ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 36 за 2023 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

А.Д. Минаков, В.А. Судаков

Методы ускорения контролируемых онлайн-экспериментов

Рекомендуемая форма библиографической ссылки: Минаков А.Д., Судаков В.А. Методы ускорения контролируемых онлайн-экспериментов // Препринты ИПМ им. М.В.Келдыша. 2023. № 36. 16 с. https://doi.org/10.20948/prepr-2023-36
https://library.keldysh.ru/preprint.asp?id=2023-36

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

А.Д. Минаков, В.А. Судаков

Методы ускорения контролируемых онлайн-экспериментов

Москва — 2023

Минаков А.Д., Судаков В.А.

Методы ускорения контролируемых онлайн-экспериментов

В работе рассматриваются актуальные методы проверки гипотез о действиях пользователей на сайте кредитной организации при тестировании на большом наборе данных. Разработаны и исследованы методы с применением последовательного подхода и байесовской статистики, сформированы рекомендации по их использованию. Применение предложенных методов позволило сократить время проведения контролируемых онлайн-экспериментов на 25-34%.

Ключевые слова: тестирование, проверка гипотез, последовательное тестирование, анализ данных.

Artyom Dmitrievich Minakov, Vladimir Anatolyevich Sudakov

Methods for accelerating controlled online experiments

The paper discusses current methods of hypothesis testing on a large set of data on user actions on a banking site. Methods have been developed and investigated using a consistent approach and Bayesian statistics, and a recommendation for use has been formed. The application of the proposed methods made it possible to reduce the time of conducting controlled online experiments by 25-34%.

Key words: testing, hypothesis testing, sequential testing, data analysis.

Введение

Повсеместная цифровизация и переход в онлайн-формат открывают невиданные ранее возможности для бизнеса. Одна из таких возможностей – проведение контролируемых онлайн-экспериментов. Эти эксперименты также называют A/B–тестами или A/B/N–тестами, можно встретить также обозначение «Сплит–тесты».

В исследовании используются термины «контролируемые эксперименты» и «A/B–тесты» как синонимы, независимо от количества тестовых групп. Важно это подчеркнуть, потому что буквы A и B обычно обозначают группы – тестовую и контрольную – но в экспериментах на самом деле может быть и более двух групп.

A/B–тестирование позволяет успешно применять статистический инструментарий для принятия наилучших решений. Например, оно может помочь выявить наилучший вариант текста или оформления сайта, почтовых рассылок, построения навигации внутри приложения. Внедрение или обновление алгоритмов машинного обучения, таких как модели рекомендаций и ранжирования, у крупной технологической компании не пройдёт без проведённого контролируемого эксперимента [1].

Не просто так корпорации инвестируют большие средства в собственные платформы для проведения контролируемых онлайн-экспериментов, а также спонсируют и всячески поощряют исследователей в области статистики. Большое количество передовых научных работ в области алгоритмов, статистики и машинного обучениями выпускается исследователями из Microsoft, Google, Яндекс и др.

Цель исследования – ускорить проведение контролируемых онлайн экспериментов на сайте кредитной организации. Исследование проводилось с использованием реальных исторических данных о действиях пользователей на сайте крупного российского банка.

В некоторых областях современные методы позволяют сокращать длительность проведения экспериментов в несколько раз, а значит, быстрее принимать правильное решение. Вложения корпораций в исследовательскую деятельность действительно окупаются, ведь ценность единицы времени в случае многомиллионного онлайн-сервиса колоссальна.

С учётом приведенных выше фактов вопрос об актуальности темы исследования не вызывает сомнений, ведь ускорение проведения экспериментов позволяет сэкономить огромное количество времени и ресурсов, в чем могут быть заинтересованы как коммерческие, так и некоммерческие организации. Экономия времени на один эксперимент позволяет запускать больше экспериментов в единицу времени, а значит тестировать больше гипотез и опережать конкурентов в новшествах.

В современном банке сайт является одной из основных точек входа, важным местом для оформления продуктов, поэтому сотни людей могут быть вовлечены

в постоянную работу над улучшением его производительности, дизайна и эффективности.

Подавляющее большинство тестов в неавторизованной зоне сайта связаны с изменением дизайна страниц или форм оформления. Меняться в форме может как внешний вид, так и наполнение, очередность шагов или даже их количество. Главное, чтобы в течение одного теста было только одно изменение, а не много сразу – ведь в таком случае невозможно будет измерить конкретный эффект от каждого.

Из-за такой специфики целевая метрика почти всегда одна и та же – конверсия из пользовательской сессии или просто захода на страницу в совершение некоторого действия, которое может быть разным. Это может быть, например, клики по кнопкам или переходы в определённые разделы, если как-то меняется навигация. Для тестов с формами логично смотреть на заполнение этих форм, полные или частичные прохождения.

Подготовка данных

В исследовании использовались настоящие данные о действиях пользователей на банковском сайте. Однако специфика не ограничивает применение результатов работы только банковским сектором, ведь данные подобного вида могут быть в любом приложении или сайте, где можно оформить некоторый продукт, подать заявку или заполнить какую-либо форму.

Для работы с массивами данных, построения графиков, реализации и сравнения методов в исследовании используется Python. Это высокоуровневый, объектно-ориентированный язык программирования, получивший особенную популярность и всеобщую любовь за его простоту и гибкость.

Одним из плюсов использования Python является огромное количество пользовательских библиотек, которые содержат в себе готовые модули и функции, которые полностью или хотя бы частично могут закрыть почти любую потребность.

В работе использовались такие библиотеки, как Pandas, Scipy, statsmodels, numpy, math, Seaborn, matplotlib и plotly. Они значительно упрощают анализ и обработку данных, помогают с реализацией математических операций и визуализацией результатов.

Обязательным этапом является обработка и агрегация данных. В изначальном виде метрики не рассчитаны, ведь каждая строка представляет из себя запись о некотором действии пользователя. Продемонстрируем, какая работа была проделана для получения качественного набора данных.

Таблица 1

Приблизительный формат изначальных данных

EVENT	User_id	DEVICE	Date	…	test_id	variation
Pageload	1234	PC	2023-01-01	…	1111	A
Menu_tap	4321	Mobile	2023-01-01	…	2222	B

Формат данных в таблице 1 изначально действительно не располагает к анализу и проведению тестов, требуется проделывать большое количество агрегаций и фильтраций для того, чтобы получить приемлемый вид, к которому можно будет применять статистические тесты и делать какие-либо расчёты.

Таблица 2

Формат набора данных после обработки

Test_id	Variation	Day_num	Duration	N_obs	N_successes
10	Test	1	20	1580	110
10	Control	1	20	1600	107

В финальном наборе данных в таблице 2 хранятся уже рассчитанные значения необходимых метрик, исключены лишние поля и обработаны аномалии. Ключевые поля: номер А/Б–теста (поле «Test_id»), тестовая группа (поле «Variation»), количество наблюдений (поле «N_obs») и успехов (поле «N_successes»).

Набор данных после обработки содержит более двух тысяч строк, в них собрана статистика по 48 А/Б–тестам. Средняя длительность одного теста в датасете составляет 25 дней, при этом минимальная – лишь 7, а максимальная – 43. Проводить тесты менее 7 дней считается плохой практикой, так как достаточно сильно изменяются эффекты в поведении пользователя в зависимости от дня недели.

Каждый из тестов содержит 2 группы, в которые были отфильтрованы тесты с большим количеством экспериментов, чтобы не вводить поправки на множественную проверку гипотез.

В одной группе в среднем 95702 наблюдения и 12954 успешные попытки. При этом минимально одна группа содержит 2701 наблюдение и 223 успеха, а максимально большая группа содержит 483912 наблюдения и 164791 успешное действие.

В исследуемой статистике тесты бывают как очень большие и долгие, так и проводимые на совсем небольшую выборку за короткий промежуток времени. Это обусловлено спецификой разных продуктов, располагаемым количеством трафика, а также характером самих тестируемых изменений.

Z-тест для пропорций

Z-тест для долей является очень распространённым выбором для проведения А/Б–тестов в случае бинарной метрики, например конверсии в некоторое действие. Это «базисный» критерий, с которым будут сравниваться остальные. Он используется при сравнении долей некоторого признака в двух генеральных совокупностях. Нулевая гипотеза выглядит следующим образом:

$$H_0: p_1 = p_2,$$

где p_1 и p_2 – доли признака в генеральных совокупностях.

Для применения критерия обязательно должно выполняться предположение о независимости выборок.

Статистика критерия вычисляется по формуле

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}},$$

где n_1 и n_2 – объёмы данных в выборках, m_1 и m_2 – количество успешных попыток в каждой выборке, $\hat{p}_1 = \frac{m_1}{n_1}$ и $\hat{p}_2 = \frac{m_2}{n_2}$ – доли успешных попыток в выборках, $\hat{p} = \frac{m_1+m_2}{n_1+n_2}$ – общая доля успехов в двух выборках.

Для применения критерия необходимо выполнение следующих условий на минимальный размер выборок:

$$n_1\hat{p}_1 \geq 10 \text{ и } n_1(1-\hat{p}_1) \geq 10, \text{а также}$$
$$n_2\hat{p}_2 \geq 10 \text{ и } n_2(1-\hat{p}_2) \geq 10.$$

Распределение z-статистики стремится к нормальному закону при $n \to \infty$, что позволяет использовать таблицу с критическими значениями стандартного нормального распределения.

Ответ на вопрос об отклонении нулевой гипотезы о равенстве долей признаков принимается с помощью рассчитанного значения p-value или доверительного интервала. Мы будем использовать p-value, оба эти подхода согласуются и дают одинаковый результат.

P-value сравнивается с заданным наперёд уровнем значимости, который обычно равен 0.05 или 0.01. Если p-value меньше уровня значимости, то можно заявлять о наличии статистически значимых различий и отклонять нулевую гипотезу H_0. В противном случае нет достаточных оснований отклонять нулевую гипотезу H_0 о равенстве долей в группах.

После применение Z-критерия ко всему набору данных выяснили, что статистически значимо лучше тестовая вариация в 16 экспериментах из 48.

Средняя длительность статистически значимых тестов 28 дней, а для незначимых тестов – 23 дня.

Посмотрим на то, как выглядят графики p-value для успешных (рис. 1) и незначимых тестов (рис. 2).

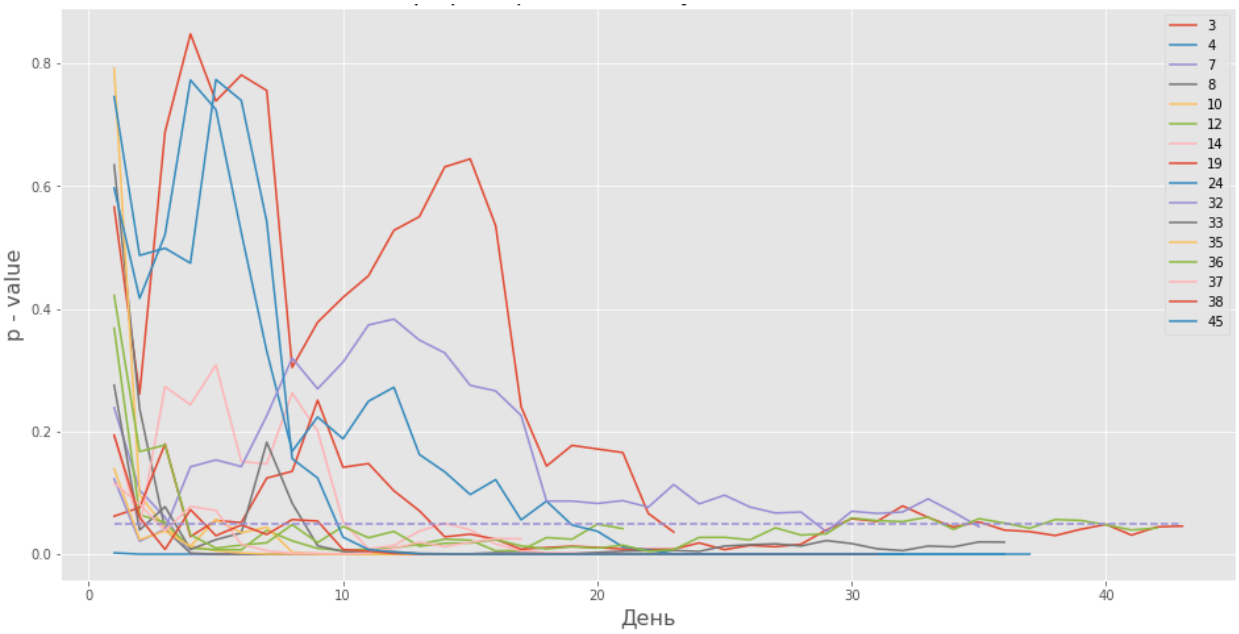

Рис. 1. Значения p-value для успешных тестов

На графике большинства успешных тестов видны «хвосты», когда график пересекает черту 0.05, а дальше уже лишь уменьшается, но экспериментаторы вынуждены всё равно ждать до конца, чтобы получить статистически достоверный вывод с заранее определённой ошибкой первого и второго рода.

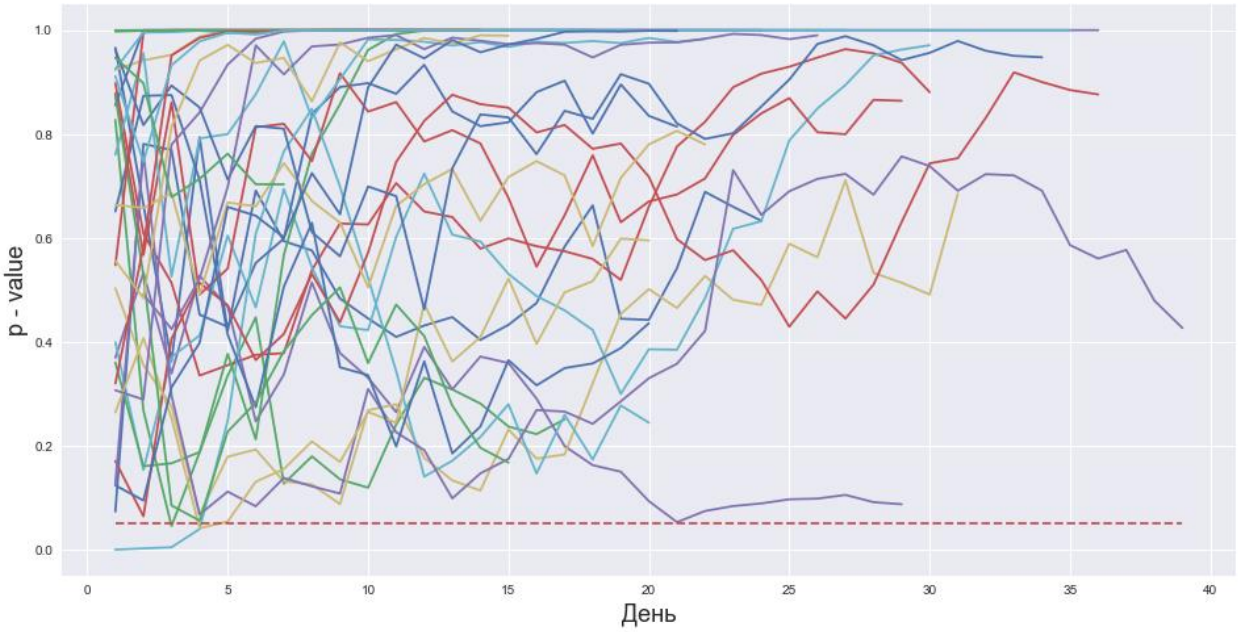

Рис. 2. Значения p-value для статистически незначимых тестов

На основе полученных графиков можно сделать вывод, что для ускорения экспериментов есть потенциал, стоит попробовать альтернативные методы, так как видны «длинные хвосты».

Байесовский подход

Применение формулы условной вероятности и методов байесовской проверки гипотез открывает много новых возможностей в А/Б–тестах. Чтобы описать отличие байесовского подхода от частотного, начнём с типов неопределенности. Существует алеаторная и эпистемическая неопределённость. Алеаторная неопределённость возникает из-за случайности. Эпистемическая неопределённость возникает из-за недостаточного знания.

Например, алеаторная неопределённость возникает в случае, когда мы знаем, сколько шаров разных цветов в мешке, но всё же не знаем, какой именно будет следующим. Примером же эпистемической неопределённости может служить ситуация, когда мы не знаем точное распределение шаров с разными цветами в мешке. При этом если мы будем вытаскивать случайный шар и класть его обратно, то со временем всё лучше будем понимать долю шаров разных цветов, а значит, эпистемическая неопределённость будет уменьшаться.

Алеаторная неопределённость всегда и везде выражается с помощью вероятностей, тут расхождений между подходами нет, а вот эпистемическая неопределённость понимается по-разному.

С точки зрения частотного подхода, вероятность – это некоторый предел частот успехов в неограниченно большой последовательности повторений эксперимента, то есть значение некоторого параметра неизвестно, но конкретно.

Байесовский подход смотрит на эпистемическую неопределённость иначе, в нём допускается её выражение в терминах вероятностных распределений. Эпистемическая неопределённость выражается с помощью функций плотности, при этом более правдоподобные значения соответствуют большим значениям плотности. С увеличением количества наблюдений выражение плотности будет меняться, учитывая особенности данных.

Частотный подход понимает вероятность как некоторую частоту, откуда и пошло название. «Байесовский», в свою очередь, понимает под вероятностью некоторую субъективную оценку неопределённости.

В частотном взгляде принято, что даже если некоторый параметр распределения неизвестен, то он всё равно константен, а «байесовский» подход допускает восприятие параметров как случайных величин.

Говоря о байесовском подходе, необходимо ввести понятия априорного и апостериорного распределений. Априорное распределение характеризует знания о неопределённости до эксперимента. В результате эксперимента мы получаем данные, которые меняют неопределённость. Распределение, которое соответствует изменённой неопределённости, называется апостериорным.

Формула Байеса для перехода от априорного к апостериорному распределению:

$$p(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)},$$

где $p(\theta|X)$ – апостериорное распределение, $P(\theta)$ – априорное распределение, $P(X|\theta)$ – правдоподобие данных, $P(X)$ – маргинальное правдоподобие.

В случае нашей метрики конверсий мы имеем бинарный результат, следовательно, это по определению испытание Бернулли. Поэтому в таком случае можно не вычислять маргинальное правдоподобие, а воспользоваться сопряжённым априорным распределением, в нашем случае таким будет являться Бета–распределение. Сопряжённое распределение – это распределение из того же параметрического семейства, что и изначальное.

На примере наших бинарных данных:

$$X \sim Bin(p, n), \text{мы наблюдаем } X = k, \text{тогда}$$
$$p \sim Beta(k - 1, n - k - 1), \text{где } k - \text{количество успехов.}$$

Таким образом, мы можем перейти к апостериорному распределению и вычислить, насколько вероятность одной вариации может быть лучше другой, с помощью полученных плотностей вероятностей конверсии в группах.

Полученное в результате значение показывает степень уверенности в результате, но неизвестно, по какому критическому значению останавливать тест.

В качестве критерия остановки был использован метод ROPE с метрикой «lift», которая является разницей средних в выборках. Якоб Кохен в своей работе [2] говорит, что исследователь определяет область практической эквивалентности (ROPE) вокруг нулевого значения, которая охватывает те значения параметра, которые считаются незначительно отличающимися от нулевого значения для практических целей. Размер ROPE будет зависеть от специфики области применения. В качестве общего примера, поскольку размер эффекта, равный 0,1, обычно считается небольшим, зависимость от размера эффекта может варьироваться от -0,1 до +0,1.

Таким образом, нет чётких рекомендаций по выбору метрики и самих границ региона, авторы статей советуют подбирать их эмпирически.

Рассмотрим результаты применения реализованного критерия с использованием «байесовского» подхода, где решение об окончании теста принимается на основании 95%-го доверительного интервала апостериорного распределения метрики разницы средних конверсий.

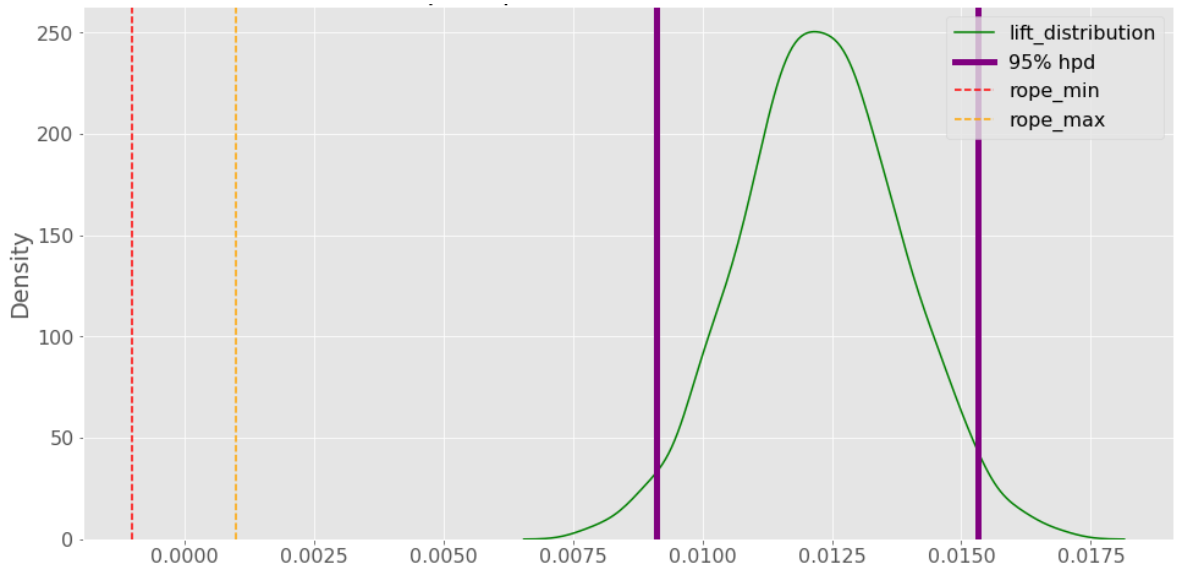

Рис. 3. Вид распределения и положение относительно интервала ROPE для успешного теста

На рисунке 3 весь 95%-й доверительный интервал находится справа от границ интервала Rope. Это значит, что мы можем останавливать тест и говорить о наличии статистически значимого улучшения в средней конверсии. Для каждого из тестов с увеличением данных выстраиваем распределение и смотрим на его положение, принимая решение последовательно.

Однако не все тесты показывают статистическую значимость. Если применение байесовского подхода не показывает значимость различий, то стоит приостанавливать тест, когда его длительность сравняется со временем проведения эксперимента с применением Z-теста.

На рис. 4 ROPE лежит внутри доверительного интервала, само распределение более плоское, обладает большей дисперсией, чем в первом примере.

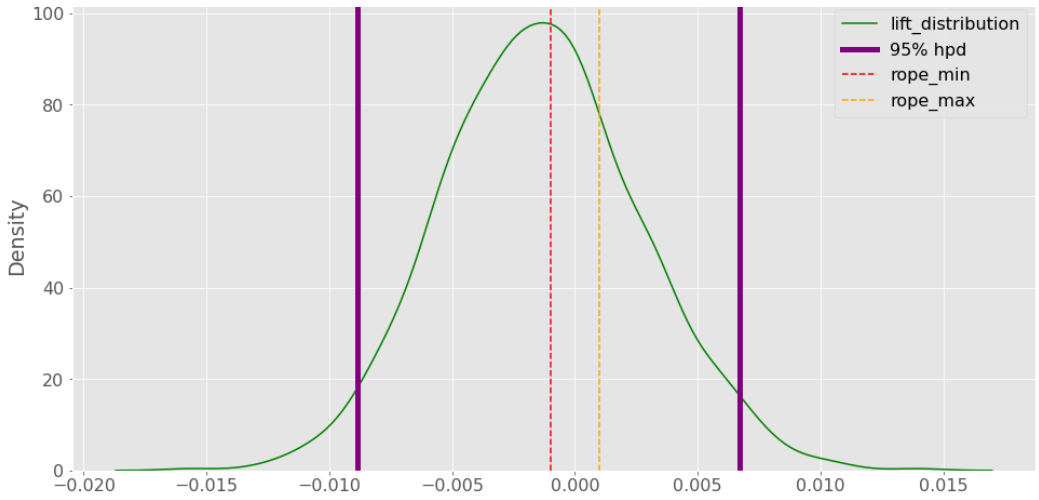

Рис. 4. Вид распределения и положение относительно интервала ROPE для теста без значимых различий

Применим реализованный метод и сравним с результатами Z-теста для пропорций с помощью матриц ошибок (см. рис. 5).

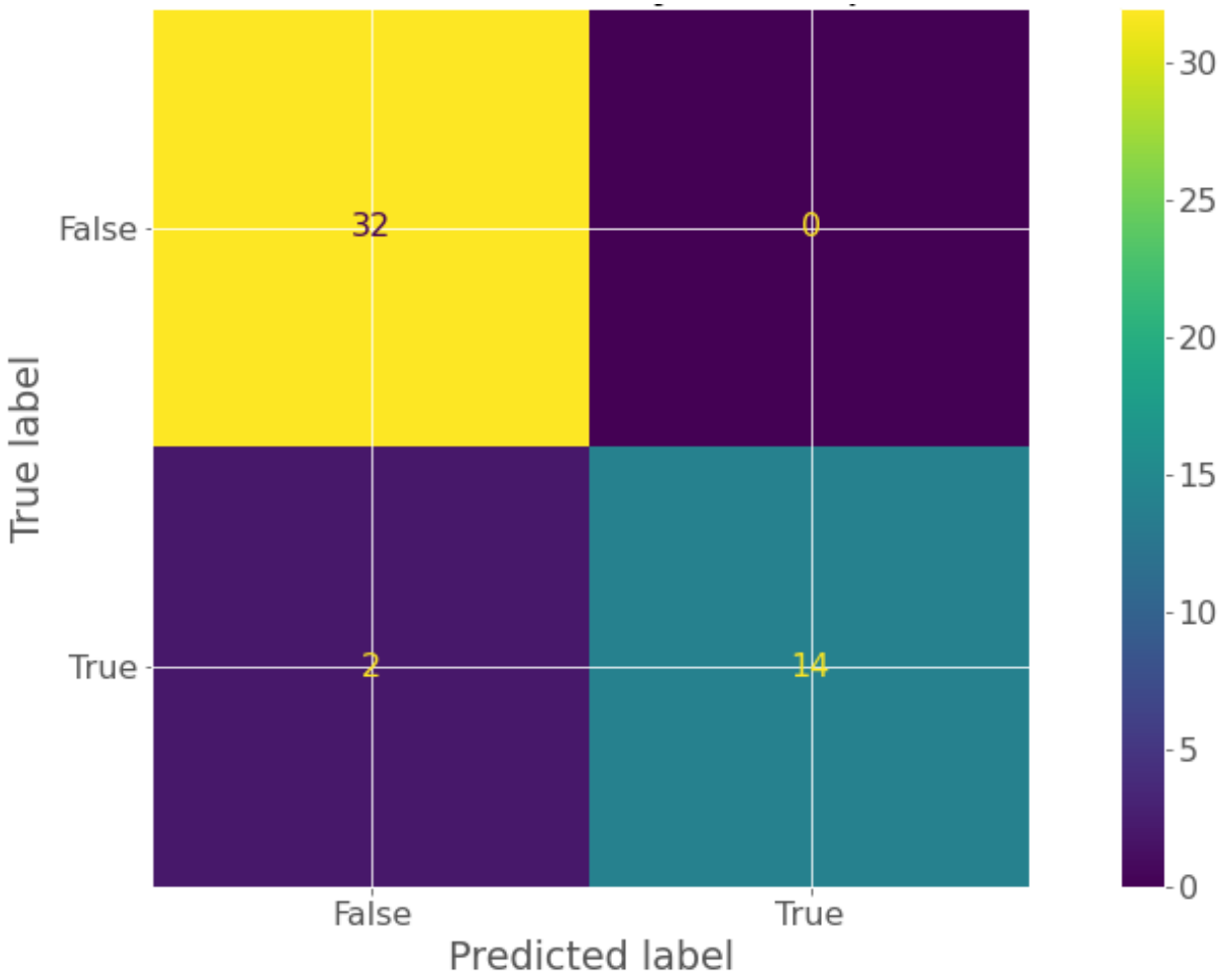

Рис. 5. Матрица ошибок для «байесовского» подхода в сравнении с Z-тестом.

Байесовский подход лишь в двух случаях разошёлся с Z-тестом. При этом он склонен реже заявлять о наличии различий, то есть является более строгим и «консервативным» методом.

Последовательный подход

Последовательный подход к проверке гипотез – это подход, при котором размер выборки заранее не определён. Данные оцениваются по мере их поступления, а решение может быть вынесено в любой момент в соответствии с заранее определённым правилом остановки. Преимущество метода заключается в том, что иногда вывод может быть сделан на гораздо более ранней стадии, чем это было бы возможно при более классической проверке гипотез или оценке, следовательно, с меньшими финансовыми или человеческими затратами. Основоположником подхода является Абрахам Вальд [3], а развивают его многие ученые современности.

В качестве предложения по борьбе с «проблемой подглядывания» исследователи из компании Optimizely предложили свой подход, который назвали «Mixture Sequential probability ratio test» или просто mSPRT [4]. Метод mSPRT основан на теории случайных процессов, позволяет останавливать тест сразу, когда «always valid p-value» будет ниже предварительно определенного порогового значения α.

При этом гарантировано, что частота ложноположительных (fdr) результатов будет не более α.

Критерий последовательного анализа для нескольких гипотез характеризуется смешанным распределением H по θ. Представим, что мы проверяем разницу в метрике θ, и, имея выборку s_n по времени n, получим смешанную функцию отношения правдоподобия

$$\Lambda_n^H(s_n) = \int_\theta \left(\frac{f_\theta(s_n)}{f_{\theta_0}(s_n)}\right)^n \mathrm{d}H(\theta),$$

где θ – разница между конверсиями групп в момент времени n.

Воспользуемся этой формулой для случайных величин, являющихся реализациями распределения Бернулли, т. к. наши конверсионные данные как раз являются реализацией этого распределения.

После приведения к Гауссу и подстановки в формулу выше получаем

$$\Lambda_n^H = \sqrt{\frac{\sigma_n^2}{\sigma_n^2+n\tau}}\exp\left\{\frac{n^2\tau^2(\theta_n)^2}{2\sigma_n^2(\sigma^2+n\tau^2)}\right\}.$$

Параметр τ^2 авторы статей советуют подбирать эмпирически, основываясь на том же порядке, что и дисперсия. В работе мы пробовали несколько значений этого параметра и также эмпирически выбрали тот, который позволял допускать меньше ошибок, но при этом давал прирост в скорости принятия решения по сравнению с Z-тестом.

Имея смешанное распределение $H(\theta)$, можно вычислить функцию правдоподобия, и если её значение выше α, то допустимый уровень FP ошибки p_n превышает α, можно останавливать тест и отклонять нулевую гипотезу.

P-value для оценки момента остановки теста в этом случае будет вычисляться следующим образом:

$$P_n = min\left(P_{n-1}; \frac{1}{\Lambda_n^H}\right), P_0 = 1.$$

Рассмотрим результаты применения «последовательного» метода mSPRT и сравним результаты.

Здесь решение применяется на основании аналога p-value «always valid p-value». Сравнивается значение с некоторым уровнем значимости, а отличие от обычного p-value в том, что при пересечении этой границы можно останавливать эксперимент.

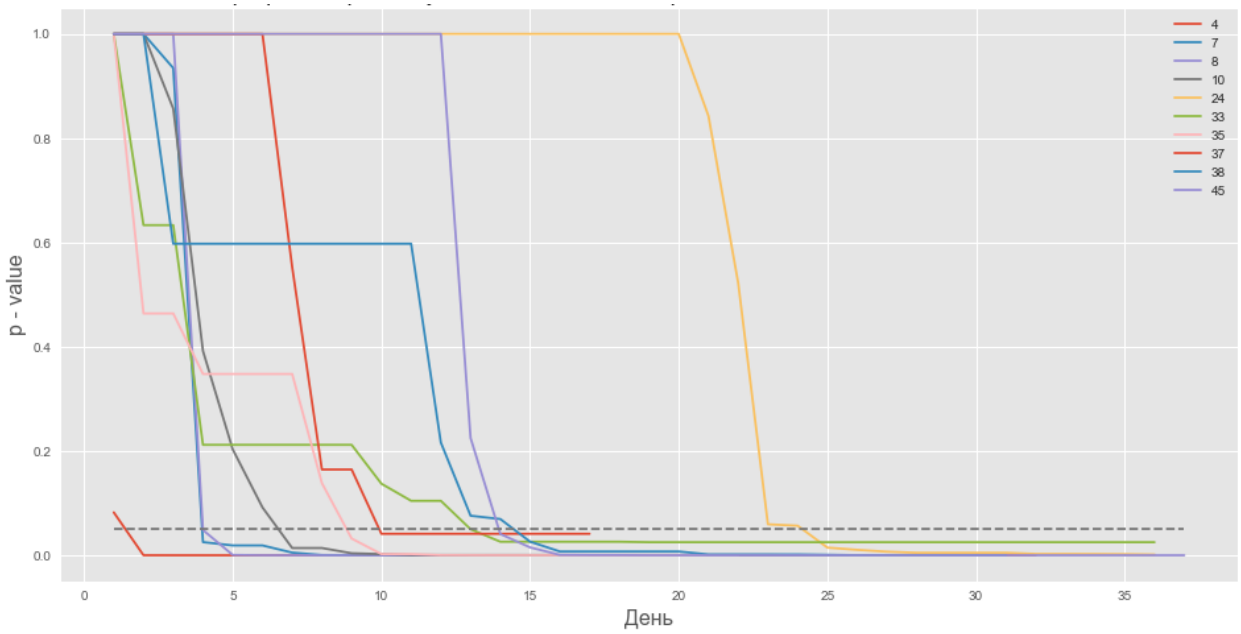

Рис. 6. Always valid p-value для успешных тестов, на которых решение метода совпало с Z-тестом для долей

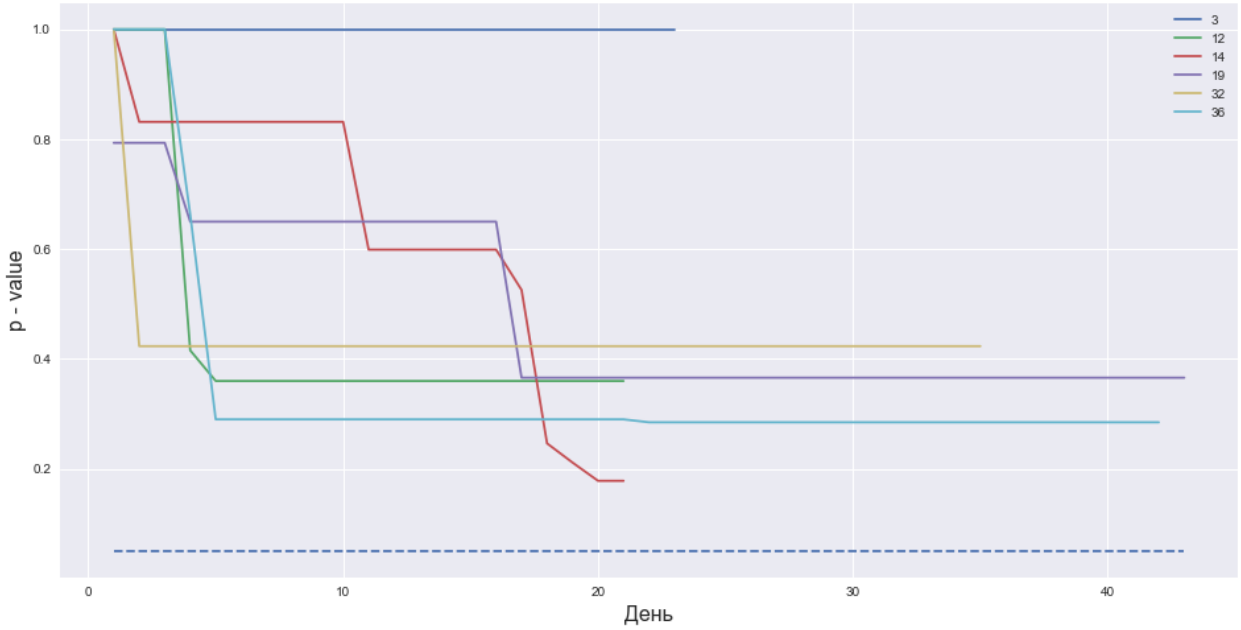

Рис. 7. Always valid p-value для неуспешных тестов, где решение метода совпало с Z-тестом для долей

Сразу бросается в глаза отличительная особенность последовательных тестов – все графики тут невозрастающие. Это обеспечивает преимущество фреймворка, дающее возможность останавливать тесты, как только линия зайдёт за черту, а также продолжать их вместо перезапуска.

Нужно отметить, что для реализации последовательного подхода необходимо осуществить подбор параметра τ^2, причём делать это действительно можно только эмпирически. Для наших данных оптимальным оказался

параметр, равный 0.001, при нём минимально количество ошибок в сравнении с Z-тестом, а также сохраняется выигрыш по времени.

Применим реализованный метод и сравним с результатами Z-теста для пропорций с помощью матриц ошибок.

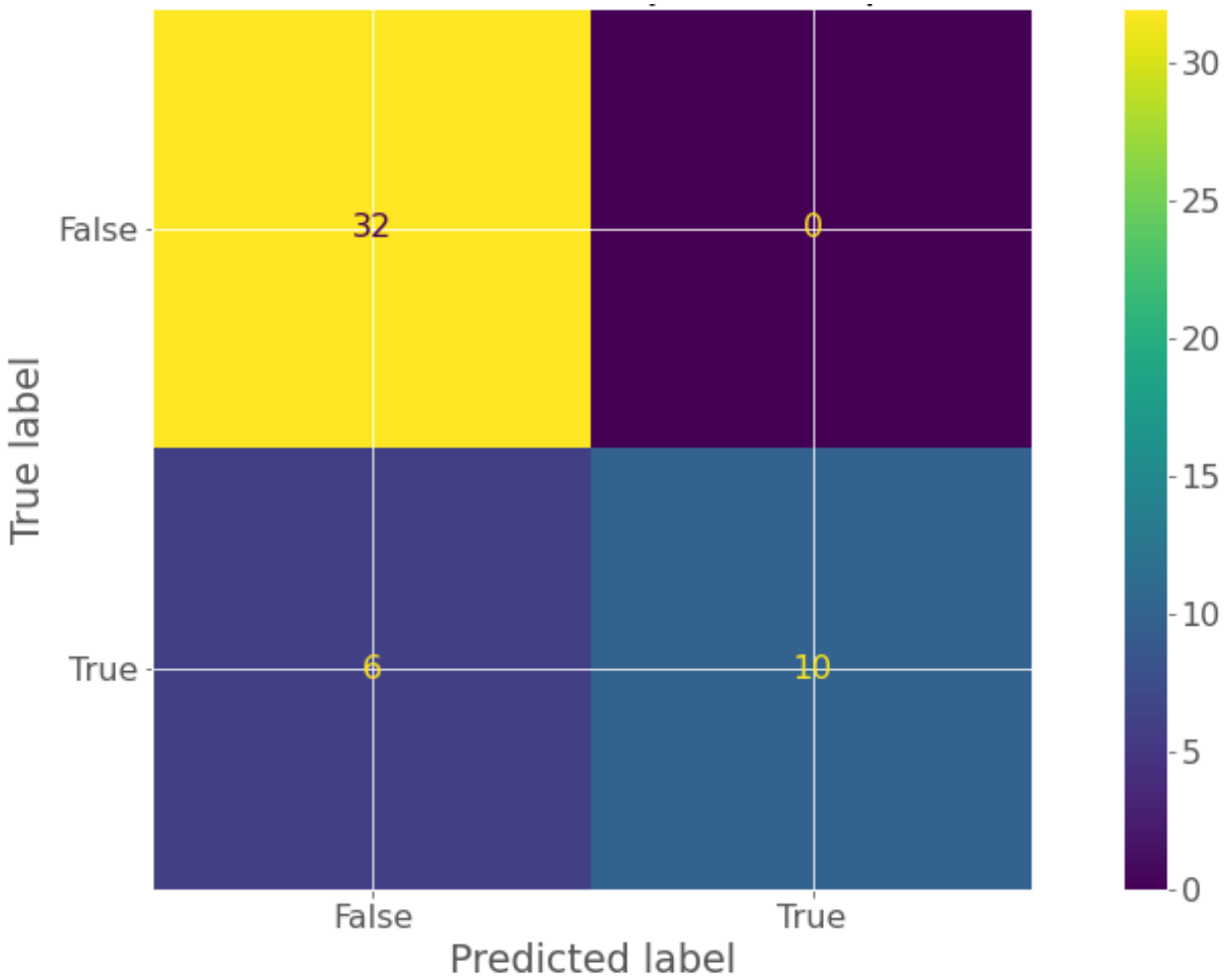

Рис. 8. Матрица ошибок для mSPRT в сравнении с Z-тестом

Последовательный подход расходится в решении в шести случаях, он значительно меньше признает значимость различий, нежели Z-тест для долей.

Интересно, но ни одной ошибки другого рода допущено не было. Исходя из этих результатов, можно сделать вывод, что оба альтернативных метода признают значимые отличия только там, где они действительно есть. Соответственно, они дают наибольший прирост по времени там, где видно сильное различие, а вот в неоднозначных случаях они менее склонны заявлять о наличии значимого различия в средних.

Посмотрим теперь, как меняются необходимая длительность теста с использованием альтернативных подходов (см. табл. 3) и ускорение в процентах (см. табл. 4) по отношению к исходной процедуре с Z-тестом.

Таблица 3

Среднее сокращение длительности экспериментов в днях

	«Байесовский» метод	«Последовательный» метод mSPRT
Для всех тестов	9	6
Для статистически значимых (по Z-тесту)	16	10
На статистически незначимых	6	5

Таблица 4

Среднее сокращение длительности экспериментов в процентах

	«Байесовский» метод	«Последовательный» метод mSPRT
Для всех тестов	34 %	25 %
Для статистически значимых (по Z-тесту)	54 %	35 %
На статистически незначимых	24 %	20 %

Байесовский подход обнаруживает изменения быстрее, а также ошибается реже. Реализация этого метода также позволяет добавить интерпретируемости результатам, поэтому рекомендуем его как наилучшую альтернативу в случае бинарной метрики.

Заключение

Целью работы являлось ускорение статистического анализа контролируемых онлайн-экспериментов. В качестве данных использовались маскированные настоящие данные о действиях пользователей на сайте кредитной организации. Решение обусловлено отсутствием необходимых данных в открытых источниках, а возможность использования таких данных подчеркивает уникальность работы.

В ходе выполнения работы был проделаны анализ и обработка исходных данных, подготовлен набор данных, состоящий из информации о 48 экспериментах. К этим данным были применены разработанные методы принятия решений. Проведено сравнение результатов Z-теста для долей с

альтернативами, рассчитаны потенциальное ускорение и вероятность расхождений в результатах.

Рекомендовано использование критерия с применением байесовской статистики, который потенциально ускоряет проведение экспериментов на сайте на 34%, что может быть актуально компаниям, которые проводят много экспериментов, потому что они должны быть готовы реализовывать и поддерживать более сложные методы ради прироста по времени.

Библиографический список

1. Kohavi R., Tang D., Xu Y. Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge: Cambridge University Press. 2020. URL: https://doi.org/10.1017/9781108653985.
2. Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd ed.). New York: Imprint Routledge. 1988. 567 p. URL: https://doi.org/10.4324/9780203771587.
3. Wald A. Sequential Tests of Statistical Hypotheses. The Annals of Mathematical Statistics. 16 (2), 1945, pp. 117–186. URL: https://www.jstor.org/stable/2235829
4. Johari R., Pekelis L., Walsh D. Always Valid Inference: Bringing Sequential Analysis to A/B Testing. 2019. URL: https://arxiv.org/pdf/1512.04922.pdf.

Оглавление