


АБ-ТЕСТИРОВАНИЕ

Байдина Ксения, Яндекс Такси

Обо мне



Яндекс, аналитик

 @kbaidina

 kseniia.baidina@gmail.com

■ 2013-2017

Бакалавриат *НИУ ВШЭ* |
Факультет Менеджмента

■ 2017-2019

Магистратура *Университета Кёльна, Германия* |
Факультет Бизнес администрирования |
Специализация Статистика и эконометрика

■ 2018-2019

Университет Кельна, Германия |
Научный сотрудник

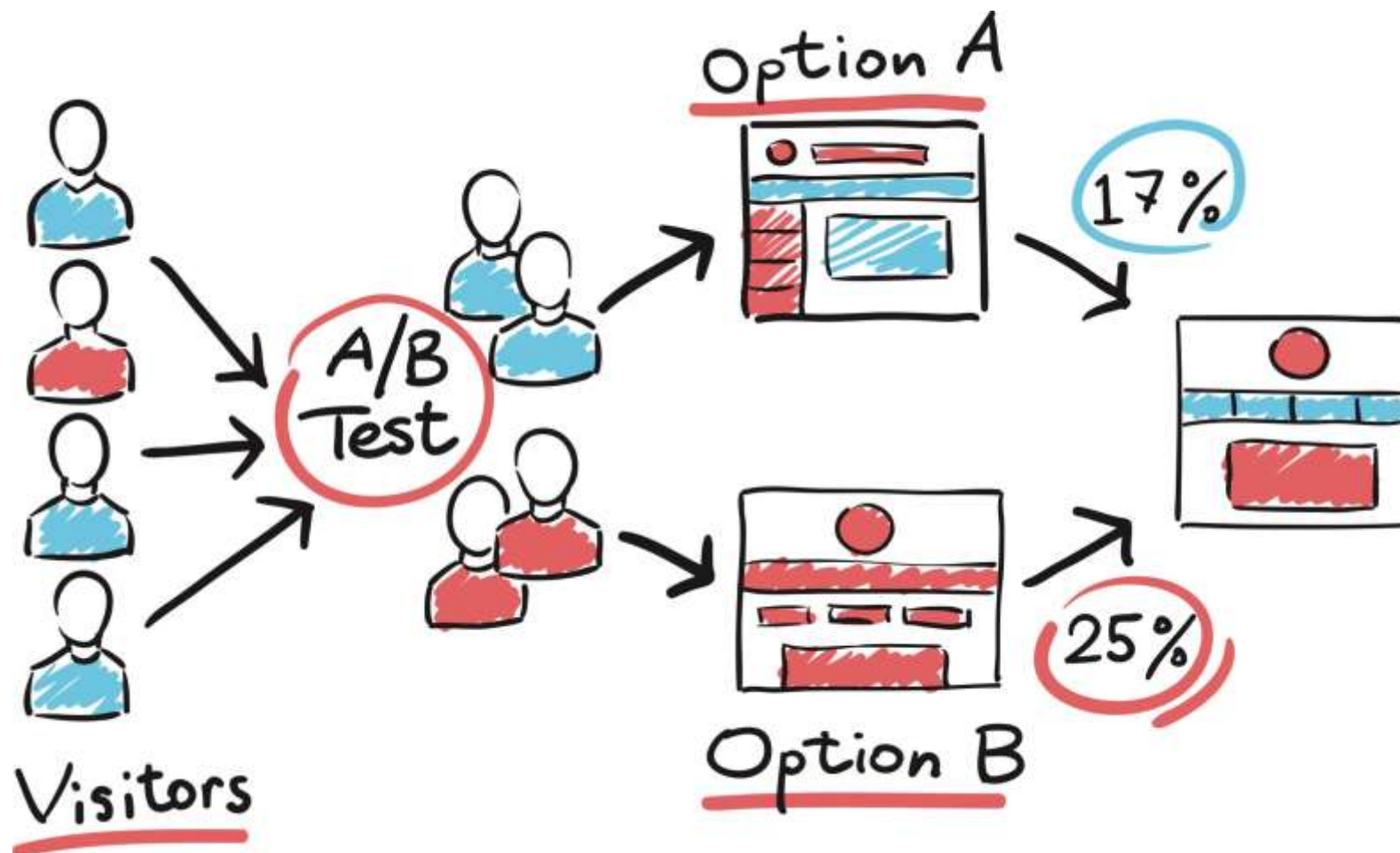
■ 2019

ОВИ |
Дата сайентист

■ С 2019

Яндекс |
Аналитик

Что такое АБ-тест



Зачем нужен АБ-тест?

Алгоритм

- Сформулировать гипотезу – что хотим проверить
- Выбрать метрики (*абсолютные метрики и конверсии*)
- Рассчитать необходимый размер выборки
- Разбить пользователей группы
- Каждой группе дать разные варианты тритмента
- Проанализировать результаты

Алгоритм

- Сформулировать гипотезу – что хотим проверить
- Выбрать метрики (*абсолютные метрики и конверсии*)
- Рассчитать необходимый размер выборки
- Разбить пользователей на группы
- Каждой группе дать разные варианты тритмента
- Проанализировать результаты

Разбить пользователей на группы

А что сложного?

Вопросы

- Как проводить несколько экспериментов?
- Как можно детектировать, что разбиение плохое?

АА-тестирование

Помогает проверить:

- Качество разбиения
- Алгоритмы расчета результатов

Виды разбиения

- По визитам
- По пользователям

Проанализировать
результаты

✓ 1 item added to Cart



Fire TV Stick with Alexa Voice Remote | Streaming...

\$29.99

Quantity added: 1

☐ This is a gift
Why is this important?

Order subtotal: **\$29.99**

1 item in your Cart

Edit your Cart

Proceed to checkout

Add \$5.01 of eligible items to your order to qualify for **FREE Shipping**. (Some restrictions apply)



Get a **\$50 Amazon.com Gift Card** instantly upon approval for the Amazon Rewards Visa Card

Current Total: \$ 29.99

Savings: **- \$ 50.00**

Cost After Savings: **\$ 0.00**

Savings Remaining: **\$ 20.01**

Apply now

Amazon.com \$25 Gift Card in...

★★★★☆ (14)

\$25.00

Add to Cart

fire tv stick
Protection Plan

provided by SquareTrade

2-Year Protection Plan for Amazon...

★★★★☆ (1,557)

\$4.99

Add to Cart

fire tv stick
Protection Plan

provided by SquareTrade

3-Year Protection Plan for Amazon...

★★★★☆ (1,557)

\$6.99

Add to Cart

Mission Cables USB Power Cable...

★★★★☆ (418)

\$18.99

Add to Cart

Nupro Travel Case for Fire...

★★★★☆ (366)

\$12.99

Add to Cart



Добавили рекомендации

- Средний чек старой версии – \$34.5
- Средний чек старой версии – \$37.2

Как могут выглядеть распределения

Статистические гипотезы

H₀:

- Утверждение о параметрах генеральных совокупностей, которые необходимо проверить
- *«Средний чек в двух группах не отличается»*

H₁:

- Утверждение, противоположное нулевой гипотезе
- *«Средний чек в двух группах отличается»*

Ошибки I и II рода

Ошибка I рода:

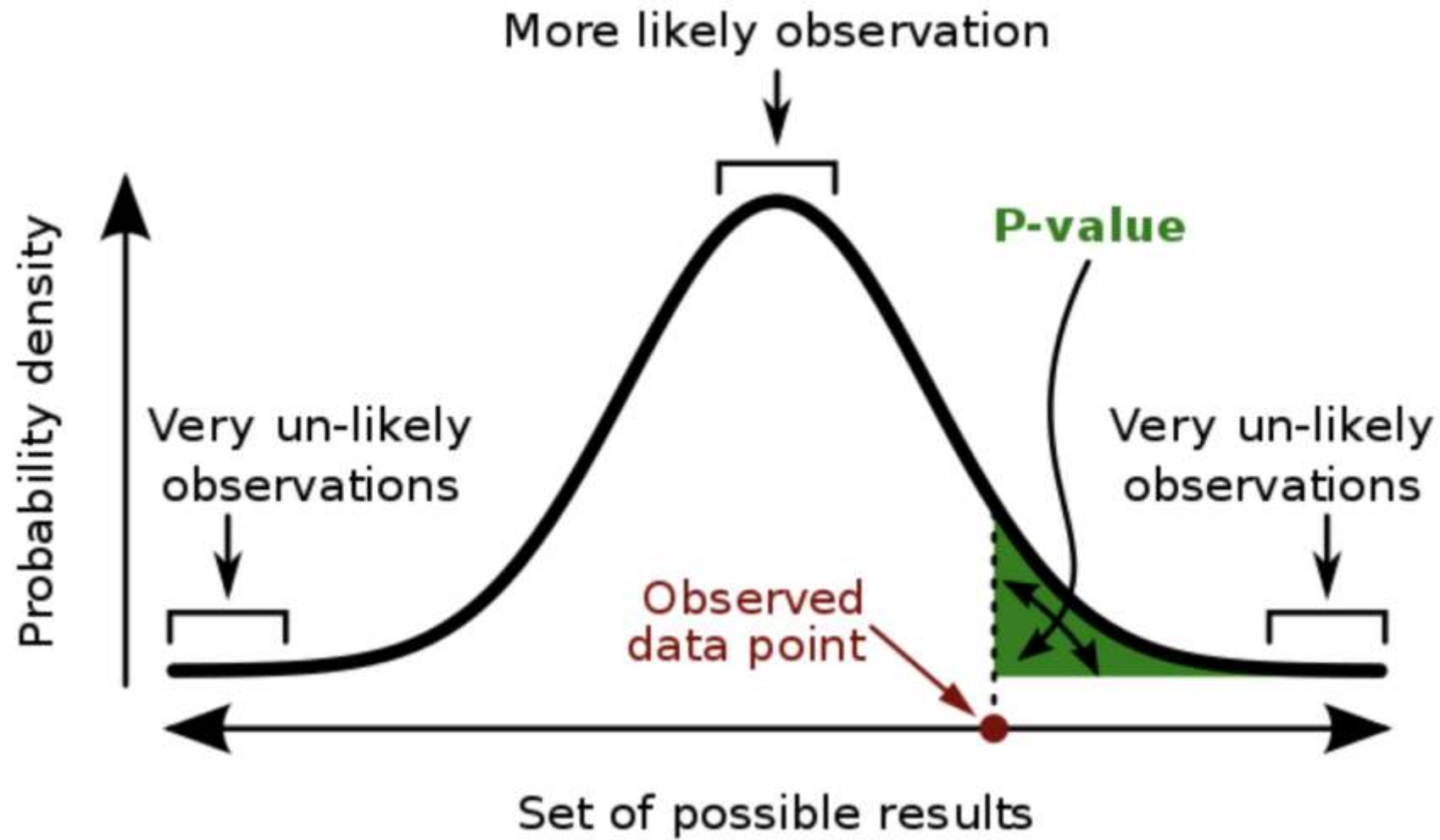
- Вероятность отвергнуть H_0 , когда она на самом деле верна (α).

Ошибка II рода:

- Вероятность не отвергнуть H_0 , когда она на самом деле не верна (β).

Определения

- *Статистический критерий* — правило, которое позволяет делать вывод о том, стоит ли на основе имеющихся данных отвергать нулевую гипотезу или нет.
- *P-значение (P-value)* — вероятность ошибки, при условии, что нулевая гипотеза верна.



Статистические критерии

- Для метрик-конверсий:
 - *Z-тест на равенство долей*
- Для абсолютных метрик:
 - *T-тест, критерий Манна-Уитни*

Z-тест на равенство долей

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1 - \widehat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

- p_1 – конверсия в первой группе
- p_2 – конверсия во второй группе
- p – конверсия для всех наблюдений
- n_1, n_2 – количество наблюдений в первой/второй группе

T-тест на равенство средних

$$s^2 = \frac{\sum_{t=1}^n (X_t - \bar{X})^2}{n - 1}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- $H_0: \bar{x}_1 = \bar{x}_2$
- $|t| > t_{кр} - H_0$ отклоняется

U-критерий Манна-Уитни

- $H_0: P(X < Y) = 0.5$
- $U > U_{\text{кр}} - H_0$ не отклоняется

U-критерий Манна-Уитни

Алгоритм:

1. Собираем наблюдения в одну выборку $n = n_1 + n_2$
2. Ранжируем наблюдения, считаем сумму рангов отдельно для каждой группы: R_x и R_y .
3. Считаем статистики:

$$U_x = R_x - \frac{n_1(n_1 + 1)}{2}$$

$$U_y = R_y - \frac{n_2(n_2 + 1)}{2}$$

$$U = \min(U_x, U_y)$$

Table 3 Critical values of U (5% significance).

$n_1 \backslash n_2$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2								0	0	0	0	1	1	1	1	1	2	2	2	2
3					0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4				0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5			0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7			1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8		0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9		0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10		0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11		0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12		1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13		1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14		1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15		1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16		1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17		2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18		2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19		2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20		2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Что вы помните?

1. Как проверить качество разбиения пользователей на группы?

Что вы помните?

1. Как проверить качество разбиения пользователей на группы?
2. Какой стат. тест используется для долей?

Что вы помните?

1. Как проверить качество разбиения пользователей на группы?
2. Какой стат. тест используется для долей?
3. Когда применять Т-тест, а когда тест Манна-Уитни?

Размер выборки

Размер выборки

- Для метрик-конверсий – калькулятор размера выборки

<https://www.evanmiller.org/ab-testing/sample-size.html>

- Для абсолютных метрик из формулы Т-теста

$$t = \frac{\overline{x_1} - \overline{x_2}}{S_x \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Множественное тестирование

Чем плохо множественное тестирование?

Множественное тестирование: поправка Бонферрони

$$\alpha^* = \alpha / m$$

- α – первоначальный уровень значимости
- m – количество проверяемых гипотез

Множественное тестирование: поправка Холма-Бонферрони

- Отранжировать p-value в порядке возрастания

$$\alpha^* = \alpha / (m + 1 - k)$$

- α – первоначальный уровень значимости
- m – количество проверяемых гипотез
- k – ранг

Усложнения

- Нарушение независимости
 - *Network-эффекты*
 - *Единица анализа \neq единица рандомизации*

Усложнения

- Нарушение независимости
 - *Network-эффекты*
 - *Единица анализа \neq единица рандомизации*
- Продвинутый анализ экспериментов
 - *Анализ по бакетам*
 - *Повышение чувствительности*

Усложнения

- Нарушение независимости
 - *Network-эффекты*
 - *Единица анализа \neq единица рандомизации*
- Продвинутый анализ экспериментов
 - *Анализ по бакетам*
 - *Повышение чувствительности*
- АБ без АБ
 - *Difference in difference*
 - *Causal impact*

Поездки в Такси

- Маркетплейс такси в городе: водители и пассажиры
- У водителей разный приоритет, который зависит от поведения водителя в сервисе
- Результат аналитики: Чем выше у водителя приоритет, тем выше у него метрики в системе (поездки, заработок, рейтинг и пр.)
- Гипотеза: если увеличим приоритет ВСЕМ водителям – сделаем систему эффективнее
- Что покажет АБ-тест?

Каннибализация

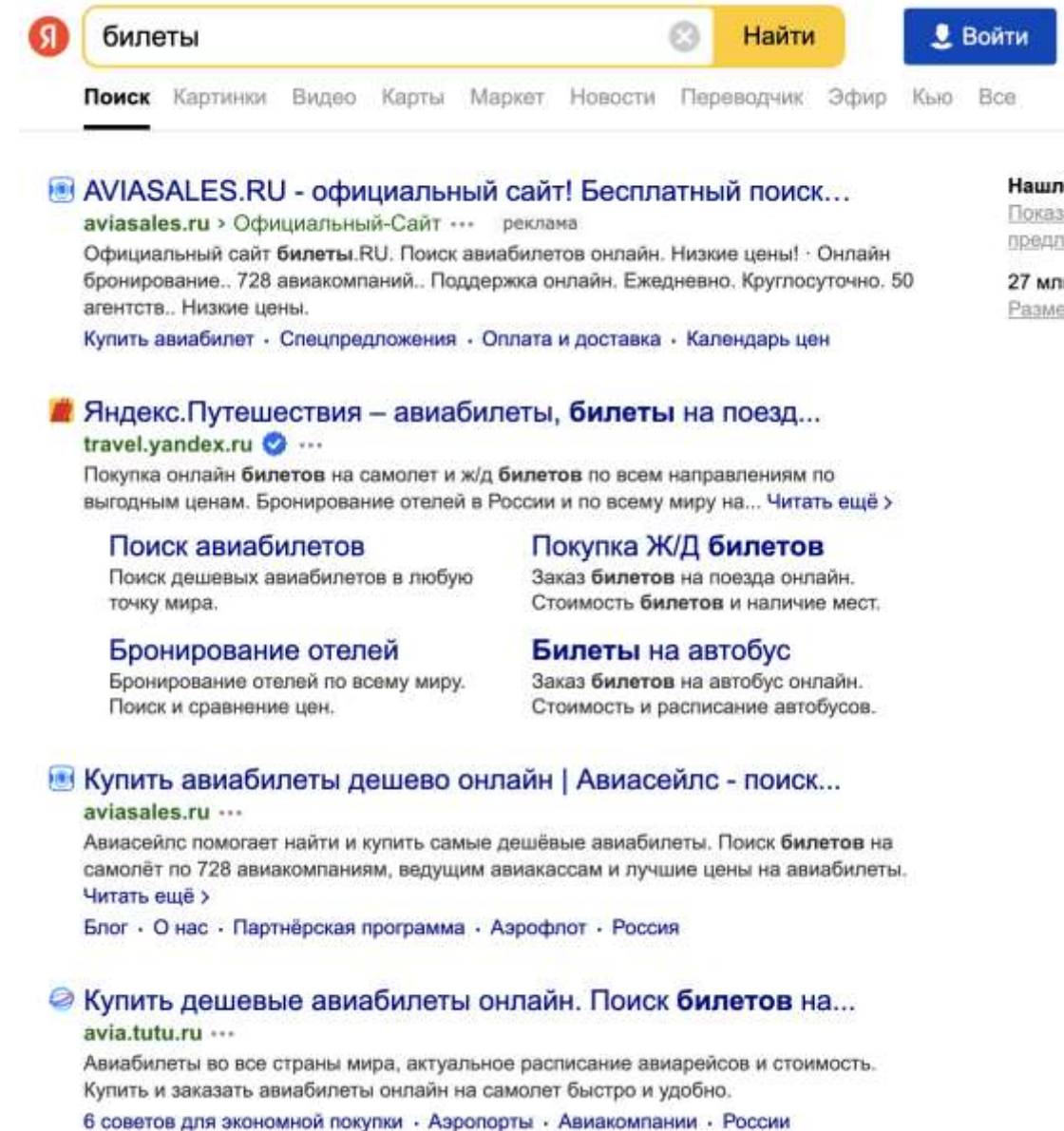
■ Реклама включена:

– 500 визитов/заказов в день с рекламы

– X визитов/заказов в день с органики

■ Реклама выключена

– Y визитов/заказов в день с органики



Вопросы

- Про АБ?

Вопросы

- Про АБ?
- Про Германию?

Вопросы

- Про АБ?
- Про Германию?
- Про Яндекс?