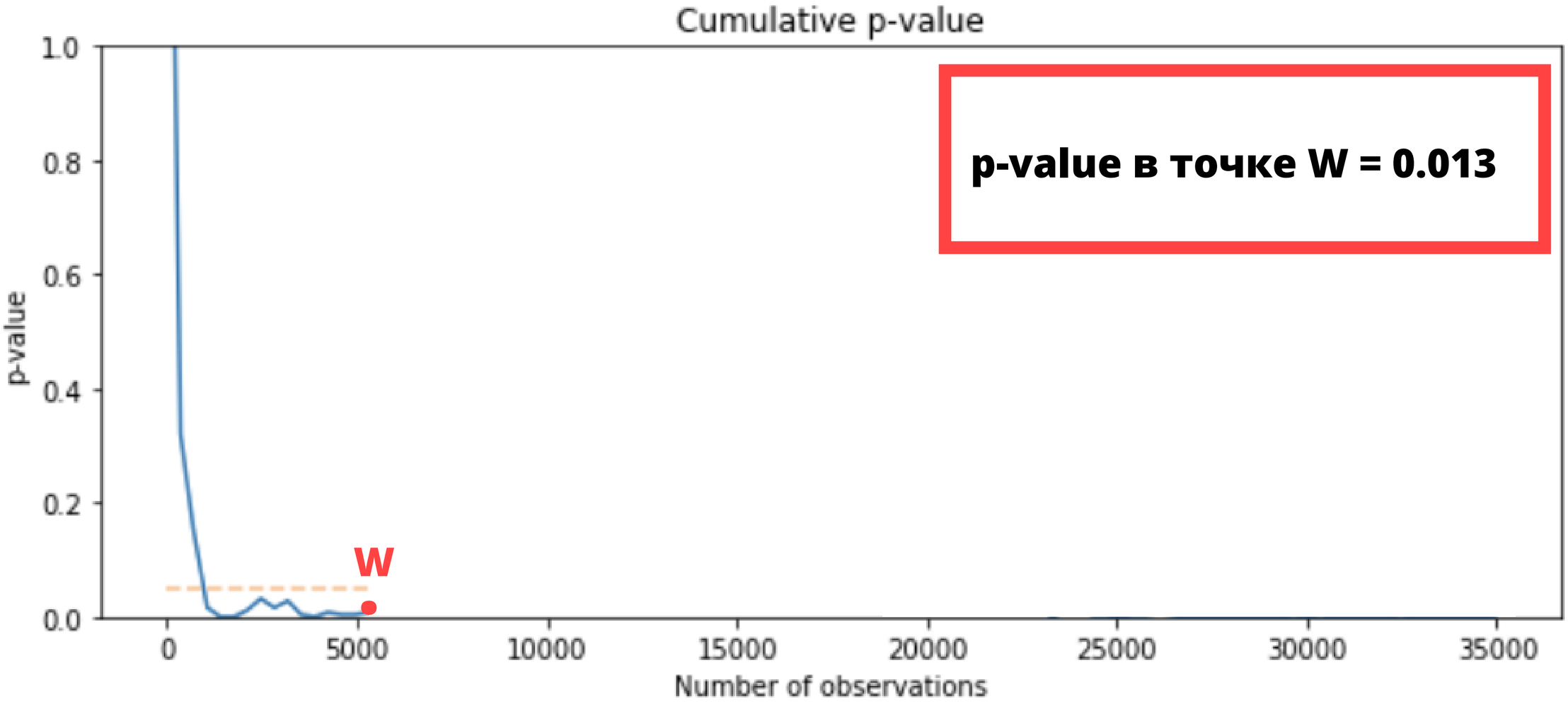
ПОДГЛЯДЫВАНИЯ, BOOTSTRAP И P-VALUE

Продуктовый анализ данных. Семинар 5. Mirzoian Shahane

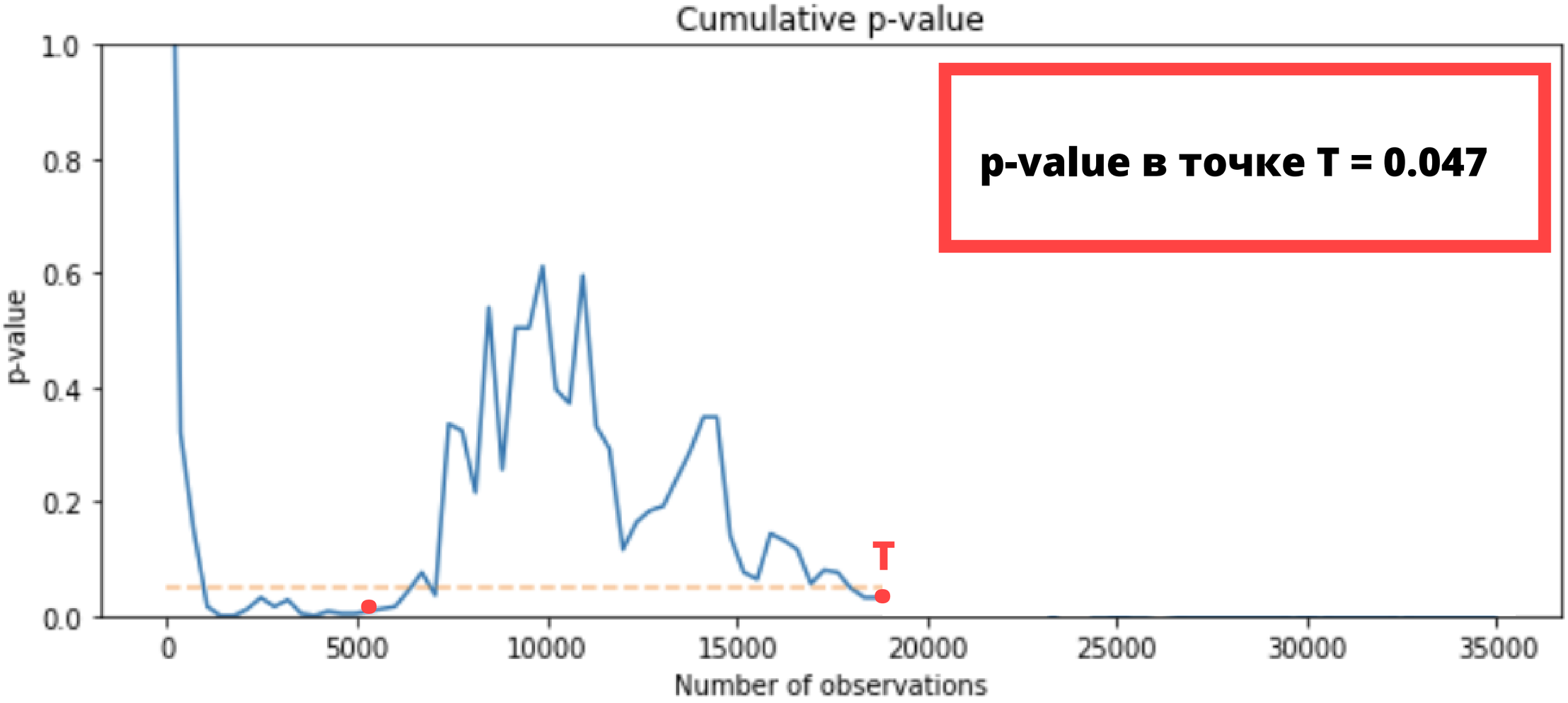
Очень кратко про анализ ab-теста
(рамках frequency analysis)

- Проводим ab-тест, собираем данные про наши группы
- Формулируем **нулевую гипотезу** о том, что группы между собой не отличаются (в зависимости от теста, мы можем проверять равенство средних или то, что выборки взяты из одного и того же распределения)
- Вычисляем **p-value**, то есть вероятность получить наблюдаемую в эксперименте или большую разницу между группами при условии верности нулевой гипотезу
- Если p-value больше **уровня значимости** (наш порог, обычно 0.05), то у нас нет оснований отклонять нулевую гипотезу
- Для окончательного вывода вычисляем **мощность** теста, то есть нашу способность обнаружить отличия там, где они есть

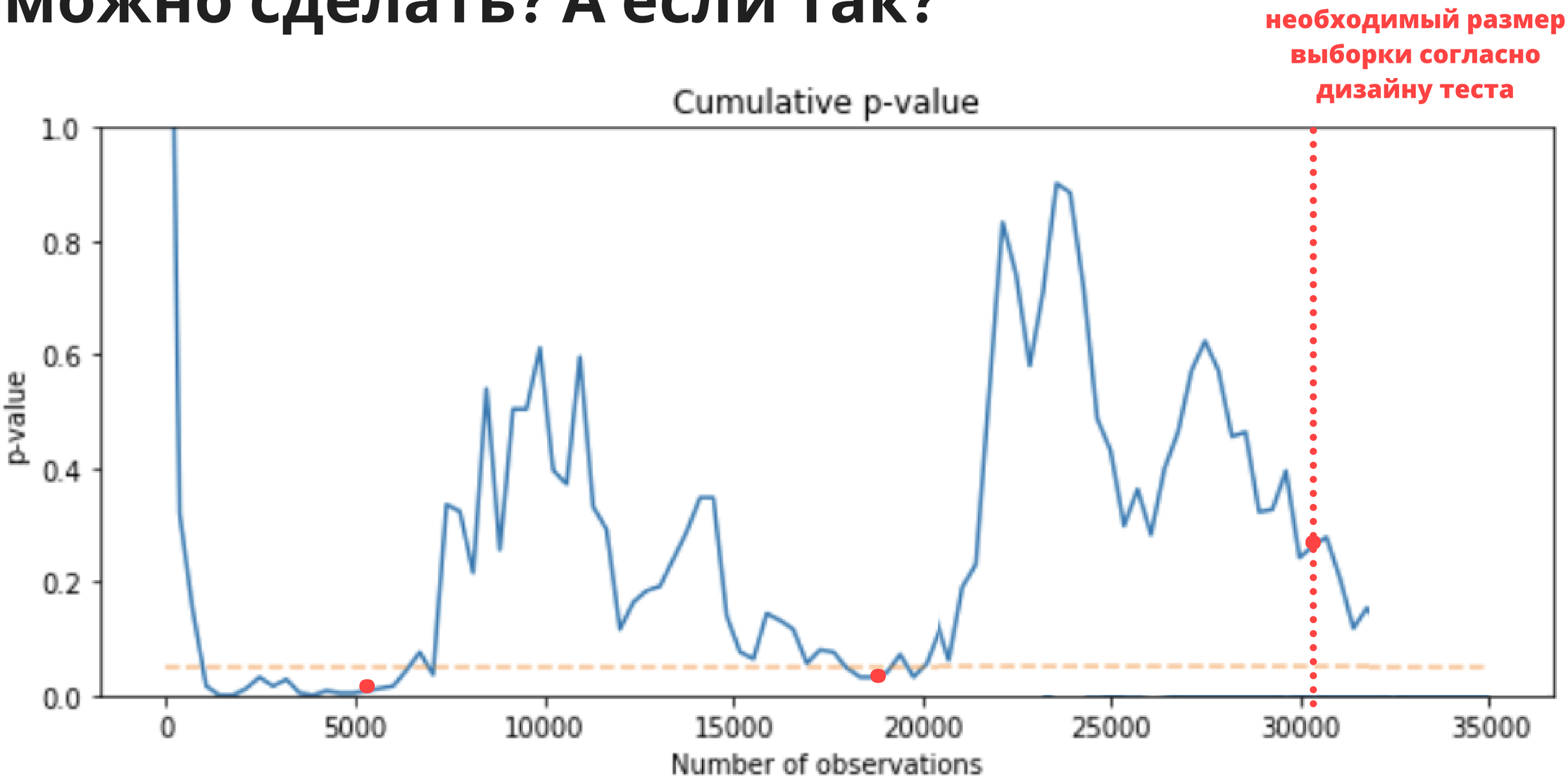
Какой вывод об отличиях между группами можно сделать?



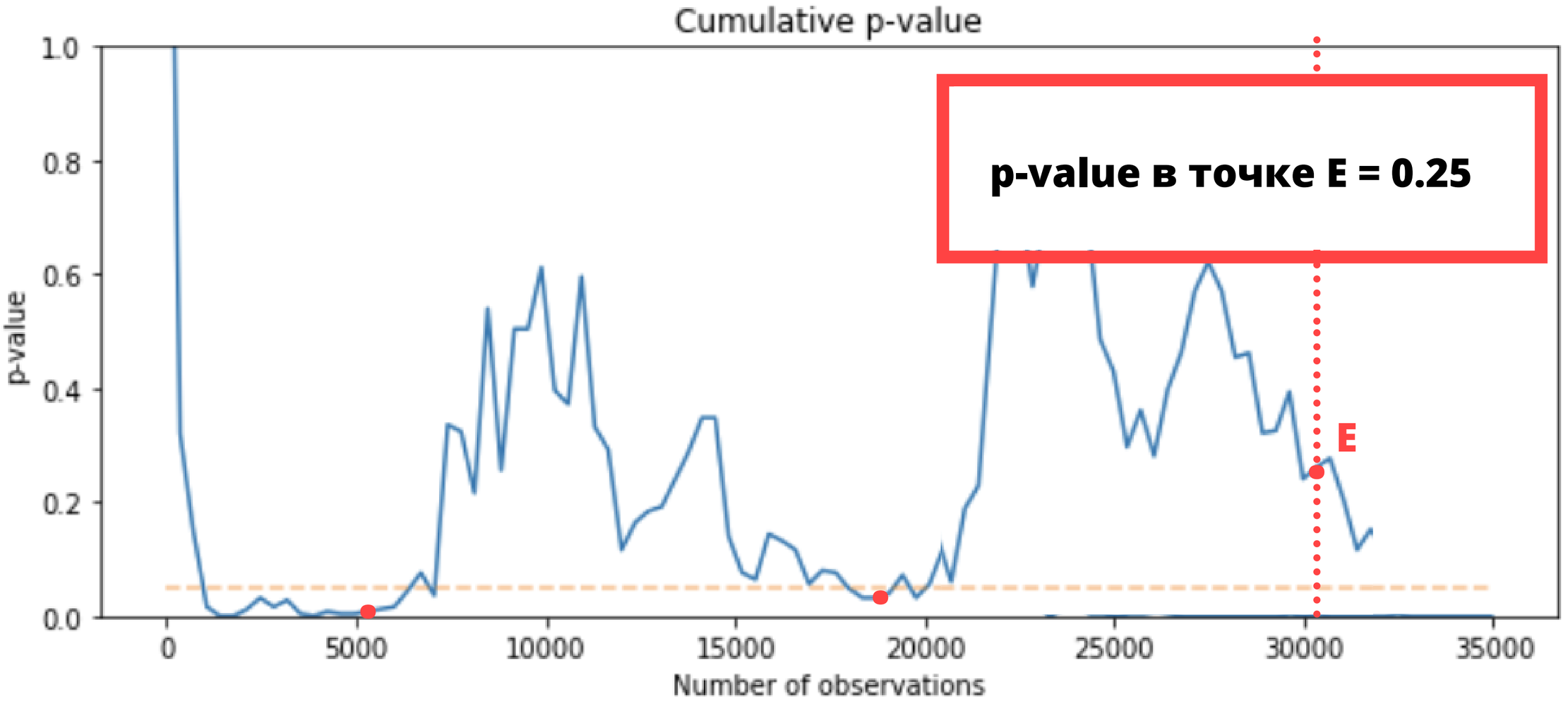
Какой вывод об отличиях между группами можно сделать? А так?



Какой вывод об отличиях между группами можно сделать? А если так?



Какой вывод об отличиях между группами можно сделать?



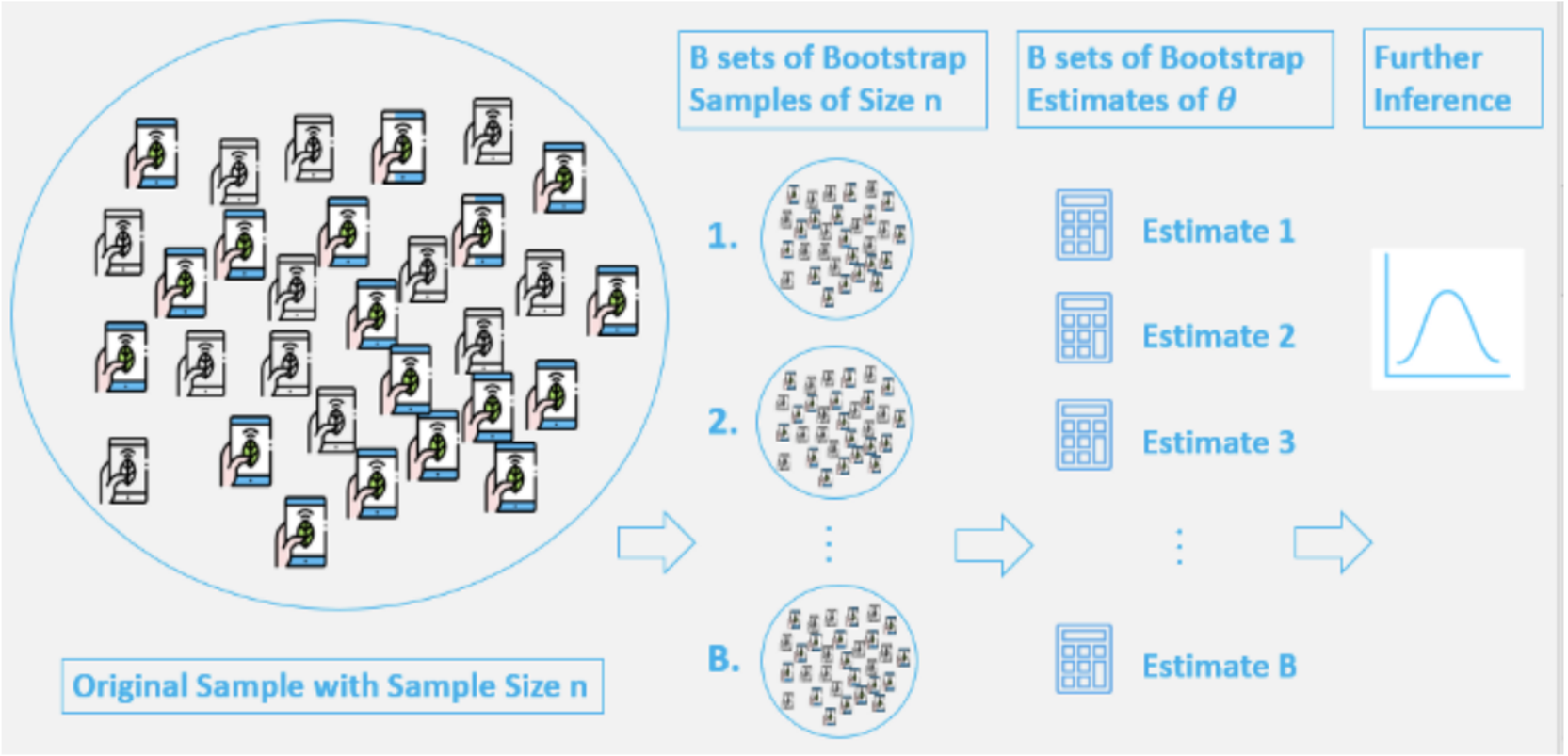
Какой вывод об отличиях между группами можно сделать?

Проблема подглядывания

При правильном процессе A/B тестирования надо **заранее определить количество пользователей**, на основе которых будет оцениваться результат, собрать наблюдения, посчитать результаты и сделать вывод

Даже если две группы идентичны, то **разница может периодически выходить за границы зоны неразличимости** по мере накопления наблюдений. Это совершенно нормально, так как границы сформированы так, чтобы при тестировании одинаковых версий лишь в 95% случаев разница оказывалась в их пределах.

Поговорим о bootstrap?



Чем хорош непараметрический bootsrtap?

- Стат. критерий, который позволяет нам не завязываться на параметры наших выборок (распределение, среднее)
- Позволяет оценить точность наших оценок по выборке (bias, variance, confidence intervals энд соу он)
- Прост в реализации

Степ - бай - степ

- Берем наши исходные выборки (суммы первых оплат тестовой и контрольной группы эксперимента, например)
- Выбираем метрику, которую хотим оценить (среднее, например)
- Для каждой группы выбираем подвыборку в возвращением (т.е. получим 2 "псевдовыборки"
- Считаем на псевдовыборках нашу метрику
- Складываем разность метрик на псевдовыборках в массив diff

Степ - бай - степ

Берем наши исходные выборки (суммы первых оплат тестовой и контрольной группы эксперимента, например)

Выбираем метрику, которую хотим оценить (среднее, например)

повторяем 1000 раз

Для каждой группы выбираем подвыборку в возвращением (т.е. получим 2 "псевдовыборки"

Считаем на псевдовыборках нашу метрику

Складываем разность метрик на псевдовыборках в массив diff

Степ - бай - степ

Берем наши исходные выборки (суммы первых оплат тестовой и контрольной группы эксперимента, например)

Выбираем метрику, которую хотим оценить (среднее, например)

повторяем 1000 раз

Для каждой группы выбираем подвыборку в возвращением (т.е. получим 2 "псевдовыборки"

Считаем на псевдовыборках нашу метрику

Складываем разность метрик на псевдовыборках в массив diff

Дальше можем использовать 2.5-ый и 97.5-ый квантили массива diff для построения доверительного интервала разности метрик или считать p-value по определению

Практика!

- Давайте **сами напишем функции** для bootstrap и отрисовки p-value в динамике
- from scratch, никаких "заготовок" :)

Домашка!

- Пройти статью про подглядывания
- Пройти курс по статистике