Математическая статистика

и основы анализа данных

Введение

Организационная информация

Первый семестр:

- Статистика
- Введение в анализ данных
- Введение в машинное обучение

Второй семестр:

- Машинное обучение

Лектор: Ольга Калиниченко

Организационная информация

Профили:

- Физика;
- Биология;
- Педагогика;
- Теоретический (при наличии желающих).

Студенты кафедры инновационной педагогики выбирают профиль "педагогика".

Организационная информация

Структура курса:

- Домашние задания на 2 недели;
- Несколько теоретических и практических задач;
- Практика на языке Python;
- Сдача заданий через телеграм-бот (подробности до конца недели);
- Гостевые лекции про реальное применение статистики (отдельно для направлений)

Первые две лекции вводные, первые два задания даются на одну неделю и являются отборочными.

2 курс ФБМФ и ФМХФ:

Результат курса можно зачесть за курс математической статистики.

Подробности на странице курса.

Зачем это нужно?

Физика:

- Анализ физических данных
- Оценки параметров и построение моделей физических систем
- Проверка гипотез и многое другое

Биология:

- Проверка гипотез в экспериментах (почти в каждой статье)
- Анализ омиксных данных
- Классификация клеток, оценка уровня экспрессии генов, кластеризация микробиома, определение вторичной структуры ДНК и многое другое

Химия:

- Нахождение химического соединения с заданными свойствами
- Оценка параметров и построение моделей реакций
- Проверка гипотез и многое другое

Работа с погрешностями тоже основана на статистике.

Основные темы (весна)

1. Простые методы анализа данных и машинного обучения;
2. Точечное оценивание;

   Свойства и методы поиска оценок; качество оценок и поиск наилучших оценок; оценки специального вида.
3. Доверительные интервалы;
4. Непараметрический подход;

   Бутстреп; ядерная оценка плотности.
5. Проверка гипотез;

   Критерии; p-value; множественная проверка гипотез; критерии согласия.
6. Линейная регрессия;

   МНК; отбор признаков.
7. Корреляционный анализ;
8. Дисперсионный анализ.
9. Байесовский подход, метод Монте-Карло;

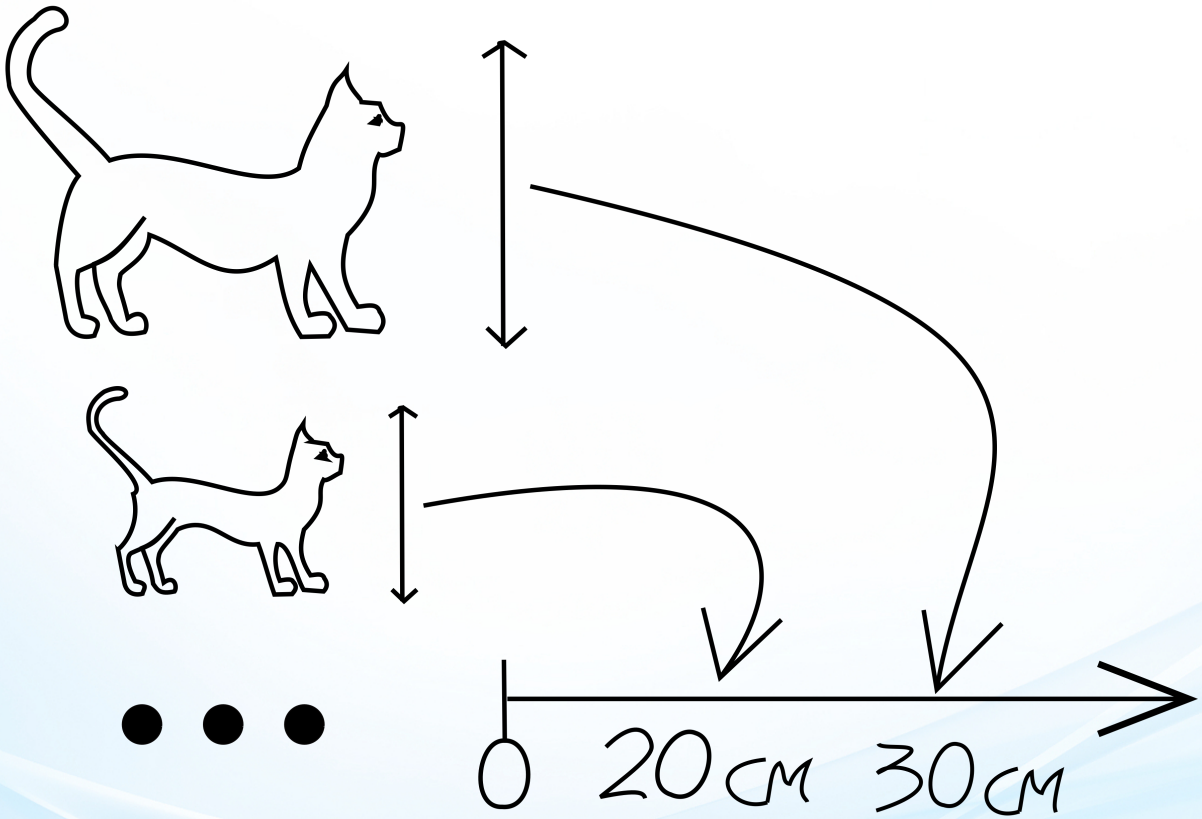
Основные книги по курсу

- Лагутин М.Б., Наглядная математическая статистика;
- L. Wasserman, All of Statistics;
- Russell B. Millar, Maximum Likelihood Estimation and Inference;
- Боровков А.А., Математическая статистика.

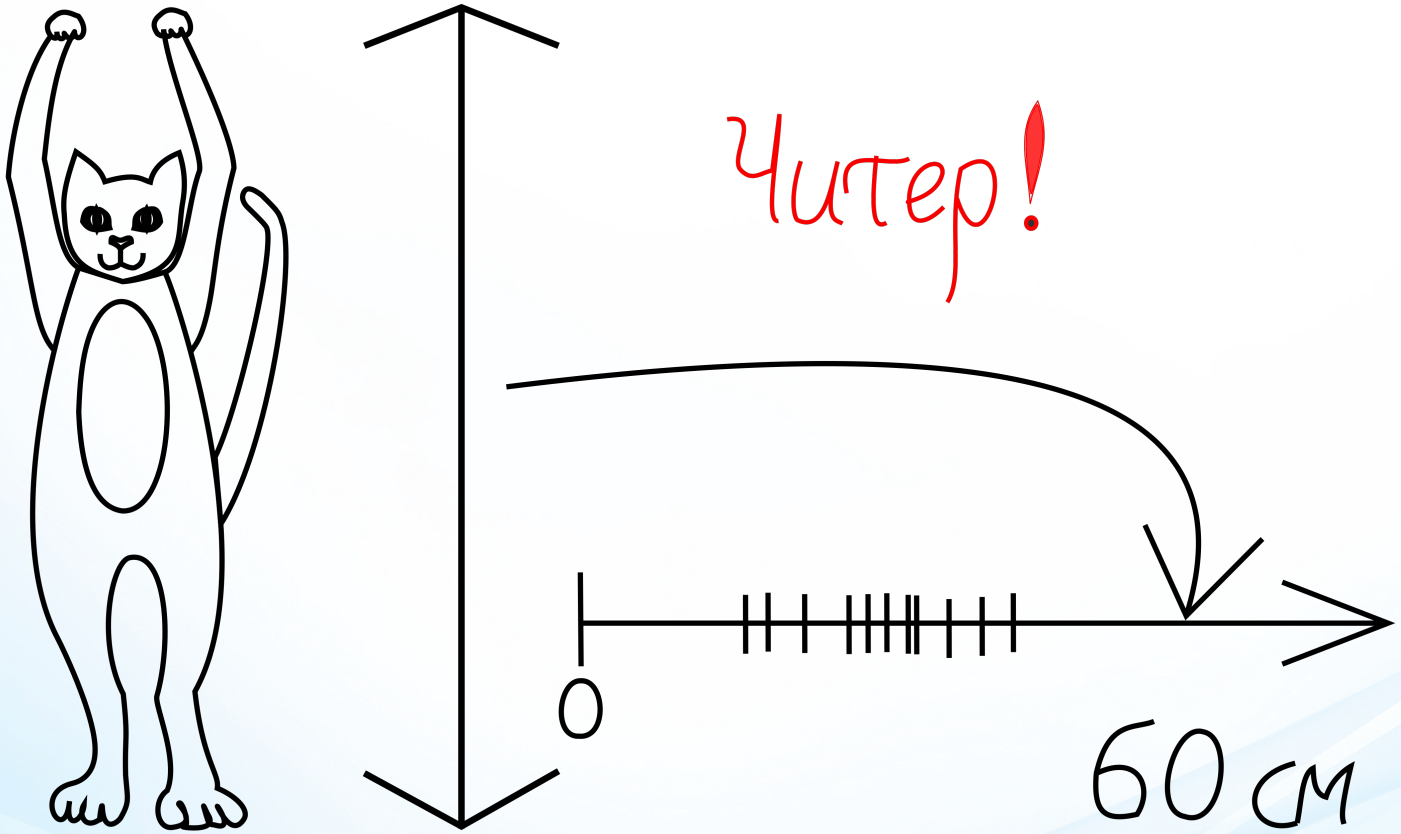
Обзор статистики

Мурмурландия

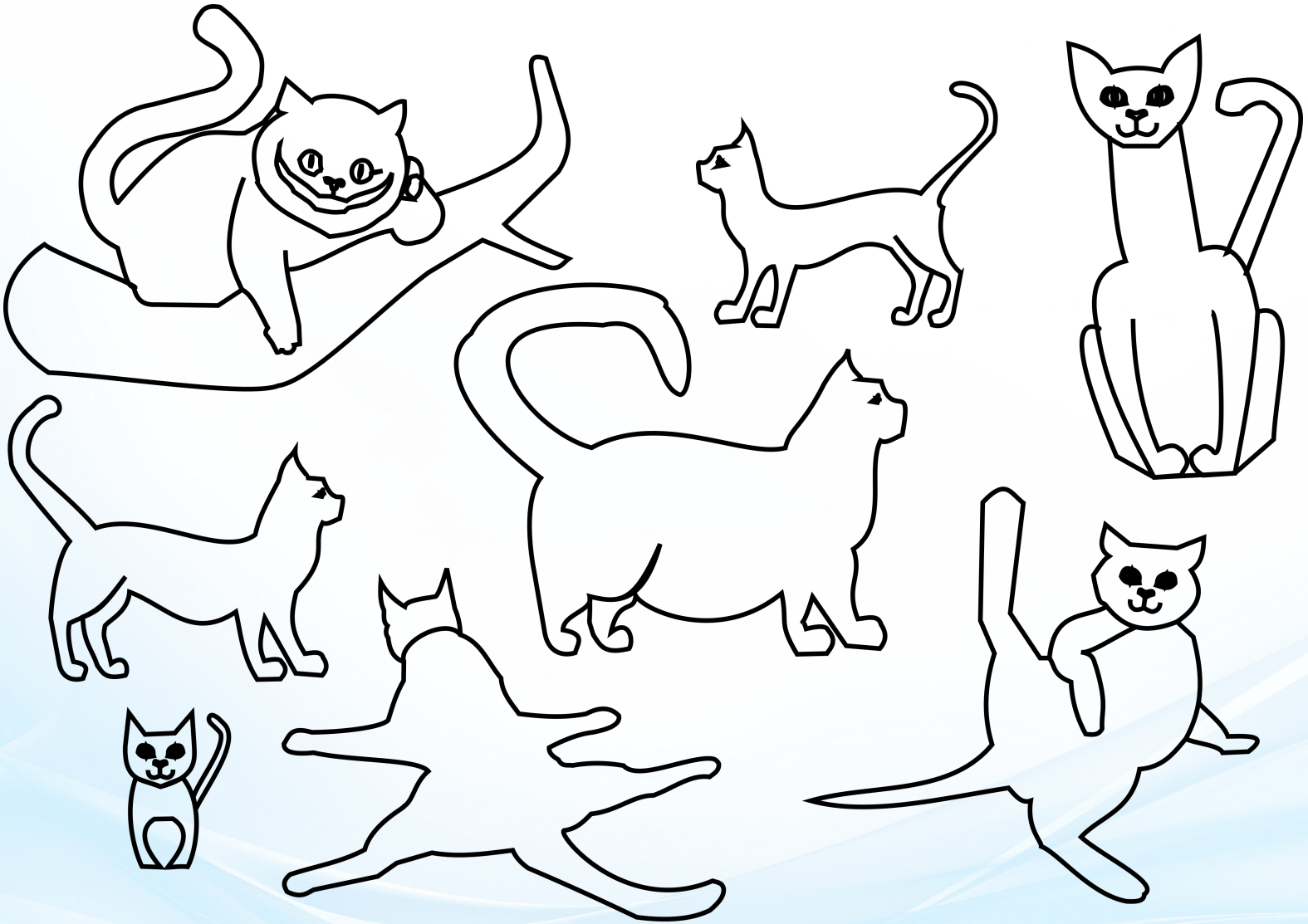
Каков средний рост котиков?



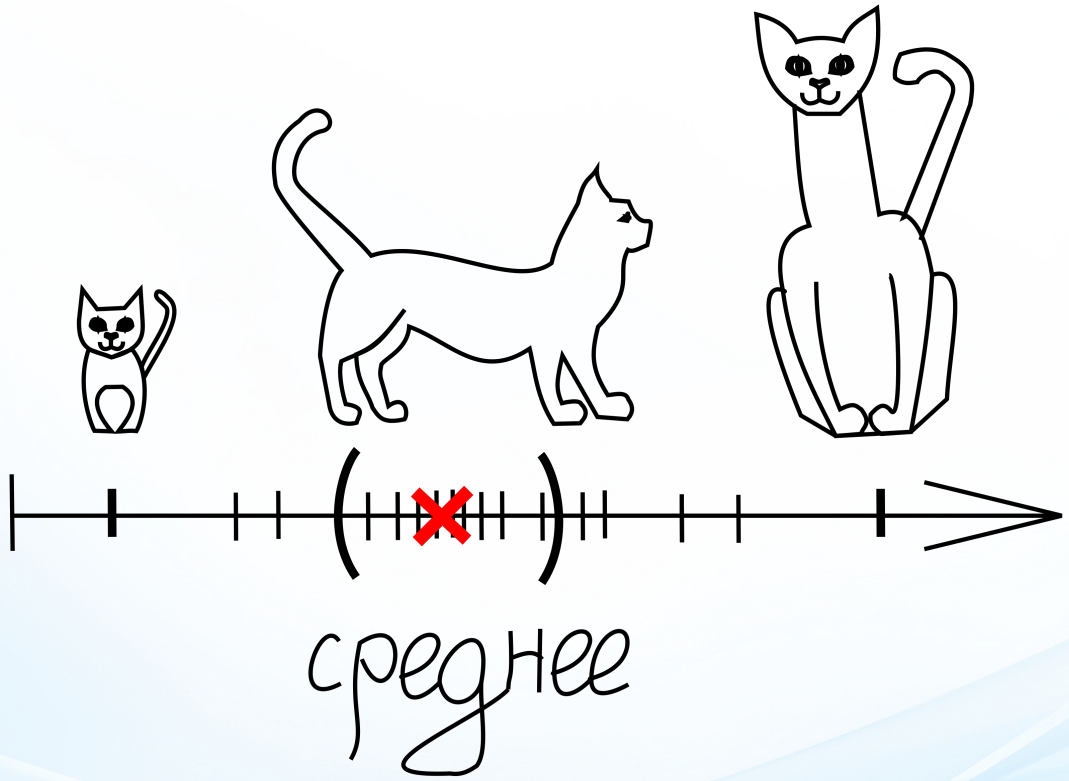
Точечное оценивание



Выбросы

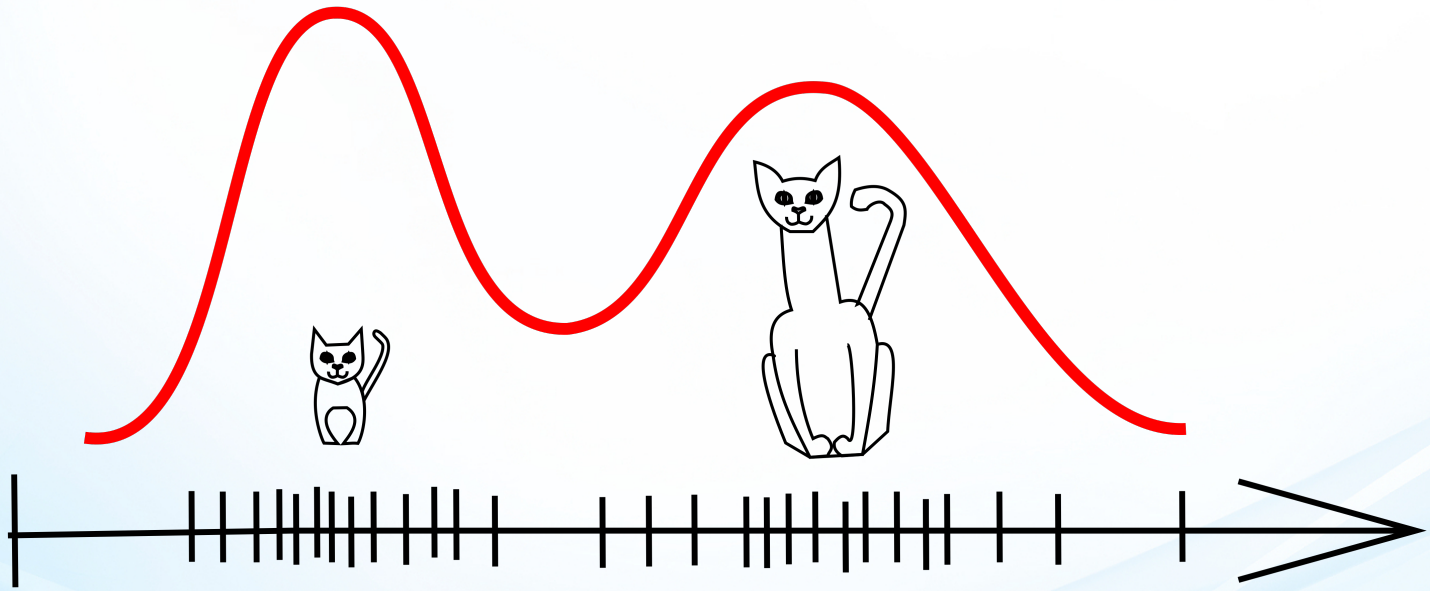


Среднее определяется неточно



Интервальное оценивание

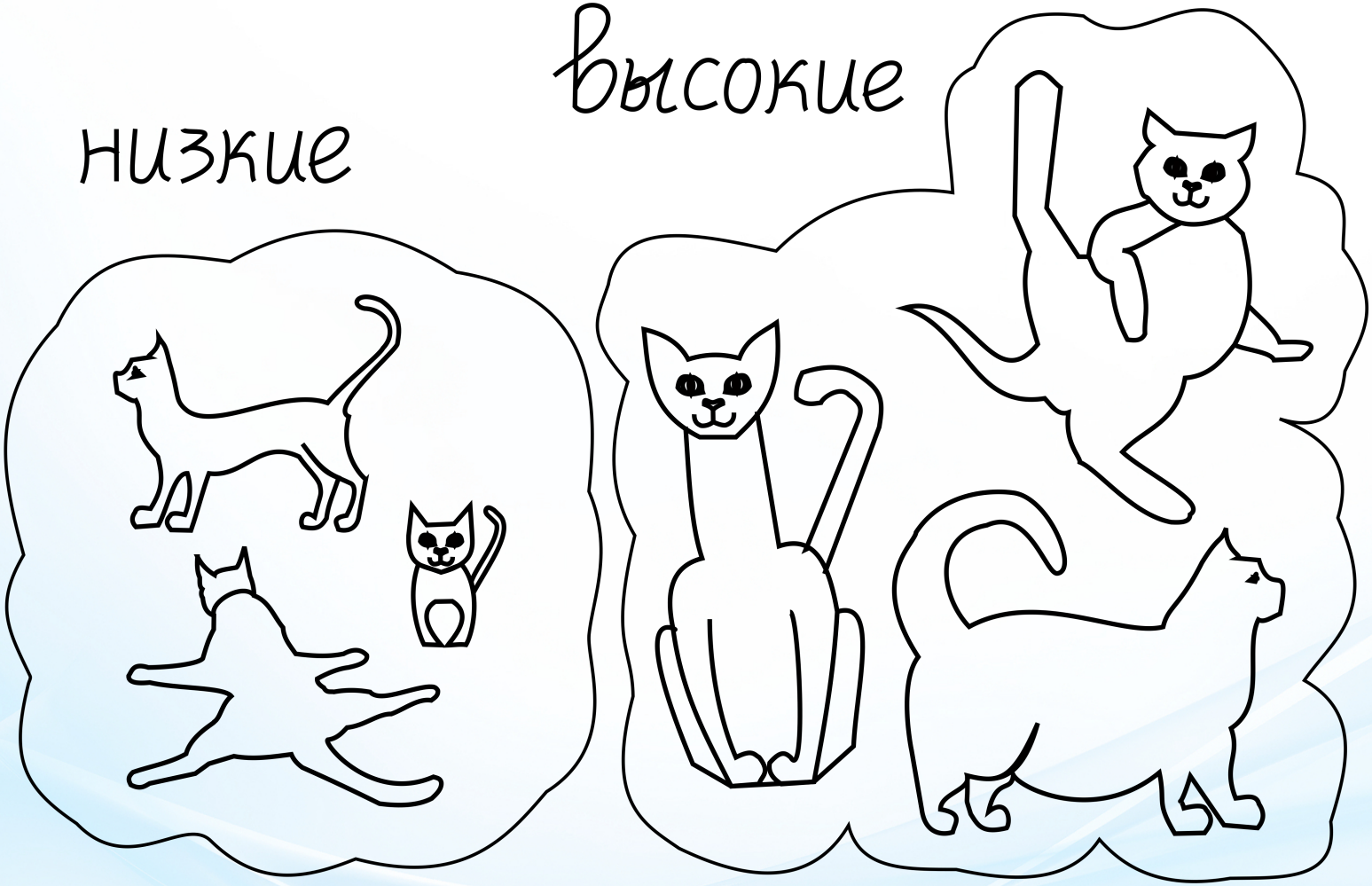
Характер распределения



Непараметрическое оценивание

низкие

высокие

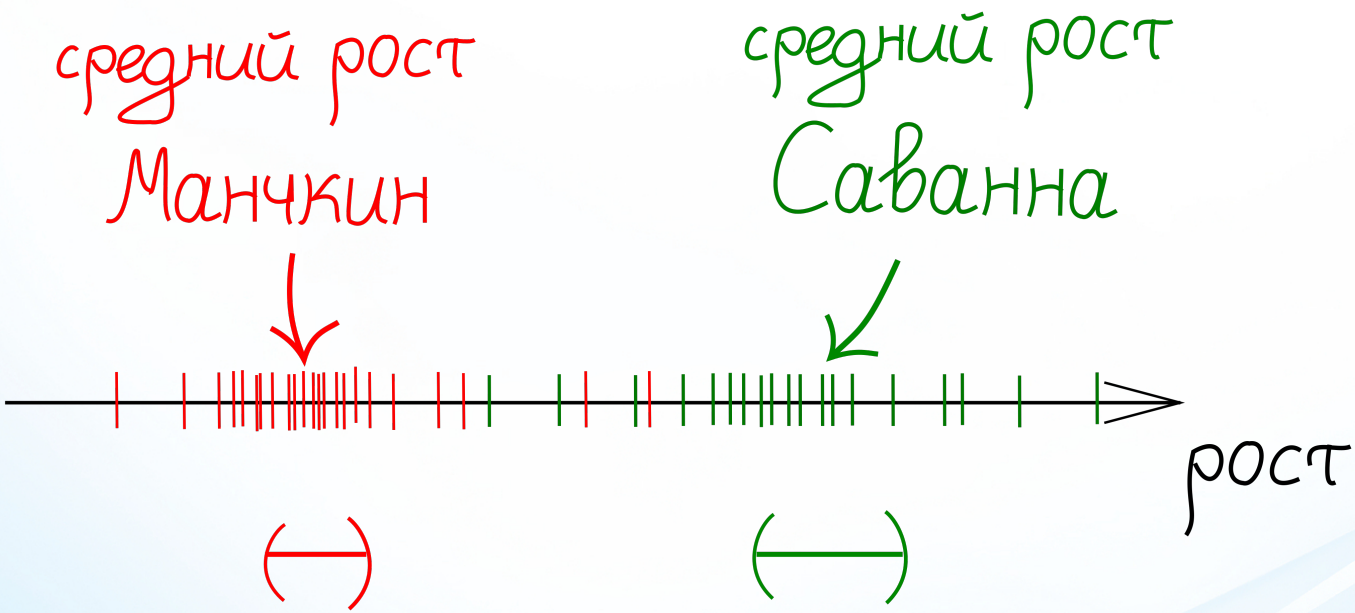


Саванна

Манчкин

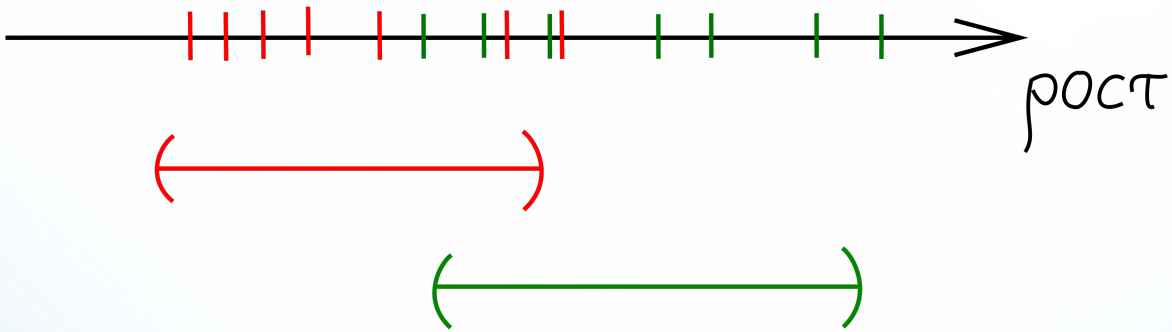
Отличается ли их средний рост?

Собираем данные

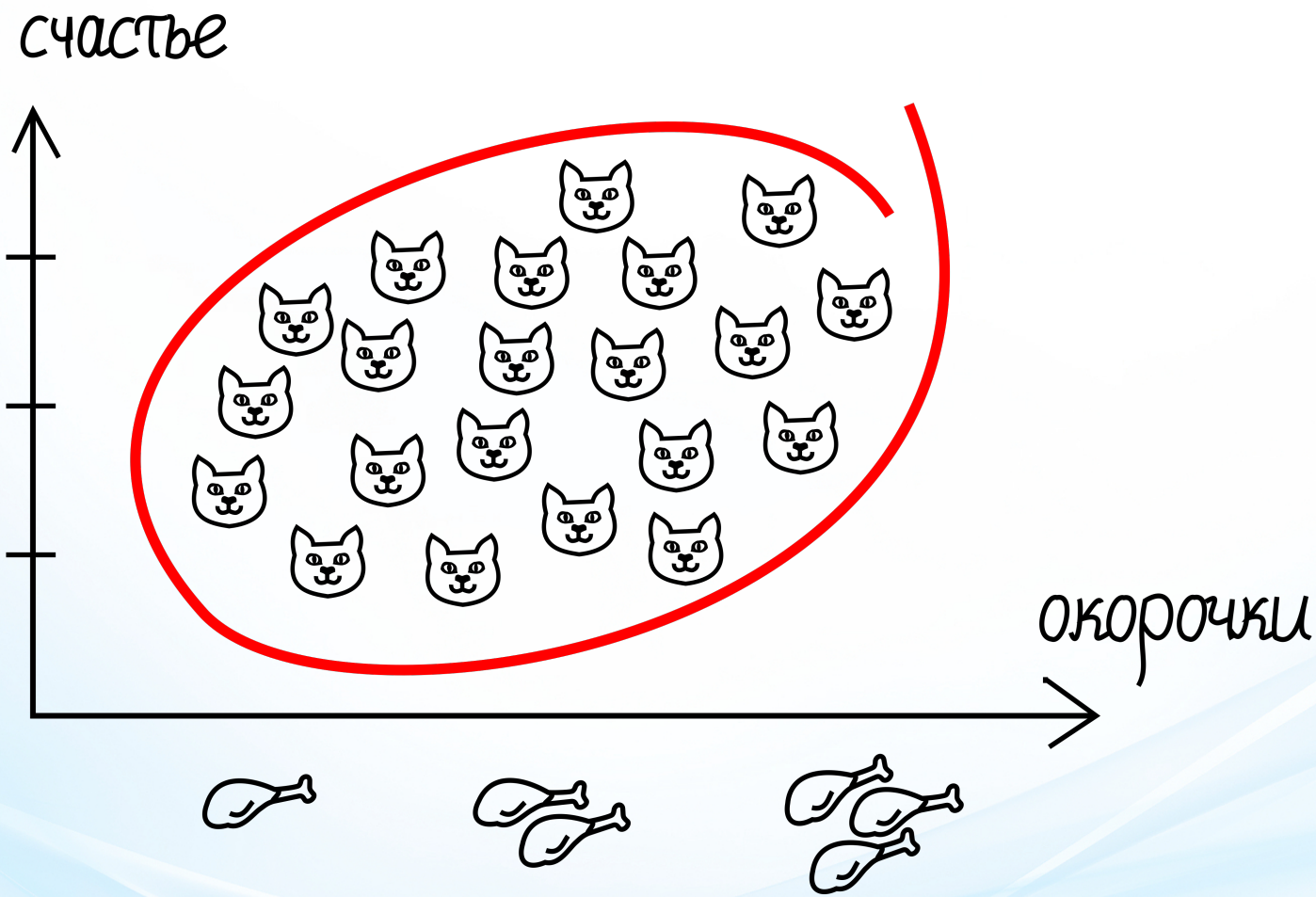


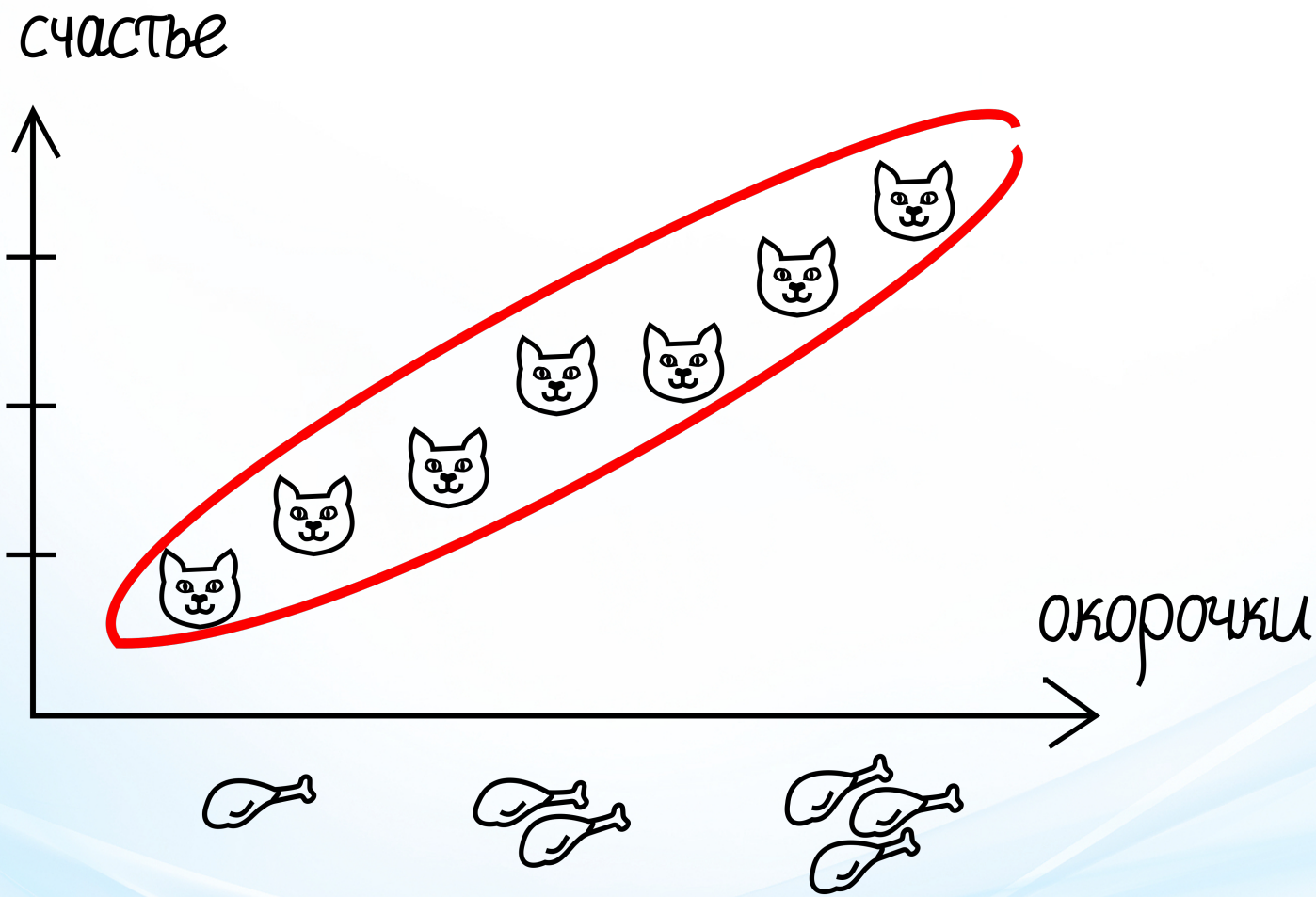
отличается

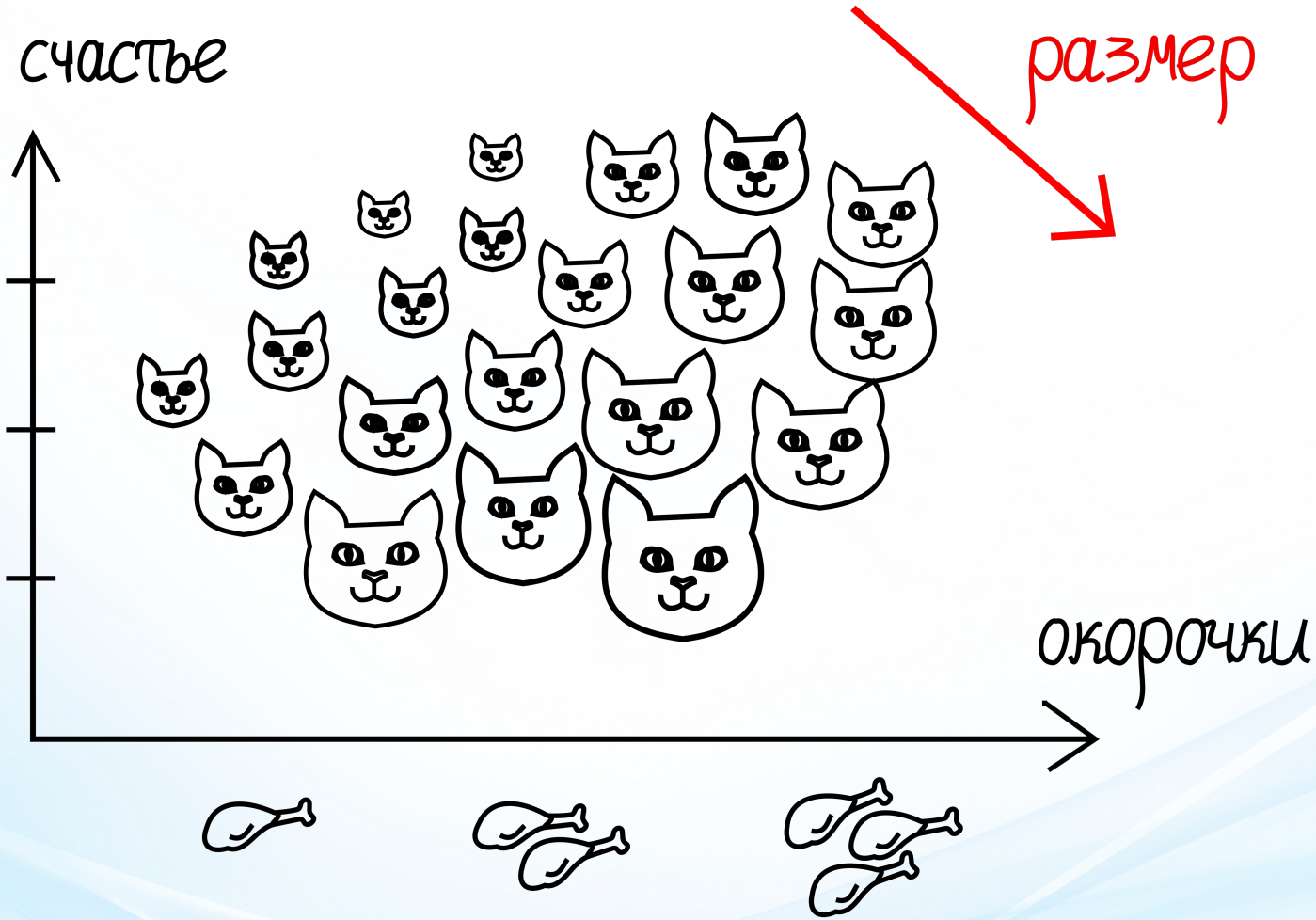
Если данных мало



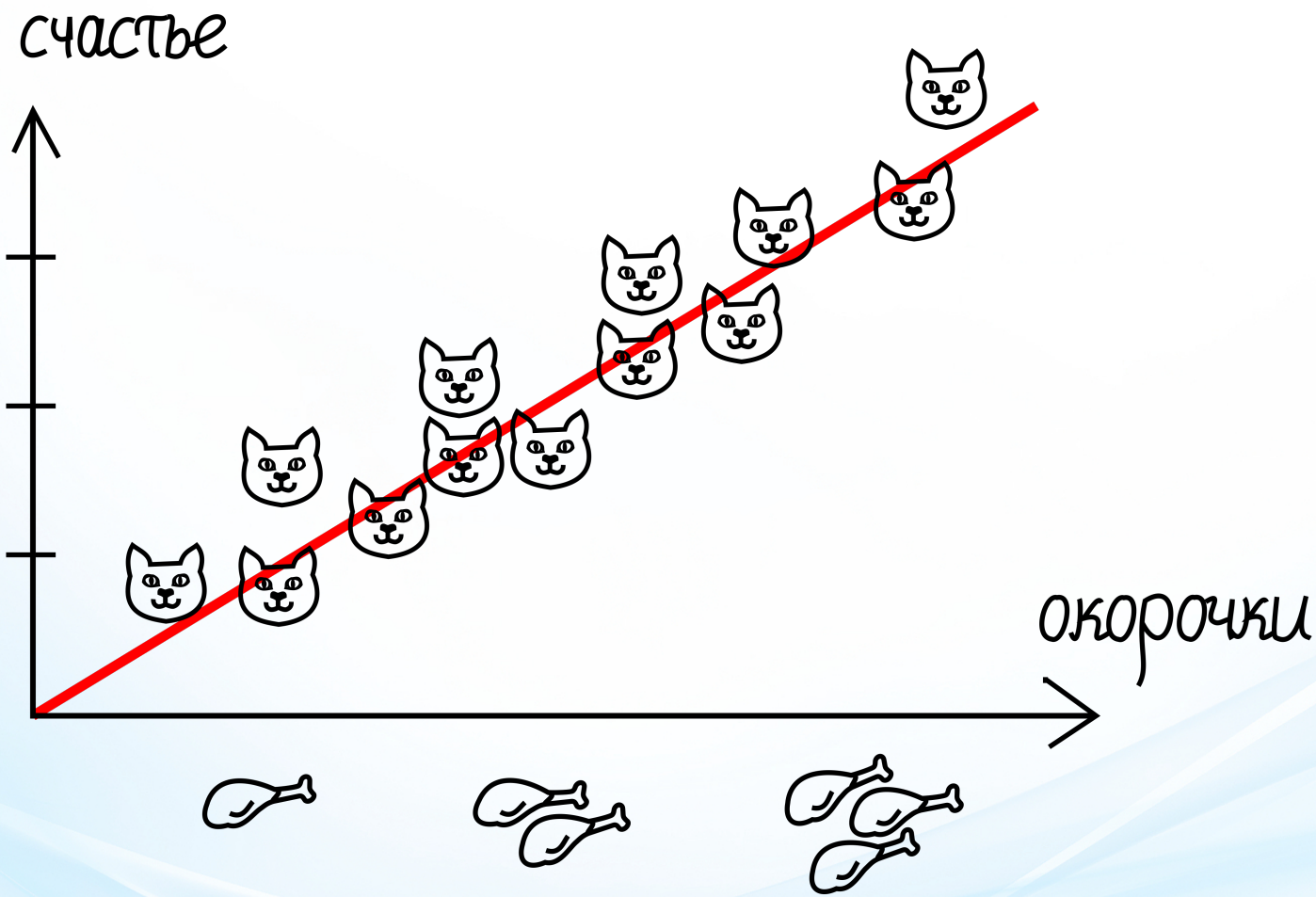
непонятно


счастье
окорочки

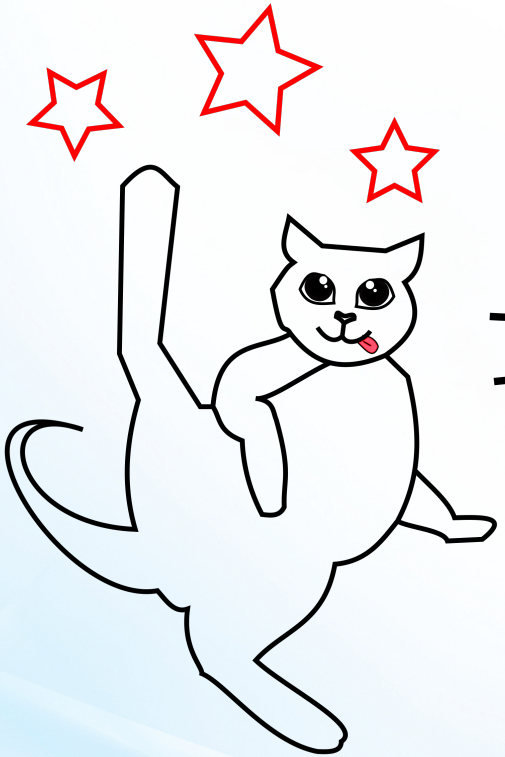

счастье
окорочки



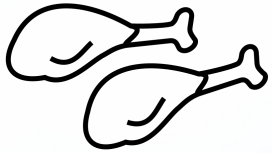
Корреляционный анализ


счастье
окорочки

Формула счастья

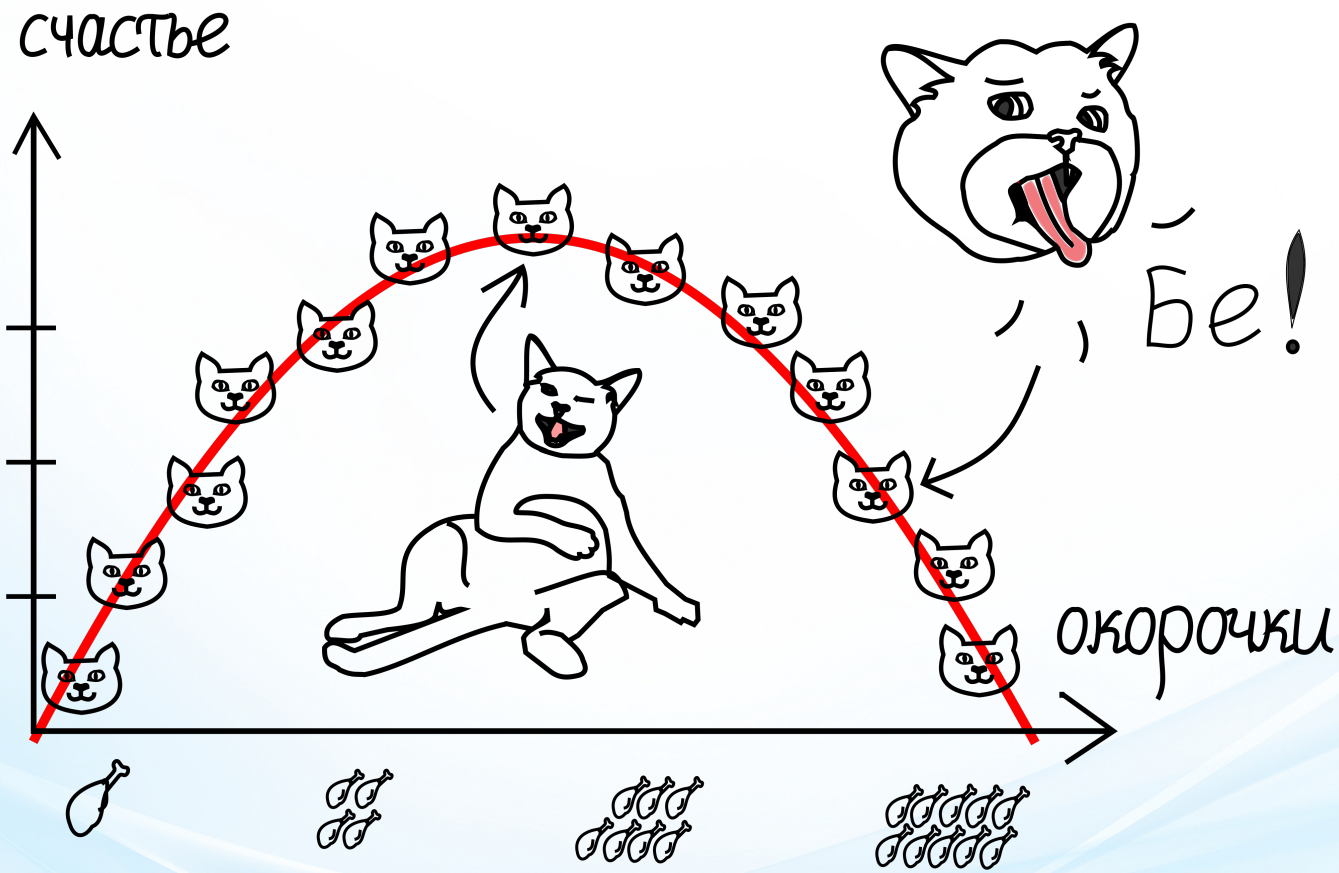


$$= \theta_0 + \theta_1 \times \text{кол-во}$$

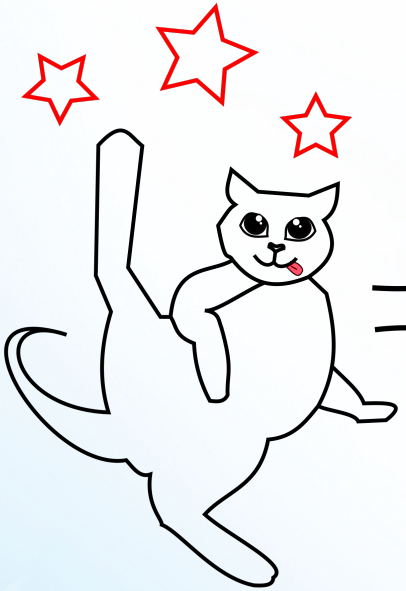


$$+ \text{погрешности}$$

Больше окорочков



Формула счастья

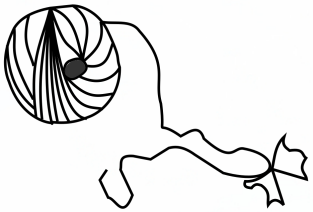


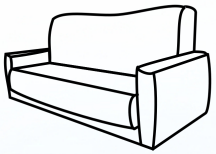
$$= \theta_0 + \theta_1 \times (\text{кол-во}) - \theta_2 \times (\text{кол-во})^2$$

$$+ \text{погрешности}$$

Другие факторы

$$= \theta_0 + \theta_1 \times \text{  } - \theta_2 \times \left(\text{  }\right)^2$$

$$+ \theta_3 \times \text{  }$$

$$+ \theta_4 \times \text{  }$$

$$+ \text{погрешности}$$

Регрессионный анализ

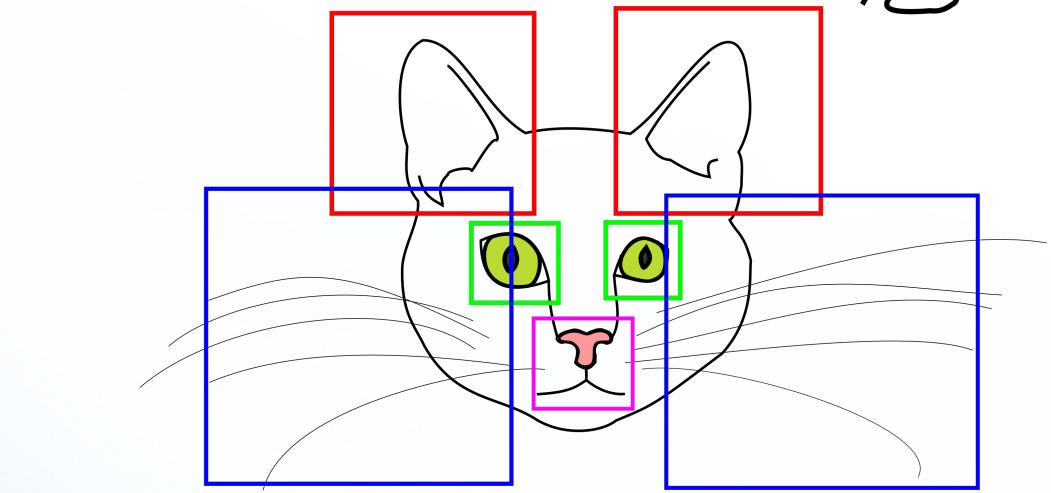
Классификация котиков

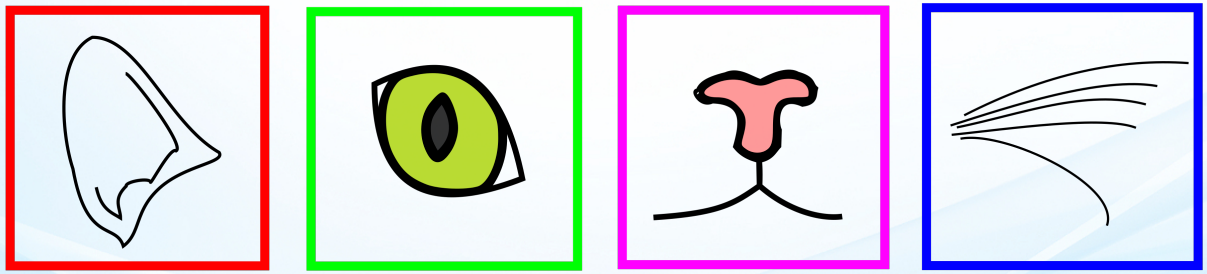
Классификация

Собираем данные

котик	порода	рост	шерсть
	Саванна	50 см	да
	Сфинкс	30 см	нет
	Манчкин	15 см	да
	Саванна	40 см	да

Распознавание мордочек





Нейронные сети

Художник курса

Примеры реальных задач

Оценки в экспериментах

Часто в экспериментах измеряется какая-либо величина.

Пример: концентрация вещества, свечение флуоресцентного маркера, энергия частицы.

Есть данные эксперимента — измерения каких-то характеристик.

Нужно сравнить полученные распределения с предсказаниями теории или найти зависимость.

- Оценить параметры — точечные оценки
- Оценить погрешности — доверительные интервалы
- Проверить, насколько хорошо данные согласуются с теорией
- Восстановить неизвестную зависимость от параметров (простейший пример — МНК)

  Все эти процедуры основываются на статистике.

Проверка гипотез

Почти в каждом эксперименте надо проверять различные предположения.

Например, является трек треком нейтрино или другой частицы.

Часто гипотез не одна, а несколько.

Необходимо поддерживать небольшой "суммарный" уровень "ошибки" для того, чтобы можно было доверять выводам ученых.

Метод решения: множественная проверка гипотез

Пример: изучить роль интерлейкина-10 в подавлении воспаления.

В ходе работы неоднократно проверяются гипотезы различного рода о влиянии IL-10 на различные метаболические процессы.

Eddie et al. Anti-inflammatory effect of IL-10 mediated by metabolic reprogramming of macrophages,

https://science.sciencemag.org/content/356/6337/513

Лекарства против респираторного заболевания

- 2 лекарства;
- несколько испытуемых.

Каждый испытуемый

- вдыхает первое лекарство с помощью ингалятора;
- проходит упражнение беговой дорожке. Измеряется время достижения максимальной нагрузки;
- после периода восстановления эксперимент повторяется со вторым лекарством.

Отличается ли время достижения максимальной нагрузки для исследуемых лекарств?

Метод решения: дисперсионный анализ.

Kanji G.K. 100 statistical tests, 2006.

Эффективность жаропонижающих средств

- 4 лекарства, в составе которых один и тот же активный ингридиент присутствует в разных дозировках;
- 4 группы из 15 морских свинок;
- В каждой группе фиксируется изменение температуры после введения жаропонижающего.

Есть ли различия в действии препаратов?

Метод решения: дисперсионный анализ.

Bonnini S., Corain L., Marozzi M., Salmaso S. Nonparametric Hypothesis Testing - Rank and Permutation Methods with Applications in R, 2014.

Диагностика диабета

Около 500 пациентов, для которых известна информация:

- Пол, возраст;
- Индекс массы тела;
- Среднее кровяное давление;
- Измерения сыворотки крови;

Нужно предсказать риск развития диабета в течении года.

Метод решения: регрессионный анализ

Медицина — диагностика рака кожи

Нейронную сеть обучили на 130 000 изображений,

в которых выделено 2000 различных заболеваний кожи.

Точность диагностики 91%, как и у коллектива из 21 дерматолога из Стэнфордской медицинской школы.

https://geektimes.ru/post/285154/

Предсказание генов

Задача: по последовательности генома найти места нахождения потенциальных генов, а также их смысловых частей.

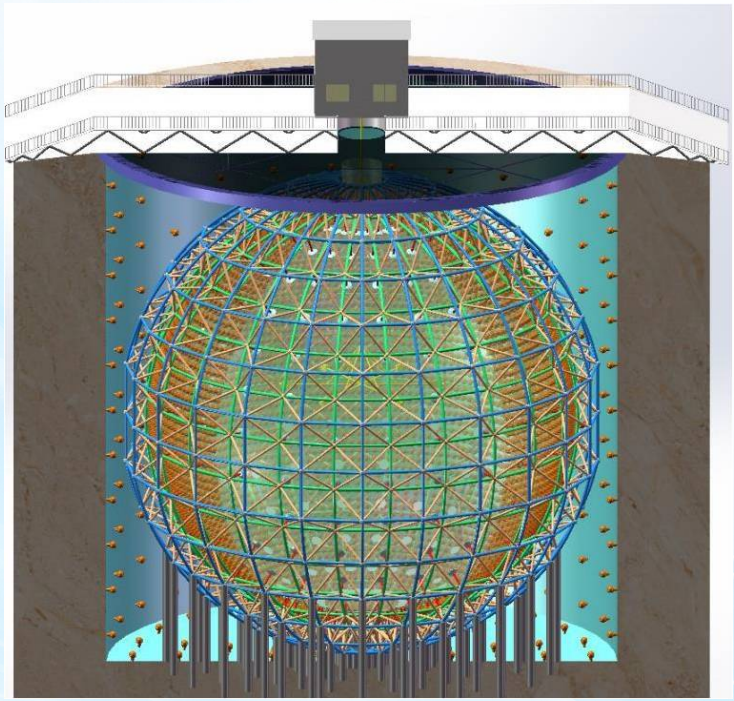
Решение: марковские модели, байесовский подход

Анализ данных в физике частиц

Пример: эксперимент JUNO.

Нужно по данным количества и времена детекции нейтрино на каждом фотодетекторе определить координаты и энергию частицы.

Метод решения: машинное обучение.



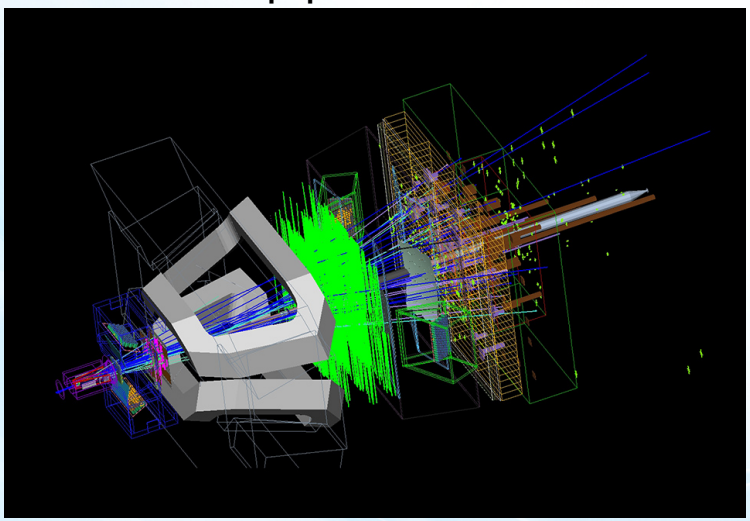
https://neutel11.wordpress.com/2015/03/04/j-cao-juno-a-multi-purpose-neutrino-observatory/

Большой адронный коллайдер

Данные, полученные на детекторе LHCb, проходят через несколько триггеров, которые оставляют только потенциально интересные события.

Классификацию событий можно делать методами машинного обучения.

Используя возможности машинного обучения (CatBoost), ученым удалось повысить эффективность системы триггеров на 40-50 процентов.

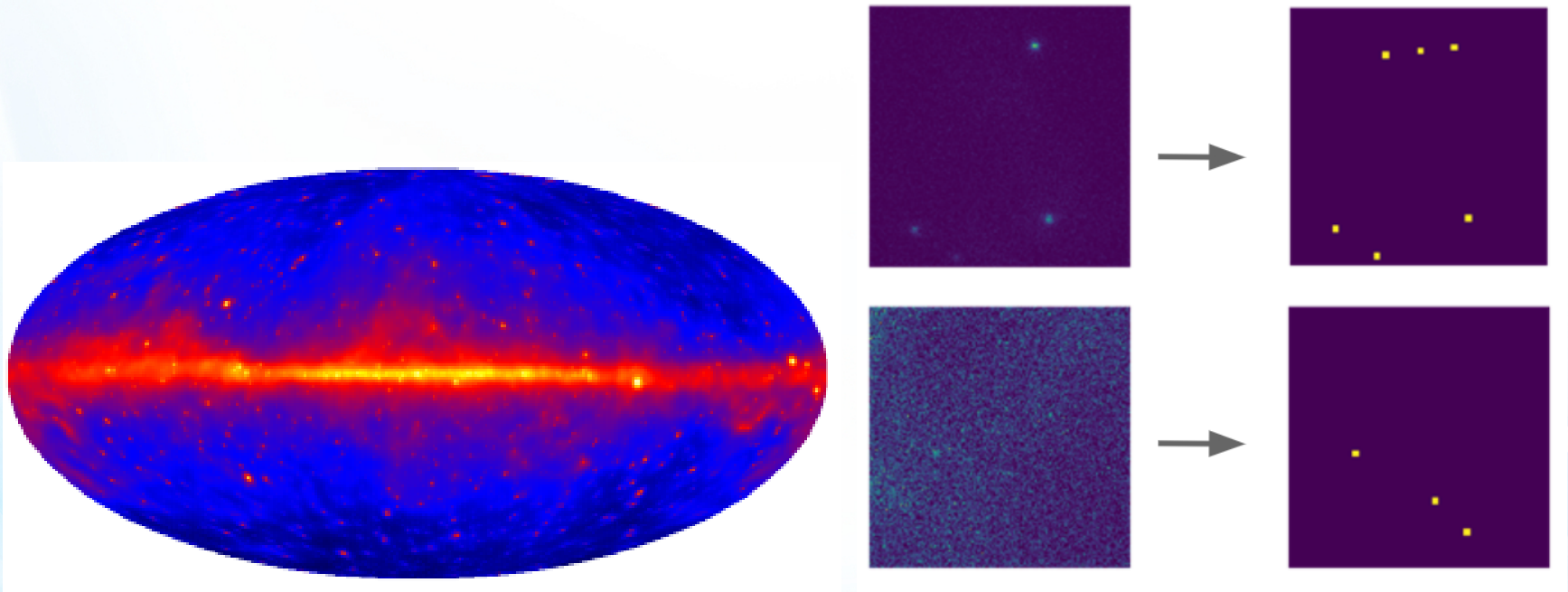


LHCb collaboration, CERN

https://nplus1.ru/material/2017/10/25/cern-yandex

Шумоподавление изображений телескопа Ферми

Найти точечные источники на изображениях гамма-телескопа Fermi.



Метод решения: байесовский подход

Создан каталог источников и открыты "гигантские пузыри Ферми".

The Denoised, Deconvolved, Decomposed Fermi γ-ray sky (Selig et al. 2015)

Альтернативный подход: нейронные сети

Аналитическая химия

Для того, чтобы новая методика количественного измерения была принята для использования, необходимо, чтобы погрешности, получаемые с помощью этой методики, были распределены нормально.

Метод решения: проверка гипотезы о нормальности выборки

Сравнение воспроизводимости результатов или методик:

Чем меньше погрешность, тем точнее методика.

Необходимо сравнить погрешности, получаемые разными методиками.

Метод решения: дисперсионный анализ

Обучение с подкреплением для стимулирования химических реакций

В ходе проведения химических реакций можно во времени менять различные параметры: *температура, добавление катализаторов, и т.д.*

Цель: получить искомое вещество

Варианты: получить в-во с наименьшими затратами;

получить вещество с заданными свойствами. *Например, лекарство.*

Для решения задачи может применяться обучение с подкреплением - раздел машинного обучения, с помощью которого можно "обучать" принятие решений.

Практическая часть

Языки для анализа данных

Слабо специализированный, проприетарный.

Недостаточно гибкий в современных задачах. Тем не менее, незаменим для профессионалов.

Достаточно гибкий, большое число библиотек.

Слишком молодой.

Jupyter Notebook — инструмент для создания документов-ноутбуков, которые сочетают в себе код и его выход, графики, текстовое описание, формулы.

Возможности:

- редактирование кода в браузере, с подсветкой синтаксиса, автоотступами и автодополнением;
- запуск кода в браузере;
- отображение результатов вычислений с медиа представлением (схемы, графики);
- работа с языком разметки Markdown и LaTeX.

Поддержка ядер: Python (по умолчанию), R, Julia, C++, Java и др.

Пример

localhost:8888 Пример

jupyter Пример (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

Пример ноутбука в виде ноутбука

Ячейка с кодом, в которой импортируются библиотеки:

In [1]:
```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

Текстовая ячейка в формате **Markdown**. Например, можно сделать *маркерованный список*

- один
- два
- три

Или написать `latex` -формулу:

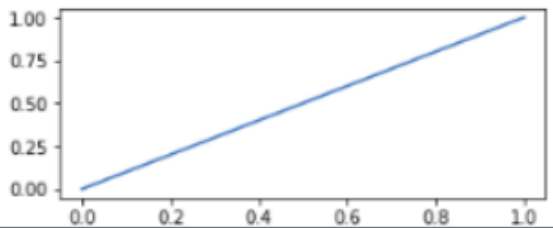
$$P(A \cap B) = P(A)P(B)$$

Далее снова код и его вывод:

In [2]:
```
print(np.arange(5))
```

```
[0 1 2 3 4]
```

In [3]:
```
plt.figure(figsize=(5, 2))
plt.plot([0, 1], [0, 1]);
```



Пример

localhost:8888 Пример

jupyter Пример (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

Пример ноутбука в виде ноутбука

Ячейка с кодом, в которой импортируются библиотеки:

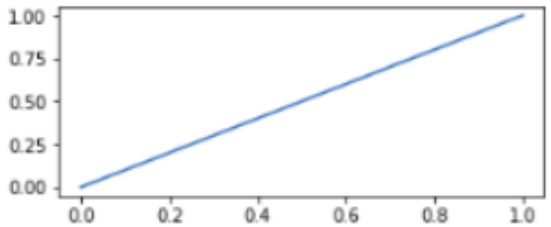
In [1]:
```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

```
Текстовая ячейка в формате **Markdown**. Например, можно сделать
*маркерованный список*

* один
* два
* три

Или написать `latex`-формулу:
$$P(A \cap B) = P(A) P(B)$$

Далее снова код и его вывод:
```

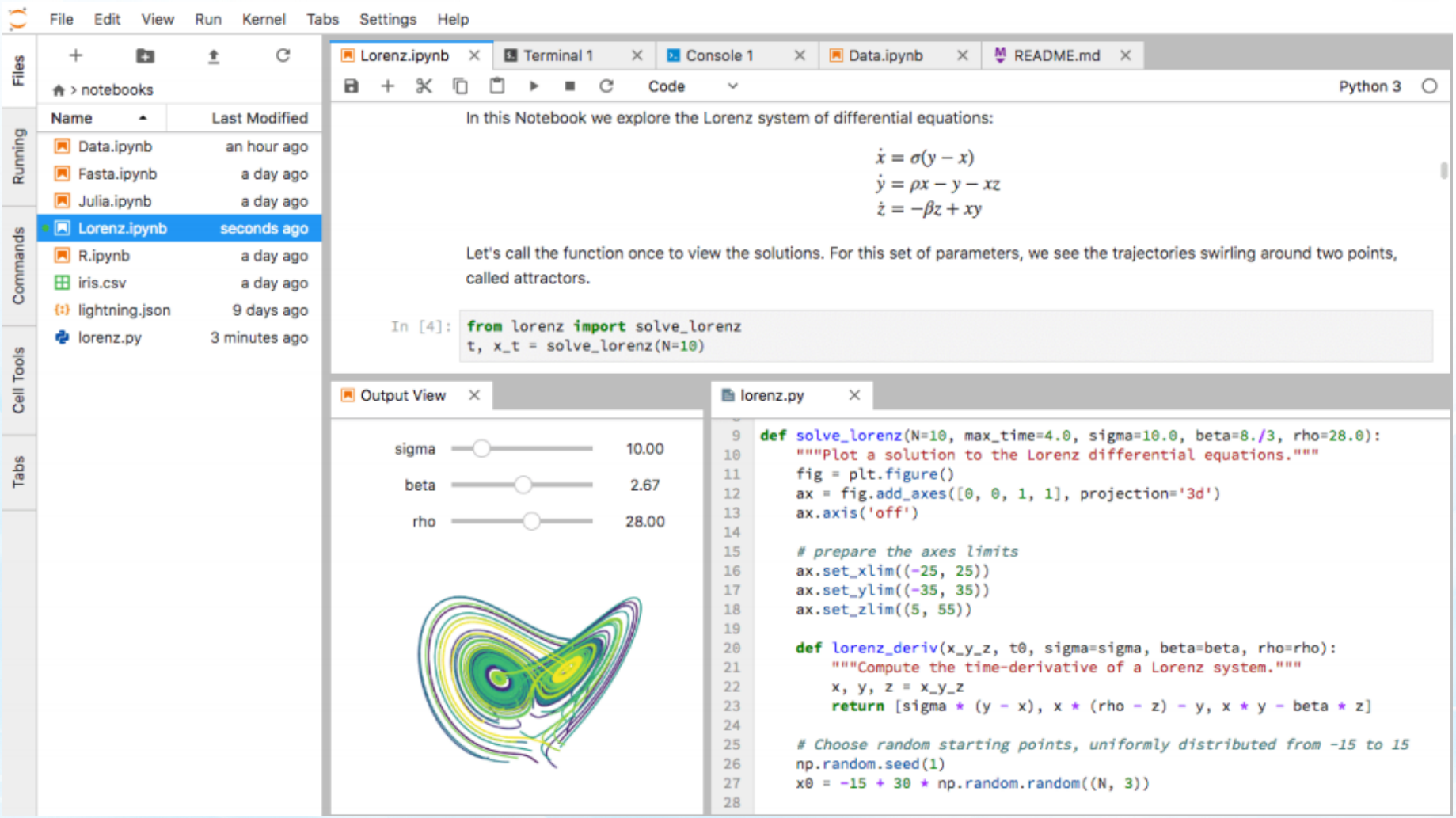
In [2]:
```
print(np.arange(5))
```

```
[0 1 2 3 4]
```

In [3]:
```
plt.figure(figsize=(5, 2))
plt.plot([0, 1], [0, 1]);
```



Jupyter Lab


File Edit View Run Kernel Tabs Settings Help
Files
Running
Commands
Cell Tools
Tabs
> notebooks
Name
Last Modified
Data.ipynb an hour ago
Fasta.ipynb a day ago
Julia.ipynb a day ago
Lorenz.ipynb seconds ago
R.ipynb a day ago
iris.csv a day ago
lightning.json 9 days ago
lorenz.py 3 minutes ago
Lorenz.ipynb
Terminal 1
Console 1
Data.ipynb
README.md
Code
Python 3
In this Notebook we explore the Lorenz system of differential equations:
$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$
Let's call the function once to view the solutions. For this set of parameters, we see the trajectories swirling around two points, called attractors.
In [4]:
from lorenz import solve_lorenz
t, x_t = solve_lorenz(N=10)
Output View
sigma 10.00
beta 2.67
rho 28.00
lorenz.py
def solve_lorenz(N=10, max_time=4.0, sigma=10.0, beta=8./3, rho=28.0):
"""Plot a solution to the Lorenz differential equations."""
fig = plt.figure()
ax = fig.add_axes([0, 0, 1, 1], projection='3d')
ax.axis('off')
# prepare the axes limits
ax.set_xlim((-25, 25))
ax.set_ylim((-35, 35))
ax.set_zlim((5, 55))
def lorenz_deriv(x_y_z, t0, sigma=sigma, beta=beta, rho=rho):
"""Compute the time-derivative of a Lorenz system."""
x, y, z = x_y_z
return [sigma * (y - x), x * (rho - z) - y, x * y - beta * z]
# Choose random starting points, uniformly distributed from -15 to 15
np.random.seed(1)
x0 = -15 + 30 * np.random.random((N, 3))

Установка Jupyter

- Ubuntu и т.п. (сначала установите пакет pip3, если его нет):

  sudo pip3 install jupyter

  Установка Jupyter Lab:

  sudo pip3 install jupyterlab

  jupyter serverextension enable --py jupyterlab --sys-prefix

  Запуск: jupyter notebook или jupyter lab

- В остальных случаях просто поставьте Anaconda:

  https://www.anaconda.com/download/

- Попробовать онлайн:

  https://jupyter.org/try

- Не получилось поставить => работа онлайн в Google Colab:

  https://colab.research.google.com/notebooks/intro.ipynb - введение в Python и Colab

Библиотеки для анализа данных

Гарантированно будем использовать:

- numpy — быстрые численные и матричные вычисления;
- scipy (stats) — статистический пакет;
- matplotlib — построение графиков;
- pandas — работа с табличными данными;
- seaborn — высокоуровневое построение графиков;
- sklearn — методы машинного обучения.

Пример установки на Ubuntu:

```
sudo pip3 install numpy
```

Также полезны:

- statsmodels — профессиональный стат. пакет;
- plotly — интерактивные графики;
- ipywidgets — виджеты.

Для освоения библиотек смотрите видеозаписи от нашей команды!

Введение

Основная задача математической статистики

Введение

Теория вероятностей

Зная природу случайного явления,

посчитать характеристику этого явления.

Математическая статистика

По результатам экспериментальных данных

высказать суждение о том, какова была природа этого явления.

Классический пример

На курсе N студентов; из них M выбирает спецкурс по анализу данных.

Задача в теории вероятностей

P(среди случайных n чел. ровно m слушателей спецкурса)$-$?

Предполагается, что M известно.

Задача в математической статистике

Среди случайных n чел. есть m слушателей спецкурса.

Оценить M.

Предполагается, что M не известно.

Еще пример

$\xi \sim \mathcal{N}(a, \sigma^2)$ — случайная величина

Задача в теории вероятностей

Известно, что $a = 2.3, \sigma = 7.1$

$$\mathrm{P}(\xi \in [0, 1]) - ?$$

$$\mathrm{E}\xi - ?$$

Задача в математической статистике

$x_1, ..., x_n$ — независимые реализации случайной величины ξ.

Оценить a и σ.

Вспоминаем оценки и погрешности в лабах!

Задача математической статистики

Пусть $x_1, ..., x_n$ — численные характеристики n-кратного повторения некоторого явления.

Будем их воспринимать как независимые реализации $\xi \sim \mathrm{P}$.

Задача: по значениям $x_1, ..., x_n$ высказать некоторое суждение о распределении P.

Решение: *статистический вывод* или *обучение*.

Основные понятия

Последовательность независимых одинаково распределенных случайных величин $X_1, \dots X_n$ называется выборкой.

Их значения $x_1, \dots, x_n$ как числа (на конкретном исходе) называются реализацией выборки.

Интуитивно: x_i - различные "измерения" какой-то величины.

Это имеющиеся у нас данные.

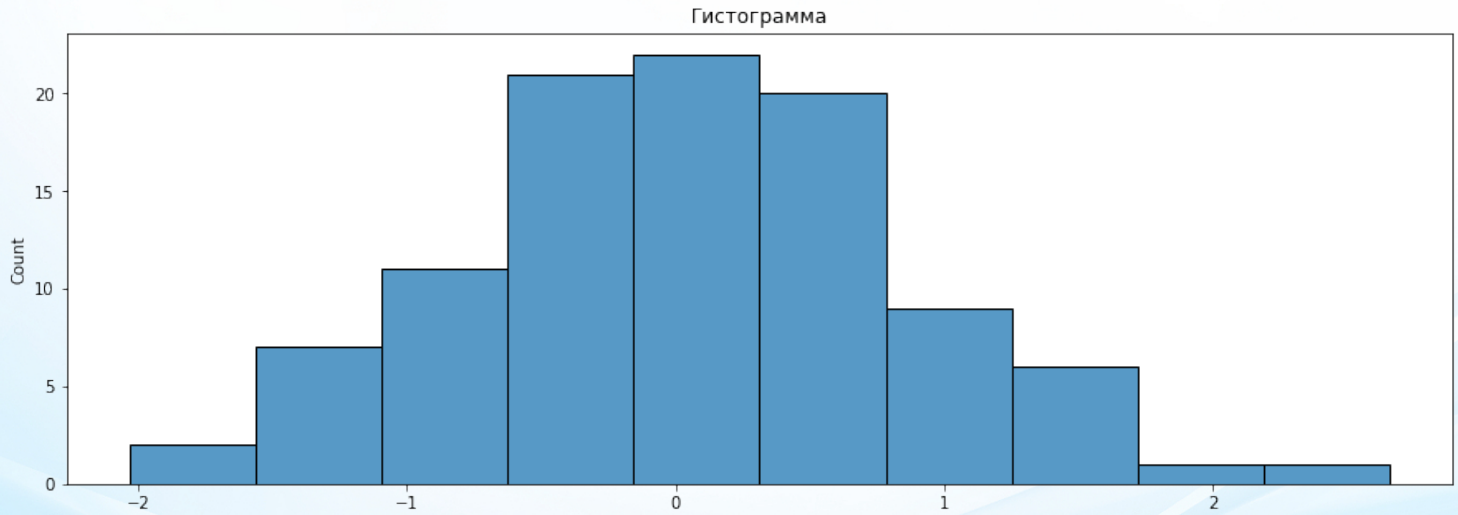
Давайте посмотрим, что вообще можно делать с данными!

Гистограмма

Пусть у нас есть реализация выборки $x_1, \dots x_n \in \mathbb{R}$.

Идея: разделим всю числовую прямую на несколько "корзин"

и посмотрим, сколько объектов (иксов) попало в каждую.

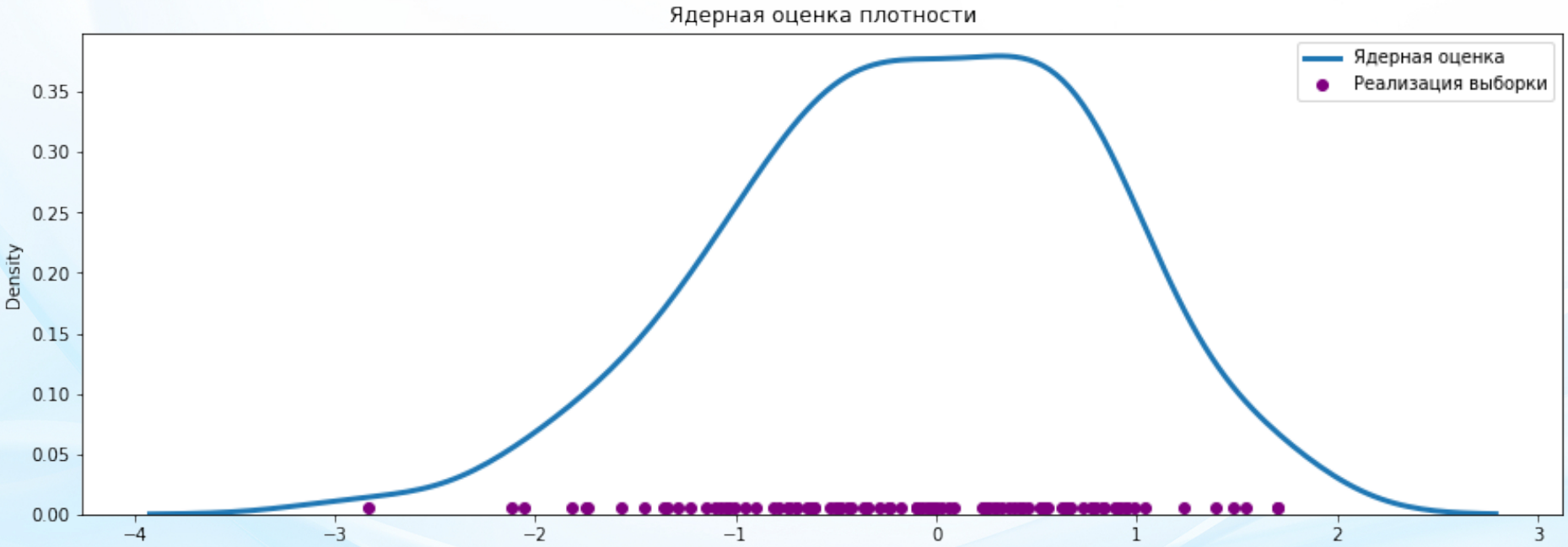


Можно построить график в виде столбиков, где высота столбика показывает, сколько объектов попало в соответствующую корзину.

Этот график по форме похож на график плотности распределения.

Ядерная оценка плотности

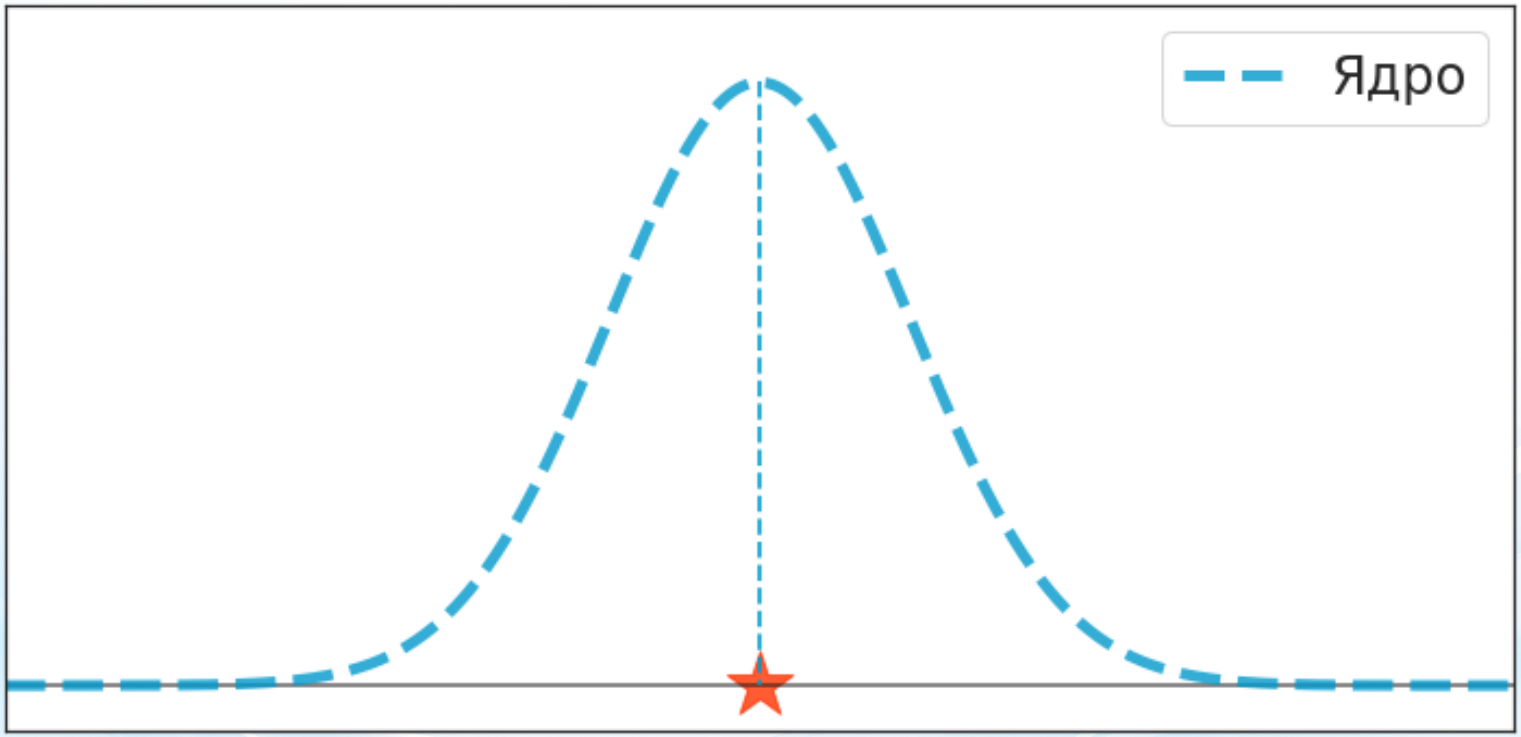
Идея: как-то оценить плотность распределения.



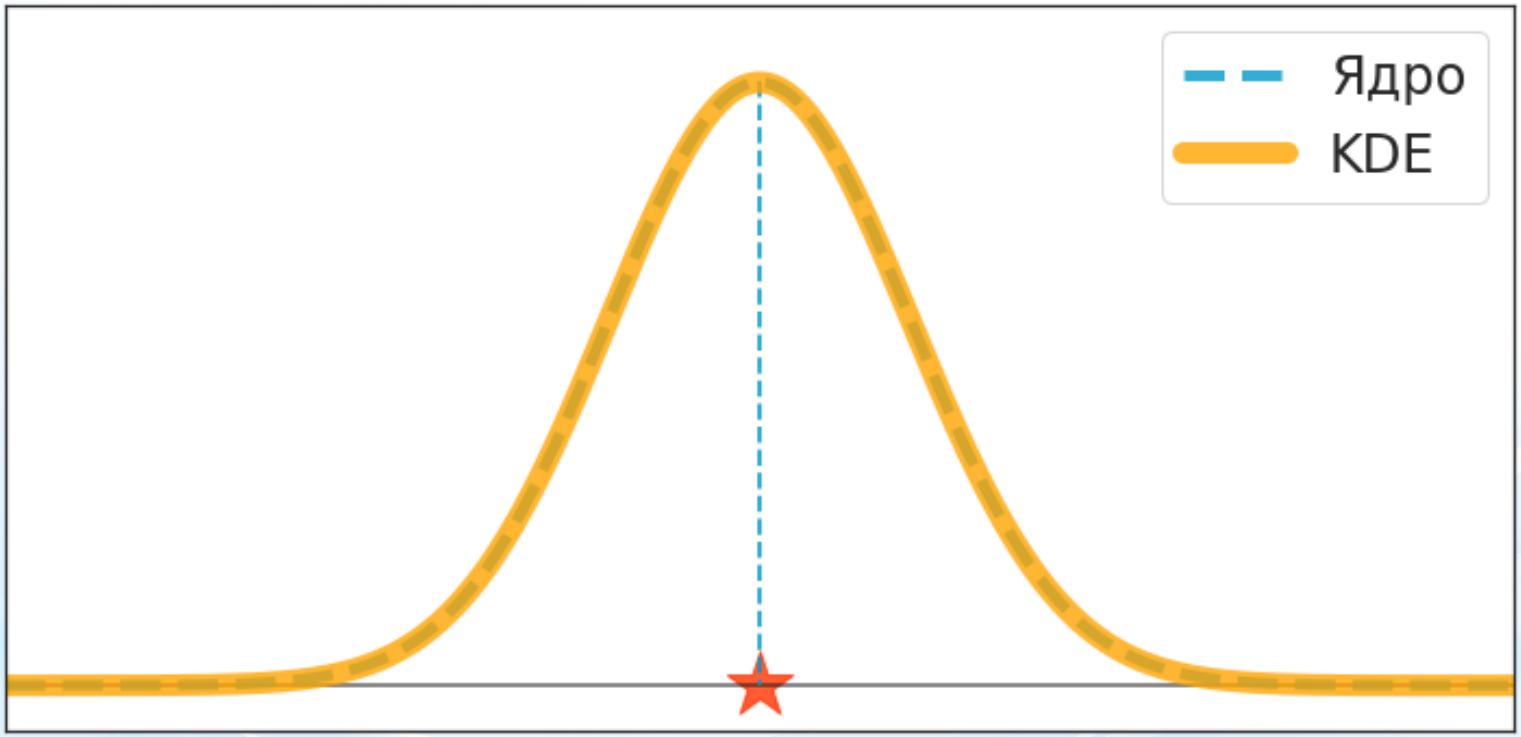
Как? Сейчас узнаем!

Ядерная оценка плотности: простые примеры

Ядерная оценка плотности: простые примеры

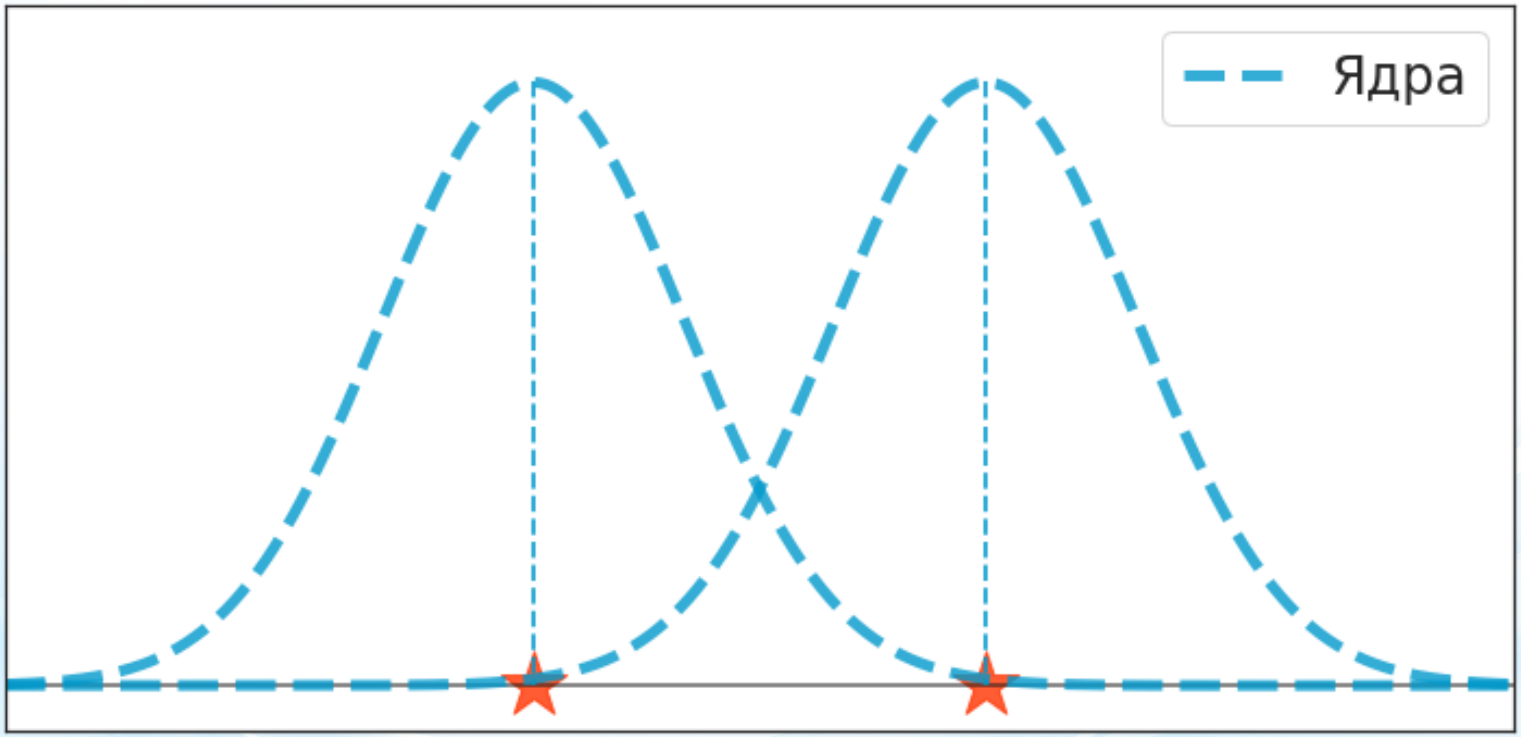


Ядерная оценка плотности: простые примеры

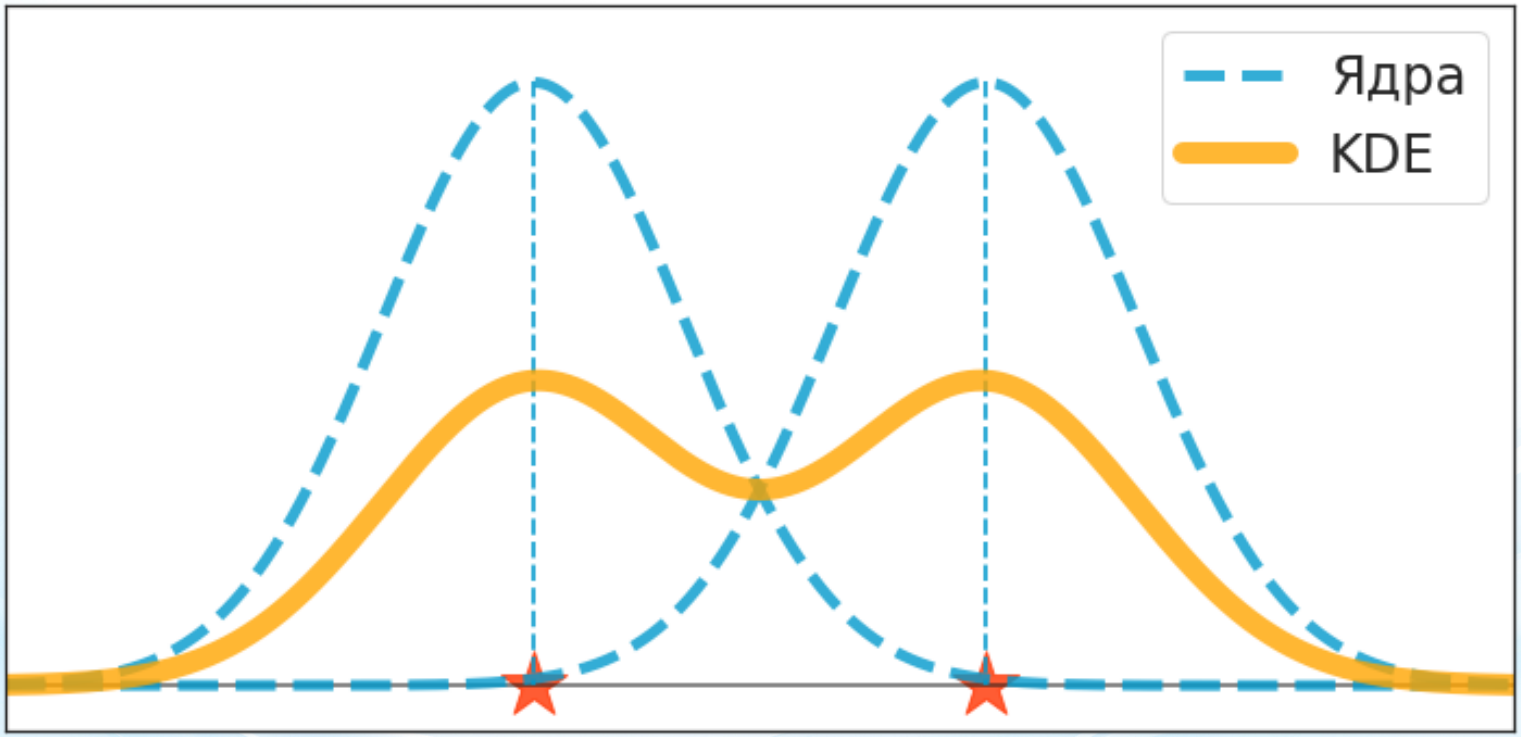


Ядерная оценка плотности: простые примеры

Ядерная оценка плотности: простые примеры

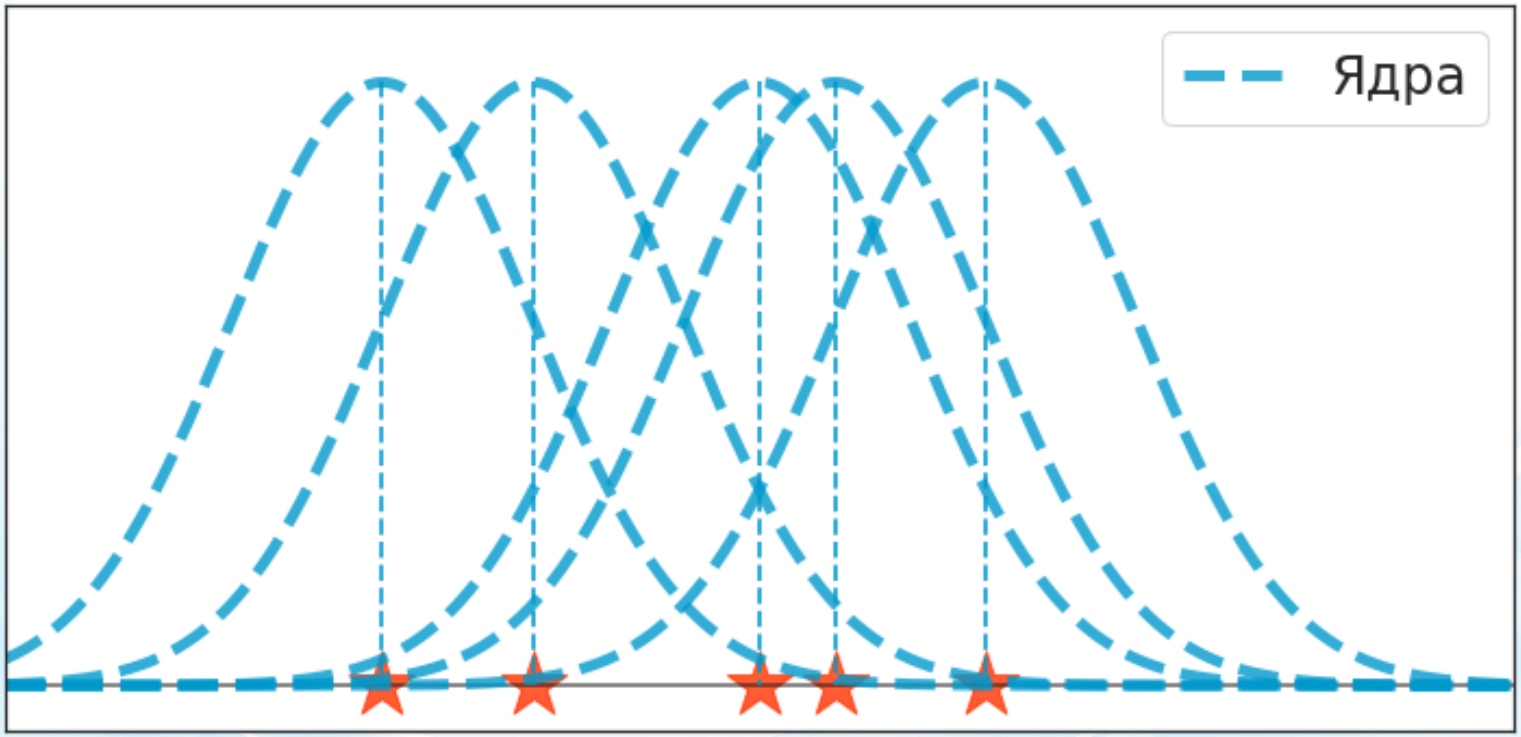


Ядерная оценка плотности: простые примеры

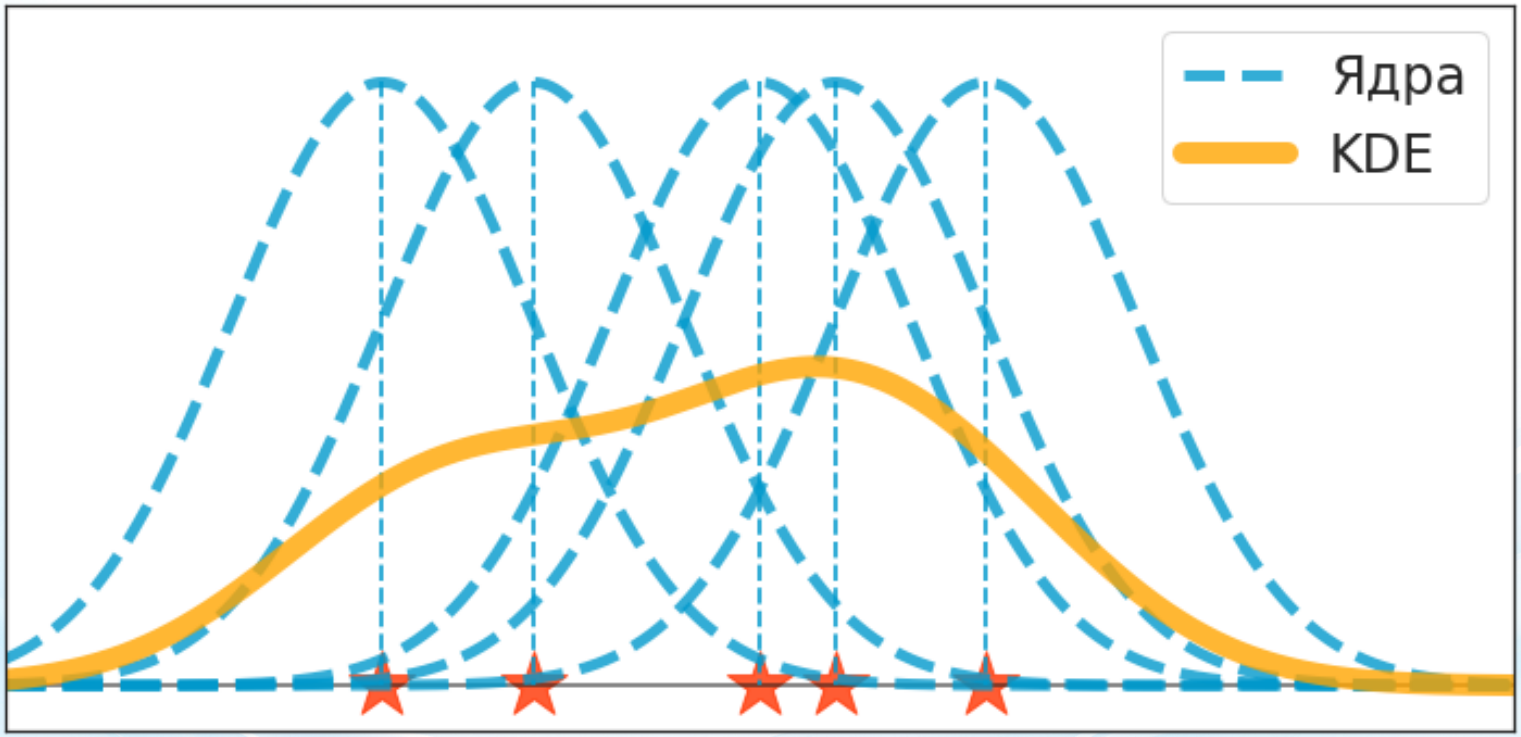


Ядерная оценка плотности: простые примеры

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

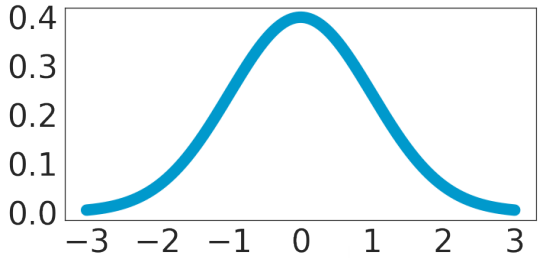
$$\widehat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} q\left(\frac{x - X_i}{h}\right)$$

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

Виды ядер

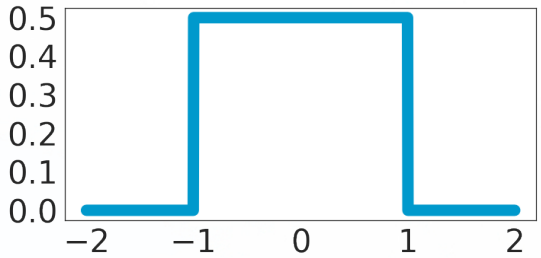
Гауссовское

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



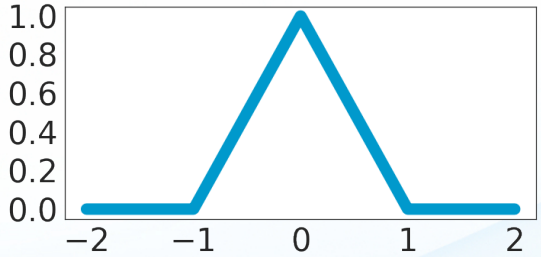
Прямоугольное

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



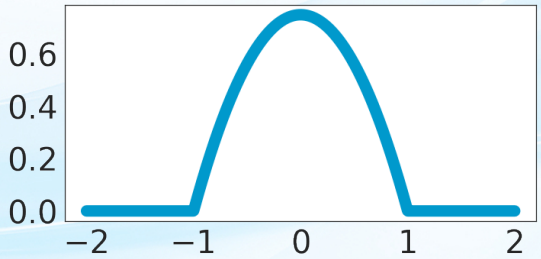
Треугольное

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



Епанечникова

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



Давайте посмотрим на все это на практике!

Статистики и оценки

Параметрический подход к статистике

Пусть у нас есть реализация $x_1, \dots x_n$ выборки $X_1, \dots X_n$.

Мы предполагаем, что элементы выборки имеют распределение из множества $\mathscr{P}$, где $\mathscr{P}$ - какое-то заданное множество распределений.

Однако мы не знаем, какое именно, и хотим это как-то оценить по имеющейся реализации.

В параметрическом подходе мы также будем предполагать, что на множестве $\mathscr{P}$ введен параметр: $\mathscr{P} = \{\mathrm{P}_\theta \mid \theta \in \Theta\}$, где Θ – множество параметров. Такое семейство $\mathscr{P}$ называется параметрическим.

Задача об оценке истинного распределения сводится к задаче об оценке значения параметра.

Статистики и оценки

Пусть $X = (X_1, ..., X_n)$ — выборка из неизвестного распределения $\mathrm{P} \in \mathscr{P}$.

Функция от выборки $S(X)$ называется статистикой.

Примеры статистик:

- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ — выборочное среднее;
- $\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ — выборочный момент k-го порядка;
- $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ — выборочная дисперсия.

Формально: Пусть $(\mathscr{X}, \mathscr{B}_{\mathscr{X}}, \mathscr{P})$, *– вероятностно-статистическая модель, где* $\mathscr{P} = \{\mathscr{P}_\theta | \theta \in \Theta\}$ *– некоторое параметрическое семейство распределений, а* Θ *– множество параметров.*

Если пара $(E, \mathscr{E})$ *— некоторое измеримое пространство, а* $S : \mathscr{X}^n \to E$ *— измеримое отображение, то* S *называется статистикой.*

Статистики и оценки

Пусть $X = (X_1, ..., X_n)$ — выборка из неизвестного распределения $\mathrm{P} \in \mathscr{P}$.

Функция от выборки $S(X)$ называется статистикой.

Если при этом S принимает значения во множестве Θ, то $S(X)$ называется точечной оценкой или просто оценкой параметра θ.

Обычно обозначается $\widehat{\theta}(X)$ *или просто* $\widehat{\theta}$.

Интуиция: пытаемся догадаться об истинном значении параметра на основании данных.

Иногда требуется оценить значение некоторой функции $\tau(\theta)$. В таком случае рассматривается $E = \tau(\theta)$, то есть образ множества параметров.


ВСЁ!