Ускорение А/Б тестов

Ильдар Сафило

Senior Machine Learning Researcher, Booking.com

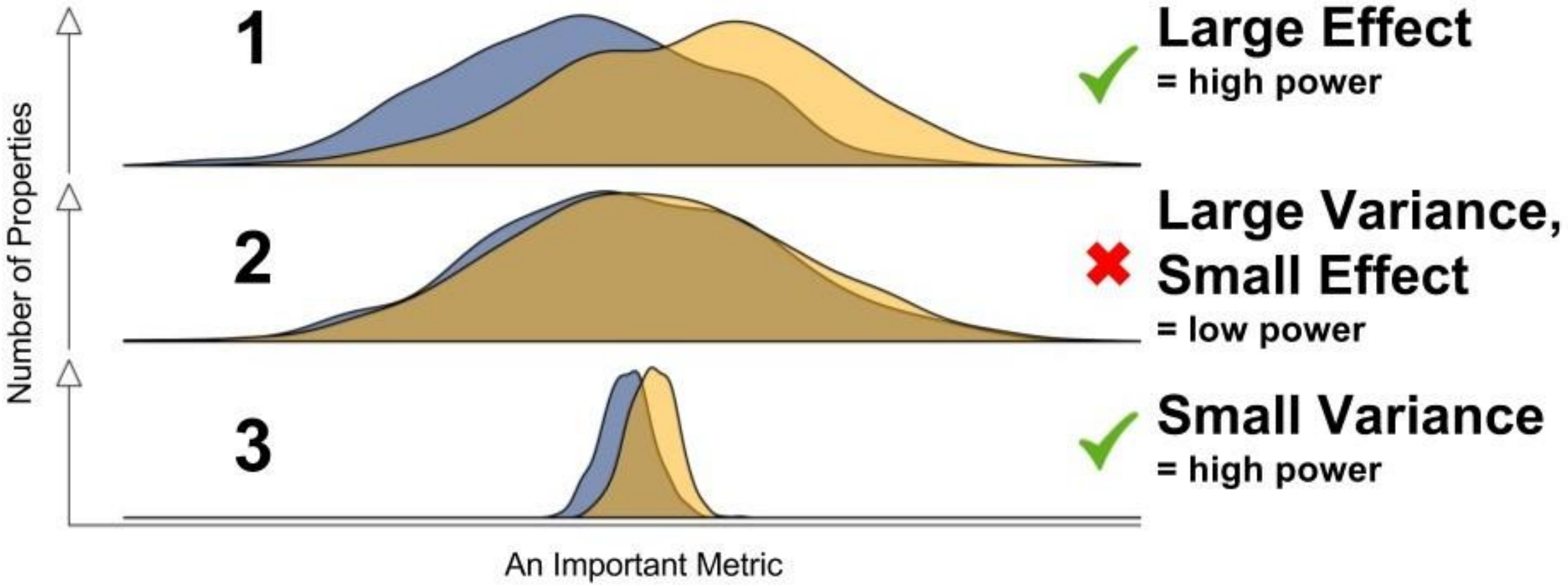
Как делать А/Б тесты, когда это кажется невозможным?

План

1. Методы ускорения А/Б тестов
2. Проблемы с маленькими выборками
3. Проблемы со сплитованием групп
4. Разбор кейса с регионами

Методы ускорения А/Б тестов

Иллюстрация от Booking.com



Как это работает?

Метрика CUPED считается следующим образом:

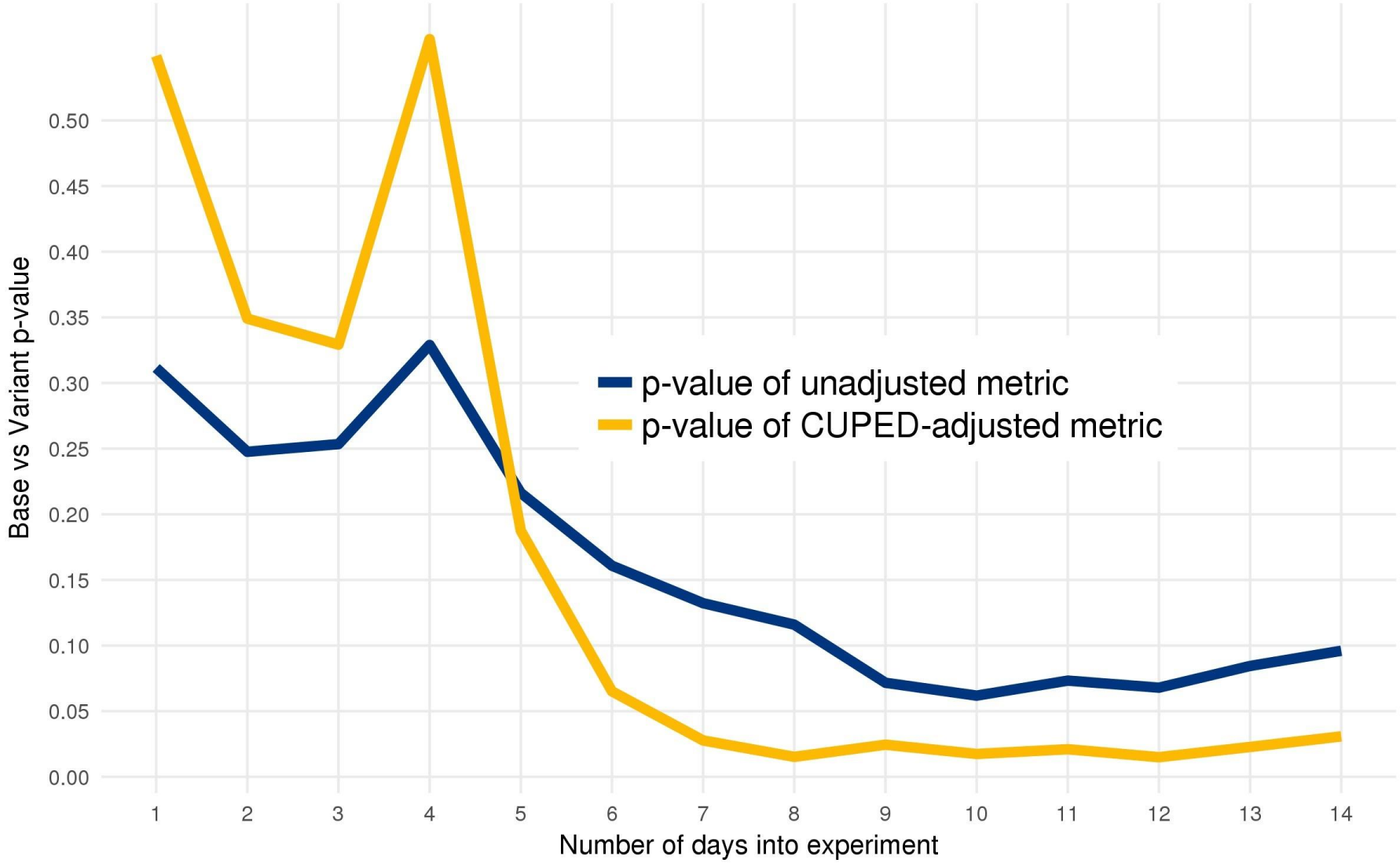
$$CUPED = metric - (covariate - mean(covariate)) * theta$$

- **covariate** — *метрика до эксперимента*
- **metric** — *метрика после эксперимента*
- **theta** вычисляется как

$$\frac{covariance(metric, covariate)}{variance(covariate)}$$

Именно за счет этого куска формулы и меняется дисперсия. Если ковариация большая, то дисперсия сократится значительно.

Пример Booking.com



Почему это работает?

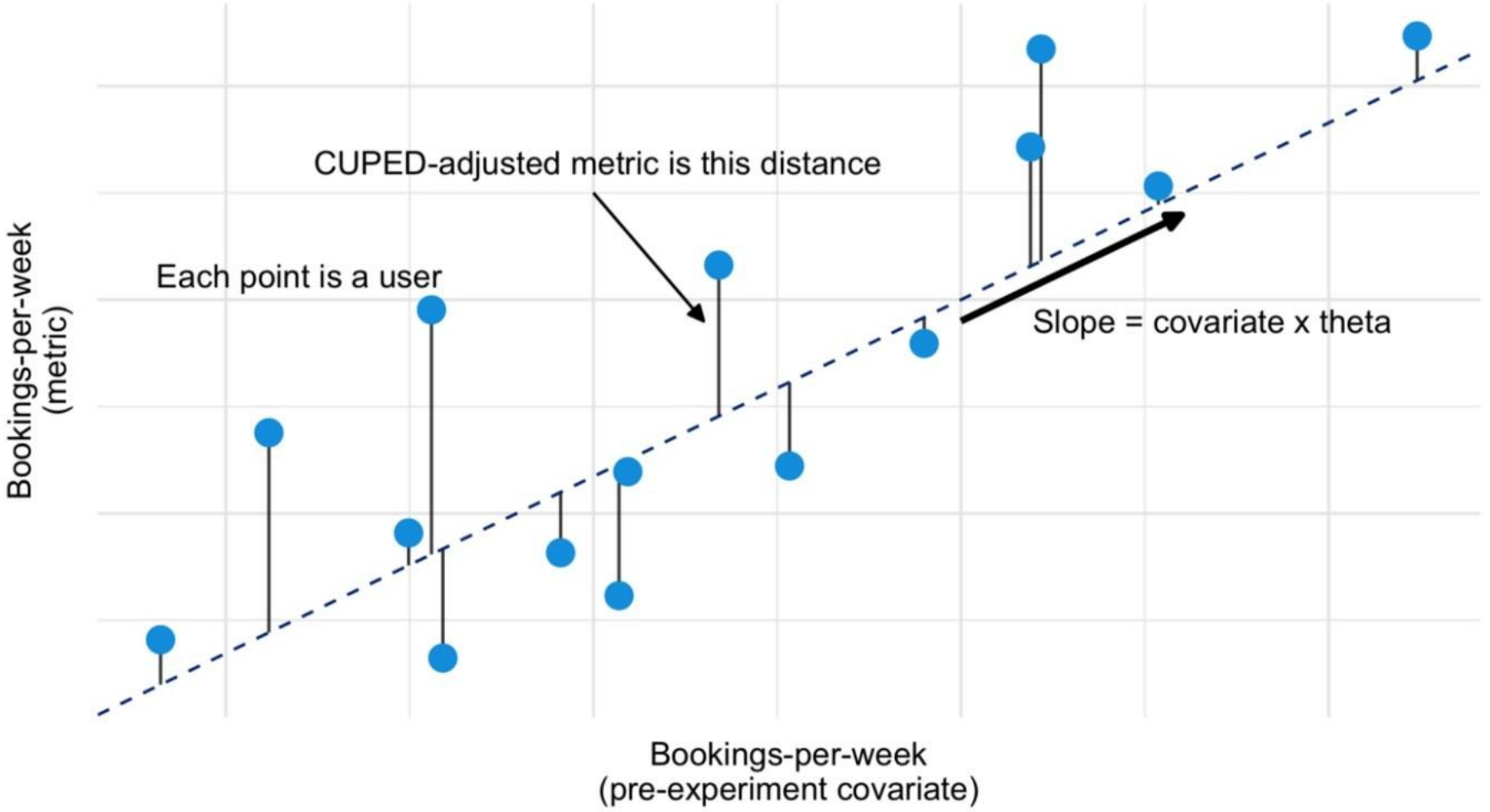


Figure 3. A visual example of how to compute a CUPED-adjusted metric for mean-centered metrics

CUPED

$$Y_{CUPED} = Y - \theta X$$

$$var\,(Y_{CUPED}) = var\,(Y) + \theta^2 var(X) - 2\theta cov(X, Y)$$

Пусть X – некоторая случайная величина, являющаяся пре-экспериментальными данными, которые не зависят от эксперимента. Y – случайная величина, интересующая нас метрика,
θ – некоторый параметр модели, который мы будем подбирать, пытаясь минимизировать дисперсию

CUPED

Оптимальный параметр : $\theta_{min} = cov(X, Y)/var(X)$

Итоговая дисперсия : $var(Y_{CUPED})_{min} = var(Y)(1 - p^2)$

$p = corr(X, Y)$

Как это работает?

$$CUPED = metric - (covariate - E(covariate)) * theta$$

$covariate$ – метрика до эксперимента
$metric$ – метрика после эксперимента

$$= E\left(metric * theta * (\text{covariate} - E(\text{covariate}))\right)$$

$$Var(covariate) = E((covariate - E(covariate))^2)$$

$$E(CUPED) = E(metric) - theta * \left(E(\text{covariate}) - E(\text{covariate})\right) = E(metric)$$

$$E(CUPED^2) = E(metric^2) - theta^2 * E((\text{covariate} - E(\text{covariate}))^2) - 2 * E(metric * theta * (\text{covariate} - E(\text{covariate})))$$

$$Var(CUPED) = E(CUPED^2) - (E(CUPED))^2$$

$$Var(CUPED) = Var(metric) + Var(covariate)\,(theta - \frac{Cov(metric, covariate)}{Var(covariate)})^2 - \frac{Cov(metric, covariate)^2}{Var(covariate)}$$

Оптимальная $theta$: $\frac{covariance(metric, covariate)}{variance(covariate)}$

Что делать с новыми пользователями?

CUPED = metric - (mean(covariate) - mean(covariate)) x theta

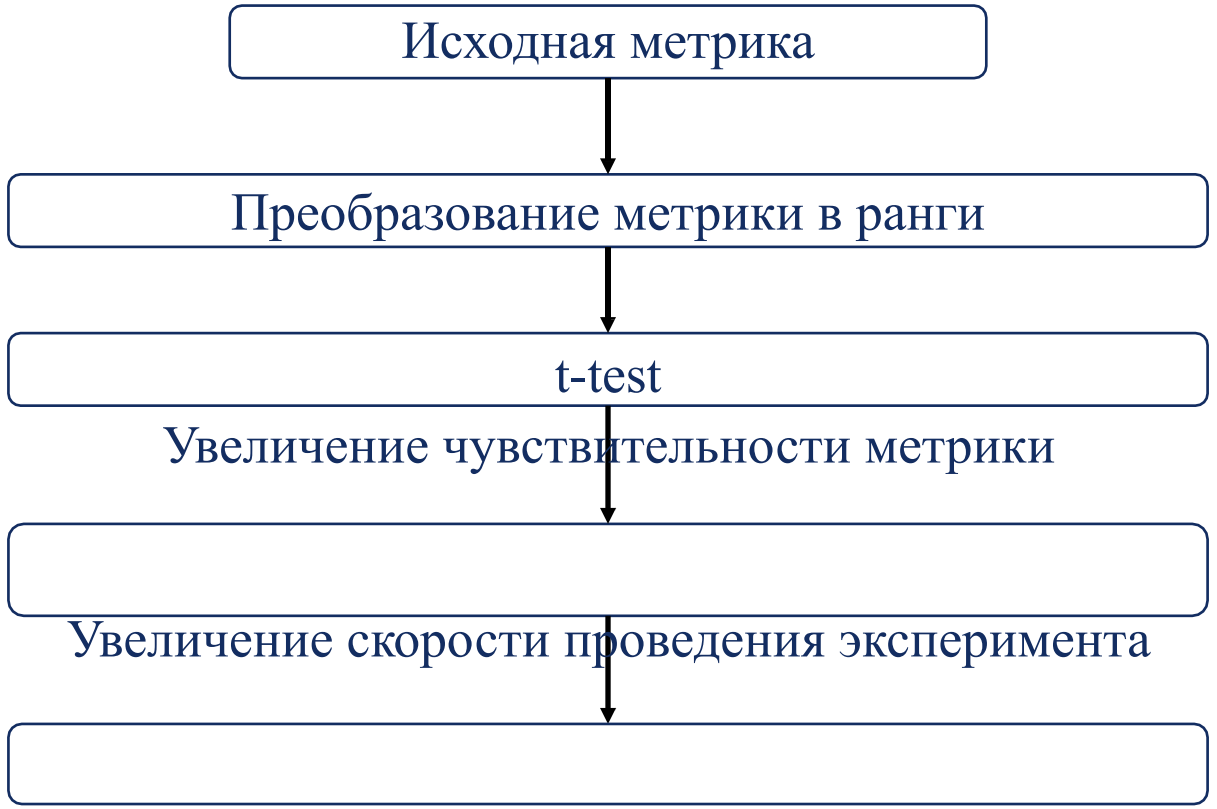
CUPED = metric - (0) x theta

CUPED = metric

Особенности

- Следует применять для непрерывных метрик, которые можно агрегировать по юзерам или клиентам за определенный период времени
- Можем исследовать изменение первоначальной метрики в группах без потери интерпретации
- Период времени, ковариату нужно подбирать по историческим данным
- Тета должна быть единой для групп
- Часто ковариата – метрика на прошлом периоде

Ранговая трансформация



https://booking.ai/increasing-sensitivity-of-experiments-with-the-rank-transformation-draft-c01aff70$_1$b$_3$255

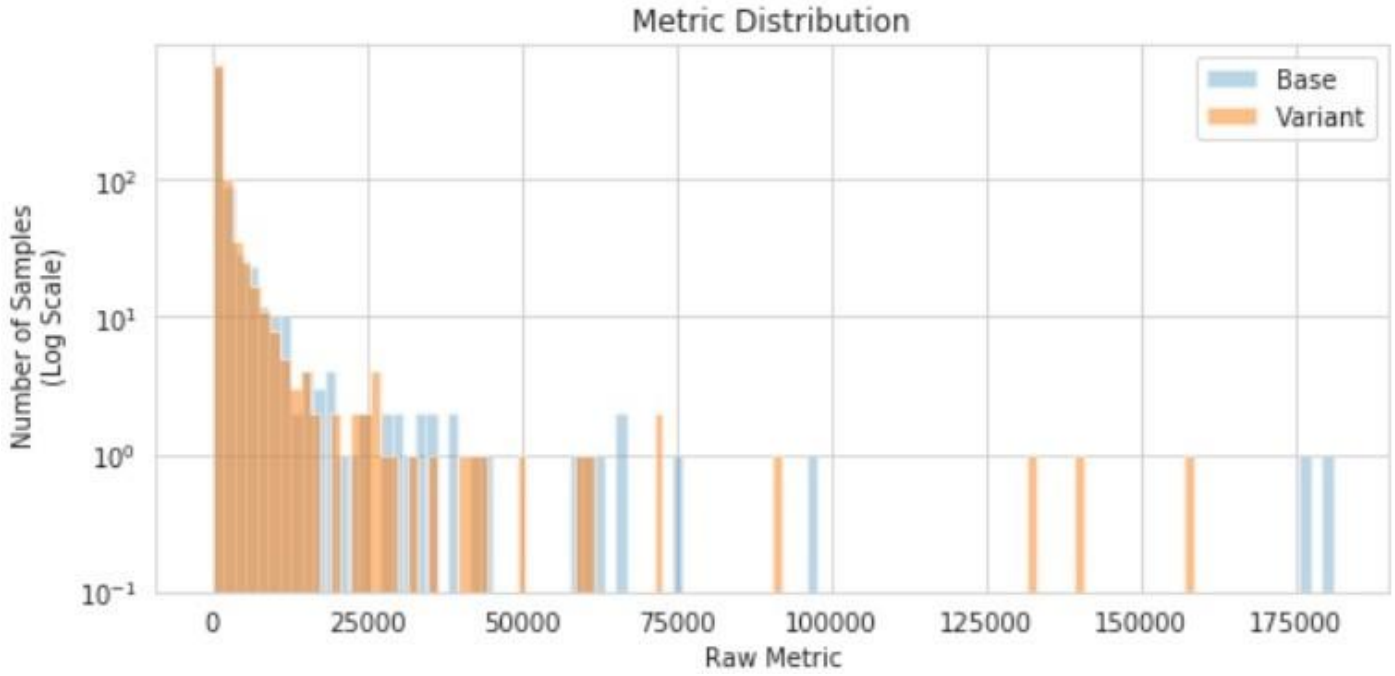
Еще пример Booking.com

Variant	Raw Metric	Rank Transformed Metric
1	0.2	1
1	0.3	2.5
2	0.3	2.5
1	0.4	4
2	2	5
2	100	6

$$M_{\text{rank transformed}} = \frac{2}{N} * rank(M_{\text{raw}}) - \frac{1}{2}$$

Данные из контрольной и тестовой выборки:

- Обьединяются
- Ранжируются
- Одинаковое значение – берем средний ранг соседей



Сравнение методов повышения чувствительности на данных booking.com

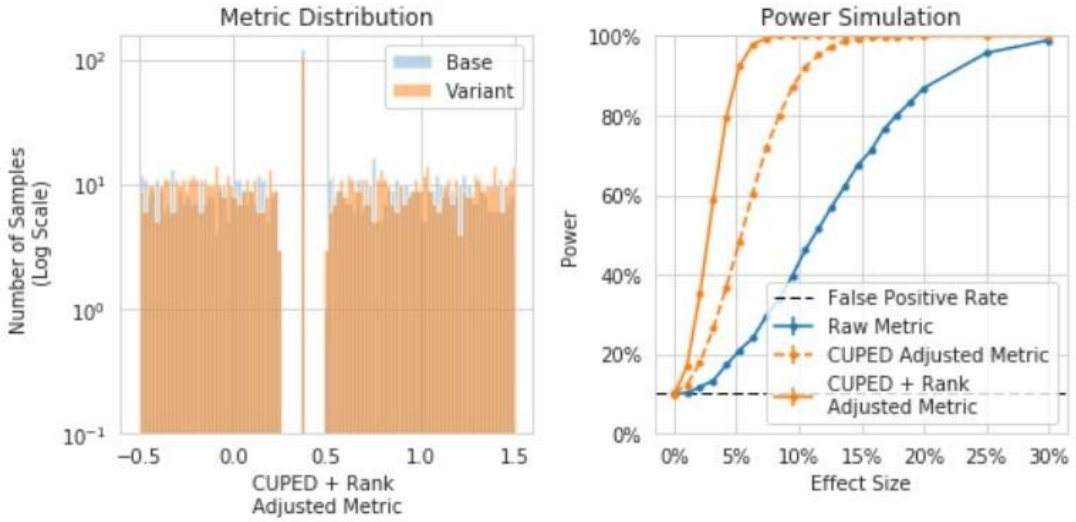


Рис. 8. Распределение метрики после ранговой трансформации (слева) и смоделированная мощность метрики (справа). Ранговая трансформация проводилась после корректировки по CUPED. В результате минимально-обнаруживаемый размер эффекта значительно снизился.

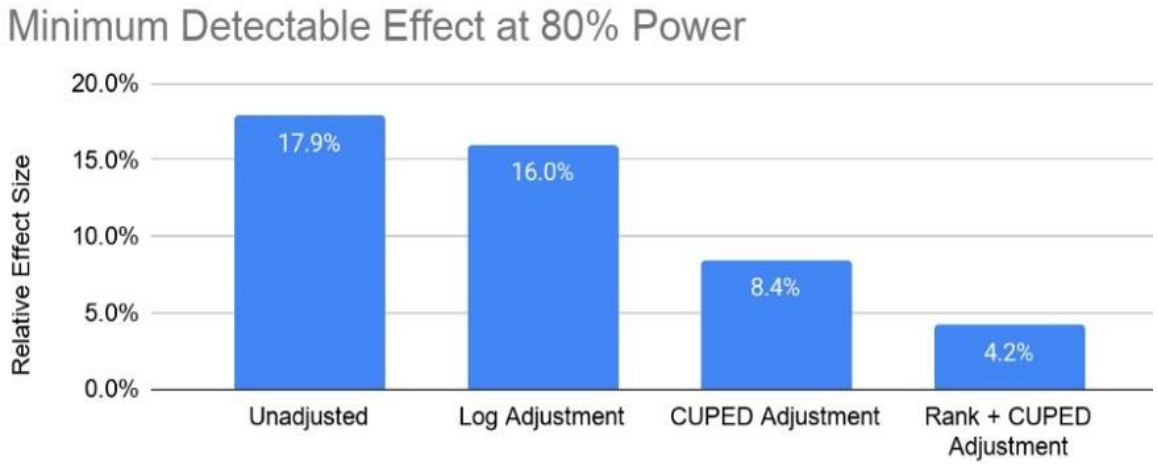


Рис. 9. Минимально-обнаруживаемый эффект при мощности 80% после разных корректировок.

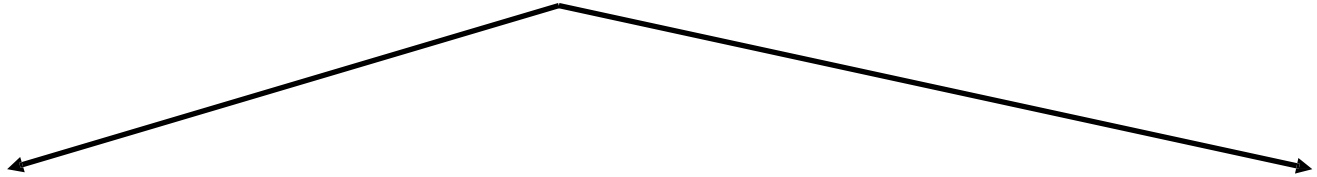
Сравнение с критерием Манна-Уитни

T-test		Манна-Уитни
	Гипотеза	
$H_0 : \overline{A} = \overline{B}$		$H_0 : P(A > B) = P(B > A)$
	Бизнес	
Среднее изменение		**Медиана**
	Статистика	
$t_r = \frac{\frac{S}{n} - \frac{1}{m}(\frac{N(N+1)}{2} - S)}{[(\sum_{i=1}^{N} R_i^2 - \frac{S^2}{n} - \frac{1}{m}(\frac{N(N+1)}{2} - S)^2)\frac{N}{nm(N-2)}]^{1/2}}$		$W = \frac{S - n(N+1)/2}{[\frac{nm}{N(N-1)}\sum_{i=1}^{N} R_i^2 - \frac{nm*(N+1)^2}{4(N-1)}]^{1/2}}$

Что выяснил booking.com?

Подходы примерно одинаковы на больших выборках вне зависимости от величины эффекта и распределения данных.

Пробуем использовать?



Да

- Когда нам больше интересно выловить выявить не среднее изменение, а медиану.
- Когда распределение данных сильно искажено и далеко от нормального

Нет

- Когда нам больше интересно выловить среднее изменение.
- Когда распределение данных близко к нормальному.
- В таких случаях лучше использовать CUPED или другие методы обработки возможных выбросов.

Стратификация

Y - наша метрика , которую мы хотим улучшать.
Также у нас есть некоторые параметры, которые не зависят от наших изменений,
но влияют на нашу метрику.
Например - регион, тип системы телефона, возраст.
Мы можем разбить наши обьекты на K-групп.

$$Y_{strat} = \sum_{k=1}^{K} w_k Y_k \ , w_k \ - \text{вес групп } K \text{(например, вероятность попадания в эту группу)},$$

$$Y_k \ - \text{наша метрика в группе } K$$

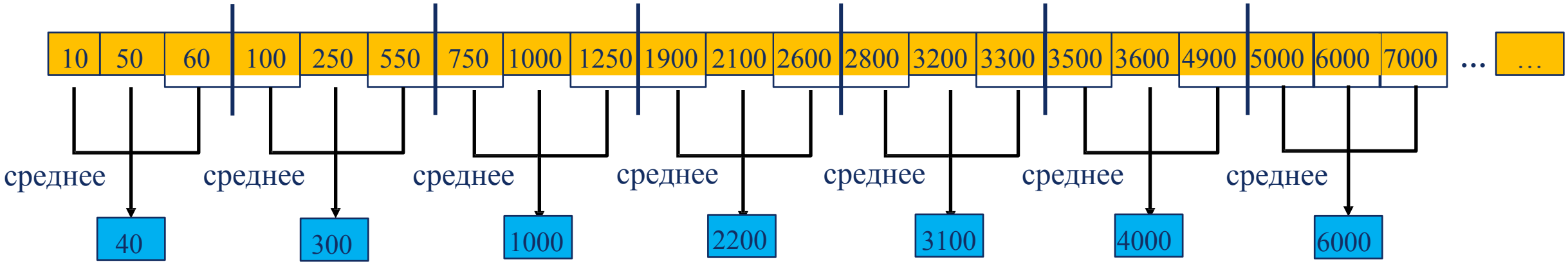
https://www.exp-platform.com/Documents/2013-02-CUPED-ImprovingSensitivityOfControlledExperiments.pdf

Особенности

- Простая реализация
- Нужно подбирать группы для стратификации с помощью признаков, на которые наши изменения при А/Б тестировании не повлияют
- Метод хорошо подходит для маленьких или средних выборок
- Сильный прирост, если текущее сэмплирование групп смещено по каким-либо признакам

Бакетирование

$X = (X_1, X_2, \dots, X_n)$ – **выборка объектов**



Бакетное сэмплирование – обобщение алгоритма

- $X = (X_1, X_2, \dots, X_n)$ – выборка обьектов, metric – наша метрика, которую мы оцениваем
- обьекты делятся на m бакетов
- в каждом бакете считается некоторая статистика – *stat* по наблюдениям, находящимся в бакете
- получается новая выборка $Y = (Y_1, Y_2, \dots, Y_m)$, где $Y_l = stat(X_{i0}, X_{i1}, \dots, X_{il})$, l – кол-во обьектов в бакете i
- К новой выборке применяем наш общий пайплайн A/B тестов

Бакетное сэмплирование – обобщение алгоритма

- Тип бакетов :
  - смысловой - объединение зависимых или коррелирующих объектов в один бакет
  - численный – сортируем объекты и разбиваем на бакеты по возрастающим интервалам

- Параметр - число бакетов

- Параметр - статистика *stat* – медиана, среднее, квантили

Что получаем?

- Можем избавляться от зависимых событий, переходя к независимым
- более аккуратно работаем с выбросами
- распределение больше похоже на нормальное
- Меньше шума в данных
- Можем сравнивать почти любые метрики

Особенности

- Нужно варьировать параметр – размер бакета
- Лучше не применять на небольшом количестве данных
- Группы А и Б должны быть одинаковыми по размеру
- Нужно аккуратно выбирать статистику для бакета

Пример

Мы хотим оценить экономический эффект нашей модели при внедрении ее в мобильное приложение или сайт. Мы можем посчитать для каждого захода пользователя некоторую метрику. Например – среднее время сессии. Как провести А/Б?

Пример

Получается, пользователь – это множество сессий или заходов.
Пользователи независимы друг от друга. Значит, мы можем представить пользователя как бакет, в котором будем брать среднее метрик его заходов.
Тогда у нас получится новая метрика – среднее время нахождения пользователя в приложении или сайте. Дальше будем по общему пайплайну А/Б тестов оценивать эту новую метрику.

Линеаризация

$$CTR = \frac{\sum_u C(u)}{\sum_u S(u)}$$ - поюзерная метрика

$KS(u)$ – грубо считаем, сколько было бы кликов в группе experiment без наших изменений.

$L(u) = C(u) - KS(u)$ – абсолютная разница в кликах

$$L'(u) = \frac{\sum_u L(u)}{|U|}$$ - усреднение по пользователям

Дробь преобразуется в поюзерную-метрику

https://www.researchgate.net/publication/322969314_Consistent_Transformation_of_Ratio_Metrics_for_Efficient_Online_Controlled_Experiments

Линеаризация

Пусть $C(u)$ – клики пользователя u – в группе experiment,
S(u) – показы пользователю u в experiment,

K – средний CTR в группе control,

$$CTR = \frac{\sum_u C(u)}{\sum_u S(u)}$$

$$L(u) = C(u) - KS(u)$$

$$\text{LCTR - } L'(u) = \frac{\sum_u L(u)}{|U|}$$

https://www.researchgate.net/publication/322969314_Consistent_Transformation_of_Ratio_Metrics_for_Efficient_Online_Controlled_Experiments

Для каких метрик работает?

$$R(U) := \frac{\sum_{u \in U} \sum_{w \in \Omega_u} x(w)}{\sum_{u \in U} |\Omega_u|} = \frac{\sum_{u \in U} X(u)}{\sum_{u \in U} Y(u)}$$ Ratio метрика

$$L_{x,k}(u) = X(u) - kY(u)$$

k-среднее R(U) в контроле

$$\frac{\sum_{u \in U} X(u)}{\sum_{u \in U} Y(u)} \to avg_U L_{x,k}$$

линеаризованная метрика

$$L(U) = avg_{u \in U} X(u) - k * avg_{u \in U} Y(u)$$

https://www.researchgate.net/publication/322969314_Consistent_Transformation_of_Ratio_Metrics_for_Efficient_Online_Controlled_Experiments

Особенности

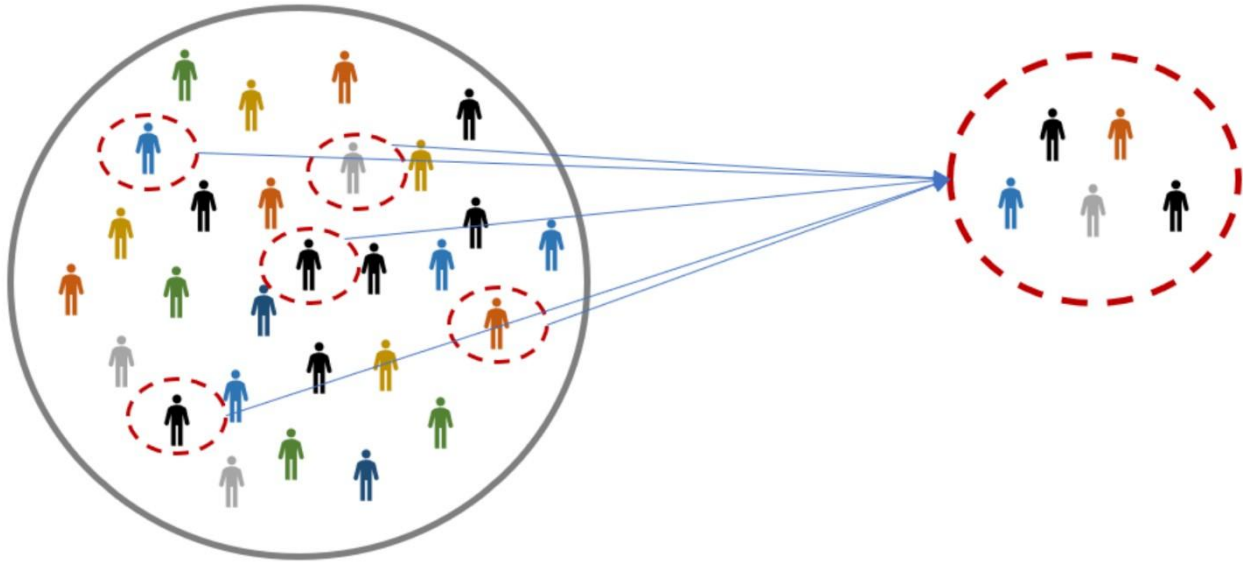
- Линеаризованная метрика сохраняет направленность и значимость
- Увеличивает чувствительность метрики, переходя к линеаризованной пользовательской метрике
- Позволяет бороться с неоднородностью метрики

•Разница между группами по линеаризованной метрике больше, чем разница в группах по ratio-метрике

Денежные метрики и маленькие выборки

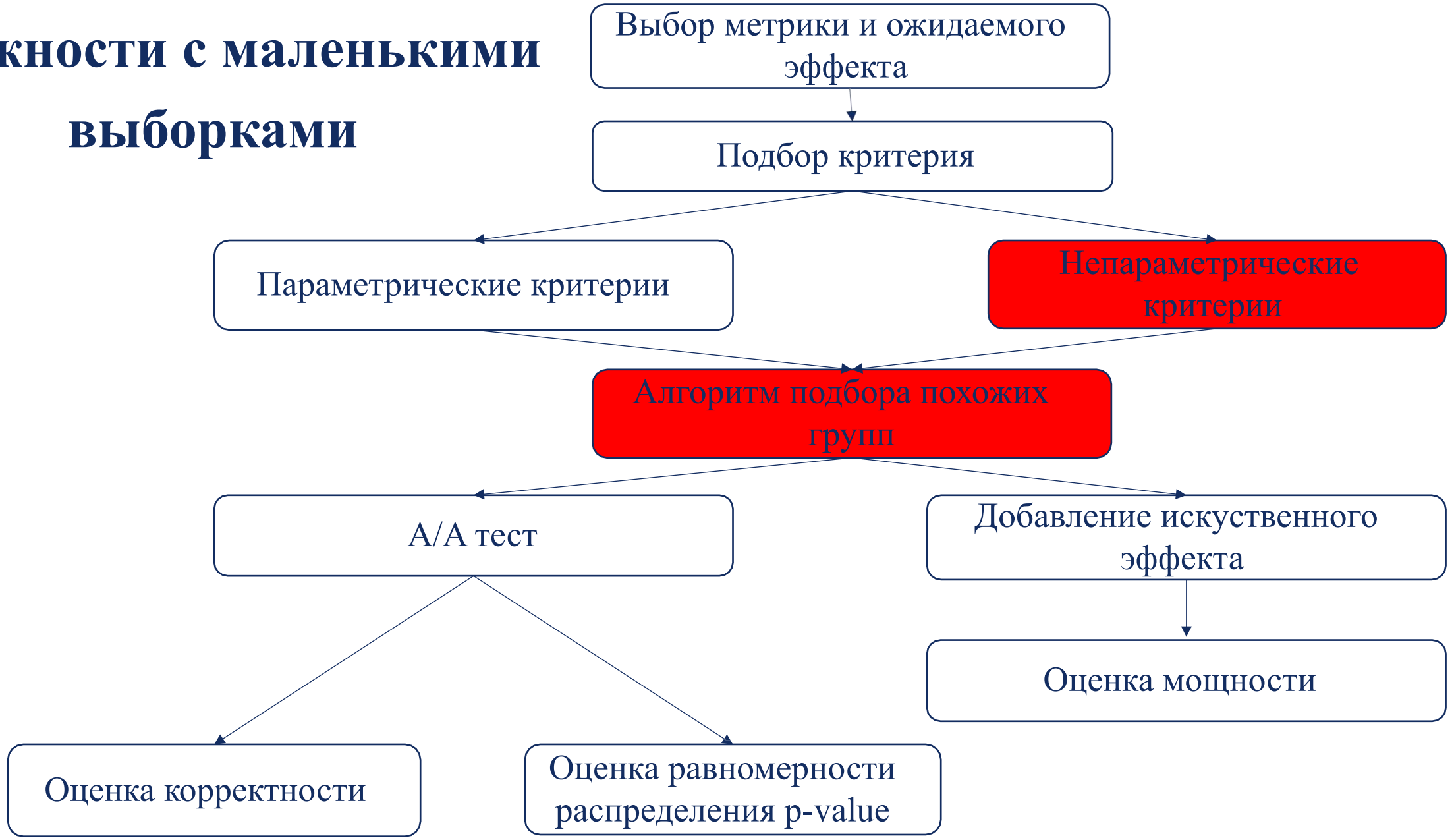
A/B тест на маленьких выборках

- Сложно подобрать группы для теста или разбить все доступные объекты на две “честные” группы для A/B.
- Плохо применимы параметрические критерии.
- Слабая чувствительность, особенно для высоко-дисперсионных денежных метрик.



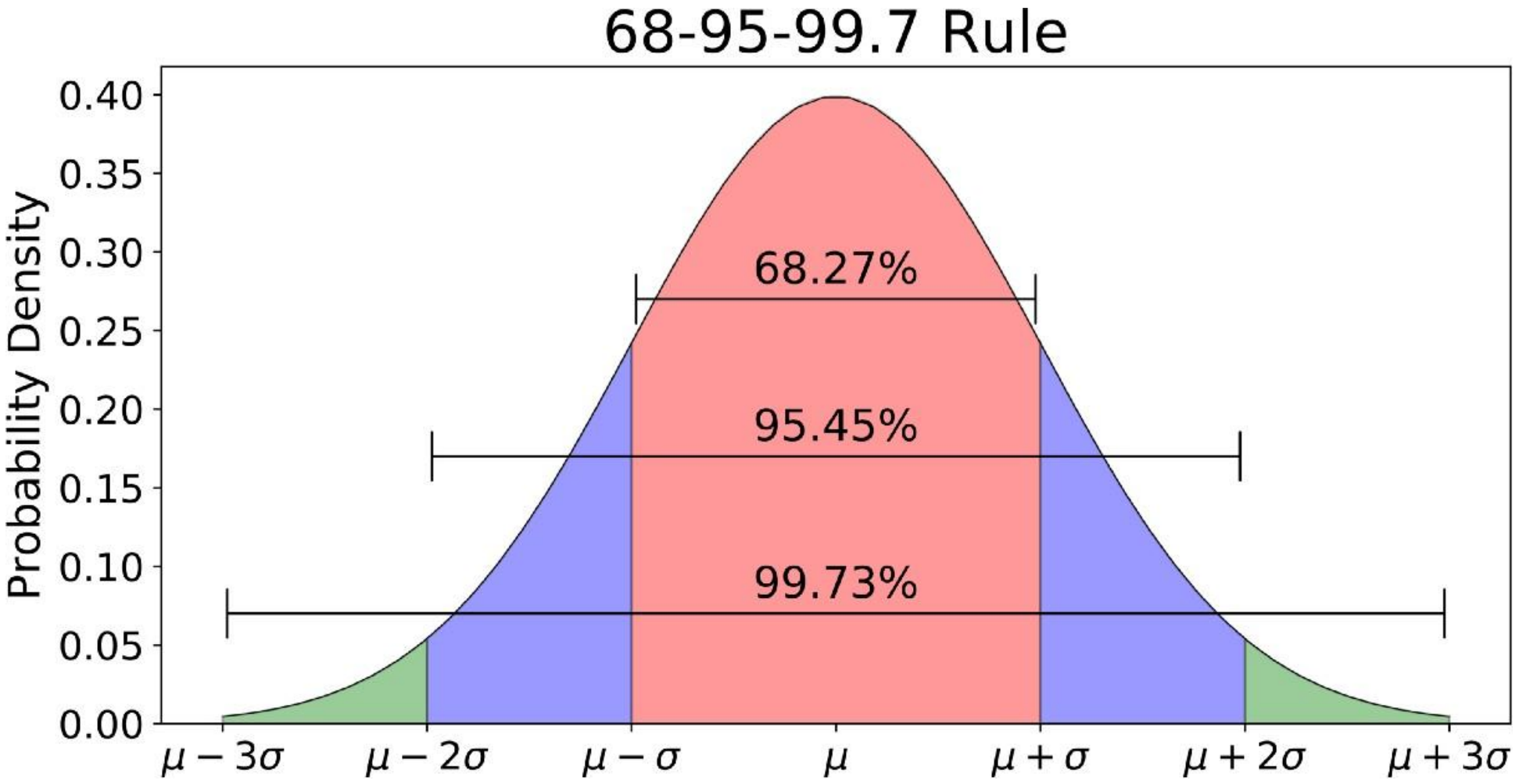
- Бутстрап применять не совсем корректно.
- Сложно измерить чувствительность и мощность перед началом теста.

Сложности с маленькими выборками

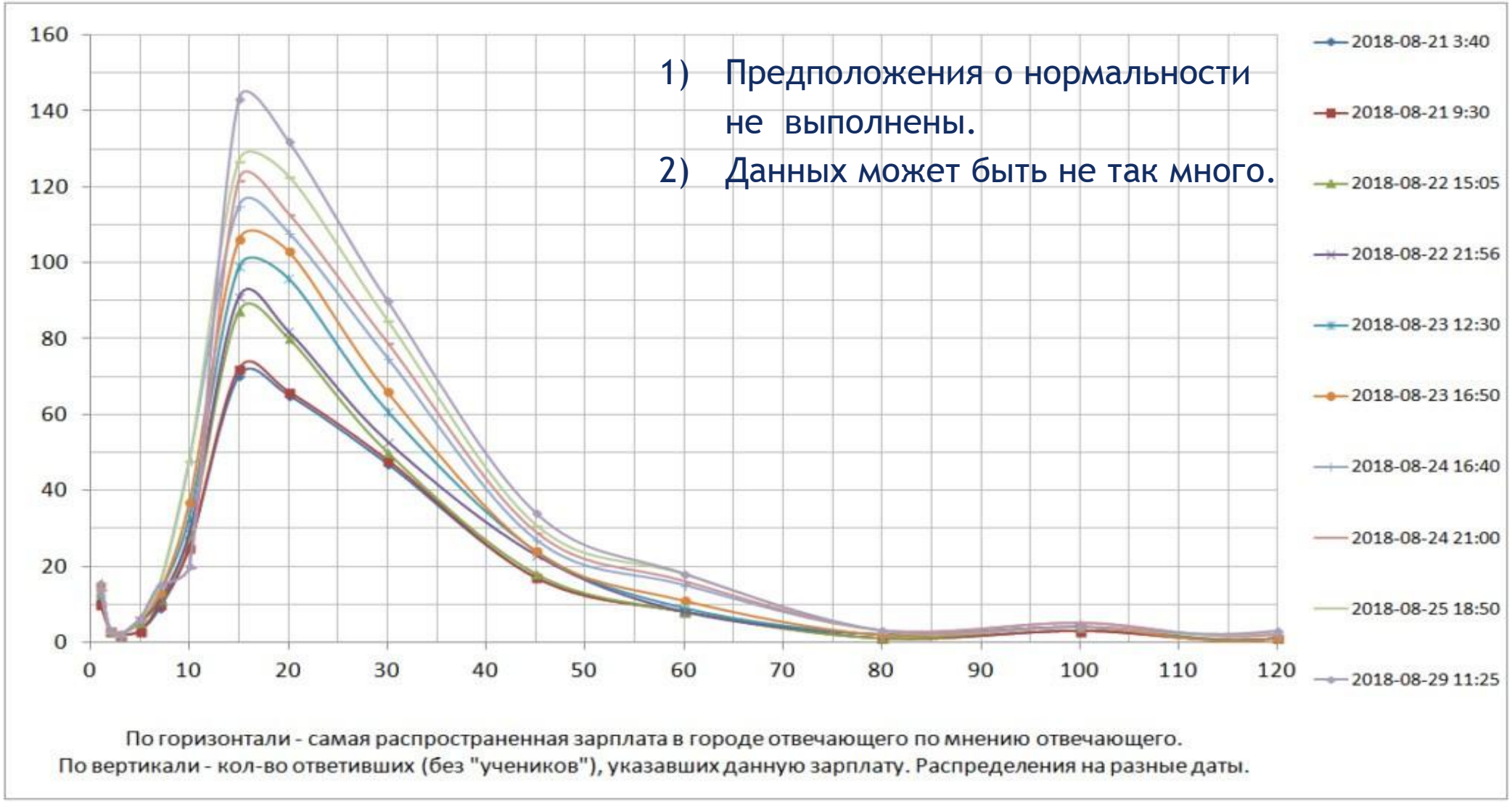


Что интересует бизнес и что с этим делать?

О чем мечтает каждый?



Что дает нам реальность?



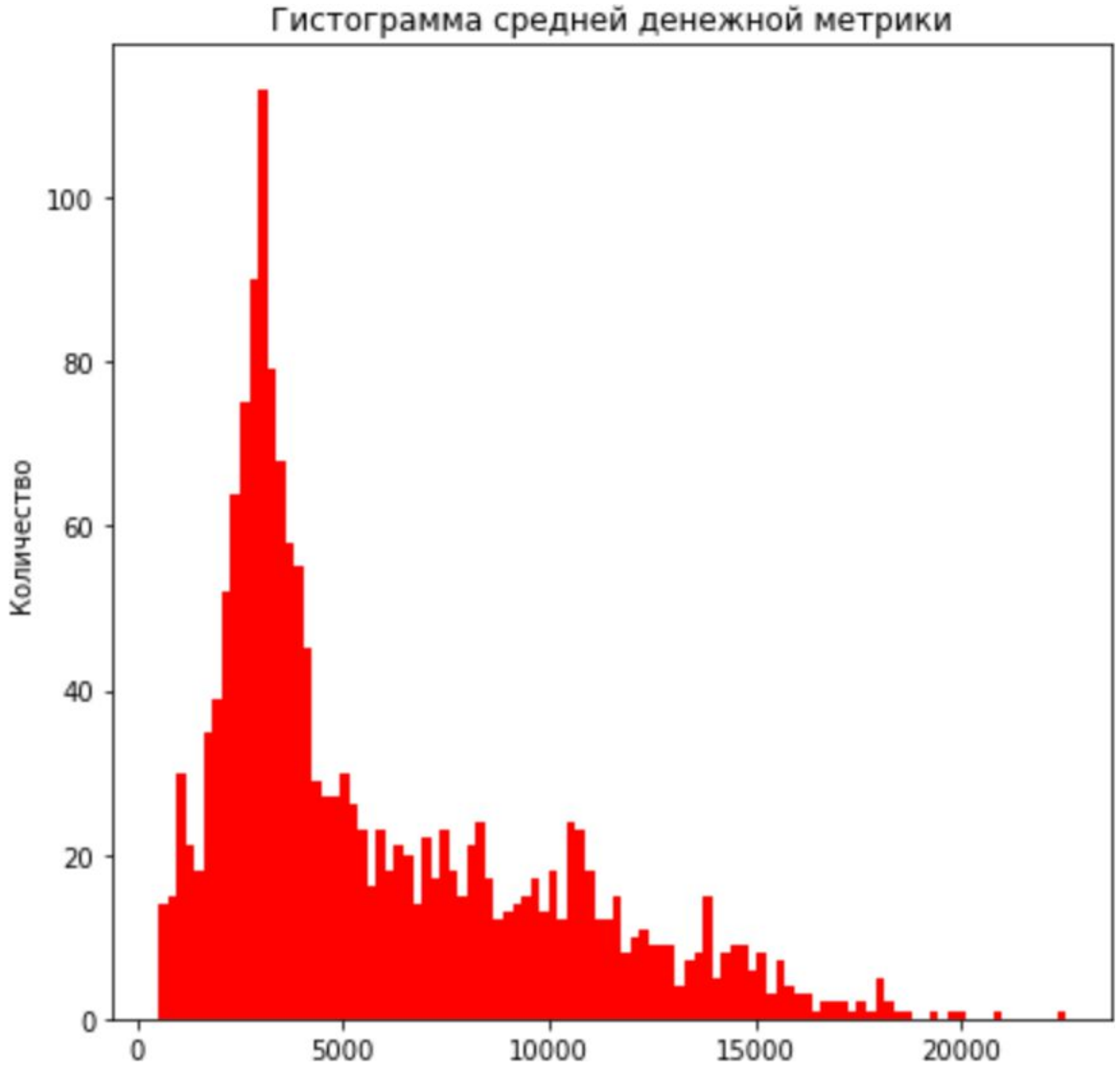
Сложности с денежными метриками

- Низкая чувствительность метрик.
- Часто есть необычные выбросы, которые нужно учитывать.
- Метрики могут быть сильно зашумлены.
- Почти всегда метрики распределены не нормально.
- Бизнесу нужна интерпретация.
- Может не быть хороших исторических данных или они будут некорректными.

Возможные решения

- Переход в другую метрику.
    - Усреднение по пользователю или другим объектам
    - Логарифмирование или другие преобразования
    - Преобразования бокса-кокса - общий вид
- CUPED для уменьшения дисперсии
- Бутстрап, но не на маленьком количестве данных
- Децильный метод – также не на маленьких выборках.
- Удаление выбросов

Децильный метод



Среднее – 6000, стандартное отклонение соизмеримо со средним.

↓

Разобьем на 10% перцентили

Децильный метод

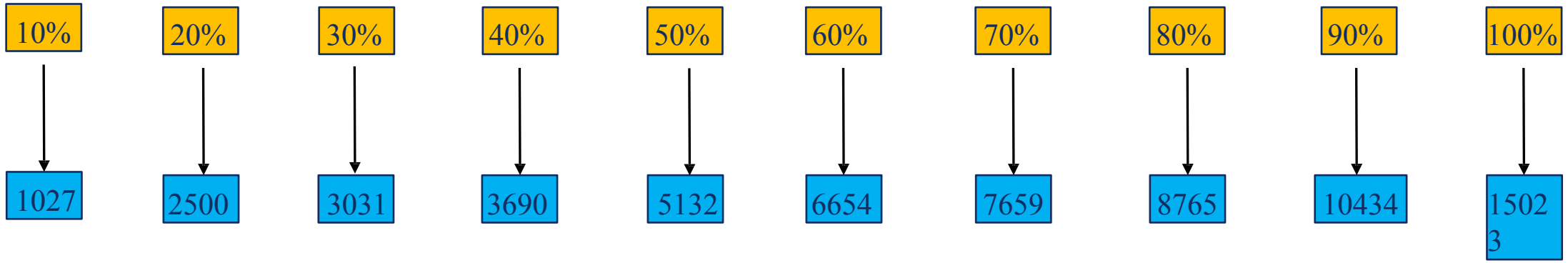
Среднее – 6000, стандартное отклонение соизмеримо со средним.

↓

Разобьем на 10% перцентили

↓

Будем оценивать среднее в каждом из перцентилей

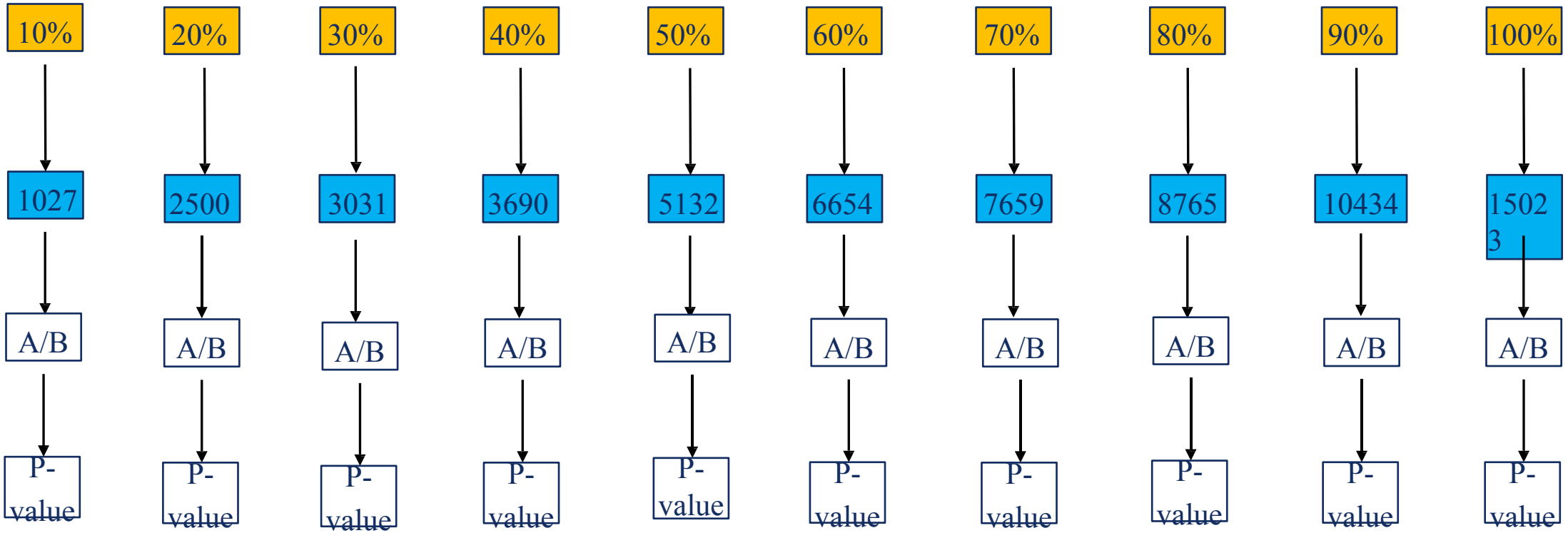


Децильный метод

Среднее – 6000, стандартное отклонение соизмеримо со средним.

Разобьем на 10% перцентили

Будем оценивать среднее в каждом из перцентилей

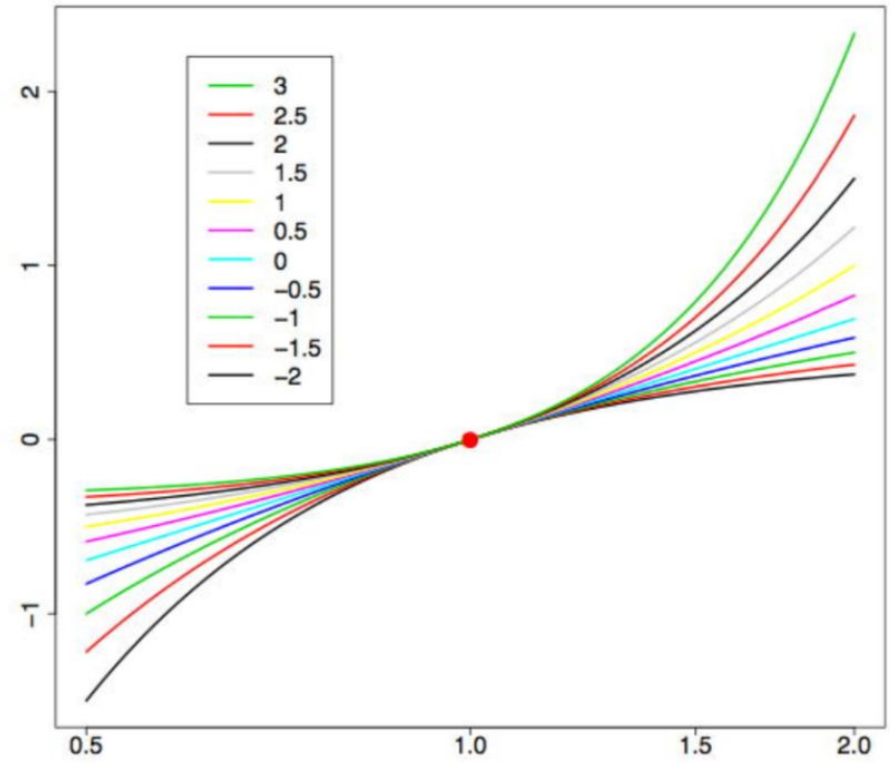


Особенности

- Хорошо подойдет для денежных метрик
- Понижает дисперсию, как следствие увеличивает чувствительность а/б теста
- Нужно помнить про поправку на множественную проверку гипотез
- Можно вылавливать изменения у определенной когорты пользователей

Преобразование Бокса-Кокса

Для выборки $X_i^n = (X_{i1}, \dots, X_{2n})$



$$x_i(\lambda) = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda}, \lambda \neq 0 \\ \ln(x_i), \lambda = 0 \end{cases}$$

Цель : хотим чтобы выполнялся тест на нормальность.

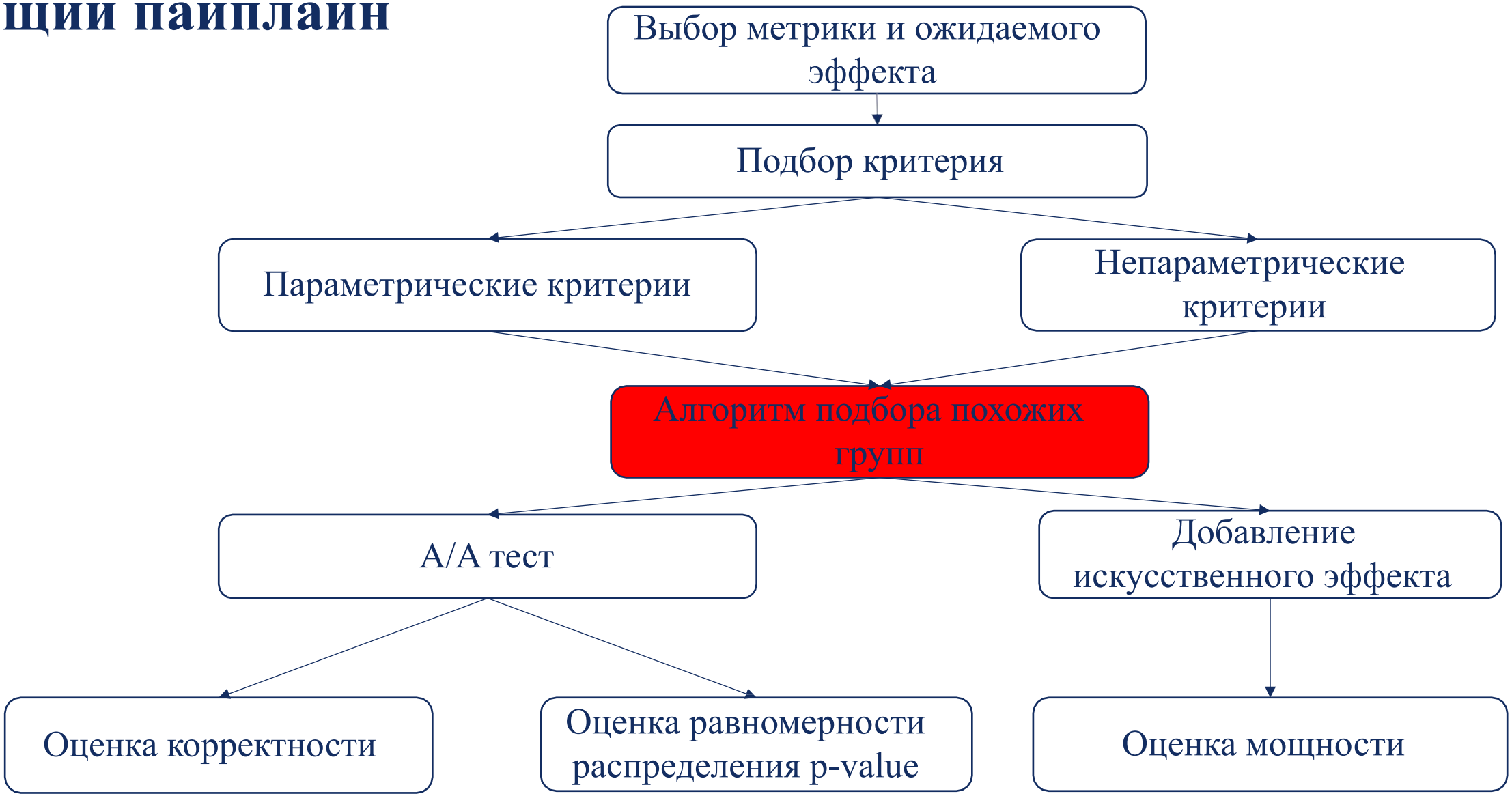
При использовании Бокс-Кокс преобразования нужно, чтобы все значения были положительными и больше 0, но можно преобразовать к такому виду.

Оптимальное значение параметра находится методом максимального правдоподобия .

https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Box-Cox_Transformation.pdf

Разбиение на группы

Общий пайплайн



Отбор групп – простой случай

Для разбиения используют хэш от id-шника пользователя с солью.

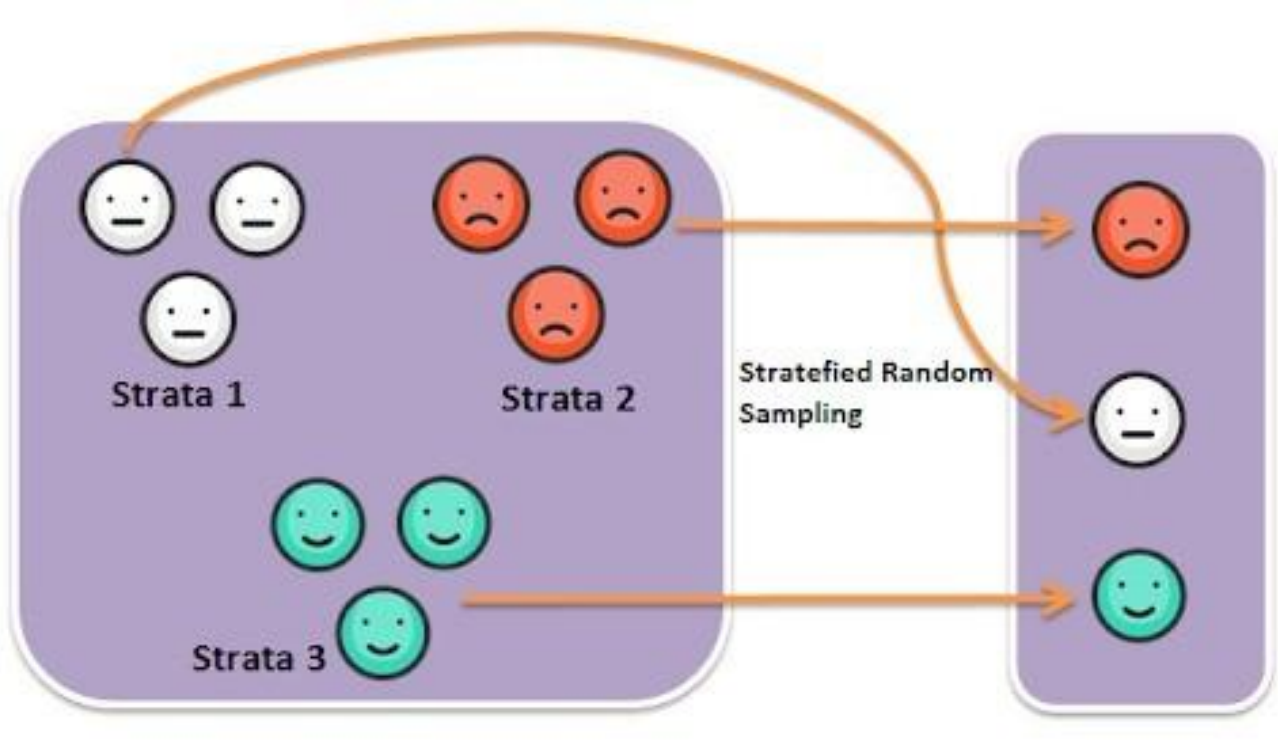
Варианты хэширования:

- SHA-2
- Стрибог
- MD5

Соль позволяет:

- избавится от зависимости от id-шника
- При сэмплировании получать разные группы для A/B

Отбор групп – сложный вариант



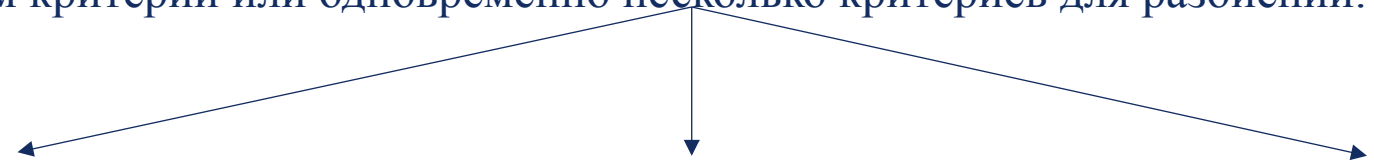
- Что делать, если ваша метрика - деньги?
- Как сделать группы в A/B тесте похожими?
- Как сделать A/B тест честным?
- Что делать, если группы маленькие?
- Что делать, если признаки - это тоже деньги?

Пример 1: разбивка магазинов/салонов по прибыльности.
Пример 2: разбивка регионов по доходу с продаж.
Пример 3: разбивка компаний по прибыльности.

Подбор групп по критериям однородности

Шаг 1: Выбираем целевую метрику и/или любой другой исторический признак, по которым хотим найти похожие группы.

Шаг 2: Выбираем критерий или одновременно несколько критериев для разбиений.



Критерий Колмогорова-Смирнова Критерий Манна-Уитни Перестановочный критерий

Шаг 3: Случайным или жадным поиском ищем разбиения, для которых выполняются критерий или критерии.

Пример: разбиения регионов на группы – использование критериев Манна-Уитни и Колмогорова-Смирнова одновременно для целевой метрики.

Подбор групп по функции стоимости или метрике похожести

Шаг 1: Выбираем целевую метрику и/или исторические признаки, по которым хотим искать похожие группы.

Шаг 2: Выбираем метрику похожести : l1-норма, l2-норма, скалярное произведение или люб другую. Также можем составить функцию стоимости из метрик с некоторыми параметрами Например, следующую:

$$cost(A,B) = \sum_{m \in M, d \in D} \lambda_m (X^A_{m,d} - X^B_{m,d})^2 ,$$

M — множество всех метрик и признаков, D — множество дат.

Шаг 3: Считаем метрику похожести или функцию стоимости для групп A и B, где каждая координата – это значение в конкретную дату.

Шаг 4: Ищем похожие группы.

https://engineering.nordeus.com/perfecting-a-b-testing/

Алгоритм подбора групп по прогнозам целевой метрики

Шаг 1: Выбираем целевую метрику, по которой хотим искать похожие группы.

Шаг 2: Прогнозируем целевую метрику на период пилота

Шаг 3: Должна быть хорошая простая модель – линейная регрессия.
Важно: Ошибки прогноза модели должны быть распределены нормально.

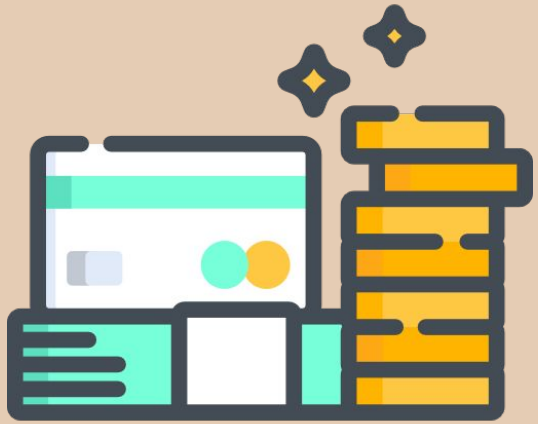
Шаг 4: Выбираем тех, кто будет вести себя одинаково в будущем

По метрике похожести

По критериям однородности

https://www.tripadvisor.com/engineering/reducing-a-b-test-measurement-variance- by-30/

Кейсы из практики



Как провести A/B тест на регионах?

Основная сложность - разбить регионы на две похожие группы.

В чем задача ?

МТС

Офис продаж

Предприниматели

$$$

Маркетинг

Как проверить эффект?

Новый продукт от BigData

МТС

Офис продаж

Предприниматели

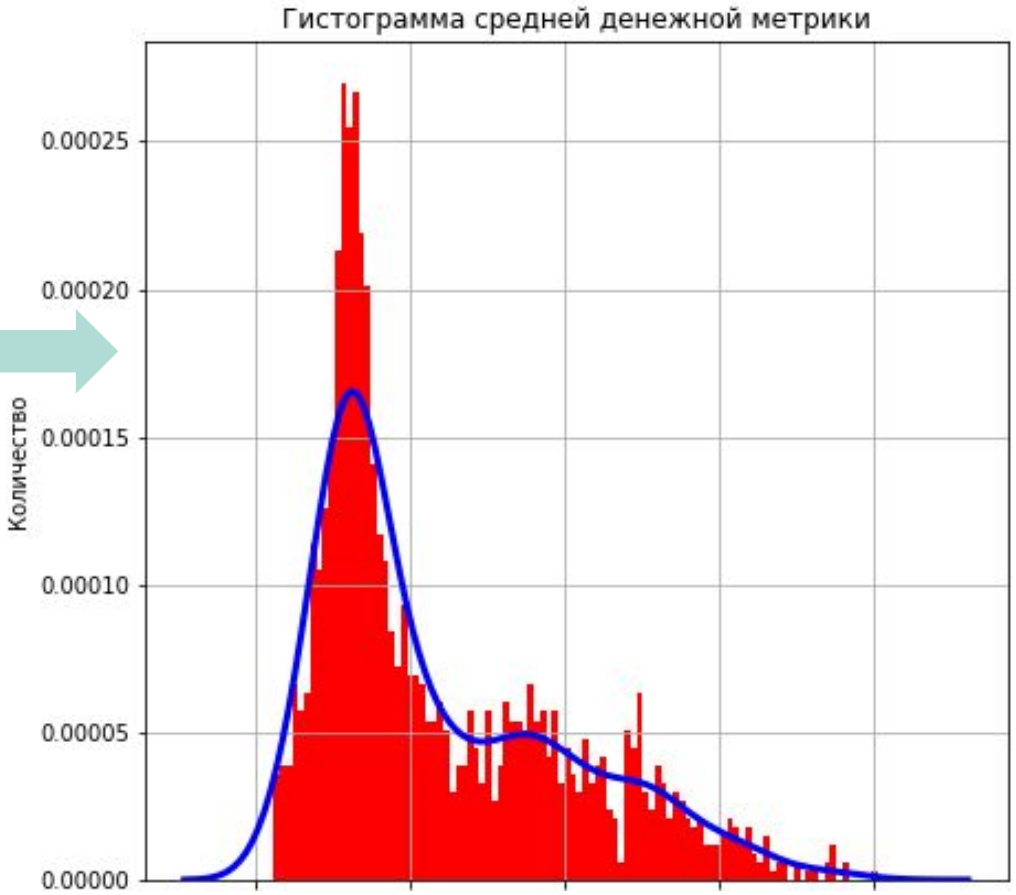
Маркетинг

$$$

$$$

Какие есть данные?

Доход, разбитый по бизнес-категориям, региона с продаж в каждый месяц за последние 2 года

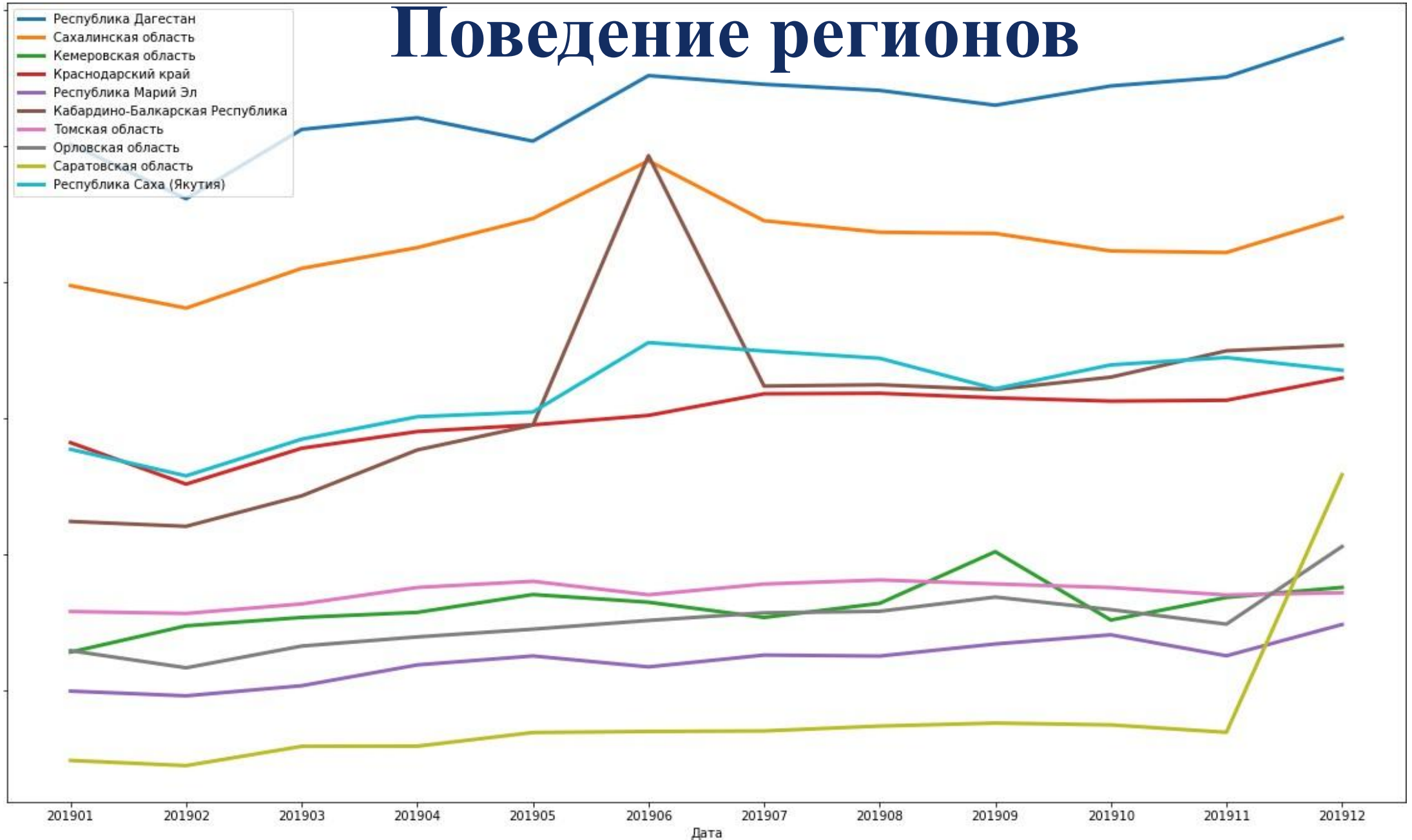


Средний доход приходящийся на одну компанию в каждом месяце

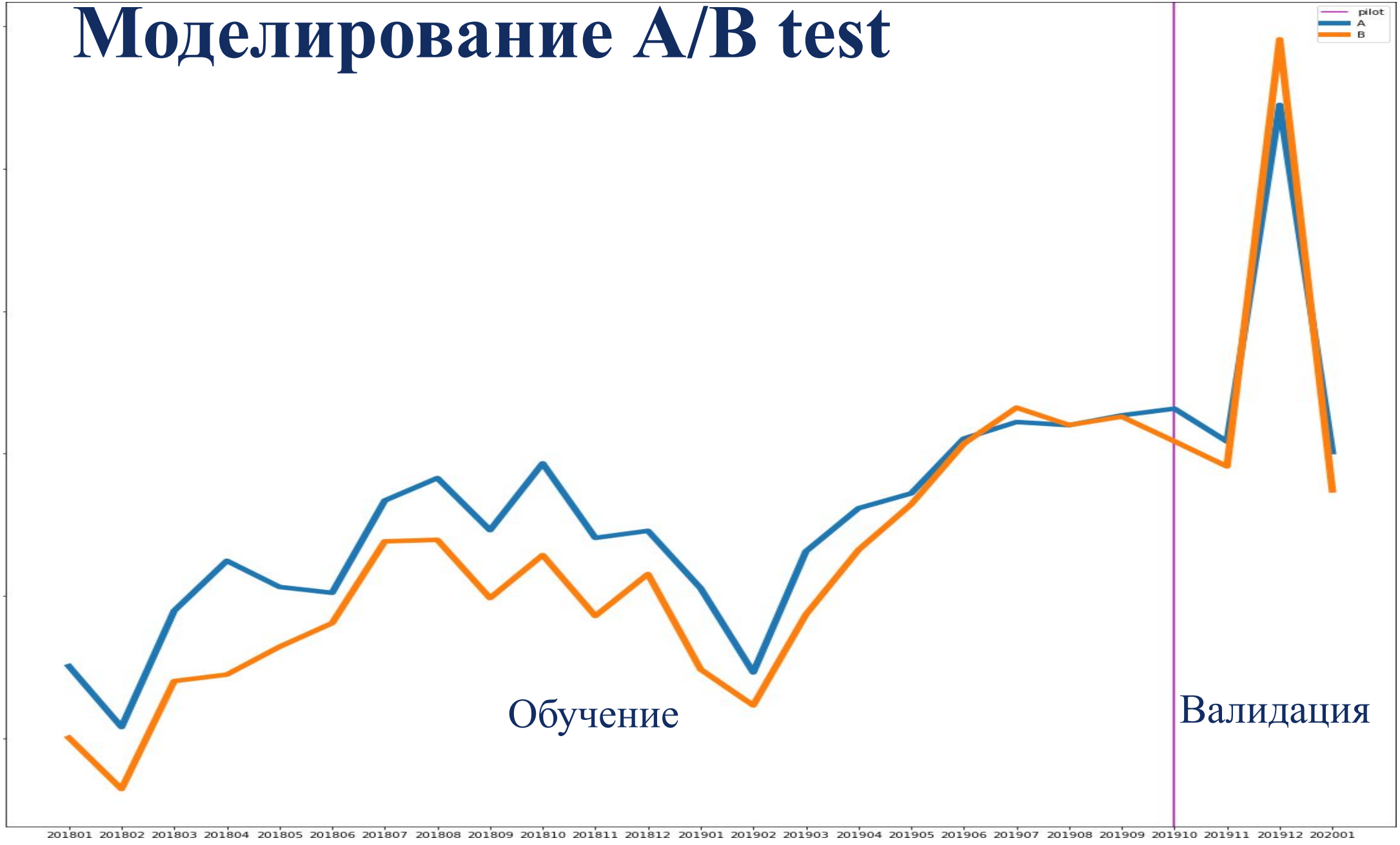
Число наших клиентов-компаний в каждом месяце

Наблюдение: регион в конкретный месяц
Наблюдение: усреднение метрики региона за несколько месяцев

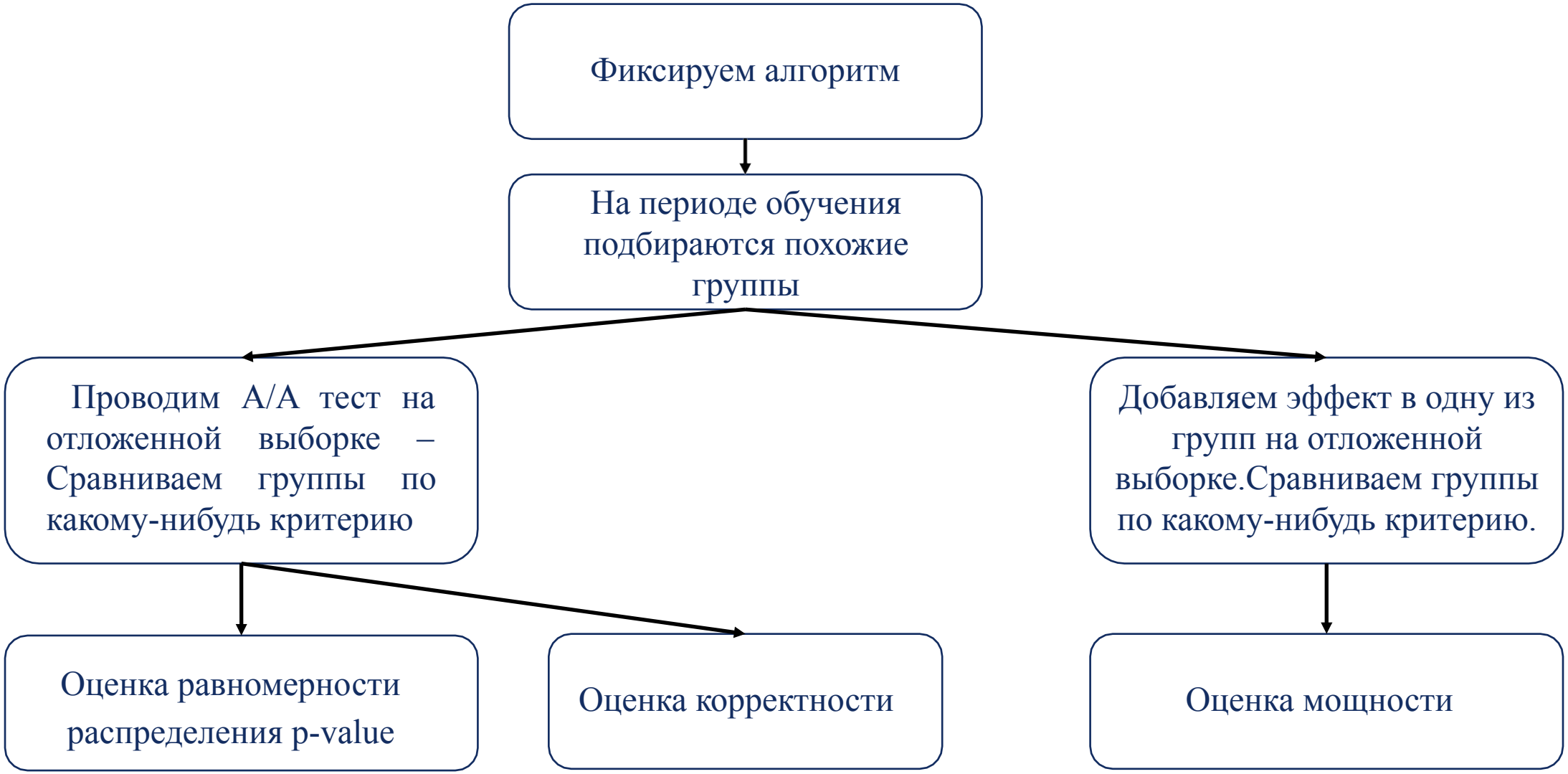
Поведение регионов



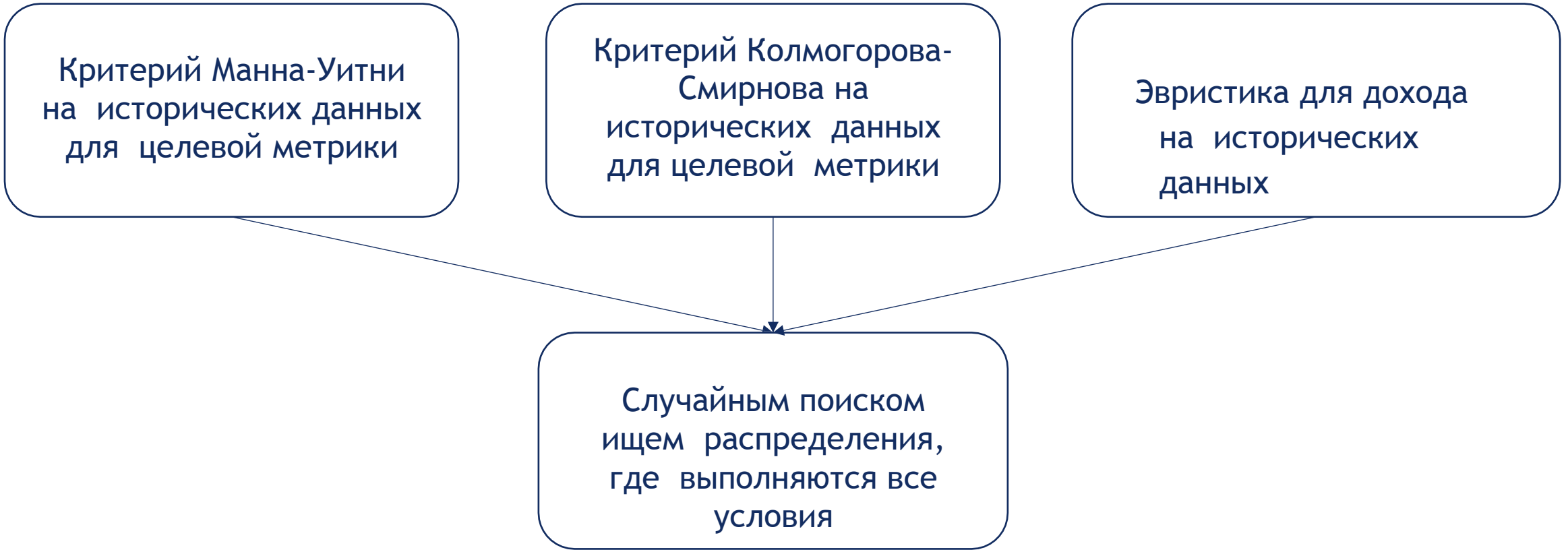
Моделирование A/B test



Как оцениваем алгоритм подбора групп?



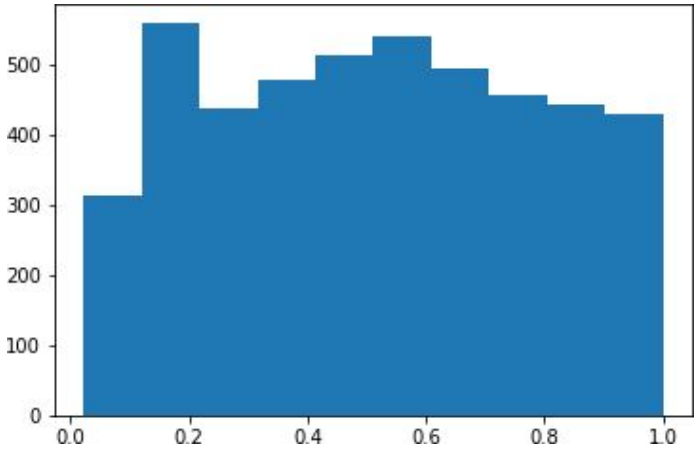
Что сработало хорошо?



Критерии оценивания пилота

	Критерий Манна-Уитни на отложенной выборке	T-test для целевой метрики на отложенной выборке	Перестановочный критерий на отложенной выборке
Корректность	0.5 %	Не корректен	2.5 %
Мощность	85% при эффекте в 7.5%	-	85 % при эффекте в 8 %

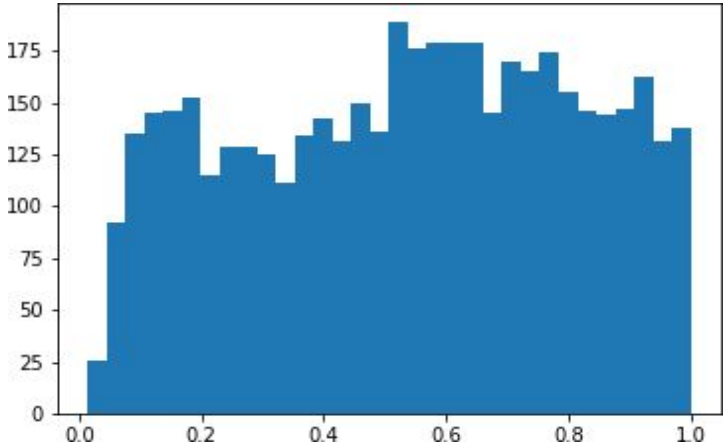
Критерии оценивания пилота



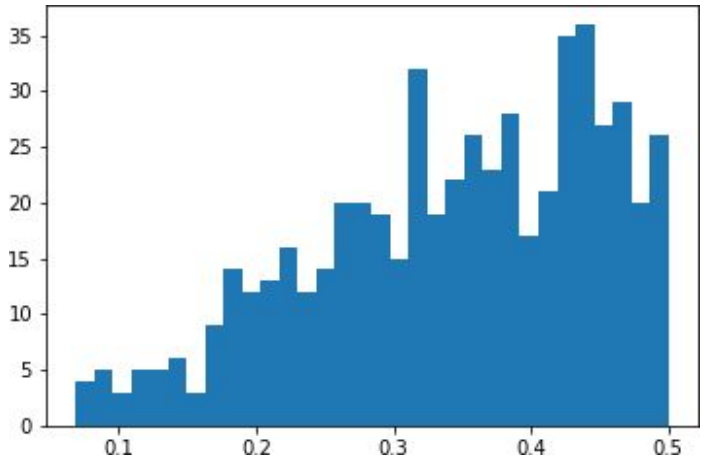
t-test

p-value распределено не равномерно

Критерии не корректные

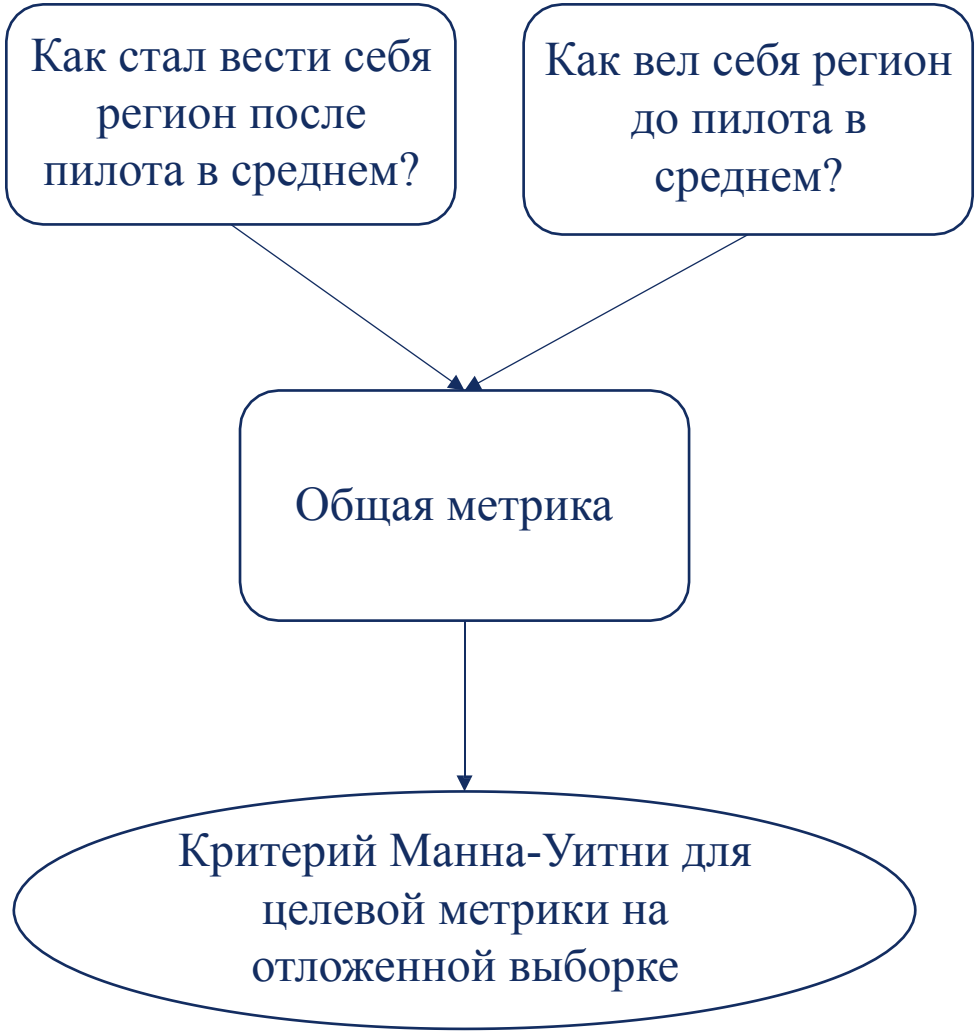


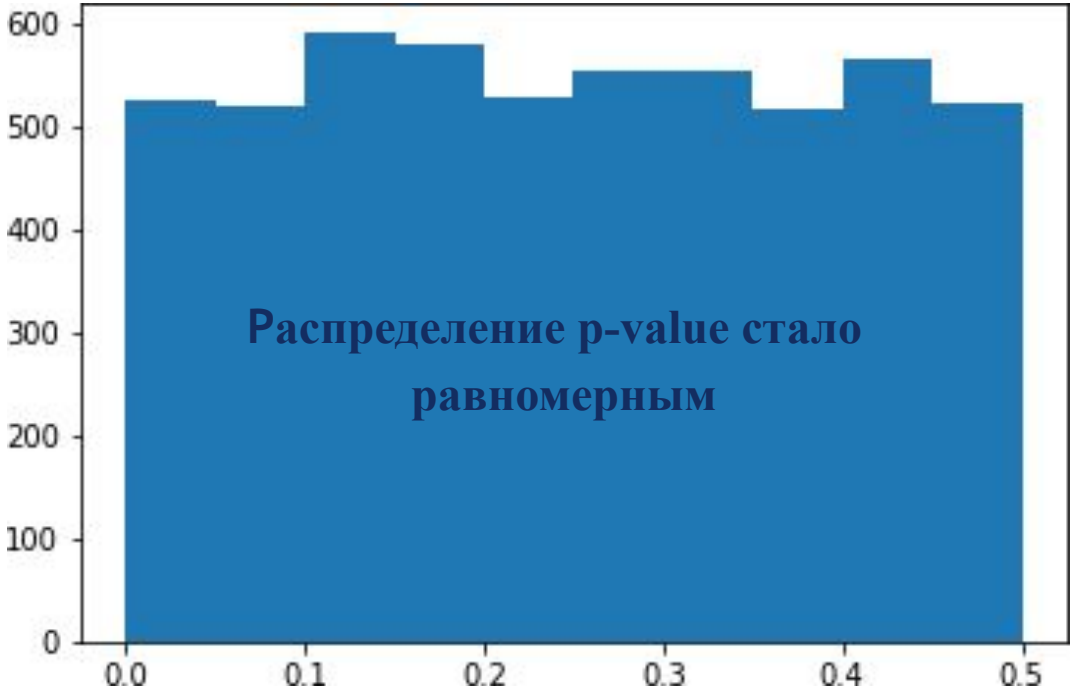
Перестановочный критерий



Критерий Манна-Уитни

Cuped на регионах





Корректность: 6 %

Мощность: в 85% при эффекте в **4.5%**

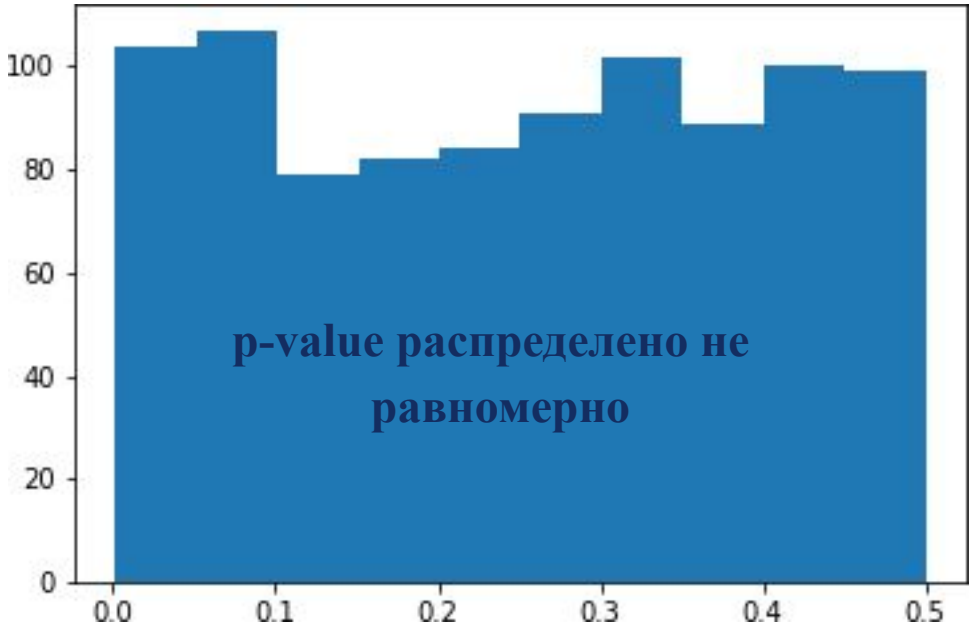
https://booking.ai/how-booking-com-increases-the-power-of-online-experiments-with-cuped-995d186fff1d

Итоговый вариант по критериям однородности

Итоговый вариант по критериям однородности

Распределение p-value оценочных критериев через метрику схожести

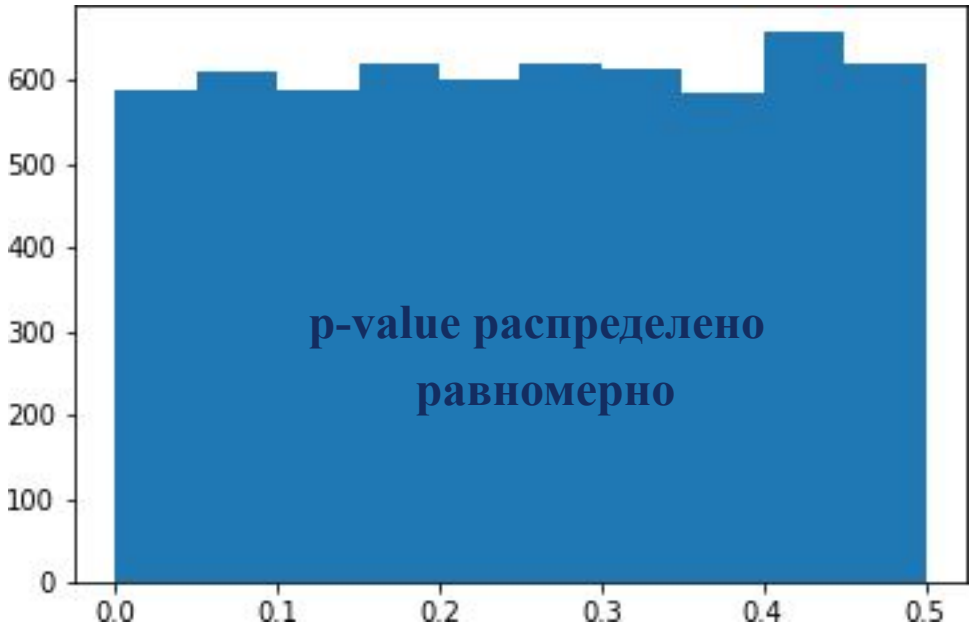
L1-норма



Корректность: 9.5%

Мощность: ловим эффект размера 9.5% в 85% случаях

L2-норма



Корректность: 7%

Мощность: ловим эффект размера 9% в 85% случаях

Разбиение регионов по l2-норме

Вылавливаем 9% эффекта в 85% случаях

Итог

Как повышать чувствительность критериев?

- Если есть хорошие исторические (пре-экспериментальные) данные, то выбрать ковариату и применить CUPED и ранговое преобразование
- Если мы можем разбить наши обьекты по независящим от AB теста признакам на сегменты, то применить стратификацию
- Если мы имеем дело с поюзерными ratio-метриками , то для повышения чувствительности метрик хорошо подойдет линеаризация
- Если у нас есть зависимые события или выбросы, то мы можем попробовать улучшить ситуацию, применив бакетное сэмплирование.

Итог

Что важно если мы работаем с денежными метриками или маленькими выборками:

- Выбрать правильную метрику или перейти к другой желательно интерпретируемой метрике с более низкой дисперсией.
- Максимально снизить дисперсию
- Быть аккуратными при преобразованиях метрики
- Оценить алгоритм подбора групп и убедиться в его корректности.
- Осторожно использовать методы увеличения чувствительности тест.

Спасибо за внимание!

Ildar Safilo

@Ildar_Saf

irsafilo@gmail.com

https://www.linkedin.com/in/isafilo/