
Статистическая проверка гипотез для связанных выборок. A/B тесты и как их проводить

Миленькин Александр
Биоинформатик в Insilico Medicine





Александр Миленькин

Биоинформатик в Insilico Medicine

О спикере:

- Преподаю в Нетологии
- Активно участвую в соревнованиях по Data Science.
- Окончил МФТИ в 2019 году
- Обожаю сноуборды

Аккаунт в Slack и Telegram



@Aleron75infskin



@Aleron



План Урока



A/B тестирование



Статистическая проверка гипотез для связанных выборок



Дисперсионный анализ

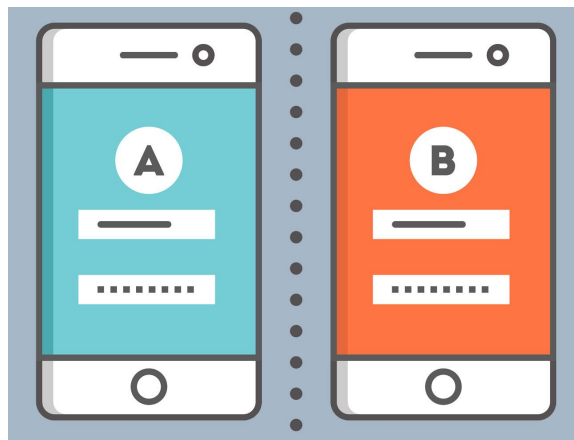


Проблема множественного сравнения

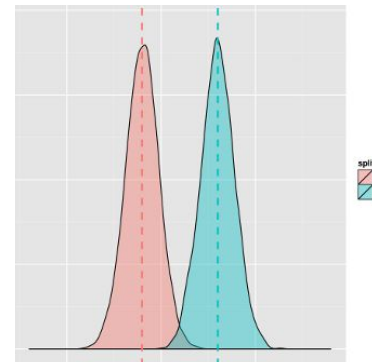


Практика на Python

Допустим у нас есть предположения, что если поменять цвет сайта, то конверсия увеличится. Как проверить?



Еще примеры!



A/B тестирование



Требования к тестам

- **Понимание метрики/цели**
- **Одновременность**
- **Случайность**
- **Достаточности объёма выборки**
- **Независимость**

Требования к тестам



Понимание метрики/цели



Одновременность



Случайность



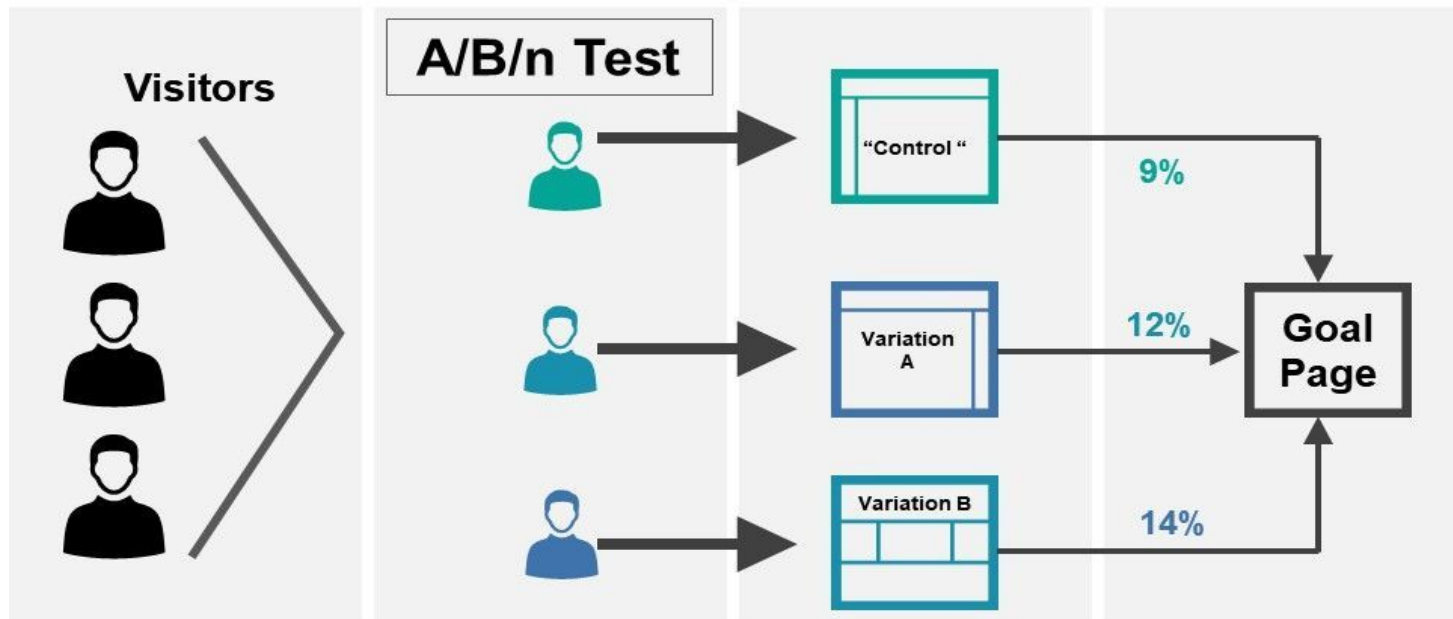
Достаточности объёма выборки



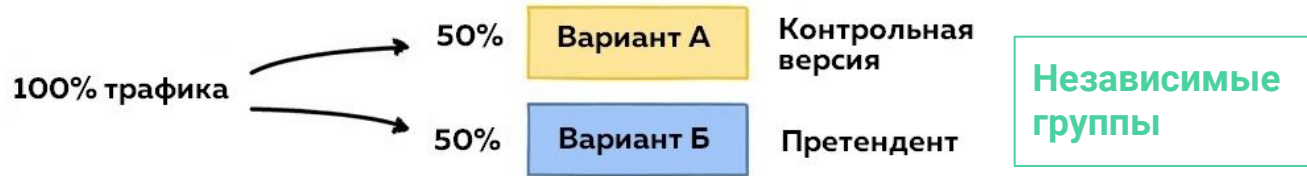
Независимость

$$n = \left(\frac{z_{\alpha} \sigma}{\Delta} \right)^2$$

Multi A/B тесты и как их проводить



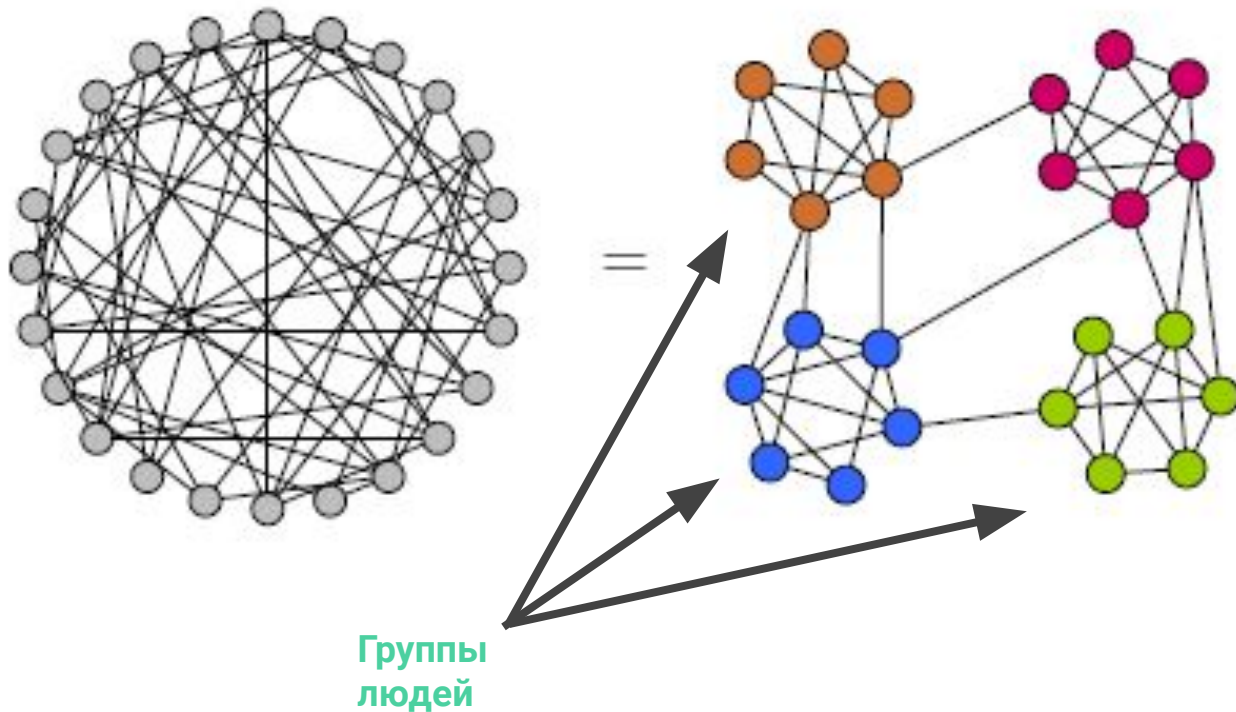
А/В тесты при зависимых группах



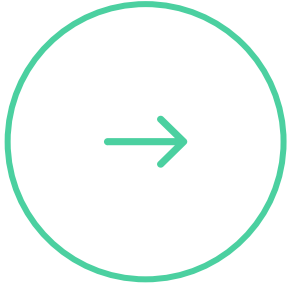
А/В тесты при зависимых группах



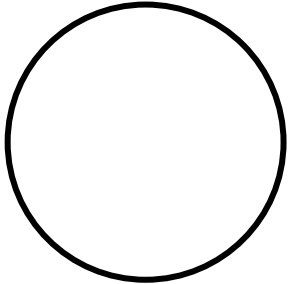
Multi A/B тесты при зависимых группах



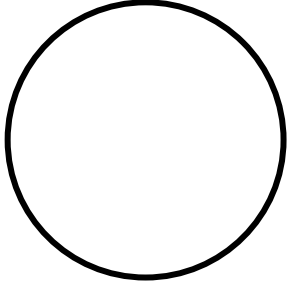
С А/В тестами ты один в поле воин!



В 2012 году сотрудник Microsoft, работавший над поисковой системой Bing, провел эксперимент по тестированию различных способов отображения рекламных заголовков. В течение нескольких часов альтернативный формат привел к увеличению доходов на 12% без влияния на показатели взаимодействия с пользователем.

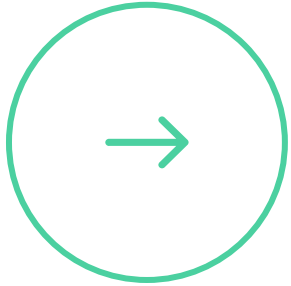


По итогам 2012 года выручка Microsoft составила \$74 млрд, чистая прибыль — около \$17 млрд.



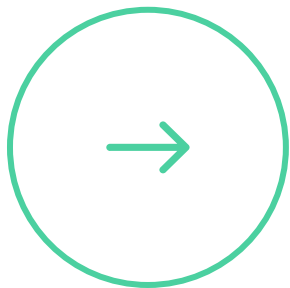
Прирост чистой прибыли \$2,04 млрд — круто?

Статистическая проверка гипотез

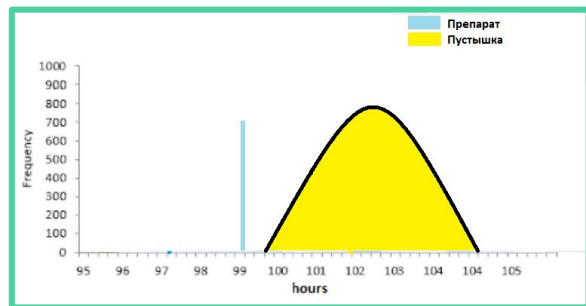


- $p \leq \alpha$: отклонить H_0 .
- $p > \alpha$: не отклонить H_0 .

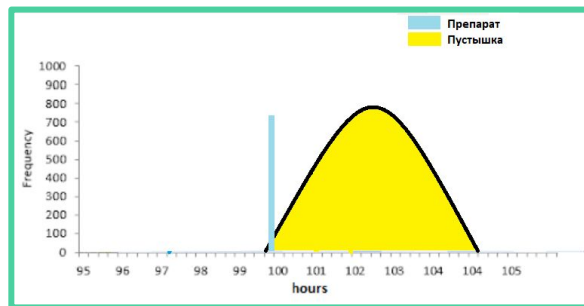
Статистическая проверка гипотез



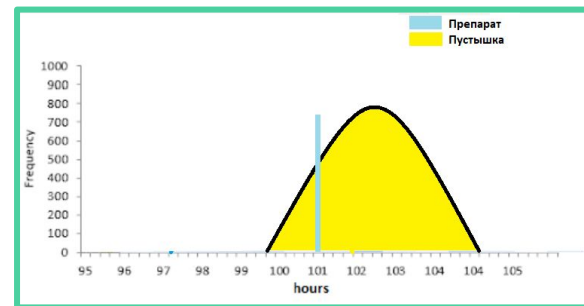
- $p \leq \alpha$: отклонить H_0 .
- $p > \alpha$: не отклонить H_0 .



p-value < 5%



p-value = 5%



p-value > 5%

Карта тестов гипотез

Сравнение двух выборок
(за исключением номинальных данных)

Параметрические
(только количественные)

t-тест Стьюдента

Непараметрические
(количественные или порядковые)

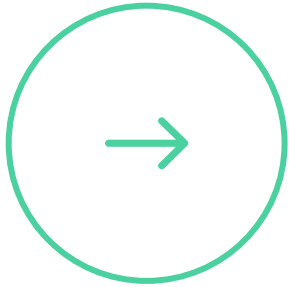
Зависимые

Критерий
Уилкоксона

Независимые

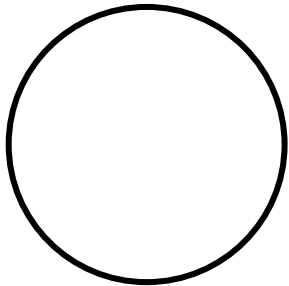
U-критерий
Манна-Уитни

t -критерий Стьюдента

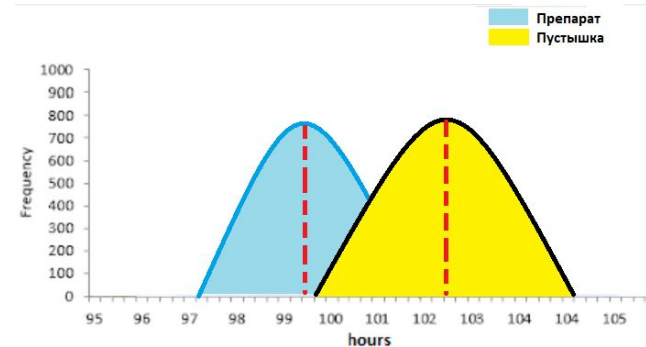


Рассмотренный ранее t -критерий Стьюдента предназначен для сравнения исключительно двух совокупностей

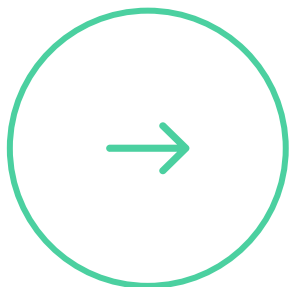
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



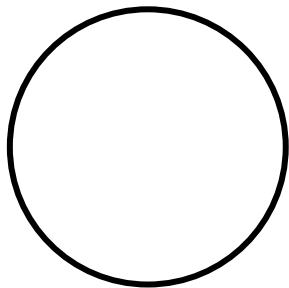
Невозможно отклонить $H_0: \mu_1 = \mu_2$
Отклонить H_0 : распределения выборок не равны.



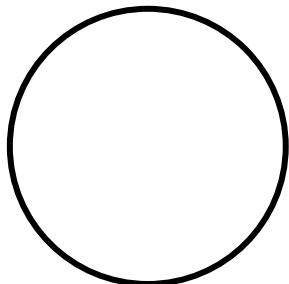
Т-критерия Вилкоксона



Используемый для проверки различий между двумя выборками зависимых или независимых по количественному признаку (непрерывного или в порядкового шкале)



Невозможно отклонить H_0 :
распределения выборок равны
Отклонить H_0 : распределения
выборок не равны.



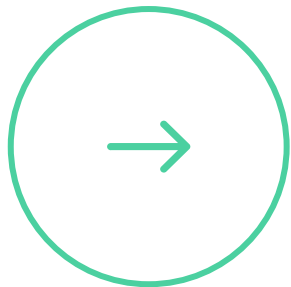
Тест Вилкоксона для
независимых выборок
называется критерием Манна-
Уитни

$$T = \sum R_r$$

где R_r - ранговые значения сдвигов с
более редким знаком

| № | Уровень тревожности (до тренинга) | Уровень тревожности (после тренинга) | Шаг 2: Разность (после-до) | Шаг 3: Значение разности по модулю | Шаг 4: Ранг разности |
|---|-----------------------------------|--------------------------------------|----------------------------|------------------------------------|----------------------|
| 1 | 15 | 14 | -1 | 1 | 3 |
| 2 | 14 | 11 | -3 | 3 | 8 |
| 3 | 16 | 17 | 1 | 1 | 3 |

Дисперсионный анализ

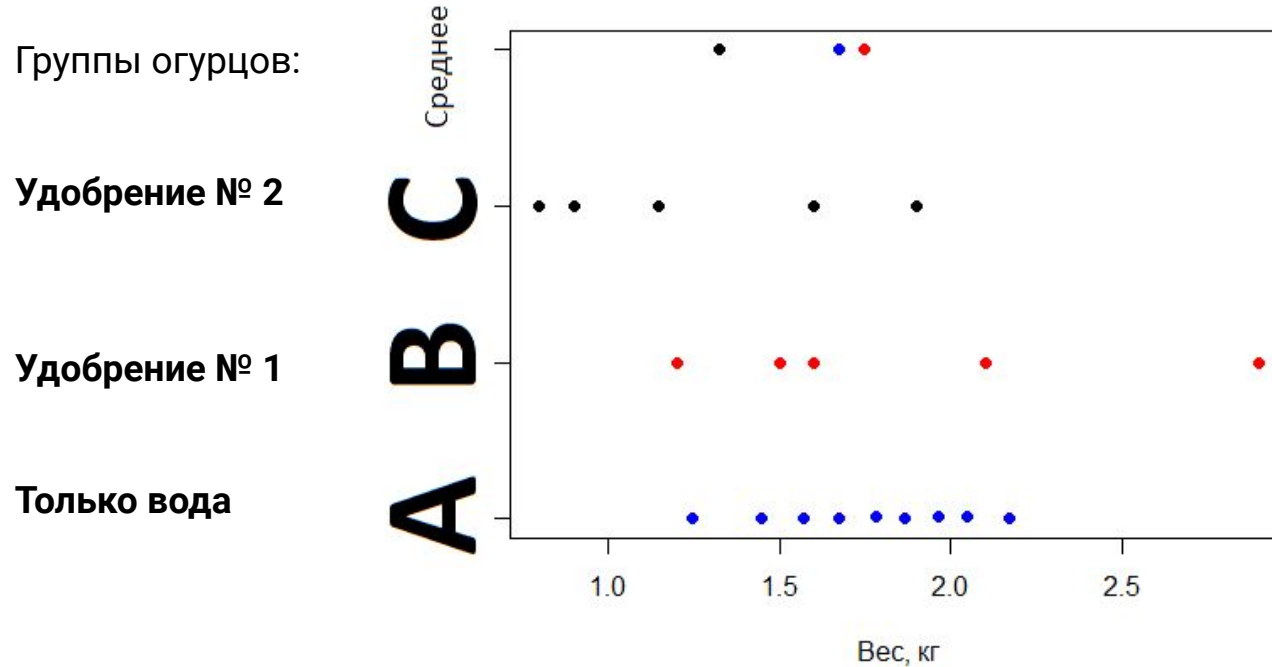


В случае, когда нужно сравнить две или больше выборок. Именно тогда и целесообразно применение **дисперсионного анализа**.

Название метода указывает на то, что выводы делают на основе исследования составляющих дисперсии.

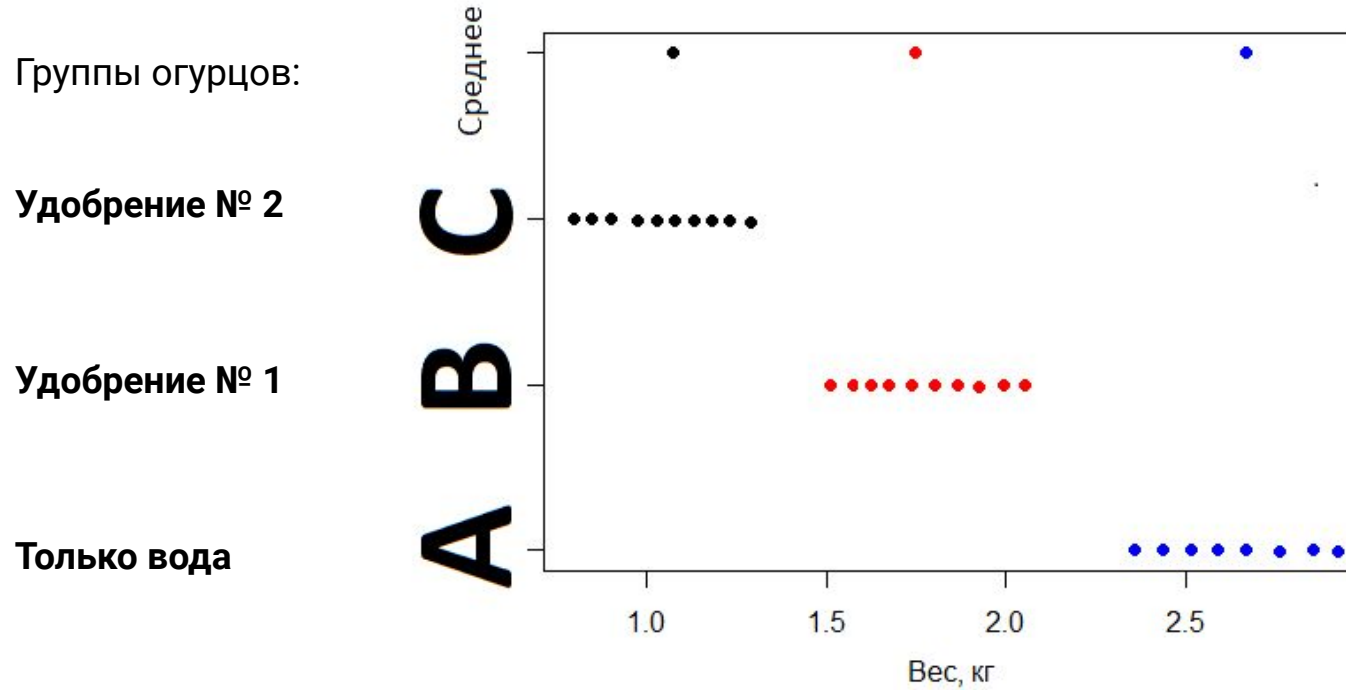
Суть изучения состоит в том, что общее изменение показателя разбивают на составляющие части, которые соответствуют действию каждого отдельно взятого фактора.

Дисперсионный анализ



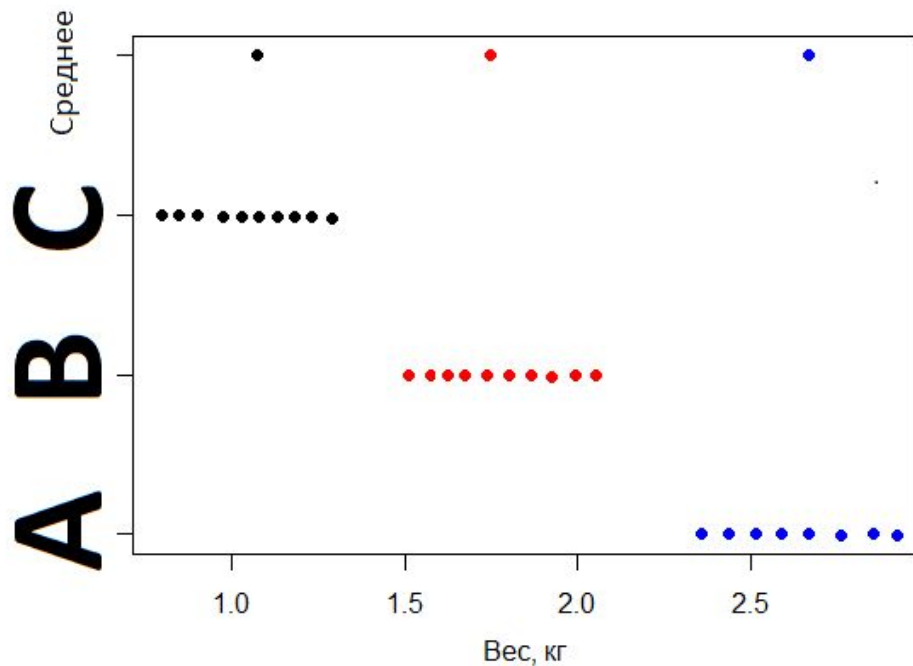
Что можно сказать про эффективность удобрений?

Дисперсионный анализ



А теперь что можно сказать про эффективность тестов?

Чем будем оценивать?

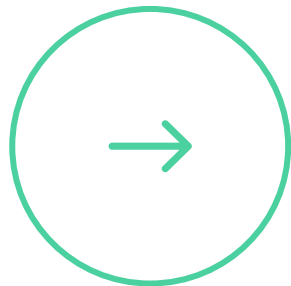


Нам хорошо знаком t -критерий Стьюдента, который мы часто с легкостью рассчитывают для каждой пары сравниваемых групп.

Получив достаточно высокое значение t в *каком-либо* из этих сравнений, исследователь сообщает, что " $P < 0.05$ ".

Это утверждение означает, что вероятность ошибочного заключения о существовании различий между групповыми средними не превышает 5% .

Проблема множественного сравнения



Однако, вероятность ошибки гораздо выше 5%. Очевидно, что дальнейшее увеличение числа проверяемых гипотез будет неизбежно сопровождаться возрастанием ошибки первого рода.

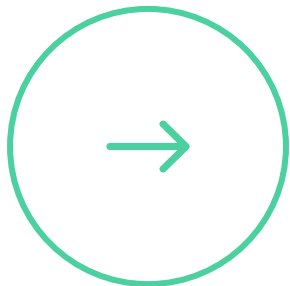
Выполняя тест Стьюдента, проверяем нулевую гипотезу об отсутствии разницы между генеральными средними двух сравниваемых групп.

Сравнивая группы А и В, он рискует ошибиться с вероятностью 5%. Точно такая же вероятность ошибки при сравнении В с С и А с С. Соответственно, вероятность ошибиться *хотя бы в одном* из этих трех сравнений составит

$$P' = 1 - (1 - \alpha)^m = 1 - (1 - 0.05)^3 = 0.143,$$

Проблема множественного сравнения

К чему приводит тестирования множества гипотез?

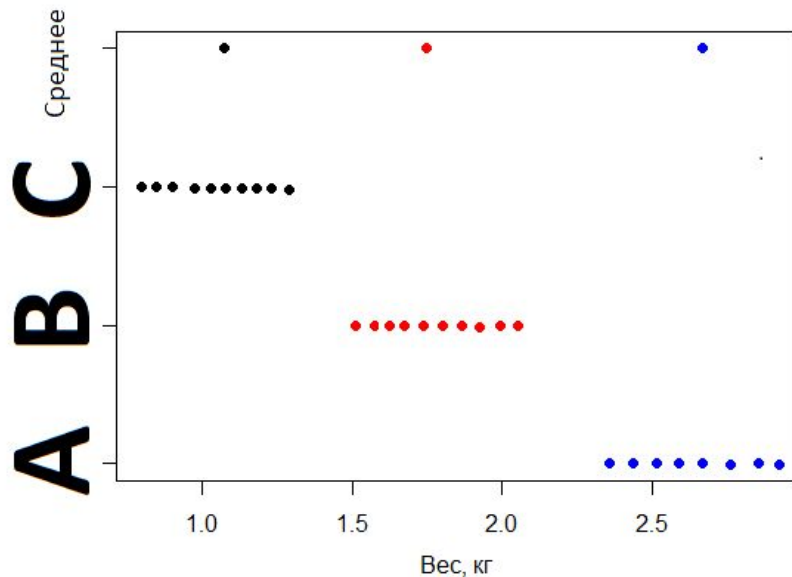


Очевидно, что дальнейшее увеличение числа проверяемых гипотез будет неизбежно сопровождаться возрастанием ошибки первого рода.

$$P' = 1 - (1 - \alpha)^m = 1 - (1 - 0.05)^3 = 0.143,$$

t-критерий Стьюдента не подходит для попарного сравнения большего количества групп что вызывает т.н. *эффект множественных сравнений*

Дисперсионный анализ

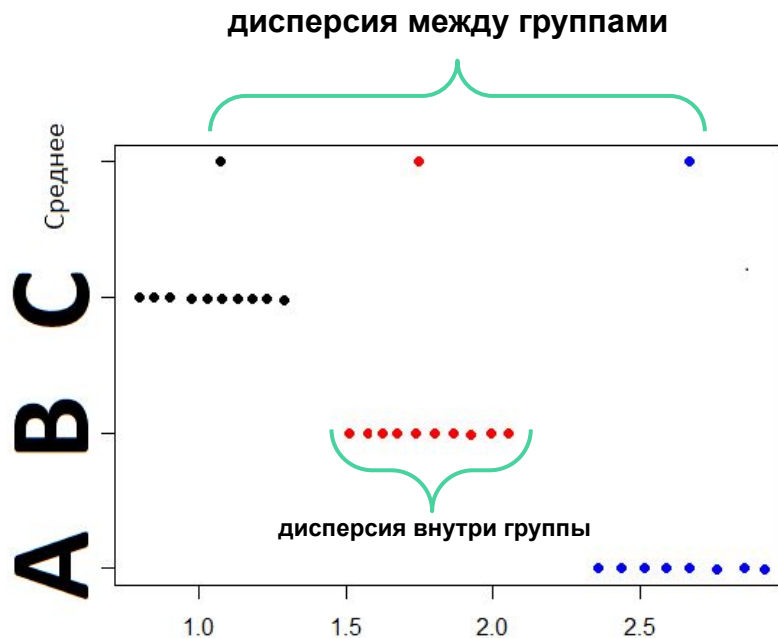


Сравните разброс значений **внутри** экспериментальных групп с **разбросом** трех **групповых средних**: разброс групповых средних в целом превышает разброс значений в экспериментальных группах.

Для оценки различий между группами следует сравнить разброс **групповых средних** с разбросом значений **внутри групп**.

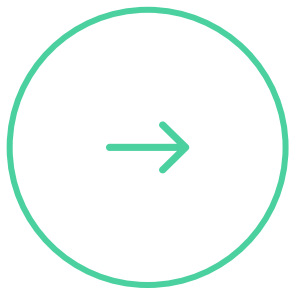
Это ключевая идея дисперсионного анализа!

Идея дисперсионного анализа

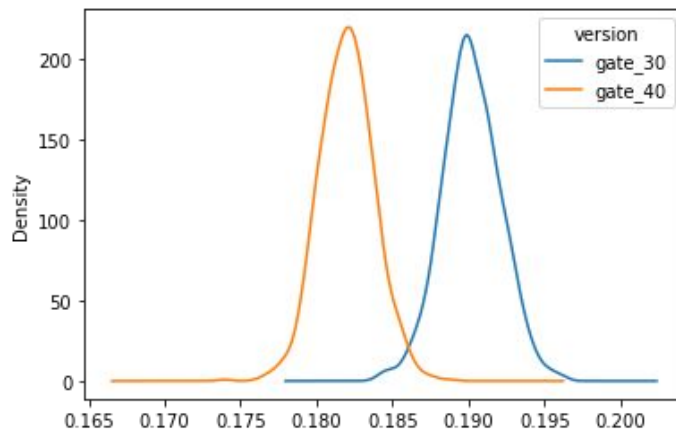


Получается, для ДА важна
дисперсия внутри группы и
дисперсия между группами

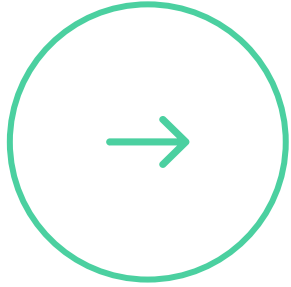
Чем будем оценивать?



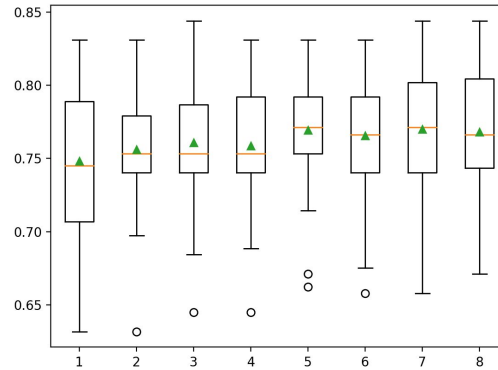
Для устранения эффекта множественных сравнений существует большой арсенал методов.



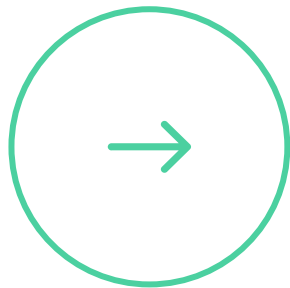
Дисперсионный анализ



принципы *однофакторного дисперсионного анализа*, как раз предназначенного для *одновременного сравнения средних значений* двух и более групп.

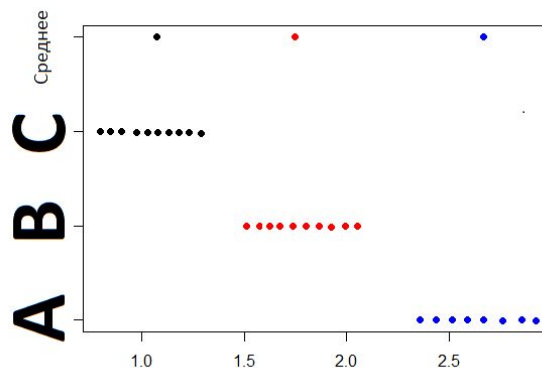


Виды и критерии дисперсионного анализа



Этот метод используется для исследования связи между качественными (номинальными) признаками и количественной (непрерывной) переменной.

В случае двух выборок, результаты дисперсионного анализа будут идентичны результатам t-критерия Стьюдента. Однако, в отличие от других критериев, это исследование позволяет изучить проблему более детально.



Мотивирующие примеры применения ДА

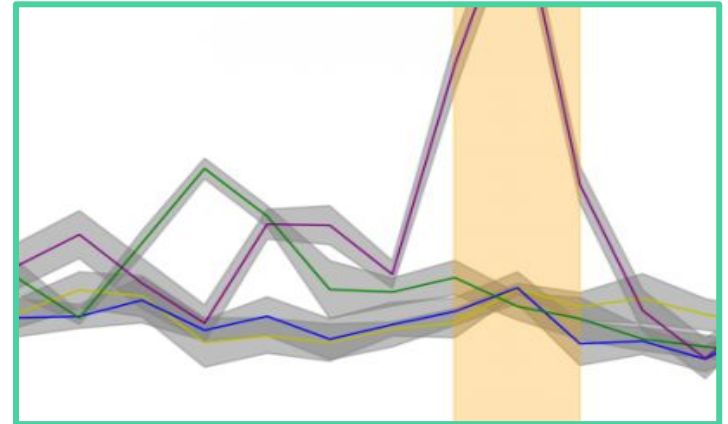
Троллейбусное депо обслуживает несколько троллейбусных маршрутов. На них работают троллейбусы различных типов, и оплату за проезд собирают 125 контролеров.

Как сравнить экономические показатели работы каждого контролера (выручку) учитывая различные маршруты, различные типы троллейбусов?



Есть ряд станков - которые изготавливают деталь. Размер каждой детали - это случайная величина, которая зависит от настройки каждого станка и случайных отклонений, возникающих в процессе изготовления деталей.

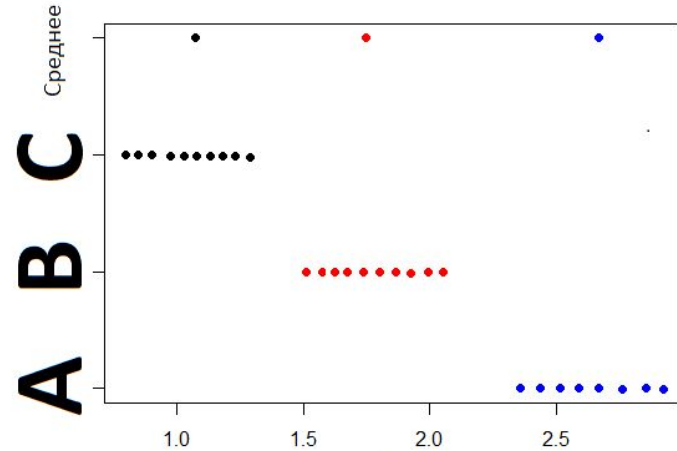
Нужно по данным измерений размеров деталей определить, одинаково ли настроены станки.



Мотивирующие примеры применения ДА

Троллейбусное депо обслуживает несколько троллейбусных маршрутов. На них работают троллейбусы различных типов, и оплату за проезд собирают 125 контролеров.

Как сравнить экономические показатели работы каждого контролера (выручку) учитывая различные маршруты, различные типы троллейбусов?



Дисперсионный анализ

каждое наблюдение мы можем разложить на следующие составляющие:

$$x_{ij} = \bar{X} + (\bar{x}_i - \bar{X}) + (x_{ij} - \bar{x}_i)$$

разброс наблюдений *внутри групп*

$$SS_W = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$$

разброс *между группами*
(разброс групповых средних)

$$SS_B = \sum_i \sum_j (\bar{x}_i - \bar{X})^2$$

$$(\bar{x}_i - \bar{X})$$



это отклонения групповых средних от общего среднего значения

$$(x_{ij} - \bar{x}_i)$$



это отклонения отдельных наблюдений от среднего значения группы, к которой они принадлежат.

Дисперсионный анализ

Внутригрупповую дисперсия

$$MS_W = SS_W / (N - k)$$

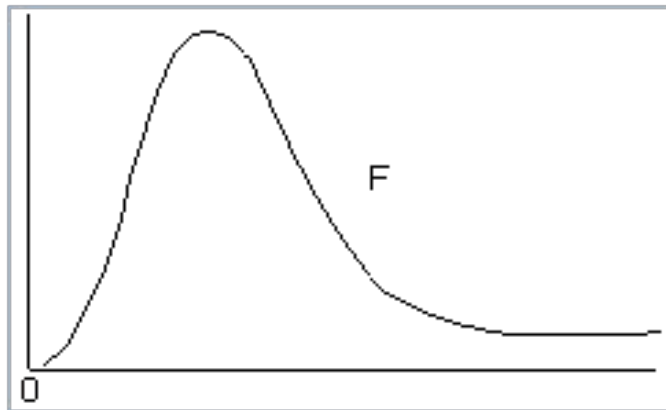
Межгрупповую дисперсия

$$MS_B = SS_B / (k - 1)$$

Критерия Фишера

$$F = MS_B / MS_W$$

Ну и где тут p-value 5% ?



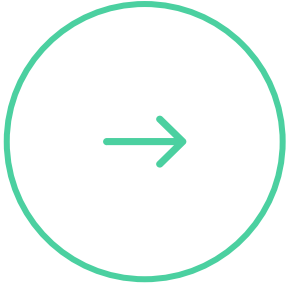
$$F = MS_B / MS_W$$

Ну и где тут p-value 5% ?



$$F = MS_B / MS_W$$

Дисперсионный анализ = ANOVA



Замечание: в специальной литературе его часто называют **ANOVA** (от англоязычного названия Analysis of Variations).

Впервые этот метод был разработан Р. Фишером в 1925 г.

А как же
Python?



Домашнее задание

1. Выполнить простые задачи в python-ноутбуке.
2. Решить приложенные Cookie Cats и Pokemons кейсы.



Спасибо за внимание!

Миленькин Александр
Биоинформатик в Insilico Medicine



fb.com/



sasha.milenkin@yandex.ru

 **НЕТОЛОГИЯ**