**Санкт–Петербургский государственный университет**

**Ненахов Иван Владимирович**

# Выпускная квалификационная работа
# Методы A/B тестирования веб-форм

Уровень образования: бакалавриат

Направление 02.03.02 «Фундаментальная информатика и информационные технологии»

Основная образовательная программа «Программирование и информационные технологии»

Научный руководитель:
доцент кафедры технологии программирования,
кандидат технических наук
Блеканов Иван Станиславович

Рецензент:
кандидат физико-математических наук,
доцент кафедры компьютерных технологий и систем
Коровкин Максим Васильевич

Санкт-Петербург

2024

**Saint-Petersburg State University**

**Nenakhov Ivan Vladimirovich**

# Final qualifying work
# Methods of A/B testing of web forms

Education level: bachelor's degree

Education program 02.03.02 «Fundamental computer science and information technology»

Scientific supervisor:
Candidate of Engineering, Associate Professor
Blekanov I.S.

Reviewer:
Candidate of Physics and Mathematics, Associate Professor
Korovkin M.V.

Saint-Petersburg

2024

# Введение

В бизнесе постоянно рождается множество идей. Главное предназначение данных идей - улучшение ключевых показателей эффективности компании. Главная цель компании, которая тестирует эти идеи - внедрение только успешных концепций. В современном мире бизнеса и интернет-маркетинга решения принимаются на основе научных данных и исследованиях пользовательского поведения. A/B тестирование является дисциплиной, которая позволяет компаниям получать эти научные данные. A/B тестирование - это эксперимент, проводимый с целью определения, какая из двух (или более) вариантов страницы сайта, электронной коммерции, приложения или рекламы улучшает целевую метрику.

Кратко процесс можно описать так:

1. Разбиение выборки пользователей на группы, обычно на контрольную и тестовую.
2. Демонстрация каждой группе одной из вариаций продукта
3. Получение результата и его интерпретация

Веб-форма - часть веб-сайта или приложения, которая позволяет пользователям вводить и отправлять данные на сервер. Веб-форма, как правило, состоит из нескольких последовательных страниц. Она может содержать поля для ввода текста, переключатели для выбора вариантов, кнопки для загрузки файлов и другие элементы управления. Веб-формы часто используются в интернет-магазинах при оформлении заказов. Ввиду этого компании стараются сделать их максимально удобными и простыми, чтобы клиенты не уходили на этапе заполнения заявки. Это очень важно, потому что заполнение веб-формы часто воспринимается как что-то рутинное. Любые сложности и неудобства будут отталкивать потенциальных клиентов. Для создания и оптимизации такой формы и применяется A/B тестирование. Пример того, как выглядит веб-форма, представлен на рисунке 1.

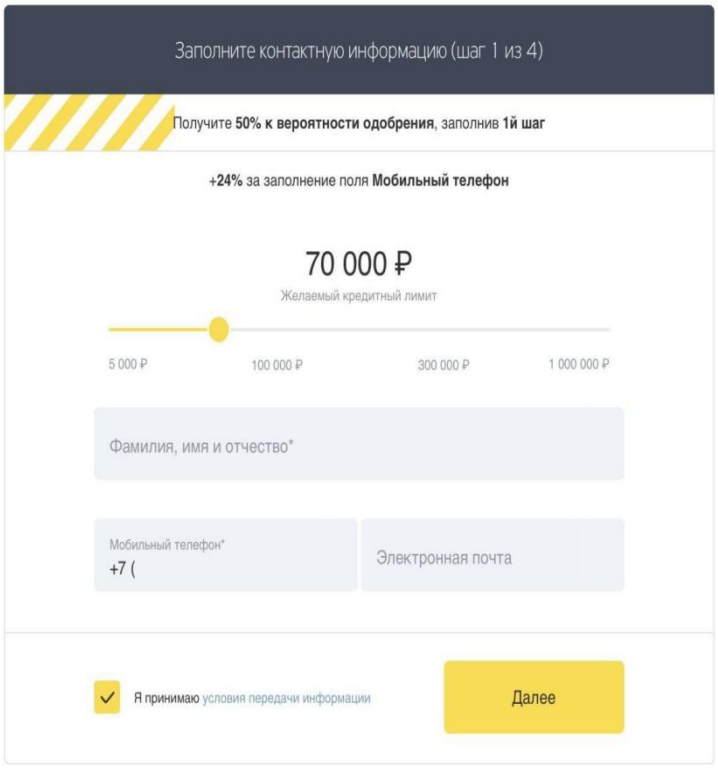


Рисунок 1 - пример веб-формы

Для веб-форм тестируется множество метрик, например:

1. Средняя длительность заполнения формы
2. Средняя длительность заполнения определённой страницы веб-формы
3. Конверсия в заполнение формы
4. Конверсия в заполнение какой-то конкретной страницы веб-формы
5. Конверсия в покупку(то есть доля форм, которые завершились покупкой, от всех форм)
6. Частота появления какого-либо алерта на форму или страницу

# Актуальность работы

A/B тестирование даёт возможность разработчикам и маркетологам находить самые эффективные методы привлечения трафика на сайт и увеличения количества клиентов, а также повышения удовлетворённости пользователей и улучшения пользовательского опыта. В наши дни A/B тестирование стало обязательным этапом создания и оптимизации сайтов, приложений, электронных магазинов, социальных сетей и многих других онлайн-платформ. Данное тестирование позволяет избавиться от догадок при выборе наиболее подходящей идеи и даёт чёткий ответ, обоснованный математически. Сейчас в крупных компаниях существуют целые аналитические команды, которые этим занимаются, что подтверждает актуальность метода. Можно выделить наиболее популярные примеры A/B тестирования:

1. A/B тестирование цен - магазины используют A/B тестирование для определения наиболее эффективной цены на какой-либо продукт. Для этого тестируются несколько разных цен на товар, чтобы увидеть, какую цену лучше всего принимают клиенты.
2. A/B тестирование изображений - маркетинговые компании проводят A/B тестирование, чтобы понять, какое изображение привлекает больше клиентов.
3. A/B тестирование заголовков - компании тестируют заголовки объявлений и статей, чтобы определить, какой заголовок лучше привлекает внимание.
4. A/B тестирование веб-форм - компании тестируют веб-формы с целью исследовать ключевые метрики как на каждой отдельной странице, так и на всей форме в целом. Важность их усовершенствований была описана прежде.

Данная тема имеет множество научных исследований. Ярким примером является работа Ron Kohavi “Controlled experiments on the web: survey and practical guide. Data Mining and Knowledge Discovery”. В ней описывается практическое руководство по проведению контролируемых экспериментов в Интернете. Авторы рассматривают методологии проведения A/B тестирования на сайтах и приложениях, подробно описывают основные понятия в этой области и приводят многочисленные рекомендации и инструменты для успешного проведения A/B тестирования. Статья также содержит наглядные примеры исследований крупных мировых компаний, подтверждая этим актуальность метода.

# Цели и задачи работы

Целью этой работы является исследование возможностей современных подходов A/B тестирования и их применение в отношении веб-форм.

Для достижения вышеуказанной цели были поставлены следующие задачи:

- Анализ актуальных исследований в области A/B тестирования
- Анализ пользовательских данных веб-форм, которые планируется исследовать в рамках данной работы.
- Программная реализация изученных методов
- Тестирование собранных данных
- Подведение итогов

# Глава 1. Анализ современных методов A/B тестирования

## *1.1* Процесс тестирования

A/B тестирование является надёжным и эффективным инструментом для оптимизации взаимодействия бренда с потребителями в цифровой среде. Путём проведения экспериментов и проверки гипотез бизнеса мы оптимизируем различные точки контакта с потребителями, такие как сайт, приложение, рекламные баннеры, тексты рекламных сообщений и другие. В данной работе будем рассматривать A/B тестирование с использованием двух вариаций.

A/B тест начинается с обозначения проблемы, которую мы хотим решить с его помощью. Пример может быть следующим. У нас есть веб-форма, которая состоит из нескольких страниц. На первой странице перечислены тарифы мобильного интернета, а также поля для ввода персональных данных. На последующих страницах также находятся поля для ввода некоторых данных, а на последней имеется кнопка, при нажатии которой форма считается заполненной. Заполненную форму будем называть полной заявкой. Проблема заключается в том, что конверсия из визита (открытия веб-формы) в полную заявку сильно ниже, чем на других формах на нашем сайте.

Далее формируются гипотезы. Гипотеза подразумевает решение нашей проблемы. Зачастую компании используют определённые шаблоны для её формирования. Например: “если мы сделаем что-то, произойдёт это, и поэтому на наших метриках это отразиться таким образом”. Продолжая пример с веб-формой это может выглядеть следующим образом. Если мы добавим навигацию в верхней части формы, клиенту станет удобнее понимать, сколько осталось этапов до конца формы, и конверсия в полную заявку вырастет.

Для формирования гипотез прибегают к следующим вещам: анализ данных, проведение опросов, UX-исследования, CustDev-исследования,

анализ рынка и конкурентов, анализ отзывов и обращений в техническую поддержку, и прочее. Очевидно, что для этого нужны не только аналитики, но и другие команды.

Следующим этапом является продумывание возможных вариантов неудачи выбранного решения. Например, может увеличиться время загрузки страницы. Все варианты не всегда возможно предугадать заранее, но хоть какие-то стоит выделить, чтобы упростить себе работу в будущем.

Далее мы определяем метрики, они разделяются на основные и дополнительные. Основная — это та, по который мы определим, какой из вариантов победил. Дополнительные метрики позволяют проверить, что тест не ухудшил другие процессы. В примере с веб-формой основной метрикой является конверсия в полную заявку. В качестве дополнительной можно рассмотреть конверсию в утилизацию. Конверсия в утилизацию — это отношения количества “утилизированных пользователей” к общему числу пользователей. Понятие утилизированного пользователя может различаться в зависимости от ситуации. В нашем случае это человек, который заполнил и отправил форму, но дальше не воспользовался продуктом, а общее число пользователей — это все люди, которые заполнили веб-форму. Данная метрика позволяет проверять, не привлёк ли новый эффект незаинтересованную аудиторию.

Существует огромное количество метрик, которые используются в A/B тестировании. Всё это зависит от специфики продукта. Можно выделить основные:

1. ARPU, Average Revenue Per User - средний доход с одного клиента за некоторый промежуток времени.
2. ARPPU, Average Revenue Per Paying User - средний доход с одного платящего клиента за некоторый промежуток времени.
3. Различные конверсии
4. Длительность сессии

Следующим шагом принято производить приоритизацию теста. В больших компаниях проводится огромное количество A/B тестов. Для одного продукта может быть запланировано несколько A/B тестов, но одновременно на одних и тех же пользователях их проводить нельзя, потому что в таком случае мы не поймём какая именно гипотеза повлияла на результат. В связи с этим мы приоритизируем тесты, чтобы понимать в каком порядке их запускать. Примером может служить методология RICE. В ней для каждого теста просчитывается формула (1). Первым запускается тот тест, у которого получается самое большое значение. Reach - количество пользователей, на которых мы хотим повлиять. Impact - оценка того, насколько сильно повлияет наше изменение на поведение пользователя. Оценка может равняться одному из следующих значений: 3 – «массовое влияние», 2 – «высокое», 1 – «среднее», 0.5 – «низкое», 0.25 – «минимальное». Confidence – оценка того, насколько сильно мы уверены в своём решении. Данная оценка имеет значение в диапазоне [0.01, 10]. Effort показывает сколько человеко/часов займёт подготовка изменения.

$$\frac{REACH \times IMPACT \times CONFIDENCE}{EFFORT} \quad (1)$$

Далее нужно определить статистический критерий и длительность теста. Под длительностью теста подразумевается количество пользователей, которые будут участвовать в тесте. Чтобы определить статистический критерий, нужно провести некоторый анализ на исторических данных. Нужно понимать какой характер они имеют. Для этого проводят симуляции A/A и A/B тестирования. Описание этих процессов будет дано в практической части. Длительность теста рассчитывается только после определения статистического критерия. Чтобы её рассчитать, нужно определить минимальный обнаруживаемый эффект и некоторые другие параметры, которые зависят от выбранной метрики. Минимальный

обнаруживаемый эффект в A/B тестировании - это наименьшее значение, которое компания ожидает получить от изменений, внесённых в тестируемую версию сайта или приложения. Этот показатель обычно задаётся заранее и определяется на основе исторических данных, бизнес-целей и ожиданий. Таким образом, выдвигается две гипотезы. Нулевая гипотеза теста гласит, что между контрольной и тестовой вариациями нет статистически значимой разницы, а альтернативная - что разница есть. В A/B тестировании применяется множество статистических критериев, основными являются:

1. Z-критерий для двух пропорций
2. T-критерий Стьюдента
3. U-критерий Манна-Уитни
4. Permutation тест
5. Bootstrap тест

Помимо этого, используются и другие критерии математической статистики. Например, тест Шапиро-Уилка на нормальность распределения, тест Фишера на равенство дисперсий двух нормальных распределений, и другие.

Далее собирается необходимое число данных и считается статистика. Полученный результат нужно интерпретировать, для этого используют p-value. P-value (уровень значимости) - это вероятность получить результат, который более экстремальный, чем тот, который был получен, при условии, что нулевая гипотеза верна. В контексте A/B тестирования, p-value используется для определения статистической значимости различий между контрольной и тестовой группами. Если p-value ниже выбранного уровня значимости (обычно равняется 0,05), то считается, что различия между группами являются статистически значимыми (есть значимое отличие между двумя группами). Если p-value выше уровня значимости, то нулевая гипотеза не может быть отвергнута и различия между группами не статистически значимы.

## 1.2 Современные подходы в A/B тестировании

### 1.2.1 Снижение дисперсии

В A/B тестировании большое внимание уделяют уменьшению дисперсии тестируемой метрики. Чем меньше дисперсия, тем более чувствительным будет наш тест. Также это уменьшает необходимое количество наблюдений в выборках. Существуют различные подходы для уменьшения дисперсии, наиболее актуальные - стратификация и CUPED. При этом важно отметить, что данные приёмы сохраняют несмещённость оценок. Посвящённые этим методам статьи показывают, что в худшем случае дисперсия не изменится, а в большинстве случаев уменьшится. Влияние на чувствительность оказывается крайне положительным.

Суть стратификации заключается в том, что мы предполагаем, что пользователей можно разбить на страты по каким-то признакам. Например, по регионам. Если мы полагаем, что между такими стратами есть разница в исследуемой метрике, то стратификация сильно уменьшает дисперсию. В данном методе для тестовой и экспериментальной выборок мы берём представителей каждой страты в количестве пропорциональном их долям в общей популяции. Это называется стратифицированным семплированием. Далее мы считаем для каждой выборки не обычное среднее, а стратифицированное, которое в свою очередь является просто взвешенным, где вес страты - доля страты в популяции. Популяция — это все пользователи, на которых мы можем повлиять нашим экспериментом. Уже для полученных метрик можно применить, например, критерий Стьюдента. Оценка стратифицированного среднего является несмещенной, оценка дисперсии этой оценки известна. Наиболее подробно подход описан в статье Huizhi Xie “Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix”.

В CUPED для уменьшения дисперсии мы стараемся избавиться от эффекта предэкспериментальных данных. Мы не хотим, чтобы они как-то

влияли на текущий тест. Делаем это с помощью некоторых математических преобразований, основанных на свойствах дисперсии и ковариации. Для этого для каждого пользователя, участвующего в A/B эксперименте нужно знать данную метрику на предэкспериментальном периоде и на текущем. Y - метрика пользователя на текущем эксперименте, X - метрика пользователя на предэкспериментальном периоде. В формуле (2) показано необходимое преобразование для получения новой метрики. В формуле (3) выведена дисперсия новой метрики. Видно, что это дисперсия меньше или равна дисперсии исходной метрики. После проведённых преобразований можно применить критерий Стьюдента. Более подробное описание подхода описано в статьях Alex Deng “Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix” и Simon Jackson “How Booking.com increases the power of online experiments with CUPED”.

$$Y_{new} = Y - \theta X \tag{2}$$

$$\theta = \frac{cov(Y, X)}{var(X)}$$

$$var(Y_{new}) = var(Y) * (1 - corr^2(Y, X)) \tag{3}$$

### 1.2.2 Работа с ratio-метриками

Ratio-метрика - это метрика отношения непользовательского уровня с зависимыми наблюдениям, но которая явным образом выражается через отношение сумм соответствующих пользовательских сигналов. Проблема такой метрики заключается в том, что мы не можем оценивать её дисперсию среднего привычным способом. Существует специальный дельта метод, который делает эту оценку. В формуле (4) показана эта оценка дисперсии. Используя данные оценки можно выполнить пропорциональный z-тест. Более подробно метод описан в статьях Roman Budylin “Consistent

Transformation of Ratio Metrics for Efficient Online Controlled Experiments” и Alex Deng “Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas”.

$$var(\frac{X}{Y}) \approx \frac{1}{E[Y]^2} Var(X) + \frac{E[X]^2}{E[Y]^4} var(Y) - 2 * \frac{E[X]}{E[Y]^3} cov(X, Y) \tag{4}$$

Следующий подход называется линеаризацией. Он преобразует ratio-метрику с зависимыми наблюдениями в среднюю пользовательскую с независимыми. В формуле (5) показано это преобразование, где буквой u обозначается уникальный пользователь. Разница линеаризованных метрик всегда сохраняет сонаправленность с изменением в целевой ratio-метрике. Линеаризованные пользовательские сигналы уже можно считать независимыми и для них определять статистическую значимость тестом Стьюдента. При этом, значения p-value для линеаризованной метрики будут консистентны значениям, полученным с помощью дельта-метода на исходной ratio-метрике.

$$\frac{\sum\limits_{u} X(u)}{\sum\limits_{u} Y(u)} \approx avg_u(L(X(u), Y(u))) \tag{5}$$

$$L(X(u), Y(u)) \approx X(u) - kY(u)$$

$$k = \frac{\sum\limits_{u} X(u)}{\sum\limits_{u} Y(u)} \tag{6}$$

Примечательно, что в формуле (6) мы используем данные только из контрольной выборки. Можно также отметить, что значение k можно

вычислить с помощью регрессионной модели, в которой X(u) - независимая переменная, Y(u) - зависимая переменная, k - коэффициент перед независимой переменной.

Ещё одним подходом является бутстрап ratio-метрики. Это такой же A/B тест с использованием статистического метода “бутстрап”, только на каждом шаге мы теперь вычисляем ratio-метрики для контрольной и тестовой выборок. Важно отметить, что теперь для каждого шага цикла мы сэмплируем выборки с повторениями по объектам, а не по наблюдениям. Если объект оказался в тестовой или контрольной группе, то и все его действия будут в этой группе.

В статье Keyu Nie “Dealing With Ratio Metrics in A/B Testing at the Presence of Intra-User Correlation and Segments” описывается алгоритм работы с ratio-метриками и предлагается их несмещенная оценка математического ожидания. Главное отличие от предыдущих способов заключается в том, что в статье доказано, что данная оценка обладает наименьшей дисперсией, то есть оценка является эффективной. Также в данной статье приведены их оценки дисперсий.

### 1.2.3 Применение метода многоруких бандитов

Суть метода заключается в том, что мы динамически в процессе A/B теста определяем в какую из групп отправить текущего пользователя. Делается это, потому что мы не хотим отправлять большое количество трафика на заведомо плохие варианты. Наибольшую востребованность данный метод получил в тестах конверсии. Суть заключается в следующем. Распределение конверсии удобно приближать с помощью бета-распределения. В самом начале для каждого варианта теста мы вводим два параметра бета-распределения и приравниваем их к единице (7). При таких параметрах бета-распределение становится обычным равномерным. Мы задаём такие параметры, потому что изначально мы не знаем как ведёт себя конверсия каждого варианта. Далее для каждого нового пользователя мы

повторяем следующие шаги. Сначала для каждого варианта теста сэмплируем конверсию из бета-распределения с параметрами, которые актуальны на текущий момент (8). Далее мы выбираем вариант с наибольшей конверсией и показываем его текущему пользователю. Далее мы получаем либо y=1, что означает, что пользователь “прошёл” конверсию, либо y=0, что означает иное. Мы обновляем параметры данного варианта теста по формуле (9). Когда мы исчерпаем количество пользователей, которое мы можем себе позволить для проведения A/B теста, тест завершается. Побеждает вариант с большей конверсией на данный момент. Варианта “статистически значимой разницы не наблюдается” здесь нету.

$$\forall i, \alpha_i = \beta_i = 1 \tag{7}$$

$$\theta_i \sim Beta(\alpha_i, \beta_i) \tag{8}$$

$$k = argmax(\theta_i)$$

$$\alpha_i = \alpha_i + y; \quad \beta_i = \beta_i + 1 - y \tag{9}$$

Доказано, что в случае с бета-распределением, как описано выше, при количестве пользователей в A/B тесте, стремящемся к бесконечности, вероятность того, что алгоритм сойдётся к верному решению равна единице.

Так же существуют модификации метода, позволяющие следить и за остальными группами, и в случае, если нам вначале не повезло, или сменилась мода на выбор варианта, мы бы могли почувствовать это и исправить ситуацию. Полное описание алгоритма содержится в статье Maryam Aziz “On Multi-Armed Bandit Designs for Dose-Finding Clinical Trials.

Главным недостатком данного метода является невозможность получить точное значение количества пользователей, необходимое для

фиксации статистически значимой разницы при заданном минимальном обнаруживаемом эффекте, уровне значимости и мощности.

### 1.2.4 Проблема подглядывания

Если мы имеем возможность наблюдать за динамикой изменения p-value в процессе A/B теста, то может появиться желания преждевременно остановить эксперимент всякий раз, когда мы видим статистически значимый результат. Но это будет ошибкой. Чем чаще мы смотрим на промежуточные результаты A/B теста с готовностью принять на их основе решение, тем выше становится вероятность, что критерий покажет значимую разницу, когда её нет. Доказано теоретически, что если запустить A/A тест (то есть показывать обоим группам одну и ту же версию страницы), то p-value хотя бы раз пересечёт порог при количестве пользователей, стремящемся к бесконечности. Однако, существуют методы, которые позволяют не ждать до конца теста.

В статье Peter O'Brien "A Multiple Testing Procedure for Clinical Trials" описываются методы Pocock и O'Brien-Fleming. Для обеих поправок нужно заранее знать максимальный срок теста и количество проверок p-value между запуском и окончанием теста. Причём проверки должны происходить через равные количества наблюдений. В статье Wenru Zhou "Approaches to analyzing binary data for large-scale A/B testing" на исследованиях показано, что поправки работают корректно. В методе Pocock мы подводим итоги тестов через равные количества наблюдений, но при некотором сниженном (более строгом) уровне значимости. В методе O'Brien-Fleming же уровень значимости изменяется в зависимости от номера проверки. Чем раньше мы пытаемся завершить тест, тем более жёсткий порог будет применён.

В статье Ramesh Johari "Peeking at A/B Tests" описан метод SPRT и его модификация mSPRT. Это последовательный тест, который в своём алгоритме использует функцию максимального правдоподобия. Данная поправка является наиболее актуальной на данный момент в индустрии.

### 1.2.5 Анализ p-value на исторических A/B тестах

Когда в компании проведено уже большое количество тестов, можно оценить общую картину и понять то, на сколько качественно они проходят.

Формула (10) показывает, какое по определению мы должны иметь распределение p-value на конкретном тесте, где гипотеза H0 верная. Видно, что это функция равномерного распределения. То есть, если взять все p-value в тестах, где мы не получили статистически значимого результата, мы должны увидеть их равномерное распределение.

$$P(pvalue < \alpha | H_0) = \alpha \quad \forall \alpha \in [0,1] \tag{10}$$

$$P(pvalue \geq A | H_1) = B \Rightarrow P(pvalue < A | H_1) = 1 - B \tag{11}$$

С другой стороны, в формуле (11) показано распределение p-value при условии, что гипотеза H1 верная. A - вероятность ошибки первого рода, B - вероятность ошибки второго рода. Можно сделать вывод, что если взять все p-value в тестах, где мы получили статистически значимый результат, то эти p-value должны “жаться” к нулю.

# Глава 2. Программная реализация A/B тестирования

## 2.1 A/B тестирование времени заполнения формы

Для данного теста используется датасет, который содержит данные о том, как пользователи ведут себя в процессе заполнения веб-формы некоторого сервиса. Данную веб-форму они заполняют каждый месяц, она обеспечивает актуализацию некоторых пользовательских данных. Были предложены некоторые улучшения, которые предлагается проверить в A/B тесте. Датасет состоит из следующих столбцов: device - какое устройство использует пользователь при заполнении веб-формы, может быть 2 варианта – “desktop” или “mobile”; source - источник, из которого пользователь пришёл в сервис изначально; ab_version - может быть либо “test” (в данном случае пользователям отображается версия веб-формы с изменениями), либо “control” (в данном случае пользователям отображается версия веб-формы без изменений); value - количество времени в секундах, которое понадобилось пользователю для заполнения веб-формы; value_before - среднее количество времени в секундах за 3 предыдущих месяца, которое пользователю понадобилось для заполнения веб-формы. Гипотеза заключается в том, что новые изменения уменьшат среднее время заполнения веб-формы. Весь код реализации написан на языке программирования Python.

На рисунке 2 показана разбивка по тесту и контролю с количеством пользователей, средним значением времени и стандартным отклонением времени. На рисунках 3 и 4 дополнительно добавлены разбивки по source и device соответственно. Можно заметить зависимость среднего времени как от источника, из которого пришли пользователи, так и от устройства. Это говорит о том, что стратифицированный тест Стьюдента может давать лучшие результаты на данном датасете.

| | value | | |
|---|---|---|---|
| | count | mean | std |
| ab_version | | | |
| control | 1533 | 331.198450 | 57.421284 |
| test | 1475 | 314.222026 | 54.423169 |

Рисунок 2 - разбивка на тест и контроль

| | | value | | |
|---|---|---|---|---|
| | | count | mean | std |
| ab_version | source | | | |
| control | Source1 | 204 | 337.202754 | 60.568547 |
| | Source2 | 595 | 338.789279 | 58.684842 |
| | Source3 | 734 | 323.376350 | 54.464641 |
| test | Source1 | 209 | 322.723489 | 57.599522 |
| | Source2 | 523 | 321.306666 | 57.514582 |
| | Source3 | 743 | 306.843732 | 50.163992 |

Рисунок 3 - дополнительная разбивка по источнику

| | | value | | |
|---|---|---|---|---|
| | | count | mean | std |
| ab_version | device | | | |
| control | Desktop | 1238 | 332.209315 | 57.490743 |
| | Mobile | 295 | 326.956244 | 57.030480 |
| test | Desktop | 1161 | 315.145845 | 54.264718 |
| | Mobile | 314 | 310.806247 | 54.957239 |

Рисунок 4 - дополнительная разбивка по устройству

На рисунке 5 показаны распределения value_before для теста и контроля. Они выглядят достаточно схоже. Оно и очевидно, так как на этих периодах у пользователей были одинаковые версии веб-формы. Обратное можно заметить на рисунке 6, на котором изображены распределения value для теста (красный цвет) и контроля (синий цвет). Но показать является ли результат статистически значимым сможет только A/B тест.


Рисунок 5 - распределения value_before

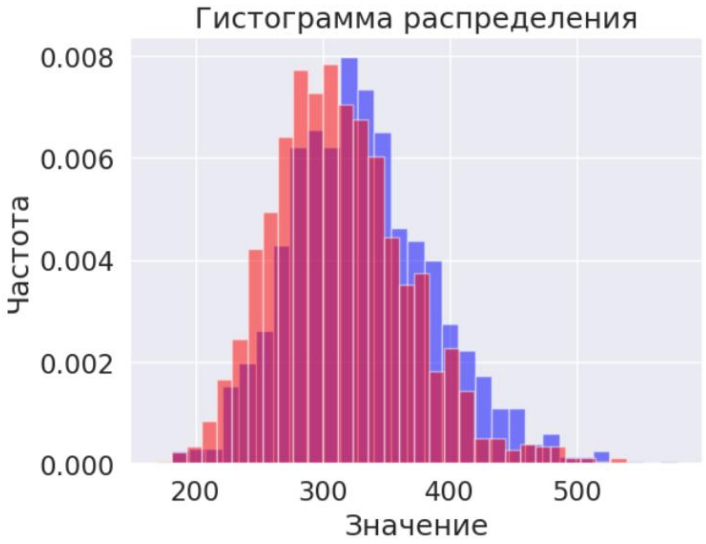


Рисунок 6 - распределения value

Проверим тест и контроль на нормальность распределения и равенство дисперсий. Для проверки на нормальность использовался qq-plot, показанные на рисунках 7 и 8. Оба графика показывают, что распределения нормальные. Для проверки на равенство дисперсий использовали тест Фишера на исторических данных(value_before) с нулевой гипотезой о том, что дисперсии равны, получили p-value равный 0.31, который больше, чем 0.05. Соответственно, дисперсии равны. Это дает нам право использовать обычную реализацию теста Стьюдента с предположением о нормальных распределениях и равных дисперсиях.

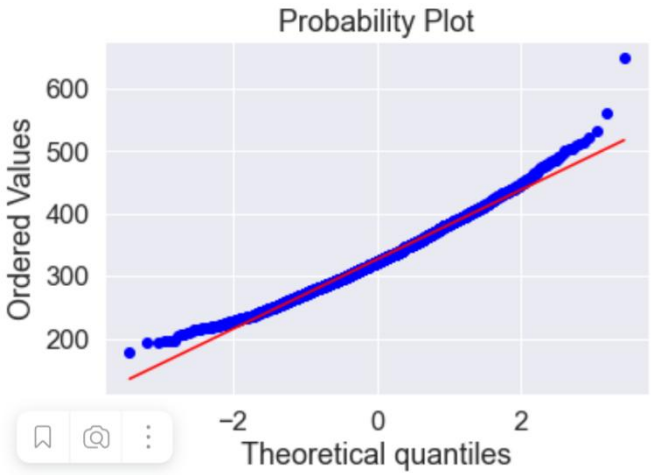


Рисунок 7 - qq-plot для теста

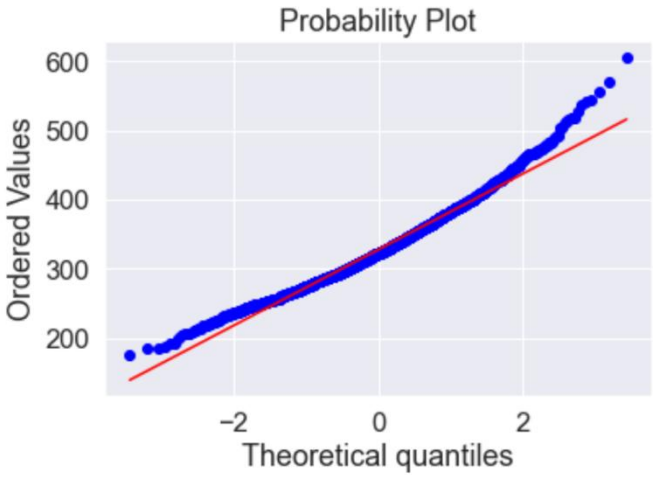


Рисунок 8 - qq-plot для контроля

Для всех статистических критериев, кроме msprt была реализована следующая симуляция A/B теста. В цикле перебираются несколько минимальных обнаруживаемых эффектов(mde), которые должны считаться статистически значимыми. В данном A/B тесте это: 0.06, 0.05, 0.04, 0.03, 0.02. На каждой такой итерации цикла создаётся искусственная тестовая выборка, которая состоит из значений контроля, от которых вычли среднее значение контроля, умноженное на mde, а сам контроль остаётся прежним. Далее рассчитывается минимальный размер выборки, необходимый для детектирования заданного эффекта. Ошибку первого рода приравняем к 0.05, ошибку второго рода к 0.2. После этого создаётся цикл и в нем N раз для контроля и искусственного теста берутся сэмплы с повторениями, которые

используются для расчёта статистики определённого критерия, далее на каждом из N шагов рассчитываются p-value, которые показывают являются ли результаты статистически значимыми. Все N результатов усредняются и получается эмпирическая мощность, которая при корректности работы теста должна равняться теоретической.

Отличие симуляции A/A теста от симуляции A/B теста заключается только в том, что искусственный тест здесь просто равняется контролю. В A/A тесте проверяется равна ли эмпирическая ошибка первого рода теоретической. Для msprt используются отдельные симуляции, поскольку это последовательный критерий и там невозможно заранее определить размер выборки. Также в симуляции A/A теста для msprt не перебираются различные mde. В этом нет необходимости, так как размер выборки заранее мы не рассчитываем.

Далее для всех статистических критериев будут показаны результаты симуляций A/A и A/B тестов в виде таблиц. Успешными считаются те варианты, в которых эмпирические уровни значимости и мощности равны теоретическим. Обозначения полей в таблице: sig - теоретическая ошибка первого рода, empirical_sig - эмпирическая ошибка первого рода, полученная в симуляции A/A теста; empirical_power - эмпирическая мощность, полученная в симуляции A/B теста; left_real_level и right_real_level - соответственно, левая и правая границы доверительного интервала либо empirical_sig, либо empirical_power; mde - минимальный обнаруживаемый эффект; sample_size - размер семпла, который используется при данном mde.

На рисунках 9 и 10 показаны результаты для теста Стьюдента. Эмпирическая ошибка первого рода очень близка к теоретической, эмпирическая мощность же показывает результаты более высокие, чем теоретическая.

A/A simulation for t-test:

| | sig | empirical_sig | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| **0** | 0.05 | 0.09 | 0.06 | 103 | 0.048073 | 0.162262 |
| **1** | 0.05 | 0.05 | 0.05 | 149 | 0.021544 | 0.111750 |
| **2** | 0.05 | 0.04 | 0.04 | 232 | 0.015663 | 0.098371 |
| **3** | 0.05 | 0.08 | 0.03 | 413 | 0.041093 | 0.149981 |
| **4** | 0.05 | 0.05 | 0.02 | 929 | 0.021544 | 0.111750 |

Рисунок 9 - A/A симуляция для теста Стьюдента

A/B simulation for t-test:

| | power | empirical_power | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| **0** | 0.8 | 0.81 | 0.06 | 103 | 0.722212 | 0.874852 |
| **1** | 0.8 | 0.80 | 0.05 | 149 | 0.711171 | 0.866633 |
| **2** | 0.8 | 0.85 | 0.04 | 232 | 0.767164 | 0.906940 |
| **3** | 0.8 | 0.78 | 0.03 | 413 | 0.689296 | 0.849987 |
| **4** | 0.8 | 0.81 | 0.02 | 929 | 0.722212 | 0.874852 |

Рисунок 10 - A/B симуляция для теста Стьюдента

На рисунках 11 и 12 показаны результаты для CUPED теста Стьюдента. Как было сказано ранее, у нас имеются усреднённые данные для каждого пользователя на периодах, предшествующих A/B тесту. Это можно использовать в методе CUPED для увеличения эмпирической мощности. Эмпирическая ошибка первого рода получилась приблизительно равной теоретической, эмпирическая мощность же показывает результат равный единице, что говорит о том, что тест не совершает ошибок второго рода.

A/A simulation for cuped t-test:

| | sig | empirical_sig | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| **0** | 0.05 | 0.06 | 0.06 | 103 | 0.027786 | 0.124768 |
| **1** | 0.05 | 0.03 | 0.05 | 149 | 0.010255 | 0.084519 |
| **2** | 0.05 | 0.05 | 0.04 | 232 | 0.021544 | 0.111750 |
| **3** | 0.05 | 0.07 | 0.03 | 413 | 0.034319 | 0.137495 |
| **4** | 0.05 | 0.09 | 0.02 | 929 | 0.048073 | 0.162262 |

Рисунок 11 - A/A симуляция для теста CUPED

A/B simulation for cuped t-test:

| | power | empirical_power | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.8 | 1.0 | 0.06 | 103 | 0.963007 | 1.0 |
| 1 | 0.8 | 1.0 | 0.05 | 149 | 0.963007 | 1.0 |
| 2 | 0.8 | 1.0 | 0.04 | 232 | 0.963007 | 1.0 |
| 3 | 0.8 | 1.0 | 0.03 | 413 | 0.963007 | 1.0 |
| 4 | 0.8 | 1.0 | 0.02 | 929 | 0.963007 | 1.0 |

Рисунок 12 - A/B симуляция для теста CUPED

Ранее было отмечено, что в данных есть некоторая зависимость времени прохождения веб-формы от типа устройства и источника. Это говорит о том, что стратифицированный тест Стьюдента может дать более лучшие результаты чем обычный. На рисунках 13 и 14 показаны результаты теста, в котором устройство используется как страта. Заметны улучшения в эмпирической ошибке первого рода. На рисунках 15 и 16 показаны результаты теста, в котором источник используется как страта. Заметны улучшения как эмпирической ошибки первого рода, так и эмпирической мощности.

A/A simulation for strat t-test:

| | sig | empirical_sig | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.05 | 0.02 | 0.06 | 103 | 0.005502 | 0.070012 |
| 1 | 0.05 | 0.05 | 0.05 | 149 | 0.021544 | 0.111750 |
| 2 | 0.05 | 0.01 | 0.04 | 232 | 0.001767 | 0.054486 |
| 3 | 0.05 | 0.01 | 0.03 | 413 | 0.001767 | 0.054486 |
| 4 | 0.05 | 0.07 | 0.02 | 929 | 0.034319 | 0.137495 |

Рисунок 13 - A/A симуляция для стратифицированного теста Стьюдента, где устройство является стратой

A/B simulation for strat t-test:

| | power | empirical_power | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.8 | 0.82 | 0.06 | 103 | 0.733326 | 0.882998 |
| 1 | 0.8 | 0.74 | 0.05 | 149 | 0.646290 | 0.815953 |
| 2 | 0.8 | 0.76 | 0.04 | 232 | 0.667677 | 0.833087 |
| 3 | 0.8 | 0.79 | 0.03 | 413 | 0.700200 | 0.858343 |
| 4 | 0.8 | 0.83 | 0.02 | 929 | 0.744520 | 0.891064 |

Рисунок 14 - A/B симуляция для стратифицированного теста Стьюдента, где устройство является стратой

A/A simulation for strat t-test:

| | sig | empirical_sig | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.05 | 0.08 | 0.06 | 103 | 0.041093 | 0.149981 |
| 1 | 0.05 | 0.05 | 0.05 | 149 | 0.021544 | 0.111750 |
| 2 | 0.05 | 0.09 | 0.04 | 232 | 0.048073 | 0.162262 |
| 3 | 0.05 | 0.06 | 0.03 | 413 | 0.027786 | 0.124768 |
| 4 | 0.05 | 0.07 | 0.02 | 929 | 0.034319 | 0.137495 |

Рисунок 15 - A/A симуляция для стратифицированного теста Стьюдента, где источник является стратой

A/B simulation for strat t-test:

| | power | empirical_power | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.8 | 0.82 | 0.06 | 103 | 0.733326 | 0.882998 |
| 1 | 0.8 | 0.82 | 0.05 | 149 | 0.733326 | 0.882998 |
| 2 | 0.8 | 0.80 | 0.04 | 232 | 0.711171 | 0.866633 |
| 3 | 0.8 | 0.78 | 0.03 | 413 | 0.689296 | 0.849987 |
| 4 | 0.8 | 0.76 | 0.02 | 929 | 0.667677 | 0.833087 |

Рисунок 16 - A/B симуляция для стратифицированного теста Стьюдента, где источник является стратой

На рисунках 17 и 18 показаны результаты для теста Бутстрап. Обычно он используется в ситуациях, когда мы не можем использовать обычный тест Стьюдента. Например, если наши данные не подчиняются нормальному распределению или дисперсии теста и контроля не равны. Поэтому очевидно, что эмпирические результаты так же близки к теоретическим, но заметных улучшений не выделяется. Также можно отметить, что здесь была использована формула расчёта размера выборки для обычного теста Стьюдента. Так как симуляции показали удовлетворительные результаты, то можно использовать эту формулу с имеющимся датасетом.

| | sig | empirical_sig | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.05 | 0.06 | 0.06 | 103 | 0.027786 | 0.124768 |
| 1 | 0.05 | 0.09 | 0.05 | 149 | 0.048073 | 0.162262 |
| 2 | 0.05 | 0.09 | 0.04 | 232 | 0.048073 | 0.162262 |
| 3 | 0.05 | 0.09 | 0.03 | 413 | 0.048073 | 0.162262 |
| 4 | 0.05 | 0.08 | 0.02 | 929 | 0.041093 | 0.149981 |

Рисунок 17 - A/A симуляция для теста Бутстрап

A/B simulation for t-test:

| | power | empirical_power | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.8 | 0.81 | 0.06 | 103 | 0.722212 | 0.874852 |
| 1 | 0.8 | 0.80 | 0.05 | 149 | 0.711171 | 0.866633 |
| 2 | 0.8 | 0.85 | 0.04 | 232 | 0.767164 | 0.906940 |
| 3 | 0.8 | 0.78 | 0.03 | 413 | 0.689296 | 0.849987 |
| 4 | 0.8 | 0.81 | 0.02 | 929 | 0.722212 | 0.874852 |

Рисунок 18 - A/B симуляция для теста Бутстрап

Последний тест - msprt. На рисунках 19 и 20 показаны результаты. В поле sample_size здесь указывается средний размер выборки из всех выборок, которые были использованы в N итерациях симуляции. Видно, что эмпирическая ошибка первого рода равна теоретической, а мощность близка к единице. Размеры выборок также оказываются приемлемыми. Однако у данного метода есть недостаток - невозможность заранее определить размер выборки, необходимый для фиксации какого-то конкретного mde.

A/A simulation for msprt test:

| | sig | empirical_sig | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|
| 0 | 0.05 | 0.046 | 1428.242 | 0.034662 | 0.060812 |

Рисунок 19 - A/A симуляция для теста msprt

A/B simulation for msprt test:

| | power | empirical_power | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.8 | 1.00 | 0.06 | 165.17 | 0.963007 | 1.000000 |
| 1 | 0.8 | 1.00 | 0.05 | 252.64 | 0.963007 | 1.000000 |
| 2 | 0.8 | 1.00 | 0.04 | 306.83 | 0.963007 | 1.000000 |
| 3 | 0.8 | 0.98 | 0.03 | 556.70 | 0.929988 | 0.994498 |
| 4 | 0.8 | 0.74 | 0.02 | 943.79 | 0.646290 | 0.815953 |

Рисунок 20 - A/B симуляция для теста msprt

По результатам проведённых симуляций было принято решение использовать критерий CUPED для A/B теста. Он показывает наилучшие результаты в A/A и A/B симуляциях, а также требует приемлемые размеры

выборок. В нашем случае минимальный обнаруживаемый эффект будем считать равным 0.03. Тест показал наличие статистически значимой разницы, p-value = 0.01 (меньше 0.05). Для этого ему потребовалось 413 элементов каждой выборки.

## 2.2 A/B тестирование конверсии из визита веб-формы в отправление заявки

Для данного теста используется датасет, который также содержит данные о том, как пользователи ведут себя в процессе заполнения веб-формы некоторого сервиса. Датасет состоит из следующих столбцов: hit-dttm – дата и время совершения действия пользователем, source – источник, из которого пришёл пользователь; ab_test_version – версия A/B теста (нас интересуют значения “2459:control” и “2459:test”); device_type – какое устройство использует пользователь при заполнении веб-формы, может быть 2 варианта – “desktop” или “mobile”; visitor_id – уникальный номер пользователя; visit_num – номер визита пользователя, в который было совершено данное действие; event_name – тип действия, совершенного пользователем (один из следующих вариантов: “pageview”, “formStep2”, “ formStep3”, “ formStep4”, “formSubmit”). Метрика, которая будет тестироваться – конверсия из визита веб-формы в отправление заявки, под этим подразумевается деление общего количества заявок на общее количество визитов. Визиту веб-формы соответствует действие “pageview”, а отправлению заявки – “formSubmit”. Важно отметить, что один пользователь может открывать веб-форму несколько раз, но в разных сессиях. Соответственно, мы имеем дело с ratio-метрикой.

На рисунке 21 показана разбивка по тесту и контролю с количеством пользователей, количеством визитов, количеством отправленных заявок, и конверсиями. На рисунках 22 и 23 дополнительно добавлены разбивки по source и device соответственно. Можно заметить зависимость конверсии от устройства.

| ab_test_version | amount | visits | submit_visits | CR |
|---|---|---|---|---|
| **2459:control** | 27682 | 29109 | 639 | 0.021952 |
| **2459:test** | 27784 | 29257 | 583 | 0.019927 |

Рисунок 21 - разбивка на тест и контроль

| source | ab_test_version | amount | visits | submit_visits | CR |
|---|---|---|---|---|---|
| **source1** | **2459:control** | 1 | 2 | 0.0 | 0.000000 |
| | **2459:test** | 1 | 1 | 0.0 | 0.000000 |
| **source2** | **2459:control** | 16 | 20 | 0.0 | 0.000000 |
| | **2459:test** | 28 | 32 | 0.0 | 0.000000 |
| **source3** | **2459:control** | 27665 | 29087 | 639.0 | 0.021969 |
| | **2459:test** | 27756 | 29224 | 583.0 | 0.019949 |

Рисунок 22 - дополнительная разбивка по источнику

| device_type | ab_test_version | amount | visits | submit_visits | CR |
|---|---|---|---|---|---|
| **Desktop** | **2459:control** | 4940 | 5208 | 100 | 0.019201 |
| | **2459:test** | 4778 | 4999 | 85 | 0.017003 |
| **Mobile** | **2459:control** | 22745 | 23901 | 539 | 0.022551 |
| | **2459:test** | 23009 | 24260 | 498 | 0.020528 |

Рисунок 23 - дополнительная разбивка по устройству

На рисунке 24 показаны конверсии в разбивке по датам. Примечательно, что в последний день тестирования поведение пользователей было отлично от поведения пользователей в предыдущие дни. Вероятно, это связано с относительно маленьким количеством пользователей: 1784 в контрольной выборке и 1842 в тестовой.

| | CR | |
|---|---|---|
| ab_test_version | 2459:control | 2459:test |
| first_date | | |
| 2021-01-11 | 0.023009 | 0.020957 |
| 2021-01-12 | 0.023711 | 0.020810 |
| 2021-01-13 | 0.019347 | 0.016054 |
| 2021-01-14 | 0.010650 | 0.016830 |

```
Text(0, 0.5, 'CR')
```

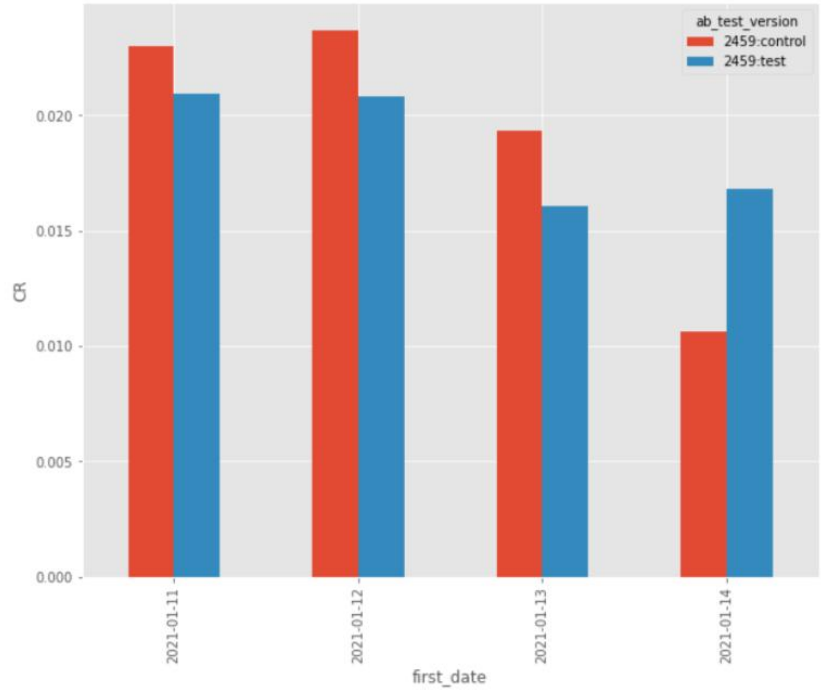


Рисунок 24 - конверсия в разбивке по датам

Симуляции A/A и A/B тестов будут происходить по аналогичному алгоритму, описанному в главе 2.1. Первым рассмотрим пропорциональный z-test, его симуляции показаны на рисунках 25 и 26 соответственно. Видно, что эмпирическая мощность в среднем меньше теоретической (которая равна 0.8). Вероятно, это объясняется тем, что мы имеем ratio-метрику. Теоретически считается некорректным использовать этот тест для такой метрики.

| | sig | empirical_sig | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.05 | 0.036667 | 0.18 | 8294 | 0.020595 | 0.064454 |
| 1 | 0.05 | 0.070000 | 0.17 | 9246 | 0.046237 | 0.104636 |
| 2 | 0.05 | 0.066667 | 0.16 | 10379 | 0.043568 | 0.100723 |
| 3 | 0.05 | 0.056667 | 0.15 | 11742 | 0.035677 | 0.088866 |
| 4 | 0.05 | 0.053333 | 0.14 | 13402 | 0.033092 | 0.084869 |

Рисунок 25 - A/A симуляция для z-test

| | power | empirical_power | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.8 | 0.643333 | 0.18 | 8294 | 0.587630 | 0.695413 |
| 1 | 0.8 | 0.606667 | 0.17 | 9246 | 0.550375 | 0.660261 |
| 2 | 0.8 | 0.660000 | 0.16 | 10379 | 0.604674 | 0.711280 |
| 3 | 0.8 | 0.723333 | 0.15 | 11742 | 0.670130 | 0.770889 |
| 4 | 0.8 | 0.683333 | 0.14 | 13402 | 0.628659 | 0.733372 |

Рисунок 26 - A/B симуляция для z-test

Следующим будет бутстрап ratio-метрики. A/A и A/B симуляции показаны на рисунках 27 и 28. Заметных улучшений по сравнению с предыдущим способом не наблюдается.

| | sig | empirical_sig | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.05 | 0.060000 | 0.18 | 8294 | 0.038286 | 0.092839 |
| 1 | 0.05 | 0.056667 | 0.17 | 9246 | 0.035677 | 0.088866 |
| 2 | 0.05 | 0.046667 | 0.16 | 10379 | 0.027999 | 0.076797 |
| 3 | 0.05 | 0.050000 | 0.15 | 11742 | 0.030532 | 0.080847 |
| 4 | 0.05 | 0.063333 | 0.14 | 13402 | 0.040917 | 0.096791 |

Рисунок 27 - A/A симуляция для теста Бутстрап

| | power | empirical_power | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.8 | 0.590000 | 0.18 | 8294 | 0.533548 | 0.644176 |
| 1 | 0.8 | 0.670000 | 0.17 | 9246 | 0.614936 | 0.720766 |
| 2 | 0.8 | 0.623333 | 0.16 | 10379 | 0.567268 | 0.676280 |
| 3 | 0.8 | 0.590000 | 0.15 | 11742 | 0.533548 | 0.644176 |
| 4 | 0.8 | 0.596667 | 0.14 | 13402 | 0.540271 | 0.650618 |

Рисунок 28 - A/B симуляция для теста Бутстрап

На рисунках 29 и 30 показаны результаты для дельта метода, который является корректным аналогом пропорционального z-теста. Однако заметных улучшений не наблюдается.

| | sig | empirical_sig | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.05 | 0.056667 | 0.18 | 8294 | 0.035677 | 0.088866 |
| 1 | 0.05 | 0.063333 | 0.17 | 9246 | 0.040917 | 0.096791 |
| 2 | 0.05 | 0.053333 | 0.16 | 10379 | 0.033092 | 0.084869 |
| 3 | 0.05 | 0.056667 | 0.15 | 11742 | 0.035677 | 0.088866 |
| 4 | 0.05 | 0.076667 | 0.14 | 13402 | 0.051628 | 0.112410 |

Рисунок 29 - A/A симуляция для дельта метода

| | power | empirical_power | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.8 | 0.613333 | 0.18 | 8294 | 0.557125 | 0.666676 |
| 1 | 0.8 | 0.640000 | 0.17 | 9246 | 0.584229 | 0.692231 |
| 2 | 0.8 | 0.623333 | 0.16 | 10379 | 0.567268 | 0.676280 |
| 3 | 0.8 | 0.683333 | 0.15 | 11742 | 0.628659 | 0.733372 |
| 4 | 0.8 | 0.640000 | 0.14 | 13402 | 0.584229 | 0.692231 |

Рисунок 30 - A/B симуляция для дельта метода

Учитывая, что мы имеем дело с ratio-метрикой, теоретически лучшим методом в данном случае будет тест, предложенный в статье Keyu Nie “Dealing With Ratio Metrics in A/B Testing at the Presence of Intra-User Correlation and Segments”. Результаты показаны на рисунках 31 и 32. Как видно, улучшения отсутствуют, есть даже небольшие ухудшения. Можно предположить, что причиной этому является то, что число пользователей, у которых больше одного визита, очень мало, и нам не удаётся корректно уловить межпользовательскую корреляцию. На рисунке 33 показана разбивка пользователей по количеству визитов.

| | sig | empirical_sig | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.05 | 0.046667 | 0.18 | 8348 | 0.027999 | 0.076797 |
| 1 | 0.05 | 0.026667 | 0.17 | 9306 | 0.013573 | 0.051729 |
| 2 | 0.05 | 0.063333 | 0.16 | 10447 | 0.040917 | 0.096791 |
| 3 | 0.05 | 0.043333 | 0.15 | 11818 | 0.025496 | 0.072717 |
| 4 | 0.05 | 0.026667 | 0.14 | 13489 | 0.013573 | 0.051729 |

Рисунок 31 - A/A симуляция для теста из статьи Keyu Nie

| | power | empirical_power | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.8 | 0.593333 | 0.18 | 8348 | 0.536908 | 0.647398 |
| 1 | 0.8 | 0.560000 | 0.17 | 9306 | 0.503422 | 0.615061 |
| 2 | 0.8 | 0.556667 | 0.16 | 10447 | 0.500087 | 0.611813 |
| 3 | 0.8 | 0.613333 | 0.15 | 11818 | 0.557125 | 0.666676 |
| 4 | 0.8 | 0.540000 | 0.14 | 13489 | 0.483452 | 0.595537 |

Рисунок 32 - A/B симуляция для теста из статьи Keyu Nie

```
17683  пользователя(ей) имеют  1 визита(ов)
588  пользователя(ей) имеют  2 визита(ов)
54  пользователя(ей) имеют  3 визита(ов)
9  пользователя(ей) имеют  4 визита(ов)
5  пользователя(ей) имеют  5 визита(ов)
2  пользователя(ей) имеют  6 визита(ов)
```

Рисунок 33 - разбивка пользователей по количеству визитов

Для проверки работоспособности данного теста предлагается сгенерировать датасет как показано на рисунке 34. Количество пользователей и конверсия имеют такие же значения, как и в нашем датасете. Отличие в том, что количество визитов у пользователей имеет более разнообразное распределение, это показано на рисунке 35. На рисунке 36 мы видим результаты симуляции A/B теста для данного датасета. Теперь эмпирическая мощность стала близка к единице, это может быть подтверждением предположения из предыдущего абзаца.

```
#data generation
mu = 1
sigma2 = 1
#генерация просмотров для каждого пользователя:
views_A = np.absolute(np.exp(sps.norm(mu, sigma2).rvs(27682)).astype(np.int64)+1)
success_rate = 0.0219
beta = 1000
alpha = success_rate * beta / (1 - success_rate)
"""генерация для каждого пользователя индивидуальной вероятности того, что
пользователь отправит заявку:"""
success_rate_A = sps.beta(alpha, beta).rvs(27682)
#генерация для каждого пользователя количества заявок:
clicks_A = sps.binom.rvs(views_A, success_rate_A)


df = pd.DataFrame({'X':clicks_A,'Y':views_A})
```

Рисунок 34 - генерация датасета

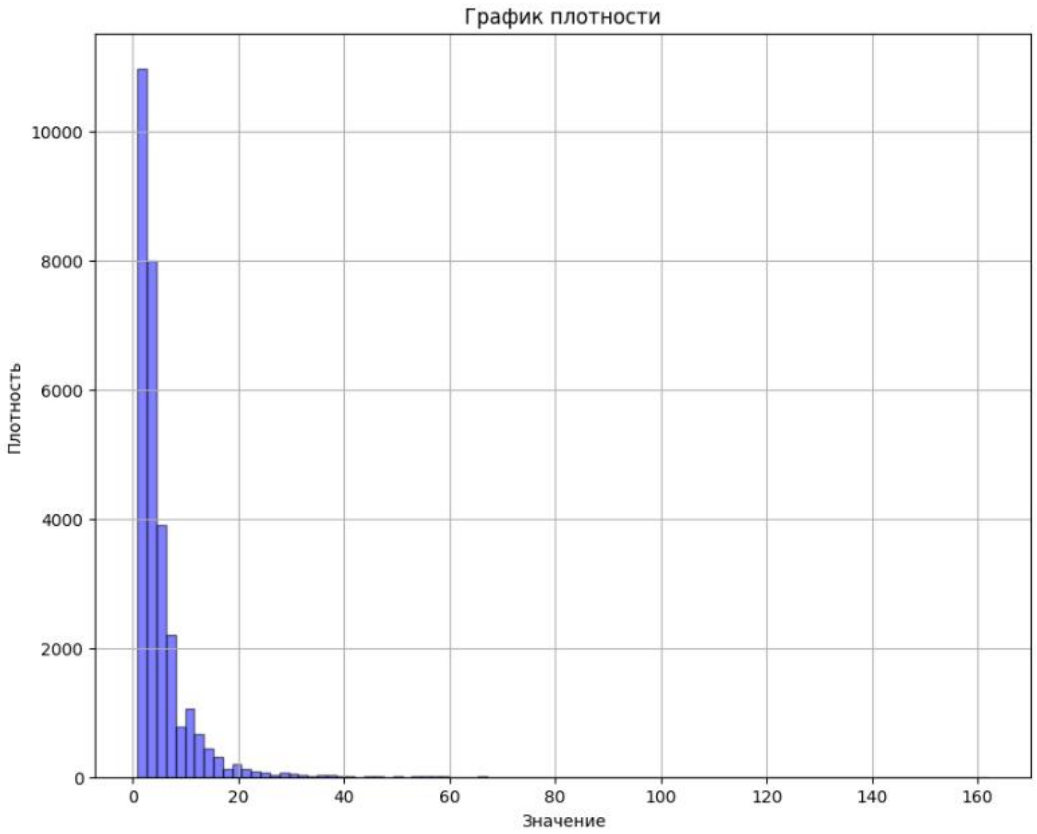


Рисунок 35 - распределение визитов в сгенерированном датасете

| | power | empirical_power | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.8 | 0.966667 | 0.18 | 9337 | 0.939738 | 0.981795 |
| 1 | 0.8 | 0.956667 | 0.17 | 10408 | 0.927283 | 0.974504 |
| 2 | 0.8 | 0.930000 | 0.16 | 11684 | 0.895364 | 0.953763 |
| 3 | 0.8 | 0.910000 | 0.15 | 13218 | 0.872223 | 0.937410 |
| 4 | 0.8 | 0.920000 | 0.14 | 15086 | 0.883727 | 0.945653 |

Рисунок 36 - A/B симуляция для сгенерированного датасета

Следующий подход, для которого были проведены симуляции на исходном датасете, - линеаризация. Результаты видны на рисунках 37 и 38, они также не имеют особых улучшений.

| | sig | empirical_sig | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.05 | 0.043333 | 0.18 | 8294 | 0.025496 | 0.072717 |
| 1 | 0.05 | 0.043333 | 0.17 | 9246 | 0.025496 | 0.072717 |
| 2 | 0.05 | 0.063333 | 0.16 | 10379 | 0.040917 | 0.096791 |
| 3 | 0.05 | 0.030000 | 0.15 | 11742 | 0.015862 | 0.056023 |
| 4 | 0.05 | 0.056667 | 0.14 | 13402 | 0.035677 | 0.088866 |

Рисунок 37 - A/A симуляция для линеаризации

| | power | empirical_power | mde | sample_size | left_real_level | right_real_level |
|---|---|---|---|---|---|---|
| 0 | 0.8 | 0.64 | 0.18 | 8294 | 0.584229 | 0.692231 |
| 1 | 0.8 | 0.65 | 0.17 | 9246 | 0.594439 | 0.701768 |
| 2 | 0.8 | 0.60 | 0.16 | 10379 | 0.543637 | 0.653835 |
| 3 | 0.8 | 0.63 | 0.15 | 11742 | 0.574045 | 0.682668 |
| 4 | 0.8 | 0.64 | 0.14 | 13402 | 0.584229 | 0.692231 |

Рисунок 38 - A/B симуляция для линеаризации

Также были проведены симуляции для метода многоруких бандитов, эмпирические мощности близки к единице для всех mde. Однако здесь мы не можем контролировать то, какой mde критерий должен считать статистически значимым. Эмпирический уровень значимости получается приблизительно равен 0.5, так как при отсутствии разницы между выборками метод выбирает победителя фактически случайно.

Ввиду всего вышеописанного для A/B теста был выбран дельта метод из-за его теоретической корректности, mde было принято приравнять 0.15. В результате тест показал, что статистически значимое увеличение конверсии отсутствует. И более того, если бы наша альтернативная гипотеза была обратной текущей, то контрольный вариант был бы статистически значимо лучше тестового, так как p-value равен 0.953. Для получения этого результата потребовалось по 11742 пользователя для каждой выборки.

# Выводы

Таким образом, по результатам проведённого исследования:

- Проведён анализ последних исследований в области A/B тестирования различных типов метрик. Были сделаны выводы о преимуществах тех или иных методов.
- Реализованы описанные методы на программном уровне.
- Изучены наиболее важные показатели, используемые в A/B тестировании веб-форм.
- Проведён анализ и предобработка используемых датасетов, выделение их особенностей. Данные особенности были применены для достижения лучших показателей в A/B тестах.
- Реализованы симуляции A/B и A/A тестирования для проверки работоспособности различных методов на программном уровне.
- Проведены полные циклы A/B тестов для двух датасетов с различными метриками.

# Заключение

A/B тестирование является надёжным инструментом для оптимизации взаимодействия бренда с потребителями в современном мире цифровых технологий. В современном бизнесе наиболее востребованным подходом является научно обоснованный, каким и является A/B тестирование. Веб-формы присутствуют на множестве сайтов, и компании заинтересованы в создании наиболее удобных для пользователей вариантов.

В результате проведения двух A/B тестов был получен один статистически значимый результат. Также были отмечены сильные и слабые стороны различных методов. В ходе анализа датасетов были выделены различные особенности, которые являются важными в контексте A/B тестирования.

# Библиографический список

1. Kohavi, R., Longobardi, R., & Sommerfield, D. (2016). Controlled experiments on the web: survey and practical guide. Data Mining and Knowledge Discovery
2. Lincoln Murphy: How to Use RICE to Prioritize Features and Product Enhancements
3. Nadim Nachar(2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution
4. Rice, John A. (2006). Mathematical Statistics and Data Analysis (3rd ed.). Duxbury Advanced.(420-443, 477-485)
5. Siroker, Koomen. A/B Testing, The most powerful way to turn clicks into customers
6. Кокрен У. Методы выборочного исследования - М., 1976.
7. Huizhi Xie, Juliette Aurisset. Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix.
8. Alex Deng, Ya Xu, Ron Kohavi. Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data
9. Simon Jackson. How Booking.com increases the power of online experiments with CUPED
10. Maryam Aziz. On Multi-Armed Bandit Designs for Dose-Finding Clinical Trials
11. Keyu Nie, Ted Tao Yuan, Yinfei Kong, Pauline Berry Burke. Dealing With Ratio Metrics in A/B Testing at the Presence of Intra-User Correlation and Segments
12. Roman Budylin, Alexey V.Drutsa , Ilya Katsev, Valeriya Tsoy.Consistent Transformation of Ratio Metrics for Efficient Online Controlled Experiments
13. Alex Deng, Ulf Knoblich.Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas

14. Ron Kohavi, Roger Longbotham, Dan Sommerfield, Randal M.Henne. Controlled experiments on the web: survey and practical guide
15. Ramesh Johari, Leonid Pekelis, Pete Koomen, David Walsh.Peeking at A/B tests
16. Peter O'Brien, Thomas R.Fleming. A Multiple Testing Procedure for Clinical Trials
17. Wenru Zhou, Miranda Kroehl, Maxene Meier, Alexander Kaizer. Approaches to analyzing binary data for large-scale A/B testing