KARPOV.COURSES >>>
КОНСПЕКТ

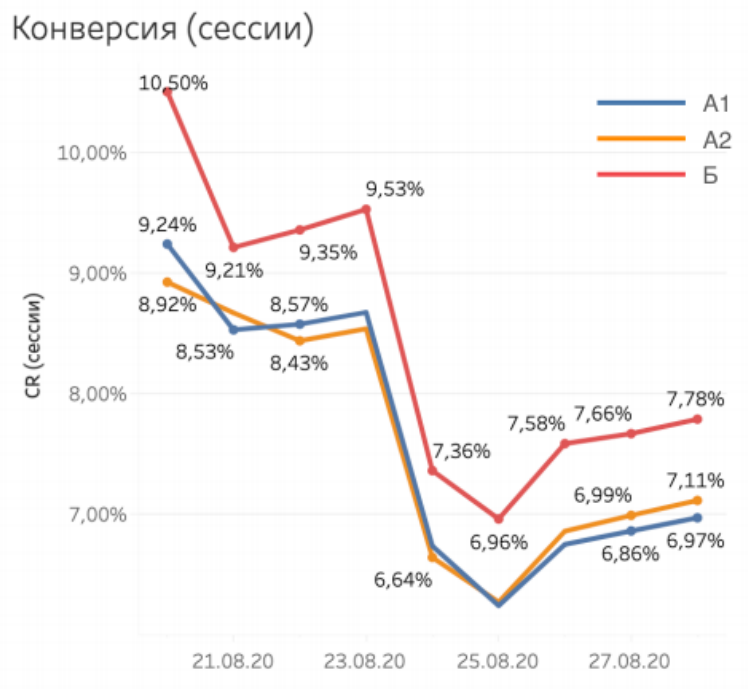
> Конспект > 10 урок > СТАТИСТИКА

> Оглавление

> Для чего A/A?

Преимущественно, задача A/A тестов заключается в том, чтобы понять, работает ли система сплитования корректно или нет.

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).



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.

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 ложных прокрасов метрики (например, конверсия и средний чек) не должна быть выше заданного уровня α

> Этапы проверки A/A

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 < α, и если условие соблюдается, то сплитовалка работает корректно

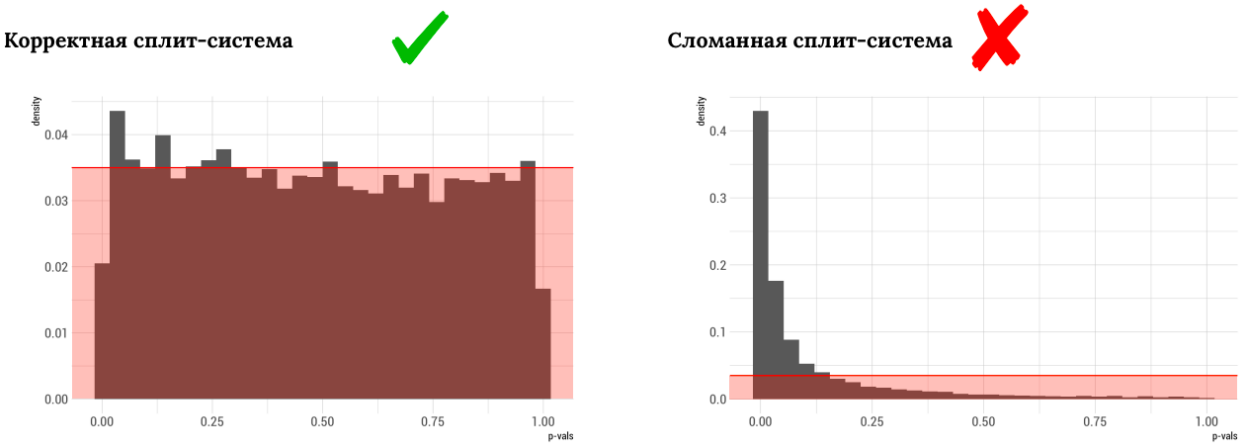
> Показатель FPR

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.

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



Красная закрашенная область – однородное 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:

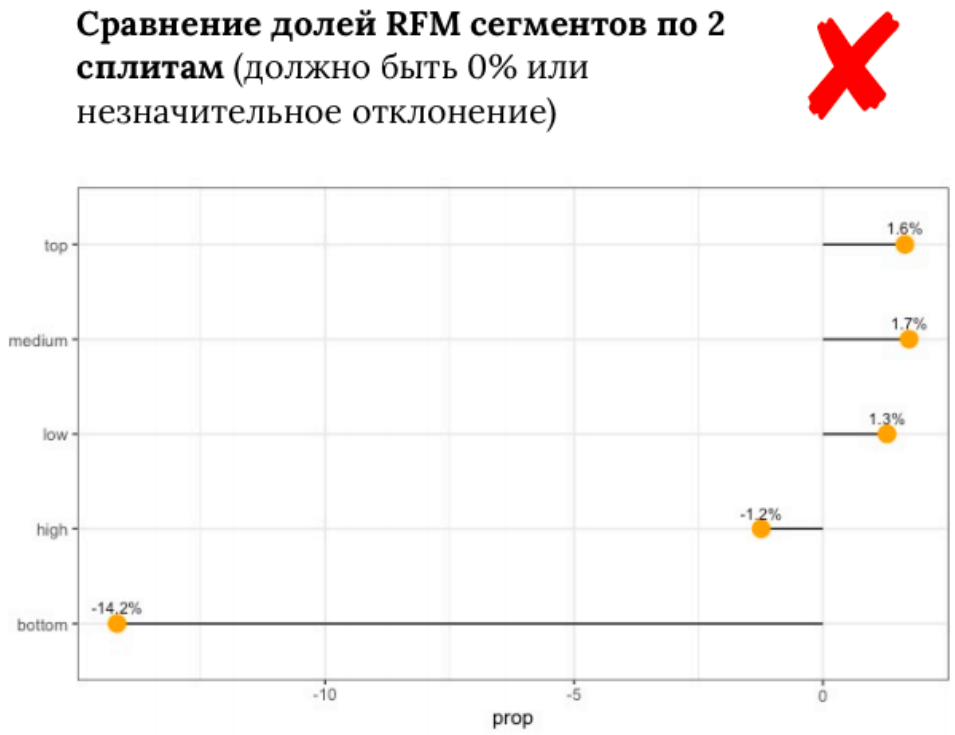
- Долгое ожидание ответа сервера по присвоению id эксперимента и сплита
- Приоритет той или иной группе
- Не на всех страницах / кейсах реализован сплит-алгоритм
- Банально «сломан» рандом (остаток от деления по сумме хеша?)

> Поиск возможной причины

Дисбаланс в группах по описательным признакам.

Первая возможная причина нарушения условия FPR < α. Для поиска дисбаланса необходимо сравнить распределившиеся доли между группами по признакам. Вполне подойдут:

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



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 python – `statsmodels.stats.contingency_tables.StratifiedTable`

Сильное отличие конверсии внутри группировок

Вторая возможная причина нарушения условия FPR < α. Для поиска причины необходимо сравнить конверсию внутри градаций между контролем и тестом:

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

> Как пользоваться StratifiedTable

Как пользоваться StratifiedTable?

По своей сути стратифицированные таблицы представляют собой разновидность т.н. **таблиц сопряжённости -** методов проверки взаимосвязи между двумя номинативными переменными. Называются они так потому, что частоты комбинаций между двумя переменными удобно укладываются в табличный формат, и именно распределение этих частот мы и анализируем. Например, мы хотим понять, влияет ли пол человека на выбор животного. У нас есть переменная пола (мужчина-женщина) и переменная вида животного (кошка-собака). Их таблица сопряжённости будет выглядеть вот так:

	Dog	Cat	Total
Male	42	10	52
Female	9	39	48
Total	51	49	100

Соответственно, через распределение этих частот встречаемости мы можем заключить, есть эффект или нет. Здесь, например, явно видно, что мужчины чаще заводят собак, а женщины кошек - однако обычно вывод совершается с помощью формальных методов вроде **теста хи-квадрат Пирсона** или **точного теста Фишера**, которые дают уже знакомое нам p-значение.

Стратифицированные таблицы вносят третье измерение в процесс. Представьте, что мы проверяем ту же гипотезу о связи пола и вида животного, однако нам интересно: а одинакова ли эта взаимосвязь в разных странах? Соответственно, страна здесь будет стратифицирующей переменной, и у нас сразу две нулевые гипотезы (если грубо):

- распределение частот значений переменных пола человека и вида животного равномерно (переменные независимы друг от друга)
- это распределение одинаково во всех странах

Использовать StratifiedTable можно через метод `from_data()`:

```
from statsmodels.stats.contingency_tables import StratifiedTable

StratifiedTable.from_data('первая_переменная', 'вторая_переменная', 'стратифицирующая_переменная', данные).summary()
```

Получится вывод, похожий на такой:

```
                    Estimate    LCB     UCB
-----------------------------------------
Pooled odds            2.174    1.984 2.383
Pooled log odds        0.777    0.685 0.868
Pooled risk ratio      1.519

                  Statistic P-value
-------------------------------------
Test of OR=1        280.138   0.000
Test constant OR      5.200   0.636

-----------------------
Number of tables     8
Min n              213
Max n             2900
Avg n             1052
Total n           8419
-----------------------
```

Здесь нас интересует p-value. Первая строчка (где *Test of OR = 1*) - это проверка гипотезы о независимости переменных, вторая строчка (где *Test constant OR*) - проверка гипотезы об одинаковости связи по всем уровням стратифицирующей переменной. Так, в этой таблице видно, что между переменными есть связь, и эта связь одинакова во всех стратах.

> 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, CTR)
- 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)

> Слайды и дополнительные материалы

- Слайды с лекции
- Статья на medium