

CUPED

АЛЕКСАНДР САХНОВ
[linkedin.com/in/amsakhnov](https://www.linkedin.com/in/amsakhnov)

Staff MLE at Alibaba Group

2 сентября 2021 г.

- 1 Оценка эффекта. Как повысить чувствительность?
- 2 CUPED
- 3 Независимость ковариаты
- 4 CUPED. Проблемы и решения
- 5 Обобщения идеи CUPED и их применение

Оценка эффекта

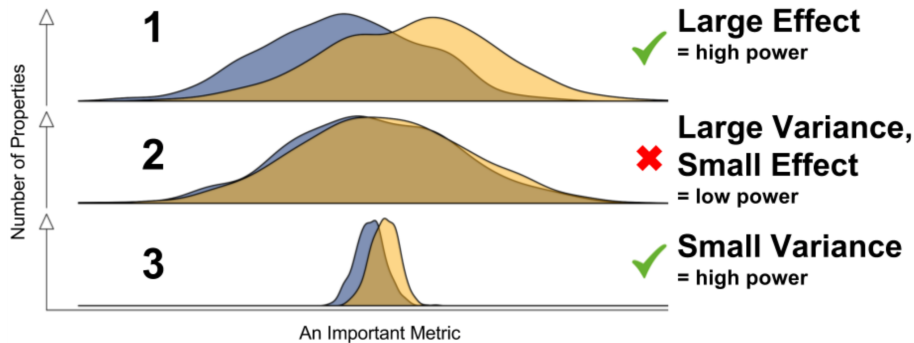


Figure 1. Distributions of properties in base (blue) and variant (yellow) for an important metric in three experiments

Оценка эффекта

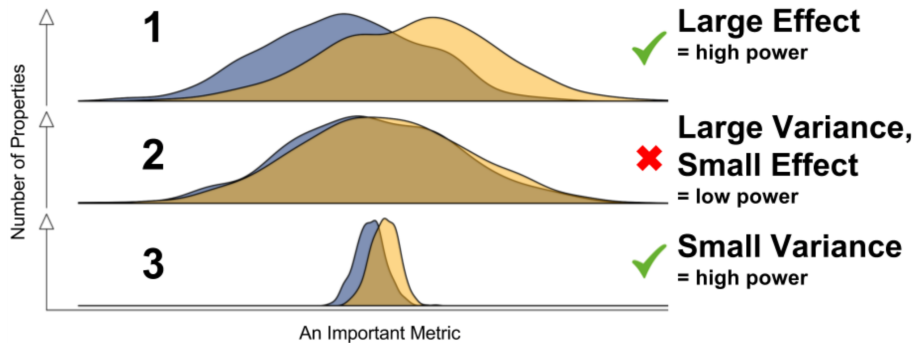


Figure 1. Distributions of properties in base (blue) and variant (yellow) for an important metric in three experiments

Даже маленькие эффекты могут нести **большую прибыль!**

Оценка размера теста

Размер теста

По величине эффекта, дисперсии, уровням значимости и мощности теста мы можем определить необходимый **размер теста**:

$$n > \frac{[\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)]^2 (\sigma_X^2 + \sigma_Y^2)}{\varepsilon^2}$$

Проблемы размера:

- проводить большие эксперименты долго и дорого
- на маленьких экспериментах можно не увидеть эффект

Мы хотим научиться отлавливать маленькие эффекты на сравнительно небольших объемах данных: **снизим дисперсию!**

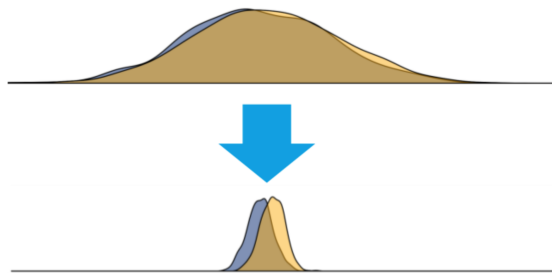


Figure 2. Example of how distribution can change when applying CUPED

За всё приходится платить

Чтобы уменьшить дисперсию нам необходимо получить дополнительную информацию.

Методы снижения дисперсии

- **Получить больше данных.** Самый надежный, но и самый дорогой способ. Мы хотим от него уйти.
- **Фильтрация выбросов.** Нужно делать обязательно!
- **Стратификация.** Просто напомним, что есть и такое!

Что мы не использовали ещё?

За всё приходится платить

Чтобы уменьшить дисперсию нам необходимо получить дополнительную информацию.

Методы снижения дисперсии

- **Получить больше данных.** Самый надежный, но и самый дорогой способ. Мы хотим от него уйти.
- **Фильтрация выбросов.** Нужно делать обязательно!
- **Стратификация.** Просто напомним, что есть и такое!

Что мы не использовали ещё?

Исторические данные

Если до начала пилота мы логировали данные по магазинам, то у нас есть данные:

- Хорошо структурированные!
- Бесплатные!
- Связанные с экспериментом!

Что такое CUPED?

Controlled-experiment Using Pre-Experiment Data

CUPED — техника А/Б-экспериментов, позволяющая увеличить чувствительность за счёт использования данных, полученных ранее.

Основная идея CUPED

1. Наблюдаемая дисперсия частично обусловлена неустранимым разбросом, а частично связана с влиянием ненаблюдаемых нами факторов.
2. Если есть основания полагать эти факторы постоянными, то они так же влияли на исторические данные.
3. Если определить связь между историческими данными и данными в эксперименте, то дисперсию можно уменьшить.

Изменение целевой метрики

Преобразование метрик

Нас интересует значение целевой метрики Y . При этом часто мы можем рассмотреть совместное распределение метрики Y и какой-то другой случайной величины X .

Таким образом, нам поступают независимые пары случайных величин (Y_i, X_i) . На основе этих пар мы можем определить новую метрику:

$$\hat{Y}_{cv} = \bar{Y} - \theta\bar{X} + \theta\mathbb{E}X.$$

Несмещенная оценка

В силу того, что $-\theta\mathbb{E}(\bar{X}) + \theta\mathbb{E}X = 0$, новая метрика даст несмещенную оценку для \bar{Y} :

$$\mathbb{E}\hat{Y}_{cv} = \mathbb{E}\bar{Y}.$$

Если в качестве X брать данные о историческом поведении пользователей, то мы анализируем изменение их поведения.

С точки зрения бизнеса

Идею можно объяснить бизнесу нормальным языком.

Вы говорите: "Нас интересует не только то, как в среднем ведут себя пользователи, а то, как **изменилось их поведение**."

Это позволяет вычитать усредненную базу.

Как вычитание может уменьшать дисперсию?

Новая метрика

Если у нас есть две случайные величины, то мы можем рассмотреть их разность:

$$Z = Y - X$$

Независимые с.в.

Для независимых с.в. дисперсия возрастет:

$$\mathbb{V}Z = \mathbb{V}Y + \mathbb{V}X$$

Зависимые с.в.

Всё меняет корреляция. Это похоже на формулу квадрата разности:

$$\mathbb{V}Z = \mathbb{V}Y - 2\text{cov}(Y, X) + \mathbb{V}X$$

Минимальная дисперсия

Дисперсия зависит от θ квадратично:

$$\begin{aligned}\mathbb{V}\hat{Y}_{cv} &= \mathbb{V}(\bar{Y} - \theta\bar{X}) = \mathbb{V}(Y - \theta X)/n \\ &= (\mathbb{V}Y - 2\theta\text{cov}(Y, X) + \theta^2\mathbb{V}X)/n.\end{aligned}$$

Это простая парабола с точкой минимума:

$$\theta_0 = \frac{\text{cov}(Y, X)}{\mathbb{V}X}$$

Тогда минимальная дисперсия равна:

$$\min(\mathbb{V}\hat{Y}_{cv}) = \mathbb{V}\bar{Y} \cdot (1 - \rho^2), \quad \rho = \frac{\text{cov}(Y, X)}{\sqrt{\mathbb{V}Y\mathbb{V}X}}$$

Любая коррелирующая с.в. может быть использована для уменьшения дисперсии.

Геометрическая интерпретация

Ковариация и скалярное произведение

Ковариация по своим свойствам напоминает скалярное произведение.

- Симметричность:

$$(a, b) = (b, a).$$

- Линейность:

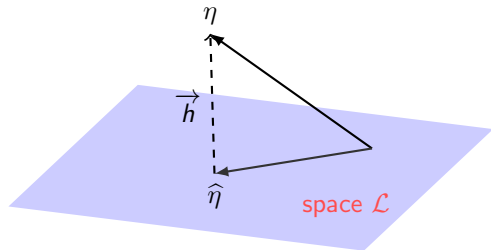
$$(\alpha a_1 + \beta a_2, b) = \alpha(a_1, b) + \beta(a_2, b).$$

- Неотрицательная определенность:

$$(a, a) \geq 0.$$

Всеми этими свойствами обладает и ковариация.

Процесс снижения дисперсии аналогичен проектированию вектора на подпространство ортогональное \vec{h} :



$$\vec{h} = \eta - \hat{\eta}$$

$$(\eta - \hat{\eta}, \xi) = 0, \forall \xi \in \mathcal{L}$$

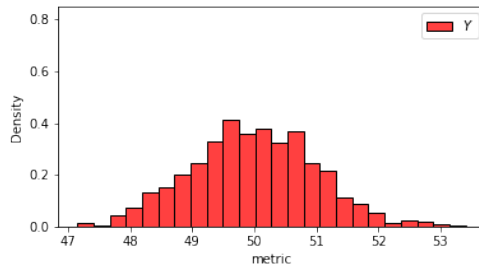
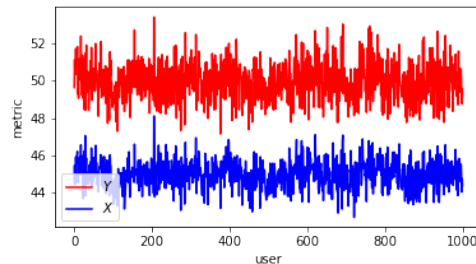
Коррелирующая с.в. позволяет снижать дисперсию

Снижение дисперсии без смещения

Переход к $\hat{Y} = Y - \theta \cdot (X - \mathbb{E}X)$ позволяет избежать смещения оценки среднего. При этом любая коррелирующая с.в. снизит дисперсию:

$$\mathbb{V}\hat{Y}_{cv} = (1 - \rho^2) \cdot \mathbb{V}Y.$$

UserId	Y	X	\hat{Y}
363	51.367	46.021	50.183
591	50.443	44.717	50.758
605	49.985	44.978	49.999
701	49.520	44.670	49.888
963	49.732	44.455	50.347
\mathbb{E}	49.983	44.991	49.983
\mathbb{V}	1.017	0.535	0.311



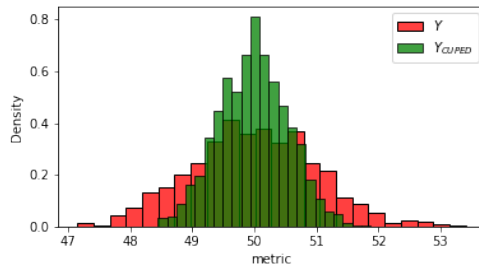
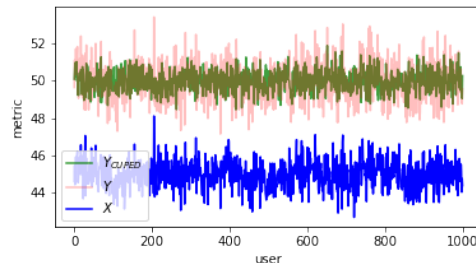
Коррелирующая с.в. позволяет снижать дисперсию

Снижение дисперсии без смещения

Переход к $\hat{Y} = Y - \theta \cdot (X - \mathbb{E}X)$ позволяет избежать смещения оценки среднего. При этом любая коррелирующая с.в. снизит дисперсию:

$$\mathbb{V}\hat{Y}_{cv} = (1 - \rho^2) \cdot \mathbb{V}Y.$$

UserId	Y	X	\hat{Y}
363	51.367	46.021	50.183
591	50.443	44.717	50.758
605	49.985	44.978	49.999
701	49.520	44.670	49.888
963	49.732	44.455	50.347
\mathbb{E}	49.983	44.991	49.983
\mathbb{V}	1.017	0.535	0.311



Применение CUPED в AB экспериментах

Онлайн эксперимент

В AB-эксперименте мы работаем с экспериментальным и контрольным рядами. Нас прежде всего интересует разница:

$$\Delta = \bar{Y}^{(t)} - \bar{Y}^{(c)}.$$

Для несмещенных оценок мы можем заменить это на:

$$\Delta_{cv} = \hat{Y}_{cv}^{(t)} - \hat{Y}_{cv}^{(c)}.$$

Требования к ковариате

Нам необходимо найти ковариату X сильно коррелированную с Y . Одновременно с этим, необходимо знать мат. ожидание $\mathbb{E}X$. Всего этого достичь непросто.

Однако, требования можно ослабить. Для нахождения разности нам действительно важно только то, что $\mathbb{E}X^{(t)} - \mathbb{E}X^{(c)} = 0$. Тогда Δ_{cv} будет давать несмещенную оценку.

ВАЖНО: отсутствие влияния эксперимента на X в совокупности со случайным семплированием обеспечивают это условие:

$$\mathbb{V}\Delta_{cv} = (1 - \rho^2)\mathbb{V}\Delta.$$

Пара слов о независимости

Доведем до абсурда

Мы знаем, что чем выше коэффициент корреляции, тем ниже дисперсия. Но ведь лучше всего с.в. коррелирует сама с собой!

$$\tilde{Y} = Y - \theta \cdot (Y - \mathbb{E}Y) = Y - \frac{\text{cov}(Y, Y)}{\mathbb{V}Y} \cdot (Y - \mathbb{E}Y) = Y - 1 \cdot (Y - \mathbb{E}Y) = \mathbb{E}Y.$$

И мы получаем нулевую дисперсию.

Но так делать НЕЛЬЗЯ!

Нет ничего важнее причины

Есть масса способов обмануть себя и "уменьшить" дисперсию. Большинство из них связано с использованием зависимых данных.

Если вы действительно хотите добиться хороших результатов, то, прежде всего, убедитесь, что достигаете этого за счет добавления новой информации!

CUPED

Где брать независимые, но связанные данные?

Y не должна оказывать воздействия на X .

Принцип причинности гарантирует отсутствие такого воздействия в том случае, если X собран до начала эксперимента.

Обратим внимание: **Pre-Experiment Data**

Что может выступить ковариатой?

- Хорошей ковариатой будет та же самая метрика, но собранная на периоде до эксперимента.
- Ковариатой может выступать любая величина, на которую эксперимент не мог повлиять: пол, возраст, страна пользователя. В этом смысле метод идейно близок к стратификации.
- Можно использовать статистики первого посещения пользователем нашего эксперимента. Например, день или время когда он первый раз попал в эксперимент. Ведь на тот момент пользователь про эксперимент ничего не знал. Это может быть хорошей идеей, особенно если мы внесли в дизайн изменение, связанное со временем или добавили товары особенно популярные в выходные.

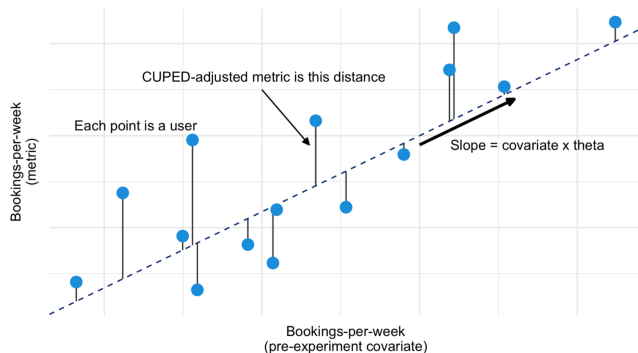
CUPED и регрессия

Регрессия

Если по одной оси отложить данные старого периода, а по другой — нового, то регрессия показывает наследуемость признаков, а остаток — изменчивость.

Именно эту наследуемость мы хотим убрать.

А изменчивость выделить. В ней содержится основной эффект.



Многопараметрический CUPED

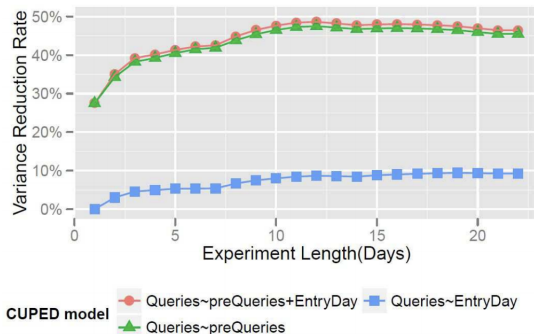
Что дает применение нескольких ковариат?

- Как в регрессии несколько признаков позволяют добиться меньшей ошибки, так и в CUPED несколько ковариат позволяют сильнее снизить дисперсию.
- Мы можем сравнить какое снижение дисперсии приносит, как каждая ковариата в отдельности, так и они все вместе.

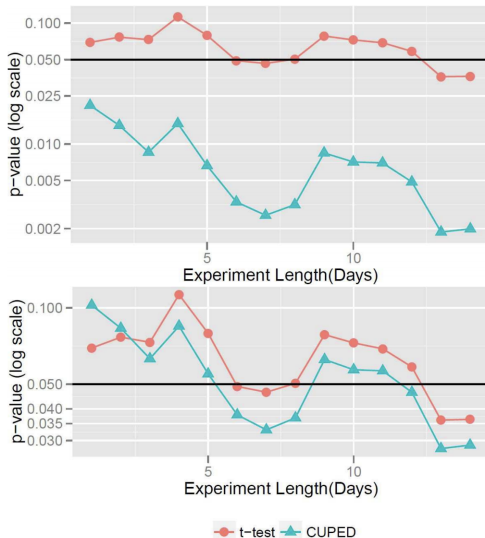
Microsoft исследовал изменение queries-per-user метрики для запросов в Bing. При этом использовались две ковариаты.

1. EntryDay — категориальный признак, показывающий день первого появления пользователя в эксперименте;
2. Число запросов за 1 неделю предэкспериментального периода.

Большая часть снижения приходится на предэкспериментальную метрику.



Какой выигрыш дает CUPED?



Microsoft провели ухудшающий эксперимент, который исследовал как снижается CTR при искусственном замедлении ответа сервера на 250 мс.

Выигрыш в чувствительности

Применение CUPED позволило получить значимый результат уже в первый день эксперимента. Можно не ждать две недели.

Выигрыш в размере эксперимента

Применение CUPED на вдвое меньшей группе пользователей позволило получить результаты сопоставимые (и даже чуть лучше), чем без CUPED.

Работа с пропущенными значениями

Лучшая ковариата

Лучшая ковариата для CUPED — та же самая метрика на предыдущем периоде.

Но мы не можем гарантировать, что пользователь из эксперимента приходил к нам в предэкспериментальный период. Значение ковариаты может отсутствовать

exp_id	unit_id	grp	metric	covariate
1	1	base	7	6
1	2	variant	12	NULL
1	3	variant	11	7
2	1	base	8	10
2	2	base	6	5
2	3	variant	9	NULL

Дополнительный признак

Мы можем ввести дополнительный признак: присутствует ли пользователь в предэкспериментальных данных. Такой подход эквивалентен разбиению всех пользователей на две страты по этому бинарному признаку.

Заполнение пропущенных значений

Введение дополнительного бинарного признака в задачу позволяет заполнить пропущенные значения любой константой по нашему выбору.

Такая техника позволяет эффективно работать с пропущенными значениями.

Корреляция и каузальность

Ложная корреляция

Мы пытаемся в массе данных отыскать скрытые взаимосвязи.

В больших объемах данных удастся встретить коррелирующие, но не имеющие реальной взаимосвязи ряды данных. Предсказательной силы это не имеет.



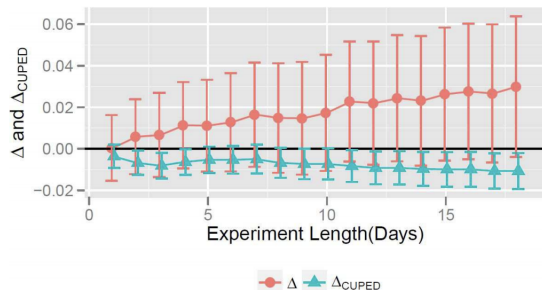
Каузальность

Важна взаимосвязь между случайными величинами.

- Мы не хотим, чтобы X находился в причинной зависимости от Y .
- Мы хотим, чтобы Y и X имели общую причину. Наш метод призван удалить влияние этой общей причины.

Нельзя увлекаться поиском дополнительных ковариат. Это может привести к переобучению.

Казусы при использовании зависимых данных



В Bing посчитали результаты для метрики queries-per-user оценку и доверительные интервалы с помощью CUPED с использованием высокоррелированной метрики distinct-queries-per-user. Метрики оказались противоположно направленными!

Наиболее коррелированная метрика

В топ по коэффициенту корреляции может пролезть метрика зависящая от эксперимента. Такие надо беспощадно отсекать!

Недооценка эффекта

Если мы используем связанную ковариату, то можем недооценить эффект эксперимента.

Известно, что число кликов коррелирует со скоростью загрузки веб-страницы. Быстрее загрузка — больше кликов.

Если выбрать число кликов как ковариату в эксперименте по улучшению скорости загрузки страницы, получим недооцененный эффект.

Когда CUPED не работает?

Что посчитать не получится?

Если бизнес задача требует подсчета квантилей или каких-то иных метрик, изменяющихся при линейных преобразованиях, то CUPED не подойдет.

Uber любит смотреть на квантили. Например, среднее время ожидания может остаться неизменным, а 99-процентный квантиль при этом возрастет. И начнут уходить пользователи. В анализе этого эффекта CUPED не поможет.

Обобщения идеи CUPED

Использование нескольких ковариат

Мы можем использовать несколько ковариат. Другой вопрос, откуда брать независимые данные. Попробуйте предложить идеи!

Немного машинного обучения

Мы говорили о прямом использовании исторических данных в качестве ковариаты. Можно поступить иначе.

1. На основе исторических данных обучить модель, предсказывающую значения метрики на период эксперимента.
2. Получить результаты предсказания.
3. Эти результаты использовать в качестве ковариаты.

CUPED и консервативный прогноз погоды

Погода на завтра

Нам надо знать, что надеть. Брать ли с собой зонт, надевать ли свитер. Или выходить в шортах и босоножках.

Мы прекрасно знаем, что сегодня погода, скорее всего, будет такая же, как вчера. А завтра, как сегодня.

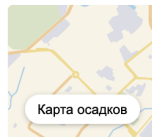
Погода в Москве





Яндекс.Погода ▾

По часам · На 10 дней · На месяц · Осадки на карте

+18° 

Облачно с прояснениями
6,3 м/с ветер
+10 утром, +11 днем



сегодня	вс 18	пн 19	вт 20	Прогноз на 10 дней
+18 	+13 	+10 	+10 	>
+9	+7	+7	+5	

CUPED и консервативное прогнозирование

Используя в качестве ковариаты исторические значения метрики на предыдущем периоде мы неявно делаем **консервативный прогноз**: мы ожидаем, что на неделе эксперимента будет всё в точности как на предыдущей неделе.

Консервативный прогноз убирает большую часть изменчивости. Но он не оптимален. Оптимум достигается тогда, когда ошибку нельзя уменьшить.

Учет сезонности

Как улучшить прогноз?

Мы можем посмотреть на исторические данные и тренды.

- В апреле можно ожидать, что каждый следующий день будет чуть теплее предыдущего.
- Наоборот, в октябре каждый следующий день скорее окажется чуть холоднее.

Климат Москвы (данные по температуре воздуха за последние 10 лет (август 2006 — июль 2016 гг.))													
Показатель	Янв.	Фев.	Март	Апр.	Май	Июнь	Июль	Авг.	Сен.	Окт.	Нояб.	Дек.	Год
Средний максимум, °C	-5,5	-3,5	3,2	11,6	20,1	22,7	25,8	23,5	16,5	8,6	2,6	-1,7	10,3
Средняя температура, °C	-7,4	-5,8	-0,1	7,1	14,7	17,7	20,6	18,6	12,6	6,1	1,1	-3,2	6,8
Средний минимум, °C	-9,2	-8,1	-3,5	2,5	9,4	12,6	15,4	13,8	8,7	3,7	-0,3	-4,6	3,4
Норма осадков, мм	49	44	39	39	62	61	85	78	73	68	57	54	708

Сезонность в бизнес-метриках

Многие товары в отдельности имеют выраженную сезонность: цветы лучше всего продаются 14 февраля и 8 марта, на майские праздники хорошо продаются шашлыки. Большинство бизнес-метрик тоже имеют сезонность.

Почему учет сезонности снижает дисперсию?



Из чего состоят ошибки

Если бы неучет сезонности и трендов приводил к ошибке на константу, то это не изменило бы дисперсию.

Дополнительное снижение дисперсии возможно за счет разнонаправленных трендов. Для одних объектов метрика растет, для других уменьшается. Где-то она остается неизменной.

Погода в разных полушариях

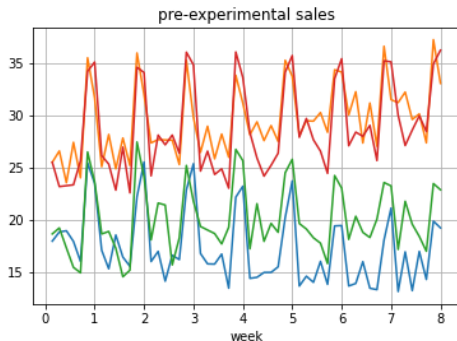
Если мы хотим предсказать погоду для городов по всему миру, то в каждом случае поправка будет своя:

- в Москве в апреле теплеет;
- в это время в Сиднее холодает;
- экватор вообще отдельный разговор

Предсказание продаж. Планирование

Исторические значения метрики

В разных магазинах мы можем наблюдать различные тренды. Где-то продажи растут, где-то они могут сокращаться.



Что мы предсказываем?

Мы планируем двухнедельный эксперимент.

Для анализа можно выбрать значения продаж в 100 магазинах. По результатам продаж за 2 предыдущих месяца мы можем предсказать продажи в следующие две недели.

Независимость данных

Важно помнить, что для стат. тестов нужны независимые данные.

Каждый из временных рядов может дать нам только **одно** независимое значение.

Например, мы можем вычислить общий объем продаж.

Предсказание продаж. Результаты

Ошибки остаются

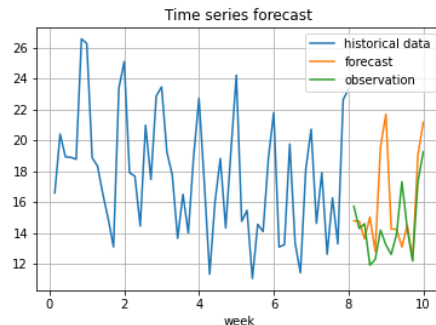
Вряд ли удастся построить абсолютно точный прогноз.

Мы видим, что наша модель не угадала драматическое падение продаж в первые выходные. Могла вмешаться погода или какие-то другие непредвиденные факторы.

Но даже при этом результат модели ближе к реальности, чем исторические данные.

Боритесь за качество модели

Чем точнее будет ваша модель предсказания, тем больше удастся снизить дисперсию.



Общий объем продаж:

- 2 недели до эксперимента: 230.33
- прогноз: 220.81
- реальные продажи: 203.08

Предсказание продаж. Достигнутые результаты

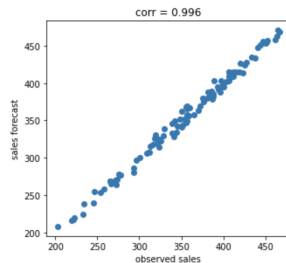
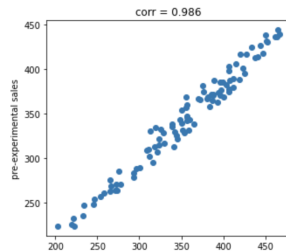
Увеличение взаимной корреляции

Можно заметить, что результаты нашего предсказания гораздо лучше коррелируют с настоящими значениями продаж. 0.9960 против 0.9862. За счет этого мы можем снизить дисперсию в 3.45 раза.

Эффект \Rightarrow разладка, смена тренда

Положительный эффект приводит к изменению поведения покупателей.

Это изменение мы и хотим детектировать. Фактически, успех эксперимента будет означать, что мы задали новый тренд. Покупатели стали вести себя иначе.



Бизнес-примеры

CUPAC

Control Using Predictions as Covariate — немножко ML на исторических данных позволяет, не написав ни единой математической выкладки, получить новый интересный метод.

- Ошибка предсказания хорошо обученной модели временного ряда имеет нулевое мат. ожидание. Иначе мы можем её ещё дообучить.
- По историческим данным научились хорошо предсказывать средний чек.
- Такое предсказание отвечает всем требованиям к ковариате в методе CUPED. Оно высоко коррелировано и независимо от эксперимента, т.к. строится на исторических данных.
- Результат предсказания применяется как ковариата в методе CUPED.

Достигнутые результаты

- Обычно можно ожидать 2-4 кратное снижение дисперсии, что в 2-4 раза сократит продолжительность теста. Или увеличит чувствительность.
- Лучший достигнутый результат — снижение дисперсии в ~ 40 раз.

Резюме

Мы научились

- Уменьшать дисперсию случайной величины с помощью использования коррелированной случайной величины
- Обсудили применение к AB-тестированию
- Изучили метод CUPED
- Прикрутили немного ML и получили CUPAC

А ещё CUPED позволяет пересчитать результаты по ранее проведенным экспериментам и увеличить чувствительность.

Дополнительные материалы

Ссылки для самостоятельного изучения

1. Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data
2. Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix
3. How Booking.com increases the power of online experiments with CUPED
4. Увеличение чувствительности А/Б-тестов с помощью Cuped. Доклад в Яндексе
5. Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix
6. Improving Experimental Power through Control Using Predictions as Covariate (CUPAC)