Вариации A/B-тестирования и его приложения

A/B/N-мультитестирование. Проблема множественного тестирования и дисперсионный анализ

Цель урока

Никто не запрещает нам тестировать сразу много гипотез, но тут можно допустить и много ошибок. Обсудим тонкости множественного тестирования.

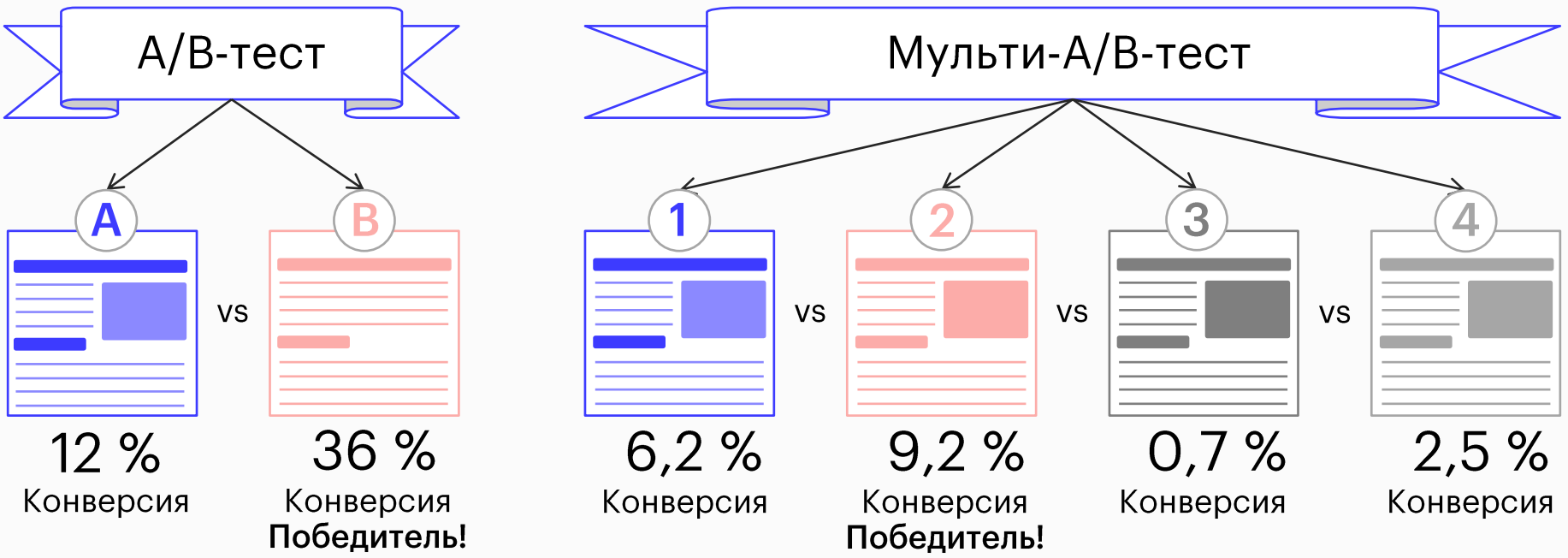
Задачи урока

- Узнаем, как применять A/B/N-, или множественное тестирование
- Научимся применять поправку Бонферрони
- Посмотрим на идею дисперсионного анализа
- Разберём, что делать, если распределение ненормальное, и изучим критерий Краскела — Уоллиса

A/B/N-мультитестирование. Проблема множественного тестирования и дисперсионный анализ

A/B/N-тестирование и A/B-тестирование

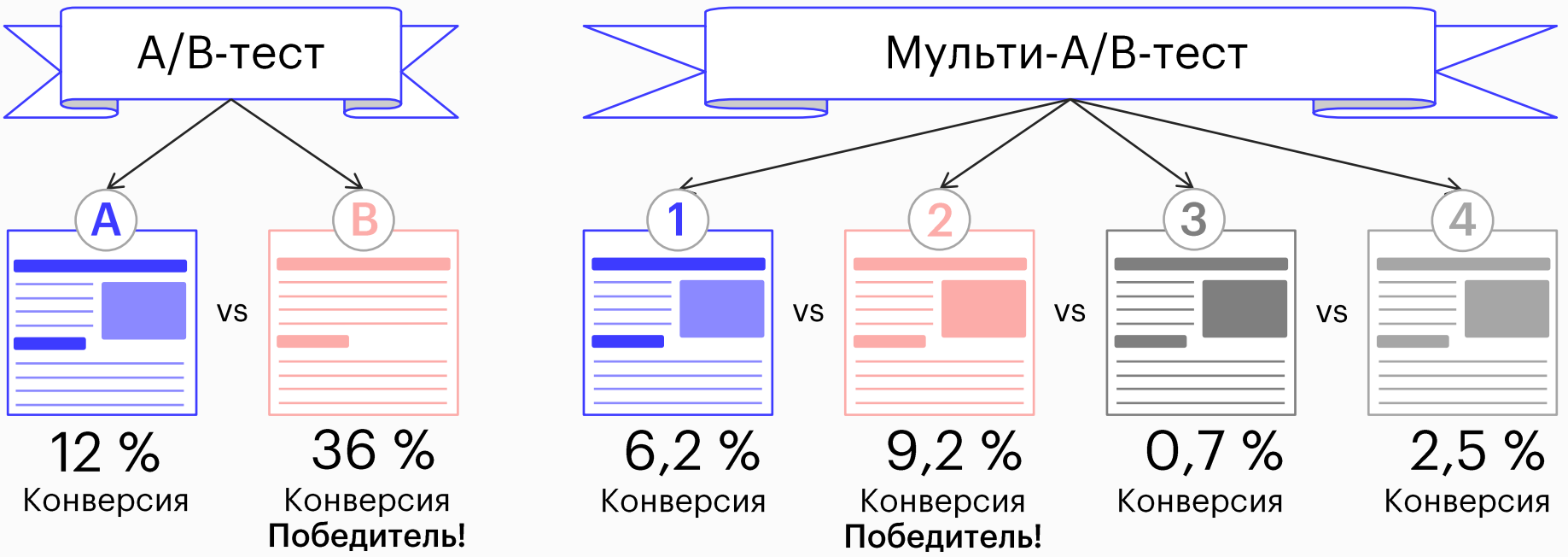
A/B-тестирование не ограничивается сравнением 2 групп. Можно сравнить N групп, но потребуется больше трафика для надёжности информации.



A/B/N-мультитестирование. Проблема множественного тестирования и дисперсионный анализ

A/B/N-тестирование и A/B-тестирование

A/B/N-тестирование позволяет протестировать группу изменений. Например, изменение не только надписи, но и цвета кнопки. Тогда каждый вариант содержит уникальное сочетание текста и оформления кнопки, что позволяет выделить оптимальный вариант.



A/B/N-мультитестирование. Проблема множественного тестирования и дисперсионный анализ

A/B/N-тестирование и A/B-тестирование

Так, ну здорово, давайте тогда одновременно будем проверять сразу много гипотез, чтобы экономить время!

Не спешите! Тут есть несколько тонкостей, на которых можно споткнуться. Давайте разбираться!

A/B/N-мультитестирование. Проблема множественного тестирования и дисперсионный анализ

A/B/N-тестирование и A/B-тестирование

Допустим, мы тестируем сразу 3 гипотезы для какой-нибудь витрины продуктов.

Мы знаем, что вероятность неверно отклонить нулевую гипотезу — это α = 5 % для одного случая.

Тогда при тестировании сразу трёх гипотез:

P (неверно отклонить нулевую гипотезу хотя бы для одного случая) = 1 – P (неверно отклонить нулевую гипотезу для 3-х случаев) = $1 - (1 - \alpha)^3$ = **14,3 %.**

И это только для трёх гипотез!

A/B/N-мультитестирование. Проблема множественного тестирования и дисперсионный анализ

Множественное тестирование

Получается, что при тестировании одновременно нескольких гипотез хотя бы по одной метрике **вероятность маловероятных событий** (например, различие между двумя группами при условии нулевой гипотезы) **увеличивается**!

Следовательно, увеличивается вероятность ошибки 1-го рода, то велика вероятность найти значимые результаты там, где их нет.

Посмотрим, что будет, если мы продолжим увеличивать количество одновременно проверяемых гипотез.

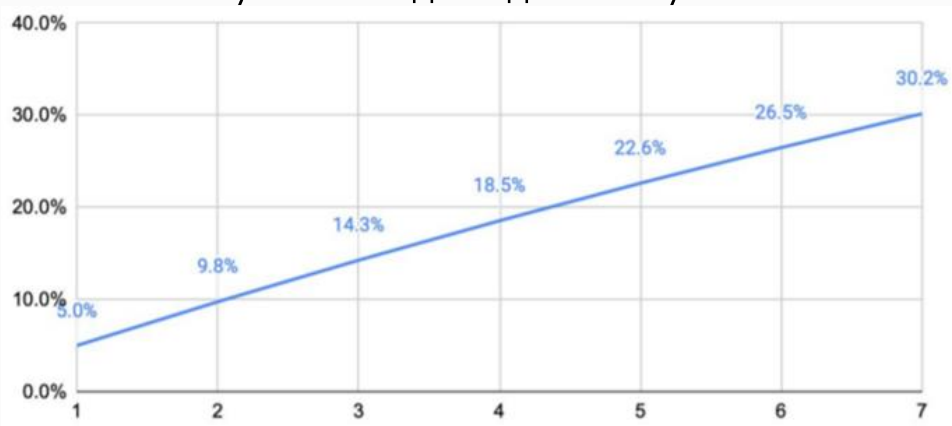
A/B/N-мультитестирование. Проблема множественного тестирования и дисперсионный анализ

Множественное тестирование и гипотезы

P (неверно отклонить нулевую гипотезу хотя бы для одного случая) = 1 – (1 – α)n.

Для α = 5 % график роста ошибки от количества тестируемых гипотез следующий.

Вероятность неверно отклонить нулевую гипотезу хотя бы для одного случая

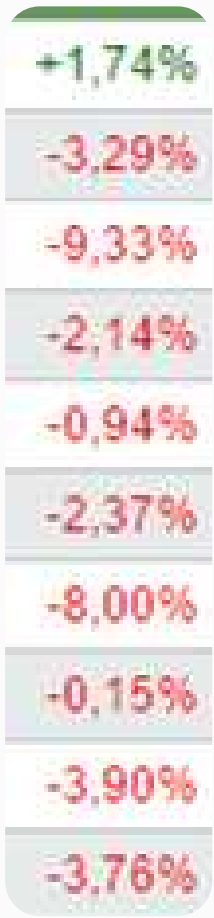


Вероятность ошибки

	Количество тестируемых гипотез						
alpha	1	2	3	4	5	6	7
1%	1%	2%	3%	4%	5%	6%	7%
5%	5%	10%	14%	19%	23%	26%	30%

Множественное тестирование и метрики

Обратите внимание, что, если у вас может быть всего **одна гипотеза**, но при этом вы отслеживаете сразу **много метрик** для этой гипотезы, что тогда?



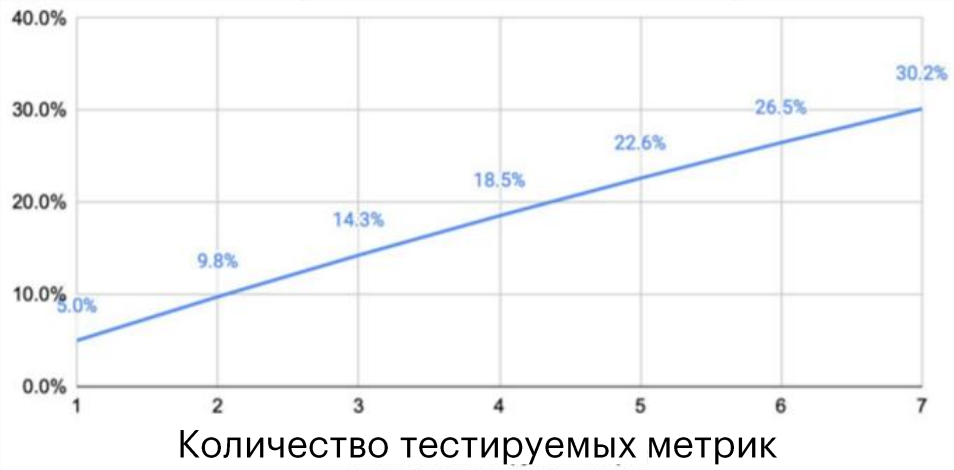
Подсказка

Множественное тестирование и метрики

Тогда задачу легко переформулировать в предыдущую и получить на выходе ту же ситуацию: по сути у вас много одинаковых гипотез, но для разных метрик. Тогда ошибка ведётся себя также:

- P (неверно отклонить нулевую гипотезу хотя бы для **одного метрики**) = 1 – (1 – α)n. Для α = 5 % график роста ошибки от **количества тестируемых метрик**

Вероятность неверно отклонить нулевую гипотезу хотя бы для одного случая



Вероятность ошибки

	Количество тестируемых метрик						
alpha	1	2	3	4	5	6	7
1%	1%	2%	3%	4%	5%	6%	7%
5%	5%	10%	14%	19%	23%	26%	30%

Множественное тестирование: решение

Увеличивая α, мы негативно влияем на мощность, отсюда имеется больший риск назвать неоднородные группы однородными.

Вывод: сравнивайте по меньшему **количеству метрик**, но с приемлемыми вероятностями ошибок 1-го и 2-го рода.

И что тогда делать, если проверять сразу много гипотез хочется, а ошибаться не хочется?

Ответ прост: использовать поправку на уровень значимости, например, поправку Бонферрони.

Множественное тестирование: поправка Бонферрони

Пусть

V — количество ложных отклонений нулевых гипотез.

FWER — вероятность хотя бы одной ошибки первого рода.

m — количество тестируемых гипотез.

Будем отвергать нулевую гипотезу, если **p-value** < α/**m**

$$\text{Тогда: } \mathrm{FWER} = P(V \geq 1) = P\left\{\bigcup_{i=1}^{m}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{i=1}^{m} P\left(p_i \leq \frac{\alpha}{m}\right) \leq m\frac{\alpha}{m} = \alpha$$

Получили то, что хотели в самом начале
для вероятности хотя бы одной ошибки первого рода.

Рассмотрим случай, когда мы проверим n гипотез.

	Верных H_{0i}	Неверных H_{0i}
Не отвергнутых H_{0i}	*U*	***T***
Отвергнутых H_{0i}	***V***	*S*

A/B/N-мультитестирование. Проблема множественного тестирования и дисперсионный анализ

Множественное тестирование: поправка Бонферрони

Вывод и смысл:

- проверять все гипотезы по критерию на уровне значимости α/m
- в таком случае FWER < α, что нам и нужно
- это простая, но грубая поправка, т. к. мощность критерия падает с увеличением числа m
- для небольшого количества групп такая поправка обычно работает

Множественное тестирование и другие поправки

Для общности скажем, что существуют и другие поправки, с помощью которых мы можем ограничить число ошибок первого рода. Для каждой из них используются определённые статистики. То есть, помимо **FWER = P (V ≥ 1)** (вероятность хотя бы одной ошибки первого рода), есть ещё:

- **FDP = V/D (D > 0)** — (доля ошибок первого рода среди всех отклонённых гипотез)
- **RDR = E(FDP)** — (средняя доля ошибок первого рода) и другие

На практике часто используют **FWER**.

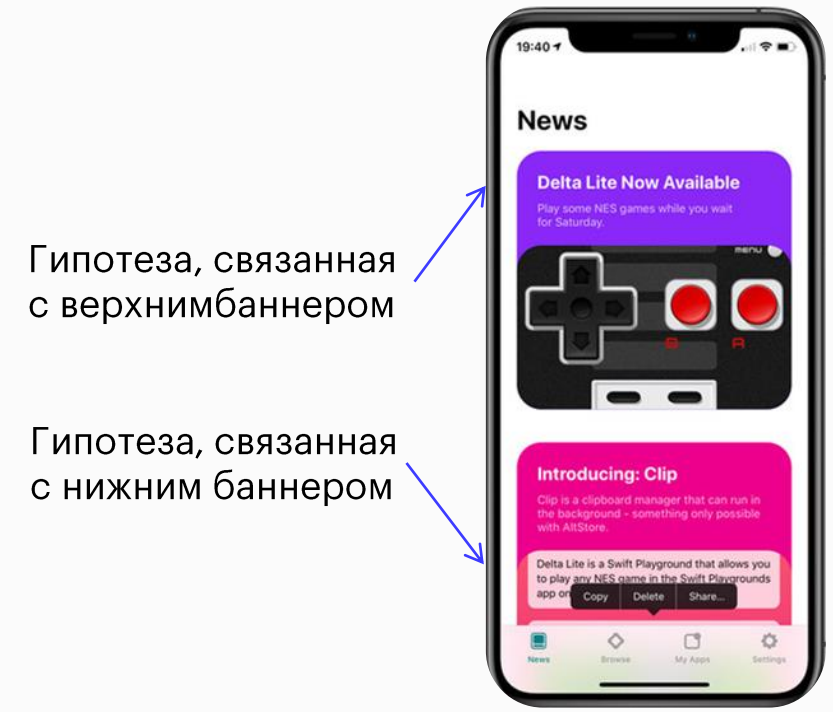
Множественное тестирование

Затронем теперь такую тему,
как независимость между тестами.

Зададимся вопросом: «А можем ли мы тестировать сразу **несколько гипотез** на **одном пользователе?»**

Если да, то мы сможем **ещё сильнее увеличить количество гипотез**, которое проверяем одновременно, а значит быстрее улучшать наш сервис!

Разберём пару примеров.



Множественное тестирование: один пользователь

Может ли один пользователь
участвовать в двух тестах?

- **Тест 1.** Первичные метрики —
  конверсия в покупку пакета
- **Тест 2.** Первичные метрики —
  конверсия в подписку
  на обновления

Множественное тестирование: один пользователь

Может ли один пользователь
участвовать в двух тестах?

- **Тест 1**. Первичные метрики —
  конверсия в покупку пакета
- **Тест 2**. Первичные метрики —
  конверсия в подписку
  на обновления

Ответ: может, так как пакет не влияет
на желание отслеживать обновления.

Множественное тестирование: один пользователь

Курьеру предлагается заказ. У него есть 5 минут на то, чтобы его принять, отклонить или проигнорировать. Если заказ был принят, то курьер может его отменить.

- **Фича 1**: предлагаем курьерам из тестовой группы 10 минут
  - **Ожидание 1**: увеличение доли принятых заказов
- **Фича 2**: увеличиваем штраф за отмену заказа в 2 раза
  - **Ожидание 2**: уменьшение отмен

Вопрос: может ли один пользователь участвовать в двух тестах?

Множественное тестирование: один пользователь

Курьеру предлагается заказ. У него есть 5 минут на то, чтобы его принять, отклонить или проигнорировать. Если заказ был принят, то курьер может его отменить.

- **Фича 1**: предлагаем курьерам из тестовой группы 10 минут
  - **Ожидание 1**: увеличение доли принятых заказов
- **Фича 2**: увеличиваем штраф за отмену заказа в 2 раза
  - **Ожидание 2**: уменьшение отмен
- **Вопрос**: может ли один пользователь участвовать в двух тестах?
  - **Ответ**: нет. Фича 1 увеличивает долю принятых заказов, а фича 2 понижает, курьеры принимают осторожнее, отсюда вывод: эффект от фичи 1 невозможно определить

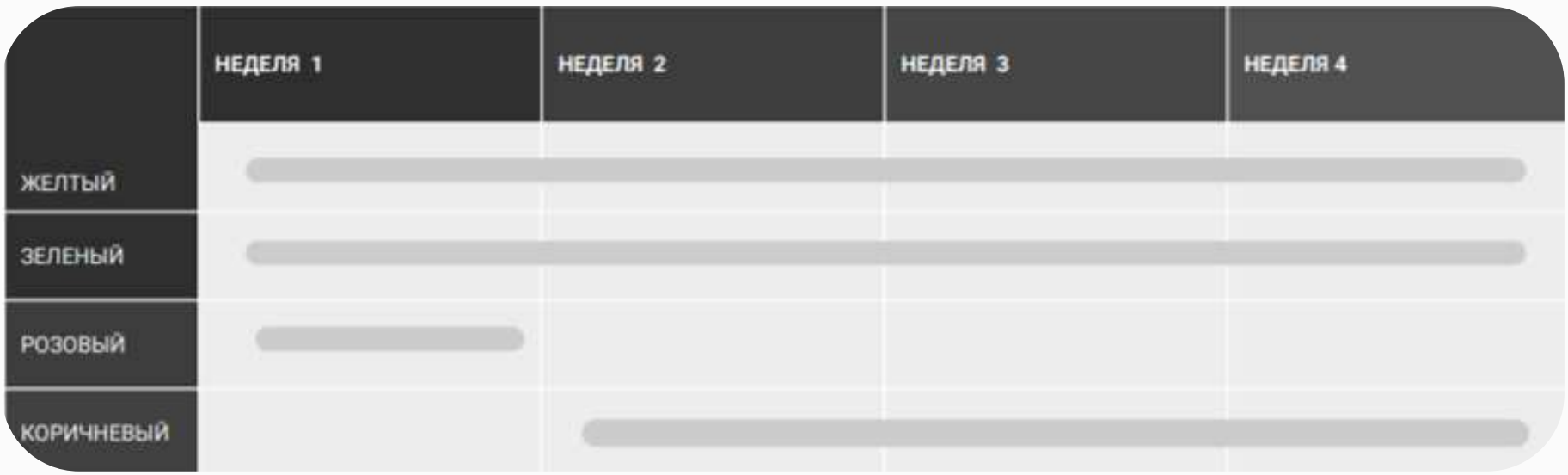
Не допускайте пересекающихся тестовых выборок.

A/B/N-мультитестирование. Проблема множественного тестирования и дисперсионный анализ

Множественное тестирование: изменение вариаций

Тестируем три вариации цвета кнопки. После первой недели конверсия по розовому цвету была ниже всех, и поэтому решили заменить его на коричневый.

Можем ли сравнить три вариации?



Множественное тестирование: изменение вариаций

Ответ: нет.

Данных по первой неделе коричневого цвета нет.

Отсюда результаты этой недели имеют меньший вес в результатах. Отсюда невалидные выводы.

Вдруг на первой неделе был аномально высокий спрос. Отсюда цены выше. Отсюда конверсии уменьшились у всех вариантов одинаково? Коричневый цвет этого уменьшения не учитывает.

Хорошо, ясно теперь, что надо снижать уровень значимости. А есть ли какие-то критерии, но не для двух групп, а сразу для нескольких?

Да, такие есть! Так же, как для двух групп у нас параметрический **критерий Стьюдента** и непараметрический **критерий Манна — Уитни**, как и для нескольких групп есть ещё непараметрический **критерий Крускала — Уоллиса** и параметрический **дисперсионный анализ.**

Далее подробнее поговорим про эти критерии. А пока только скажем, что эти критерии позволяют смотреть наличие статистически значимого различия без накопления ошибки
при множественном тестировании.

```
from scipy import stats

# Используем дисперсионнай анализ
statistic, pvalue = stats.f_oneway(A, B, C, D, F)

# Теперь вычисляем Крускал Уоллис тест
statistic, pvalue = stats.kruskal(A, B, C, D, F)
```

Дисперсионный анализ

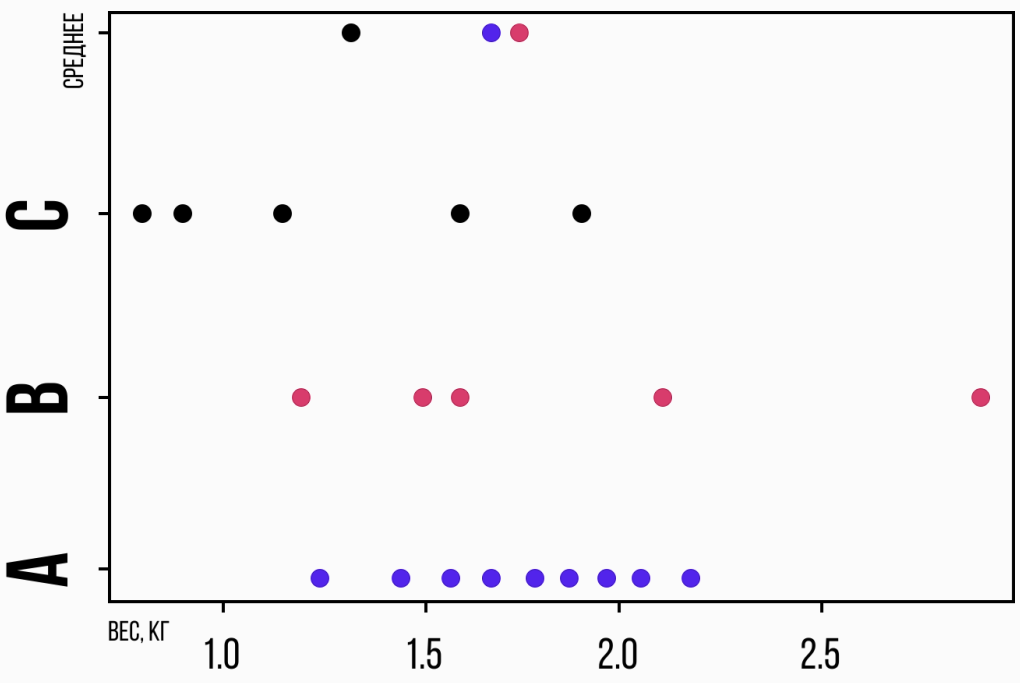
Допустим, мы хотим определить, какой способ выращивания арбузов лучше.

Группы арбузов:

Удобрение № 2

Удобрение № 1

Только вода



A/B/N-мультитестирование. Проблема множественного тестирования и дисперсионный анализ

Дисперсионный анализ

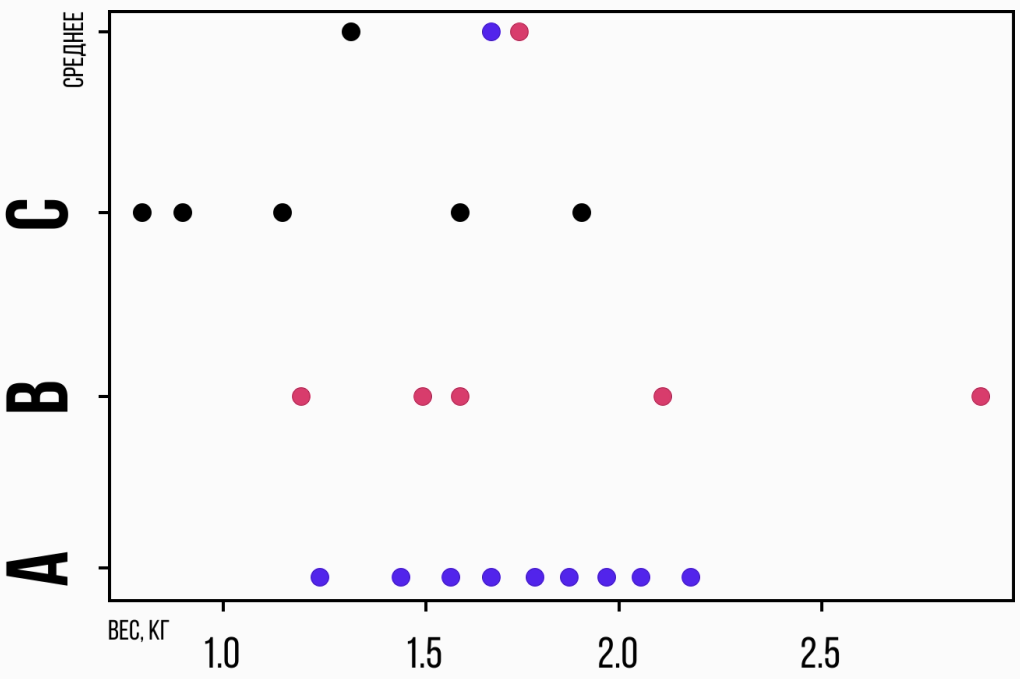
Что можно сказать про эффективность удобрений? Какое лучше?

Группы огурцов:

Удобрение № 2

Удобрение № 1

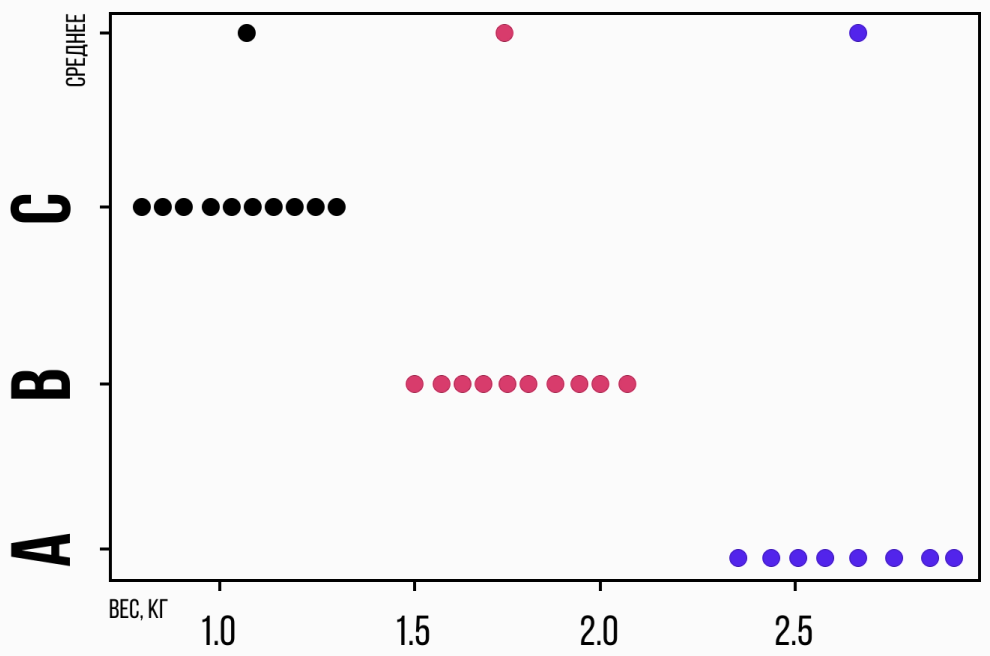
Только вода



Идея дисперсионного анализа

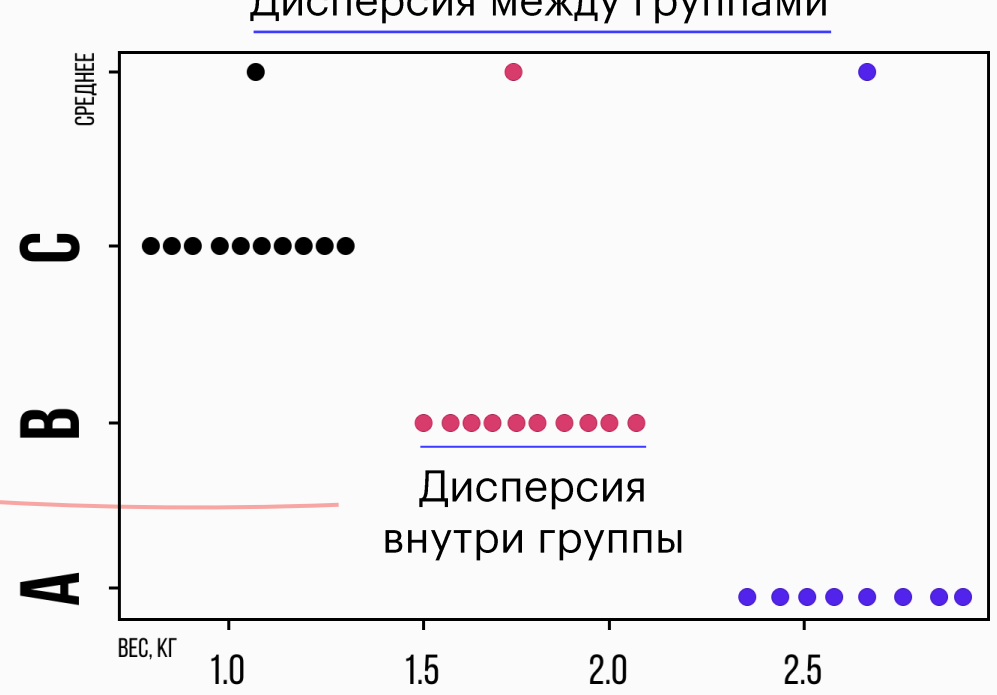
А что теперь можно сказать
про эффективность
удобрений? Какое лучше?

А что изменилось?



Идея дисперсионного анализа

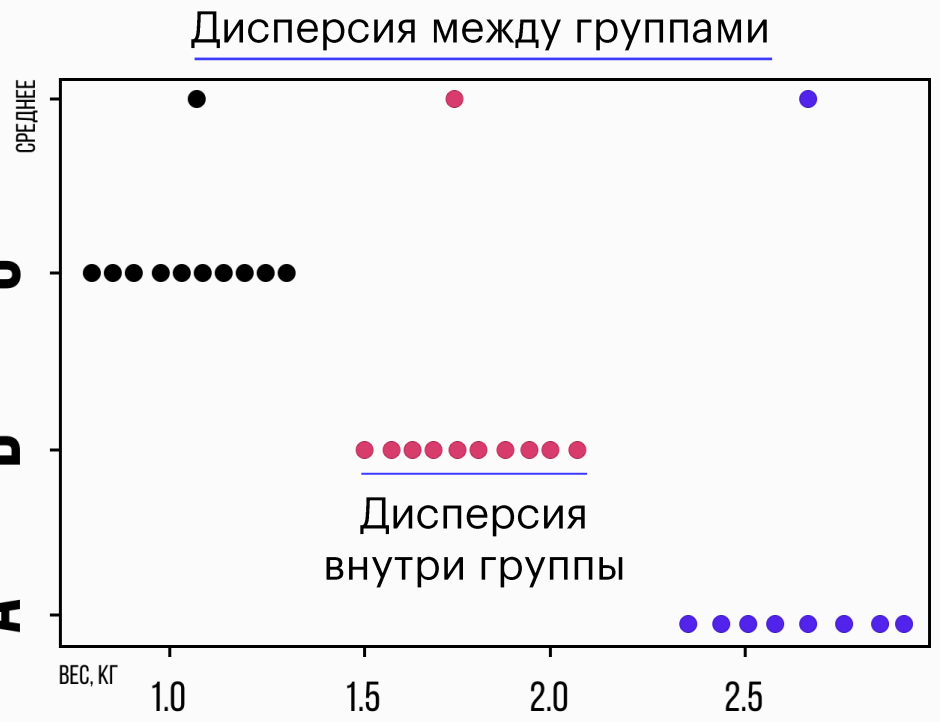
Изменились всего две составляющие!



Это и есть ключевая идея дисперсионного анализа!

Идея дисперсионного анализа

Получается,
для **дисперсионного анализа** важны
дисперсия внутри групп
и дисперсия между
группами!

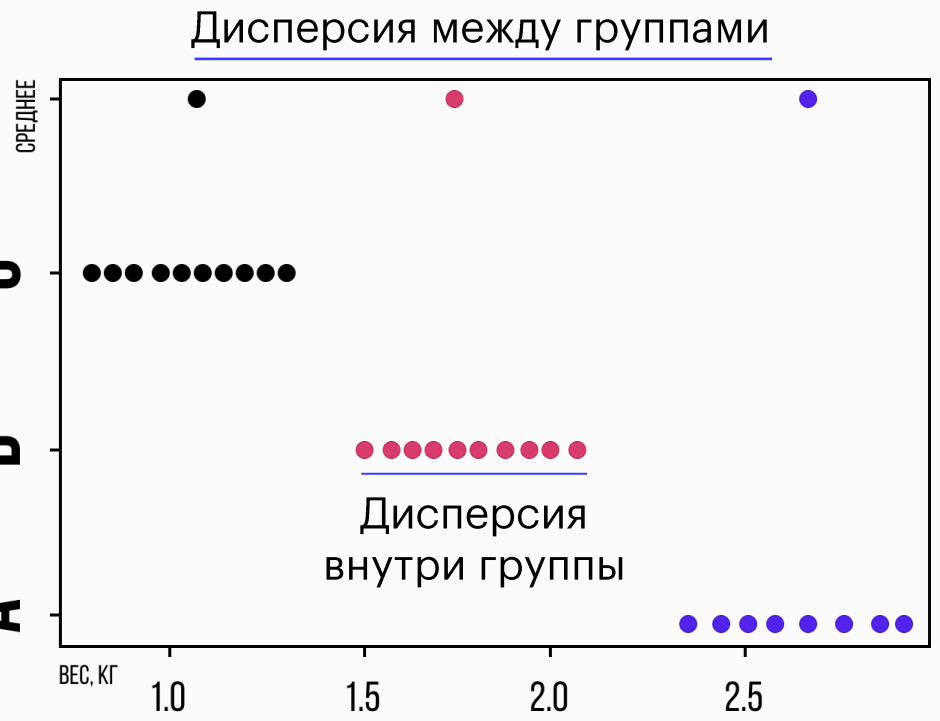


A/B/N-мультитестирование. Проблема множественного тестирования и дисперсионный анализ

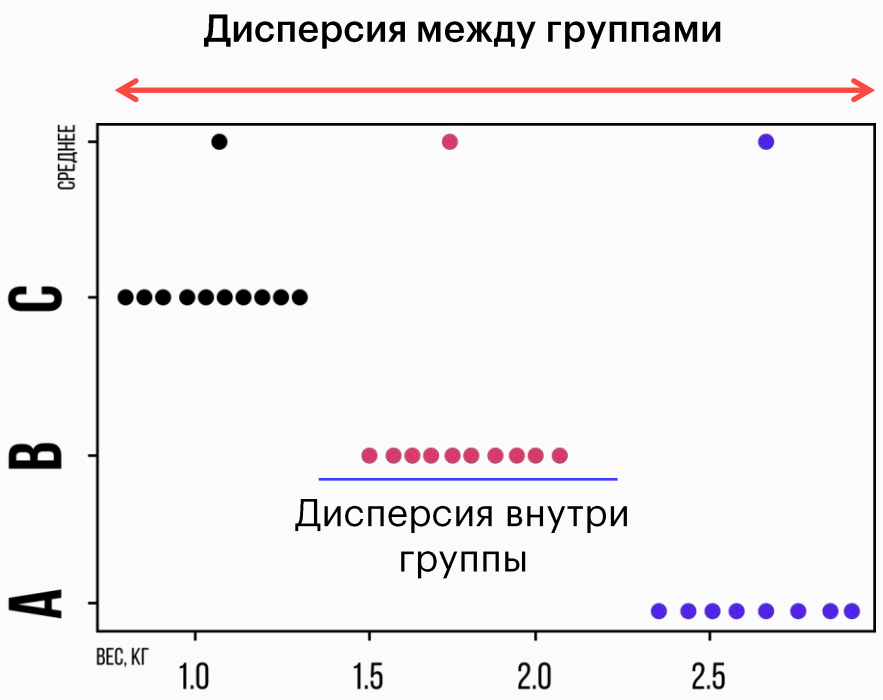
Идея дисперсионного анализа

$$F = \frac{\text{Дисперсия между группами}}{\text{Дисперсия внутри группы}}$$

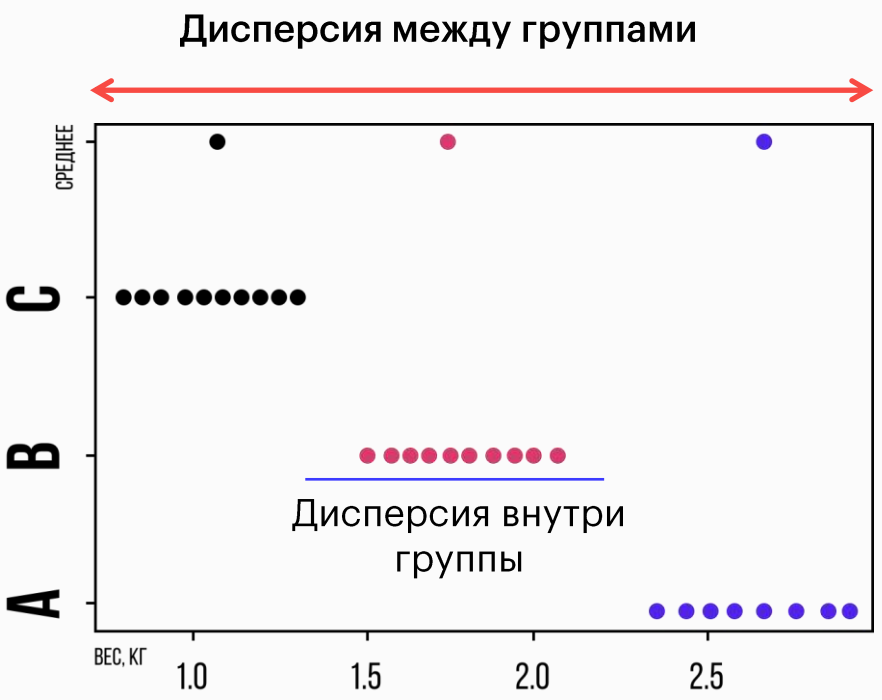
Чем больше F,
тем проще различить
выборки



Идеи дисперсионного анализа



F = 10



F = 5

Чем больше F, тем проще различить выборки

Критерий Краскела — Уоллиса

Дисперсионный анализ по Краскелу — Уоллису относится к группе непараметрических методов статистики. Это значит, что он не зависит от распределения. В нём используются ранги исходных значений и их суммы в сравниваемых группах. В частности, метод Краскела — Уоллиса основан на вычислении т. н. *H*-критерия:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{m} \frac{R_i^2}{n_i} - 3(N+1),$$

Критерий Краскела — Уоллиса

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{m} \frac{R_i^2}{n_i} - 3(N+1),$$

Где n^i — число наблюдений в группе i, N = общее число наблюдений во всех m-группах, а R^i — сумма рангов наблюдений в группе i. Ранг представляет собой порядковый номер конкретного наблюдения в ряду упорядоченных по возрастанию наблюдений. Аналогично F-критерию, чем больше значение H-критерия, тем больше у нас оснований отклонить нулевую гипотезу об отсутствии разницы между сравниваемыми группами.

Итоги и выводы урока

Множественное A/B-тестирование может сильно ускорить количество гипотез, тем самым ускорить развитие вашего сервиса.

- Помните, чем больше гипотез вы одновременно проверяете, тем более строгие условия должны быть на уровне значимости. Например, можно использовать уровень значимости, делённый на количество тестирований
- Задавайтесь вопросом, являются ли ваши тестирования зависимыми, так вы сможете либо ускорить ваше тестирование рамках одного пользователя, либо исключить возможные ошибки
- В случае множественного тестирования могут помочь специальные критерии, решающие проблему роста ошибки