

УДК 004.85
ББК 32.971.3
К75

Рон Кохави, Диана Тан, Я Сюй

К75 Доверительное А/В-тестирование. Практическое руководство по контролируемым экспериментам / пер. с англ. В. С. Яценкова. – М.: ДМК Пресс, 2021. – 298 с.: ил.

ISBN 978-5-97060-913-2

Сложно понять ценность идеи, пока она не опробована на практике. В этой книге рассказывается о том, как контролируемые онлайн-эксперименты (или, как их еще называют, А/В-тесты) позволяют оценить эффективность тех или иных идей по оптимизации веб-сайтов и добиться максимальной отдачи от их использования. Вы узнаете, как правильно подобрать инструменты для тестирования, провести сбор данных и обеспечить измеримость результатов. На конкретных примерах показано, как при помощи А/В-тестов были улучшены веб-ресурсы известных компаний.

Контролируемые онлайн-эксперименты широко применяются в Amazon, Booking.com, eBay, Facebook, Google, LinkedIn, Microsoft, Twitter, Яндекс и других компаниях. Эта методика становится неотъемлемой частью культуры бизнеса, основанной на данных.

Издание адресовано техническим специалистам и менеджерам, заинтересованным в увеличении прибыльности своих онлайн-проектов.

УДК 004.85
ББК 32.971.3

Copyright Original English language edition published by Cambridge University Press is part of the University of Cambridge. Copyright © 2020 by Ron Kohavi, Diane Tang, Ya Xu. Russian-language edition copyright © 2021 by DMK Press. All rights reserved.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

ISBN (анг.) 978-1-108-72426-5
ISBN (рус.) 978-5-97060-913-2

© 2020 Ron Kohavi, Diane Tang, and Ya Xu
© Оформление, издание, перевод,
ДМК Пресс, 2021

Оглавление

Отзывы рецензентов о книге	5
Предисловие от издательства	18
Вступление	19
Предисловие	21
Благодарности	23
ЧАСТЬ I. ВВЕДЕНИЕ ДЛЯ ВСЕХ	25
Глава 1. Введение и мотивация	27
1.1. Терминология контролируемых онлайн-экспериментов.....	29
1.2. Зачем нужны эксперименты? Корреляции, причинно-следственная связь и достоверность.....	33
1.3. Необходимые ингредиенты для проведения эффективных контролируемых экспериментов.....	35
1.4. Постулаты	36
1.5. Постепенные улучшения	39
1.6. Примеры интересных контролируемых онлайн-экспериментов	41
1.7. Стратегия, тактика и их связь с экспериментами.....	46
1.8. Дополнительное чтение	50
Глава 2. Проведение и анализ экспериментов. Пример полного цикла	52
2.1. Условия демонстрационного эксперимента	52
2.2. Проверка гипотез: установление статистической значимости.....	56
2.3. Разработка эксперимента	58
2.4. Проведение эксперимента и сбор данных	61
2.5. Интерпретация результатов.....	61
2.6. От результатов к решениям	63
Глава 3. Закон Тваймана и надежность экспериментов	66
3.1. Неправильная интерпретация статистических результатов.....	67
3.1.1. Нехватка статистической мощности	67
3.1.2. Неправильная интерпретация p -значений.....	67
3.1.3. Отслеживание p -значений	69
3.1.4. Множественные проверки гипотез.....	69
3.2. Достоверные интервалы.....	70

3.3. Угрозы внутренней достоверности.....	70
3.3.1. Нарушения правила SUTVA	70
3.2.2. Ошибка выжившего	71
3.2.3. Вынужденное воздействие.....	71
3.2.4. Несоответствие коэффициента выборки	72
3.4. Угрозы внешней достоверности	76
3.4.1. Эффекты первичности.....	76
3.4.2. Эффекты новизны.....	76
3.4.3. Выявление эффектов первичности и новизны.....	78
3.5. Разделение по сегментам	78
3.5.1. Сегментированное представление показателя	79
3.5.2. Сегментированное представление эффекта (гетерогенность эффекта).....	80
3.5.3. Анализ эффекта по сегментам, вводящий в заблуждение.....	81
3.6. Парадокс Симпсона	82
3.7. Поощряйте здоровый скептицизм.....	84

Глава 4. Платформы и культура экспериментов 85

4.1. Модели зрелости экспериментов.....	85
4.1.1. Лидерство	87
4.1.2. Процесс	88
4.1.3. Разработать самим или купить готовый продукт?	91
4.2. Инфраструктура и инструменты.....	94
4.2.1. Разработка, настройка и управление экспериментом	96
4.2.2. Развертывание эксперимента.....	97
4.2.3. Инструменты для экспериментов.....	100
4.2.4. Масштабирование экспериментов: тонкости назначения вариантов	101
4.2.5. Параллельные эксперименты	103
4.2.6. Анализ экспериментов	105

ЧАСТЬ II. ИЗБРАННЫЕ ТЕМЫ ДЛЯ ВСЕХ 107

Глава 5. Скорость имеет значение!..... 111

5.1. Ключевое предположение: локальная линейная аппроксимация	113
5.2. Как измерить быстродействие веб-сайта.....	114
5.3. Схема эксперимента по замедлению	116
5.4. Влияние различных элементов страницы	118
5.5. Экстремальные результаты	119

Глава 6. Организационные показатели 121

6.1. Таксономия показателей	121
6.2. Выработка показателей: принципы и методы	125
6.3. Оценка показателей.....	128

6.4. Развивающиеся показатели	129
6.5. Дополнительное чтение	130
6.6. Примечание: ограничительные показатели	130
6.7. Примечание: преднамеренная манипуляция показателями.....	132

Глава 7. Показатели экспериментов и общий критерий оценки..... 135

7.1. От бизнес-показателей к показателям, подходящим для экспериментов.....	136
7.2. Объединение ключевых показателей в ОЕС	138
7.3. Пример: ОЕС для электронной почты на Amazon	140
7.4. Пример: ОЕС для поисковой системы Bing.	141
7.5. Закон Гудхарта, закон Кэмпбелла и замечание Лукаса	143

Глава 8. Институциональная память и метаанализ..... 145

8.1. Что такое институциональная память?.....	145
8.2. Почему полезна институциональная память?.....	146

Глава 9. Этика контролируемых экспериментов 150

9.1. Что лежит в основе этики	150
9.1.1. Риски	152
9.1.2. Преимущества и выгоды	153
9.1.3. Возможность выбора	155
9.2. Сбор данных	155
9.3. Культура и процессы	156
9.4. Примечание: идентификация пользователей	157

ЧАСТЬ III. ДОПОЛНИТЕЛЬНЫЕ И АЛЬТЕРНАТИВНЫЕ МЕТОДЫ КОНТРОЛИРУЕМЫХ ЭКСПЕРИМЕНТОВ..... 159

Глава 10. Дополнительные методы..... 163

10.1. Пространство дополнительных методов.....	163
10.2. Анализ на основе журналов	164
10.3. Экспертная оценка.....	166
10.4. Исследование пользовательского опыта.....	167
10.5. Фокус-группы	168
10.6. Обзоры	169
10.7. Внешние данные.....	170
10.8. Подведем итог главы.....	172

Глава 11. Наблюдательные исследования причинно-следственных связей 174

11.1. Когда контролируемые эксперименты невозможны	174
--	-----

11.2. Планы для наблюдательных исследований причинно-следственных связей	176
11.2.1. Прерывистый временной ряд	176
11.2.2. Эксперименты с чередованием	178
11.2.3. Метод разрывной регрессии	178
11.2.4. Инструментальные переменные и естественные эксперименты.....	180
11.2.5. Отбор подобного по склонности.....	180
11.2.6. Дифференциальная разница.....	181
11.3. Ловушки причинно-следственных связей	182
11.4. Приложение: опровергнутые исследования причинно-следственных связей	185

ЧАСТЬ IV. ПЛАТФОРМЫ ДЛЯ ЭКСПЕРИМЕНТОВ: УГЛУБЛЕННОЕ ИЗУЧЕНИЕ.....

189

Глава 12. Эксперименты на стороне клиента.....

193

12.1. Различия между серверной и клиентской стороной.....	193
12.1.1. Отличие №1: процесс выпуска.....	194
12.1.2. Отличие №2: обмен данными между клиентом и сервером	195
12.2. Следствия из компромиссов	197
12.3. Выводы.....	201

Глава 13. Инструментарий экспериментов.....

202

13.1. Инструменты на стороне клиента и сервера	202
13.2. Обработка журналов из нескольких источников.....	204
13.3. Культура измерений.....	205

Глава 14. Выбор единицы рандомизации.....

206

14.1. Единица рандомизации и единица анализа	208
14.1 Рандомизация на уровне пользователя	209

Глава 15. Развитие эксперимента: компромисс между скоростью, качеством и риском.

212

15.1. Что такое рампинг?.....	212
15.2. Шаблон SQR для рампинга	213
15.3. Четыре фазы рампинга.....	214
15.3.1. Первая фаза рампинга: до MPR.....	215
15.3.2. Вторая фаза рампинга: MPR.....	216
15.3.3. Третья фаза рампинга: пост-MPR	216
15.3.4. Четвертая фаза рампинга: длительное удержание или репликация.....	216
15.4. Что после рампинга?.....	218

Глава 16. Анализ масштабных экспериментов.....	219
16.1. Подготовка данных	219
16.2. Вычисление данных	220
16.3. Формирование сводки и визуализация результатов.....	222

ЧАСТЬ V. РАЗВЕРНУТОЕ ОПИСАНИЕ АНАЛИЗА ЭКСПЕРИМЕНТОВ225

Глава 17. Статистика контролируемых онлайн-экспериментов.....	229
17.1. Двухвыборочный t -тест	229
17.2. p -значение и доверительный интервал	230
17.3. Предположение о нормальности.....	231
17.4. Ошибки типа I/II и статистическая мощность	233
17.5. Смещение.....	235
17.6. Множественное тестирование.....	235
17.7. Метаанализ Фишера	236

Глава 18. Оценка дисперсии и повышение чувствительности: подводные камни и решения	238
18.1. Распространенные ошибки	239
18.1.1. Дельта или процентная дельта?	239
18.1.2. Показатели отношения: когда уровень анализа отличается от уровня эксперимента	239
18.1.3. Выбросы.....	241
18.2. Повышение чувствительности	242
18.3. Дисперсия других статистических данных	244

Глава 19. А/А-тестирование.....	246
19.1. Почему нужны А/А-тесты?	246
19.1.1. Пример 1: уровень анализа отличается от уровня рандомизации	247
19.1.2. Пример 2: поощрение остановки эксперимента при достижении статистической значимости	249
19.1.3. Пример 3: переадресация браузера	249
19.1.4. Пример 4: неравное распределение по группам	250
19.1.5. Пример 5: различия в оборудовании.....	251
19.2. Как проводить А/А тесты	251
19.3. Когда А/А-тест не подходит	252

Глава 20. Включение по условию для повышения чувствительности	254
20.1. Примеры включения по условию.....	254

20.1.1. Пример 1: преднамеренно частичное воздействие	255
20.1.2. Пример 2: условное воздействие	255
20.1.3. Пример 3: Увеличение охвата	256
20.1.4. Пример 4: изменение покрытия	256
20.1.5. Пример 5: контрфактическое включение для моделей машинного обучения	257
20.2. Числовой пример	258
20.3. Оптимальное и консервативное включение	258
20.4. Общий эффект воздействия	259
20.5. Достоверность включения	261
20.6. Распространенные ошибки	261
20.7. Открытые вопросы	263

Глава 21. Несоответствие коэффициента выборки и другие ограничительные показатели

21.1. Несоответствие коэффициента выборки (SRM)	264
21.2. Причины возникновения SRM	266
21.3. Устранение SRM	268
21.4. Другие ограничительные показатели, связанные с доверием	269

Глава 22. Утечка и интерференция между вариантами

22.1. Примеры	272
22.2. Некоторые практические решения	275
22.2.1. Полезное правило: ценность действия в экосистеме	276
22.2.2. Изоляция	277
22.2.3. Анализ на уровне ребер графа	279
22.2.4. Обнаружение и мониторинг взаимовлияния	280

Глава 23. Измерение долгосрочных эффектов

23.1. Что такое долгосрочные эффекты?	281
23.2. Причины, по которым могут различаться краткосрочные и долгосрочные эффекты	282
23.4. Зачем измерять долгосрочные эффекты?	284
23.5. Длительные эксперименты	285
23.6. Альтернативные методы для длительных экспериментов	288
23.6.1. Метод №1: когортный анализ	288
23.6.2. Метод № 2: постпериодный анализ	288
23.6.3. Метод №3: воздействие с интервалом во времени	290
23.6.4. Метод №4: сдерживание и обратный эксперимент	292

Предметный указатель