Лекция 1

Введение в статистику и ML

Курс: Введение в DS на УБ и МиРА (весна, 2022)
Преподаватель: Владимир Омелюсик

4 апреля 2022 г.

Предыстория

- Всевышкинский экзамен по анализу данных.
- Три уровня: начальный, **базовый**, продвинутый.
- На базовом уровне нужно знать основы статистики (проверка гипотез, линейные модели), машинного обучения (постановка задач, простые модели) и анализа данных (визуализация, обработка).

О курсе

- 10 лекций (онлайн), 14 семинаров (онлайн / очно).
- На лекциях – теория, на семинарах – практика на компьютерах. Программировать будем только на Python.
- Лекции – в Тимс, с записью. Записи будут в Тимс и на YT (ссылку сообщим). Конспекты лекций будут (где – чуть позже). Лекции где-то парные, где-то одинарные.
- Семинары – на усмотрение семинаристов.
- Семинары рассинхронизированы, и мы будем это учитывать.

Формула оценки

$$\text{Итог} = 0.5 \times \text{ДЗ} + 0.2 \times \text{Квизы} + 0.3 \times \text{Экзамен}$$

- 5 домашних заданий на компьютере, оценка за ДЗ – это средняя оценка по ним.
- Квизы проводятся на лекциях в Google-формах, предупреждаем заранее (первый квиз будет на лекции 11 апреля).
- Письменный экзамен в конце курса, не блокирующий. Автоматов не предусмотрено.
- Любая форма контроля на усмотрение преподавателя может быть представлена к устной защите.
- Округление арифметическое, округляется только итог.

Коммуникация

- Страничка на Wiki (здесь агрегируются вся информация и материалы).
- GitHub с материалами (сюда ведут ссылки со странички на Wiki).
- Телеграм-канал (тут выкладываются важные объявления).
- Телеграм-чаты групп (ссылки выложим в канал).
- Анонимная Google-форма (сюда можно писать текущий фидбэк).

И ещё две ссылки, чтобы всё было в одном месте:

- Ссылка на Тимс.
- Ссылка на плейлист с записями лекций.

План лекций

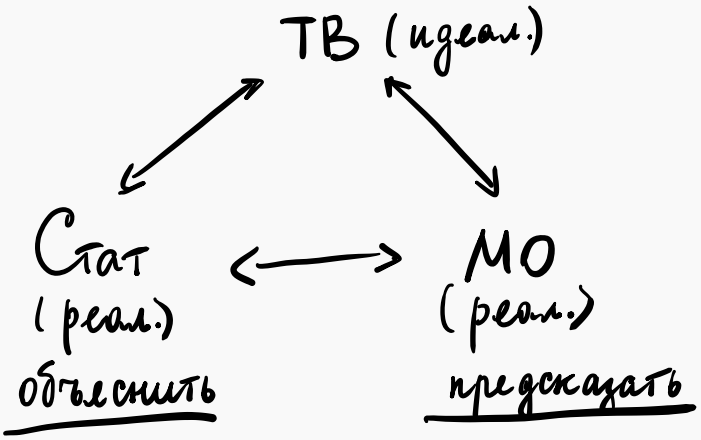
1. (4 апреля) Введение в статистику и ML. Выборка.
2. (4 апреля) Теория вероятностей. Распределения. Описательные статистики.
3. (11 апреля) Ковариация и корреляция. Кейсы по визуализации.
4. (18 апреля) Тестирование гипотез (Z- и t-тесты для одной и двух выборок). p-value.
5. (25 апреля) Непараметрика. χ^2-критерий согласия.
6. (25 апреля) Линейная регрессия (оценка и метрики).
7. (16 мая) Линейная регрессия (интерпретация и гипотезы).
8. (16 мая) Задачи машинного обучения. Кросс-валидация. Градиентный спуск.
9. (30 мая) Задача классификации. Метод k ближайших соседей.
10. (6 июня) Логистическая регрессия.

Don't Panic

- Выглядит технично, но это так кажется.
- Будем максимально использовать математику 9-10 класса.
- Много практических кейсов, чтобы было интереснее.
- Running gag про то, что машинное обучение – это `.fit()` и `.predict()`.
- Understanding Machine Learning through Memes.

Теория вероятностей, статистика и машинное обучение

- Теория вероятностей: идеальный мир, кубики и урны.
- Статистика: цель – построить модель, которая объясняет мир вокруг нас.
- Машинное обучение: цель – построить модель, которая хорошо предсказывает какую-то величину.



Играл. кубик, 6 граней, все грани равнов-о

$P(\text{грань } \cdot) = \frac{1}{6}$

$P(\text{грань } \cdot\cdot) = \frac{1}{6}$

$\vdots$

$P(\text{грань } \begin{smallmatrix}\cdot\cdot\\\cdot\cdot\end{smallmatrix}) = \frac{1}{6}$

? А если 30 раз, то сколько в среднем

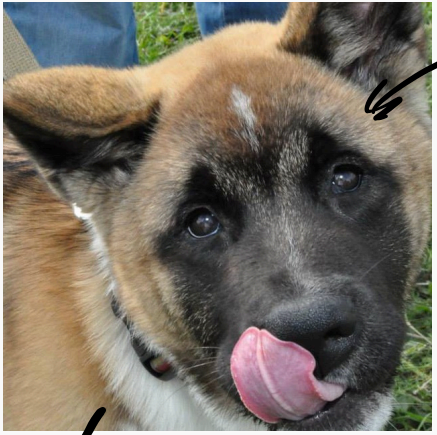
? А если 2 раза подряд, то $\begin{smallmatrix}\cdot\cdot\\\cdot\cdot\end{smallmatrix}$, $\begin{smallmatrix}\cdot\cdot\\\cdot\cdot\end{smallmatrix}$

? ...

Статистика: пример с ростом собак

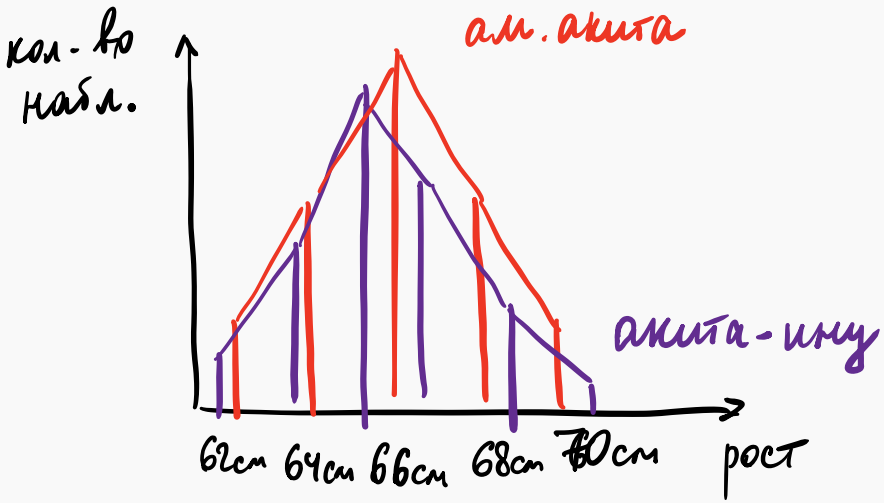
- Выборка из 1000 независимых измерений роста собак породы американская акита. Средний рост – 66 см.
- Выборка из 1000 независимых измерений роста собак породы акита-ину. Средний рост – 65 см.

акита-ину



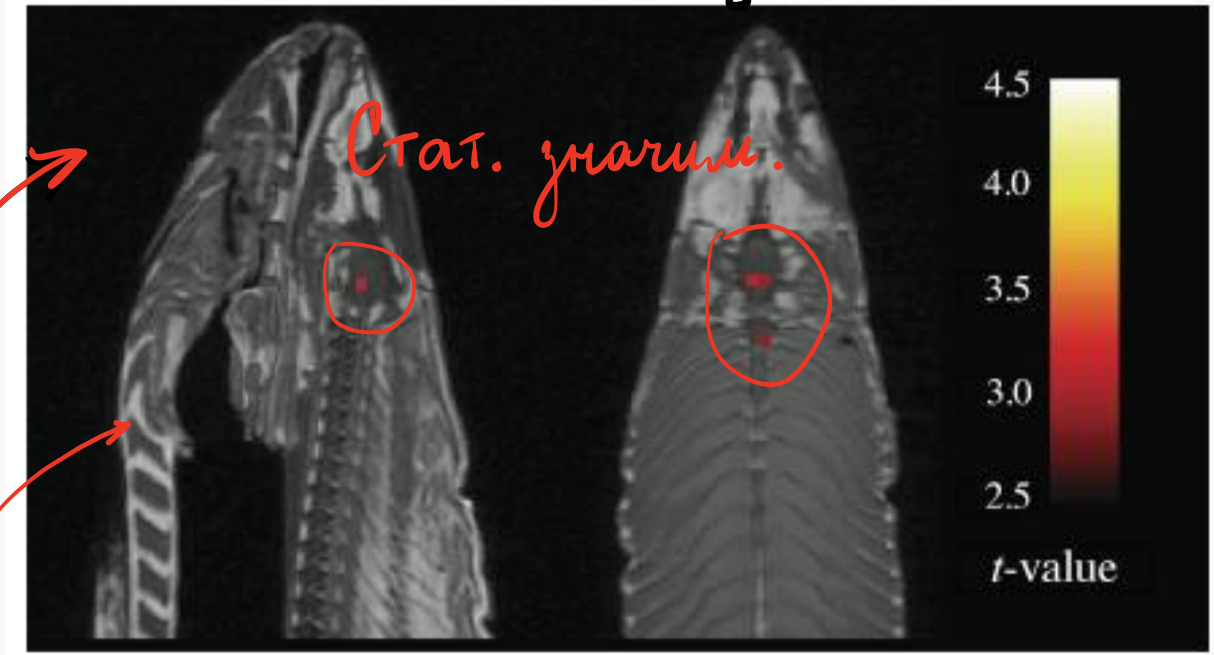
амер. акита

Статистика: пример с ростом собак

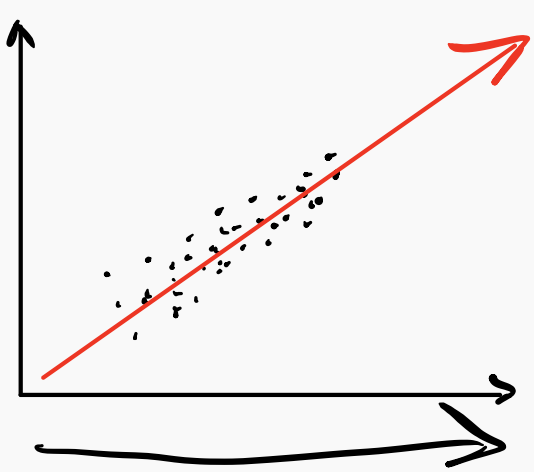


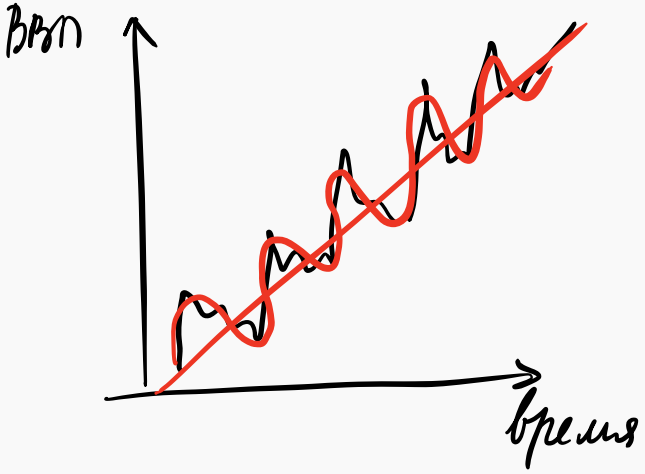
Статистика: пример с лососем

- Почитать Dead Salmon Study (Bennet et.al.)

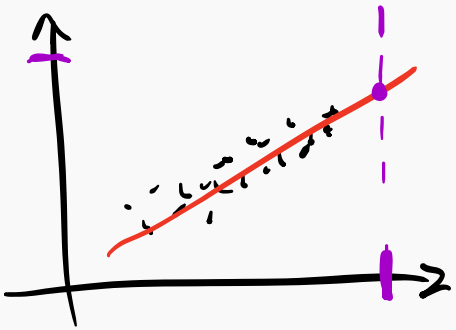


Статистика: пример с регрессией и временными рядами





Машинное обучение: пример со стоимостью квартиры

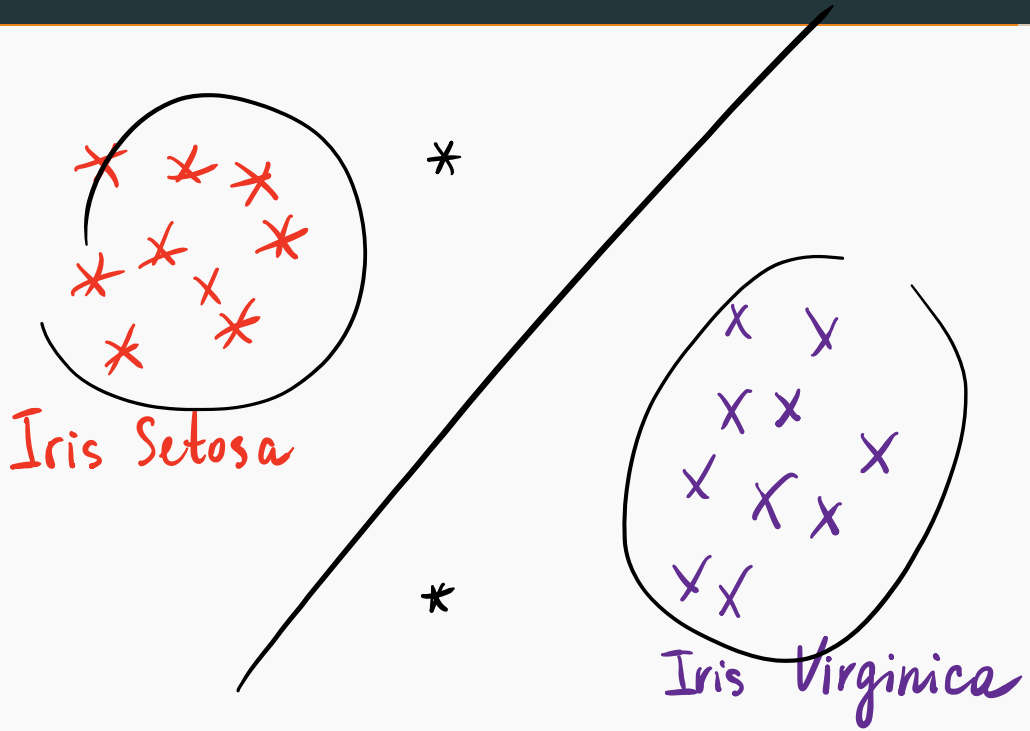


Набор данных:

стоим. кв. (у.е.)	$S, м^2$	расст. до метро
10	15	.
20	45	.
70	100	.
15	.	
⋮	.	
	.	
	⋮	
	⋮	

1000 наб.

Машинное обучение: пример с классификацией цветков ириса



ТВ
(инструм.)
стат ⇔ МО
(объяс.) (предск.)

Генеральная совокупность и выборка

Генеральная совокупность
Совокупность всех мыслимых объектов, относительно которых предстоит делать выводы в исследуемой задаче.

Выборка
Часть генеральной совокупности, используемая в исследовании.

Обозначения

$X = [3, 4, 5, 1, 7, 8, \ldots, 9, 10]$

$Y = [1000, 1500, 700, 900]$ "вектора" ≡ "списки чисел"

$Z =$

$A =$

$B =$

Примеры

Глубоков. кольмагоп – ген. Сов.

↳ ① К. в Fix. ок.

② К. с глуб. булк.

Все ли выборки хороши?

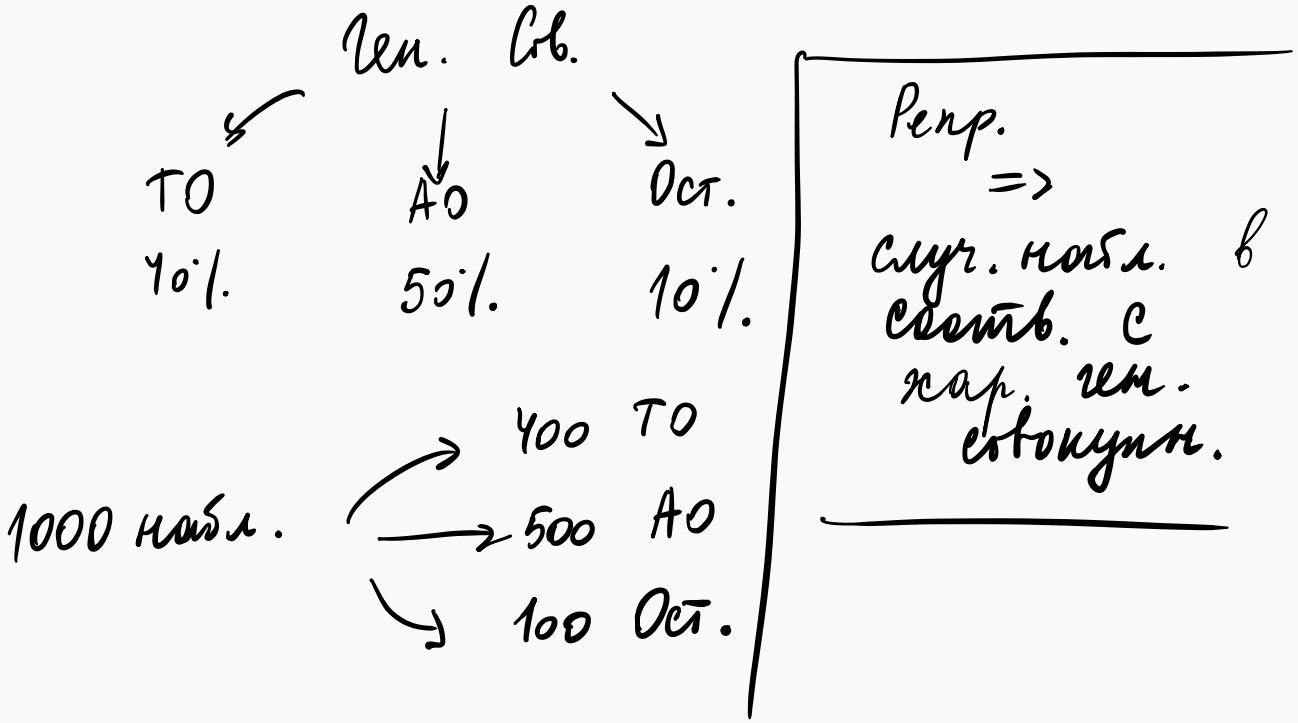
Репрезентативность выборки

Репрезентативность выборки
Соответствие характеристик выборки характеристикам генеральной совокупности.

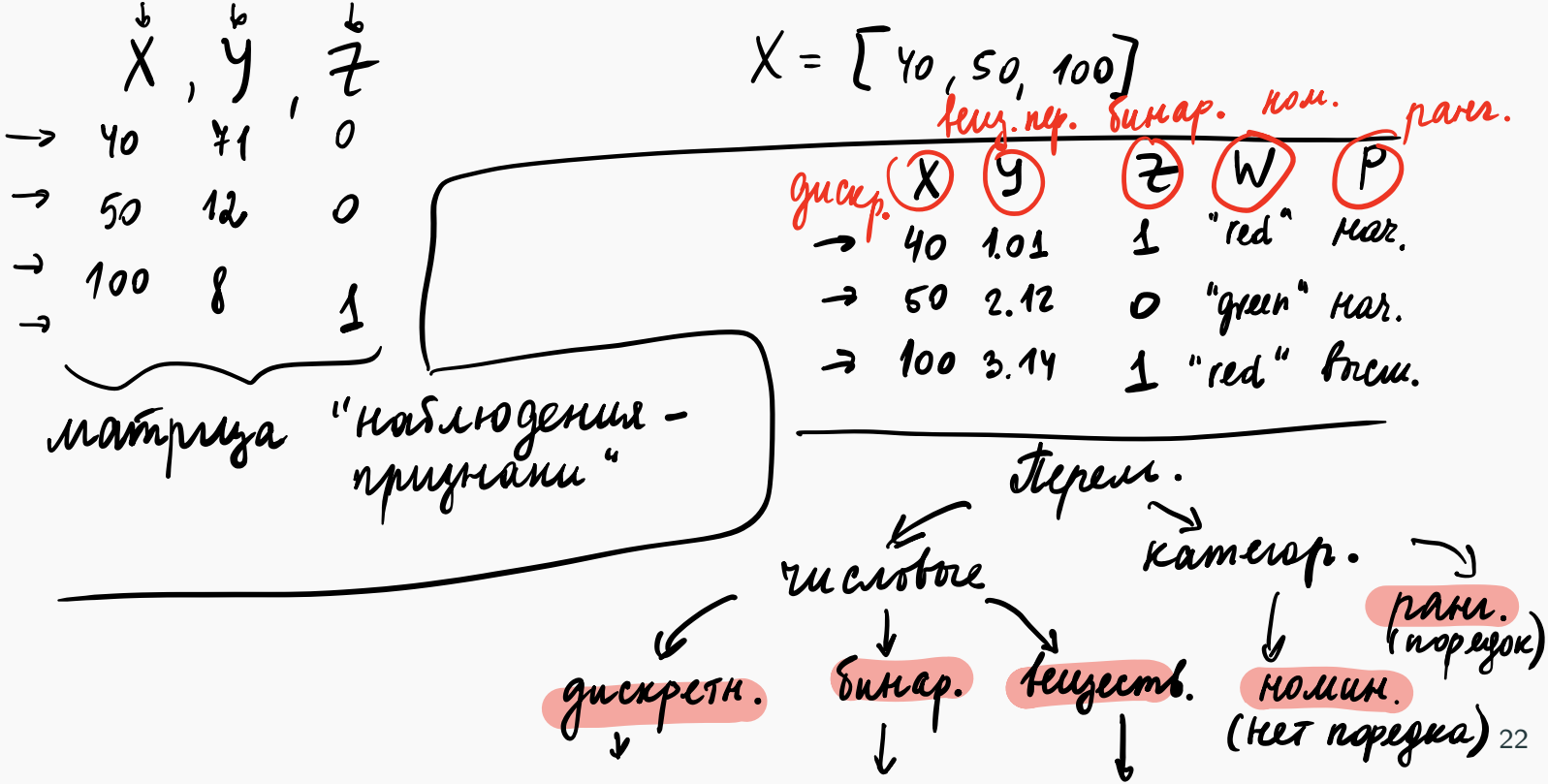
Способы построения выборки

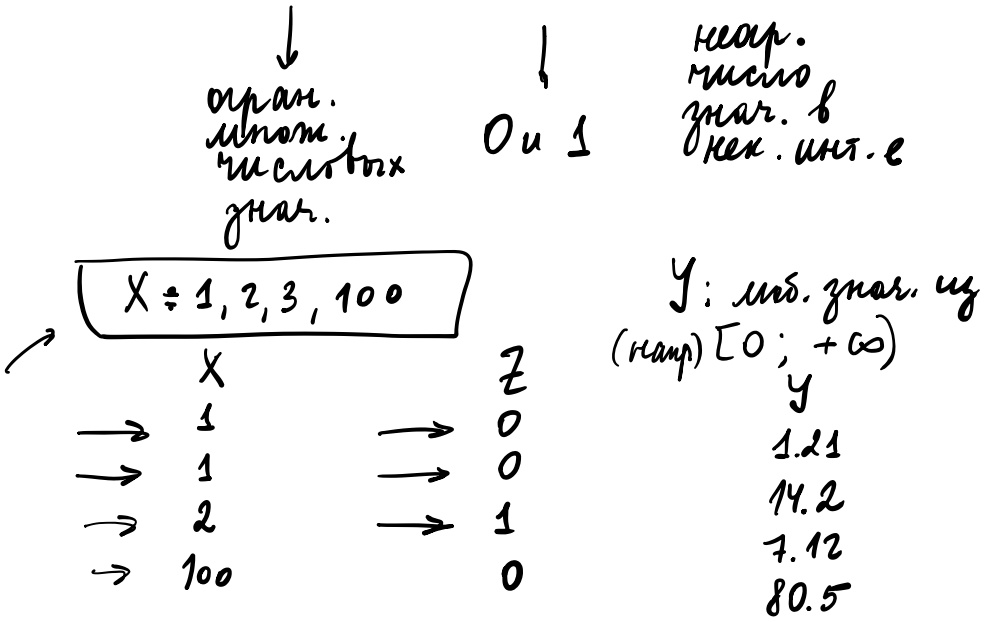
* Опред. факторы / хар-и гс
* Сегм. гс
* Случ.

Случайная выборка



Типы данных


X, y, z
40 71 0
50 12 0
100 8 1
матрица "наблюдения - признаки"
X = [40, 50, 100]
чещ.пр. бинар. ном. ранг.
дискр. X y z W P
40 1.01 1 "red" нач.
50 2.12 0 "green" нач.
100 3.14 1 "red" высш.
Перем.
числовые
категор.
дискретн.
бинар.
веществ.
ранг. (порядок)
номин. (нет порядка)



Номин: цвета, материалы, назв. городов

Ранг: степени отр., воинск. зв.