Продвинутые методы A/B тестирования

Команда курса

- Оля Кравцова @oakravts
- Юлиан Сердюк @jserdyuk

Зачем проходить этот курс?

1. Узнаете, какие проблемы стоят перед бизнесом и поймете, как аналитики их решают
2. Погрузитесь в продвинутые методы оценки экспериментов, которые являются best-practices в настоящее время в индустрии
3. Основные методы (t-test, бутстрап, тест манна-уитни) известны всем, но их возможностей порой недостаточно.

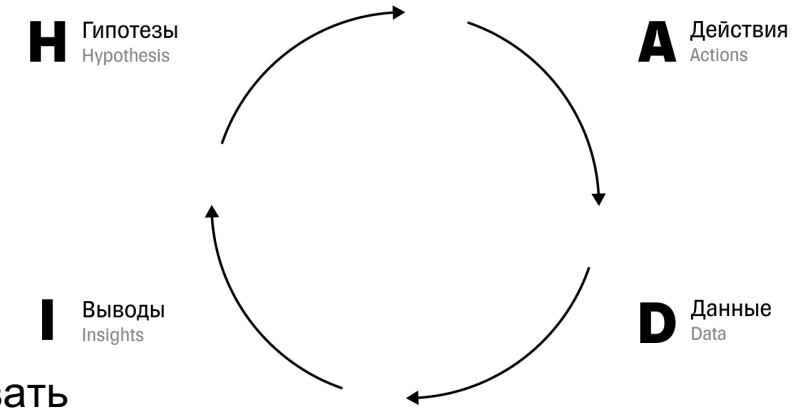
Data-driven подход

В каждом решении так или иначе должна фигурировать причинно-следственная связь, которая поможет надежно оценить эффективность того или иного изменения в процессах или воздействии. И именно поэтому мы обращаемся к A/B тестированию, как механизму оценки всех наших инициатив/идей/гипотез.

HADI цикл

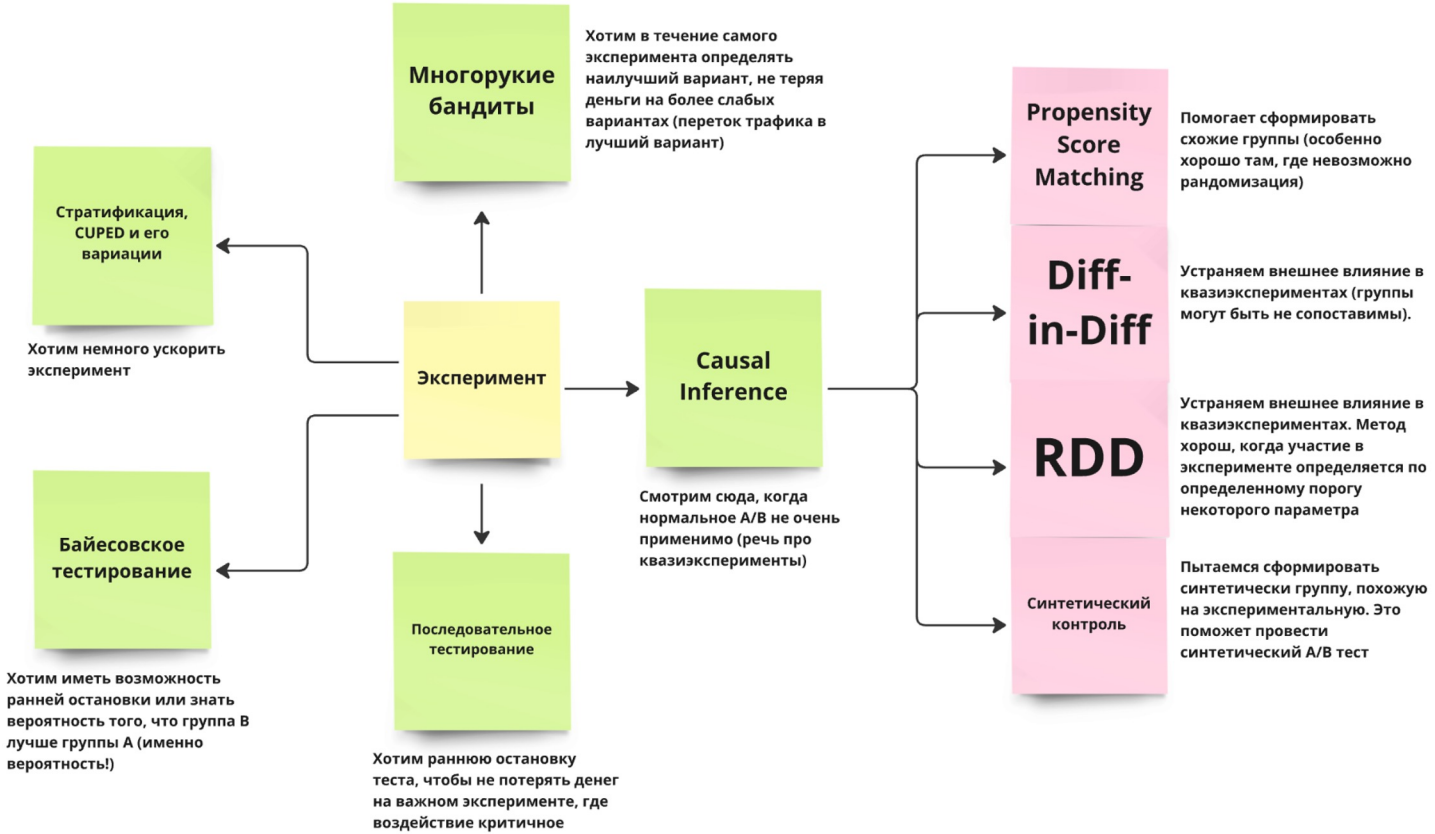
HADI (Hypothesis, Action, Data, Insight) — это цикл итеративного улучшения и оптимизации продукта.

1. **Hypothesis** (Гипотеза)
2. **Action** (Действие)
3. **Data** (Данные)
4. **Insight** (Выводы)



Цикличность HADI позволяет быстро тестировать идеи, минимизируя затраты времени и ресурсов
=> эффективное развитие продукта

Карта методов курса



Кейсы применения методов

Ускорение тестов

Ускорение тестов

Цель: Подтверждение эффекта на GMV для защиты бюджета продукта Видеоаналитика на инвестиционном комитете.

Описание продукта: Видеоаналитика X5 предотвращает отсутствие товаров на полках, обеспечивая их наличие и снижая потери продаж (недополученная выручка).

Проблема: Товар есть на складе, но отсутствует на полке => снижение продаж и отток клиентов.

Ускорение тестов

Проблема классического A/B теста:
Требуется более 5 недель для детекции эффекта в 2.5%.

Решение: Применение ускоренных методов:

- Стратификация, CUPED и ML-подходы.
- Сокращение длительности эксперимента до 3 недель.

Результат: Обнаружен эффект в 2.5% на GMV за 3 недели.

Итог: Успешная защита бюджета на следующий цикл финансирования.

Последовательное тестирование

Последовательное тестирование

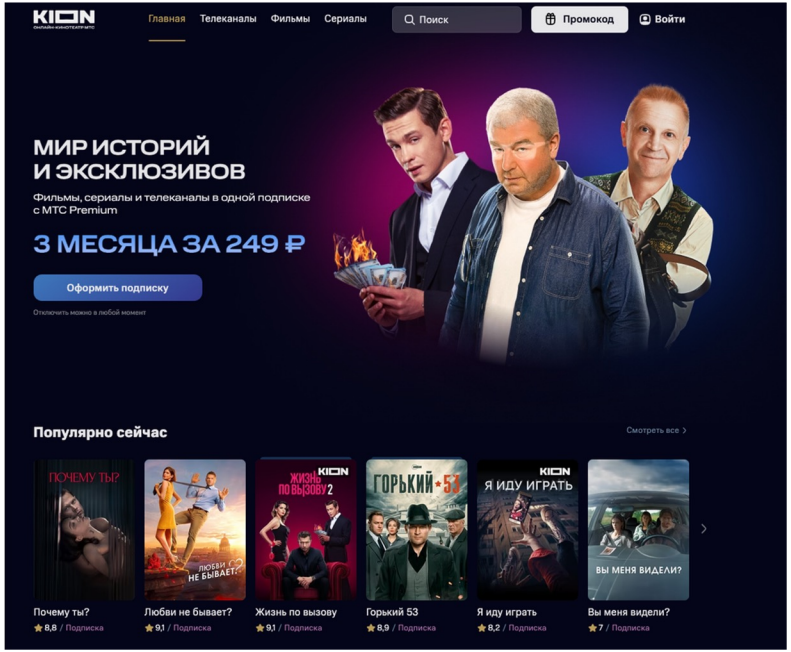
Контекст: Использование последовательных тестов для контроля эксперимента.

Задача: Обеспечить раннюю остановку теста при ухудшении метрик в целевой группе.

Ситуация: На 3-й день эксперимента новая рекомендательная модель значительно просадила TVTu.

Действия: Оперативная реакция и остановка теста.

Результат: Минимизация убытков и предотвращение долгосрочных потерь из-за багов в модели.



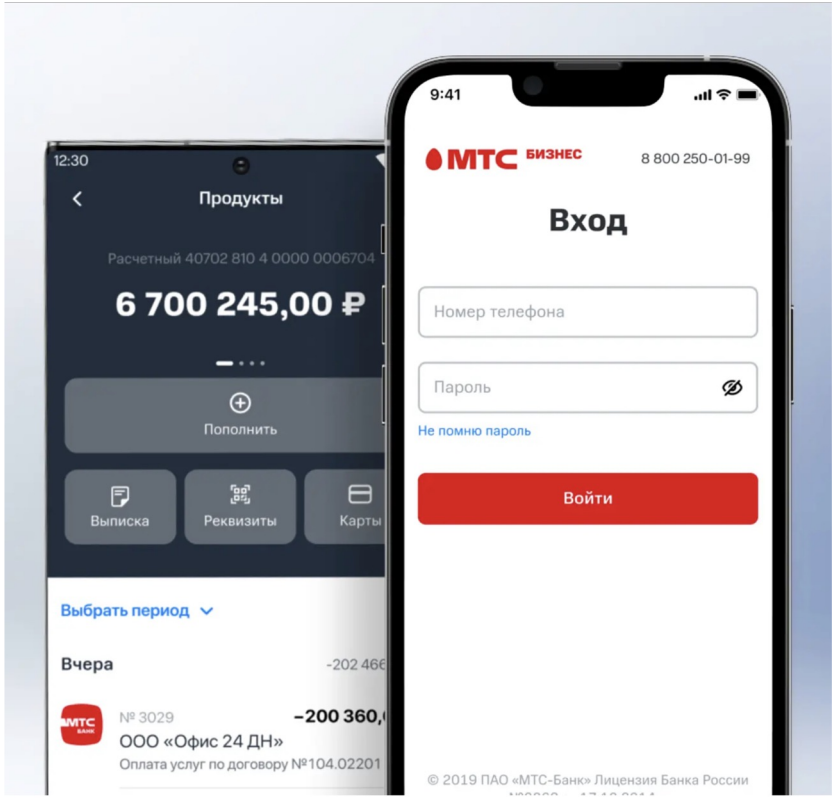
Байесовское тестирование

Байесовское тестирование

Цель: Проверить влияние нового UI мобильного приложения на удержание и вовлеченность пользователей через увеличение количества транзакций.

Метрика для принятия решения: Среднее количество транзакций в приложении на пользователя.

Вопрос бизнеса: Какова вероятность, что новый дизайн лучше текущего, и каковы ожидаемые потери, если дизайн окажется хуже?



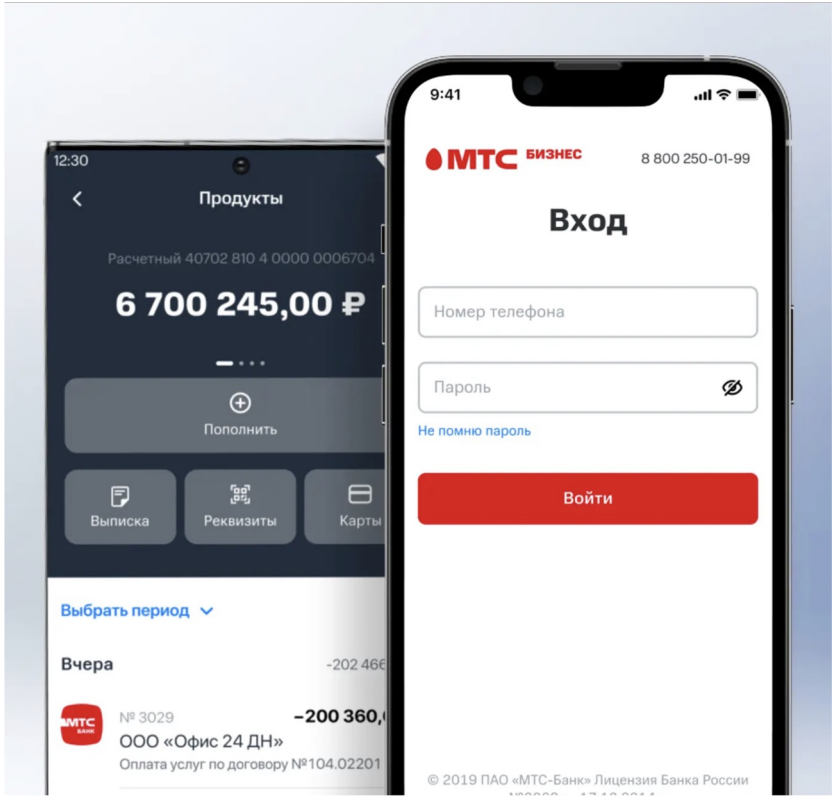
Байесовское тестирование

Подход: Байесовский метод для моделирования распределения эффекта с расчетом вероятности и риска.

Результаты:

- Вероятность того, что новый дизайн лучше — **90%** (оценено с использованием MCMC).
- При 10% вероятности неудачи **ожидаемые потери** составят **в среднем 0.5%** транзакций.
- В случае успеха — прирост в **5%** транзакций.

Вывод: Потенциальная прибыль существенно перевешивает возможные потери, что делает новый дизайн экономически целесообразным.



Многорукие бандиты

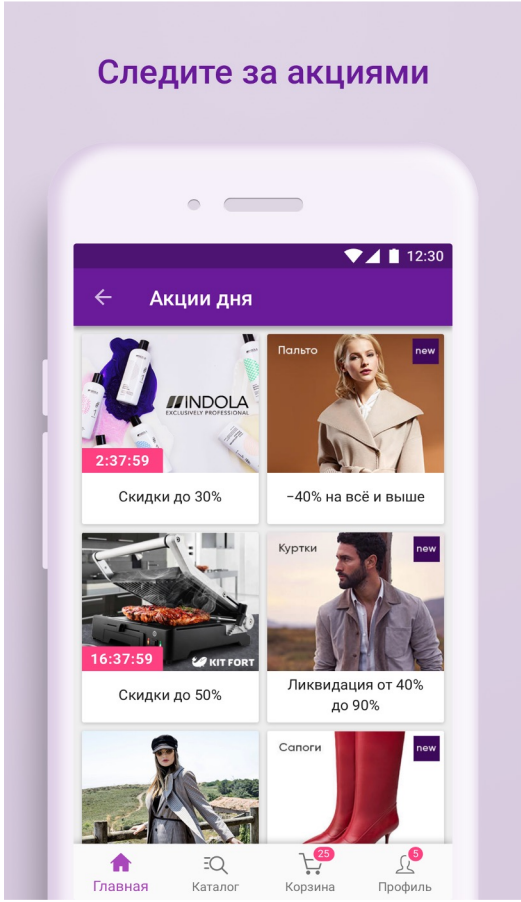
Многорукие бандиты

Цель: Оптимизация внешнего вида корзины в мобильном приложении для увеличения GMV.

Исходные гипотезы: 8 комбинаций изменений в корзине (размещение чекбоксов, блок предложений, быстрый просмотр товара).

Проблемы классического подхода:

- Поправки на множественное тестирование снижают чувствительность и замедляют получение результатов.
- Тестирование каждой гипотезы по отдельности требует слишком много времени.



Многорукие бандиты

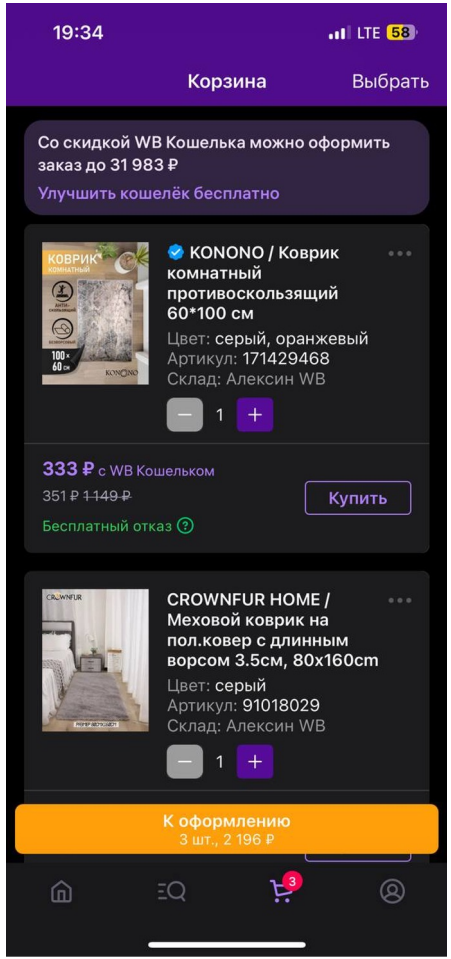
Решение: Запуск многорукого бандита, который динамически перераспределяет трафик к лучшим вариантам.

Целевая метрика: GMV (Gross Merchandise Value).

Процесс тестирования:

- Через 3 дня бандит начал направлять трафик к лидирующему варианту.
- Через 2 недели почти весь трафик перешел на победившую версию: левый чекбокс, доп. рекомендации и быстрый просмотр товара.

Результат: Победившая версия успешно раскатана на всю аудиторию мобильного приложения.



Propensity Score Matching

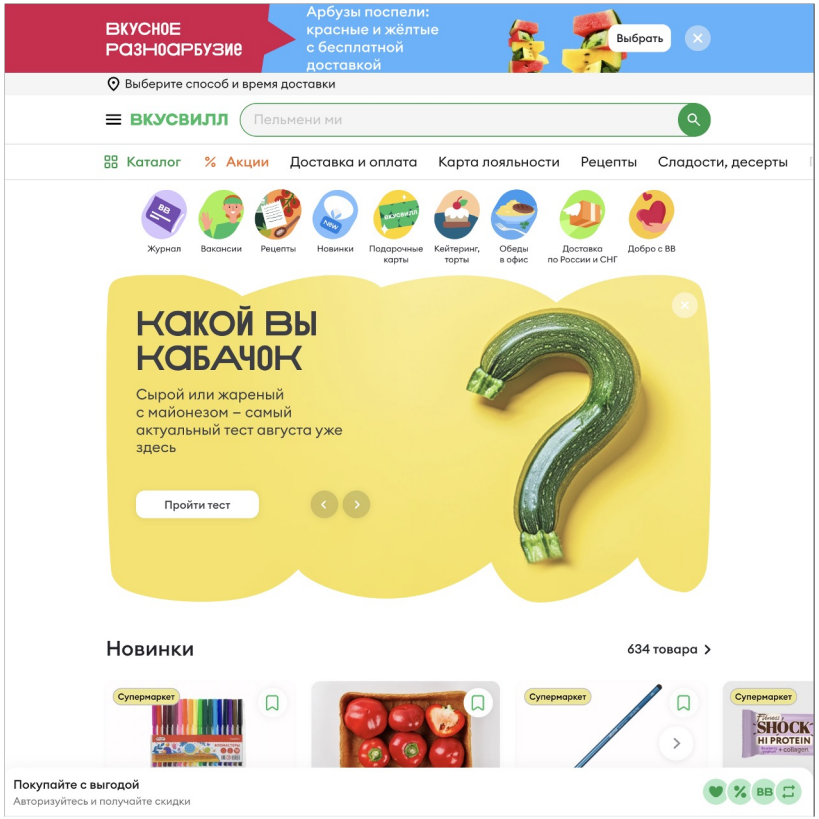
Causal Inference. PSM

Цель: Оценить эффективность новой программы лояльности (предложение со скидкой на новинки за каждый 10-й чек).

Задача: Проверить влияние программы на метрики: выручка, количество чеков и средний чек.

Проблема: Невозможность частичного развертывания программы — она внедрена массово на регионы (Москва и СПб).

- Прямое сравнение с другими регионами (например, Краснодар) некорректно из-за различий в характеристиках.



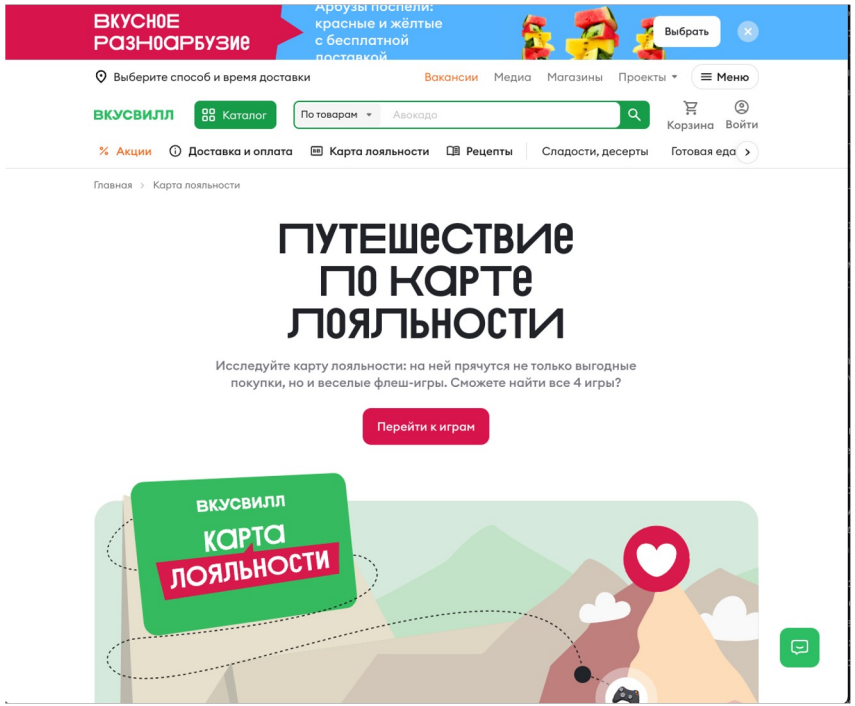
Causal Inference. PSM

Решение: Применение метода **Propensity Score Matching (PSM)** для подбора похожих покупателей из других регионов.

- Подбор контрольной группы с аналогичной склонностью к участию в программе на основе демографии, поведения и других факторов.

Результаты: Сравнение выручки в целевой и контрольной группах.

- **Итог:** Новая механика не дала статистически значимого роста выручки.
- **Вывод:** Возможная слабость предложения — дальнейшая проработка программы необходима.



Difference-in-Difference

Causal Inference. Diff-in-Diff

Цель: Увеличение трафика и выручки в маленьких магазинах через внедрение кофе- и сокоматов.

Гипотеза: Наличие автоматов с кофе и соком побудит покупателей зайти и дополнительно приобрести другие товары.

Дизайн эксперимента:

- Тестирование на всех магазинах в Санкт-Петербурге.
- Магазины в СПб получили автоматы, а в Москве мы ничего не меняем.

Causal Inference. Diff-in-Diff

Проблема с подбором контрольной группы:

- Различия между магазинами в Москве и СПб делают подход PSM неэффективным.

Решение: Применение метода **Diff-in-Diff**.

- Учет различий в регионах и сезонных факторов через сравнение изменений показателей за разные периоды.

Результаты:

- Через 4 недели после внедрения автоматов в магазины СПб:
  - **Выручка** выросла на **3%**.
  - **Количество чеков** увеличилось на **7%**.
- Эти результаты свидетельствуют о дополнительном притоке трафика в магазины.

Regression Discontinuity Design

Causal Inference. RDD

Цель: Оценить влияние начисления 2% кэшбека на покупателей, которые тратят более 50 тыс рублей в месяц.

Задача: Повышение лояльности среди высокодоходных клиентов.

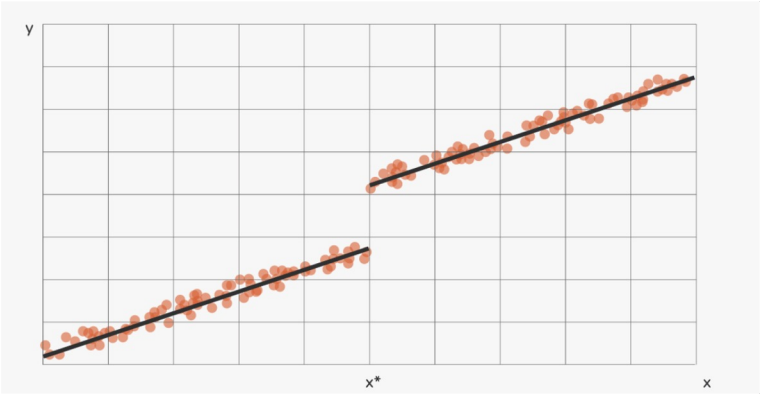
Проблема: Трудности с подбором контрольной группы из-за различий в поведении между высокодоходными и менее доходными покупателями.

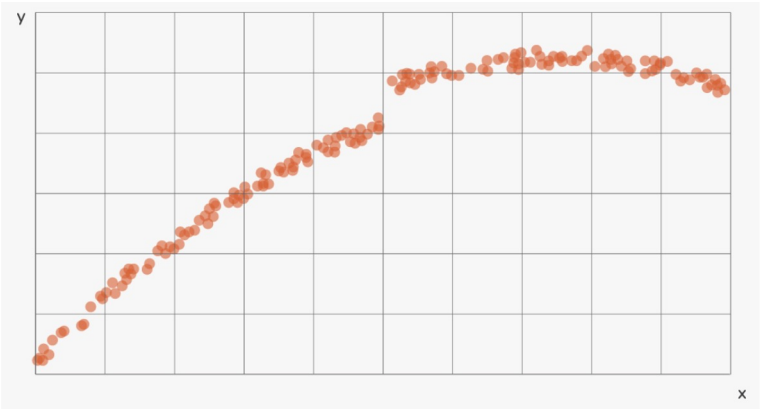
Causal Inference. RDD

Решение: Применение метода RDD:

- **Подход:** Сравнение покупателей, чьи траты близки к порогу 50 тыс рублей (например, 49 500 и 50 500 руб.).
- Построение регрессионных моделей на целевую метрику (выручку) для групп с расходами < 50 тыс и > 50 тыс рублей.

Ключевая идея: Разрыв в прогнозируемых значениях регрессий около порога 50 тыс и будет оценкой эффекта кэшбека.
Результат: Определение размера эффекта на целевую метрику, связанного с внедрением кэшбека.





Мы ожидаем, что вы уже:

- ✓ Умеете оценивать эксперименты классическими методами (t-test, тест манна-уитни, бутстрап)
- ✓ Знаете математику, заложенную в базовые статистические подходы
- ✓ Можете сформулировать и объяснить каждый этап жизненного цикла A/B тестов, почему они нужны и важны

Программа курса

- ✓ Курс охватывает современные методы ускорения и оптимизации A/B-тестов, включая стратификацию, CUPED, байесовское тестирование и алгоритм «Многорукий бандит». Студенты изучат последовательное тестирование, методы «A/B без A/B», такие как Matching, Diff-in-Diff и Synthetic control, что позволит эффективно решать сложные бизнес-задачи.

Блок 1	**Обзор методов**
1.1	Примеры проблем и их решений с помощью методов из курса
1.2	Карта методов
Блок 2	**Ускорение A/B-тестов**
2.1	Стратификация/Постстратификация
2.2	CUPED
2.4	Связь CUPED и линейной регрессии
2.3	CUPAC
Блок 3	**Последовательное тестирование (Sequential testing)**
3.1	Последовательный анализ Вальда
3.2	Отношение правдоподобия
3.3	SPRT, T-SPRT, mSPRT
3.4	Pros and cons
Блок 4	**Байесовское тестирование**
4.1	Отличия частотного и байесовского подхода
4.2	Алгоритм проведения байесовского A/B-теста
4.3	Метод Монте-Карло для марковских цепей
4.4	Pros and cons
Блок 5	**Многорукие бандиты**
5.1	Концепция многоруких бандитов
5.2	Отличия MAB/CMAB и зона применимости
5.3	MAB-подход, Thompson sampling
5.4	CMAB-подход, Linear Thompson sampling
5.5	Pros and cons
Блок 6	**«A/B без A/B» 1**
6.1	Matching methods, Propensity score matching
6.2	Diff in Diff
Блок 7	**«A/B без A/B» 2**
7.1	Regression Discontinuity Design
7.2	Synthetic control

Оценка

Активность	Вес	Описание
Домашние задания	70%	6 домашних заданий. За каждую домашнюю работу можно набрать 10 баллов
Экзамен	30%	Экзамен — это набор задач, которые нужно решить за отведенное время. Максимально можно набрать 10 баллов

Формула расчета итоговой оценки: 0,7 × среднее за домашние задания + 0,3 × экзамен.