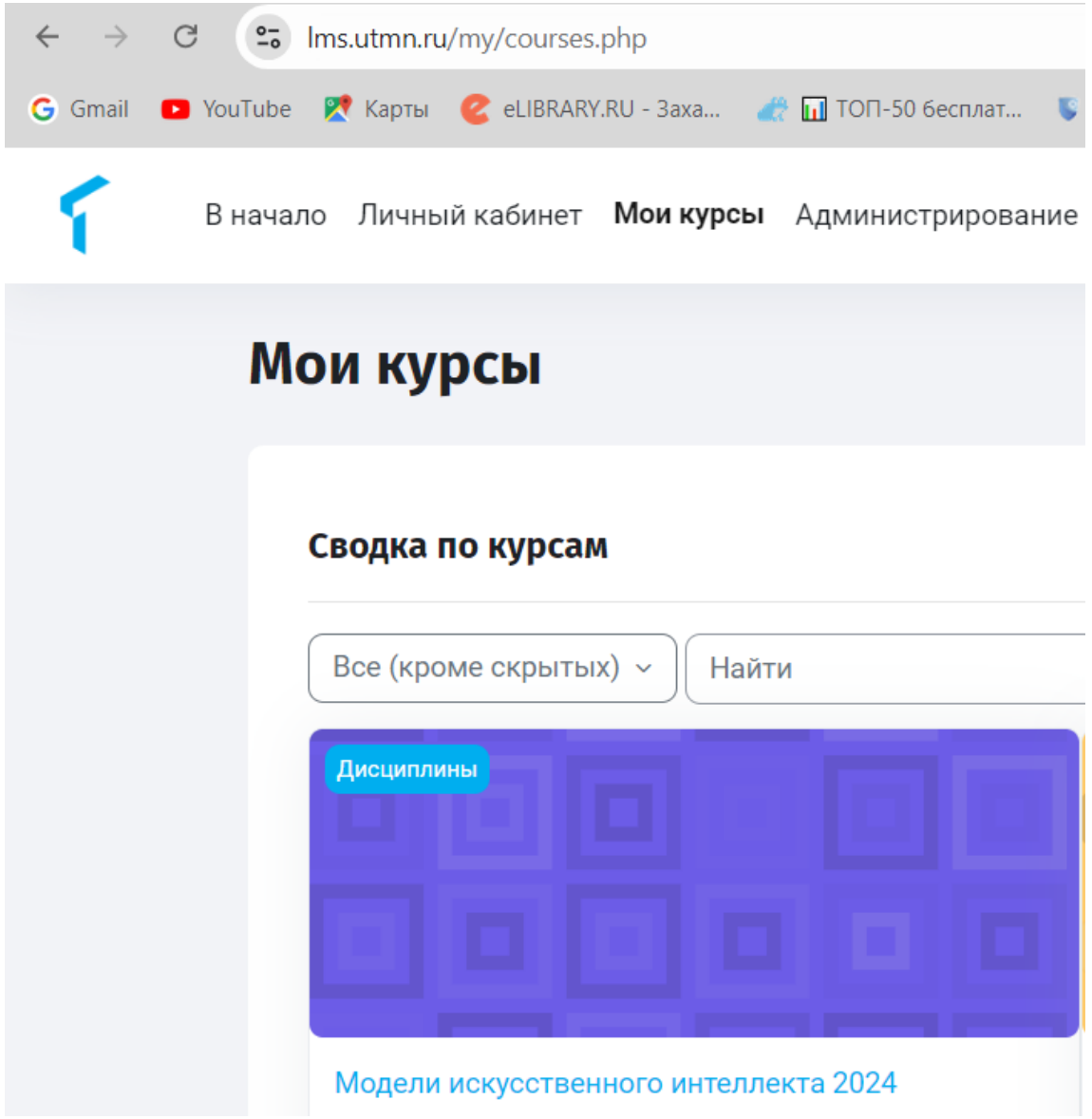
Основные цели (в рамках проекта)

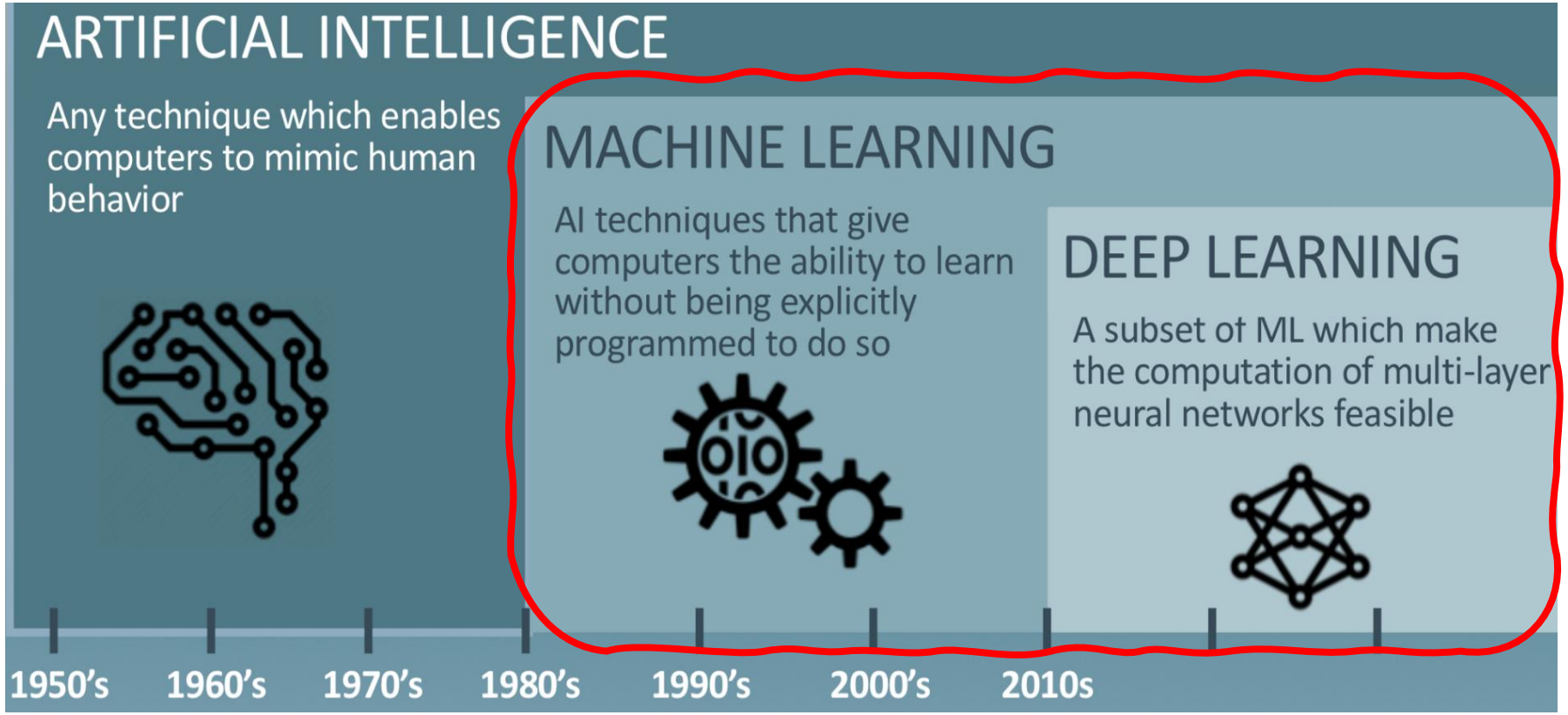
- Адекватное применение математических методов для предварительного анализа данных.
- Создание и оценка моделей для решения задач интерпретации и прогнозирования.
- Оформление и публичное представление результатов анализа данных.

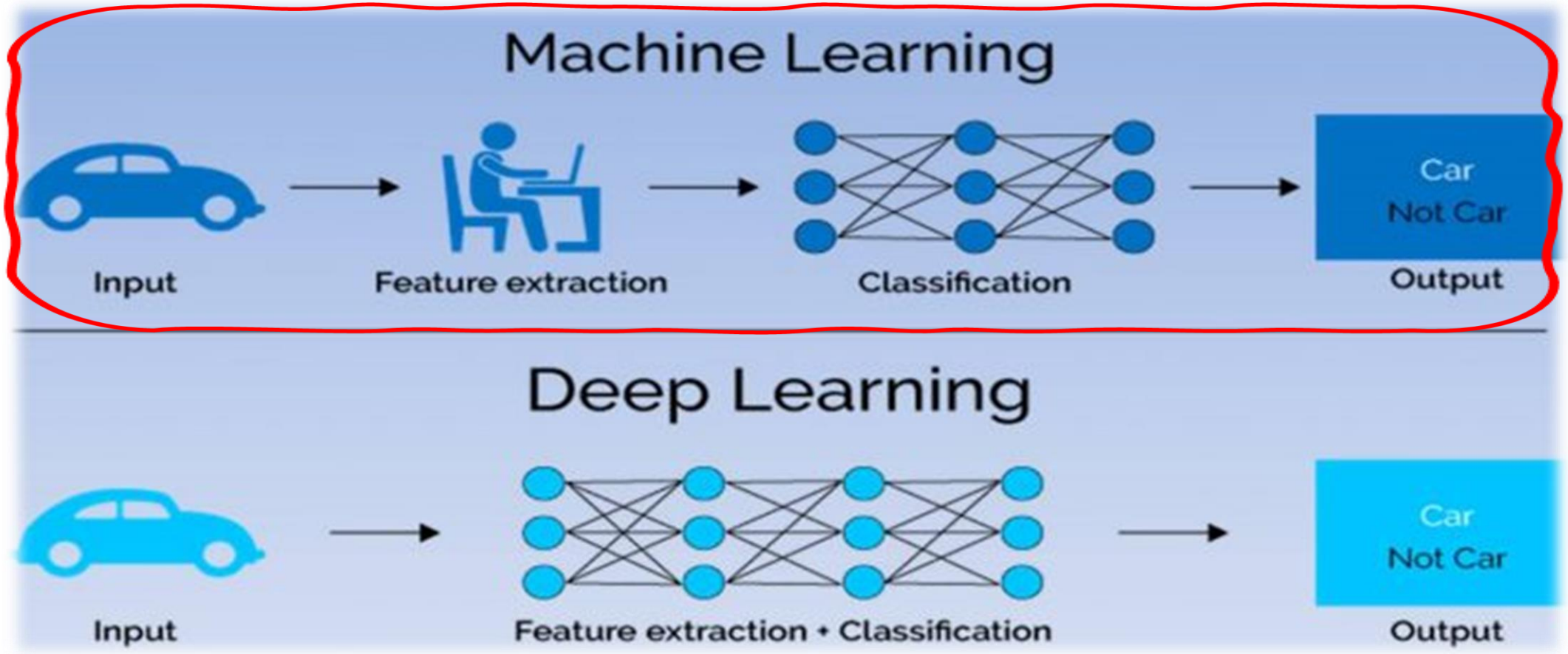
Оценивание

- Выполнение практических заданий на данных проекта в течение семестра.
- Защита отчета по проекту в конце семестра:
  - блокноты/приложение,
  - текст до 10 страниц, включающий *Введение, Материалы и методы, Результаты, Выводы*.

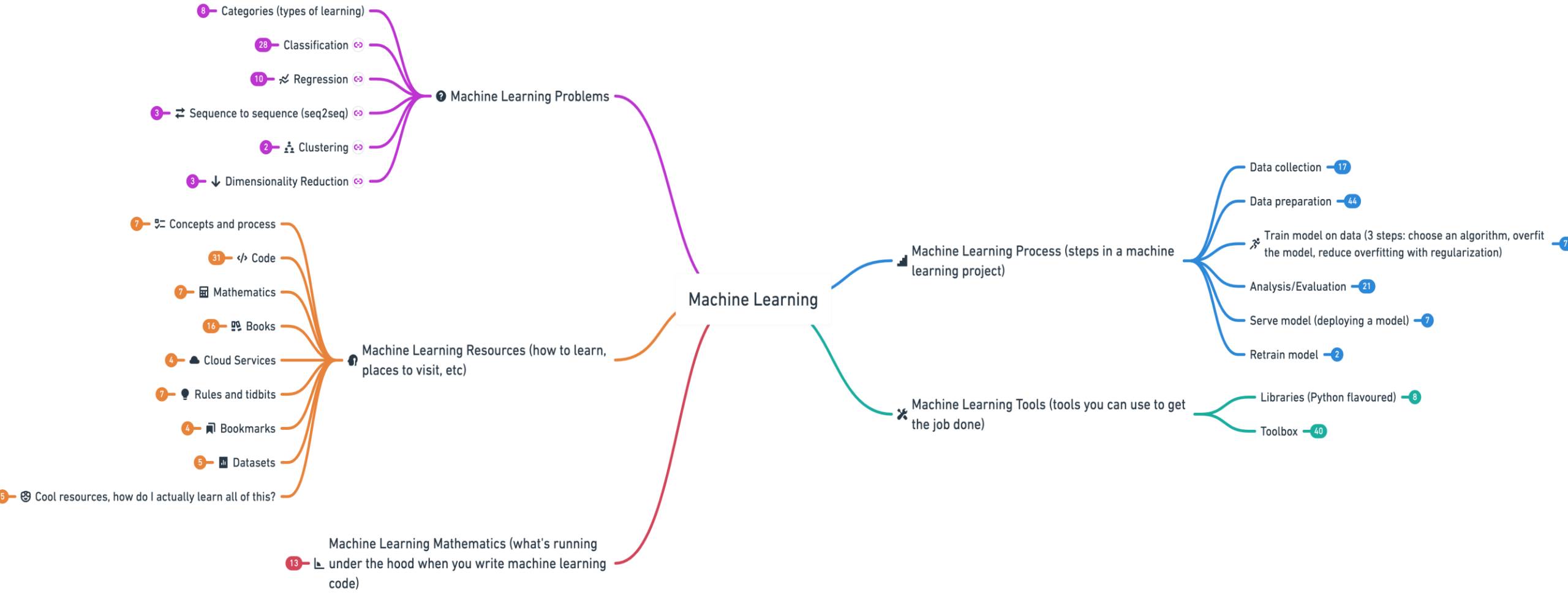


AI, искусственный интеллект - история




Machine Learning
Input
Feature extraction
Classification
Car
Not Car
Output
Deep Learning
Input
Feature extraction • Classification
Car
Not Car
Output

2020 Machine Learning Roadmap (90% для 2024) -> ->

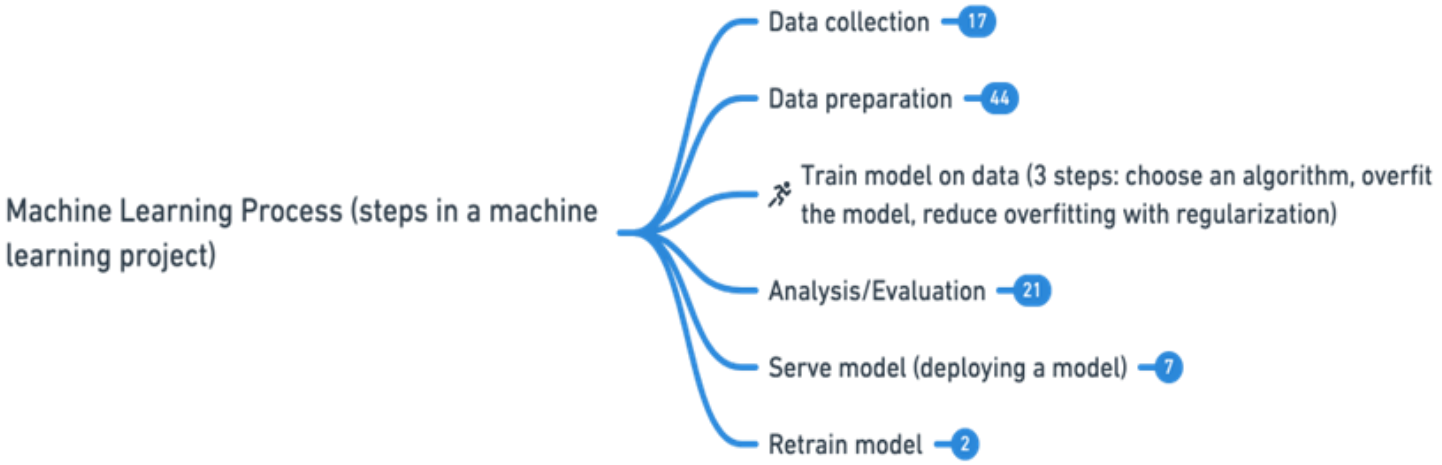


Задачи машинного обучения

- Классификация
- Регрессия
- Seq2seq
- Кластеризация
- Понижение размерности

Процесс машинного обучения (подробности в ML_roadmap)

- Сбор данных
- Подготовка данных
- Обучение модели
  - Выбор алгоритма
  - Обучение
  - Регуляризация
- Анализ и оценка модели
- Внедрение модели
- Пере(до-)обучение модели



Основные задачи Data Science

Анализ данных для интерпретации ситуации (объекта) **descriptive analysis**

Анализ данных для прогноза **predictive analysis**

Анализ данных для выработки рекомендаций **prescriptive analysis**

Методология CRISP DM

CRISP DM - **стандартизированный** подход к реализации проектов, связанных с **анализом данных**.

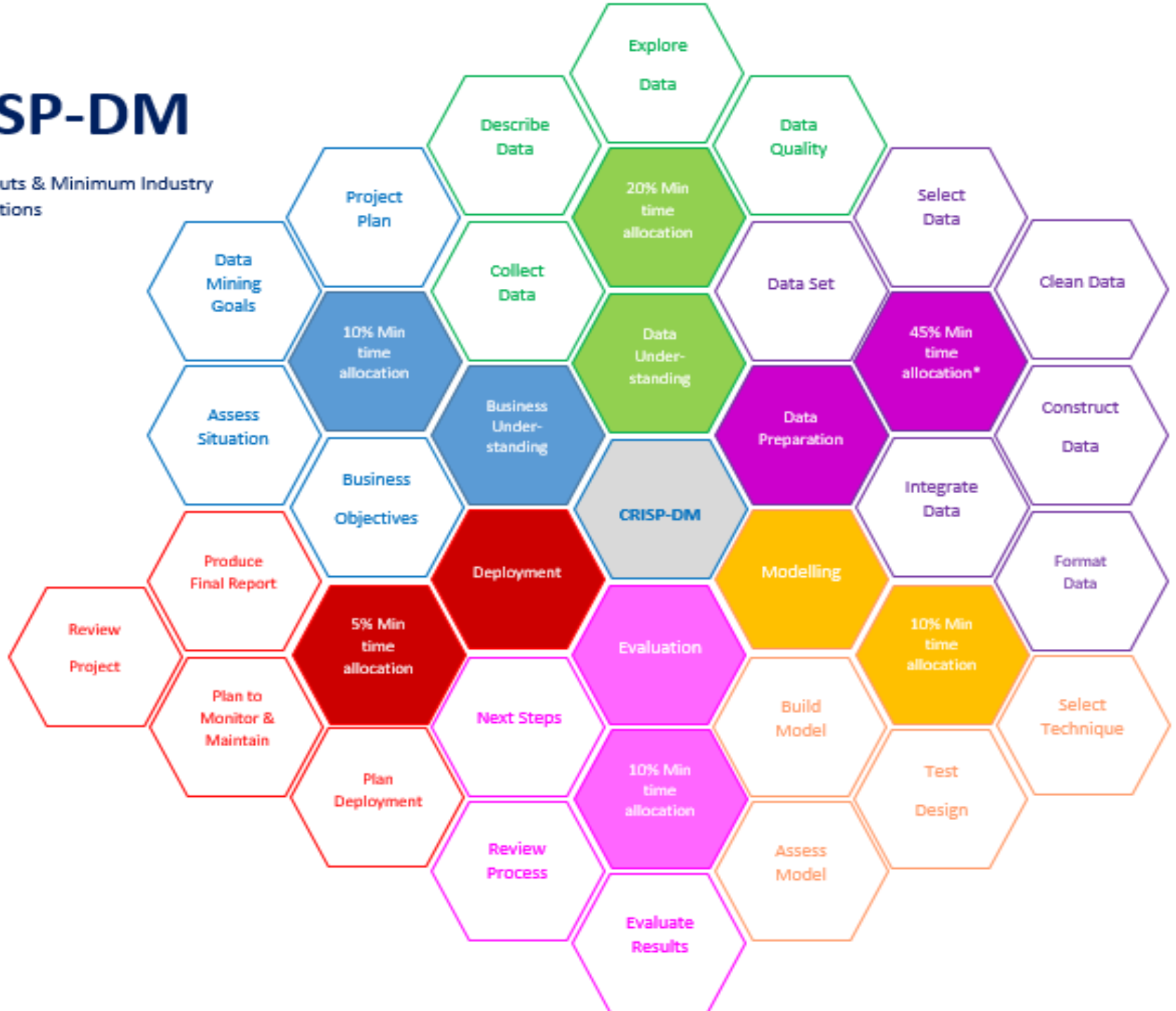
Cross-Industry Standard Process for Data Mining, межотраслевой стандартный процесс для исследования данных.

- **Процесс анализа данных**:
  - ✓ **6 стадий** со своими задачами,
  - ✓ описывается **жизненным циклом**

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth (2000);

CRISP-DM

Tasks, Outputs & Minimum Industry Time Allocations



Key:

Phase

Time %

Output

Notes:

* Data preparation minimum time allocation 50% (Shearer 2000) adjusted to 45% to enable total minimum time allocations to equal 100%.

Reference:

Shearer, C. 2000, "The CRISP-DM Model: The New Blueprint for Data Mining", *Journal of Data Warehousing*, vol. 5, no. 4 Fall, pp. 13.

1. Понимание бизнеса (Business Understanding)
10%

Определить бизнес цели

Оценить ситуацию (ресурсы, риски, требования, ограничения)

Определить цели анализа данных

Составить план проекта

2. Понимание данных (Data Understanding) **20%**

Собрать исходные (сырые, raw) данные

Описать данные

Исследовать данные (например, найти явные несоответствия, недостаток данных для бизнес цели)

Проверить качество данных

3. Подготовка и предварительная обработка данных (Data Preparation & Preprocessing)
45%


Отобрать данные для анализа (dataset)


Извлечь информацию


Очистить/восполнить данные


Привести данные в нужный формат

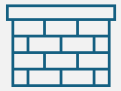

Объединить и/или агрегировать данные


Сформулировать гипотезу о решении проблемы

4. Построение моделей (Modeling) **10%**

Выбрать методы/технику моделирования

Подготовить тесты для оценки модели



Построить модель (модели)

Оценить качество модели

5. Оценка результатов (Evaluation) **10%**

Оценить результаты с точки зрения достижения бизнес целей

Оценить нерешенные бизнес задачи

Описать полученные результаты

Определить следующие шаги

6. Внедрение (Deployment) **5%**

Разработать план внедрения

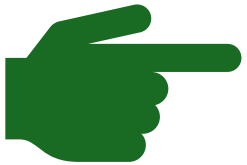
Разработать план сопровождения

Подготовить финальный отчет

Сформулировать общие выводы по результатам проекта

Зачем знать математику?

Выдвижение корректных гипотез и идей.



Выбор адекватных методов.

Обоснование результатов.

Математический анализ

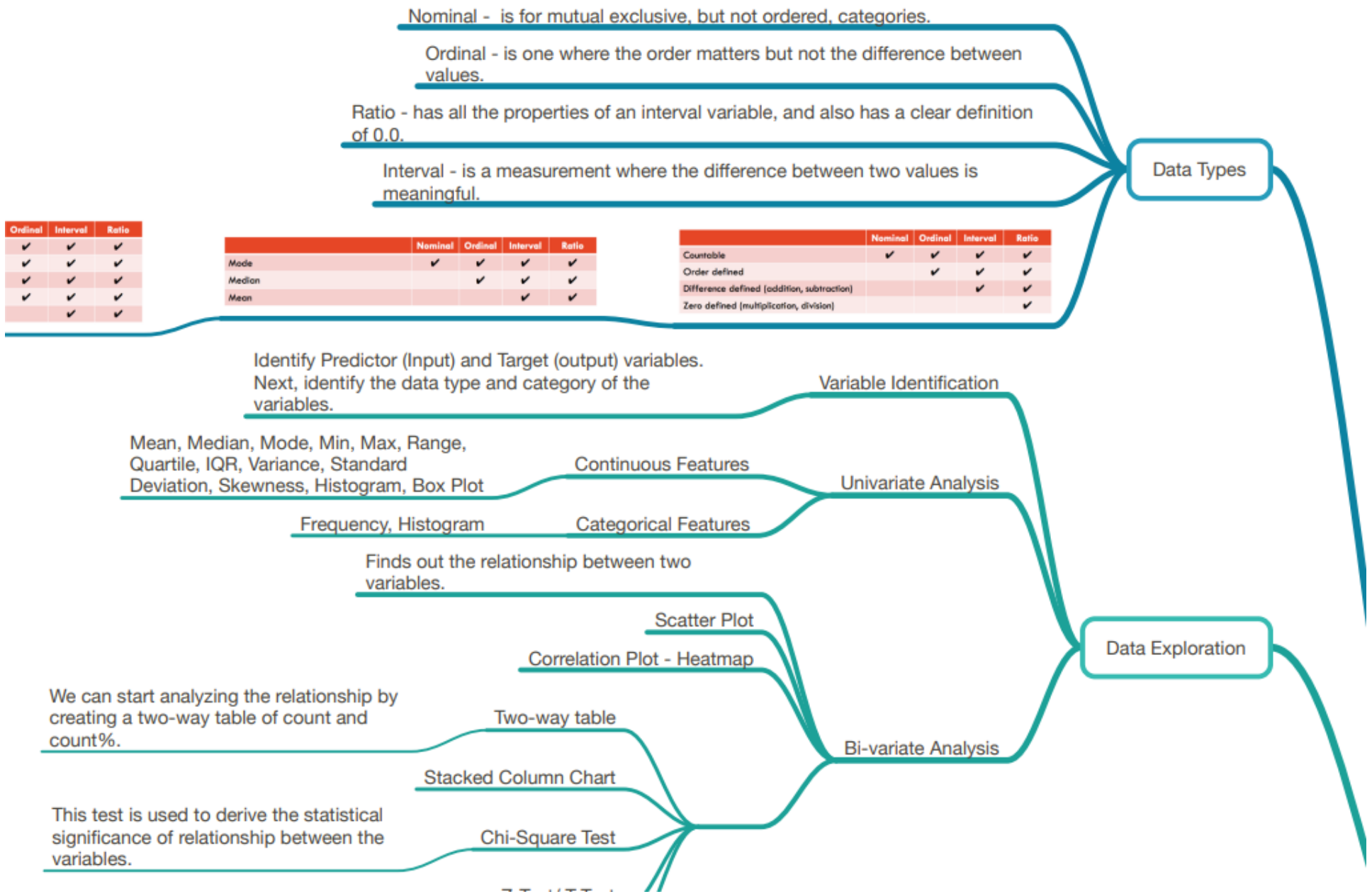
- Функции одной переменной, пределы, дифференцируемость;
- Основы интегрального исчисления, оценка определенных интегралов;
- Специальные функции;
- Функции многих переменных, частные производные, градиент;
- Основы обыкновенных и дифференциальных уравнений в частных производных.

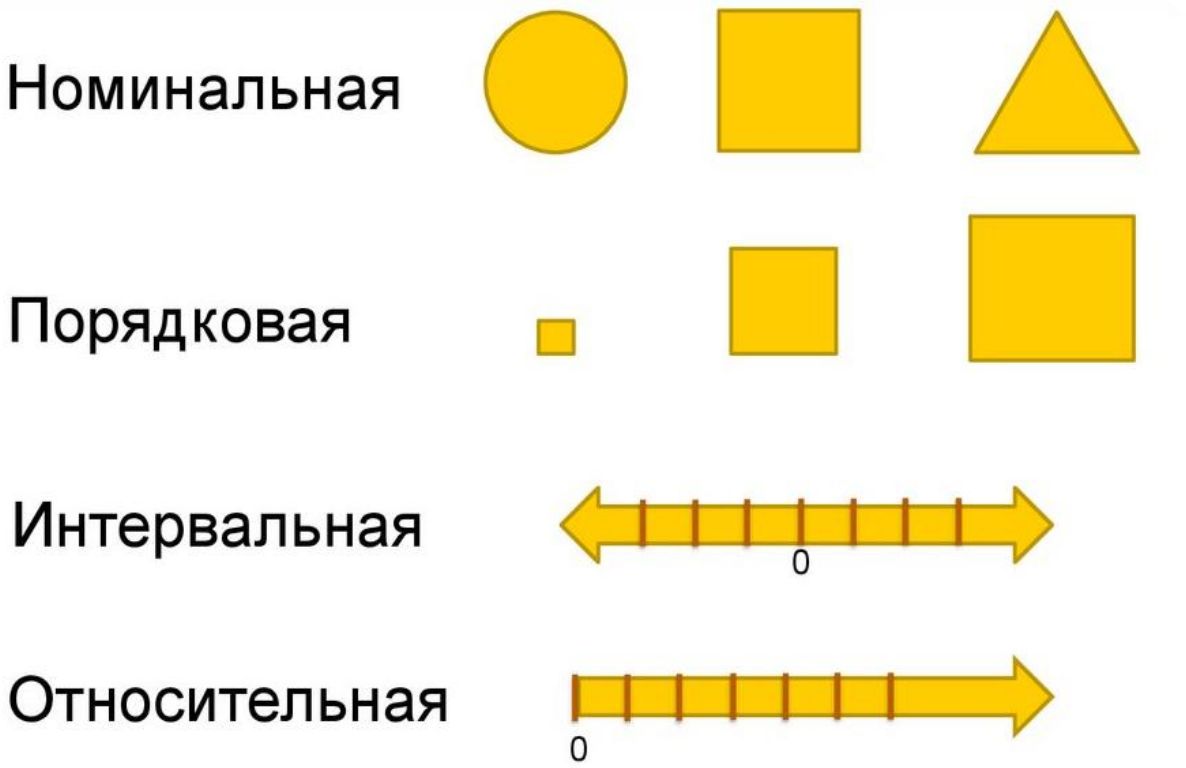
Линейная алгебра

- Основные свойства матриц и векторов: скалярное умножение, линейное преобразование, транспонирование, ранг, детерминант;
- Внутренние и внешние произведения, правило умножения матриц и различные алгоритмы, обратные матрицы;
- Специальные матрицы: квадратная, единичная и треугольная матрицы, представление о разреженной и плотной матрице;
- Векторное пространство, базис, ортогональность, ортонормированность, метрики;
- Собственные вектора и значения;

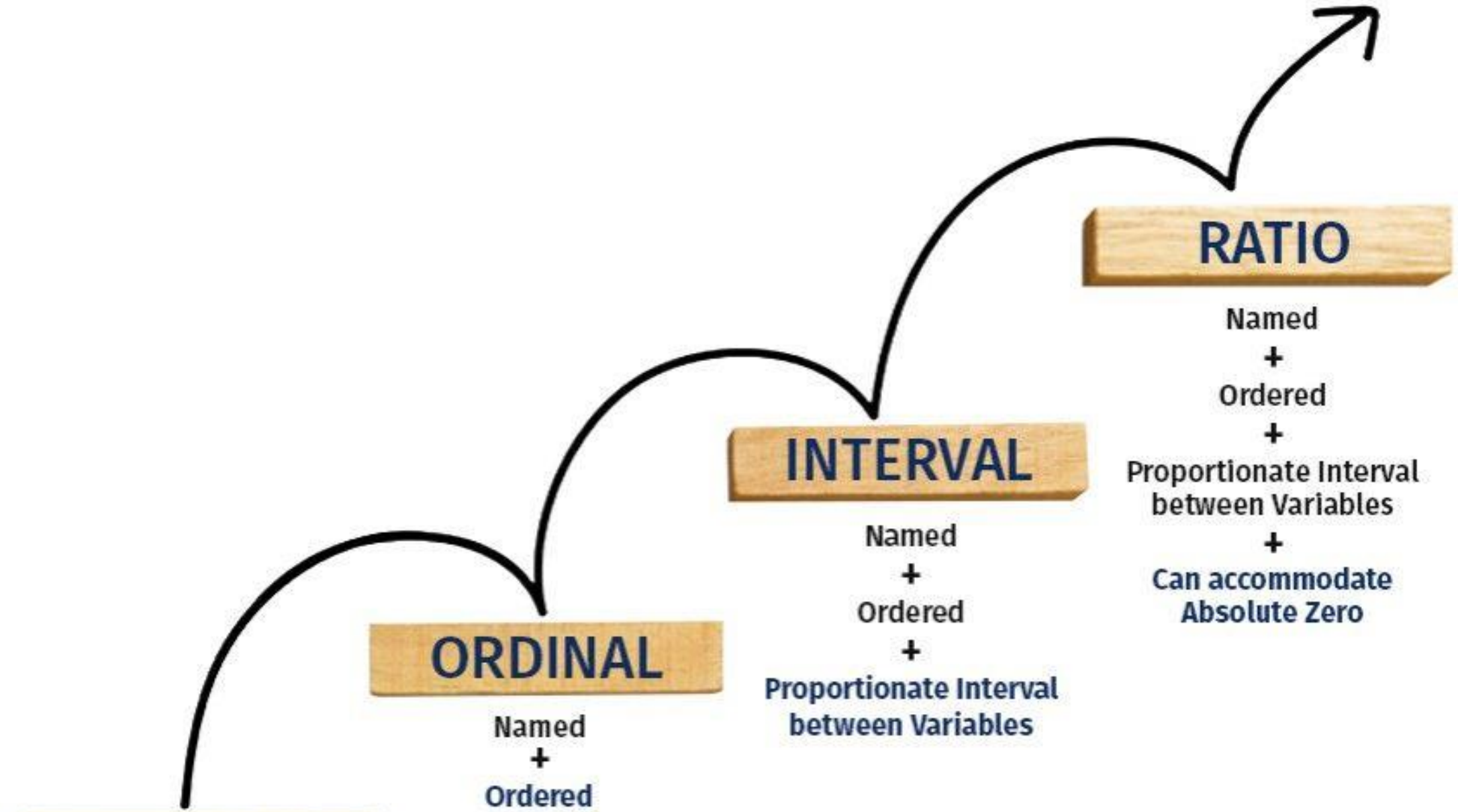
Математическая статистика

- Обобщение данных и описательная статистика, центральная тенденция, дисперсия, ковариация, корреляция;
- Базовая вероятность: основная идея, условная вероятность;
- Функции распределения вероятностей: равномерные, нормальные, биномиальные, дискретные и непрерывные;
- Выборка, измерение;
- Проверка гипотез.


Data Types
Nominal - is for mutual exclusive, but not ordered, categories.
Ordinal - is one where the order matters but not the difference between values.
Ratio - has all the properties of an interval variable, and also has a clear definition of 0.0.
Interval - is a measurement where the difference between two values is meaningful.
Ordinal Interval Ratio
✔ ✔ ✔
✔ ✔ ✔
✔ ✔ ✔
✔ ✔ ✔
✔ ✔
Nominal Ordinal Interval Ratio
Mode ✔ ✔ ✔ ✔
Median ✔ ✔ ✔
Mean ✔ ✔
Nominal Ordinal Interval Ratio
Countable ✔ ✔ ✔ ✔
Order defined ✔ ✔ ✔
Difference defined (addition, subtraction) ✔ ✔
Zero defined (multiplication, division) ✔
Data Exploration
Variable Identification
Identify Predictor (Input) and Target (output) variables. Next, identify the data type and category of the variables.
Univariate Analysis
Continuous Features
Mean, Median, Mode, Min, Max, Range, Quartile, IQR, Variance, Standard Deviation, Skewness, Histogram, Box Plot
Categorical Features
Frequency, Histogram
Bi-variate Analysis
Finds out the relationship between two variables.
Scatter Plot
Correlation Plot - Heatmap
Two-way table
We can start analyzing the relationship by creating a two-way table of count and count%.
Stacked Column Chart
Chi-Square Test
This test is used to derive the statistical significance of relationship between the variables.



Шкалы измерений – как измерены данные?



Иерархия шкал

Особенности шкал - операции

Абсолютный ноль

Равенство интервалов

Операции \ Тип шкалы	Номинальная	Порядковая	Интервальная	Отношений
= !=	x	x	x	x
< >		x	x	x
+ -			x	x
* /				x

Что можно оценить?

	НОМИНАЛЬНАЯ	ПОРЯДКОВАЯ	ИНТЕРВАЛЬНАЯ	ОТНОШЕНИЙ
порядок		x	x	x
мода	x	x	x	x
медиана		x	x	x
среднее			x	x
разность значений			x	x

Центральная тенденция и отклонение

	Номинальная	Порядковая	Интервальная	Отношений
Центр	Мода	+ Медиана	+ Среднее арифметическое	+ Среднее геометрическое или гармоническое
Отклонение	Только по отдельным значениям	+ Перцентили	+ Дисперсия (или среднеквадратич. отклонение	+ Процентные отклонения

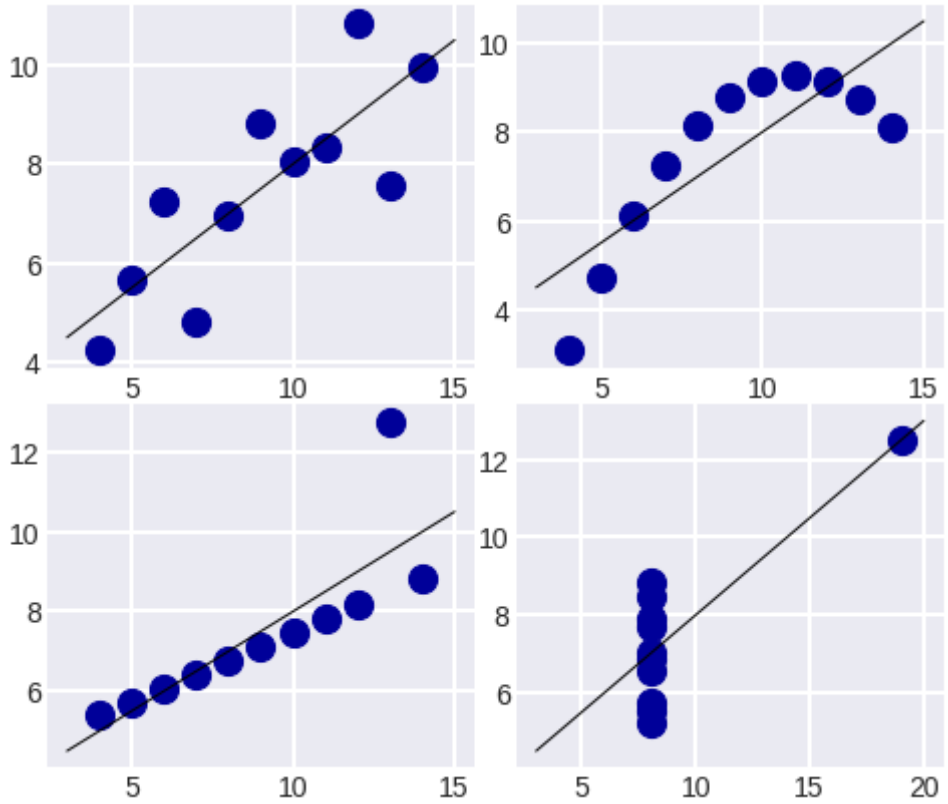
Среднеквадратическое отклонение

- Статистическая характеристика распределения случайной величины, показывающая среднюю степень разброса значений величины относительно математического ожидания (среднего).
  σ – отклонение.
  D - дисперсия.
- Случайная величина – характеристика, принимающая разные значения с определенной вероятностью.

$$D = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}}$$

Зачем визуализировать данные?
Квартет Энскомба



Характеристики наборов данных

Характеристика	Значение
Среднее значение переменной x	9.0
Дисперсия переменной x	10,0
Среднее значение переменной y	7,5
Дисперсия переменной y	3,75
Корреляция между переменными x и y	0,816
Прямая линейной регрессии	y=3+0,5x
Коэффициент детерминации линейной регрессии	0,67

Основные понятия ТВ и МС

- **Случайное событие** – событие, которое может произойти или не произойти при заданном наборе факторов.
- **Вероятность**, P – числовая мера правдоподобия случайного события.
- **Случайная величина** – характеристика, зависящая от случайного исхода некоторого опыта и принимающая разные значения с определенной вероятностью.

Свойства вероятности

- $0 \leq P(A) \leq 1$ для любого события A
- P=0: невозможное событие
- P=1: достоверное событие
- C = A или B = A + B – объединение, или сумма событий
- C = A и B = AB – пересечение, или произведение событий
- $P(A + B) = P(A) + P(B) – P(AB)$
- Если A и B несовместимы, то $P(AB) = 0$, $P(A + B) = P(A) + P(B)$

Зависимость/ независимость событий

- **О1.** События A и B называются **независимыми**, если **P(AB) = P(A)P(B)**
- P(A|B) – **условная вероятность** события A при условии, что произошло событие B:
  P(A|B) = P(AB) / P(B), при P(B) > 0
- Формула умножения вероятностей:
  P(AB) = P(A|B) P(B) = P(B|A) P(A)
- **О2.** Событие A **не зависит** от события B и наоборот, если **P(A|B) = P(A)**

Как измерить вероятность – закон больших чисел

- Прямое измерение – опыт в неизменных условиях
- N – число повторений опыта (>>1)
- N(A) – число повторений события A
- Тогда **P(A) ≈ N(A)/N** – суть теоремы Бернулли.
- **Т. Бернулли – простая формулировка закона больших чисел**:
  если вероятность события одинакова во всех испытаниях, то с увеличением числа испытаний частота события стремится к вероятности события и перестает быть случайной.

Распределение вероятностей случайной величины

- Математическое описание случайной величины – **множество ее значений и распределение вероятностей** на этом множестве.
- Пример
  - Множество значений:
    {орел, решка}
  - Распределение вероятностей:
    P(орел) = 0.5, P(решка) = 0.5

Случайные величины – дискретные и непрерывные

Дискретная случайная величина – множество возможных значений:

- **конечно** *или*
- **счётно** (можно перенумеровать 0, 1, 2, …)

Пример

- Даны две булевы случайные величины A и B.
- P(A) = 1/2, P(B) = 1/3 и P(A | не-B) = 1/4.
- P(A | B) = ?

Случайные величины – дискретные и непрерывные

- **Если случайная величина X принимает любые вещественные значения, то для любого значения x получим**
  $P(X = x) = 0$
- **Поэтому непрерывную случайную X описывают с помощью функций:**
  - **Функция распределения:**
    $F(x) = P(X \leq x)$
  - **Функция плотности вероятности:**
    $f(x) = F'(x)$

Свойства функции плотности вероятности

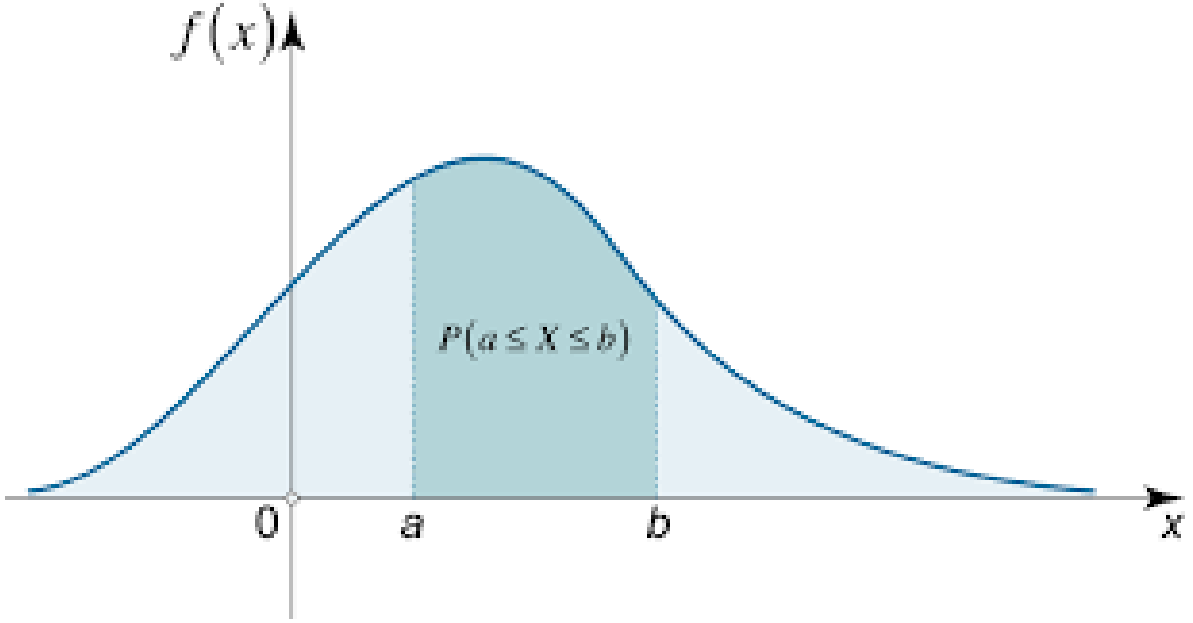
(i) $F(x) = \int_{-\infty}^{x} f(x)dx$

(ii) $f(x) = F'(x)$

(iii) $P(a < X < b) = P(a \leq X < b)$

$= P(a < X \leq b) = P(a \leq X \leq b)$

$= \int_{a}^{b} f(x)dx$



Примеры распределений

Дискретные

- **Биномиальное -** число случайных событий в серии опытов
- **Пуассона -** число случайных событий за определенное время
- …

Непрерывные

- **Нормальное -** случайные события, зависящие от многих случайных факторов
- **Экспоненциальное -** случайные временные периоды, зависящие от случайных факторов
- …

Числовые характеристики распределения вероятностей

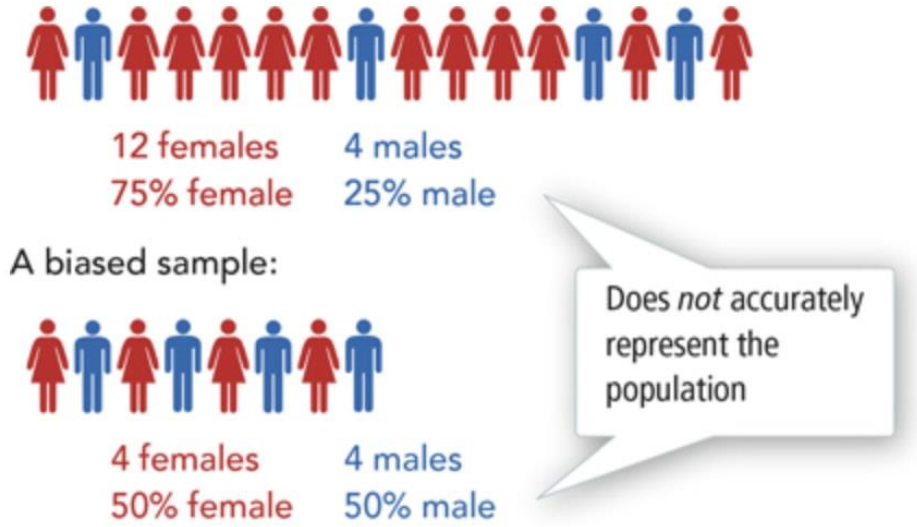
Mean - математическое ожидание, Variance - дисперсия,
Standard Deviation - стандартное отклонение

	Discrete Random Variable	Continuous Random Variable
Mean (Expected Value)	$\mu = E(X) = \sum_{i=1}^{n} xf(x)$	$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$
Variance	$\sigma^2 = V(X) = \sum_{i=1}^{n} (x-\mu)^2 f(x)$	$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$
Standard Deviation (Standard Error)	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$

Независимость случайных величин и коэффициент корреляции

- Если случайные величины X и Y независимы, то коэффициент корреляции
  $corr(X, Y) = 0$
- Обратное утверждение неверно
- $|corr(X, Y)| = 1 \Leftrightarrow X, Y$ – линейно связаны:
  то есть существуют a и b такие, что
  $P(Y = aX + b) = 1$

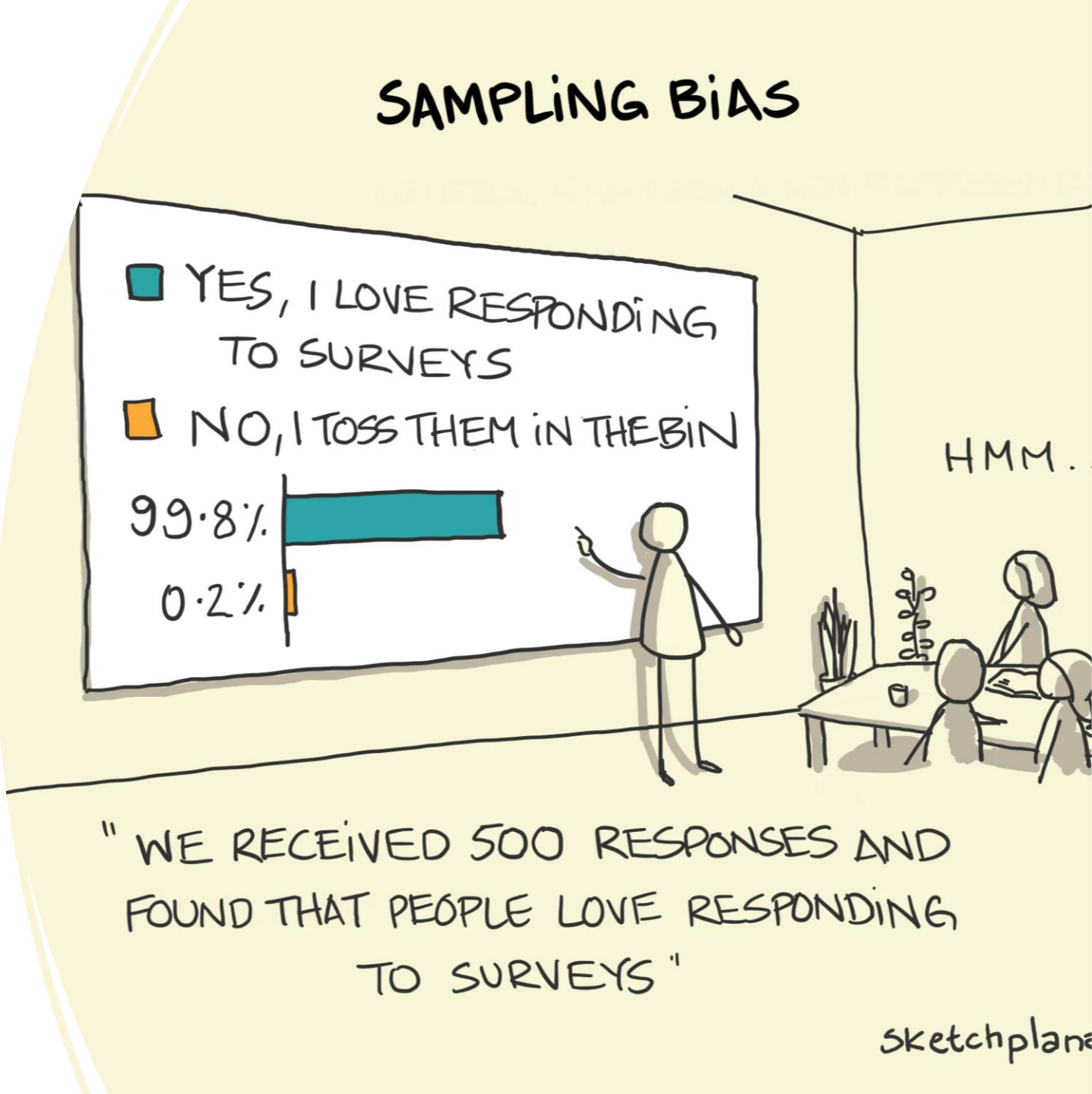
Случайные выборки



- **Генеральная совокупность** (population) – все множество изучаемых объектов
- **Выборка** (sample) – выбранная группа объектов
- **Репрезентативная** выборка наилучшим образом представляет генеральную совокупность
- **Случайный выбор** – способ получения репрезентативной выборки
- Случайный выбор n объектов из N – все наборы из n объектов могут быть выбраны с одинаковой вероятностью

Смещение выборки

- **Смещение выборки** (sampling bias) – выборка представляет не всю генеральную совокупность
- Причина смещения - некоторые выбираемые элементы совокупности имеют более низкую или более высокую вероятность выбора, чем другие
- **Смещение – основной источник ошибок при обучении моделей**



Характеристики выборки

- **Эмпирическая функция распределения** – вместо вероятности частота.
  Когда данных много - на графиках
- **Выборочное среднее** - число
- **Выборочная дисперсия** - число
- **Выборочный коэффициент корреляции** (Пирсона) - число

Нормальное распределение

- **Центральная предельная теорема**:
  - Если случайная величина является суммой многих случайных слабо взаимозависимых величин, каждая из которых вносит малый вклад относительно общей суммы,
    то центрированное и нормированное распределение такой величины при достаточно большом числе слагаемых стремится к нормальному распределению.
- Нормальное распределение описывает случайные явления, зависящие от **большого количества независимых случайных факторов, среди которых нет сильно выделяющихся**.

Нормальное распределение

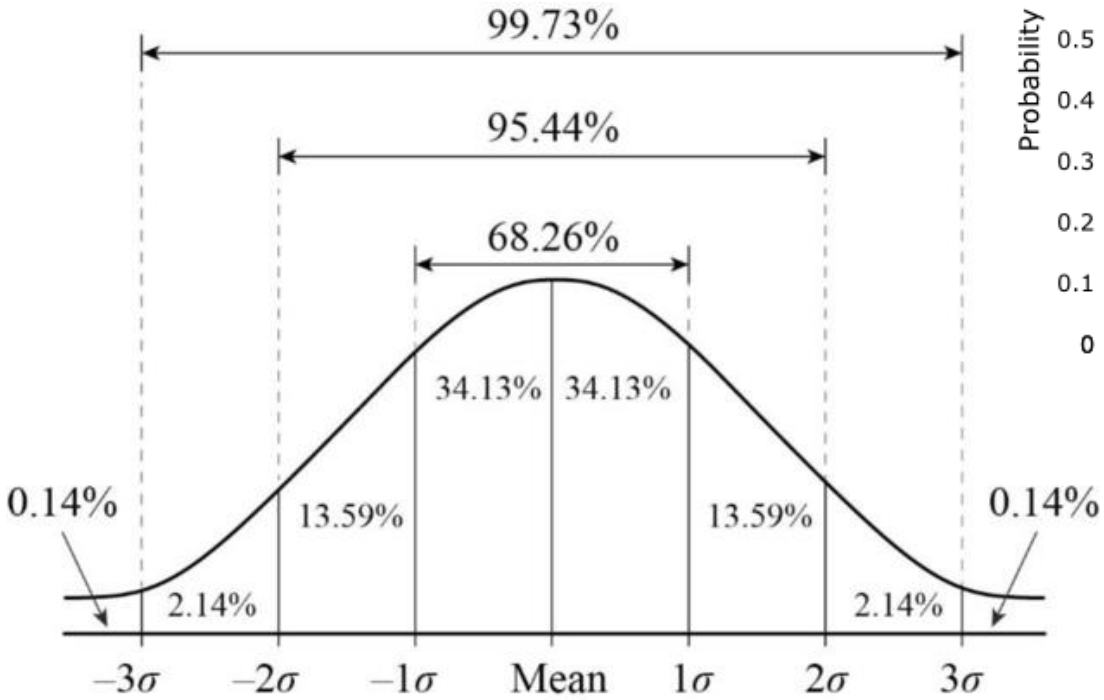
- Случайная величина X имеет нормальное распределение вероятностей с параметрами μ и σ^2:
  $\boldsymbol{X \sim N(\mu, \sigma^2)}$,
  если функция плотности распределения вероятности $f(x)$ имеет вид:

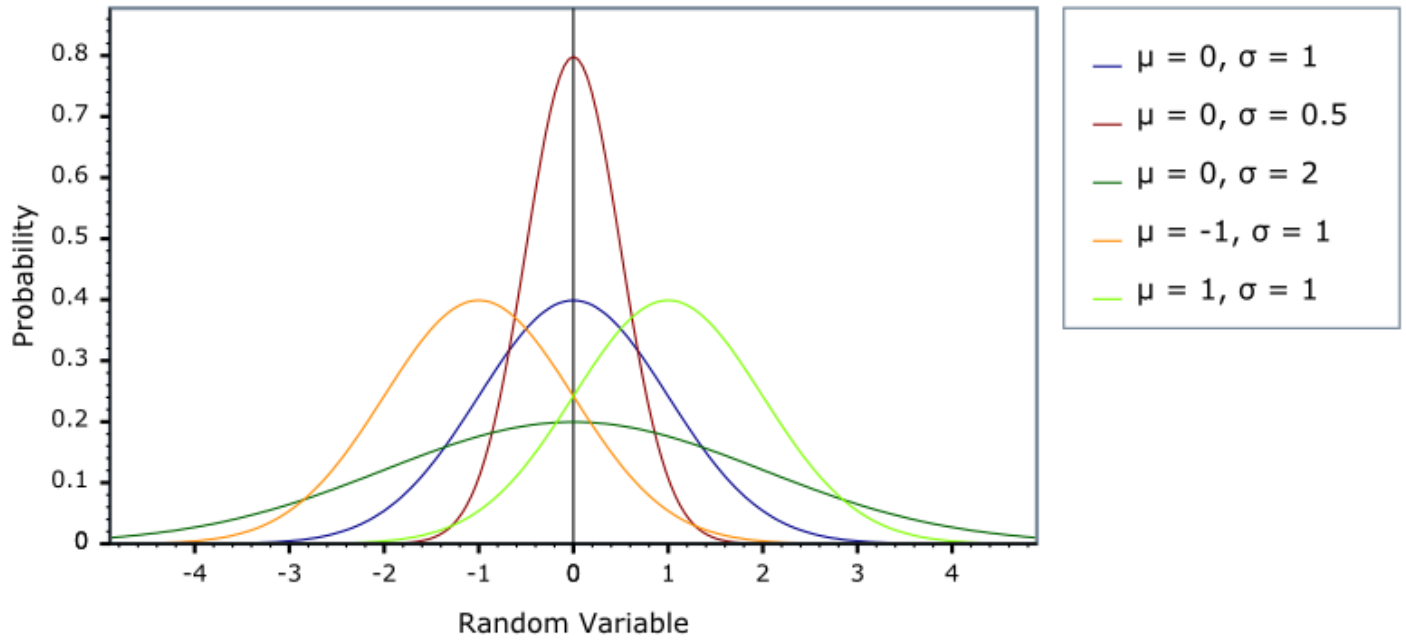
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- Стандартное нормальное распределение:

  $\mu = 0, \sigma = 1$

Свойства


99.73%
95.44%
68.26%
34.13%
34.13%
0.14%
13.59%
13.59%
0.14%
2.14%
2.14%
−3σ
−2σ
−1σ
Mean
1σ
2σ
3σ


Probability
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Random Variable
μ = 0, σ = 1
μ = 0, σ = 0.5
μ = 0, σ = 2
μ = -1, σ = 1
μ = 1, σ = 1

Проверка гипотез. Основные понятия

- **Статистическая гипотеза** – это предположение о распределении вероятностей некоторого события, которое проверяется по имеющимся случайным данным.
- **Нулевая гипотеза H_0** - предположение по умолчанию о свойствах генеральной совокупности (population) на основе данных случайной выборки (sample).
- **Альтернативная (исследовательская) гипотеза H_1** - предположение о свойствах ГС, исключающее нулевую гипотезу.
- **Проверка гипотезы** - достаточно ли аргументов, чтобы отвергнуть **H_0**.
- Можно принять только **одно из двух решений**:
  - отвергнуть нулевую гипотезу H_0 и принять альтернативную гипотезу H_1
  - остаться в рамках нулевой гипотезы H_0

Суть проверки гипотез

Естественные науки:

- Гипотезу отвергают, если происходит то, что при ее справедливости невозможно.

Статистика:

- Гипотезу отвергают, если происходит то, что при ее справедливости практически невозможно = очень маловероятно.

Значимость и проверка гипотез

- **Уровень значимости α** – пороговое значение вероятности события
- Событие А считается практически невозможным, если его вероятность меньше принятого уровня значимости: $P(A) < \alpha$
- Если произошло такое событие А, то гипотеза отвергается на уровне значимости α
- Критическое событие (критерий) в идеале:
  - $P(A) \sim 0$ при $\mathbf{H_0}$
  - $P(A) \sim 1$ при $\mathbf{H_1}$

Алгоритм тестирования (проверки) гипотезы

Выбрать критерий для проверки.

Задать уровень значимости α.

Вычислить критериальное значение для проверки $W(\alpha)$ (это случайная величина!).

Вычислить p - вероятность критериального значения при верной H_0 (p-value).

Если $p > \alpha$,

то H_0 не отклоняется,

иначе H_0 отклоняется и принимается H_1.

Возможные ситуации при принятии решения
(верный-истинный, неверный-ложный)

		Истинность гипотезы H_0	
		true	false
Результат принятия решения при принятии или отклонении гипотезы H_0	Не отклонять	Верная H_0 верно принята	Неверная H_0 неверно принята (Ошибка второго рода)
	Отклонить	Верная H_0 неверно отвергнута (Ошибка первого рода)	Неверная H_0 верно отвергнута

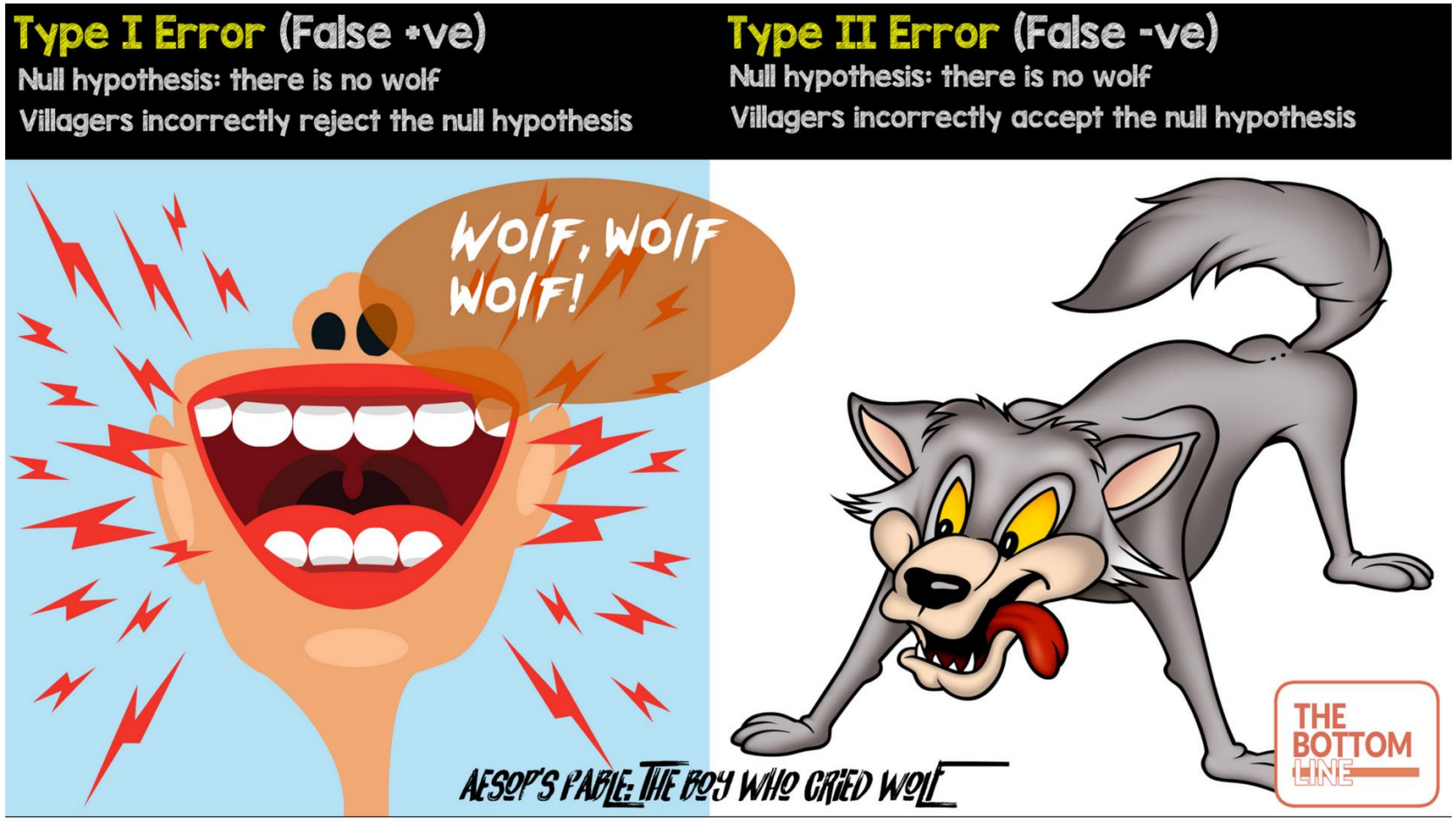
Ошибки первого рода

- **Можно ошибочно отклонить истинную H_0.**
  - Это ошибка первого рода – ложная тревога, **ложноположительное решение (false positive, FP).**
  - Распределение выборки соответствует гипотезе H_0, но она неверно отвергнута статистическим критерием.
  - Уровень значимости α – вероятность отклонить H_0, если на самом деле она верна.

Ошибки второго рода

- **Можно ошибочно оставить ложную H_0.**
  - Это ошибка второго рода – пропуск события, **ложноотрицательное решение (false negative, FN)**.
  - Распределение выборки соответствует гипотезе H_1, но она неверно отвергнута статистическим критерием.
  - Величина β - вероятность оставить ложную H_0.
- Мощность критерия – величина 1- β: вероятность отклонения ложной H_0.
- Чем выше мощность критерия, тем меньше вероятность совершить ошибку второго рода

Пример из сказки



Возможные ситуации при принятии решения

Верить ли мальчику, кричащему «Волки!
Нулевая гипотеза – волков нет.

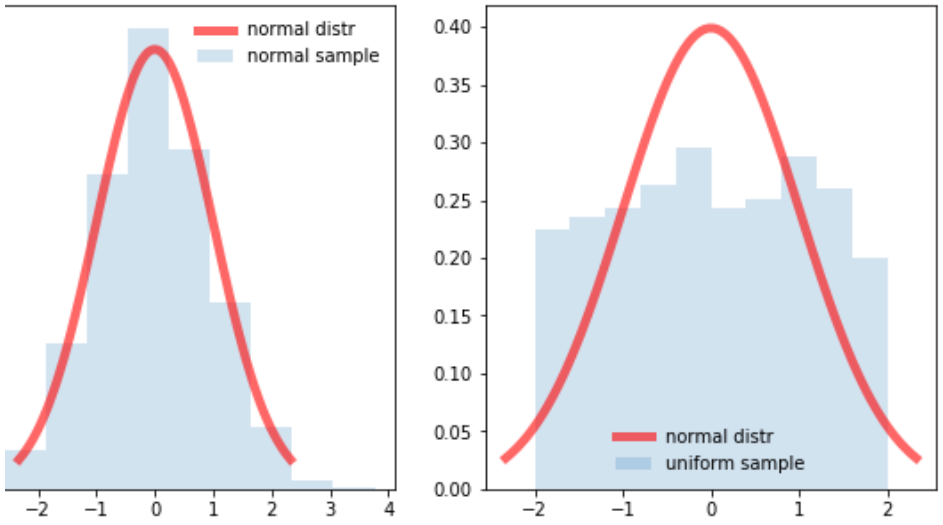
		Истинность гипотезы H_0	
		true	false
Результат принятия решения при принятии или отклонении гипотезы H_0	Не отклонять	~~Верная H_0 верно принята~~	Неверная H_0 неверно принята (Ошибка второго рода)
	Отклонить	Верная H_0 неверно отвергнута (Ошибка первого рода)	~~Неверная H_0 верно отвергнута~~

Формулировка гипотез на примере анализа распределения случайной величины

- Задача – проверить, имеют ли данные распределение Гаусса (нормальное).
- Нулевая гипотеза H_0 - данные выборки позволяют заключить, что распределение случайной величины соответствует закону Гаусса (нормальное) .
- Альтернативная гипотеза H_1 - по данным выборки нельзя сделать вывод о том, что случайная величина распределена по закону Гаусса.

Тестирование гипотезы

- Нулевая гипотеза H_0 - предположение, **противоположное** тому, что тестируется.
- Альтернативная гипотеза H_1 - предположение, которое тестируется.
- Если отвергнута H_0, то принимается H_1.
- Почему увеличение уровня значимости увеличивает риск ошибок 1-ого рода и снижает риск ошибок 2-ого рода?
  А при уменьшении уровня значимости все происходит наоборот?



Исследование зависимостей

$$Cov_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{(n-1)} = \frac{\Sigma xy - n\bar{x}\bar{y}}{(n-1)}$$

- **Параметрические методы** (нормальное распределение значений числовых переменных)
  - Коэффициент (матрица) корреляции Пирсона – **нормированная** числовая мера некоторого типа статистической связи между двумя переменными.
  - Коэффициент (матрица) ковариации - **ненормированная** числовая мера **линейной** связи между двумя переменными.

$$\sigma^2 = \sum \frac{(X-\mu)^2}{N}$$

$$Cor(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Гипотезы

Предположения

- Наблюдения в каждой выборке независимы и одинаково распределены.

- H_0: переменные независимы.
- H_1: существует зависимость между переменными.

Исследование зависимостей

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

Порядковая шкала:

Коэффициент (матрица) корреляции Спирмена – **нормированная** числовая мера некоторого типа статистической связи между двумя переменными.

di - разность между двумя рангами

n – число наблюдений

Непараметрические методы (категориальные шкалы, особые распределения)

Гипотезы

Предположения

- Наблюдения в каждой выборке независимы и одинаково распределены.
- Наблюдения могут быть ранжированы

- H_0: переменные независимы.
- H_1: существует зависимость между переменными.

Исследование зависимостей

Непараметрические методы (категориальные шкалы, особые распределения)

Номинальная шкала:

Тест хи-квадрат

Проверяет, являются ли две категориальные переменные связанными или независимыми.

Оценивает статистическую значимость различий двух или нескольких относительных показателей (частот).

Гипотезы

Предположения

- Наблюдения, использованные при расчете таблицы сопряженности, являются независимыми.
- Не менее N (20-25) значений в каждой ячейке таблицы сопряженности.

- H_0: две выборки независимы.
- H_1: существует зависимость между выборками.

Таблица сопряженности

Пол	**Увлечение**
Male	Art
Female	Math
Male	Science
Male	Math
…	…

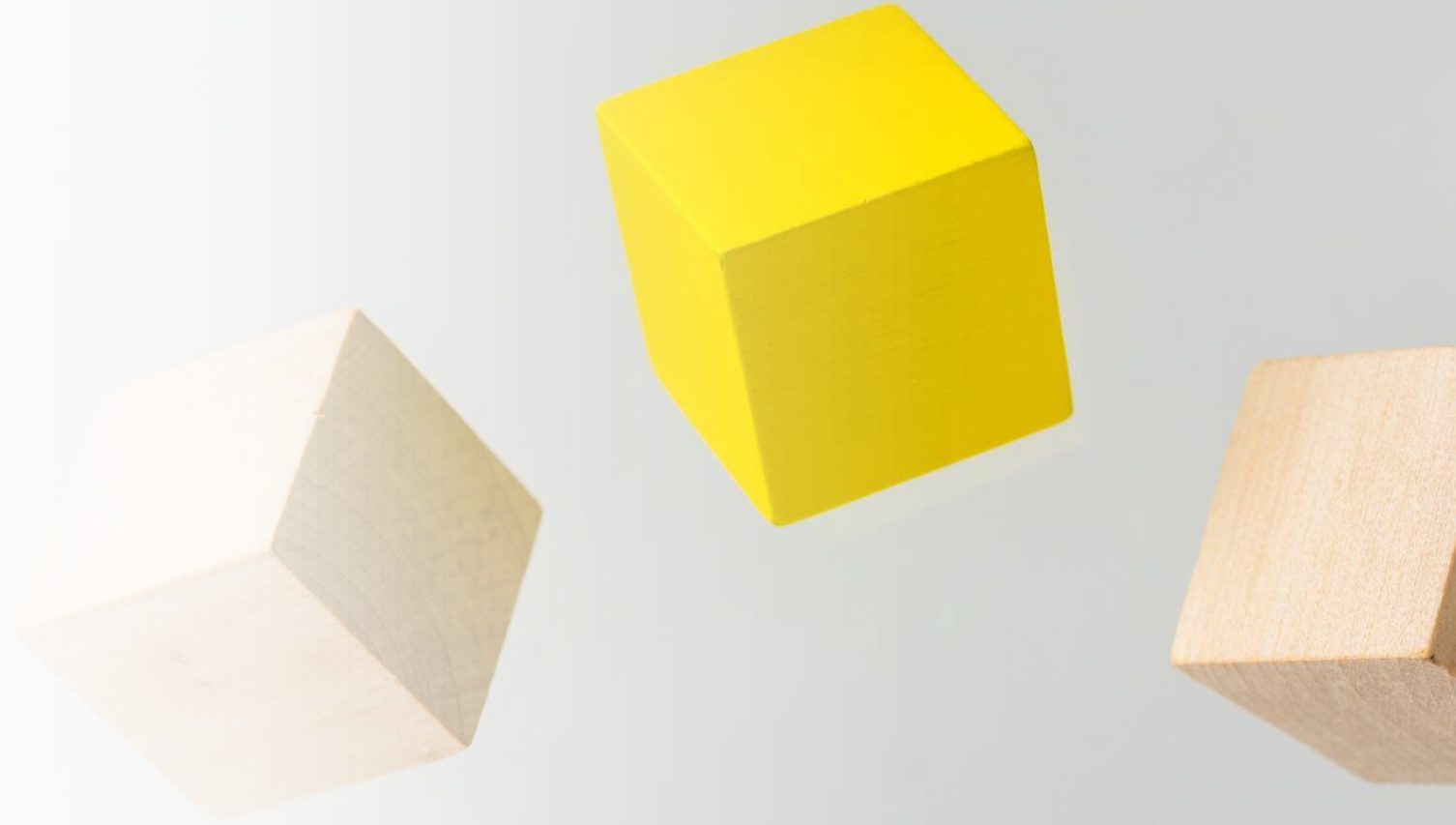
Таблица исходных данных (130x2)

	Math	Art	Science
Male	**20**	**30**	**15**
Female	**20**	**15**	**30**

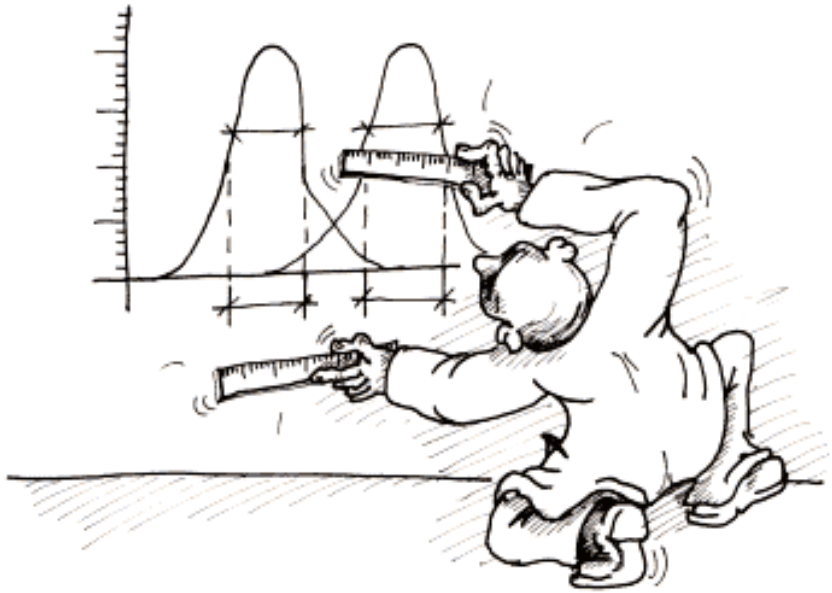
Таблица частот

Гипотезы. Сравнение выборок

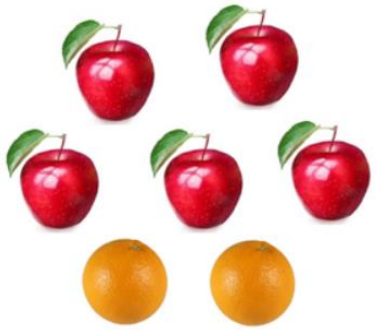
https://docs.scipy.org/doc/scipy/reference/stats.html

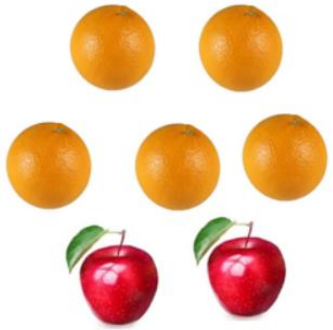


Сравнение наборов данных



- По признакам
- По объему
- **По распределению значений признаков**





Критерий Стьюдента (t-критерий) (У. Госсет)

- Проверяет, значительно ли отличаются средние для двух **независимых** выборок из набора (наборов) числовых данных определенного признака.

- Требования к данным
  - Данные в каждой выборке нормально распределены.
  - Распределения имеют одинаковую (различную) дисперсию.
- Гипотезы
  - H_0: средние равны.
  - H_1: средние неравны.

Пример

- Даны **независимые** выборки из одной (разных) совокупности с результатами определенного ЕГЭ двух школ.
- Тест определяет, сильно ли различается среднее (ожидаемое) значение по выборкам.
- Если p >0,05 (или другого заданного значения), то мы не можем отклонить нулевую гипотезу об идентичных средних оценках.
- Иначе отвергаем нулевую гипотезу равных средних значений.

Критерий Стьюдента (t-критерий)

- Проверяет, значительно ли отличаются средние для двух **зависимых** наборов числовых данных.

- Требования к данным
  - Данные в каждом наборе нормально распределены.
  - Распределения имеют одинаковую дисперсию.
  - Есть попарное соответствие значений в наборах.
- Гипотезы
  - H_0: средние равны.
  - H_1: средние неравны.

Пример

- Даны **зависимые** выборки с оценками **одного и того же набора учащихся** на **разных** экзаменах.
- Тест определяет, сильно ли различается средний балл по выборкам (например, экзаменам).
- Если p >0,05 (или 0,1), то мы не можем отклонить нулевую гипотезу об идентичных средних оценках.
- Иначе отвергаем нулевую гипотезу равных средних значений.