

# Прикладная статистика в машинном обучении

## Лекция 7

### A/B-тестирование

И. К. Козлов  
(Мехмат МГУ)

2022

# АБ тестирование

1 АБ тестирование

2 Размер выборки

3 Стратификация

4 CUPED

5 Вместо послесловия

IT — это простенькая фарма

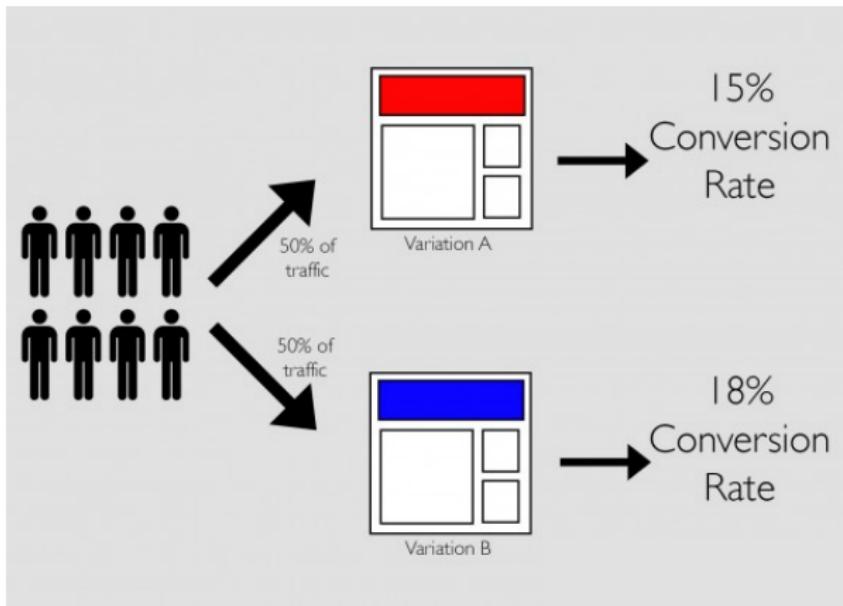


*Двойное слепое рандомизированное плацебо-контролируемое испытание.*

Сильно упрощая — АБ-тесты.

## AB tests

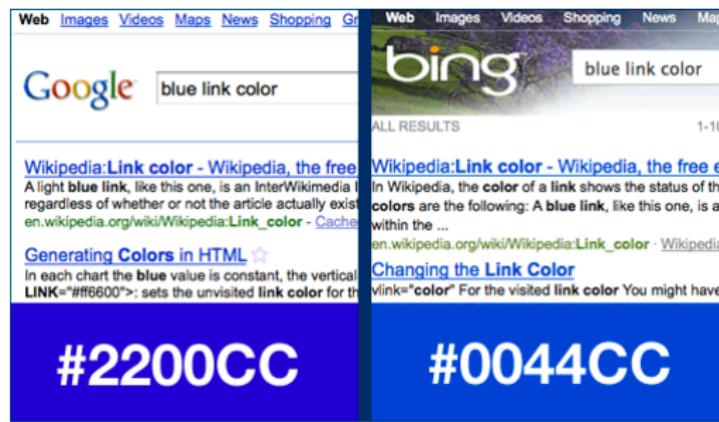
Меняем один из параметров, смотрим на выгоду.



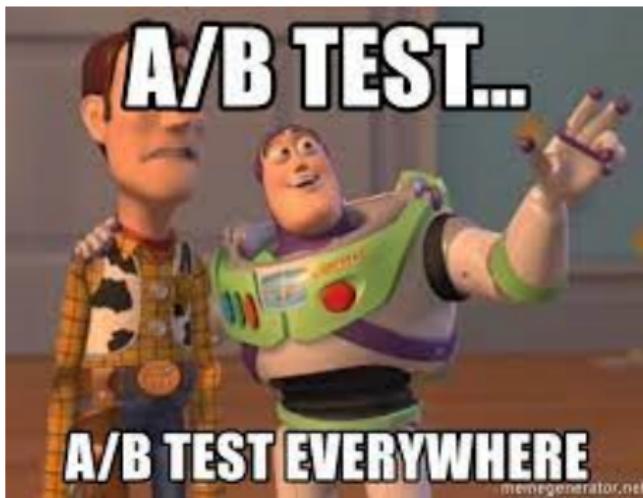
# 41 shades of blue

Особенности IT (поисковики, Netflix и т.д.):

- огромный поток пользователей,
- куча проблем вплоть до “какого цвета сделать ссылки”.



Естественно, АВ-тесты прельщают своей простотой, объективностью и клиентоориентированностью.



## AB tests

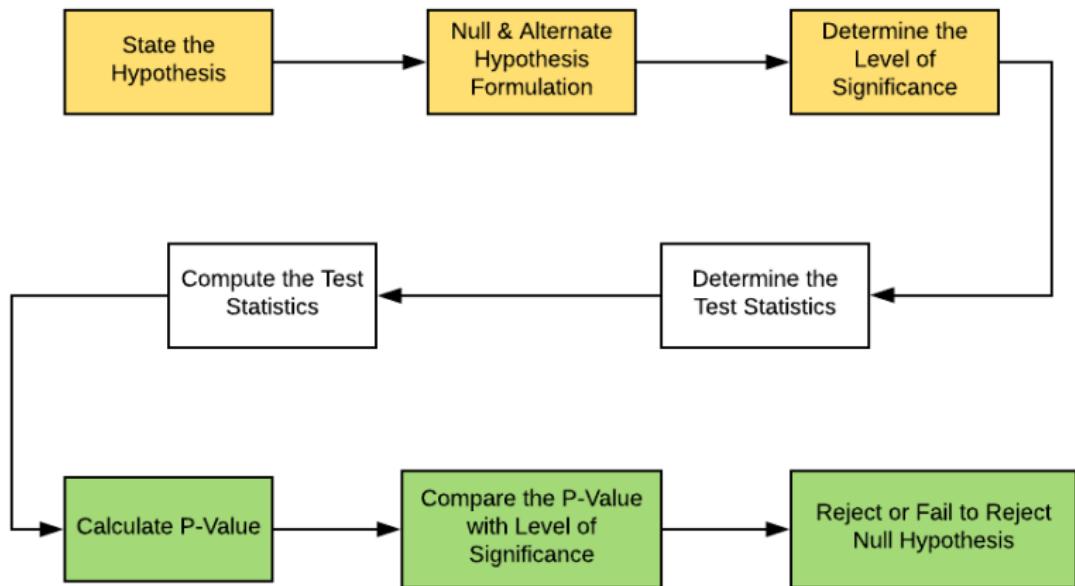
Есть *огромное* количество материалов по A/B-тестированию:

- Как устроено А/В-тестирование в Авито
- Как провести А/В-тестирование: 6 простых шагов
- It's All A/Bout Testing: The Netflix Experimentation Platform
- и всё, что Вы ещё нагуглите... :)

# Пайплайн

*A/B* тесты — частный случай тестирования гипотез.

Пайплайн по сути тот же.



# Повторим пайплайн

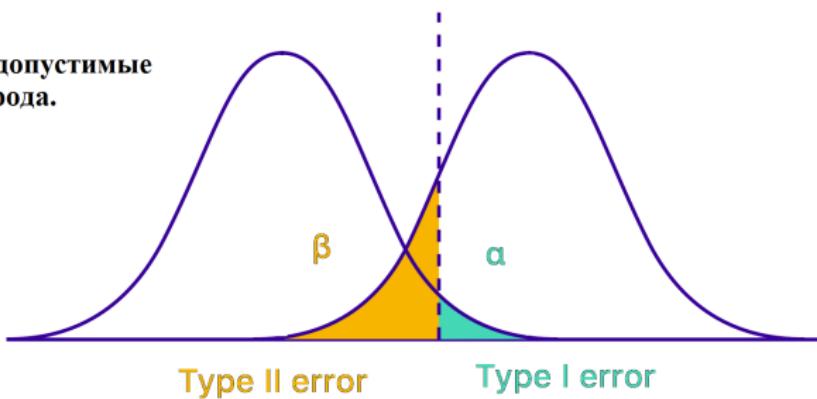
## Схематичный пайплайн

1. Фиксируем гипотезы:

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A \neq \mu_B$$

2. Фиксируем допустимые ошибки I и II рода.



3. Выбираем тест и считаем p-value

4. Если  $p\text{-value} < \alpha$ , то отклоняем  $H_0$

## Популярные тесты

Предполагаемое распределение	Примеры данных	Тест
Нормальное	Средняя выручка с пользователя (ARPU)	t-тест Стьюдента/Уэлча
Мультиномиальное	Количество различных купленных товаров	Критерий хи-квадрат
Неизвестно		Тест Манна-Уитни

Ещё можно строить доверительные интервалы при помощи бутстрепа.

Если данных мало, могут потребоваться тесты поточнее.

## Теория без практики мертва

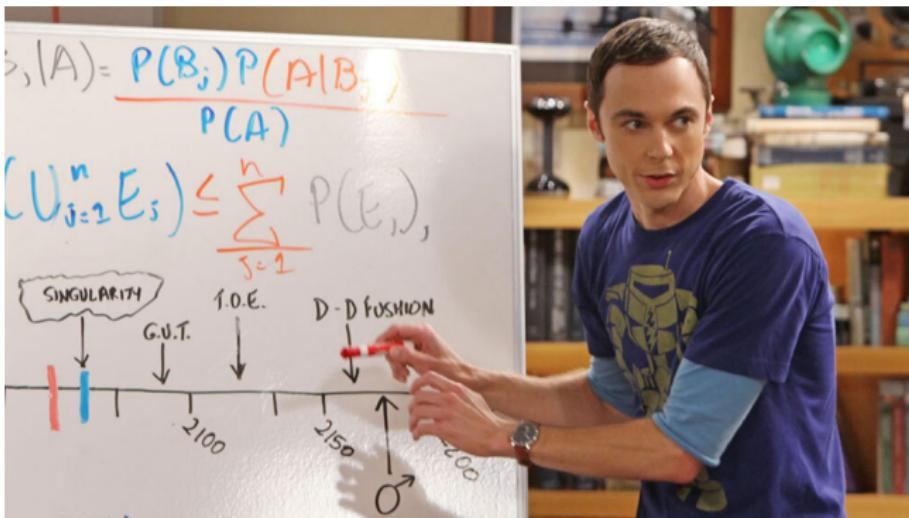
Мы не можем обнять необъятное.

Лучший способ разобраться в материале? Поработать аналитиком/DS)



# Кино и математика

Поговорим немного про Netflix и попробуем найти в этом математику)



## Статьи, которые обсудим

Мы обсудим интересные моменты из следующих статей:

- Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data
- Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix
- How Booking.com increases the power of online experiments with CUPED

## Похожий доклад

Доклад будет достаточно похож на:

<https://habr.com/ru/company/yandex/blog/497804/>

Хабр

Увеличение чувствительности А/Б-тестов  
с помощью Cupred. Доклад в Яндексе

Блог компании Яндекс

# Размер выборки

1 АБ тестирование

2 Размер выборки

3 Стратификация

4 CUPED

5 Вместо послесловия

## Хэширование

В крупных компаниях:

- огромный поток пользователей,
- одновременно проводится большое количество экспериментов.

## Повторим пайплайн

Возникают вопросы вида — как лучше и “покомпактней” захэшировать пользователей.

Как у нас устроено А/Б-тестирование. Лекция Яндекса



Рис.: Хэш - нарезал и перемешал

Естественно, хочется побольше экспериментов и побыстрее.

Для этого нужно уменьшить *минимальный размер выборки*.

На практике можно использовать калькуляторы типа [optimizely](#) или [abtasty](#).

# Калькуляторы

Вбил  
пару  
чисел.

Получил  
ответ.

A/B test sample size calculator

Powered by Intelligence Cloud's stats engine

Baseline Conversion Rate: 3%

Minimum Detectable Effect: 20%

Statistical Significance: 95% [Edit](#)

Your control group's expected conversion rate. [\[?\]](#)

The minimum relative change in conversion rate you would like to be able to detect. [\[?\]](#)

95% is an accepted standard for statistical significance, although Optimizely allows you to set your own threshold for significance based on your risk tolerance. [\[?\]](#)

Sample Size per Variation: 13,000

Рис.: Калькулятор для размера выборки

## Теорема

- Минимальный размер выборки, чтобы оценить **доли населения**:

$$n = \frac{\hat{p}(1 - \hat{p})z^2}{m^2}$$

- Минимальный размер выборки для оценки **среднего**:

$$n = \frac{\sigma^2 z^2}{m^2}$$

# Размер выборки

## Теорема

- Минимальный размер выборки, чтобы оценить **доли населения**:

$$n = \frac{\hat{p}(1 - \hat{p})z^2}{m^2}$$

- Минимальный размер выборки для оценки **среднего**:

$$n = \frac{\sigma^2 z^2}{m^2}$$

Мы строим  $1 - \alpha$  доверительный интервал (для  $p$  или  $\sigma$ ):

- $m$  — **предельная погрешность выборки** ( $= \frac{1}{2}$  длины доверительного интервала);
- $Z = Z_{\alpha/2}$  — квантиль нормального распределения.

## Размер выборки

Объясним, как получаются эти оценки.

**Доля населения.** Мы ищем параметр  $p$  для Bernoulli( $p$ ).

Строим доверительный интервал по ЦПТ:

$$\left( \hat{p} - z\sqrt{\frac{p(1-p)}{n}}, \quad \hat{p} + z\sqrt{\frac{p(1-p)}{n}} \right)$$

## Размер выборки

Объясним, как получаются эти оценки.

**Доля населения.** Мы ищем параметр  $p$  для Bernoulli( $p$ ).

Строим доверительный интервал по ЦПТ:

$$\left( \hat{p} - z\sqrt{\frac{p(1-p)}{n}}, \quad \hat{p} + z\sqrt{\frac{p(1-p)}{n}} \right)$$

Отсюда получаем требуемую формулу

$$z\sqrt{\frac{p(1-p)}{n}} = m \quad \Leftrightarrow \quad n = \frac{p(1-p)z^2}{m^2}.$$

## Размер выборки

**Пример.** Хотим в США измерить популярность президента. Хотим 95% доверительный интервал шириной не более, чем в 2 процентных пункта (0.02).



Население США 330 млн человек



51% Joe Biden

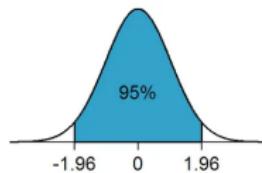
44% Donald Trump

Вероятность  $p = 0.5$



Предельная ошибка выборки  
= 0.5 доверительного интервала

$$m = 0.5 * 0.02 = 0.01$$



Квантиль  $z = 1.96$

**Ответ:** Для опроса нужно  $n = \frac{0,25 * (1,96)^2}{0,01^2} = 9604 \approx 10'000$  человек.

## Уменьшить дисперсию

**Наблюдение.** Минимальный размер выборки имеет вид

$$n = \text{Дисперсия} \cdot \frac{z^2}{m}.$$

Чтобы уменьшить выборку,  
нужно уменьшить дисперсию

# Уменьшить дисперсию

Способы уменьшить дисперсию:

- Удаление аномалий.
- Стратификация.
- CUPED.
- Применение ML моделей<sup>1</sup>.

Обсудим 2ой и 3ий способы.

<sup>1</sup> Помогают в других способах: искать выбросы и т.д.

# Стратификация

1 АБ тестирование

2 Размер выборки

3 Стратификация

4 CUPED

5 Вместо послесловия

# Стратификация

Стратификация уменьшает дисперсию.



Дисперсия всех котиков  
Большая



Дисперсия котиков дома  
Поменьше

## Стратификация

**Q:** Есть ли естественные разбиения на группы для пользователей Netflix?

## Стратификация

Q: Есть ли естественные разбиения на группы для пользователей Netflix?

A: Примеры признаков:



**Страна**

**Пол/Возраст**



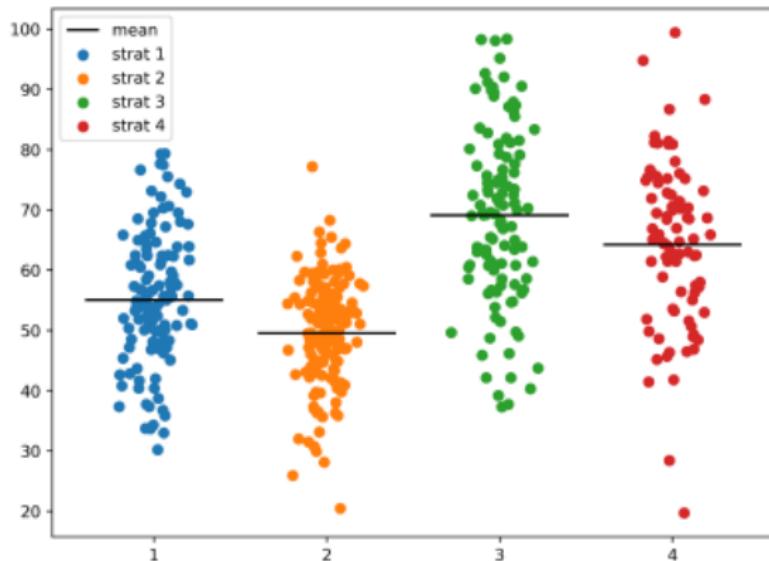
**Устройства**

Итак, Netflix.

- Есть **бизнес-метрика**  $Y$ .
- Известен набор **ковариатов**  $X$  — признаков, коррелирующих с метрикой  $Y$ .  
Они должны быть измеримы до эксперимента (пол, возраст и т.д.)

Мы разбиваем пользователей на  $K$  страт по значениям ковариат  $X$ .

В каждой страте берём среднее значение бизнес-метрики  $Y$ .



## Условное матожидание

Если занумеровать страты  $Z = 1, \dots, k$ , то мы рассматриваем условное матожидание

$$\mathbb{E}(Y | Z = k).$$

Это функция от  $k$ , которая обозначается через  $\mathbb{E}(Y | Z)$ .

Вспомним формулы для её матожидания и дисперсии.

## Law of total expectation

“Математическое ожидание убирает условие”.

### Закон полного математического ожидания

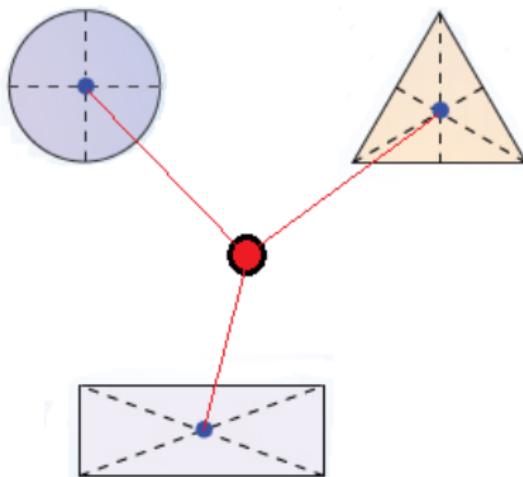
Если  $\mathbb{E}(X) < \infty$ , то для любой случайной величины  $Y$

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X | Y)).$$

Ещё английские названия: the [law of iterated expectations](#), the [tower rule](#), the [smoothing theorem](#).

## Law of total expectation

Наглядное объяснение формулы  $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X | Y))$ :



Центр масс системы - это центр масс усреднённых подсистем.

## Law of total expectation

Докажем теорему в случае, когда  $X$  и  $Y$  принимают конечное число значений.

$$\begin{aligned} E(E(X | Y)) &= E\left[\sum_x x \cdot P(X = x | Y)\right] \\ &= \sum_y \left[ \sum_x x \cdot P(X = x | Y = y) \right] \cdot P(Y = y) \\ &= \sum_y \sum_x x \cdot P(X = x, Y = y). \end{aligned}$$

## Law of total expectation

Докажем теорему в случае, когда  $X$  и  $Y$  принимают конечное число значений.

$$\begin{aligned} E(E(X | Y)) &= E\left[\sum_x x \cdot P(X = x | Y)\right] \\ &= \sum_y \left[ \sum_x x \cdot P(X = x | Y = y) \right] \cdot P(Y = y) \\ &= \sum_y \sum_x x \cdot P(X = x, Y = y). \end{aligned}$$

Остается переставить местами суммы:

$$\begin{aligned} \sum_x \sum_y x \cdot P(X = x, Y = y) &= \sum_x x \sum_y P(X = x, Y = y) \\ &= \sum_x x \cdot P(X = x) \\ &= E(X). \end{aligned}$$

## Литература

Доказательство в общем случае — см. большие книжки по Теории Вероятностей:



А. Н. Ширяев  
*Вероятность.*

## Закон полной дисперсии

### Закон полной дисперсии

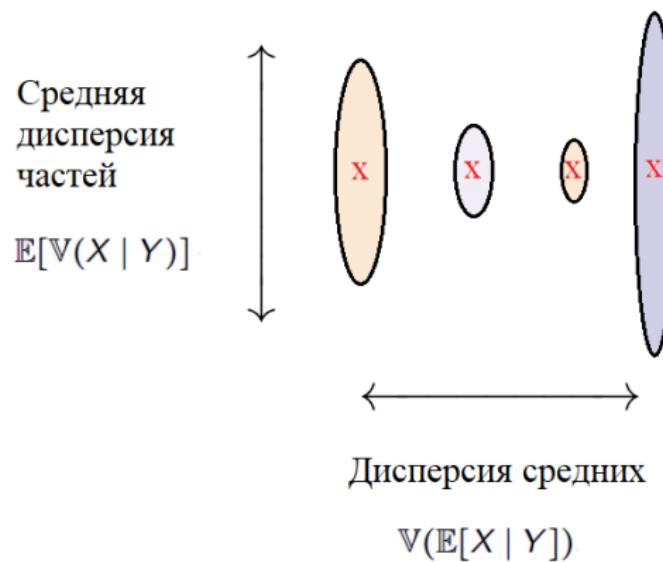
Если  $\mathbb{V}(X) < \infty$ , то для любой случайной величины  $Y$

$$\mathbb{V}(X) = \mathbb{E}[\mathbb{V}(X | Y)] + \mathbb{V}(\mathbb{E}[X | Y]).$$

## Закон полной дисперсии

Наглядное объяснение формулы. Объекты — множества вида  $Y = y$ .

Полная дисперсия = сумма дисперсии “по  $x$ ” и дисперсии “по  $y$ ”.



## Закон полной дисперсии

**Доказательство** закона полной дисперсии.

Честно расписываем дисперсию и применяем закон полного математического ожидания

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X | Y)).$$

## Закон полной дисперсии

**Доказательство** закона полной дисперсии.

Честно расписываем дисперсию и применяем закон полного математического  
матожидания

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X | Y)).$$

Получаем

$$\begin{aligned}\mathbb{E}[\mathbb{V}(X|Y)] &= \mathbb{E}[\mathbb{E}(X^2|Y) - (\mathbb{E}[X|Y])^2] = \mathbb{E}(X^2) - \mathbb{E}[(\mathbb{E}[X|Y])^2] \\ \mathbb{V}[\mathbb{E}(X|Y)] &= \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[\mathbb{E}(X|Y)])^2 = \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[X])^2\end{aligned}$$

## Закон полной дисперсии

**Доказательство** закона полной дисперсии.

Честно расписываем дисперсию и применяем закон полного математического  
матожидания

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X | Y)).$$

Получаем

$$\begin{aligned}\mathbb{E}[\mathbb{V}(X|Y)] &= \mathbb{E}[\mathbb{E}(X^2|Y) - (\mathbb{E}[X|Y])^2] = \mathbb{E}(X^2) - \mathbb{E}[(\mathbb{E}[X|Y])^2] \\ \mathbb{V}[\mathbb{E}(X|Y)] &= \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[\mathbb{E}(X|Y)])^2 = \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[X])^2\end{aligned}$$

Складывая, получаем требуемое

$$\mathbb{V}(X) = \mathbb{E}[\mathbb{V}(X | Y)] + \mathbb{V}(\mathbb{E}[X | Y]).$$

Перерыв

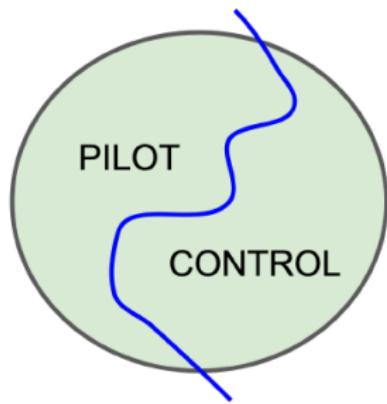
Перерыв

# Netflix стратификация

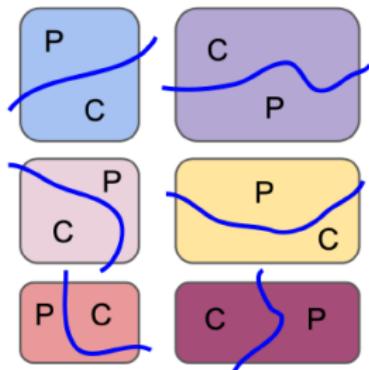
Закончим рассказ про стратификацию из



Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data



Случайное разбиение



Стратифицированное разбиение

## Сэмплирования

В статье Netflix сравниваются 2 сэмплирования:

- ① Случайное сэмплирование (simple random sampling).

Обозначения:  $Y_{srs}$ ,  $\mathbb{E}_{srs}$  и  $\mathbb{V}_{srs}$ .

## Сэмплирования

В статье Netflix сравниваются 2 сэмплирования:

- ① Случайное сэмплирование (simple random sampling).

Обозначения:  $Y_{srs}$ ,  $\mathbb{E}_{srs}$  и  $\mathbb{V}_{srs}$ .

- ② Стратифицированное сэмплирование.

При сэмплировании сохраняются доли страт  $p_k$ :

$$n_k = p_k n.$$

За бизнес-метрику берётся взвешенное среднее по стратам

$$Y_{strat} = \sum p_k \bar{Y}_k.$$

Обозначения:  $\mathbb{E}_{strat}$  и  $\mathbb{V}_{strat}$ .

## Матожидания совпали

Пусть  $\mu_k$  и  $\sigma_k^2$  — среднее и дисперсия по  $k$ -той страте.

Матожидания совпадают

$$\mathbb{E}_{srs} = \mathbb{E}_{strat} = \sum_{k=1}^K \mu_k.$$

Осталось сравнить дисперсии.

## Дисперсия уменьшилась

Стратифицированная дисперсия:

$$\mathbb{V}_{strat} = \frac{1}{n} \sum_{k=1}^K p_k \sigma_k^2.$$

## Дисперсия уменьшилась

Стратифицированная дисперсия:

$$\mathbb{V}_{strat} = \frac{1}{n} \sum_{k=1}^K p_k \sigma_k^2.$$

Для случайного сэмплирования формула

$$\mathbb{V}(X) = \mathbb{E}[\mathbb{V}(X | Y)] + \textcolor{red}{\mathbb{V}(\mathbb{E}[X | Y])}$$

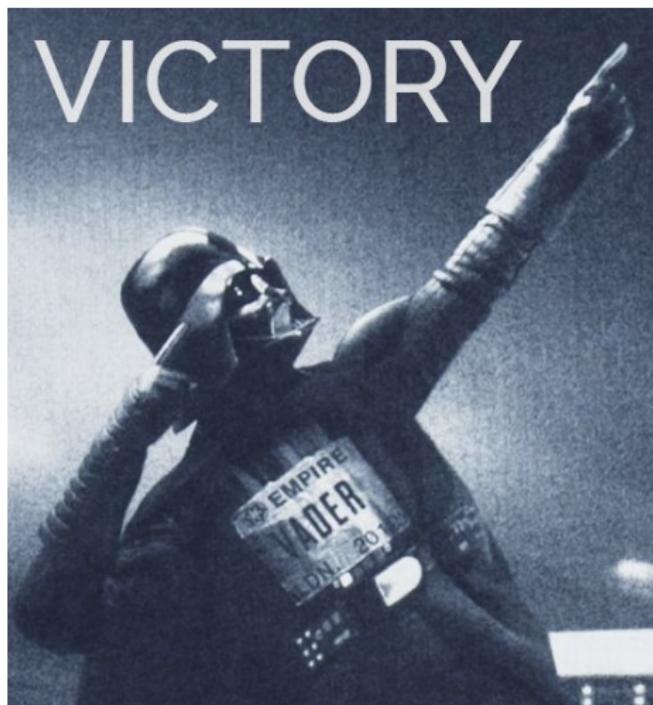
принимает вид

$$\mathbb{V}_{srs} = \frac{1}{n} \sum_{k=1}^K p_k \sigma_k^2 + \frac{1}{n} \sum_{k=1}^K p_k (\mu_k - \mu)^2.$$

Удалось **уменьшить дисперсию** — избавиться от дисперсии между стратами.

Show must go on

Успех! Но это не конец истории.



# CUPED

1 АБ тестирование

2 Размер выборки

3 Стратификация

4 CUPED

5 Вместо послесловия

## Вычитаем ковариат

Простая идея.

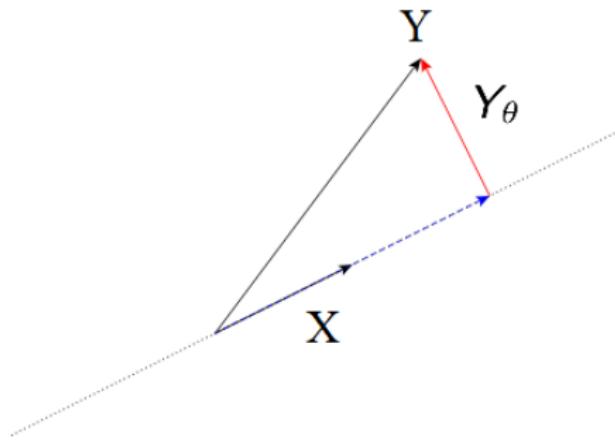
Мы хотим уменьшить дисперсию  $Y$ . Рассмотрим величину

$$Y_\theta = Y - \theta X.$$

Q: При каком  $\theta \in \mathbb{R}$  дисперсия  $\mathbb{V}Y_\theta$  минимальна?

## Проекция

Фактически мы проецируем  $Y$  на ортогональное подпространство к  $X$ :



## Уменьшение дисперсии

Формула для ковариации:

$$\mathbb{V}(Y_\theta) = \mathbb{V}(Y) - 2\theta \operatorname{cov}(X, Y) + \theta^2 \mathbb{V}(X).$$

## Уменьшение дисперсии

Формула для ковариации:

$$\mathbb{V}(Y_\theta) = \mathbb{V}(Y) - 2\theta \operatorname{cov}(X, Y) + \theta^2 \mathbb{V}(X).$$

Минимум достигается при

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

и он равен

$$(1 - \rho^2)\mathbb{V}(Y), \quad \rho = \frac{\operatorname{cov}(X, Y)}{\sqrt{\mathbb{V}(X) \mathbb{V}(Y)}}.$$

**Дисперсия уменьшилась.**

## Лучший ковариат

*Замечание.* Доведём до абсурда:

- ② Q: Какая величина лучше всех коррелирует с  $Y$ ?

## Лучший ковариат

Замечание. Доведём до абсурда:

② Q: Какая величина лучше всех коррелирует с  $Y$ ?

A:  $X = Y$ . Данные исчезают:

$$Y_\theta = 0.$$

Не то, чего мы хотим.

$X$  должен быть **коррелирован**, но не быть в причинной зависимости от  $Y$ .

## Самый близкий человек

Загадка: кто был с тобой всю жизнь и знает всё, что с тобою было?

## Самый близкий человек

Загадка: кто был с тобой всю жизнь и знает всё, что с тобою было?



Ты из прошлого.

# CUPED

Controlled-experiment Using Pre-Experiment Data

В качестве  $X$  берётся значение метрики за предыдущий период.

## Идея CUPED

### Идея CUPED.

Часть дисперсии  $Y$  вызвана постоянными факторами, которые также влияли на прошлое  $X$ .

Беря  $Y_{CUPED} = Y - \theta X$ , мы избавляемся от влияния этих факторов.

## Идея CUPED

### Идея CUPED.

Часть дисперсии  $Y$  вызвана постоянными факторами, которые также влияли на прошлое  $X$ .

Беря  $Y_{CUPED} = Y - \theta X$ , мы избавляемся от влияния этих факторов.

### Суть CUPED

Нас интересует не то, как в среднем ведут себя пользователи,  
а то, как изменилось их поведение.

## Регрессия

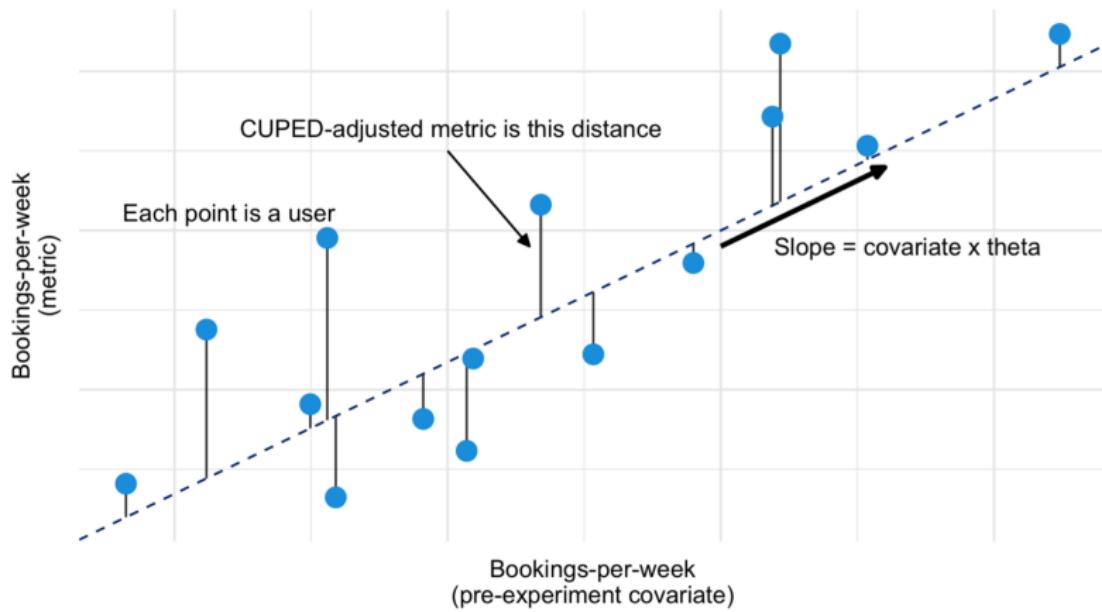
Q: Нам знакомо представление  $Y = \theta X + \varepsilon$ ?

## Регрессия

Q: Нам знакомо представление  $Y = \theta X + \varepsilon$ ?

A: Да, подбор параметра  $\theta$  — это регрессия (через начало координат)

Мы убираем корреляцию и оставляем изменения-остатки.



Просто добавь ML

Последний штрих — добавим в лекцию немного ML.

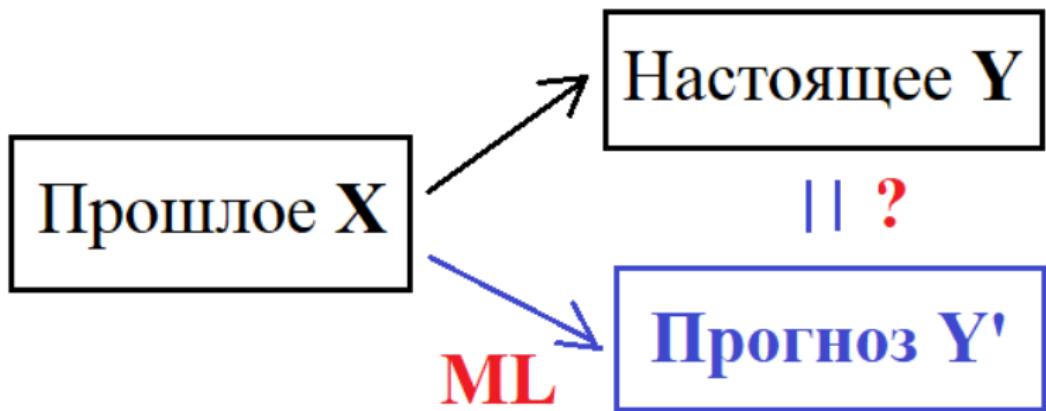


## Не живи в прошлом

Почему мы “живём прошлым”?

ML модели, работающие с *временными рядами*,  
учатся “предсказывать” будущее.

Почему бы не оставить в  $Y$  только те факторы,  
которые мы не смогли предсказать по прошлому  $X$ ?



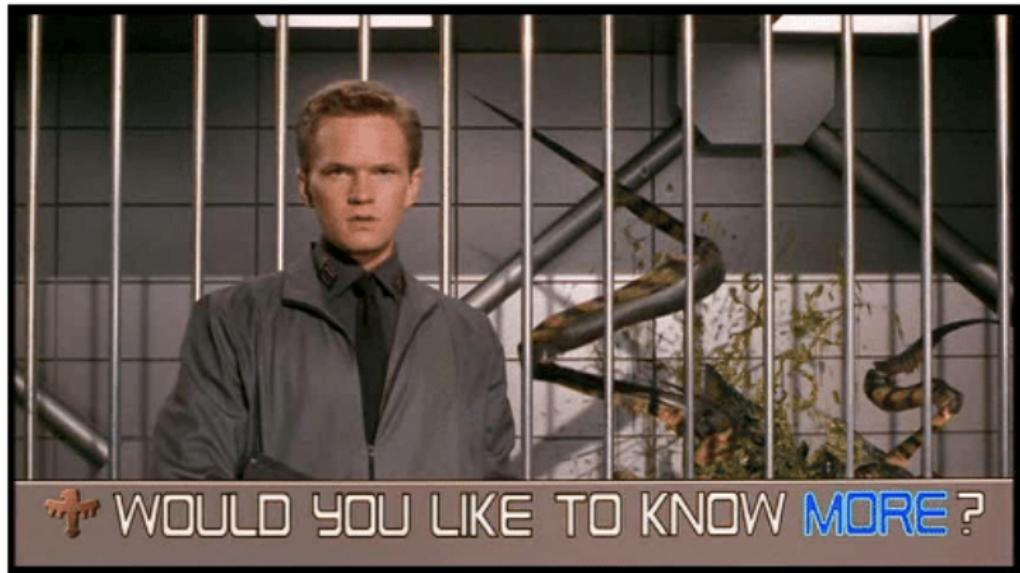
# CUPAC

Control Using Predictions as Covariate

В качестве ковариаты можно взять **предсказание ML модели** на исторических данных.

Больше ML

**Хочешь узнать про ML-модели больше?**



WOULD YOU LIKE TO KNOW MORE?

## Снова в Школу

Для начала полезно пройти курс по ML, например в



ШКОЛА АНАЛИЗА ДАННЫХ

## Вместо послесловия

1 АБ тестирование

2 Размер выборки

3 Стратификация

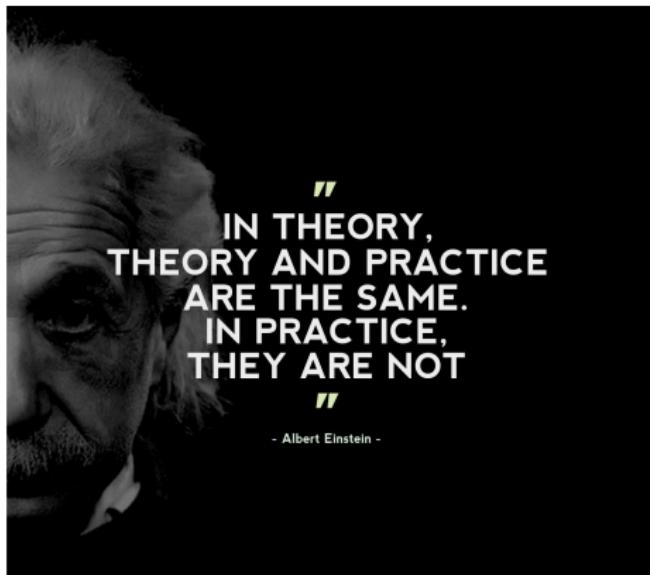
4 CUPED

5 Вместо послесловия

Лучше 1 раз увидеть, что 100 раз услышать

В заключение — пара замечаний.

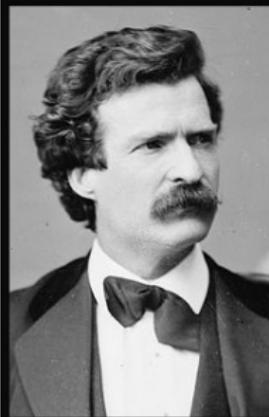
На практике Вы узнаете гораздо больше)



## 3 kinds of lies

Data-driven development - не панацея.

Никто не гарантирует, что гипотеза верна, и что дизайн эксперимента не был “подогнан под ответ”.



There are three kinds of lies — lies, damned lies and statistics.

(Mark Twain)

# Как сломать тест

## Как сломать тест:

ИСТОРИИ

**В телеграме проводят «народные исследования» российской вакцины от коронавируса** Мы спросили у организаторов исследования, зачем они это делают — и что узнали об эффективности вакцины

17:30, 18 ноября 2020 · Источник: Meduza



Наталья Колесникова / AFP / Scanpix / LETA

Никогда так не делайте!!!

Спаси котёнка!

АБ-тест — как котёнок Шрёдингера.

Нарушите правила эксперимента — котёнок умрёт!



Рис.: Не убивайте котёнка!