e^xperiment $f$est

A/A-тесты

проверка качества систем сплитования

Что узнаем?

- Определение
- Задачи A/A-тестов
- Методы расчетов A/A-тестов
- Ограничения

A/A-тесты, преимущественно, необходимы для проверки систем сплитования

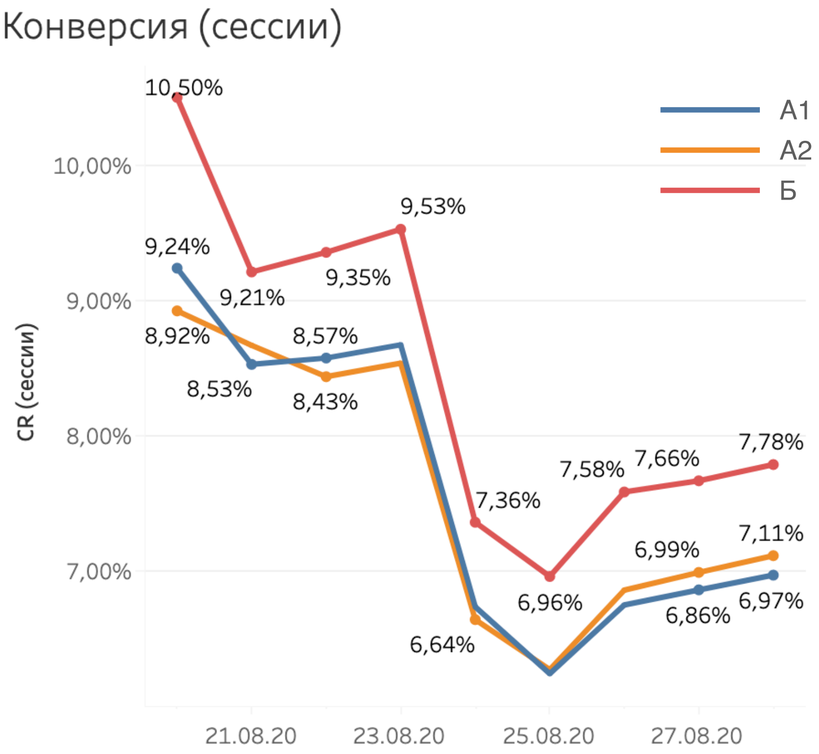
В A/A-тестах мы хотим принимать нулевую гипотезу

$$H_0 : OEC_{control_1} = OEC_{control_2}$$

а не отвергать ее, проверяя OEC (Overall Evaluation Criterion)

https://exp-platform.com/Documents/2015-08OnlineControlledExperimentsKDDKeynoteNR.pdf

A/A/Б-тест



В A/A/Б-тестах мы хотим принимать нулевую гипотезу в паре A1/A2 и отвергать на A1+A2/Б

Дополнительная контрольная ветка служит страховкой. В случае, если кто-то в компании решит запускать эксперимент с той же целевой метрикой, что у вас (но вы об этом можете не знать), вы будете уверены что все ок

Для чего A/A?

Убедиться в корректности системы сплитования можно путем двухэтапной проверки:

- **Честное деление пользователей между группами**. Сохраняется репрезентативность по долям и дисперсии: сплитовалка не должна отдавать приоритет какой-либо из групп по какому-либо признаку, в силу чего может произойти дисбаланс -> изменение дисперсии и средних
- **Проверка FPR с помощью бизнес-метрик.** Частота ложноположительных результатов при про проверкеметрики (например, конверсия и средний чек) не должна быть выше заданного уровня α

Этапы проверки A/A с помощью синтетики

1. **Проводим A/A тест.** Время на A/A определяется таким образом, чтобы охватить как можно больше факторов влияния на метрику (например, недельная сезонность)
2. **Симулируем новые A/A.** Тест пересчитывается ≥ 10 тыс. раз при помощи симуляции новых «синтетических» A/A
3. **Считаем стат. значимость.** В каждом тесте считается p-value при помощи статистического оценщика (бутстрап, т-тест и т.п.)
4. **Считаем метрику качества *FPR (False Positive Rate)***
5. **Делаем выводы.** Проверяется условие $FPR < \alpha$, и если условие соблюдается, то сплитовалка работает корректно

Показатель *FPR*

Для проверки качества сплитовалки считаем долю ложно положительных оценок (*FPR*): $\frac{FP}{N} = \frac{FP}{FP + TN} = \frac{FP}{N_{sim}}$

FP – False Positive или $I\{P \leq \alpha\}$, I – индикаторная функция, P – полученные p-value на каждой итерации синтетического теста, α – уровень альфа

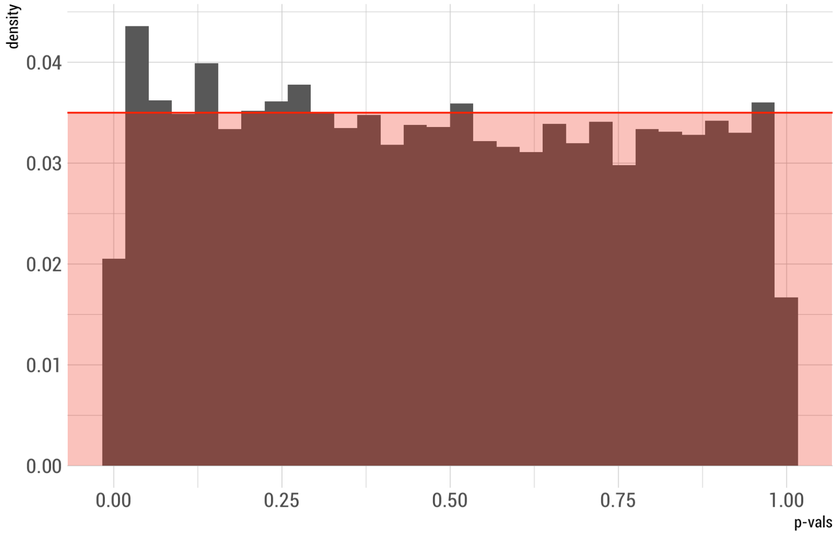
TN – True Negative или $I\{P > \alpha\}$,

N_{sim} – количество повторных экспериментов

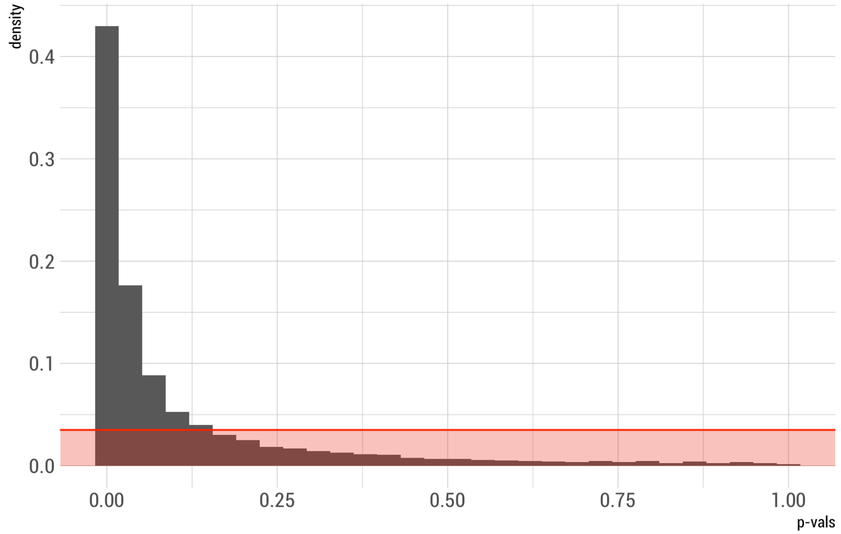
По сути, необходимо проверять *FPR* на каждом уровне значимости: частота ложных прокрасов не должна быть выше заданного уровня значимости. *FPR* не должен превышать 0.05 для α = 0.05. Соответственно и для 0.01, 0.005 и т.п.

Показатель FPR

Корректная сплит-система



Сломанная сплит-система



Красная закрашенная область – uniform теоретическое распределение α.
Если бины выше или ниже красной линии, то что-то не так и нужно искать причины.

Завышенный FPR

Техническая реализация

Основные причины кроются в сломанном сплит-алгоритме.
Причины необходимо искать на стороне где реализован скрипт и его запуск. Частые кейсы:

- Долгое ожидание ответа сервера по присвоению id эксперимента и сплита
- Приоритет той или иной группе
- Не на всех страницах / кейсах реализован сплит-алгоритм
- Банально «сломан» рандом (остаток от деления по сумме хеша?)

Поиск возможной причины

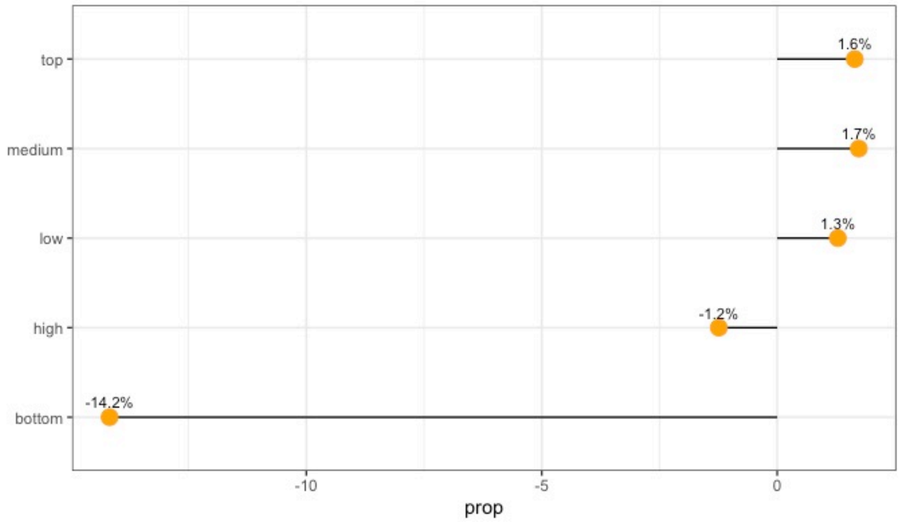
Дисбаланс в группах по описательным признакам.

Первая возможная причина нарушения условия $FPR < \alpha$.

Для поиска дисбаланса необходимо сравнить распределившиеся доли между группами по признакам. Вполне подойдут:

- регионы
- источники трафика
- браузер и т.п.

Сравнение долей RFM сегментов по 2 сплитам (должно быть 0% или незначительное отклонение)



Поиск возможной причины

Критерий Кохрана-Мантеля-Ханзеля для проверки дисбаланса

Для проверки фактических долей с их теоретическим равномерным распределением используются специализированные критерии согласия.

В ситуации с A/A подойдет критерий CMH (Cochran–Mantel–Haenszel) для проверки таблиц сопряженности 2 x 2 x K,

где K – количество градаций по анализируемому признаку (например браузер 1, браузер 2 и т.п.)

В питоне `statsmodels.stats.contingency_tables.StratifiedTable`

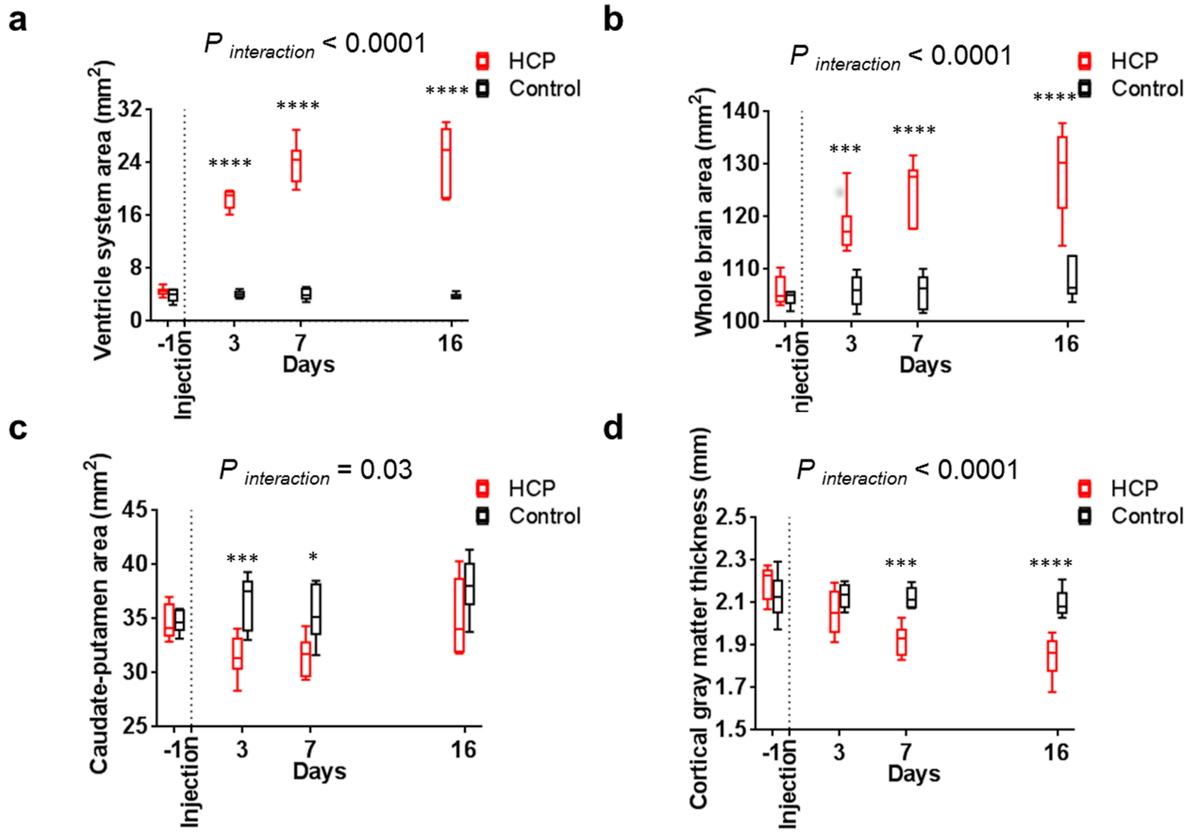
Поиск возможной причины

Сильное отличие конверсии внутри группировок

Вторая возможная причина нарушения условия $FPR < \alpha$.

Для поиска причины необходимо сравнить конверсию внутри градаций между контролем и тестом:

- Проверка pvalue на уровне альфы
- Дополнительная проверка FPR на уровне альфы (опционально)



Изображение https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0148652.g003

Ограничения и другие моменты

- A/A желательно проводить как можно дольше, чтобы достичь достаточной репрезентативности (охватить недельную сезонность и разные группы пользователей)
- В случае, если нет возможности ждать, то не рекомендуется использовать долгоиграющие метрики для проверки сплита (например, C2)
- Пост-симуляции нужно делать без возвращения наблюдений в сплитах
- Для пост-симуляций лучшим образом подойдет бутстрап, благодаря своей точности

Где еще применяются A/A-тестирования?

- A/A/B для контроля вмешательства других (параллельных) экспериментов
- Подбор релевантных групп для сравнения между собой (например, поиск близких регионов)
- Симуляции для проверки статистического оценщика (например, чтобы проверить мощность для t-теста при разных treatment эффектах)

e^xperiment $f$est

Мирмахмадов Искандер