УДК 004.932

Методологические проблемы и современные решения в области контролируемых онлайн-экспериментов

Зотов Даниил Александрович zotov.da@bmstu.ru
Пивоварова Наталья Владимировна pivovarova.natasha2013@yandex.ru

МГТУ им. Н.Э. Баумана, Москва, Россия

В данной работе рассматриваются современные подходы и проблемы, возникающие при проведении контролируемых онлайн-экспериментов (также называемых A/B-тестами), являющихся ключевым инструментом для эмпирической проверки гипотез в цифровых продуктах. Несмотря на широкое распространение методологии в индустрии, ее применение сталкивается с рядом методологических и практических проблем. Особое внимание уделено анализу существующего инструментария для проектирования, выполнения и интерпретации онлайн-экспериментов, а также проблемам измерения малых эффектов, выявления гетерогенных эффектов и оценки долгосрочного влияния изменений. Работа систематизирует основные направления текущих исследований и формулирует открытые вопросы, требующие дальнейшего изучения.

Ключевые слова: *A/B-тестирование, чувствительность экспериментов, гетерогенные эффекты, долгосрочные эффекты*

Введение. В последние десятилетия методология проведения контролируемых онлайн-экспериментов, также известная как A/B-тестирование, стала одним из ключевых инструментов эмпирической проверки гипотез в цифровых и инженерных системах. Методология позволяет исследователям и инженерам количественно оценивать влияние изменений или новых функциональных компонент на реальные показатели эффективности систем.

Первоначально методология развивалась в контексте веб-экспериментов в крупных технологических компаниях, однако с развитием цифровизации принципы контролируемых экспериментов находят применение и в производственных областях. В частности, современные CAD/CAM/CAE-системы (Autodesk Fusion 360, Siemens NX, PTC Creo) все активнее интегрируют облачные модули сбора пользовательских данных и предлагают возможности адаптивных интерфейсов и интеллектуальной поддержки проектирования. В этом контексте онлайн-эксперименты могут использоваться для оценки эффективности новых конфигураций параметрического моделирования или интерфейсных решений, влияющих на производительность инженера.

Несмотря на формальную простоту методологии, корректное проведение тестов сопряжено с рядом проблем. Современные исследования указывают на необходимость переосмысления классической методологии тестирования с учетом роста масштаба данных и динамики поведения пользователей.

Цель настоящей работы — систематизировать современные теоретические подходы и методологические проблемы, связанные с проведением контролируемых онлайн-экспериментов. Особое внимание уделяется трем ключевым аспектам: проблеме малых эффектов и чувствительности тестов, учету гетерогенности воздействий и анализу долгосрочных эффектов изменений. Работа имеет обзорный характер и призвана обобщить достижения последних лет и обозначить открытые вопросы, требующие дальнейшего изучения.

Задача контролируемых онлайн-экспериментов. Контролируемые онлайн-эксперименты представляют собой форму эмпирической проверки гипотез, основанную на идее случайного распределения пользователей или объектов между двумя и более вариантами системы (группы эксперимента) — контрольным (A) и тестовым (B). В классическом варианте эксперимент проводится с целью выявления причинно-следственного эффекта некоторого изменения на измеряемую метрику результата. Рандомизация обеспечивает эквивалентность групп и, следовательно, статистическую корректность сравнения средних значений метрик между вариантами.

Схема A/B-теста восходит к принципам статистических экспериментов, которые были сформулированы Р. Фишером и Дж. Нейманом, и реализует контролируемое воздействие на пользователей при соблюдении условий независимости наблюдений и случайности распределения участников эксперимента между исследуемыми вариантами. Формально эффект теста можно представить как разность математических ожиданий:

$$\tau = E[y_1] - E[y_0], \quad (1)$$

где y_1 — значение рассматриваемого показателя при наличии воздействия (тестовая версия); y_0 — значение рассматриваемого показателя при отсутствии воздействия (контрольная версия). На практике оценка эффекта осуществляется по разности средних выборочных значений метрик, нормированных на дисперсию и скорректированных с учетом размера выборки, что позволяет использовать стандартные критерии статистической значимости (*t*-тест, *U*-тест и др.).

Ключевым понятием в контексте онлайн-экспериментов является метрика (показатель). Метрики отражают количественные изменения в поведении пользователей или характеристиках системы и служат основой для принятия решений о внедрении изменений. Метрика должна обладать направленностью, т. е. коррелировать с целями внедрения тестируемых изменений, и чувствительностью, позволяющей фиксировать даже малые, но систематические изменения [1]. В цифровых продуктах в качестве типовых метрик выступают показатели конверсии, удержания, выручки и др. В инженерных и производственных системах в качестве метрик могут использоваться время построения модели, средняя длительность интерактивных операций и проч. Результат эксперимента оценивается исходя из влияния на целевую метрику — ключевого показателя, на улучшение которого направлено тестируемое нововведение.

Проведение корректного онлайн-эксперимента требует соблюдения ряда условий. Во-первых, должна быть обеспечена независимость распределения пользователей между вариантами и исключены систематические смещения в характеристиках пользователей. Во-вторых, эксперимент должен взаимодействовать с репрезентативной аудиторией, на которую планируется распространить выводы. В-третьих, необходимо контролировать взаимное влияние участников, поскольку в цифровых средах пользователи могут косвенно влиять друг на друга, нарушая предпосылку независимости. Существенное значение имеет длительность эксперимента, так как недостаточная продолжительность приводит к высокой вариативности результатов.

Современные системы онлайн-экспериментов функционируют как высокоавтоматизированные платформы, включающие модули для распределения трафика, регистрации событий, расчета метрик и визуализации результатов. В ведущих компаниях индустрии такие платформы работают на уровне инфраструктуры и поддерживают тысячи параллельных тестов. Методология постоянно развивается, в ряде современных платформ применяются адаптивные схемы: последовательное тестирование, байесовские подходы и прочие методы [2].

Таким образом, методология контролируемых онлайн-экспериментов представляет собой универсальный инструмент эмпирической верификации гипотез, применимый как в сфере цифровых сервисов, так и в области инженерных программных комплексов и интеллектуальных производственных систем. Однако ее практическое использование сопряжено с целым рядом методологических проблем, которые будут рассмотрены в последующих разделах.

Методологические проблемы современных решений. Несмотря на зрелость методологии контролируемых онлайн-экспериментов, ее практическое применение остается сопряженным с рядом проблем, затрагивающих корректность статистических выводов, воспроизводимость результатов и интерпретацию эффектов.

Одной из центральных проблем является чувствительность тестов к малым эффектам. В цифровых продуктах даже незначительные с точки зрения относительной разницы эффекты могут оказывать заметное влияние на общие показатели продукта. Традиционные методы, основанные на стандартных статистических критериях, часто не способны отличить слабые, но устойчивые эффекты от шумов. При этом увеличение размера выборки не всегда решает задачу: оно может повышать вероятность обнаружения статистически значимых, но практически несущественных эффектов [3]. Подобная ситуация особенно характерна для систем, в которых данные поступают непрерывным потоком, а пользователи участвуют в нескольких параллельных экспериментах, что увеличивает риск ложных открытий в длительных экспериментах.

Важной методологической проблемой остается определение корректных метрик эксперимента, способных надежно отражать влияние изменения на поведение пользователей. Многие A/B-тесты используют метрики, не обладающие причинной согласованностью, когда изменение промежуточных

показателей не гарантирует улучшения целевой метрики [4]. Неправильный выбор метрики приводит к искаженным выводам о направлении эффекта и росту ложных положительных результатов.

В современной литературе большое внимание уделяется проблеме гетерогенности эффектов, то есть различий в воздействии изменений на разные подгруппы пользователей. Классические подходы предполагают оценку усредненного эффекта, тогда как на практике воздействие часто оказывается неравномерным. Игнорирование этой неоднородности может привести к ошибочным решениям — например, новая функциональность улучшает метрику в среднем, но ухудшает ее для определенных категорий пользователей. Для решения проблемы были предложены алгоритмы с применением причинного анализа и машинного обучения, которые строят индивидуализированные оценки эффектов на основе деревьев решений и ансамблевых методов [5, 6]. Однако эти подходы требуют строгого контроля переобучения и корректной валидации статистических выводов, что до сих пор остается активной областью исследований.

Другой проблемой является прогнозирование результата во времени на основании краткосрочных результатов, полученных во время эксперимента. Классический эксперимент исходит из предположения стационарности среды, однако в реальных системах поведение пользователей, сезонные воздействия и прочие внешние факторы могут изменяться. В долгосрочных тестах часто наблюдается «дрейф» эффектов — изменение знака или величины воздействия по мере адаптации пользователей [7]. Эта проблема приводит к необходимости учета временных факторов при анализе данных, применения моделей временных рядов и методов коррекции новизны и обучаемости [8].

В последнее время особое внимание получает проблема взаимного влияния участников различных экспериментальных вариантов. В контексте онлайн-платформ пользователи могут неявно взаимодействовать друг с другом, что нарушает предпосылку независимости наблюдений. Подобные эффекты трудно устранить даже при тщательной рандомизации, особенно в многопользовательских или сетевых приложениях.

Современная практика онлайн-экспериментов сталкивается с целым спектром методологических проблем. Эти проблемы требуют комплексного подхода, включающего развитие, как статистических методов, так и инструментальных решений на уровне платформ. В последующих разделах рассматриваются наиболее значимые проблемы и направления их решения — измерение малых эффектов, анализ гетерогенных воздействий и оценка долгосрочных эффектов изменений, которые определяют дальнейшее развитие методологии.

Малые эффекты и чувствительность экспериментов. Одной из центральных проблем в настоящее время является измерение и корректная интерпретация малых эффектов. В условиях зрелых цифровых продуктов, где большинство очевидных улучшений уже реализованы, ожидаемые изменения в целевых метриках оказываются незначительными — на уровне долей про-

цента. Такие эффекты могут иметь большое практическое значение, но их статистическое выявление и воспроизводимая оценка представляют серьезную трудность. В индустриальной практике более 80 процентов A/B-тестов не дают статистически значимых различий, однако именно среди малых эффектов часто скрываются инновации с долгосрочным стратегическим потенциалом [5].

Задача чувствительности заключается в определении минимального эффекта, который статистический тест способен достоверно обнаружить при заданных уровне значимости и размере выборки. Классическая постановка задачи чувствительности базируется на определении мощности теста, т. е. на вероятности обнаружить эффект при его наличии. Для двухвыборочного *t*-теста мощность определена как

$$1 - \beta = \Phi\left(\frac{|\delta|\sqrt{n}}{\sigma} - z_{1-\frac{\alpha}{2}}\right),$$

где Φ — функция распределения стандартного нормального закона; β — вероятность совершения ошибки второго рода; δ — величина эффекта; σ^2 — дисперсия метрики; n — размер выборки; α — уровень значимости; z_η — η-й квантиль стандартного нормального распределения. В реальных экспериментах при малых значениях δ мощность теста резко снижается, что делает результаты статистически неустойчивыми. Самый простой способ решения методологической проблемы — увеличение размера выборки n, которое позволяет частично компенсировать эффект, но сопровождается ростом вычислительных и временных затрат, что является неприемлемым вариантом для многих современных компаний.

Для повышения чувствительности разработан ряд специализированных подходов, направленных на снижение дисперсии метрик. Одним из наиболее эффективных методов является ковариационная корректировка CUPED (*Controlled Pre-Experiment Data*) [9]. Его идея состоит в том, чтобы использовать поведение пользователей до эксперимента как коварианту, уменьшающую внутригрупповую вариативность. Оценка эффекта выражается как

$$y_i \ = y_i - \theta(x_i - \overline{x}),$$

где y_i — наблюдаемое значение метрики; x_i– значение метрики до начала эксперимента; $\overline{x}$ — среднее значение метрики до начала эксперимента; θ — коэффициент для минимизации дисперсии, $\theta = \frac{cov(Y_i, X_i)}{var(X_i)}$. Данный подход позволяет уменьшить дисперсию и увеличить чувствительность теста при неизменной длительности эксперимента, снижая дисперсию оценки эффекта до 50 %. CUPED и его вариации стали индустриальным стандартом и регулярно применяются в крупных цифровых компаниях.

Другой класс решений связан с последовательными и байесовскими экспериментами, которые обновляют статистику в процессе теста без увеличения

доли ошибок. В ряде работ показано, что адаптивные процедуры обеспечивают ту же достоверность, что и классические тесты, но при этом позволяют раньше выявлять малые эффекты [10]. Подобные методы особенно актуальны для потоковых систем, где данные поступают непрерывно. Дополнительно, в последние годы обсуждаются байесовские последовательные схемы принятия решений. В работе [11] представлена схема, объединяющая байесовскую оценку эффекта с механизмами обучения с подкреплением. Байесовская часть непрерывно обновляет распределение эффекта по мере поступления данных, формируя вероятностную оценку значимости малых изменений. Алгоритм обучения с подкреплением, в свою очередь, выбирает момент остановки эксперимента и объем дополнительной выборки так, чтобы минимизировать ожидаемое время теста**,** сохраняя заданный уровень ошибки и повышая вероятность обнаружения слабых эффектов.

Несмотря на значительный прогресс, существующие методы повышения чувствительности онлайн-экспериментов остаются ограниченными в применении. Индустрия сосредоточена на достижении статистической, но часто игнорирует вопрос практической значимости, что приводит к накоплению формально достоверных, но не интерпретируемых результатов [12].

Эффективность CUPED во многом зависит от стабильности ковариант и отсутствия системных сдвигов: при нестационарных данных и взаимодействии экспериментов их эффективность резко падает [9]. Авторы подхода последовательного тестирования отмечают, что даже корректно откалиброванные тесты сталкиваются с трудностью выбора оптимального момента остановки, особенно в условиях множественных параллельных проверок, где контроль уровня ошибок становится нетривиальной задачей. Кроме того, подчеркивается, что объединение байесовского вывода, последовательного анализа и методов обучения с подкреплением способно улучшить обнаружение слабых эффектов, но при этом остаются открытыми вопросы математической корректности и устойчивости этих моделей при потоковом обновлении данных [11]. Эти ограничения определяют тематику дальнейших исследований, направленных на разработку устойчивых адаптивных тестов и критериев практической значимости для динамических экспериментальных систем.

Гетерогенные эффекты. В традиционном анализе A/B-тестов предполагается, что изменение оказывает одинаковое влияние на всех пользователей, то есть эффект является гомогенным. Однако в реальных цифровых системах воздействие часто существенно варьируется между сегментами аудитории: в зависимости от типа операционной системы, географии пользователя и прочих признаков. Это явление называют гетерогенностью эффектов, и оно является одной из центральных проблем интерпретации и принятия решений на основе онлайн-экспериментов [13].

Формально индивидуальный эффект для наблюдаемого объекта i определяется по формуле (1). Тогда условный средний эффект может быть определен как

$$\tau = E[y_1 - y_0 | W = w,$$

где W — случайный вектор характеристик единицы наблюдения. На рис. 1 представлена зависимость эффекта воздействия τ от времени использования некоторой платформы пользователем: если изначально эксперимент предполагает, что эффект распределен равномерно, то на практике эффект различается в зависимости от индивидуальных особенностей пользователя.

Для оценки гетерогенности применяются два основных подхода. Первый подход является параметрическим, эффект моделируется линейной регрессией

$$Y_i = \alpha + \tau D_i + \beta^T W_i + \gamma^T (D_i \times W_i) + \varepsilon_i,$$

где D_i, $D_i \in \{0,1\}$ — индикатор воздействия; β — вектор коэффициентов при ковариатах W_i, который описывает их базовое влияние на метрику независимо от эффекта воздействия; γ — вектор коэффициентов при наличии эффекта воздействия; ε_i — случайная ошибка (предполагается, что $E[\varepsilon_i] = 0$). Параметры τ, α, β, γ оцениваются с помощью стандартной регрессии, а затем анализируется значение γ и строится прогноз индивидуальных эффектов для пользователей с определенным вектором характеристик W_i:

$$\tau_i = \tau + \gamma^T W_i.$$

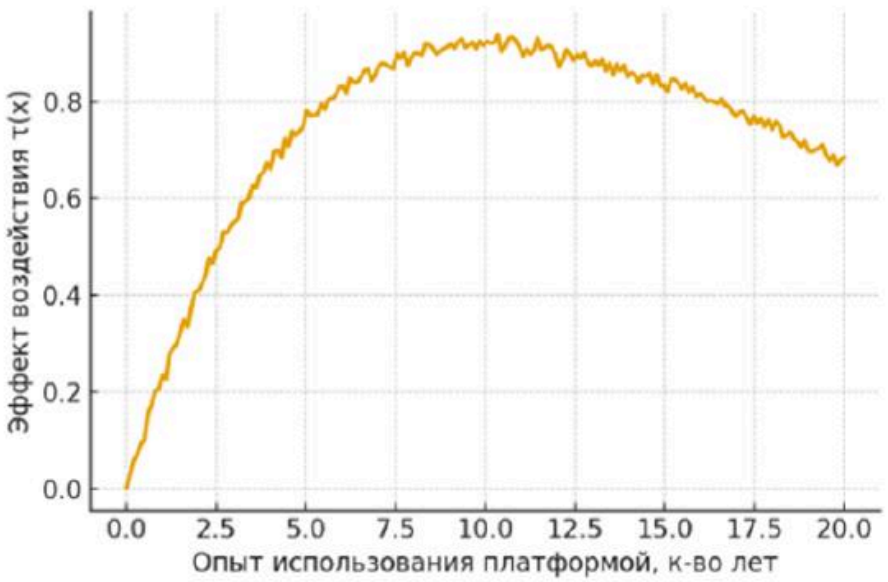

Рис. 1. Зависимость эффекта от возраста пользователя на некоторой платформе

Ограничение метода заключается в его линейности — при росте количества признаков его точность резко падает. Поэтому второй подход к оценке таких эффектов является непараметрическим и основан на применении методов машинного обучения. Наиболее известным методом стали каузальные деревья [5], идея которых заключается в разделении всей совокупности наблюдений на подгруппы (листья дерева), внутри которых эффект примерно однороден, и вычислении оценки локального эффекта внутри каждого из ли-

стьев. Для повышения устойчивости оценок подход был обобщен в каузальный лес [14], в котором используется ансамблирование — объединение множества деревьев. Оценка эффекта для наблюдения с характеристиками x вычисляется как среднее по ансамблю:

$$\hat{\tau}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{\tau}^{(b)}(x),$$

где B — число деревьев; $\hat{\tau}^{(b)}(x)$ — оценка эффекта для определенного дерева.

Несмотря на эффективность методов, их применение сталкивается с рядом ограничений: высокая вычислительная сложность, необходимость большого количества данных в выборках и непроработанная методология построения достоверных доверительных интервалов для эффектов.

Интерпретация таких моделей часто носит визуально-аналитический характер: исследователь может изучать распределение $\hat{\tau}(x)$, выделять на его основании кластеры пользователей с наиболее выраженным эффектом и формулировать гипотезы для последующих экспериментов. На рис. 2 представлено распределение $\hat{\tau}(x)$, на основании которого исследователь может заметить, что у большинства пользователей наблюдается умеренное положительное воздействие, а «хвосты» распределения указывают на подгруппы с более сильной или слабой реакцией на изменение.

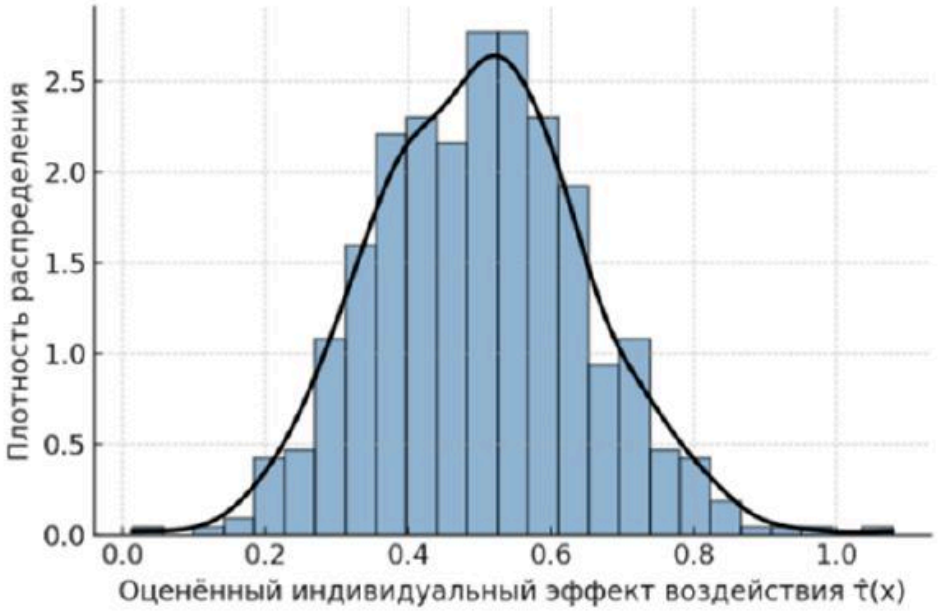

Рис. 2. Распределение индивидуальных эффектов воздействия $\hat{\tau}(x)$

На данный момент остается открытым ряд вопросов, посвященных проблеме гетерогенных эффектов. Во-первых, отсутствуют надежные методы статистической верификации моделей гетерогенности [8]. Во-вторых, остается нерешенной проблема интерпретируемости результатов: сложные модели, как, например, каузальный лес, не предоставляют понятную оценку индиви-

дуальных эффектов, что, как следствие, ограничивает применение методов в прикладных системах [5]. Подчеркивается, что рассмотрение взаимозависимых эффектов между экспериментами увеличивает риск получения ложных результатов, который необходимо контролировать [7].

Долгосрочные эффекты. Оценка долгосрочных эффектов является одной из наиболее сложных и активно обсуждаемых проблем современной экспериментальной методологии. В большинстве практических сценариев A/B-тесты проводятся в течение ограниченного времени, что позволяет достичь статистической мощности, но не отражает возможных изменений в поведении пользователей после завершения эксперимента. В ряде современных исследований на практических примерах показано, что эффект нередко изменяет знак во времени из-за адаптации пользователей, сезонных колебаний и взаимодействий между экспериментами [3, 7, 8].

Временную динамику эффекта можно описать зависимостью

$$\tau(t) = E[y_{1t} - y_{0t}],$$

где $\tau(t)$ — средний эффект во времени. При этом $\tau(t)$ может демонстрировать нестационарность: на практике наблюдается три типичных интервала динамики эффекта во времени — усиление, адаптация и стабилизация [4], представленных на рис. 3.

Одним из распространенных инструментов измерения устойчивого воздействия является модель разностей-в-разностях (*Difference-in-Differences*), позволяющая скорректировать эффект с учетом фоновых временных трендов [15]. Оценка эффекта вычисляется как

$$\tau = \left(E\left[y_{1,t=1}\right] - E\left[y_{1,t=0}\right]\right) - \left(E\left[y_{0,t=1}\right] - E\left[y_{0,t=0}\right]\right),$$

где $t, t \in \{0,1\}$ — индикаторная переменная; $t = 0$ — период до начала эксперимента, $t = 1$ — период после начала эксперимента. Метод компенсирует систематические смещения метрик между группами, которые обсуловлены сезонностью и гетерогенностью и позволяет оценить «чистый» дополнительный эффект, обусловленный исключительно тестируемым изменением.

Несмотря на простоту, метод требует соблюдения предположения, согласно которому до вмешательства тестовая и контрольная группы изменяются одинаково, различие в группах обосновано только вмешательством и добавлением новой функциональности. Нарушение этого предположения приводит к смещению оценок. Поэтому, для повышения устойчивости результатов часто применяют метод синтетического контроля, в котором контрольная группа искусственно формируется на основе комбинации «похожих» наблюдений.

В работе [8] выделяется влияние факторов «новизны» и «привыкания» к изменению, которые по-разному искажают эффект спустя время. Использование авторегрессионных и байесовских моделей позволяет описывать эти зависимости, но требует длительного горизонта наблюдения и усложняет оценку доверительных интервалов эффекта.

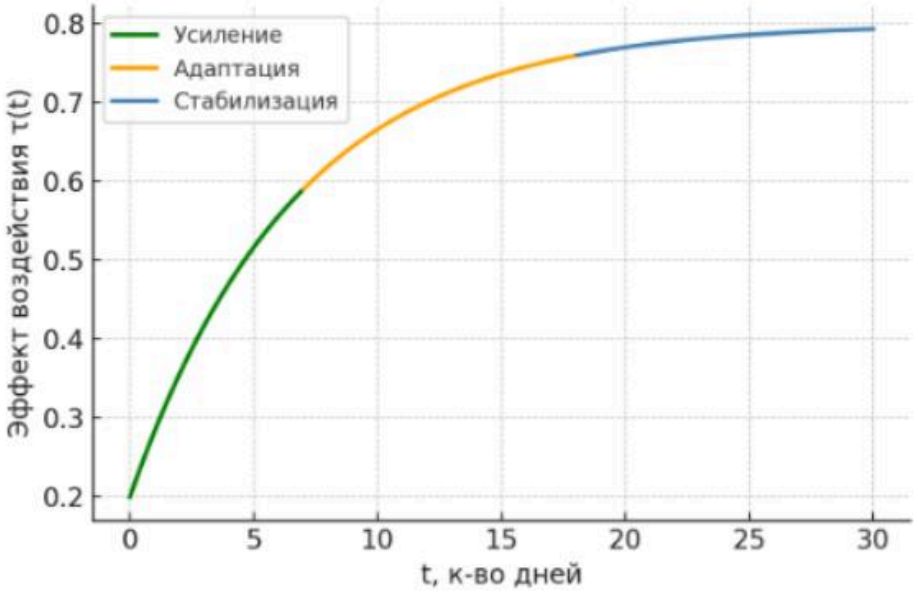

Рис. 3. Типичная динамика эффекта $\tau(t)$ во времени

Несмотря на развитие методологии, надежное измерение долгосрочных эффектов остается открытой задачей. Ряд авторов отмечает, что отсутствуют общепринятые технические средства длительных экспериментов [7]. Кроме того, подчеркивается необходимость формализации критериев перехода между фазами долгосрочного эффекта и определение «точки стабилизации» [10]. Обращается внимание на отсутствие инструментов для отделения влияния долгосрочных изменений от влияния внешних и сезонных колебаний [8]. Наконец, измерение долгосрочного воздействия ограничено со стороны инфраструктуры, так как большинство платформ ориентировано на проведение экспериментов с длительностью до нескольких недель, не предусматривающих непрерывный сбор данных после их завершения [1]. Решение проблем в этой сфере требует систематизации и переосмысления актуальной литературы для связи классических краткосрочных методов тестирования с прогнозированием долгосрочных эффектов.

Заключение. В работе представлен обзор современных методологических проблем контролируемых онлайн-экспериментов. Показано, что ключевые проблемы связаны с измерением малых эффектов, выявлением гетерогенности и учетом долгосрочных изменений. Современные методы, такие как CUPED и последовательные тесты, повышают чувствительность экспериментов, но не устраняют проблему интерпретации результатов. Подходы к анализу гетерогенных эффектов показывают ограниченность существующих методов в условиях большого количества данных. Краткосрочные результаты часто не отражают реальную динамику, что требует учета временных зависимостей и, несмотря на значительный прогресс, отсутствует единая методологическая основа, объединяющая различные направления анализа. Дальнейшие исследования предполагает интеграцию статистических, причинных и байесовских подходов для повышения надежности и воспроизводимости экспериментальных выводов.

Литература

[1] Kohavi R., Vermeer L. Top challenges from the First Practical Online Controlled Experiments Summit. *ACM SIGKDD Explorations Newsletter*, 2019, pp. 20–23. https://doi.org/10.1145/3331651.3331655

[2] Deng A., Shi X. Data-driven Metric Development for Online Controlled Experiments: Seven Lessons Learned. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16 Workshop on Large-Scale Experimentation)*, 2016, pp. 77–86.

[3] Bojinov I., Gupta S. Online Experimentation: Benefits, Operational and Methodological Challenges, and Scaling Guide. *Harvard Business School Working Paper 22-002*, 2022. https://doi.org/10.1162/99608f92.a579756e

[4] Kohavi R., Tang D., Xu. Y. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge, Cambridge University Press, 2020, 302 p.

[5] Wager S., Athey S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 2018, pp. 1228–1242.

[6] Chernozhukov V., Chetverikov D., Demirer M., Duflo E., Hansen C., Newey W. Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review: Papers and Proceedings*, 2017, pp. 261–265. https://doi.org/10.48550/arXiv.1701.08687

[7] Dmitriev P., Frasca B., Gupta S., Kohavi R., Vaz G. Pitfalls of Long-Term Online Controlled Experiments. *IEEE International Conference on Big Data*, 2016, pp. 1367–1376. https://doi.org/10.1109/BigData.2016.7840744

[8] Sadeghi S., Gupta S., Gramatovici S., Lu J., Ai H., Zhang R. Novelty and Primacy: A Long-Term Estimator for Online Experiments. *Technometrics*, 2022, № 64, pp. 524–534.

[9] Kohavi R., Deng A., Walker T. Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data. *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM '13)*, New York, ACM Press, 2013, pp. 123–132.

[10] Johari R., Pekelis L., Walsh D.J. Always Valid Inference: Bringing Sequential Analysis to A/B Testing. *Operations Research*, 2022, pp. 1590–1606. https://doi.org/10.48550/arXiv.1512.04922

[11] Wan R., Liu Y., McQueen J., Hains D., Song R. Experimentation Platforms Meet Reinforcement Learning: Bayesian Sequential Decision-Making for Continuous Monitoring. *Proceedings of the 29th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2023)*, New York, ACM Press, 2023, pp. 4120–4130.

[12] Kohavi R., Deng A., Longbotham R. Seven Rules of Thumb for Web Site Experiments.. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*, New York, ACM Press, 2014, pp. 1857–1866.

[13] Larsen N., Sengupta S., Stallrich J., Kohavi R. Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology. *The American Statistician*, 2023, № 78, pp. 1–15.

[14] Athey S., Imbens G.W. Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences (PNAS)*, 2016, pp. 7353–7360. https://doi.org/10.1073/pnas.1510489113

[15] Baker A., Callaway B., Cunningham S., Goodman-Bacon A., Sant'Anna P. H. C. Difference-in-Differences Designs: A Practitioner's Guide. *Arxiv preprint arXiv:2503.13323*, 2025. https://doi.org/10.48550/arXiv.2503.13323

Methodological challenges and modern solutions in online controlled experiments

Zotov Daniil Alexandrovich zotov.da@bmstu.ru
Pivovarova Natalya Vladimirovna pivovarova.natasha2013@yandex.ru

BMSTU, Moscow, Russia

This paper reviews contemporary approaches and challenges in conducting controlled online experiments, also known as A/B tests, which serve as a key instrument for empirical hypothesis validation in digital products. Despite the widespread adoption of this methodology in industry, its application faces a number of methodological and practical difficulties. Particular attention is given to the analysis of existing tools for the design, execution, and interpretation of online experiments, as well as to the problems of measuring small effects, identifying heterogeneous treatment effects, and assessing long-term impacts of changes. The study systematizes the main directions of current research and outlines open methodological questions that require further investigation.

Keywords: *A/B testing, experiment sensitivity, heterogeneous effects, long-term effects*