

Юстина Иванова

Программист, data scientist

Основные статистические тесты и
проверка гипотез.

Спикер



Юстина Иванова,

- PhD студент в Университете Больцано
- Data scientist по
Компьютерному зрению
в компании ОЦРВ
- Выпускница МГТУ им. Баумана
- Магистр по Artificial Intelligence
В University of Southampton

Повторение

Доверительный интервал — интервал, в котором лежит $p\%$ данных.

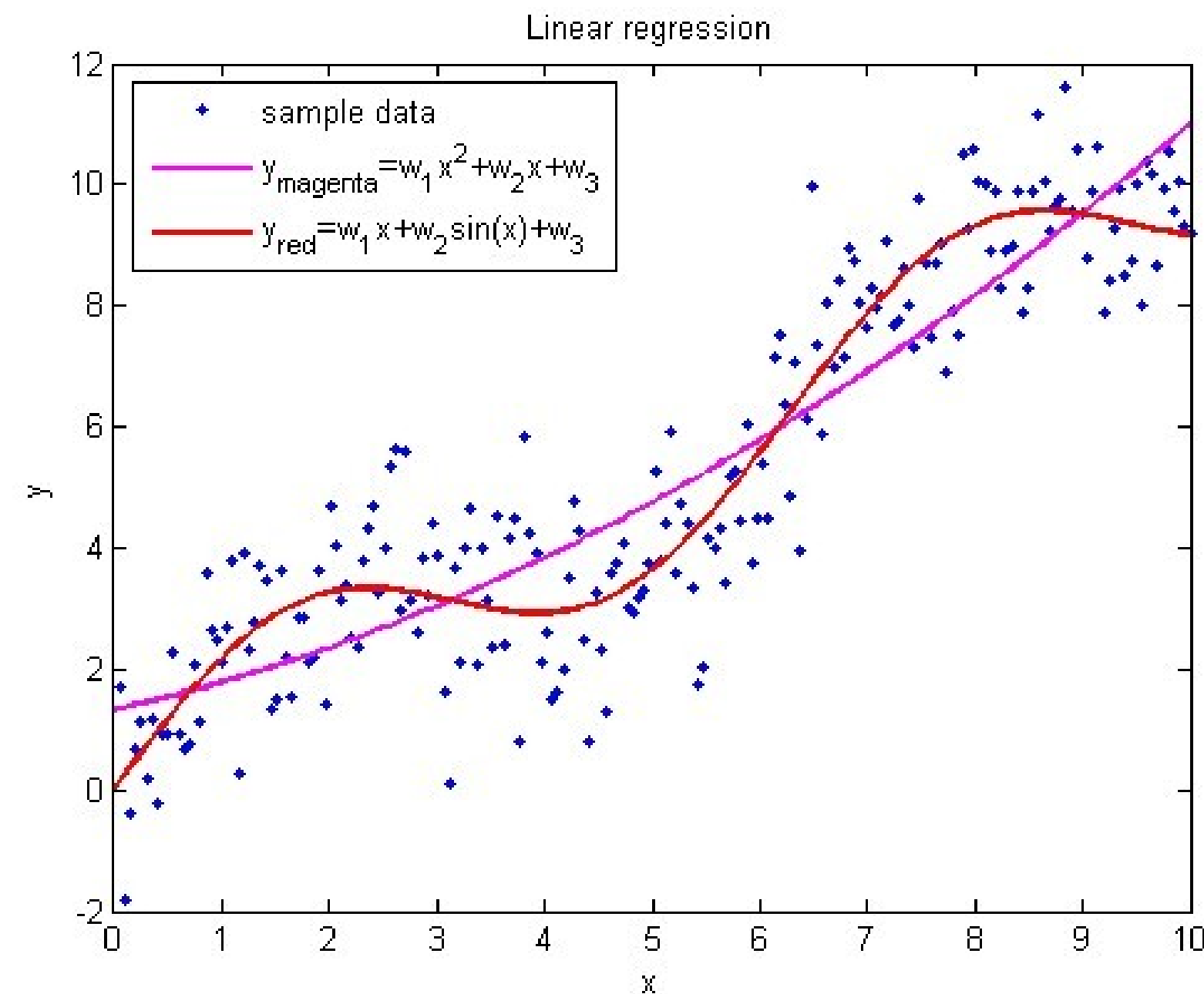
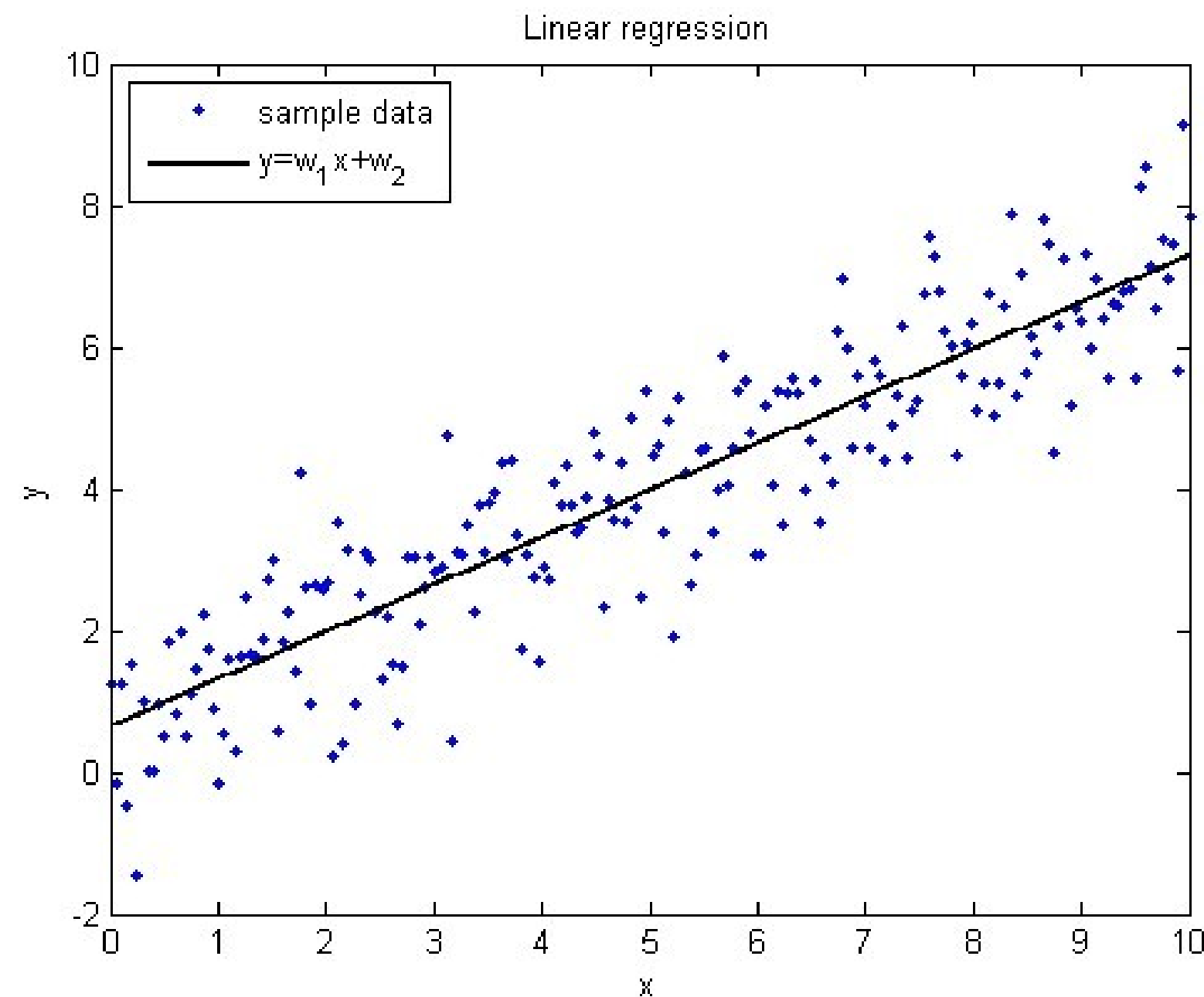
Правило трех сигм.

Виды распределений: дискретное и непрерывное.

Скалярное произведение векторов и проецирование вектора.

Линейная регрессия и классификационный анализ.

Линейная регрессия



Для заданного пространства данных найти уравнение прямой (или кривой), минимизирующую сумму расстояний от точек до нее.

Логистическая регрессия.

Задача логистической регрессии – определить вероятность принадлежности к классу.

Построена на основе линейной функции.

$$h(x) = \theta^T x$$

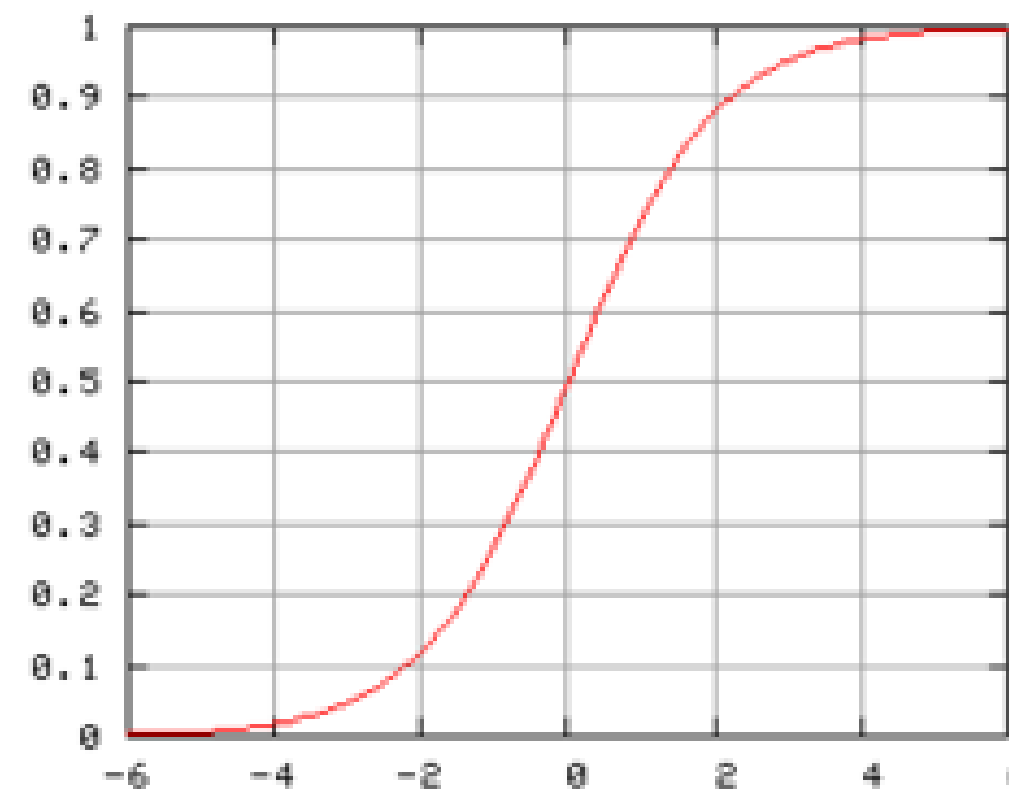
К линейной функции применяется функция активации:

$$h(x) = \sigma(\theta^T x)$$

Функция активации:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

Сигмоида.



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Производная сигмоиды:

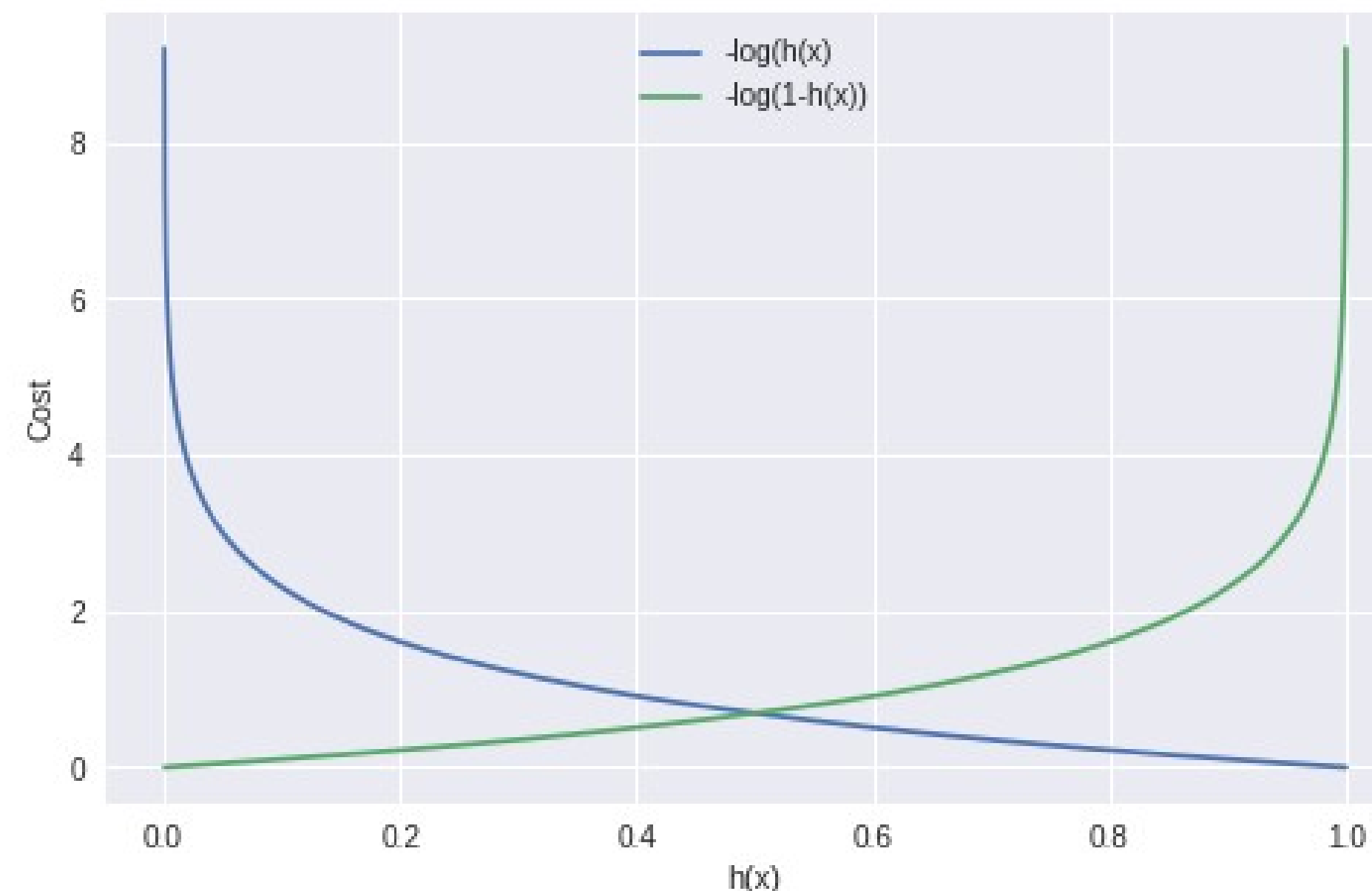
$$\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$$

Функция ошибки в логистической регрессии.

Модель ищет параметры, которые минимизируют функцию ошибки:

$$cost = \begin{cases} -\log(h(x)), & \text{if } y = 1 \\ -\log(1 - h(x)), & \text{if } y = 0 \end{cases}$$

Чем выше вероятность определения класса 1 при верном классе 0, тем выше стоимость ошибки.



Функция ошибки в логистической регрессии.

Общий вид функции ошибки для модели:

$$\text{cost}(h(x), y) = -y \cdot \log(h(x)) - (1 - y) \log(1 - h(x))$$

Ошибка для всех данных датасета:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^i \log(h(x^i)) + (1 - y^i) \log(1 - h(x^i))]$$

Где m – количество элементов.

Градиентный спуск.

Будем искать минимум функции относительно параметров:

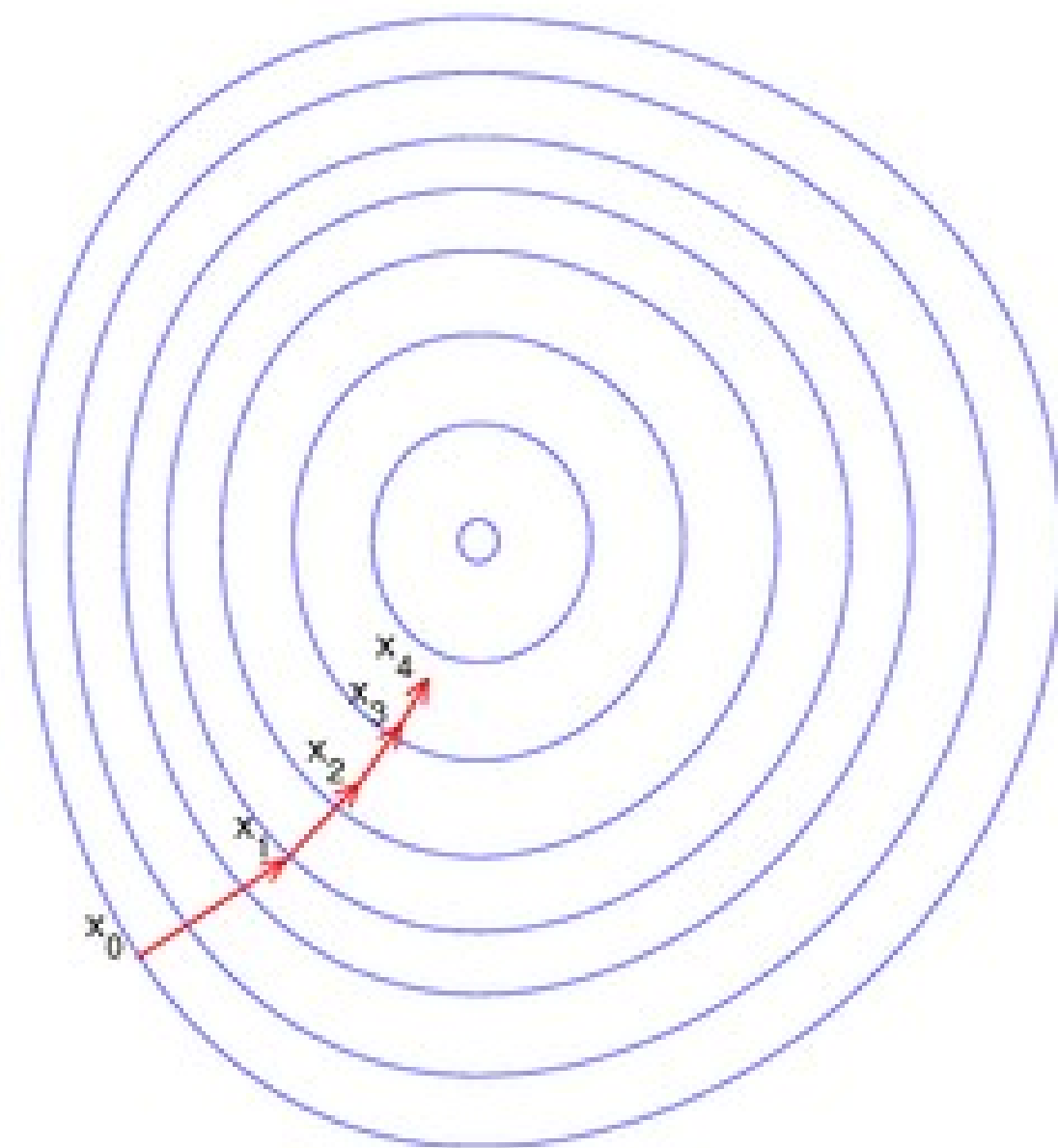
$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^m (h(x^i) - y^i) x_j^i$$

Для поиска минимума используется градиентный спуск.

Это метод нахождения локального минимума (максимума) функции с помощью движения вдоль градиента.

Градиентный спуск.

Основная идея – идти в направлении наискорейшего спуска, оно задается антиградиентом.



$$\vec{x}^{[j+1]} = \vec{x}^{[j]} - \lambda^{[j]} \nabla F(\vec{x}^{[j]})$$

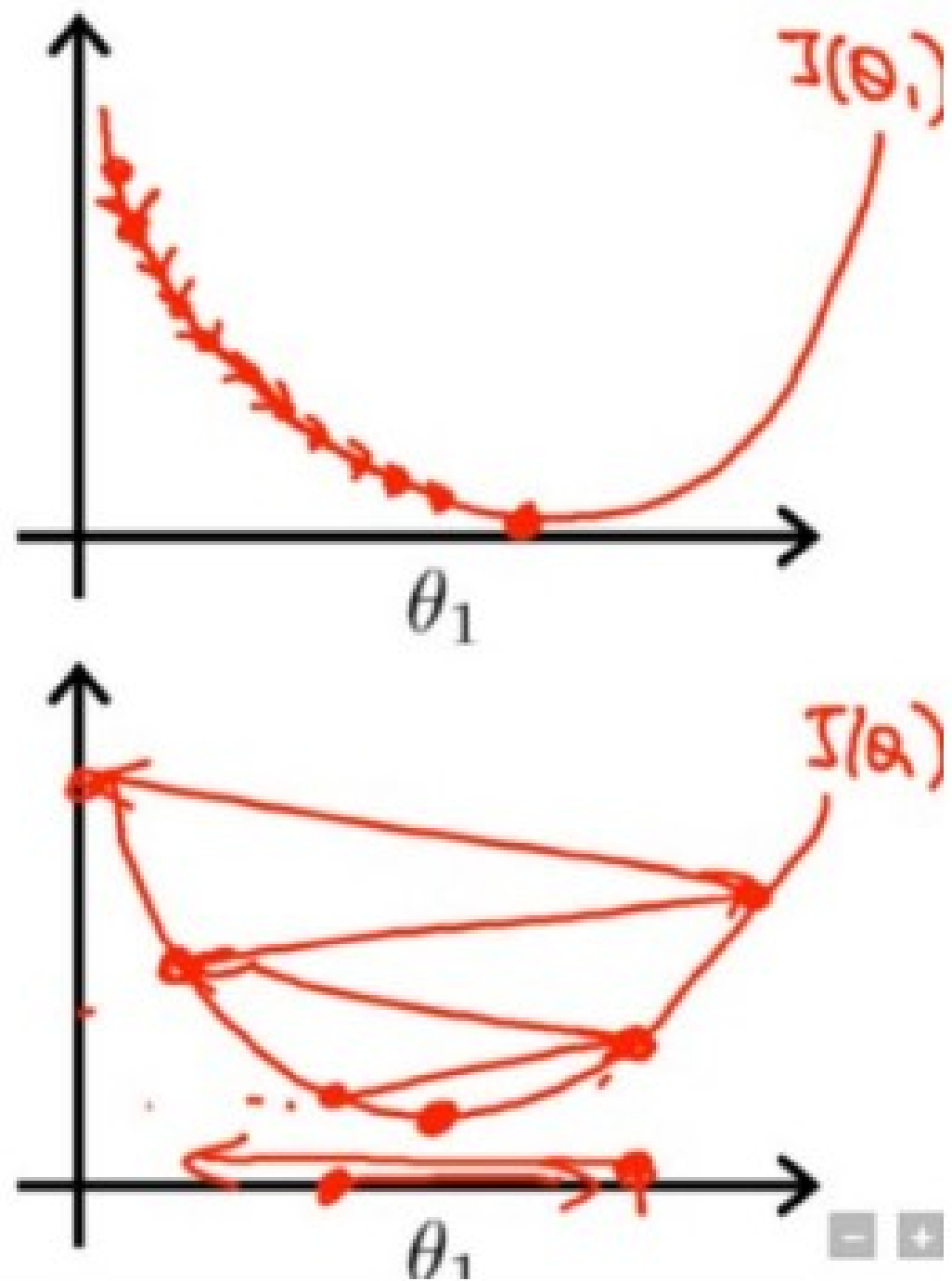
Где $\lambda^{[j]}$ - скорость градиентного спуска,
 F - целевая функция

Скорость градиентного спуска

Проблемы градиентного спуска.

$$\vec{x}^{[j+1]} = \vec{x}^{[j]} - \lambda^{[j]} \nabla F(\vec{x}^{[j]})$$

Где $\lambda^{[j]}$ - скорость градиентного спуска,
F - целевая функция



Source: Andrew Ng's course on Coursera

Квантиль

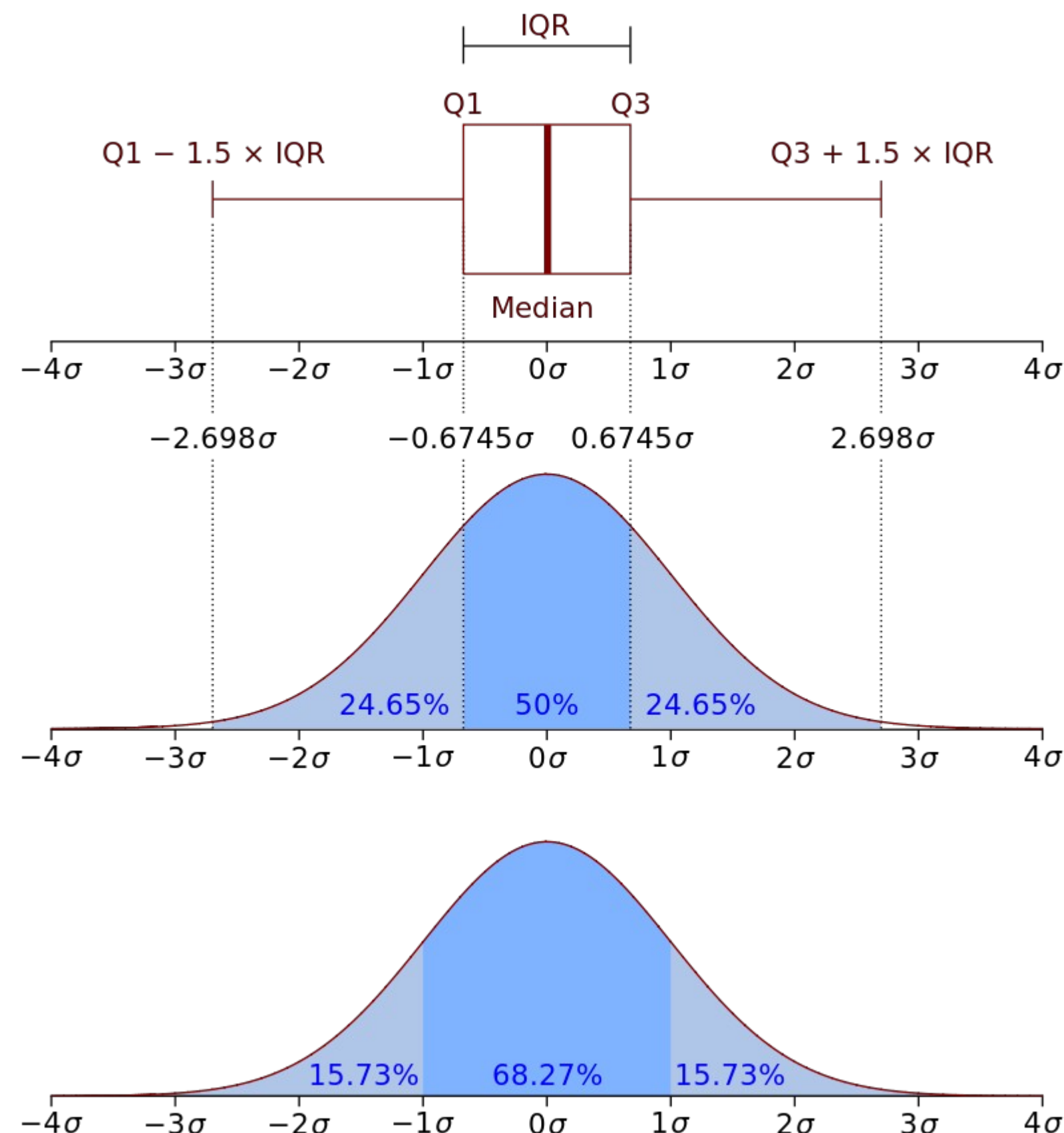
0,25-квантиль называется первым (или нижним) квартилем;

0,5-квантиль называется медианой или вторым квартилем;

0,75-квантиль называется третьим (или верхним) квартилем.

Интерквартильным размахом называется разность между третьим и первым квартилями, то есть $x_{\{0,75\}} - x_{\{0,25\}}$. Интерквартильный размах является характеристикой разброса распределения величины и является аналогом **дисперсии**. Вместе, медиана и интерквартильный размах могут быть использованы вместо **математического ожидания** и дисперсии в случае распределений с большими выбросами, либо при невозможности вычисления последних.

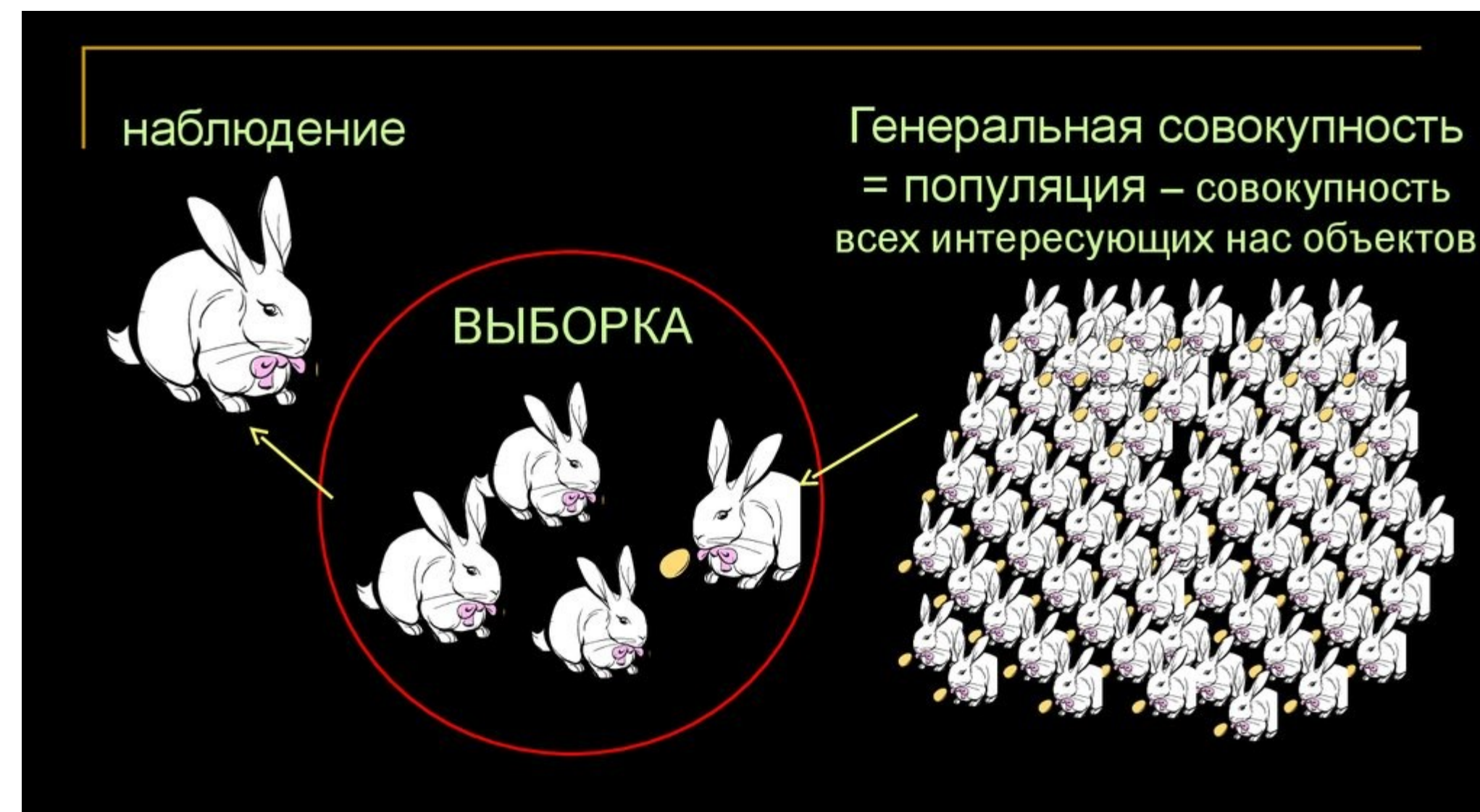
Значение квантиля



Предоставляют важную информацию о структуре **вариационного** (колонку таблицы) ряда признака. Вместе с медианой они делят вариационный ряд на 4 равные части. Квартилей две, их обозначают символами Q, верхняя и нижняя квартиль. 25% значений меньше, чем нижняя квартиль, 75% значений меньше, чем верхняя квартиль.

Генеральная совокупность и выборка.

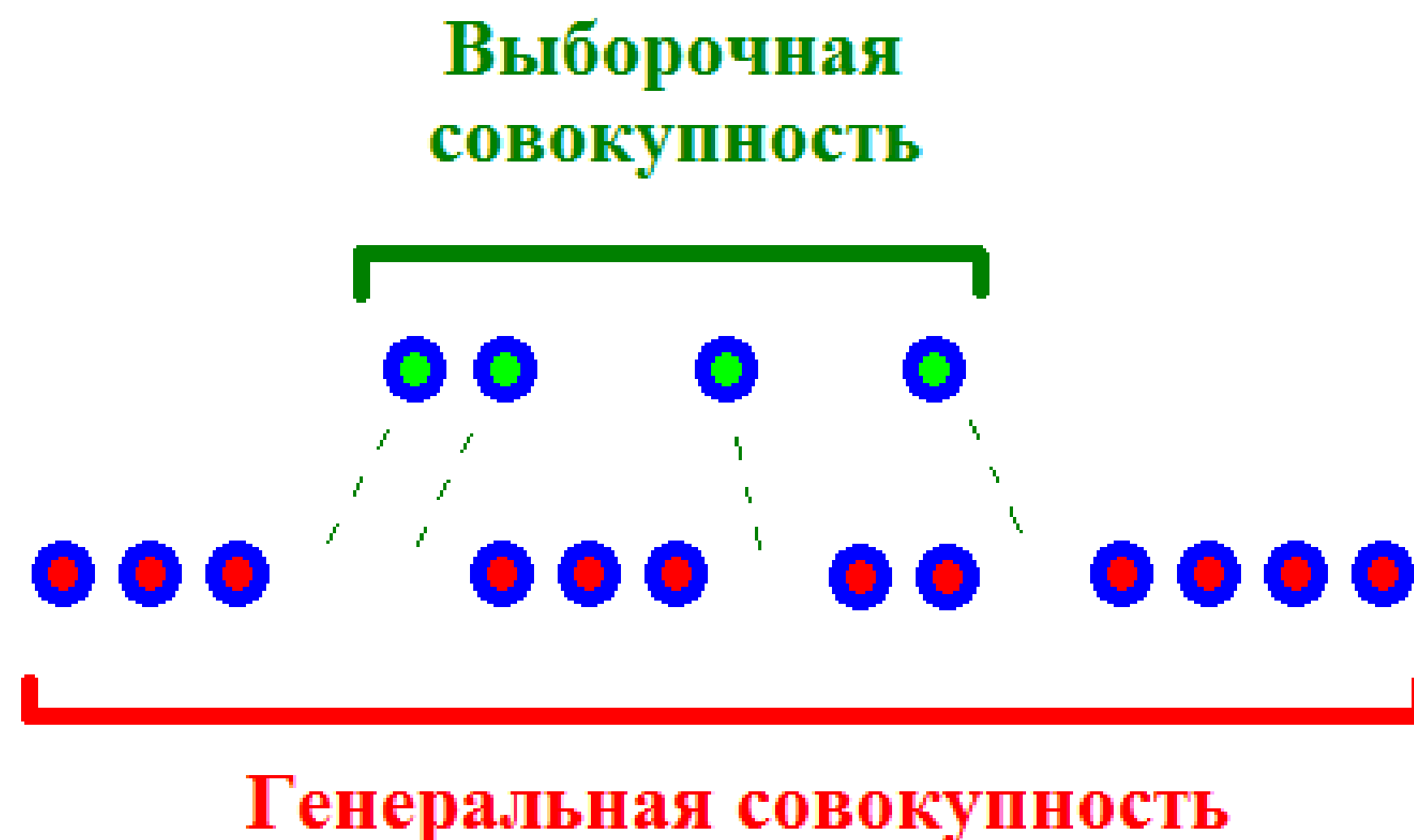
Генеральная совокупность — множество всех объектов, обладающих изучаемым признаком.



На основе свойств выборки делаем заключение о свойствах генеральной совокупности.

Статистические гипотезы о данных

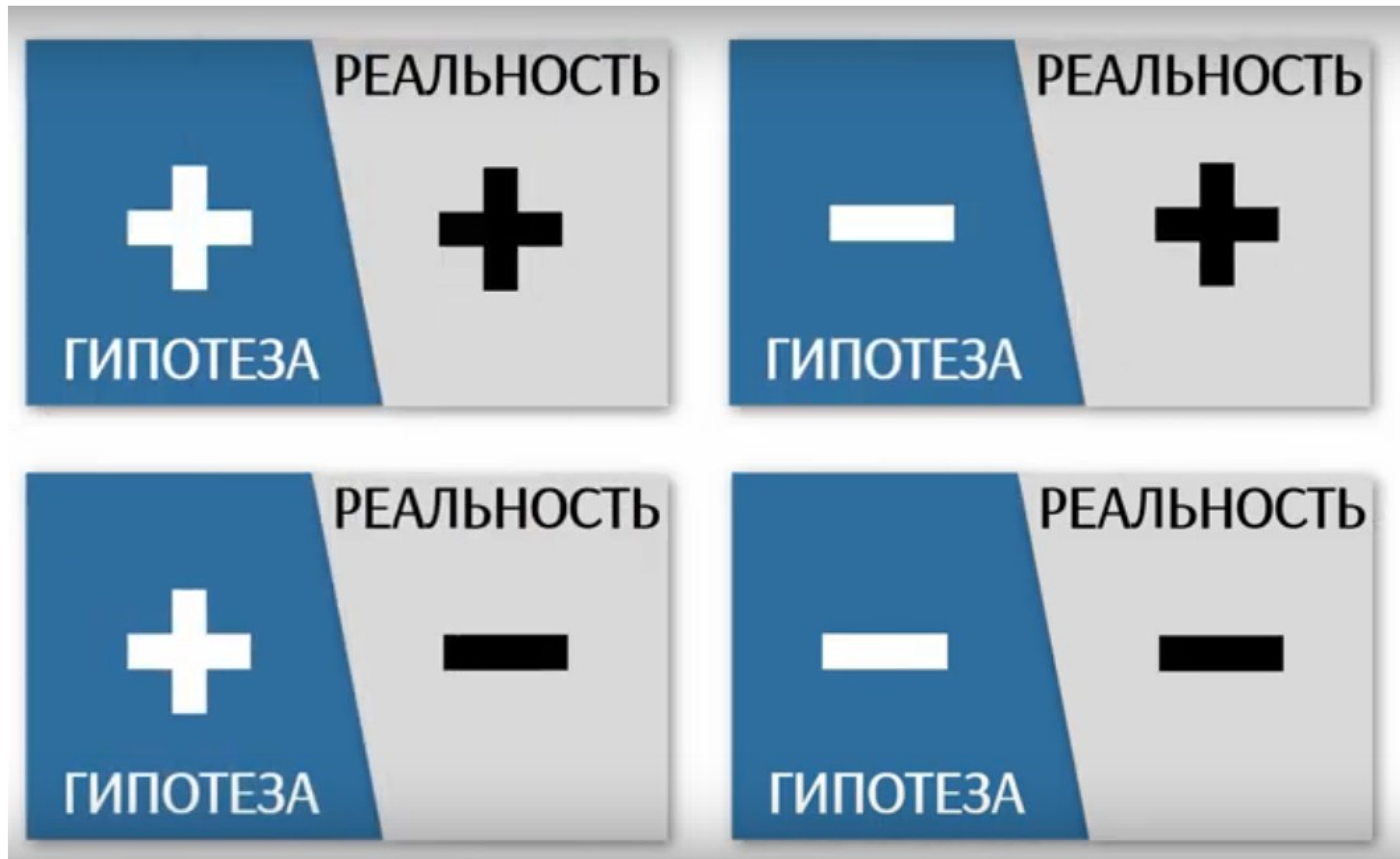
Выборочная совокупность — множество всех объектов, отобранных случайно из генеральной совокупности для изучения.



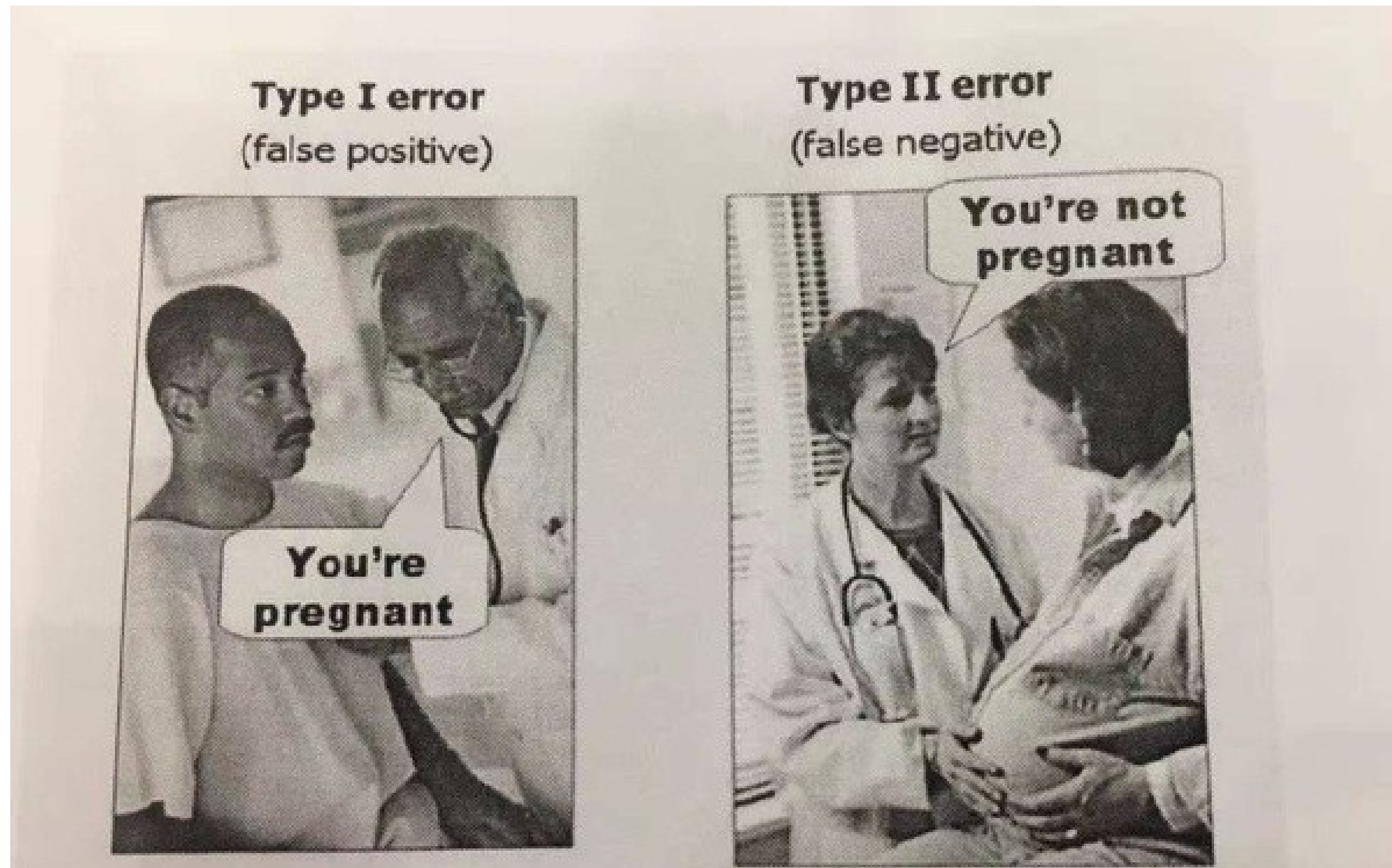
Нулевая гипотеза (H_0) — гипотеза о сходстве

Альтернативная гипотеза, конкурирующая, (H_1) — гипотеза о различиях

Статистические гипотезы о данных



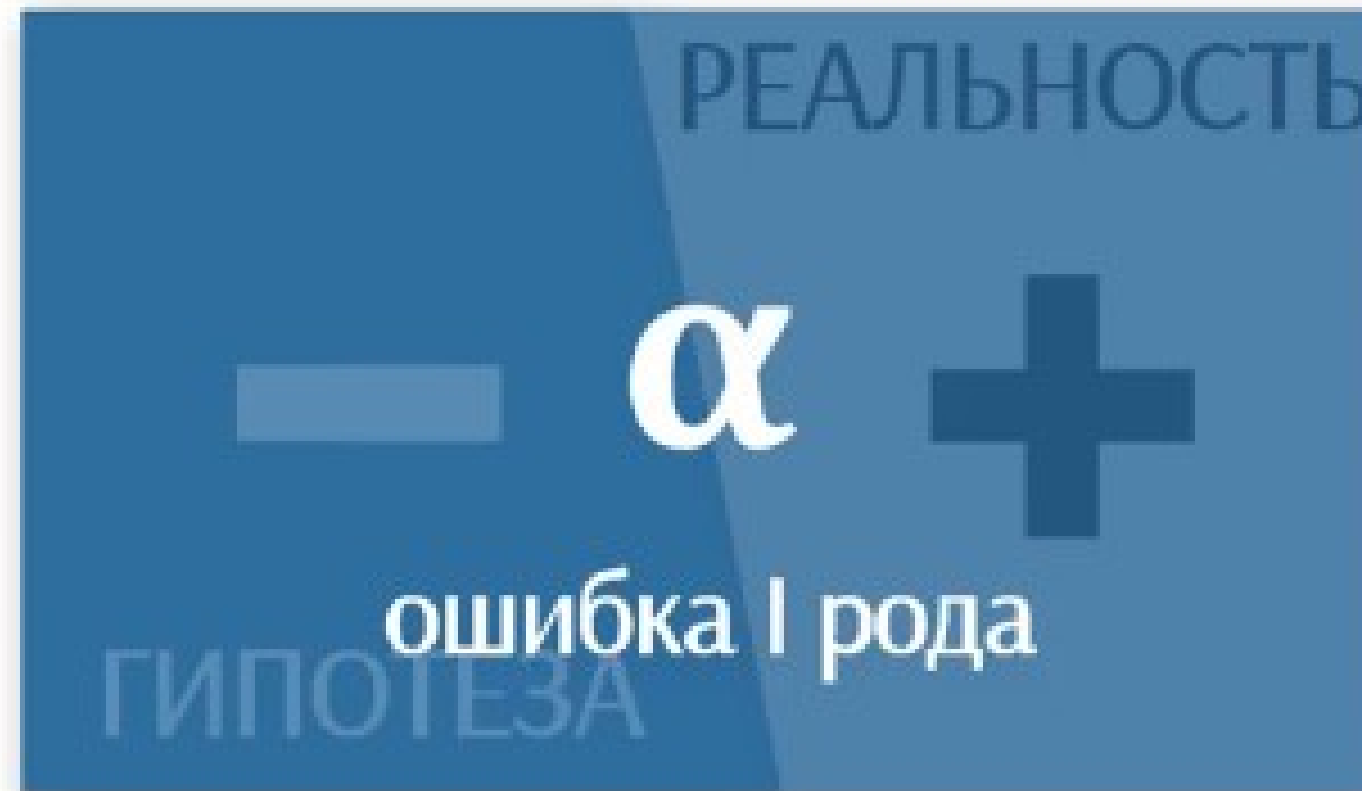
Пример: тест на наличие беременности.



Пример: тест на наличие болезни.

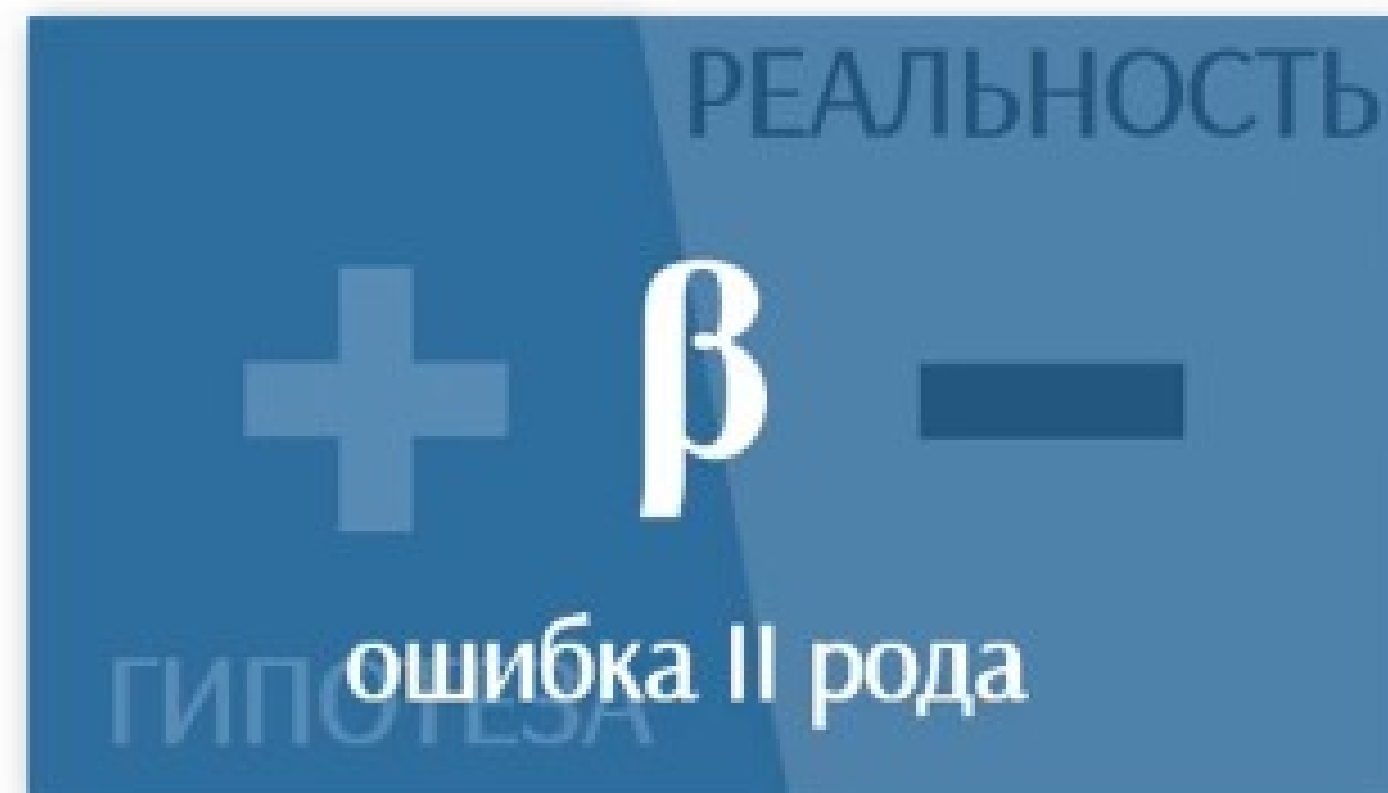
	Disease present	Disease absent
Positive	a True positive	b False positive
Negative	c False negative	d True negative

Статистические гипотезы о данных



Ошибка 1 рода:

Вероятность отвергнуть гипотезу,
Но в действительности она верна
Alpha — вероятность ошибки.
Критически значимый уровень
alpha = 0.05



Ошибка 2 рода:

Вероятность принять гипотезу,
Но в действительности она неверна
beta — вероятность ошибки.
Мощность исследования = 1-beta.

Статистическая значимость

СТАТИСТИЧЕСКАЯ ЗНАЧИМОСТЬ (ЗНАЧЕНИЕ P)
– РАСЧЕТНАЯ ВЕРОЯТНОСТЬ ОШИБКИ
ПЕРВОГО РОДА, КОТОРАЯ РАССЧИТЫВАЕТСЯ С
ПОМОЩЬЮ РАЗЛИЧНЫХ СТАТИСТИЧЕСКИХ
КРИТЕРИЕВ



$$P < 0,05$$

Подсчитывается с помощью разных критериев.

А/В тесты

А/В тестирование — это мощный маркетинговый инструмент для повышения эффективности работы вашего интернет-ресурса.

Ниже на картинках приведены примеры распределения значений показателя в сегментах.



Пример А/В теста: Wallmonkeys

Компания WallMonkeys решила оптимизировать веб-сайт на клики и конверсию.

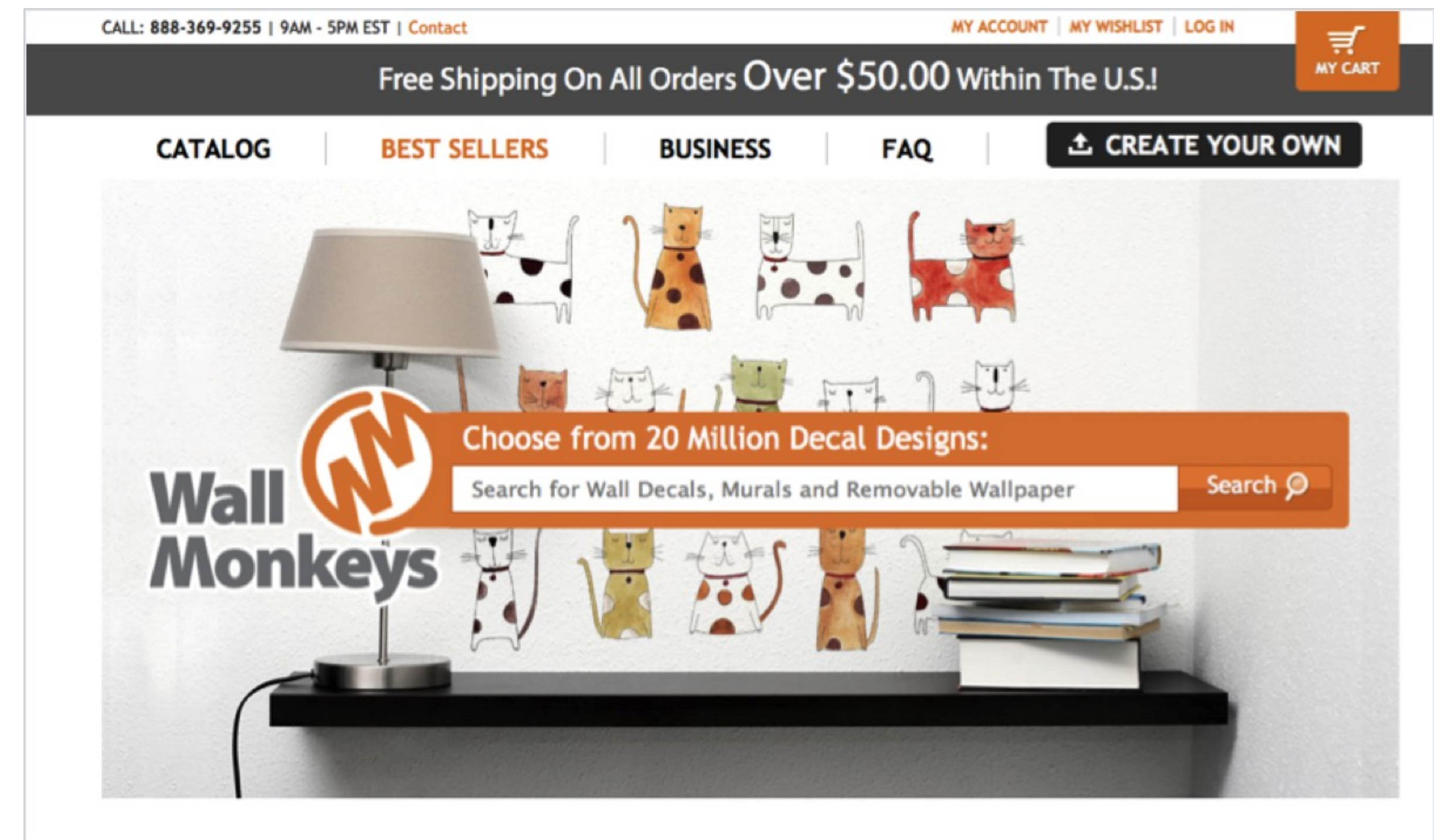


Пример А/В теста: WallMonkeys

1 тест: 27% кликов.



2 тест: 550% кликов



Виды статистических критериев

Критерии согласия -

проверка на согласие подразумевает проверку предположения о том, что исследуемая случайная величина подчиняется предполагаемому закону.

Параметрические критерии -

группа статистических критериев, которые включают в расчет параметры вероятностного распределения признака (средние и дисперсии).

Непараметрические критерии -

группа статистических критериев, которые не включают в расчёт параметры вероятностного распределения и основаны на оперировании частотами или рангами.

Параметрическая — непараметрическая гипотеза

Параметрические критерии - группа статистических критериев, которые включают в расчет параметры вероятностного распределения признака (средние и дисперсии).

t-критерий Стьюдента

Критерий Фишера

Критерий отношения правдоподобия

Критерий Романовского

Параметрическая — непараметрическая гипотеза

Непараметрические критерии

Группа статистических критериев, которые не включают в расчёт параметры вероятностного распределения и основаны на оперировании частотами или рангами.

Q-критерий Розенбаума

U-критерий Манна — Уитни

Критерий Уилкоксона

Критерий Пирсона

Критерий Колмогорова — Смирнова

Распределение Стьюдента

Мы хотим сгенерировать нормальное распределение, но по некоторым причинам не можем вычислить среднеквадратичное отклонение (например, выборка маленькая). Мы можем найти выборочное среднее и выборочную дисперсию по выборке.

Пусть x_1, \dots, x_n — выборка размером n

Выборочное среднее $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

Выборочная дисперсия $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

Распределение Стьюдента

Случайная величина t имеет распределение Стьюдента с $n-1$ степенями свободы, где n — размер выборки.

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Данный критерий был разработан Уильямом Госсетом для оценки качества пива в компании Гиннесс. В связи с обязательствами перед компанией по неразглашению коммерческой тайны (руководство Гиннеса считало таковой использование статистического аппарата в своей работе), статья Госсета вышла в 1908 году в журнале «Биометрика» под псевдонимом «Student» (Студент).

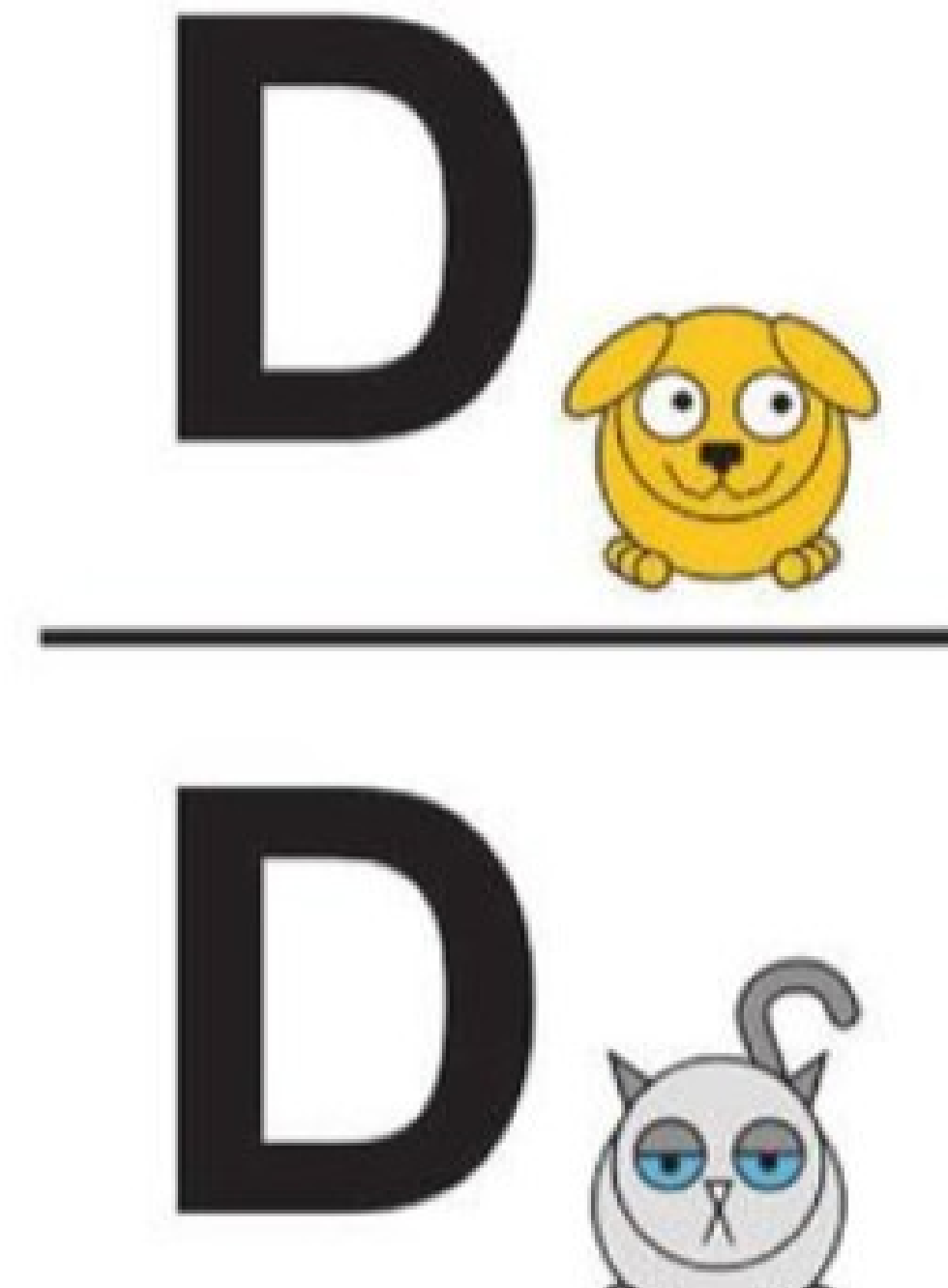
Проверка гипотез методом Стьюдента



Тесты для проверки гипотез: тест Фишера

Метод Стьюдента чувствителен к выбросам => параметрический метод: мы можем посмотреть, являются ли песики более разнообразными по размеру, чем котики, или же нет. Для этого мы можем воспользоваться F-критерием равенства дисперсий Фишера, который укажет нам, насколько различаются между собой эти показатели.

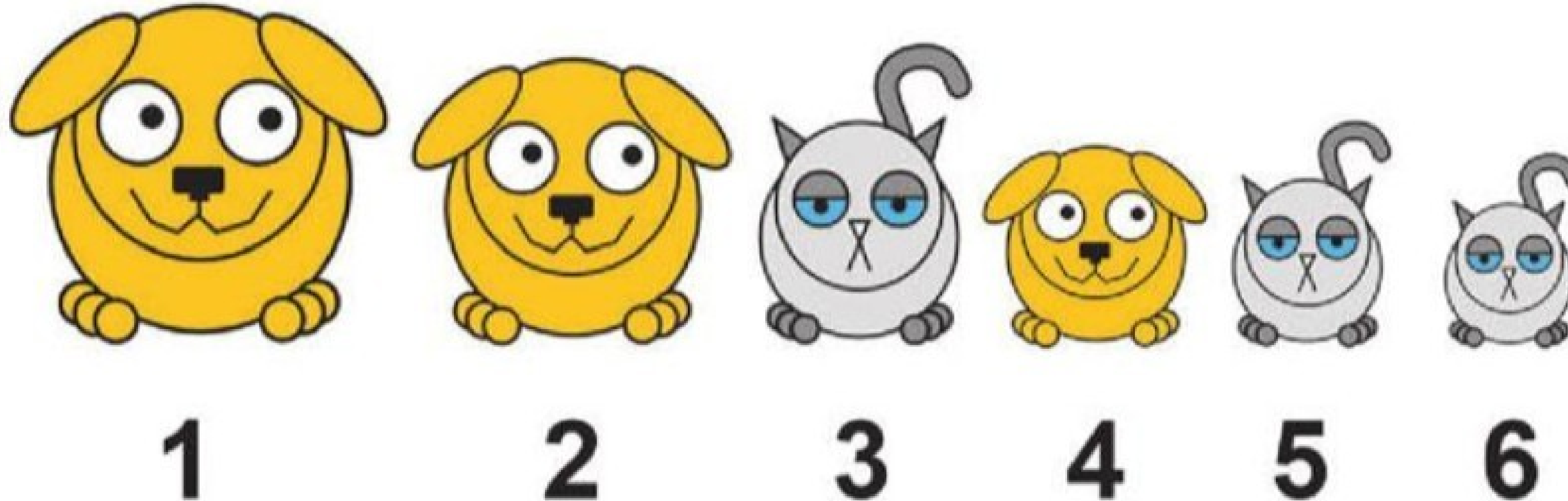
F
Фишера



Проверка гипотез методом Манна-Уитни

Метод Стьюдента чувствителен к выбросам \Rightarrow параметрический метод:

Чтобы рассчитать критерий Манна-Уитни, необходимо выстроить всех песиков и котиков в один ряд, от самого мелкого к самому крупному, и назначить им ранги. Самому большому зверьку достанется первый ранг, а самому маленькому – последний. После этого мы снова делим их на две группы и считаем суммы рангов отдельно для песиков и для котиков. Общая логика такова: чем сильнее будут различаться эти суммы, тем больше различаются песики и котики.



Вопросы?

Контакты спикера:
yustiks@gmail.com