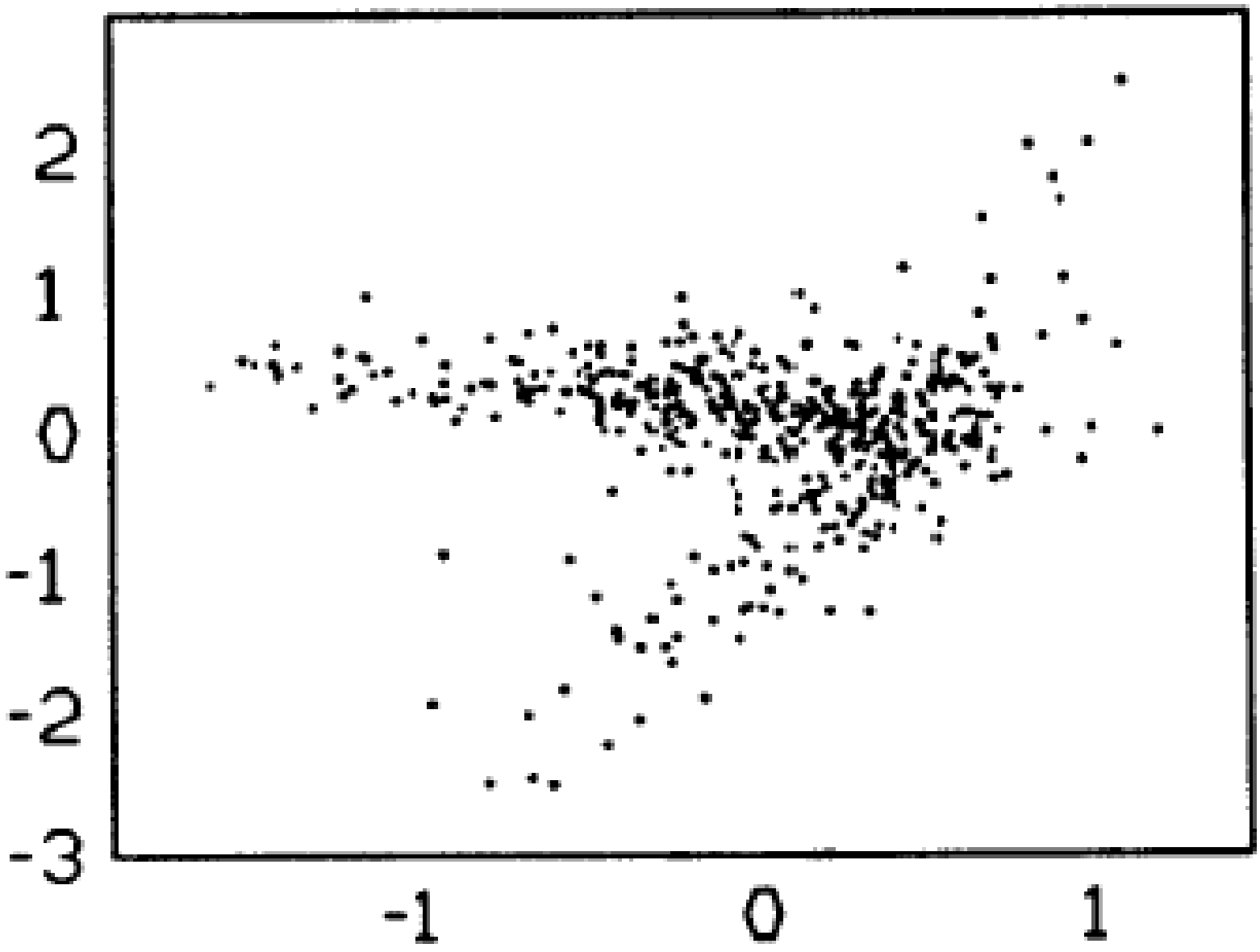
- X1 = ln(уровень преступности)
- X2 = разреженность населения (обратно к плотности)
- X3 = доля предприятий не торгового характера среди всех предприятий квартала
- X4 = район у реки (1 = да, 0 = нет)
- X5 = (загазованность)^2
- X6 = (среднее число комнат в квартире/доме среди домов этого квартала)^2
- X7 = доля старых домов (построенных до 1940)

Кварталы города Бостон 6

- X8 = ln(расстояние до промышленных/деловых районов города (взвешенное))
- X9 = ln(расстояние до радиальных автодорог)
- X10 = величина налога на недвижимость в районе в $10,000
- X11 = число учеников не одного учителя (в школах района)
- X12 = ln(0.4-(B - 0.63)^2), где B - доля афроамериканцев
- X13 = ln(процент населения с низким соц. статусом)
- X14 = ln(медиана цен жилых домов в 1000 долларов)


1
0
-1
-2
-3
-4
-3
-2
-1
0
1


4
2
0
-2
-2
-1
0
1
2


2
1
0
-1
-2
-3
-1
0
1

- Подобные графики указывают, что надо делать! (порождают гипотезы)

Почему несколько решений?

- Так как ответ получен в результате минимизации некоторой функции,
- было найдено несколько локальных минимумов.

- Локальные минимумы тоже могут быть интересны!

- При чем тут гистограммы и непараметрические оценки плотности?

- Какая проекция интересная?
- Какая проекция неинтересная?

- Неинтересна проекция, у которой нормальное распределение

- Интересна проекция,
- распределение которой максимально отклоняется от нормального

- Измерим отклонение от нормальности
- Индекс качества проекции

$$I_1 = \int_{-\infty}^{\infty} \left(f_{np}(t) - \varphi(t) \right)^2 dt$$

- При этом
- плотность нормального распределения

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \exp\left(- \frac{t^2}{2} \right)$$

- оценка плотности проекции

$$f_{np}(t)$$

- Откуда берется оценка плотности распределения $f_{np}(t)$

- Это гистограмма
- или ядерная оценка плотности

- Наилучшая проекция доставляет максимум индексу проекции.
- Процедура минимизации облегчается, когда оценка плотности гладкая.
- Отсюда преимущество ядерных оценок перед гистограммой.

Independent component analysis (ICA)

- Возможен другой индекс качества проекции, основанный на энтропии

$$H_{np} = -\int_{-\infty}^{\infty} g(s)\ln\left(g(s)\right)ds$$

Многомерное шкалирование

- Минимальное отличие суммы расстояний между точками в исходных данных и в проекции
- Минимальное отличие рангов расстояний между точками в исходных данных и в проекции

- В настоящее время информативные переменные (features) часто ищут с помощью глубокого обучения
- Для человека с кувалдой весь мир выглядит как гвоздь.