Ускорение A/B-тестов

1. Введение

В 1990-х в связи с появлением коммерческих провайдеров, веб-браузеров и поисковых систем интернет стал доступен широкой аудитории. В это же время организации начинали коммерциализацию интернета: создавали собственные веб-страницы и интернет-магазины. Перед организациями встал вопрос: как конвертировать посетителей сайта в своих клиентов? Как понять, что сайт привёл новых клиентов? Как понять, какая маркетинговая онлайн-стратегия более эффективна? Как проанализировать новый объёмный набор данных о потенциальных клиентах? Для поиска ответов на эти вопросы начали активное использование статистических тестов, а активное распространение Agile- и Lean-подходов в 2010-х годах привело к пониманию, что A/B-тесты нуждаются в ускорении.

В целом эксперименты в индустрии могут затягиваться по нескольким причинам, и их ускорение требует комплексного подхода:

1. Отложенный эффект изменений: в SaaS-секторе, где ключевой метрикой является продление подписки, получение результатов может требовать более длительный период для анализа. Для решения этой проблемы:
    - использование прокси-метрик: например, отмены подписок могут служить индикатором будущих продлений;
    - анализ краткосрочных метрик: метрики, такие как количество сессий или время, проведённое в приложении, могут предсказывать более отдалённые показатели. Это позволяет получать более быстрые результаты и делать выводы о влиянии изменений.
2. Когда мы стремимся обнаружить малозначительные изменения, необходимо привлекать больше пользователей для получения статистически значимых результатов. Для ускорения процесса можно использовать такие подходы:
    - пересмотр способа проверки гипотезы: в новых продуктах или стартапах обычно фокусируются на доработке гипотезы через UX-исследования (исследования пользовательского опыта). В таких случаях A/B-тестирование может оказаться слишком затратным, так как в этом случае требуется уже разработка продукта;
    - разработка более чувствительных метрик/моделей: это может быть длительный аналитический процесс, особенно в зрелых компаниях, где оптимизация становится приоритетом.

О последнем пункте мы поговорим подробнее на текущем уроке. В рамках этой темы мы рассмотрим, какие способы ускорения тестов используются в индустрии, какие проблемы могут возникнуть в процессе их использования и какие модификации для стандартных методов помогут их решить.

2. Основные этапы проведения A/B-теста

1. Сформулировать бизнесовую проблему: мы хотим снизить стоимость поддержки пользователей без потери качества.
2. Сформулировать гипотезу по решению бизнесовой проблемы: если мы внедрим чат-ботов, то они заберут на себя часть нагрузки по обработке обращений с оператором, и тем самым мы снизим косты и не потеряем в качестве.
3. Проверить гипотезу.
   1) Выбираем метрики: средний кост на обработку одного обращения и CSAT.
   2) Выбираем статистический критерий, с помощью которого будем проверять гипотезу, и фиксируем ошибки первого и второго рода: t-test, ошибка первого рода — 5%, ошибка второго рода — 20%.
   3) Считаем минимальный необходимый объём выборки:

$$n = 2\left(\frac{Z_{\alpha/2} + Z_{\beta}}{\delta/\sigma}\right)^2, \tag{1}$$

   где

   - n — размер выборки для каждой группы;
   - $Z_{\alpha/2}$ — критическое значение стандартного нормального распределения для уровня значимости α;
   - Z_{β} — критическое значение стандартного нормального распределения для мощности теста;
   - δ — минимальный размер эффекта, который мы хотим обнаружить;
   - σ — стандартное отклонение.

   4) Дизайним сплит: сплитовать будем по пользователям, которые обращаются в чат поддержки. Контроль — обрабатывают только операторы, тест — первая линия поддержки чат-ботов. Вероятность попадания в тест — 5%.
   5) Запускаем эксперимент.
4. Принять решение о том, подтвердилась гипотеза или нет, с помощью оценки реального эффекта воздействия (англ. average treatment effect, ATE):

$$ATE = \mathbf{E}[Y_{treatment} - Y_{control}],$$

   где Y — бизнес-метрика.

   Самым простым способом оценки эффекта воздействия является сравнение средних показателей:

$$\widehat{ATE}_{base} = \bar{Y}_{treatment} - \bar{Y}_{control.} \tag{2}$$

   В итоге мы получим несмещённую и состоятельную оценку ATE.
5. Сделать изменения.

В рамках этого урока мы сосредоточимся на этапах 3.3) и 4.

3. Способы ускорить A/B-тест

Как мы увидели из формулы 1, скорость проведения теста зависит от размера выборки и стандартного отклонения целевого показателя. Но у нас не всегда есть возможность ждать накопления необходимого минимального объёма выборки, а рабочих методов увеличить скорость накопления объёма выборки нет, поэтому мы будем оптимизировать дисперсию целевого показателя, то есть стараться повысить чувствительность теста.

Определение Чувствительность теста — способность метода находить эффект, который есть в действительности

В случае если мы работаем с поведением отдельных покупателей, мы можем **поработать с «целевой аудиторией»** нашего продукта. К примеру, влияние изменённого дизайна сайта стримингового сервиса удобнее всего смотреть на новых пользователях. Однако в таком случае мы «урезаем» основную выборку, и это может привести нас к первой проблеме — нехватке размера выборки. Но мы можем снизить стандартное отклонение с помощью разбиения наших пользователей на группы — **разделить их по некоему внешнему признаку** — например по стране проживания — такой приём называется стратификацией, и тогда нам хватит меньшего объёма выборки, чем без стратификации.

Для уменьшения вариативности в данных можно добавить имеющуюся информацию обо всех признаках в **регрессионную модель**, или линейную регрессию (англ. CUPED). Эти методы хороши своей интерпретируемостью, однако чрезвычайно требовательны к входным данным. Например, одной из таких проблем является требование к линейности отклика к предикторам. Решить эту проблему можно с помощью **методов машинного обучения** (CUPAC, MLRATE): они устойчивы к нелинейности, однако усложняют интерпретацию результатов. В следующих пунктах мы рассмотрим каждый из методов более подробно.

3.1 Стратификация

Основная идея стратификации заключается в том, чтобы разделить нашу выборку на страты, обработать выборку в каждой страте и затем объединить результаты оценок из отдельных страт для получения общей оценки.

Разобьём область выборки Y на K страт, при этом p_k — это вероятность того, что Y попадает в k-ю страту, $k = 1, ..., K$. Если мы зафиксируем количество точек, выбираемых из k-й страты, равным $n_k = n \cdot p_k$, мы можем определить стратифицированное среднее как:

$$\hat{Y}_{strat} = \sum_{k=1}^{K} p_k \bar{Y}_k, \tag{3}$$

причём $\mathbf{E}(\hat{Y}_{strat}) = \mathbf{E}(\bar{Y})$ и $var(\hat{Y}_{strat}) < var(\bar{Y})$.

Пример 1. Рассмотрим выборку, которая состоит из измерений роста 7-летних детей (123 ± 5 сантиметров), 12-летних детей (156 ± 8 сантиметров) и взрослых (175 ± 10 сантиметров). Если мы измерим стандартное отклонение выборки в целом, то она составит более 20 сантиметров, в то время как разбивая её по возрастным группам, мы получим значения около 5, 8 и 10

Хорошая сделанная стратификация — это та, что отражает основные кластеры в наших данных. Явно идентифицируя эти кластеры как страты, мы фактически устраняем дополнительную дисперсию.

Конечно, не всегда есть возможность работать со стратификацией в таком виде. Например работая с онлайн-данными, мы собираем данные по мере их поступления. Поэтому обычно мы не можем проводить выборку из заранее сформированных страт. Однако мы всё равно можем использовать переменные, собранные до эксперимента, для построения страт, после того как все данные собраны. Такой приём называется постстратификацией.

В общем случае среднее после постстратификации не обязательно будет равно средневыборочному. Разница между этими значениями может возникать из-за различных факторов:

- различия в распределении страт: если распределение страт в вашей выборке отличается от реального распределения в генеральной совокупности, это может привести к различиям между средним значением и средневыборочным;
- при постстратификации вы можете применять веса (домножать на коэффициенты) к разным стратам, чтобы скорректировать результаты. Эти веса могут изменять итоговое среднее значение, особенно если одна из страт имеет большее влияние на итоговый результат;
- размер выборки: если внутри каждой страты размер выборки мал, то результаты могут быть менее надёжными. Соответственно, мы снова получим различия в средних.

Для разбиения на страты мы используем матрицу ковариат X. Пусть Y_i бизнес-метрика для i-го наблюдения (например, сумма, потраченная юзером с ID, равным i), X_i — ковариата для i (например, страна проживания юзера с ID, равным i), тогда из 3:

$$\hat{Y}_{strat} = \sum_{k=1}^{K} p_k \Big(\frac{1}{n_k} \sum_{i:X_i=k} Y_i \Big)$$

и размер эффекта

$$ATE_{strat} = \sum_{k=1}^{K} p_k (\bar{Y}_k^{(t)} - \bar{Y}_k^{(c)}),$$

где t и c — экспериментальная и контрольная группы соответственно.

Несмотря на удобство применения, стратификация (и постстратификация) — самый нестабильный метод снижения дисперсии:

- подбор страт для A/B-тестирования может быть сложной задачей: пользователи могут менять свои страты с различной частотой. Это создаёт трудности в том, как правильно зафиксировать принадлежность пользователя к конкретной страте на момент проведения теста. Например, если пользователь изменил свою категорию после начала тестирования, это может исказить результаты. Чтобы избежать таких ситуаций, важно заранее определить критерии для стратификации и установить чёткие временные рамки, в течение которых пользователи будут закреплены за определённой стратой;
- неправильный выбор страт может привести к увеличению вероятности ошибки первого рода: для минимизации возникновения ошибки необходимо постоянно отслеживать результаты A/A-тестов. Регулярный анализ A/A-тестов помогает выявить аномалии и убедиться в том, что любые наблюдаемые эффекты в A/B-тестах не являются следствием неправильной стратификации или других систематических ошибок.

3.2 Controlled-experiment using pre-experiment data — CUPED

Как мы уже заметили, основная проблема оценки эффекта состоит в сильной вариативности данных, что может привести нас к неверным выводам и принятию неэффективных бизнес-решений. В 2008 году эту проблему подробно описал Фридман в своей статье "On regression adjustments to experimental data". В статье критикуется использование линейной регрессионной модели и доказывается, что, даже при выполнении условия рандомизации данных, высока вероятность получения смещённых оценок. Основная идея предложенной им корректировки регрессии состоит в том, чтобы использовать дополнительные данные, собранные до начала эксперимента (англ. pre-experiment data). Например, включать информацию о поведении пользователей в качестве независимых переменных в регрессионную модель. Идея Фридмана, по сути, отражает основной принцип работы метода CUPED — использовать предэкспериментальные данные в качестве ковариат.

Предположим, что предэкспериментальные данные X являются одномерной ковариатой. В CUPED вместо того чтобы рассматривать стандартную бизнес-метрику Y, мы смотрим на новую метрику, определяемую как

$$Y_{CUPED} = Y - \theta(X - \bar{X}),$$

где θ — это некоторый параметр, который мы полагаем равным $cov(X, Y)/var(X)$. Выбор параметра θ связан с тем, что при таком параметре нам удастся минимизировать дисперсию:

$$var(Y_{CUPED}) = var(Y)(1 - \rho^2),$$

где $\rho = corr_{srs}(X, Y)$ — это коэффициент корреляции Пирсона между X и Y. Интуиция такого метода заключается в том, что мы раскладываем общую дисперсию бизнес-метрики Y на две части: часть, вызванную дисперсией ковариаты X, и часть, объясняемую другими неизвестными переменными. Причём θ считается на объединении контрольной и экспериментальной групп.

Пример 2. Сгенерируем две выборки $Y^{(c)}_{before}$ и $Y^{(t)}_{before}$ размером $N = 1000$ из нормального распределения со средним 1500 и стандартным отклонением 150. Это будут значения двух выборок до начала эксперимента. Добавим в выборку $Y^{(c)}_{before}$ шум со стандартным отклонением, равным 100, в выборку $Y^{(t)}_{before}$ добавим тот же шум и к каждому значению выборки прибавим значение ATE, равное 200. Назовём их $Y^{(c)}_{after}$ и $Y^{(t)}_{after}$ соответственно. Результат генерации проиллюстрирован на рисунке 1. Для данной генерации оценка эффекта с помощью обычной разницы средних (2) даёт результат $ATE_{base} = 201.382$, в то время как CUPED даёт более далёкую от реальности оценку $ATE_{CUPED} = 207.945$. Всё дело в том, что **CUPED снижает значение дисперсии не всегда, а в среднем**. Для того чтобы увидеть, как работает метод CUPED, нужно множество раз провести наш A/B-тест, поэтому повторим наш способ генерации данных 1000 раз. На рис.2 видно, что дисперсия оценки отличается в два раза. Более подробно — смотри в файле *cuped_cupac.ipynb*.

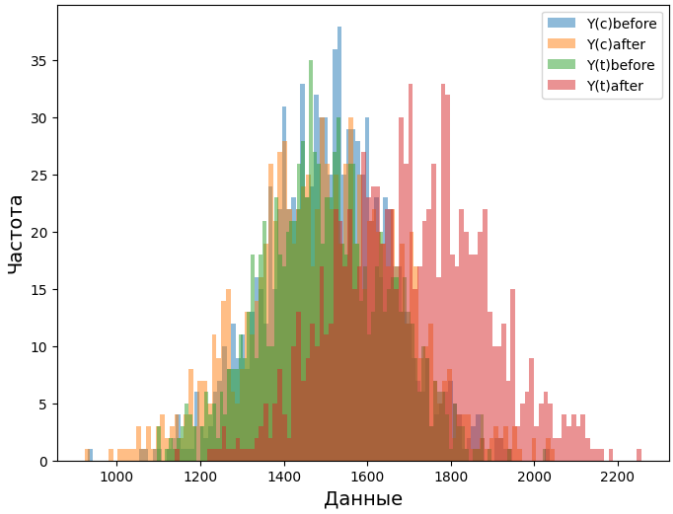

Рис. 1. Визуализация данных из примера 2

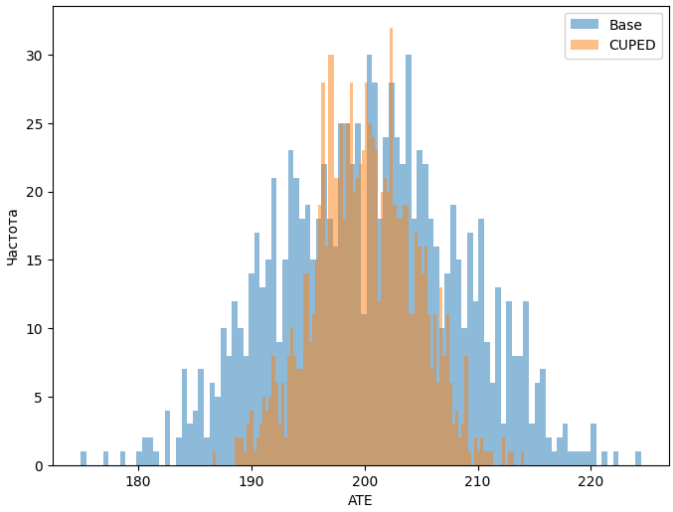

Рис. 2. Дисперсия оценки эффекта стандартного метода и метода CUPED

Для повышения точности оценки эффекта изменений используются различные способы улучшения стандартного метода CUPED.

3.3 Control using predictions as covariates — CUPAC

Метод CUPAC является своего рода расширением метода CUPED (controlled experiments using pre-experiment data). CUPAC, по сути, является CUPED, где вместо ковариаты берутся не предэксперементальные данные, а предикты модели, которая построена на предэксперементальных данных. Причём CUPAC подходит и для динамических способов принятия решений, а CUPED — для повышения статистической мощности тестов.

$$Y_{CUPAC} = Y - \theta g(X) = Y - \theta \hat{Y}, \tag{4}$$

где $\theta = \frac{cov(\hat{Y},Y)}{var(\hat{Y})}$.

Использование предиктов, полученных с помощью машинного обучения, позволяет получить значительное улучшение по нескольким причинам. Во-первых, это хороший способ для случая, если предварительные данные эксперимента нельзя получить непосредственно, а только спрогнозировать. Во-вторых, использование машинного обучения позволяет захватывать более сложные взаимосвязи между несколькими факторами, которые могут быть упущены при использовании линейных ковариат.

Пример 3. Сгенерируем нормально распределённые данные для контрольной и экспериментальной групп, где экспериментальная группа имеет смещение влево. Добавим шум к обеим группам на основе ковариаты, скрывая наш ATE. Затем проводим t-тест на этих зашумлённых данных. Проведём оценку, используя метод наименьших квадратов для каждой группы, чтобы предсказать зашумлённые данные на основе ковариаты, получим остатки. Эти остатки представляют собой часть данных, не объясняемую ковариатой. Наконец, мы проводим t-тест на остатках. В итоге метод CUPAC помогает получить в среднем более статистически значимые оценки: базовый p-value = 0.37, p-value CUPAC = 0.04. Более подробно — смотри в файле *cuped_cupac.ipynb*.

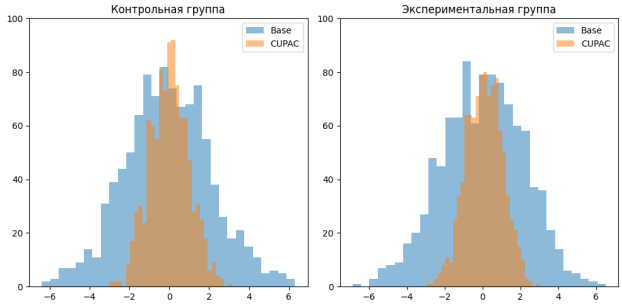

Рис. 3. Дисперсия остатков стандартного метода и метода CUPAC

CUPAC направлен на корректировку результатов с учётом различий между группами, что может помочь в выявлении более точного эффекта.

4. Выводы

- Задача ускорения A/B-тестов по фиксированному временному промежутку, по сути, является задачей повышения чувствительности.
- Повышать чувствительность теста можно с помощью методов стратификации, CUPED и CUPAC.
- Выбор способа повышения чувствительности зависит от ваших ресурсных возможностей и результатов проведения A/A-тестов.

Список литературы

1. Statistical methods for research workers. 4th ed. // ScienceDirect. 2006. URL: https://www.sciencedirect.com/science/articl (дата обращения: 15.10.2024).

2. URL: https://arxiv.org/pdf/1208.2301 (дата обращения: 15.10.2024).

3. Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data. URL: https://exp-platform.com/Documents/2013-02-CUPED-ImprovingSensitivityOfControlledExperiments.pdf (дата обращения: 15.10.2024).

4. Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix. URL: https://www.kdd.org/kdd2016/p xieA.pdf (дата обращения: 15.10.2024).

5. Deep Dive Into Variance Reduction // Experimentation Platform. 2022. URL: https://www.microsoft.com/en-us/research/group/experimentation-platform-exp/articles/deep-dive-into-variance-reduction/ (дата обращения: 15.10.2024).