ПОЧЕМУ УРОК ВАЖЕН?

Привет!

Сегодня мы разберемся с теоретической частью A/B тестирования. Даже если мы видим, что конверсия в варианте нашего эксперимента выше, чем в базе, то сразу сделать вывод о том, что вариант действительно лучше – нельзя. Ведь так могло получиться просто по случайности.

Представьте себе, что вы запустили A/A эксперимент: в обоих вариантах пользователи получают совершенно одинаковый опыт. Результат в группах A1 и A2 все равно получится разный, ведь каждая выборка случайна по природе. Чтобы определить, можно ли объяснить разницу в результатах случайностью, или нужно сделать вывод, что изменения которые мы внесли действительно увеличили конверсию, нужно применить статистический тест.

Статистический тест – это очень полезный инструмент, но даже он не может дать однозначно правильный ответ на вопрос, отличаются ли метрики выборок A и B. Чтобы пользоваться тестом эффективно нужно понимать, как с помощью уровня значимости и мощности теста можно определить, какова вероятность получить ложноположительный или ложноотрицательный результат.

СЕГОДНЯ МЫ ИЗУЧИМ

**Структура статистического теста.**Видео – 31 минута.

* Матрица ошибок теста.
* Структура теста.

**Воркшоп.** Видео – 23 минуты.

* Рассчитываем необходимый размер выборки.

СТРУКТУРА СТАТИСТИЧЕСКОГО ТЕСТА1/3

Статистический тест позволяет нам оценивать насколько вероятно, что результат нашего эксперимента – это случайность, а не следствие изменений продукта. В то же время сам тест тоже имеет вероятностную природу: каждый раз когда вы его делаете у вас есть шанс сделать неправильный вывод.

Если вы делаете много экспериментов, то количество ложных результатов будет определяться уровнем значимости и мощностью теста, которые вы устанавливаете. Но, к сожалению, чем реже вы согласны допускать ошибки, тем реже вам удастся сделать определенные выводы из теста, поэтому экспериментаторам приходится поддерживать баланс между точностью и скоростью.

Мощность теста зависит в том числе и от размера эффекта, который вам необходимо обнаружить. Если в эксперименте вам необходимо выяснить, возросла ли конверсия на 30%, то достаточно будет совсем небольшого числа наблюдений. Если же вам нужно зарегистрировать изменение конверсии на 1%, то выборка потребуется в десятки раз больше.

ВОРКШОП2/3

Лучше всего важные соотношения между уровнем доверия, статистической мощностью, размером выборки и силой эффекта можно почувствовать если просто поиграть с калькулятором результатов A/B теста и посмотреть, что меняется.

Давайте на примере онлайн калькулятора научимся рассчитывать необходимый размер выборки и подводить итоги эксперимента.

ПОДВЕДЕМ ИТОГИ3/3

* Есть два параметра тестирования, которые обычно выбираются на долгий срок исходя из стратегических соображений: уровень значимости и расчетная мощность теста.
* Уровень значимости определяет как часто мы готовы допустить ложноположительный вывод, а расчетная мощность – как часто мы готовы смириться с ложноотрицательным.
* Если уровень значимости и расчетную мощность мы зафиксировали, то для любого эксперимента мы можем установить четкое соотношение между размером выборки и минимальным эффектом, который он сможет зарегистрировать.
* Если мы хотим сделать эксперимент более чувствительным (иметь возможность отловить эффект меньшего размера), то нужно увеличить размер выборки. А если мы хотим уменьшить количество наблюдений в выборке (провести эксперимент быстрее или дешевле), то минимальный эффект, который мы сможем отловить, станет больше.

Да пребудет с тобой статистическая мощность, экспериментатор!

ПРЕЗЕНТАЦИЯ

[Презентация урока “Статистический тест для оценки результатов A/B эксперимента”](https://hb.bizmrg.com/productstar_analytics/24.%20%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9%20%D1%82%D0%B5%D1%81%D1%82%20%D0%B4%D0%BB%D1%8F%20%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B8%20%D1%80%D0%B5%D0%B7%D1%83%D0%BB%D1%8C%D1%82%D0%B0%D1%82%D0%BE%D0%B2%20AB%20%D1%8D%D0%BA%D1%81%D0%BF%D0%B5%D1%80%D0%B8%D0%BC%D0%B5%D0%BD%D1%82%D0%B0/Product%20Star_%20%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9%20%D1%82%D0%B5%D1%81%D1%82%20%D0%B4%D0%BB%D1%8F%20%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B8%20%D1%80%D0%B5%D0%B7%D1%83%D0%BB%D1%8C%D1%82%D0%B0%D1%82%D0%BE%D0%B2%20A_B%20%D1%8D%D0%BA%D1%81%D0%BF%D0%B5%D1%80%D0%B8%D0%BC%D0%B5%D0%BD%D1%82%D0%B0.pdf)

ИНСТРУМЕНТЫ

1. [Калькулятор для расчета размера выборки](https://abtestguide.com/abtestsize/)
2. [Калькулятор для расчета результатов теста](https://abtestguide.com/calc/)
3. [Полный курс от Андрея Менде «Управление продуктом на основе данных»](https://productdo.it/data_driven)
4. [Полный курс от Андрея Менде «A/B тестирование на практике»](https://productdo.it/ab_testing)

СТАТЬИ

[Statistical Significance in A/B Testing – a Complete Guide](http://blog.analytics-toolkit.com/2017/statistical-significance-ab-testing-complete-guide/)

ВИДЕО

[A conceptual introduction to power and sample size calculations](https://www.youtube.com/watch?v=QBONLUp7i28)

#### Домашнее задание

Сегодня тебя ждет необычное домашнее задание: тест в формате квиза, жми кнопку "Далее".

# Вводные данные: Команда, в который ты работаешь, установила следующие стандарты экспериментирования и всегда их придерживается: Уровень доверия 90%. Мощность эксперимента 80%. Односторонняя статистическая гипотеза. Длительность эксперимента всегда округляется вверх до целой недели ([Калькулятор для расчета размера выборки](https://abtestguide.com/abtestsize/), [Калькулятор для расчета результатов теста](https://abtestguide.com/calc/)). На лэндинг страницу сервиса приходит 5000 человек в неделю и они делают примерно 400 заказов. Количество посетителей будет сохраняться на этом же уровне в ближайшие недели. Твоя группа запускает промо-акцию, которая будет окупаться только если она позволит поднять количество заказов до 450 в неделю. Чтобы проверить это, промо-акция будет запускаться через A/B тест. Сколько недель должен длиться эксперимент?

# 1 из 3

###### 1.

Одной недели достаточно

###### 2.

Нужно как минимум две недели

###### 3.

Нужно как минимум три недели

Да, все верно!

###### 4.

Не меньше четырех недель

# Вводные данные: Команда, в который ты работаешь, установила следующие стандарты экспериментирования и всегда их придерживается: Уровень доверия 90%. Мощность эксперимента 80%. Односторонняя статистическая гипотеза. Длительность эксперимента всегда округляется вверх до целой недели ([Калькулятор для расчета размера выборки](https://abtestguide.com/abtestsize/), [Калькулятор для расчета результатов теста](https://abtestguide.com/calc/)). Результаты проведенного эксперимента получились следующие: A – 7510 посетителей, 620 покупок. B – 7495 посетителей, 665 покупок. Директор по маркетингу замечает, что конверсия выросла даже в базе эксперимента и высказывает сомнения, что промо-акция вообще что-то дала. Он говорит, что можно продолжить экспериментировать, если промо-акция дает прирост конверсии хотя бы на 7%. Вы рекомендуете продолжать эксперимент?

# 2 из 3

###### 1.

Конечно промо-акция эффективна, ведь конверсия в варианте выросла больше, чем в базе

###### 2.

Из наблюдений в эксперименте можно утверждать, что конверсия выросла как минимум на 7,47%

###### 3.

Из наблюдений в эксперименте можно утверждать, что конверсия выросла как минимум на 8.87%

Нет, конверсия в эксперименте выросла на 7,47%, а 8.87% – это ее абсолютное значение, а не относительное изменение.

###### 4.

Из наблюдений в эксперименте, к сожалению, нельзя сделать вывод о том, что промо-акция дала статистически значимое изменение конверсии

# Вводные данные: Команда, в который ты работаешь, установила следующие стандарты экспериментирования и всегда их придерживается: Уровень доверия 90%. Мощность эксперимента 80%. Односторонняя статистическая гипотеза. Длительность эксперимента всегда округляется вверх до целой недели ([Калькулятор для расчета размера выборки](https://abtestguide.com/abtestsize/), [Калькулятор для расчета результатов теста](https://abtestguide.com/calc/)). Результаты проведенного эксперимента получились следующие: A – 7510 посетителей, 620 покупок. B – 7495 посетителей, 665 покупок. Твой коллега из другой компании, где принято использовать уровень значимости (confidence level) = 95%, говорит, что у них этот эксперимент пришлось бы признать незначимым. Он прав?

# 3 из 3

###### 1.

Нет, уровень значимости не влияет на результат эксперимента

###### 2.

Нет, результат значим даже при 95% уровне значимости

###### 3.

Да, при уровне значимости 95% мы бы не смогли сделать вывод о том, что мы наблюдаем реальное увеличение конверсии

Да, верно! Можно пересчитать результат с другим уровнем значимости или просто посмотреть на расчетную p-value: она больше 5%.

# РЕЗУЛЬТАТ ЗАДАНИЯ