АБ-тестирование

Олег Хомюк

NewProLab, Специалист по большим данным
Весна 2019г.

Хомюк Олег

Yandex, Consultant+, Ezhome, Lamoda

oleg.khomyuk@gmail.com
Telegram: @khomyuk
Skype: oleg.khomyuk

https://www.facebook.com/oleg.khomyuk
https://www.linkedin.com/in/olegkhomyuk


50 % visitors
see variation A
Variation A
23%
conversion
50 % visitors
see variation B
Variation B
11%
conversion

АБ-тестирование

How to Write a Book, Fast
14 Days from Start to Finish
Unique, Step By Step Program
Write-A-Book-Faster.com

How to Write a Book Fast
14 Days from Start to Finish
Unique, Step By Step Program
Write-A-Book-Faster.com

АБ-тестирование

How to Write a Book, Fast
14 Days from Start to Finish
Unique, Step By Step Program
Write-A-Book-Faster.com

CTR = 4,4%

How to Write a Book Fast
14 Days from Start to Finish
Unique, Step By Step Program
Write-A-Book-Faster.com

CTR = 4,12%

Difference = 8%

1. Проверка статистических гипотез

Проверка статистических гипотез

Что это такое?

Статистическая гипотеза – это

Проверка статистических гипотез

Что это такое?

Статистическая гипотеза – это любое предположение о генеральной совокупности, проверяемое по выборке.

Проверка статистических гипотез

Зачем нужно **проверять** гипотезы?

Проверка статистических гипотез

Зачем нужно **проверять** гипотезы?

- **Отвечать на вопросы и принимать решения**
  (честная ли монетка, какая из монеток чаще падает орлом вверх)

Проверка статистических гипотез

Зачем нужно **проверять** гипотезы?

- **Отвечать на вопросы и принимать решения**
  (честная ли монетка, какая из монеток чаще падает орлом вверх)

Почему нельзя **просто** посмотреть на данные?

Проверка статистических гипотез

Зачем нужно **проверять** гипотезы?

- **Отвечать на вопросы и принимать решения**
  (честная ли монетка, какая из монеток чаще падает орлом вверх)

Почему нельзя **просто** посмотреть на данные?

- **Исключить влияние случайности**
  (это по случайным причинам так получилось или на самом деле так?)

Проверка статистических гипотез

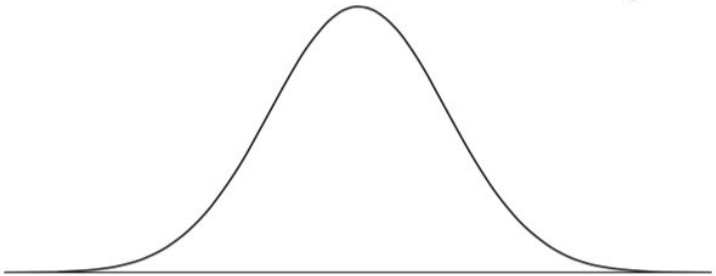
выборка: $X^n = (X_1, \ldots, X_n),\ X \sim \mathbf{P} \in \Omega$

нулевая гипотеза: $H_0\colon\ \mathbf{P} \in \omega,\ \ \omega \in \Omega$

альтернатива: $H_1\colon\ \mathbf{P} \notin \omega$

статистика: $T(X^n),\ \ T(X^n) \sim F(x)$ при $\mathbf{P} \in \omega$

$T(X^n) \not\sim F(x)$ при $\mathbf{P} \notin \omega$

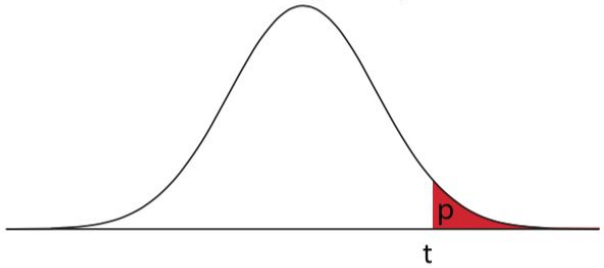


Проверка статистических гипотез

реализация выборки: $x^n = (x_1, \ldots, x_n)$

реализация статистики: $t = T(x^n)$

достигаемый уровень значимости: $p(x^n)$ — вероятность при H_0 получить $T(X^n) = t$ или ещё более экстремальное



$$p(x^n) = \mathbf{P}(T \geqslant t \,|\, H_0)$$

Гипотеза отвергается при $p(x^n) \leqslant \alpha$, α — уровень значимости

Проверка статистических гипотез

Основная идея:

- **Понять, насколько невероятны наблюдаемые данные с точки зрения нулевой гипотезы**
  (10 орлов из 10 бросков)

Проверка статистических гипотез

Основная идея:

- **Понять, насколько невероятны наблюдаемые данные с точки зрения нулевой гипотезы**
  (10 орлов из 10 бросков)

- **Формально с этим могут помочь значения статистики**
  (если средняя доля орлов для честной монетки сильно отклоняется от 0.5 - 0.9 например - то можно назвать такое значение “радикальным” или нетипичным для H0)

Проверка статистических гипотез

p-value - это

Проверка статистических гипотез

p-value - вероятность получить наблюдаемые данные (или более радикальные), при условии того, что верна нулевая гипотеза

Проверка статистических гипотез

p-value - вероятность получить наблюдаемые данные (или более радикальные), при условии того, что верна нулевая гипотеза

Общая схема:

- Получаем реализацию выборки

Проверка статистических гипотез

p-value - вероятность получить наблюдаемые данные (или более радикальные), при условии того, что верна нулевая гипотеза

Общая схема:

- Получаем реализацию выборки
- Считаем, насколько она невероятная, вычислив p-value

Проверка статистических гипотез

p-value - вероятность получить наблюдаемые данные (или более радикальные), при условии того, что верна нулевая гипотеза

Общая схема:

- Получаем реализацию выборки
- Считаем, насколько она невероятная, вычислив p-value
- Если p-value меньше некоторого заранее заданного значения, отвергаем нулевую гипотезу

Проверка статистических гипотез

Гипотеза H_0	Принимается	Отвергается
Верна	Правильное решение	Ошибка 1-го рода
Неверна	Ошибка 2-го рода	Правильное решение

Обозначим через α – вероятность допустить ошибку 1-го рода, через β – вероятность ошибки 2-го рода.

Вероятность α допустить ошибку 1-го рода, то есть отвергнуть верную гипотезу H_0, называют ***уровнем значимости***.

Проверка статистических гипотез

Задача асимметрична в смысле важности ошибок:

- Ошибки первого рода ограничиваются уровнем значимости
- Ошибки второго рода минимизируются путем выбора лучшего (более мощного) критерия

Свойства критериев:

- Корректность критерия
- Мощность критерия

Проверка статистических гипотез

- Если величина p достаточно мала, то данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

Проверка статистических гипотез

- Если величина p достаточно мала, то данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

- Если величина p недостаточно мала, то данные не свидетельствуют против нулевой гипотезы в пользу альтернативы.

Проверка статистических гипотез

- Если величина p достаточно мала, то данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

- Если величина p недостаточно мала, то данные не свидетельствуют против нулевой гипотезы в пользу альтернативы.

При помощи инструмента проверки гипотез нельзя доказать справедливость нулевой гипотезы.

Отсутствие доказательств ≠ доказательство отсутствия.

Проверка статистических гипотез

При любой проверке гипотез нужно оценивать размер эффекта — степень отличия нулевой гипотезы от истины, и оценивать его практическую значимость.

Проверка статистических гипотез

При любой проверке гипотез нужно оценивать размер эффекта — степень отличия нулевой гипотезы от истины, и оценивать его практическую значимость.

(Lee et al, 2010): за три года женщины, упражнявшиеся не меньше часа в день, набрали значимо меньше веса, чем женщины, упражнявшиеся меньше 20 минут в день ($p < 0.001$).

Разница в набранном весе составила 150 г. Практическая значимость такого эффекта сомнительна.

Проверка статистических гипотез

Утверждается, что осьминог предсказывает результаты матчей с участием сборной Германии на чемпионате мира по футболу 2010 года, выбирая кормушку с флагом страны-победителя. По результатам 13 испытаний ему удаётся верно угадать результаты 11 матчей.
Критерий даёт достигаемый уровень значимости $p \approx 0.0112$.

Проверка статистических гипотез

Утверждается, что осьминог предсказывает результаты матчей с участием сборной Германии на чемпионате мира по футболу 2010 года, выбирая кормушку с флагом страны-победителя. По результатам 13 испытаний ему удаётся верно угадать результаты 11 матчей.
Критерий даёт достигаемый уровень значимости $p \approx 0.0112$.

Достигаемый уровень значимости нельзя интерпретировать как вероятность справедливости нулевой гипотезы!

0.0112 — не вероятность того, что осьминог выбирает кормушку наугад! Эта вероятность равна единице.

Проверка статистических гипотез

Мощность - вероятность отвергнуть H0, если верна альтернатива.

Мощность критерия зависит от следующих факторов:

- размер выборки;
- размер отклонения от нулевой гипотезы;
- чувствительность статистики критерия;
- тип альтернативы.

Проверка статистических гипотез

На выборке из 10 бросков монетки вы не отличите честную от смещённой с вероятностью орла 0.51.

Проверка статистических гипотез

На выборке из 10 бросков монетки вы не отличите честную от смещённой с вероятностью орла 0.51.

Обеспечение требуемой мощности: размеры выборки подбирается так, чтобы при размере отклонения от нулевой гипотезы не меньше заданного (например, вероятность орла не меньше 0.51) мощность была не меньше заданной.

Проверка статистических гипотез

На выборке из 10 бросков монетки вы не отличите честную от смещённой с вероятностью орла 0.51.

Обеспечение требуемой мощности: размеры выборки подбирается так, чтобы при размере отклонения от нулевой гипотезы не меньше заданного (например, вероятность орла не меньше 0.51) мощность была не меньше заданной.

Руководствуясь этим правилом, оценивается время АБ тестирования. Например, вы хотите показать увеличение конверсии с 0.05 до 0.053, значит, нужно собрать столько событий, чтобы при конверсии не менее 0.053 гипотеза о её равенстве 0.05 отвергалась с вероятностью более 85%.

2. Параметрические критерии

Параметрические критерии

Особенности:

- Предполагают дополнительные знания о характере распределения в выборке
- Это позволяет использовать более мощные критерии для конкретных случаев

Параметрические критерии

Особенности:

- Предполагают дополнительные знания о характере распределения в выборке
- Это позволяет использовать более мощные критерии для конкретных случаев

К сожалению, реальные данные очень редко распределены как табличные распределения. Но есть ряд популярных случаев, часто применимых на практике.

Параметрические критерии

Биномиальный критерий:

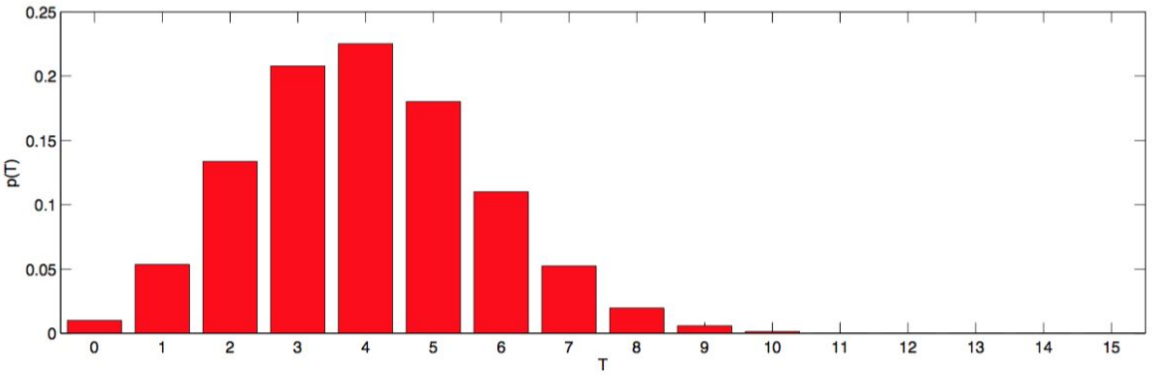
выборка: $X^n = (X_1, \ldots, X_n), X \sim Ber(p)$

нулевая гипотеза: $H_0 : p = p_0$

альтернатива: $H_1 : p <\neq> p_0$

статистика: $T(X^n) = \sum_{i=1}^{n} X_i$

нулевое распределение: $Bin(n, p_0)$



Параметрические критерии

Z-критерий для доли:

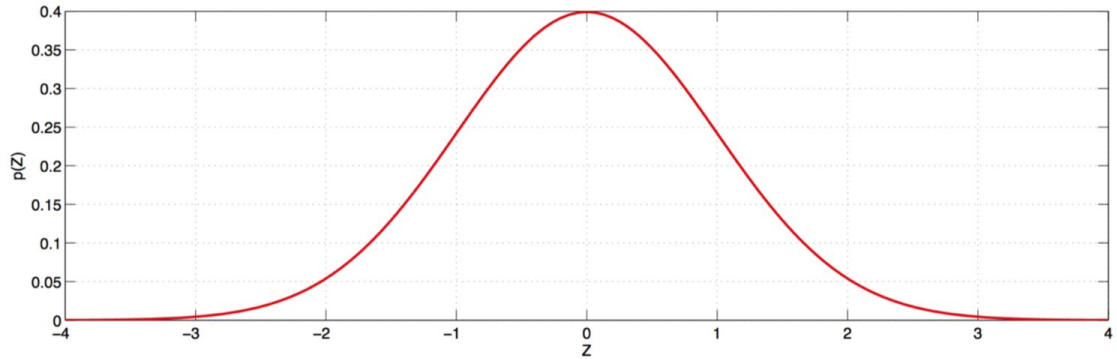
выборка: $X^n = (X_1, \ldots, X_n), X \sim Ber(p)$

нулевая гипотеза: $H_0 : p = p_0$

альтернатива: $H_1 : p <\neq> p_0$

статистика: $Z_S(X^n) = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

нулевое распределение: $N(0,1)$



Параметрические критерии

Z-критерий разности долей (независимые выборки):

выборки: $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1}), X_1 \sim Ber(p_1)$
$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2}), X_2 \sim Ber(p_2)$
выборки независимы

Исход \ Выборка	$X_1^{n_1}$	$X_2^{n_2}$
1	a	b
0	c	d
$\sum$	n_1	n_2

нулевая гипотеза: $H_0 : p_1 = p_2$

альтернатива: $H_1 : p_1 <\neq> p_2$

статистика: $Z\left(X_1^{n_1}, X_2^{n_2}\right) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$P = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$, $\hat{p}_1 = \frac{a}{n_1}$, $\hat{p}_2 = \frac{b}{n_2}$

нулевое распределение: $N(0,1)$

Параметрические критерии

Z-критерий разности долей (связанные выборки):

выборки: $X_1^n = (X_{11}, \ldots, X_{1n}), X_1 \sim Ber(p_1)$
$X_2^n = (X_{21}, \ldots, X_{2n}), X_2 \sim Ber(p_2)$

выборки связанные

X_1^n \ X_2^n	1	0
1	e	f
0	g	h

нулевая гипотеза: $H_0 : p_1 = p_2$

альтернатива: $H_1 : p_1 <\neq> p_2$

статистика: $Z(X_1^n, X_2^n) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{f+g}{n^2} - \frac{(f-g)^2}{n^3}}} = \frac{f-g}{\sqrt{f+g-\frac{(f-g)^2}{n}}}$

нулевое распределение: $N(0,1)$

Параметрические критерии

Z-критерий:

выборки: $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1}), X_1 \sim N\left(\mu_1, \sigma_1^2\right)$
$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2}), X_2 \sim N\left(\mu_2, \sigma_2^2\right)$

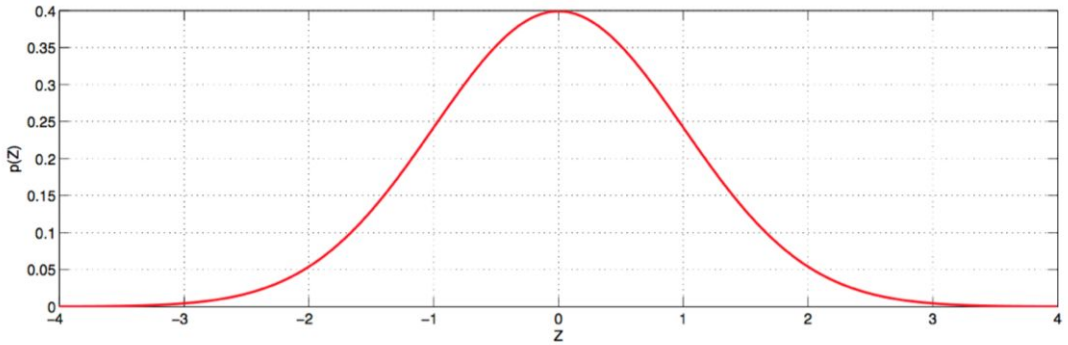
σ_1, σ_2 известны

нулевая гипотеза: $H_0 : \mu_1 = \mu_2$

альтернатива: $H_1 : \mu_1 <\neq> \mu_2$

статистика: $Z\left(X_1^{n_1}, X_2^{n_2}\right) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

нулевое распределение: $N(0, 1)$



Параметрические критерии

t-критерий Стьюдента:

выборки: $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1}), X_1 \sim N\left(\mu_1, \sigma_1^2\right)$
$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2}), X_2 \sim N\left(\mu_2, \sigma_2^2\right)$

σ_1, σ_2 неизвестны

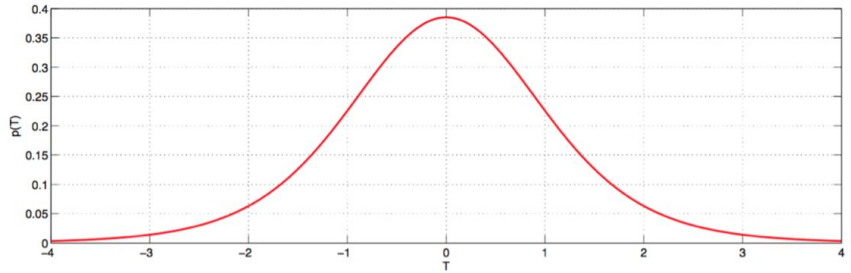
нулевая гипотеза: $H_0 : \mu_1 = \mu_2$

альтернатива: $H_1 : \mu_1 <\neq> \mu_2$

статистика: $T\left(X_1^{n_1}, X_2^{n_2}\right) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}$$

нулевое распределение: $\approx St(\nu)$



Приближение достаточно точно при $n_1 = n_2$ или $[n_1 > n_2] = [\sigma_1 > \sigma_2]$.

Параметрические критерии

- Перед использованием методов, предполагающих нормальность, стоит проверить нормальность. (критерий Харке-Бера, критерий согласия Пирсона)

Параметрические критерии

- Перед использованием методов, предполагающих нормальность, стоит проверить нормальность. (критерий Харке-Бера, критерий согласия Пирсона)

- Если гипотеза нормальности отвергается, следует использовать непараметрические методы.

3. Непараметрические критерии

Непараметрические критерии

Критерий Мана-Уитни:

выборки: $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1})$
$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2})$
выборки независимые

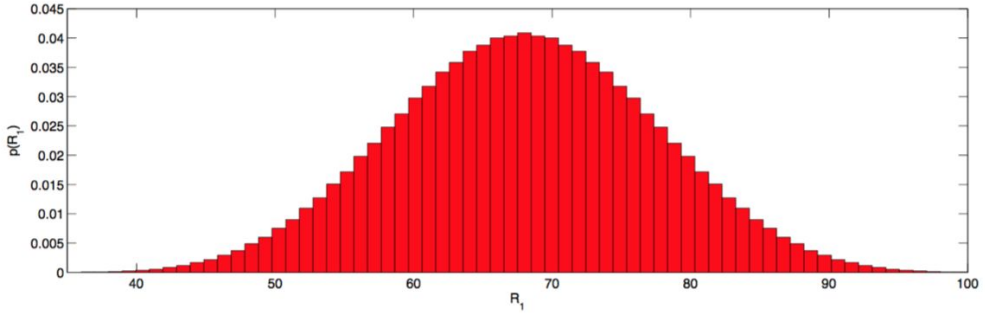
нулевая гипотеза: $H_0 \colon F_{X_1}(x) = F_{X_2}(x)$

альтернатива: $H_1 \colon F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta <\neq> 0$

статистика: $X_{(1)} \leq \ldots \leq X_{(n_1+n_2)}$ — вариационный ряд объединённой выборки $X = X_1^{n_1} \bigcup X_2^{n_2}$

$$R_1(X_1^{n_1}, X_2^{n_2}) = \sum_{i=1}^{n_1} \operatorname{rank}(X_{1i})$$

нулевое распределение: табличное



Bootstrap

Бутстрэп[1] (англ. *bootstrap*) в статистике — практический компьютерный метод исследования распределения статистик вероятностных распределений, основанный на многократной генерации выборок методом Монте-Карло на базе имеющейся выборки[2].

Позволяет просто и быстро оценивать самые разные статистики (доверительные интервалы, дисперсию, корреляцию и так далее) для сложных моделей.

4. Разбиение на группы

Разбиение на группы

Что стоит проверить:

- Пол, возраст распределены одинаково в группах (Критерии согласия)

Разбиение на группы

Что стоит проверить:

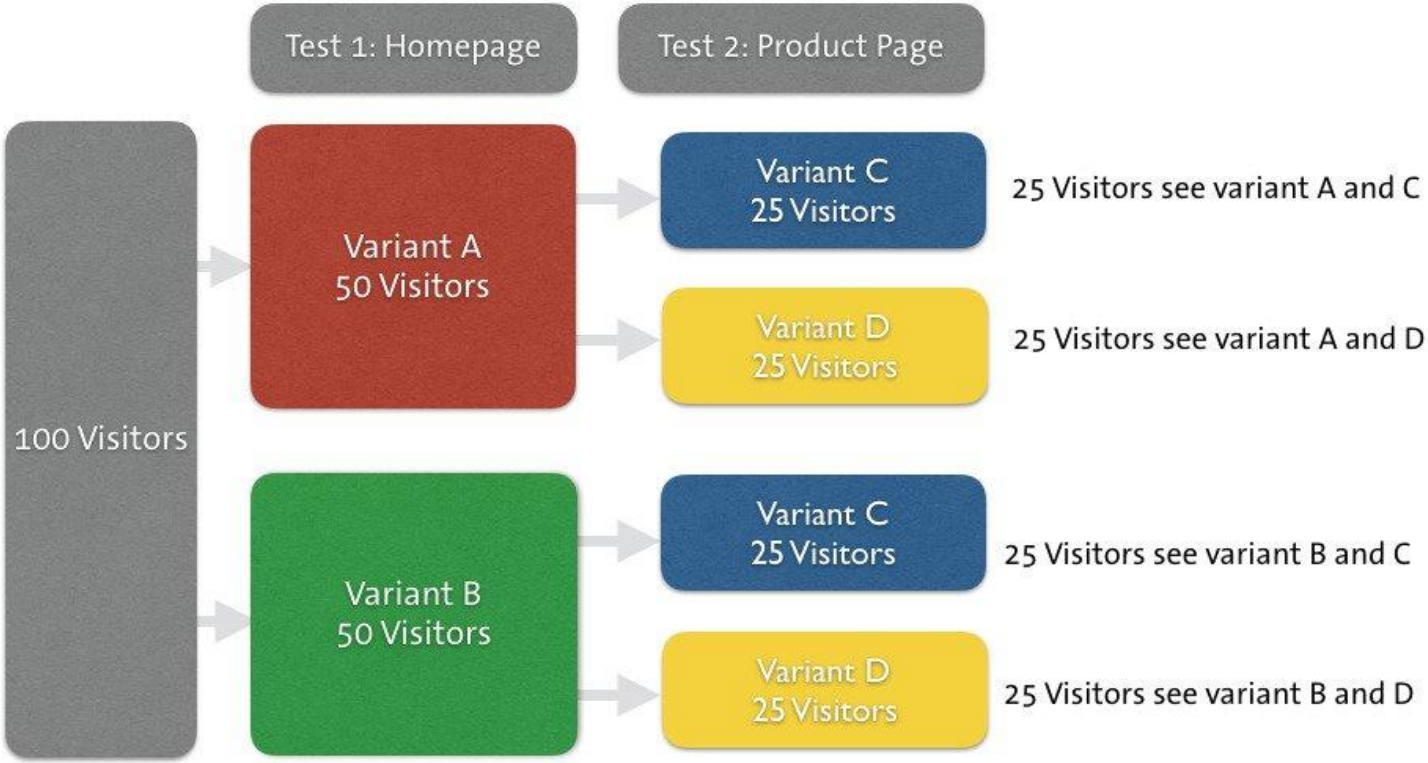
- Пол, возраст распределены одинаково в группах (Критерии согласия)
- Исторически целевые метрики (конверсия, выручка) для групп не отличаются (непараметрические критерии)

Разбиение на группы

Что стоит проверить:

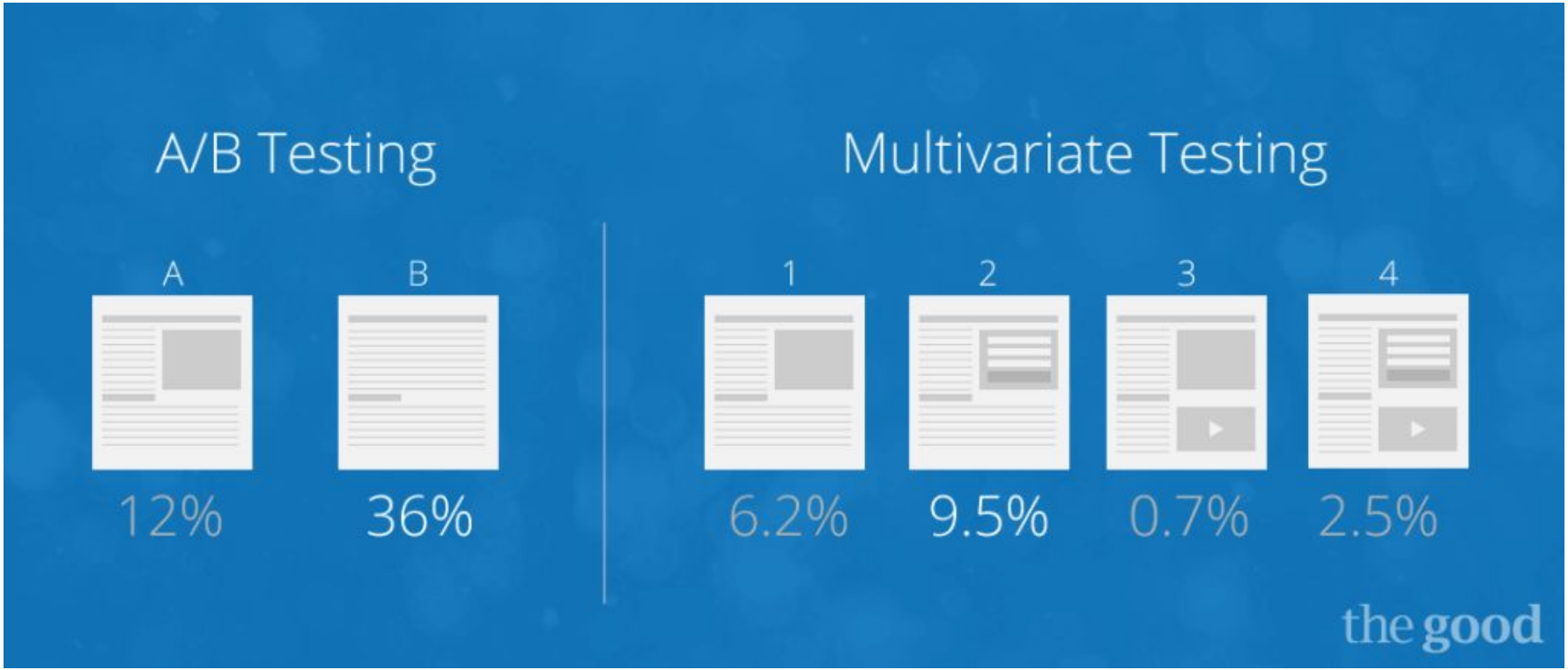
- Пол, возраст распределены одинаково в группах (Критерии согласия)
- Исторически целевые метрики (конверсия, выручка) для групп не отличаются (непараметрические критерии)
- AA-тест

Многомерные АБ-тесты

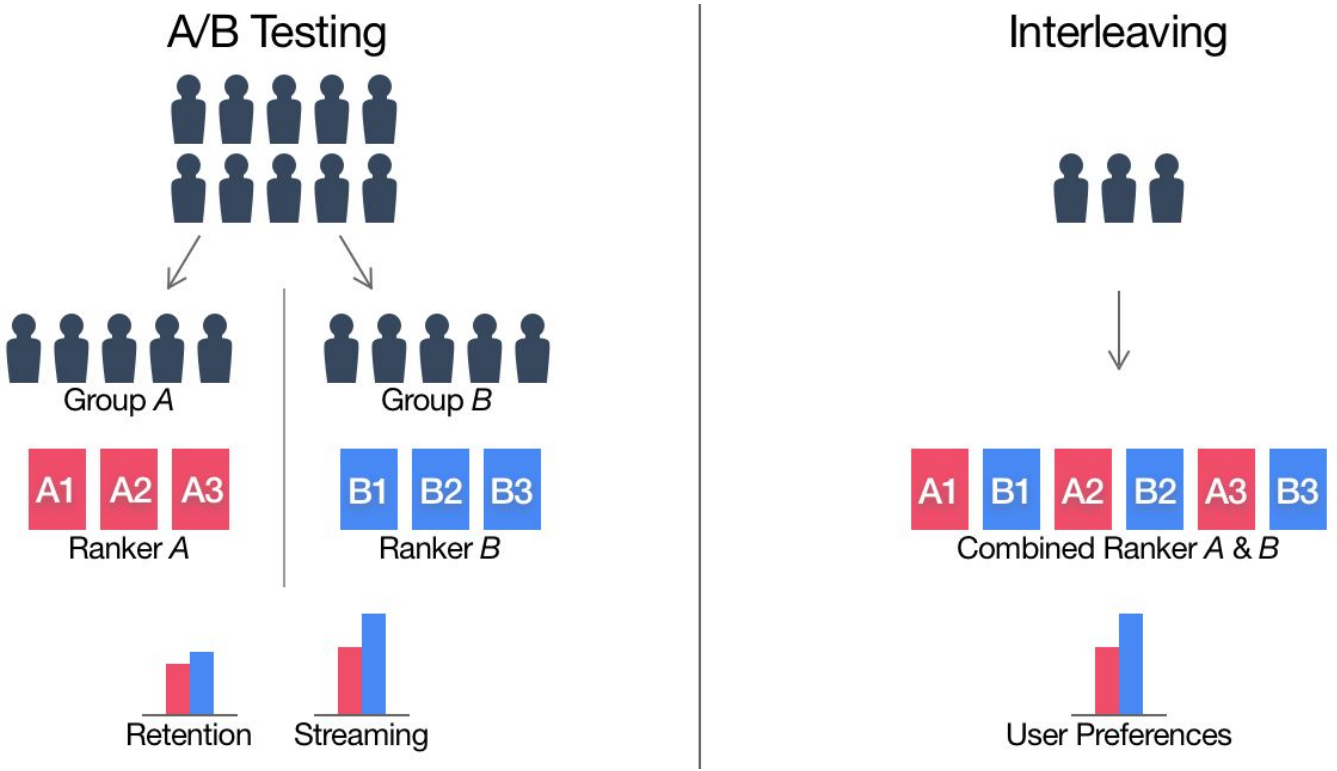


5. Другие виды контролируемых экспериментов

Multivariate Testing



Team-Draft-Interleaving



6. Что может пойти не так?

Множественная проверка гипотез

- Одновременно тестируется средний чек, среднее число товаров в чеке, среднее число аксессуаров в чеке. Для каждой из этих величин вы составили свой критерий для проверки гипотезы о наличие эффекта.

Каков **уровень значимости** для такой одновременной проверки гипотез?

Множественная проверка гипотез

Ошибка первого рода вызвана не особенностью данных, а тем, что мы несколько раз её проверяем.

Множественная проверка гипотез

Ошибка первого рода вызвана не особенностью данных, а тем, что мы несколько раз её проверяем.

Нужно применять методы **множественной проверки гипотез** (Multiple comparisons problem).

Есть реализация в Python — statsmodels.sandbox.stats.multicomp.multipletests.

Последовательный анализ

Можно ли досрочно завершить АБ-тест?

Последовательный анализ

Можно ли досрочно завершить АБ-тест?

(Например, план был на 28 дней, рассчитано по биномиальному критерию для детекции 1% роста конверсии, но через 11 дней вы видите +3% со значимостью)

Последовательный анализ

Можно ли досрочно завершить АБ-тест?

(Например, план был на 24 дня, рассчитано по биномиальному критерию для детекции 1% роста конверсии, но через 11 дней вы видите +3% со значимостью)

Впрямую нельзя, смещается нулевая статистика!

мы останавливаемся там, где данные больше всего свидетельствуют против нулевой гипотезы

Последовательный анализ

Можно ли досрочно завершить АБ-тест?

(Например, план был на 24 дня, рассчитано по биномиальному критерию для детекции 1% роста конверсии, но через 11 дней вы видите +3% со значимостью)

Впрямую нельзя, смещается нулевая статистика!

мы останавливаемся там, где данные больше всего свидетельствуют против нулевой гипотезы

Нужно применять последовательный анализ (гуглите Sequential analysis).

Он даёт во-первых корректные, а во-вторых более мощные критерии, пользуясь дополнительным знанием о потоковости данных.

Длительность эксперимента

- Бесконечные эксперименты

Длительность эксперимента

- Бесконечные эксперименты

- Для новой функциональности не стоит учитывать первые дни (выработка пользовательской привычки 21 день)

- Учет цикличности в пользовательских привычках
  (длительность кратная 7 дням даже если 4 дней достаточно)

АА-тест

- В двух группах показывается одно и то же (например, сайт и его полная копия)
- Цель - не “обнаружить” различий
- Помогает судить о корректности процедуры проведения эксперимента
- Иногда предшествует настоящему АБ тесту

Что еще стоит проверить

- Процесс раскатки
  Равные условия для тестового варианта и контрольного варианта
  (кейс про мобильные платформы)

Что еще стоит проверить

- Процесс раскатки
  Равные условия для тестового варианта и контрольного варианта (кейс про мобильные платформы)

- Условия проведения эксперимента идентичны будущим боевым условиям

Что еще стоит проверить

- Процесс раскатки
  Равные условия для тестового варианта и контрольного варианта (кейс про мобильные платформы)

- Условия проведения эксперимента идентичны будущим боевым условиям

- Влияние параллельных экспериментов друг на друга (дизайн блока рекомендаций и его наполнение)

Полезные ссылки

- Лекции ВШЭ
  http://wiki.cs.hse.ru/Прикладной_статистический_анализ_данных

- Лекции Воронцова
  http://www.machinelearning.ru/wiki/index.php?title=Статистический_анализ_данных_%28курс_лекций%2C_К.В.Воронцов%29

Спасибо за внимание!