Оценка эффективности метода CUPED при анализе продуктовых метрик цифровых сервисов

Минаков Артём Дмитриевич
аспирант, Российский экономический университет имени Г. В. Плеханова, amnkvwork@yandex.ru

Ограниченный экспериментальный трафик снижает чувствительность A/B-тестов и затрудняет обнаружение небольших продуктовых эффектов. Цель исследования состоит в оценке эффективности классического CUPED, использующего значение целевой метрики в предпериоде для снижения дисперсии оценки эффекта. Статистическое моделирование выполнено для конверсии, количества действий и выручки, а аналитические расчёты проверены методом Монте-Карло. Эффективность CUPED преимущественно определяется корреляцией между значениями метрики в предпериоде и экспериментальном периоде. При корреляции 0,4 дисперсия снижается примерно на 16%, а MDE на 8,3%; при корреляции 0,8 снижение составляет 64% и 40%. Признаков систематического смещения не выявлено. Для разреженных метрик и тяжёлых правых хвостов аналитические расчёты рекомендуется проверять на исторических данных.

Ключевые слова: CUPED, A/B-тест, снижение дисперсии, мощность статистического критерия, минимально обнаружимый эффект, продуктовые метрики, метод Монте-Карло.

Введение

A/B-тестирование применяется для оценки продуктовых изменений в цифровых сервисах. Случайное распределение пользователей между группами позволяет отделить эффект изменения от различий в составе аудитории. Однако небольшой эффект может оставаться неразличимым на фоне естественной вариативности. Его обнаружение требует увеличения выборки или продолжительности эксперимента.

В предыдущей работе нехватка трафика была выделена как одно из ограничений A/B-теста [1]. Она особенно заметна для небольших аудиторий, редких событий и множества параллельных экспериментов. Один из способов снижения потребности в новых наблюдениях состоит в использовании значений целевой метрики до начала воздействия. Такой подход расширяет информационную основу продуктовых решений [2].

Метод CUPED был предложен для повышения чувствительности контролируемых экспериментов за счёт данных предпериода [3]. Из значения метрики в экспериментальном периоде вычитается линейная составляющая, предсказываемая по поведению того же пользователя до эксперимента. Чем сильнее связь между периодами, тем больше снижение дисперсии. CUPED можно рассматривать как форму регрессионной корректировки [4, 5].

Продуктовые метрики различаются по форме распределения. Конверсия бинарна, число действий содержит много нулей, а выручка имеет тяжёлый правый хвост. Поэтому теоретическую связь между корреляцией и снижением дисперсии необходимо проверить в конечных выборках.

Цель исследования состоит в оценке эффективности классического CUPED для продуктовых метрик с различными распределениями и в формировании практических ориентиров его применения в A/B-тестах цифровых сервисов. Объектом исследования выступает статистическая оценка эффекта продуктового изменения. Предметом является изменение точности этой оценки после корректировки на значение метрики в предпериоде.

Для достижения цели фиксируется связь между корреляцией, дисперсией, MDE и требуемым трафиком. Эта связь проверяется для девяти вариантов метрик. Дополнительно оцениваются мощность, ошибка первого рода, покрытие и смещение, после чего формируются практические ориентиры.

Основная гипотеза состоит в том, что относительный выигрыш CUPED определяется прежде всего корреляцией между периодами, тогда как тип метрики влияет на надёжность аналитического приближения. Работа не предлагает нового статистического метода и имеет прикладной характер.

Теоретические основы метода CUPED

Пусть Y обозначает значение целевой метрики пользователя в экспериментальном периоде, а X её значение в предпериоде. Ковариата X измеряется до запуска изменения. Поэтому она не может быть его следствием. В классической постановке CUPED коэффициент корректировки оценивается как отношение ковариации двух величин к дисперсии ковариаты [3].

$\theta = \mathrm{Cov}(Y, X) / \mathrm{Var}(X)$

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

$\mathrm{Var}(Y_{CUPED}) / \mathrm{Var}(Y) = 1 - \rho^2$

снижение MDE $= 1 - \sqrt{(1 - \rho^2)}$

множитель эквивалентного размера выборки $= 1 / (1 - \rho^2)$

Центрирование X относительно общего среднего $\bar{X}$ сохраняет средний уровень скорректированной метрики. При случайном распределении пользователей ожидаемые значения ковариаты в контрольной и экспериментальной группах совпадают. При фиксированном θ вычитание одинаково построенной предсказуемой части не меняет ожидаемую разность средних. В конечной выборке θ оценивается по наблюдаемым данным. Поэтому отсутствие смещения не принимается как автоматическое следствие формулы и дополнительно проверяется методом Монте-Карло.

Коэффициент θ оценивается по объединённой выборке обеих групп. Для одной ковариаты и линейной связи остаточная дисперсия равна исходной, умноженной на $1 - \rho^2$. Поэтому корреляции 0,4 и 0,8 дают снижение дисперсии около 16% и 64%.

При фиксированных уровне значимости, целевой мощности и процедуре проверки MDE пропорционален стандартной ошибке [6]. Поэтому он сокращается медленнее дисперсии. При корреляции 0,4 снижение MDE составляет 8,3%.

Сокращение требуемого трафика не равно множителю эквивалентной выборки. Если дисперсия снижается на 16%, тот же уровень точности достигается на 84% исходной выборки, а фактическая выборка эквивалентна увеличению в $1 / 0{,}84 = 1{,}19$ раза.

Корректность CUPED требует, чтобы ковариата была измерена до воздействия, определялась одинаково в обеих группах и не зависела от назначения в экспериментальную группу. В исследовании используется одна полностью наблюдаемая ковариата, индивидуальная рандомизация и отсутствие взаимного влияния пользователей. Данные двух периодов связываются на уровне пользователя, что требует согласованного электронного обмена данными [7].

Выражение $1 - \rho^2$ описывает линейную проекцию при известных параметрах генеральной совокупности. В конечной выборке θ и корреляция оцениваются с ошибкой, а проверка гипотез опирается на нормальное приближение. Поэтому его точность может различаться для бинарных, разреженных и тяжелохвостных метрик.

Снижение дисперсии и надёжность статистических выводов следует рассматривать раздельно. CUPED может уменьшать дисперсию, но аналитическая мощность остается неточной, если распределение выборочного среднего ещё недостаточно близко к нормальному.

Методика статистического моделирования

Статистическое моделирование организовано по схеме, в которой фиксируются цель, способ формирования данных, эффект, методы и показатели качества [8]. Общие принципы статистических испытаний описаны также в русскоязычной литературе [9]. Все параметры заданы заранее, а начальное значение генератора равно 20260829.

Исследование включает девять вариантов продуктовых метрик. Они объединены в три семейства. Для каждого семейства задан основной вариант и две проверки устойчивости. Нормальное распределение не включено в перечень продуктовых метрик. Оно используется только для технической проверки настройки корреляции.

Таблица 1

Параметры сценариев статистического моделирования

Метрика	Вариант	Распределение	Параметры	Назначение
Конверсия	Редкое событие	Бернулли	p = 1%	Проверка
Конверсия	Основной	Бернулли	p = 10%	Основной расчёт
Конверсия	Частое событие	Бернулли	p = 30%	Проверка
Действия	Высокий разброс	Отрицательное биномиальное с избытком нулей	среднее 3; нули 30%; k = 0,5	Проверка
Действия	Основной	Отрицательное биномиальное с избытком нулей	среднее 3; нули 30%; k = 1,5	Основной расчёт
Действия	Низкий разброс	Отрицательное биномиальное с избытком нулей	среднее 3; нули 30%; k = 5	Проверка
Выручка	Основной	Нули и логнормальное	плательщики 10%; σ = 1,2; среднее 1	Основной расчёт
Выручка	Редкие плательщики	Нули и логнормальное	плательщики 1%; σ = 1,2; среднее 1	Проверка
Выручка	Тяжёлый правый хвост	Нули и логнормальное	плательщики 10%; σ = 1,8; среднее 1	Проверка

Значение конверсии формируется как случайная величина с распределением Бернулли. Основная вероятность события равна 10%. Проверки устойчивости соответствуют редкому событию с вероятностью 1% и более частому событию с вероятностью 30%. Такое построение позволяет проверить, меняется ли качество приближения при разной частоте события.

Количество действий описывается отрицательным биномиальным распределением с избыточным числом нулевых значений. В 30% случаев число действий принудительно задаётся равным нулю. Общее среднее сохраняется на уровне трёх действий на пользователя. Параметр неоднородности активных пользователей принимает значения 0,5, 1,5 и 5. Меньшее значение соответствует более выраженному разбросу между пользователями.

Выручка равна нулю для неплательщиков. Положительная часть имеет логнормальное распределение. В основном варианте доля плательщиков составляет 10%, а параметр формы σ равен 1,2. Средняя выручка по всем пользователям нормирована до одной условной единицы. В первом дополнительном варианте доля плательщиков снижена до 1% при прежнем значении σ. Во втором варианте σ увеличен до 1,8 при доле плательщиков 10%, что формирует тяжёлый правый хвост.

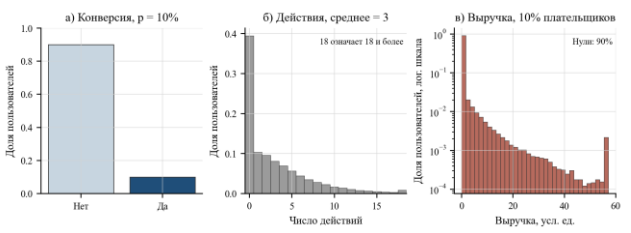


Рисунок 1. Распределения трёх базовых продуктовых метрик

Связанные значения в предпериоде и экспериментальном периоде формируются преобразованием пары нормальных случайных величин к целевым распределениям [10]. Параметр исходной зависимости подбирается численно, чтобы корреляция Пирсона между итоговыми метриками отличалась от заданной не более чем на 0,01.

Размер каждой группы равен 10 тысячам, 100 тысячам или 1 миллиону пользователей. Корреляция изменяется от 0 до 0,8 с шагом 0,1. Эффект принимает три значения. Это отсутствие эффекта, половина MDE и один MDE, рассчитанный для анализа без корректировки. Уровень значимости равен 5%, целевая мощность равна 80%, а трафик распределяется между группами поровну.

Эффект задаётся как изменение среднего значения метрики в экспериментальной группе. Для конверсии изменяется вероятность события. Для количества действий и выручки изменяется среднее при сохранении остальных параметров распределения. При разных уровнях эффекта используются одни и те же исходные псевдослучайные числа. Это уменьшает дополнительный разброс при сравнении сценариев. Сценарий без эффекта нужен для оценки ошибки первого рода и покрытия доверительного интервала.

Оценки, полученные без корректировки и после применения CUPED, сравниваются одной двусторонней процедурой с асимптотическим нормальным приближением. В обоих случаях оценка эффекта равна разности средних между экспериментальной и контрольной группами. Для CUPED сначала по объединённой выборке контрольной и экспериментальной групп оценивается θ, затем формируется скорректированная метрика. Различия в результатах поэтому связаны с корректировкой, а не с выбором разных статистических критериев.

Аналитическая сетка содержит 261 комбинацию: 243 для трёх базовых метрик и 18 для проверок устойчивости. Метод Монте-Карло применяется выборочно. Для группы из 10 тысяч пользователей выполняется 2000 повторений при ρ = 0, 0,4 и 0,8, для 100 тысяч пользователей проводится 300 повторений при ρ = 0,4. Для проверок устойчивости используется 500 повторений при ρ = 0,5, а сценарии с миллионом пользователей оцениваются аналитически.

Оцениваются дисперсия, стандартная ошибка, MDE, мощность, ошибка первого рода, покрытие 95-процентного доверительного интервала и смещение. Также рассчитываются сокращение требуемого трафика и множитель эквивалентной выборки. Смещение сопоставляется с 99-процентным интервалом Монте-Карло.

Результаты исследования

Численный подбор параметра корреляции обеспечил соблюдение заданного критерия во всех сценариях. Максимальное абсолютное отклонение фактической корреляции от целевой составило 0,0011. В технической проверке с нормальным распределением максимальное отклонение квадрата фактической корреляции от квадрата заданной корреляции составило 0,00031. Эти результаты подтверждают точность настройки зависимости между величинами.

В аналитическом приближении ожидаемый относительный выигрыш не зависит от масштаба метрики и размера выборки, если корреляция остаётся неизменной. Абсолютный MDE при увеличении каждой группы с 10 тысяч до 100 тысяч пользователей уменьшился в √10 раза. Относительное преимущество CUPED при этом сохранилось. Основные ориентиры представлены в таблице 2.

Таблица 2

Теоретические ориентиры эффективности метода CUPED

ρ	Снижение дисперсии, %	Снижение MDE, %	Сокращение выборки, %	Множитель выборки	Мощность CUPED, %
0,0	0,0	0,0	0,0	1,00	80,0
0,2	4,0	2,0	4,0	1,04	81,6
0,3	9,0	4,6	9,0	1,10	83,6
0,4	16,0	8,3	16,0	1,19	86,4
0,5	25,0	13,4	25,0	1,33	89,9
0,8	64,0	40,0	64,0	2,78	99,7

Примечание. Мощность рассчитана для эффекта, равного одному MDE анализа без корректировки. Для него исходная аналитическая мощность составляет 80%.

При корреляции 0,3 дисперсия уменьшается на 9%, а MDE на 4,6%. Такой выигрыш может быть полезен при простой реализации корректировки. Его практическая значимость зависит от стоимости внедрения и доступного трафика. При корреляции 0,4 снижение достигает 16% для дисперсии и 8,3% для MDE. Фактическая выборка становится эквивалентной выборке без корректировки, увеличенной в 1,19 раза.

При корреляции 0,5 дисперсия уменьшается на 25%, MDE на 13,4%, а множитель эквивалентного размера выборки равен 1,33. При корреляции 0,8 снижение достигает 64% для дисперсии и 40% для MDE. Тот же уровень точности без корректировки потребовал бы примерно в 2,78 раза больше пользователей.

Метод Монте-Карло в основных сценариях в целом воспроизвел эту закономерность. При корреляции 0,4 эмпирическое снижение дисперсии составило 16,52% для конверсии, 13,74% для количества действий и 16,90% для выручки. При корреляции 0,8 соответствующие значения составили 62,59%, 63,49% и 63,94%. Величина расхождений не противоречит точности метода Монте-Карло при данном числе повторений. Точечные различия между типами метрик поэтому не интерпретируются как устойчивые.

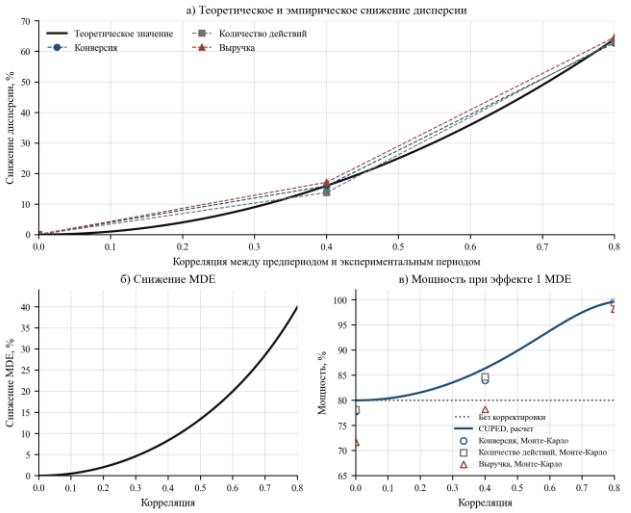


Рисунок 2. Зависимость показателей эффективности CUPED от корреляции

Рост мощности также соответствует снижению стандартной ошибки. При корреляции 0,4 и истинном эффекте величиной в один MDE анализа без корректировки мощность критерия для конверсии увеличилась с 76,8% до 84,0%. Для метрики количества действий рост составил с 79,4% до 84,7%. Для выручки он составил с 71,0% до 78,2%. При корреляции 0,8 мощность критерия после применения CUPED достигла 99,4%, 99,5% и 98,2% соответственно.

Полученная мощность критерия для выручки оказалась ниже аналитического ориентира даже без корректировки. Причина связана с формой распределения, а не с изменением целевого эффекта. Большая доля нулевых значений и редкие крупные значения замедляют приближение разности средних к нормальному закону. CUPED уменьшает дисперсию и повышает мощность по сравнению с процедурой без корректировки, но не устраняет это расхождение полностью.

В сценариях без эффекта частота ошибки первого рода составила 3,95–5,5% без корректировки и 3,95–6,2% после CUPED. Покрытие доверительного интервала равно 94,5–96,05% и 93,8–96,05% соответственно. Все значения находятся в заранее установленных диагностических пределах. При номинальной частоте 5% ошибка её оценки методом Монте-Карло составляет около 0,5 процентного пункта для 2000 повторений и 1,3 пункта для 300 повторений.

Ни для оценки без корректировки, ни для CUPED отклонение от истинного эффекта не вышло за границы 99-процентного интервала Монте-Карло. При нулевой корреляции эмпирическое изменение дисперсии было близко к нулю.

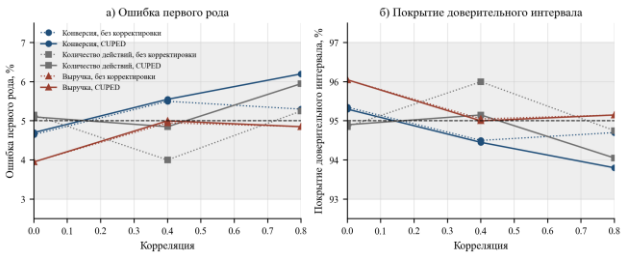


Рисунок 3. Частота ошибки первого рода и покрытие доверительного интервала по результатам метода Монте-Карло

Таблица 3
Результаты метода Монте-Карло и проверок устойчивости

Метрика и вариант	ρ	Снижение, %	Без CUPED, %	CUPED, %	Ошибка I рода, %	ДИ, %
Конверсия, 10%	0,4	16,52	76,8	84,0	5,5	94,5
Конверсия, 10%	0,8	62,59	78,9	99,4	6,2	93,8
Действия, k = 1,5	0,4	13,74	79,4	84,7	4,9	95,2
Действия, k = 1,5	0,8	63,49	77,9	99,5	5,9	94,0
Выручка, 10%; σ = 1,2	0,4	16,90	71,0	78,2	5,0	95,0
Выручка, 10%; σ = 1,2	0,8	63,94	71,8	98,2	4,9	95,2
Конверсия, 1%	0,5	23,70	72,8	81,6	6,0	94,0
Конверсия, 30%	0,5	27,61	79,0	89,0	4,6	95,4
Действия, k = 0,5	0,5	26,90	74,8	86,6	4,0	96,0
Действия, k = 5	0,5	22,66	78,0	88,4	4,8	95,2
Выручка, 1%; σ = 1,2	0,5	25,48	54,6	66,2	5,6	94,4
Выручка, 10%; σ = 1,8	0,5	33,34	77,2	88,0	4,0	96,0

Примечание. Снижение дисперсии и мощность приведены для эффекта 1 MDE. Ошибка первого рода и покрытие приведены для нулевого эффекта. Размер каждой группы равен 10 тысяч пользователей.

Проверки устойчивости подтвердили общий характер связи между корреляцией и относительным снижением дисперсии. Для конверсии с вероятностью 1% эмпирическое снижение составило 23,7% при истинном эффекте в один MDE. Для конверсии 30% оно составило 27,61%. Для вариантов количества действий значения равны 26,90% при k = 0,5 и 22,66% при k = 5. Теоретический ориентир при корреляции 0,5 равен 25%.

Наиболее заметное отклонение возникло для метрики выручки с большой долей нулевых значений. При доле плательщиков 1% эмпирическое снижение дисперсии составило 25,48%, что близко к теории, однако мощность увеличилась только с 54,6% до 66,2%. Покрытие доверительного интервала CUPED в сценарии с эффектом составило 92,8%. Это значение немного ниже принятой границы 93%. Для сценария без эффекта покрытие равнялось 94,0%.

При тяжёлом правом хвосте с σ = 1,8 эмпирическое снижение дисперсии достигло 33,34%, а мощность выросла с 77,2% до 88,0%. Средняя выборочная корреляция равнялась 0,545 при заданном значении 0,5 для генеральной совокупности. Редкие крупные наблюдения заметно изменяли выборочную оценку связи. Поэтому для подобных метрик одной точечной оценки корреляции по историческим данным недостаточно.

Во всех семействах метрик снижение дисперсии остаётся близким к квадрату корреляции. Точность аналитической мощности сильнее зависит от формы распределения, особенно для разреженной и тяжелохвостной выручки.

Обсуждение результатов и практические рекомендации

Полученные результаты позволяют использовать корреляцию как основной предварительный критерий целесообразности классического CUPED. При ρ ниже 0,3 снижение дисперсии меньше 9%, а сокращение MDE меньше 4,6%. Такой выигрыш предлагается считать слабым. При $0,3 \le \rho < 0,5$ выигрыш предлагается считать умеренным. При ρ не ниже 0,5 он предлагается считать существенным.

Эти границы не являются новым статистическим критерием. Они переводят непрерывную теоретическую зависимость в форму, удобную для планирования эксперимента. Решение остаётся экономическим и организационным. Даже небольшое сокращение MDE может быть полезно, если ковариата уже доступна и корректировка встроена в аналитическую систему. Если сбор данных предпериода требует отдельного дорогого процесса, слабый выигрыш может не окупить усложнение. При внедрении CUPED расчёт может стать частью более широкой многоуровневой информационной системы [11].

Практическая процедура начинается с проверки, что ковариата измерена до воздействия и определяется одинаково для всех пользователей. Затем на исторических данных оцениваются корреляция, снижение дисперсии, MDE и требуемый трафик. До просмотра результатов фиксируются окно предпериода, правило включения пользователей и способ оценки θ.

Для метрик без большой доли нулевых значений и тяжёлого правого хвоста при достаточном размере выборки аналитический расчёт можно использовать как первое приближение. Для редких событий, метрик выручки с большой долей нулевых значений и тяжёлым правым хвостом его нужно дополнить A/A-проверкой или моделированием на исторических данных. Такая проверка должна оценивать не только снижение дисперсии, но и ошибку первого рода, покрытие доверительного интервала и смещение.

Тип метрики не используется как самостоятельное основание для отказа от CUPED. Бинарная конверсия и метрики количества действий в проведенном исследовании дали результаты, близкие к теории. Выручка также показала снижение дисперсии, но потребовала более осторожной оценки мощности. Основанием для решения служит сочетание наблюдаемой корреляции и результатов проверки конкретной статистической процедуры.

Исследование ограничено одной полностью наблюдаемой ковариатой, измеренной в предпериоде, и индивидуальной рандомизацией без взаимного влияния пользователей. Синтетические параметры не описывают конкретный сервис. Не рассматриваются пропуски, изменение состава аудитории, неоднородность эффекта, нестабильность измерения и сценарии, одновременно меняющие долю активных пользователей и распределение их значений.

Отдельного исследования требуют метрики в виде отношений. Их оценка зависит от совместного поведения числителя и знаменателя. Для построения стандартной ошибки могут применяться линеаризация и дельта-метод [12, 13]. Выводы, полученные для средних значений на пользователя, нельзя без дополнительного обоснования переносить на такие метрики.

Дальнейшее развитие связано с несколькими ковариатами, нелинейными корректировками и признаками, построенными методами машинного обучения. Подобные подходы способны извлекать больше предсказательной информации,

чем одна метрика предпериода [14, 15, 16]. При этом усложняется защита от переобучения и возникает необходимость раздельного обучения и оценки. В данной работе эти методы не рассматриваются совместно с классическим CUPED.

Заключение

Статистическое моделирование подтвердило, что эффективность классического CUPED в первую очередь определяется корреляцией между значениями метрики в предпериоде и экспериментальном периоде. Теоретические соотношения для дисперсии и MDE подтвердились для конверсии, количества действий и базовой выручки. Систематического смещения не выявлено, а ошибка первого рода и покрытие в основных сценариях сохранились в заранее заданных диагностических пределах.

Корреляция ниже 0,3 соответствует слабому практическому выигрышу, $0,3 \le \rho < 0,5$ умеренному, а $\rho \ge 0,5$ существенному. Для метрик с большой долей нулей или тяжёлым правым хвостом теоретический расчёт следует проверять на данных, близких к реальному эксперименту. Такой подход позволяет использовать CUPED как воспроизводимый элемент системы выбора методов оценки продуктовых изменений.

Литература

1. Минаков А. Д., Вилкул А. А. О методах оценки эффективности продуктовых изменений в цифровых сервисах // Инновации и инвестиции. 2026. № 2. С. 698.
2. Уринцов А. И., Дик В. В. Система формирования и принятия решений в условиях информатизации общества: монография. М.: Евразийский открытый институт, 2008. 223 с. ISBN 978-5-374-00187-7.
3. Deng A., Xu Y., Kohavi R., Walker T. Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data // Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. 2013. P. 123-132. DOI: 10.1145/2433396.2433413.
4. Kohavi R., Tang D., Xu Y. Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge: Cambridge University Press, 2020.
5. Айвазян С. А., Мхитарян В. С. Прикладная статистика. Основы эконометрики. В 2 т. Т. 1. Теория вероятностей и прикладная статистика. 2-е изд., испр. М.: ЮНИТИ, 2001. 656 с.
6. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale: Lawrence Erlbaum Associates, 1988.
7. Уринцов А. И. Электронный обмен данными: учебно-практическое пособие. М.: Евразийский открытый институт, 2011. 181 с. ISBN 978-5-374-00463-2.
8. Morris T. P., White I. R., Crowther M. J. Using Simulation Studies to Evaluate Statistical Methods // Statistics in Medicine. 2019. Vol. 38. No. 11. P. 2074-2102. DOI: 10.1002/sim.8086.
9. Ермаков С. М. Метод Монте-Карло и смежные вопросы. 2-е изд., доп. М.: Наука, 1975. 471 с.
10. Nelsen R. B. An Introduction to Copulas. 2nd ed. New York: Springer, 2006. DOI: 10.1007/0-387-28678-0.
11. Уринцов А. И. Многоуровневые экономические информационные системы. М.: Московский международный институт эконометрики, информатики, финансов и права, 2003.
12. Deng A., Knoblich U., Lu J. Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2018. P. 233-242. DOI: 10.1145/3219819.3219919.
13. Бикмуллина И. И., Жданов А. С. Методология обработки ratio-метрик на основе линеаризации и дельта-метода для проверки гипотез в A/B-тестировании // Дидактика математики: проблемы и исследования. 2025. Вып. 3 (67). С. 44-61. DOI: 10.24412/2079-9152-2025-67-44-61.
14. Poyarkov A., Drutsa A., Khalyavin A., Gusev G., Serdyukov P. Boosted Decision Tree Regression Adjustment for Variance Reduction in Online Controlled Experiments // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 235-244. DOI: 10.1145/2939672.2939688.
15. Jin Y., Ba S. Towards Optimal Variance Reduction in Online Controlled Experiments. 2021. arXiv: 2110.13406.
16. Guo Y., Coey D., Konutgan M., Li W., Schoener C., Goldman M. Machine Learning for Variance Reduction in Online Experiments // Advances in Neural Information Processing Systems. 2021. Vol. 34.

Evaluating the Efficiency of the CUPED Method for Product Metrics in Digital Services

Minakov A.D.
Plekhanov Russian University of Economics

JEL classification: C10, C50, C60, C61, C80, C87, C90

Limited experimental traffic reduces the sensitivity of A/B tests and makes small product effects difficult to detect. This study evaluates classical CUPED, which uses pre-experiment values of the target metric to reduce the variance of the estimated effect. Statistical simulation is performed for conversion, action-count, and revenue metrics, and the analytical calculations are checked using the Monte Carlo method. CUPED efficiency is determined primarily by the correlation between metric values in the pre-experiment and experimental periods. At a correlation of 0.4, variance decreases by about 16% and MDE by 8.3%; at a correlation of 0.8, the reductions are 64% and 40%. No evidence of systematic bias is found. For sparse metrics and heavy right tails, analytical calculations should be checked against historical data.:

Keywords: CUPED, A/B test, variance reduction, statistical power, minimum detectable effect, product metrics, Monte Carlo method.

References

1. Minakov A. D., Vilkul A. A. O metodakh otsenki effektivnosti produktovykh izmenenii v tsifrovykh servisakh [Methods for Evaluating the Effectiveness of Product Changes in Digital Services]. Innovatsii i investitsii. 2026;(2):698. (In Russ.).
2. Urintsov A. I., Dik V. V. Sistema formirovaniya i prinyatiya reshenii v usloviyakh informatizatsii obshchestva [A System for Forming and Making Decisions in the Context of Society Informatization]. Moscow: Eurasian Open Institute; 2008. 223 p. (In Russ.).
3. Deng A., Xu Y., Kohavi R., Walker T. Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data. Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. 2013. P. 123–132. DOI: 10.1145/2433396.2433413.
4. Kohavi R., Tang D., Xu Y. Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge: Cambridge University Press; 2020.
5. Aivazyan S. A., Mkhitaryan V. S. Prikladnaya statistika. Osnovy ekonometriki. V 2 t. T. 1. Teoriya veroyatnostei i prikladnaya statistika [Applied Statistics. Fundamentals of Econometrics. In 2 vols. Vol. 1. Probability Theory and Applied Statistics]. 2nd ed. Moscow: UNITI; 2001. 656 p. (In Russ.).
6. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale: Lawrence Erlbaum Associates; 1988.
7. Urintsov A. I. Elektronnyi obmen dannymi [Electronic Data Interchange]. Moscow: Eurasian Open Institute; 2011. 181 p. (In Russ.).
8. Morris T. P., White I. R., Crowther M. J. Using Simulation Studies to Evaluate Statistical Methods. Statistics in Medicine. 2019;38(11):2074–2102. DOI: 10.1002/sim.8086.
9. Ermakov S. M. Metod Monte-Karlo i smezhnye voprosy [The Monte Carlo Method and Related Topics]. 2nd ed. Moscow: Nauka; 1975. 471 p. (In Russ.).
10. Nelsen R. B. An Introduction to Copulas. 2nd ed. New York: Springer; 2006. DOI: 10.1007/0-387-28678-0.
11. Urintsov A. I. Mnogourovnevye ekonomicheskie informatsionnye sistemy [Multilevel Economic Information Systems]. Moscow: Moscow International Institute of Econometrics, Informatics, Finance and Law; 2003. (In Russ.).
12. Deng A., Knoblich U., Lu J. Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2018. P. 233–242. DOI: 10.1145/3219819.3219919.
13. Bikmullina I. I., Zhdanov A. S. Metodologiya obrabotki ratio-metrik na osnove linearizatsii i delta-metoda dlya proverki gipotez v A/B-testirovanii [A Methodology for Processing Ratio Metrics Based on Linearization and the Delta Method for Hypothesis Testing in A/B Tests]. Didaktika matematiki: problemy i issledovaniya. 2025;3(67):44–61. DOI: 10.24412/2079-9152-2025-67-44-61. (In Russ.).
14. Poyarkov A., Drutsa A., Khalyavin A., Gusev G., Serdyukov P. Boosted Decision Tree Regression Adjustment for Variance Reduction in Online Controlled Experiments. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 235–244. DOI: 10.1145/2939672.2939688.
15. Jin Y., Ba S. Towards Optimal Variance Reduction in Online Controlled Experiments. 2021. arXiv:2110.13406.
16. Guo Y., Coey D., Konutgan M., Li W., Schoener C., Goldman M. Machine Learning for Variance Reduction in Online Experiments. Advances in Neural Information Processing Systems. 2021;34.