

Курс по А/Б тестированию

Даниил Потапов

Руководитель Лаборатории Искусственного Интеллекта

РСХБ



О себе



Что-то профессиональное:

- Стаж 10 лет, 4 компании
- GIS, Front-end, Full-stack
- Open source & open edu

Что-то необычное:

- 10 лет занимался шахматами
- Арахнофоб
- Писал читы для игр

Что-то личное:

- 3,5к часов в CS 1.6
- Люблю научную фантастику
- Коллекционер книг и игр

Что-то неприличное:

- ругаюсь %#&
- КМС по литрболу
- Не умею готовить еду

Команда курса

Даниил Потапов

Рук-ль Лаборатории ИИ
PCХБ



Ирина Елисова

Head of ML
Geomotive



Людмила Коновалова

Senior Data Analyst
Яндекс.GPT



Ильдар Сафило

Senior ML Researcher
Booking.com



Юрий Котов

Senior Data Engineer
Т-Банк



О курсе

- 5 лекций
- 5 семинаров
- 2 домашки
- 16 часов на занятия

О курсе

- 5 лекций
- 5 семинаров
- 2 домашки
- 16 часов на занятия

Орг. моменты

- Занятия - вторник и среда
 - Чаще всего раз в неделю
- Одна пара за раз, 18:40 - 20:10
- Аттестация - экзамен/зачет
 - Выставляем по баллам
- Канал в Telegram



Еще орг. моменты

- Материалы курса будем выкладывать в Telegram чате
- Там же мы закрепим ментора за каждым студентом
 - Тем не менее, не стесняйтесь задавать вопросы в общем чате
 - Будем начислять доп. баллы тем, кто помогает своим однокурсникам
- Будет отдельная гугл таблица с оценками по ДЗ
- Об этом всем дополнительно еще в Telegram чате проинформируем
- И еще раз QR



План курса

Лекции

1. Введение в А/Б-тестирование
2. Дизайн А/Б-теста
3. Дизайн в реальных условиях
4. Методы ускорения
5. Последовательный анализ

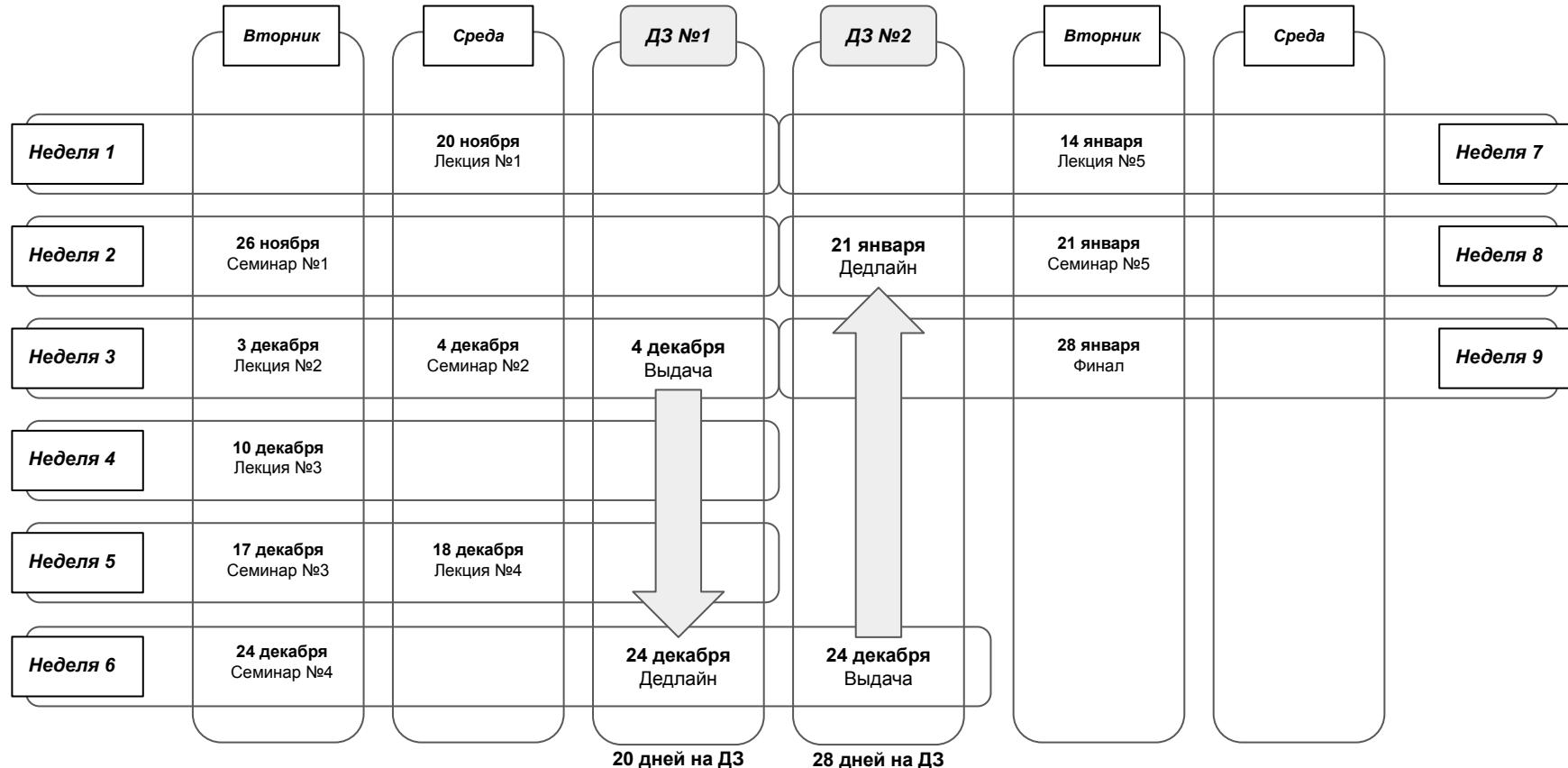
Семинары

1. Вспоминаем мат. стат
2. Практика по дизайну А/Б-теста
3. Обзор методов и фреймворков
4. Практика по ускорению
5. Практика по посл. анализу

Домашки (50 баллов каждая)

1. Самостоятельный дизайн А/Б
2. Ускорение своего дизайна

Расписание



Домашние задания

Домашки (50 баллов каждая)

1. Самостоятельный дизайн А/Б
2. Ускорение своего дизайна

Правила:

- Две попытки исправить замечания
- После дедлайна сдать/досдать нельзя
- На проверку ДЗ у ментора будет неделя
- Дедлайн второй ДЗ - 21 января, далее неделя на подведение итогов
- Оценки/зачеты будут выставлены к финальному занятию 28 января

План лекции

- Что такое А/Б-тестирование
- Основные понятия и предпосылки
- Общая схема подготовки А/Б-теста
- Статистические критерии
- Мощность и корректность

План лекции

- Что такое А/Б-тестирование
- Основные понятия и предпосылки
- Общая схема подготовки А/Б-теста
- Статистические критерии
- Мощность и корректность

Что такое А/Б-тестирование

А/Б-тестирование - это проверка гипотез.

- Две группы (но может быть и больше) - случайно/независимо полученные
- Метрика - что измеряем?
- Значимость - какова вероятность, что это случайность?

На основе анализа изменений интересующей метрики в двух группах пользователей (вызванных, например, изменениями пользовательского интерфейса, рекомендациями и т.д.) мы можем установить и подтвердить неслучайный характер этих изменений.

Что такое А/Б-тестирование

А/Б-тестирование - это метод, облегчающий принятие решений, базирующийся на данных.

- Какое влияние внесенных изменений наблюдается?
 - Положительное или отрицательное
- Каков масштаб их воздействия?
 - Сильный или слабый
- Является ли результат значимым?
 - Вероятность получить такой результат случайно
- Является ли результат практически значимым?
 - Приносит ли бизнес ценность

Что такое А/Б-тестирование

А/Б-тестирование - это исследование пользовательского опыта

О чем важно помнить:

- Отсутствие влияющих факторов, кроме самого тестируемого изменения
- Репрезентативность выборки
- Возможность ответить на бизнес-вопрос без проведения теста
- Точность оценки: за точность оценки отвечают статистические критерии, плотность и количество данных, параметры распределения метрик.

История возникновения

- Исследование гомеопатических лекарств (**1835 год**)
 - [Inventing the randomized double-blind trial: the Nuremberg salt test of 1835](#)
 - “In 1834, annoyed by homeopathy's rising popularity, Friedrich Wilhelm von Hoven, the city's highest ranking public health official and head of the local hospitals, published a devastating critique of homeopathy ... Von Hoven accused homeopathy of lacking any scientific foundation. He suggested that homeopathic drugs were not real medicines at all and alleged homeopathic cures were either due to dietetic regimens and the healing powers of nature, or showed the power of belief.”
- Конец 19 - начало 20 века - применение в рекламе
 - [Клод Хопкинс](#) на промо-купонах отслеживал эффективность своих рекламных кампаний
- 1908 год - [Уильям Госсет](#) придумал [t-тест \(Т-критерий Стьюдента\)](#)
 - Мотивация - оценка качества пива Гиннесс
 - [Guinness, Gosset, Fisher, and Small Samples](#)

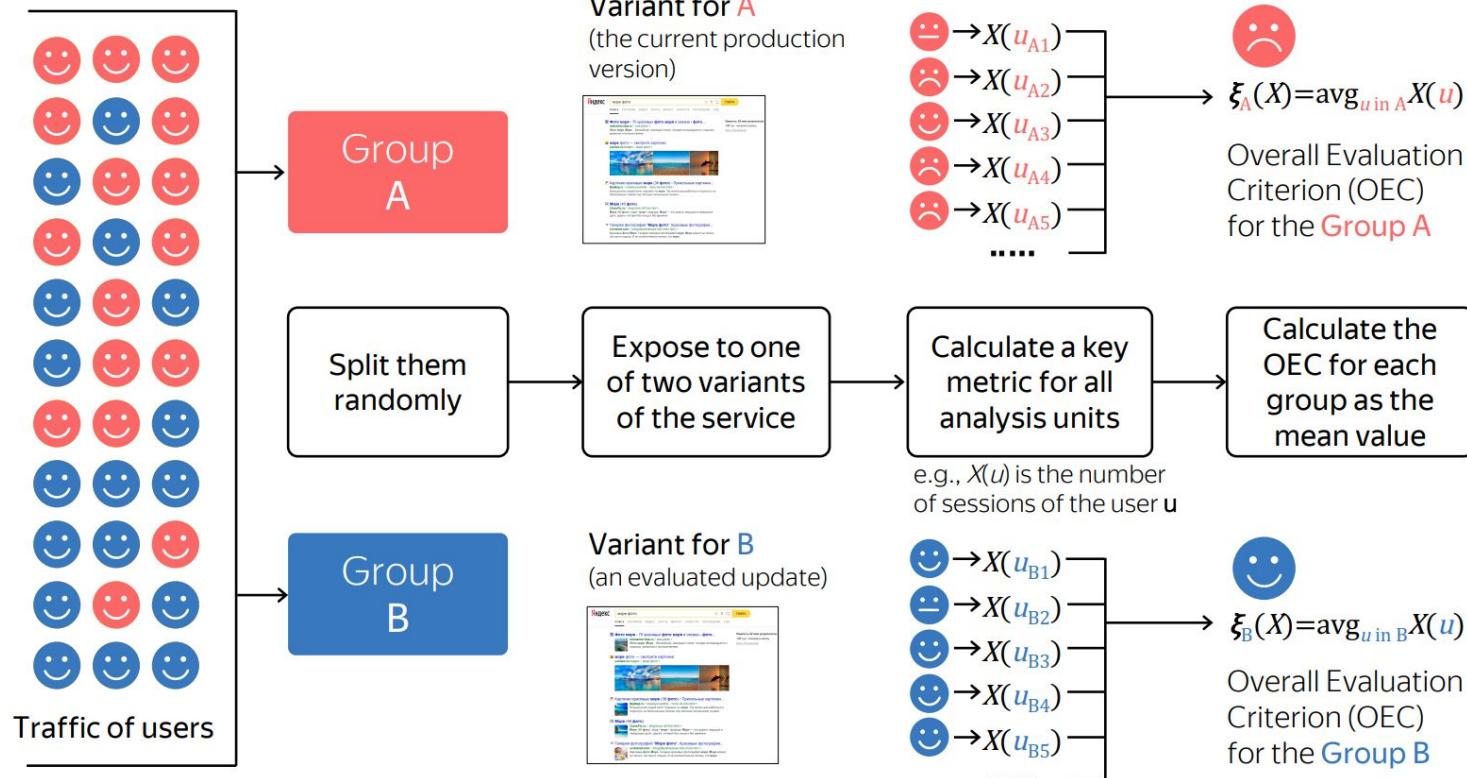
Где применяются

- Везде :)

Где применяются

- Медицина
- Банки
- E-commerce
- Онлайн кинотеатры
- Социальные сети
- Сервисы такси/доставки
- Онлайн игры
- И так далее

Как применяется



Как применяется



$$\xi_A(X) = \text{avg}_{u \text{ in } A} X(u)$$

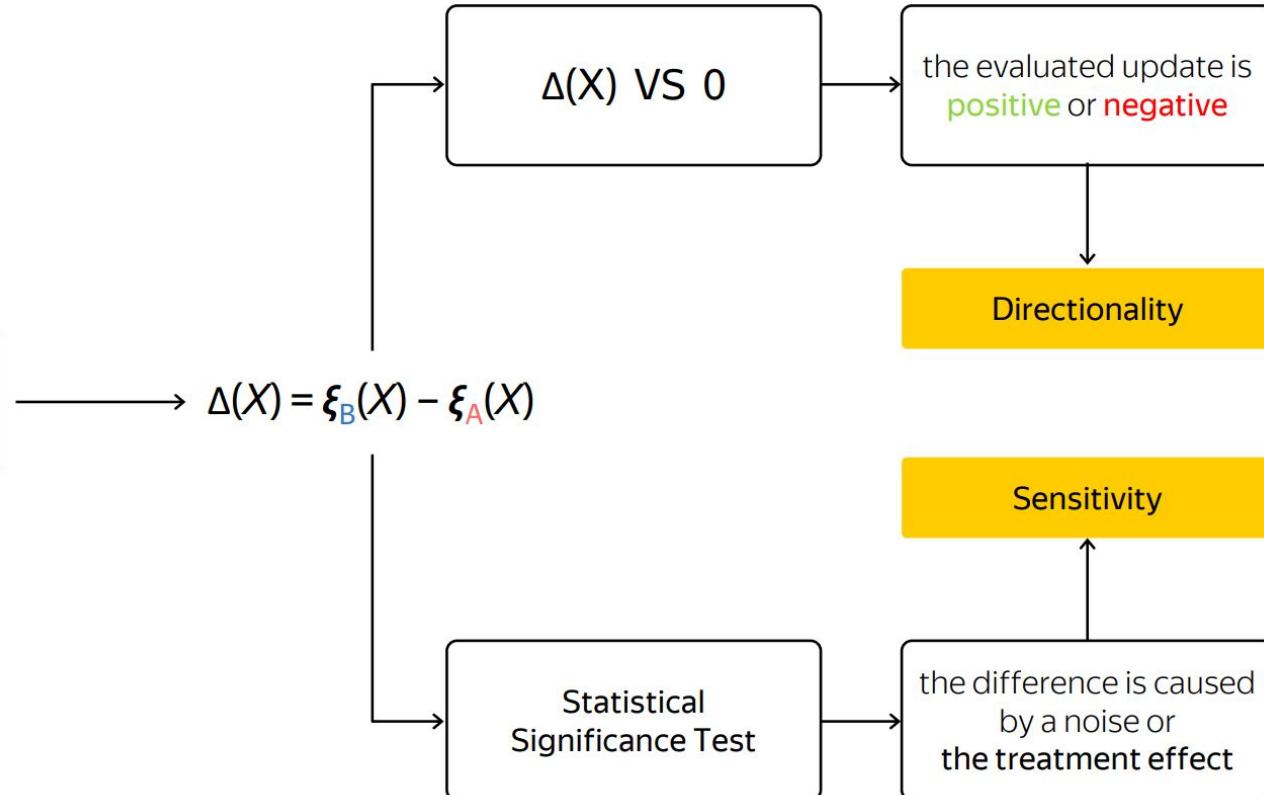
Overall Evaluation Criterion (OEC) for the **Group A**

Calculate the OEC for each group as the mean value



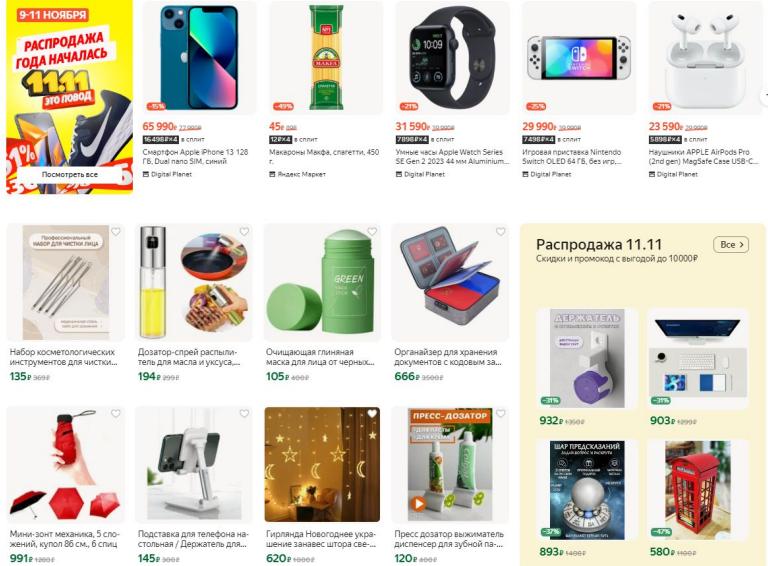
$$\xi_B(X) = \text{avg}_{u \text{ in } B} X(u)$$

Overall Evaluation Criterion (OEC) for the **Group B**

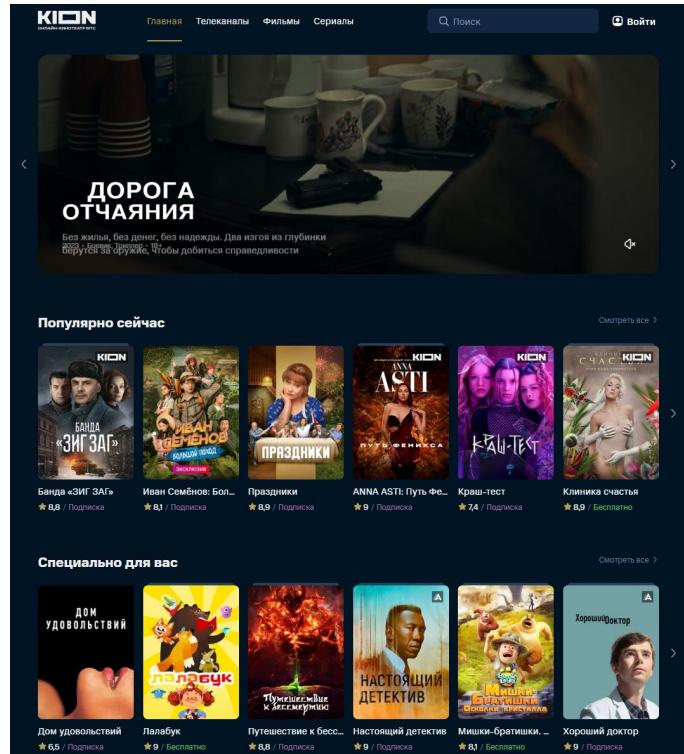


Примеры

E-commerce



Онлайн кинотеатр



Что тут можно по тестировать через А/Б?

План лекции

- Что такое А/Б-тестирование
- Основные понятия и предпосылки
- Общая схема подготовки А/Б-теста
- Статистические критерии
- Мощность и корректность

Гипотезы и метрики

Идея, лежащая в основе продуктовой гипотезы, заключается в том, чтобы решить, как "исправить" потенциальную проблему в продукте, как решение этих проблем повлияет на показатели, представляющие интерес

Метрика отражает эффективность нововведения для экспериментальной и контрольной групп теста, показывая, существует ли статистически значимая разница между этими двумя группами

Одним из способов проверить точность выбранной метрики, может быть ответ на следующий вопрос:

Если бы этот выбранный показатель значительно увеличился, в то время как все остальное осталось неизменным, достигли бы мы нашей цели и решили проблему?

Гипотезы и метрики

- Прибыль, доход
- Клики - на пользователя, в рамках сессии
- Конверсия - по каналам, по товарам
- Среднее время пользовательской сессии
- Среднее время между пользовательскими сессиями
- Retention - возвращаемость пользователя
- Средний чек, средняя маржа
- И тысячи других

Гипотезы и метрики

- Необходимо приоритезировать гипотезы для проверки и выбрать наиболее важную для продукта
- Доход не всегда является конечной целью, поэтому нужно связать метрику с прямыми целями продукта и целями более высокого уровня
- Для начала лучше выбирать одну целевую метрику для принятия решений, однако мониторить можно несколько

Основные термины

Проверка статистической гипотезы — это процесс принятия решения о том, противоречит ли рассматриваемая статистическая гипотеза наблюдаемой выборке данных, с помощью статистического критерия

В основе - не опровержение или доказательство, а проверка на вероятность наблюдаемого события.

Нулевая гипотеза H_0 - между группами значимой разницы нет

Альтернативная гипотеза H_1 - соответственно, разница есть

Выборка $X = \{X_1, \dots, X_n\}$ Статистика (функция) - $T(\cdot)$

Основные термины

Гипотезы H_0 и H_1

Выборка $X = \{X_1, \dots, X_n\}$

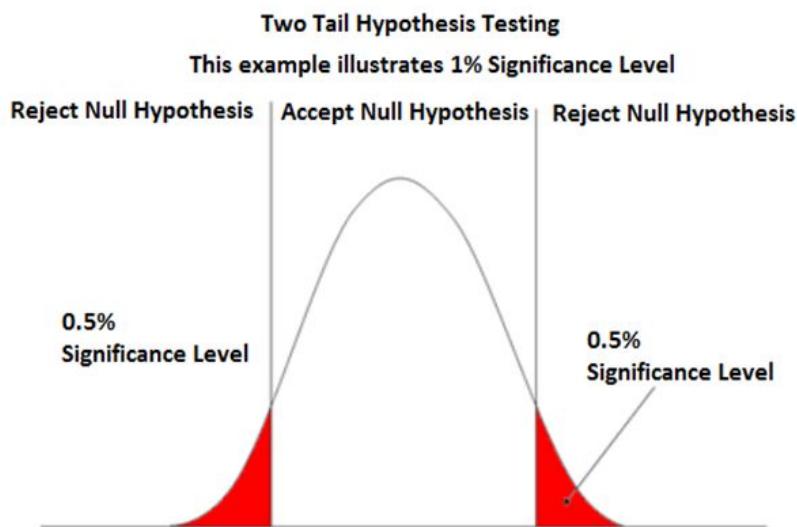
Статистика - $T(\cdot)$, $t = T(X)$

Достигаемый уровень значимости $p(X)$

- это вероятность при H_0 получить $t = T(X)$

или еще более экстремальное значение

p-value - $p(X) = \Pr(T \geq t | H_0)$

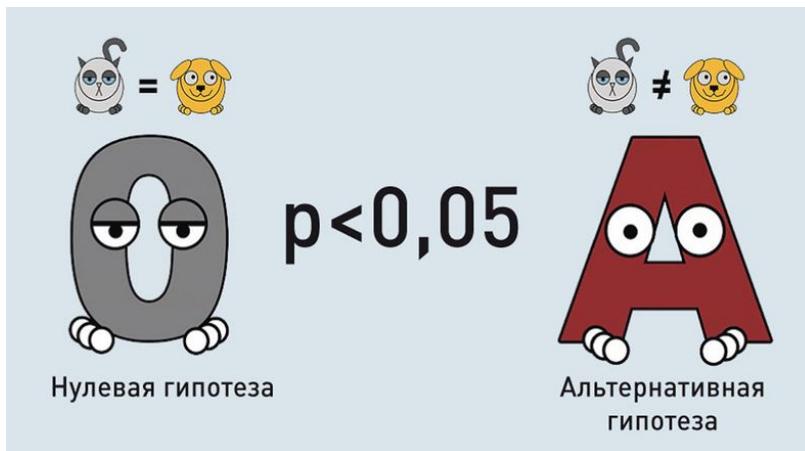


Гипотеза H_0 отвергается при $p(X) \leq \alpha$

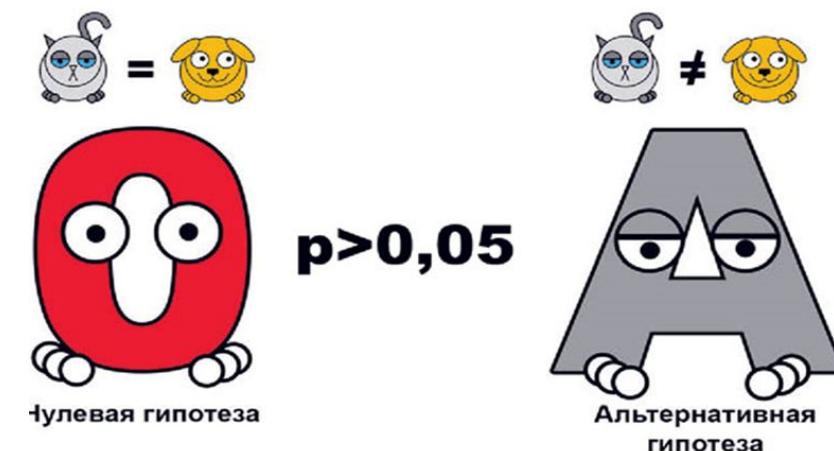
α - уровень значимости

Интерпретация результата

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



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



Интерпретация результата

Достигаемый уровень значимости (p-value) **неравен**:

- вероятности истинности нулевой гипотезы (потому что это вероятность при справедливости нулевой гипотезы получить значение статистики, такое же или ещё более экстремальное)
- вероятности ошибки первого рода
- вероятность того, что повторный эксперимент не приведёт к тому же решению

$1 -$ (достигаемый уровень значимости) не равно:

- вероятности истинности альтернативной гипотезы
- вероятности ошибки второго рода

Если данные не противоречат нулевой гипотезе, это **ещё не значит, что гипотеза верна.**

Ошибки 1 и 2 рода

H_1 : есть беременность; H_0 : нет беременности

Истинный
позитив, верна
 H_1



Ложный
позитив,
ошибка I
рода,
ложная
тревога



Истинный
негатив,
верна H_0

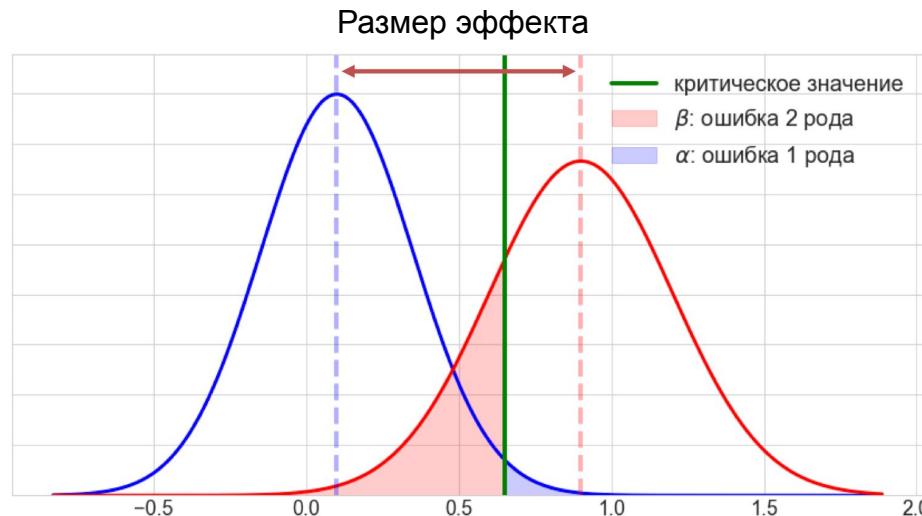


Ложный
негатив,
ошибка II
рода,
халатная
беспечность



- Ошибка 1 рода – вероятность принять альтернативную гипотезу, когда верна нулевая (α)
- Ошибка 2 рода – вероятность принять нулевую гипотезу, когда верна альтернативная (β)

Ошибки 1 и 2 рода



- Синим - гипотеза H_0
- Красным - гипотеза H_a

План лекции

- Что такое А/Б-тестирование
- Основные понятия и предпосылки
- Общая схема подготовки А/Б-теста
- Статистические критерии
- Мощность и корректность

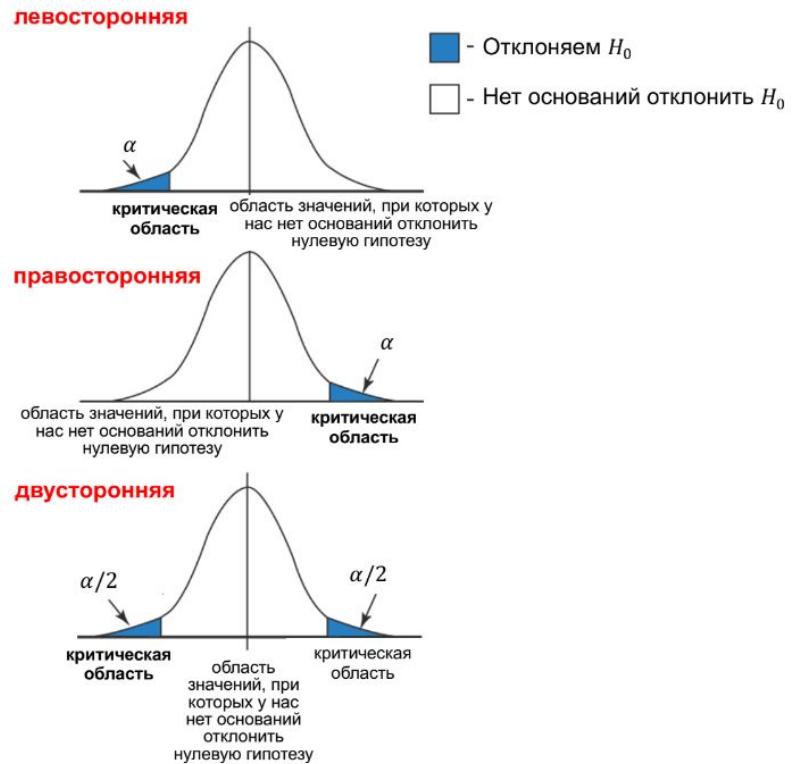
Общая схема подготовки к А/Б



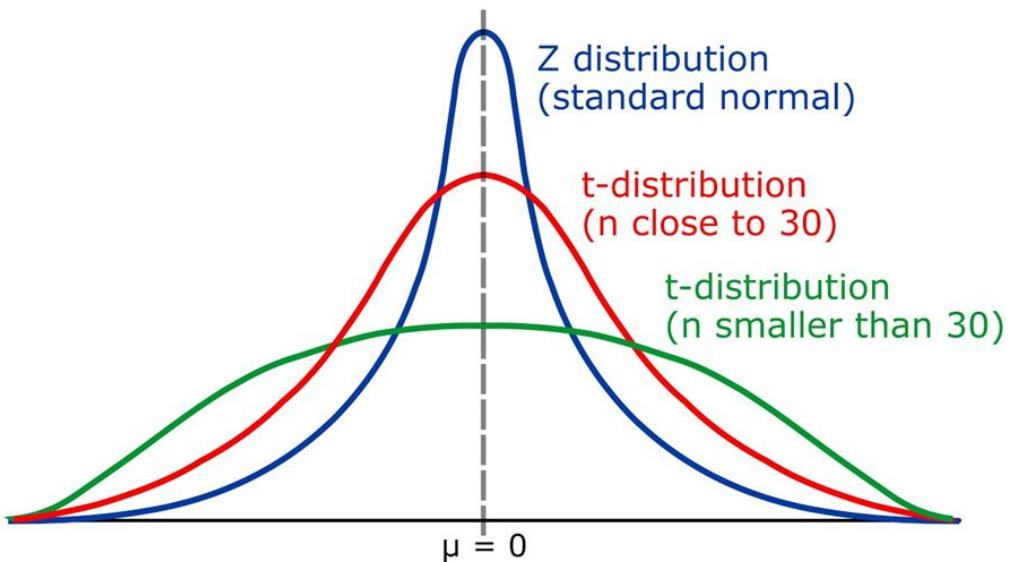
Односторонние и двусторонние критерии

Когда мы уверены в направлении ожидаемых изменений – можем выбрать односторонний критерий

Тогда вся вероятность ошибки 1-го рода располагается в выбранном «хвосте»



Z- и T-распределения



Степень свободы (Degree of Freedom, df) - разность числа наблюдений и числа оцененных параметров

Нет единой формулы, у каждого критерия своя.

TABLE E. Values of t at the .05, .01, and .001 Levels of Significance

df	Two-tailed			One-tailed			
	.05	.01	.001	.05	.01	.001	
1	12.706	63.657	636.619	1	6.314	31.821	318.309
2	4.303	9.925	31.599	2	2.920	6.965	22.327
3	3.183	5.841	12.924	3	2.353	4.541	10.215
4	2.777	4.604	8.610	4	2.132	3.747	7.173
5	2.571	4.032	6.869	5	2.015	3.365	5.893
6	2.447	3.707	5.959	6	1.943	3.143	5.208
7	2.365	3.500	5.408	7	1.895	2.998	4.785
8	2.306	3.355	5.041	8	1.860	2.897	4.501
9	2.262	3.250	4.781	9	1.833	2.821	4.297
10	2.228	3.169	4.587	10	1.813	2.764	4.144
11	2.201	3.106	4.437	11	1.796	2.718	4.025
12	2.179	3.055	4.318	12	1.782	2.681	3.930
13	2.160	3.012	4.221	13	1.771	2.650	3.852
14	2.145	2.977	4.141	14	1.761	2.625	3.787
15	2.132	2.947	4.073	15	1.753	2.603	3.733
16	2.120	2.921	4.015	16	1.746	2.584	3.686
17	2.110	2.898	3.965	17	1.740	2.567	3.646
18	2.101	2.879	3.922	18	1.734	2.552	3.611
19	2.093	2.861	3.883	19	1.729	2.540	3.579
20	2.086	2.845	3.850	20	1.725	2.528	3.552
21	2.080	2.831	3.819	21	1.721	2.518	3.527
22	2.074	2.819	3.792	22	1.717	2.508	3.505
23	2.069	2.807	3.768	23	1.714	2.500	3.485
24	2.064	2.797	3.745	24	1.711	2.492	3.467
25	2.060	2.787	3.725	25	1.708	2.485	3.450
26	2.056	2.779	3.707	26	1.706	2.479	3.435
27	2.052	2.771	3.690	27	1.703	2.473	3.421
28	2.048	2.763	3.674	28	1.701	2.467	3.408
29	2.045	2.756	3.659	29	1.699	2.462	3.396
30	2.042	2.750	3.646	30	1.697	2.457	3.385
40	2.021	2.705	3.551	40	1.684	2.423	3.307
50	2.009	2.678	3.496	50	1.676	2.403	3.261
60	2.000	2.660	3.460	60	1.671	2.390	3.232
70	1.994	2.648	3.435	70	1.667	2.381	3.211
80	1.990	2.639	3.416	80	1.664	2.374	3.195
90	1.987	2.632	3.402	90	1.662	2.369	3.183
.00	1.984	2.626	3.391	100	1.660	2.364	3.174
∞	1.960	2.576	3.292	∞	1.645	2.327	3.091

adapted from A. L. Sockloff and J. N. Edney, Some extension of Student's t and Pearson's r central distributions, Technical Report (May 1972), Measurement and Research Center, Temple University, Philadelphia.

План лекции

- Что такое А/Б-тестирование
- Основные понятия и предпосылки
- Общая схема подготовки А/Б-теста
- Статистические критерии
- Мощность и корректность

Статистические критерии

Параметрические критерии основаны на том, что распределение данных известно. То есть, при применении какого-нибудь параметрического критерия нужно всегда следить за тем, что главное допущение критерия – тип распределения – выполняется. Как правило, многие параметрические критерии предполагают нормальность распределения данных.

Непараметрические критерии исходят из того, что распределение данных неизвестно. Поэтому при использовании этих критериев часто действия производятся не с самими значениями в выборке/выборках, а с их рангами/частотами.

В общем случае параметрические критерии обладают большей мощностью

Z-test для долей

Условия применения:

- Выполняется ЦПТ для средних (распределения асимптотически нормальны)
- Большой размер выборки
- Подходит для случайных величин с распределением Бернулли

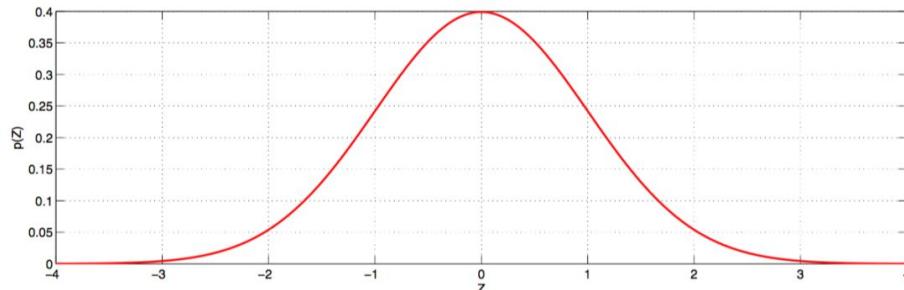
выборка: $X^n = (X_1, \dots, X_n), X \sim Ber(p)$

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: $N(0, 1)$



Z-test для долей. Пример

Задача

Пропорция любителей Доброколы в прошлом году была 20%. По опросам 500 людей в этом году получилась доля в 30%.

Изменился ли значимо (уровень ошибки 5%) процент любителей Доброколы?

Вычисления

$$n = 500, p_0 = 0.2, p = 0.3 \quad Z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.3 - 0.2}{\sqrt{\frac{0.2(1-0.2)}{500}}} = 5.59$$
$$Z_{0.05} = \pm 1.96 < 5.59$$

Итог

Процент значимо изменился

Z-test для разности долей (независимые выборки)

Условия применения:

- Выполняется ЦПТ для средних (распределения асимптотически нормальны)
- Большой размер выборки
- Подходит для случайных величин с распределением Бернулли
- Группы (подвыборки) независимы

Пример расчета критерия

выборки:

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

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

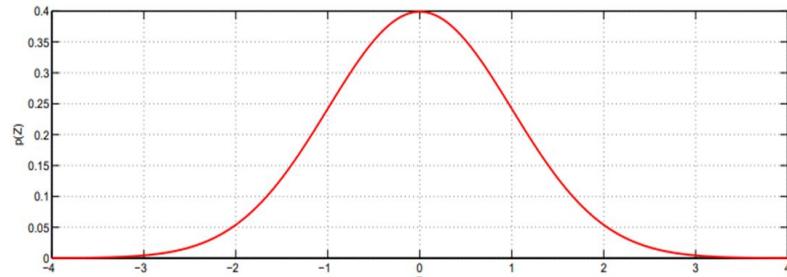
$$p_1 = \frac{\mathbb{E}A}{n_1}, \quad \hat{p}_1 = \frac{a}{n_1}, \quad p_2 = \frac{\mathbb{E}B}{n_2}, \quad \hat{p}_2 = \frac{b}{n_2};$$

$$H_0: p_1 = p_2;$$

$$H_1: p_1 \neq p_2;$$

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

$$Z(X_1^{n_1}, X_2^{n_2}) \sim N(0, 1) \text{ при } H_0.$$



Z-test для разности средних (независимые выборки)

Условия применения:

- Выполняется ЦПТ для средних (распределения асимптотически нормальны)
- Большой размер выборки
- Группы (подвыборки) независимы
- Известно стандартное отклонение генеральной совокупности

Критерий является асимптотическим

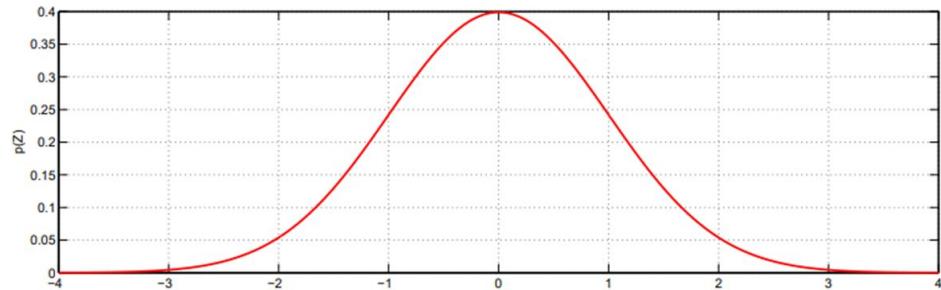
выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2),$
 $X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma_2^2),$
 σ_1, σ_2 известны;

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

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

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

$Z(X_1^{n_1}, X_2^{n_2}) \sim N(0, 1)$ при $H_0.$



t-test для разности средних (независимые выборки)

Условия применения:

- Предположение о нормальности выборки
- Может быть малый размер выборки
- Группы (подвыборки) независимы
- Неизвестно стандартное отклонение генеральной совокупности
- Предположение о равенстве дисперсий в группах (но есть модификация – t-test Уэлша)

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

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

$T(X_1^{n_1}, X_2^{n_2}) \approx St(\nu)$ при H_0 .

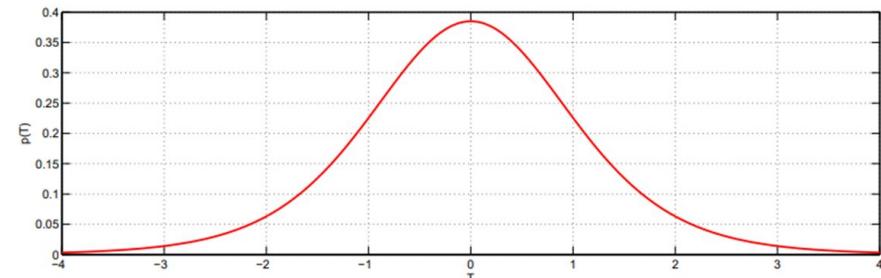
выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma^2)$,
 $X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma^2)$,
 σ неизвестна;

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

альтернатива: $H_1: \mu_1 \neq \mu_2$;

статистика: $T(X_1^{n_1}, X_2^{n_2}) = \frac{\bar{X}_1 - \bar{X}_2}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$,
 $S = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$;

$T(X_1^{n_1}, X_2^{n_2}) \sim St(n_1 + n_2 - 2)$ при H_0 .



Пример расчета критерия

t-test для разности средних (зависимые выборки)

Условия применения:

- Скоррелированность сравниваемых значений
- Предположение о нормальности распределения выборки

выборки: $X_1^n = (X_{11}, \dots, X_{1n}), X_1 \sim N(\mu_1, \sigma_1^2),$
 $X_2^n = (X_{21}, \dots, X_{2n}), X_2 \sim N(\mu_2, \sigma_2^2),$

выборки связанные;

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

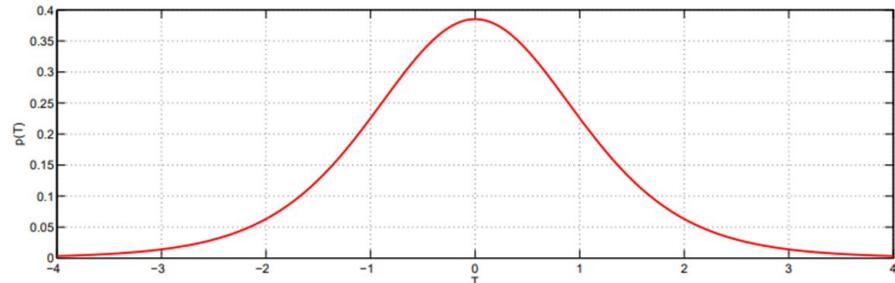
альтернатива: $H_1: \mu_1 < \neq > \mu_2;$

статистика: $T(X_1^n, X_2^n) = \frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}},$

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2},$$

$$D_i = X_{1i} - X_{2i};$$

$$T(X_1^n, X_2^n) \sim St(n-1) \text{ при } H_0.$$



Хи-квадрат критерий для дисперсии

выборка: $X^n = (X_1, \dots, X_n), X \sim N(\mu, \sigma^2)$;

нулевая гипотеза: $H_0: \sigma = \sigma_0$;

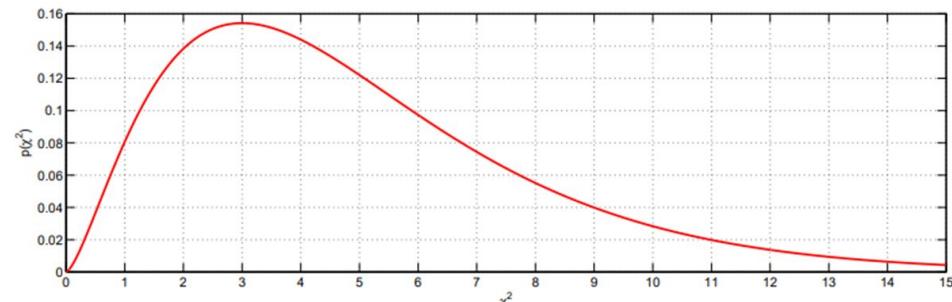
альтернатива: $H_1: \sigma < \neq > \sigma_0$;

статистика: $\chi^2(X^n) = \frac{(n-1)S^2}{\sigma_0^2}$;

$\chi^2(X^n) \sim \chi_{n-1}^2$ при H_0 ;

Условия применения:

- Предположение о нормальности выборки
- Известен параметр μ



Пример расчета критерия

F-критерий Фишера для отношения дисперсий (независимые выборки)

Условия применения:

- Предположение о нормальности выборки
- Подходит для сравнения дисперсий двух выборок

Неустойчив к отклонениям от нормальности!

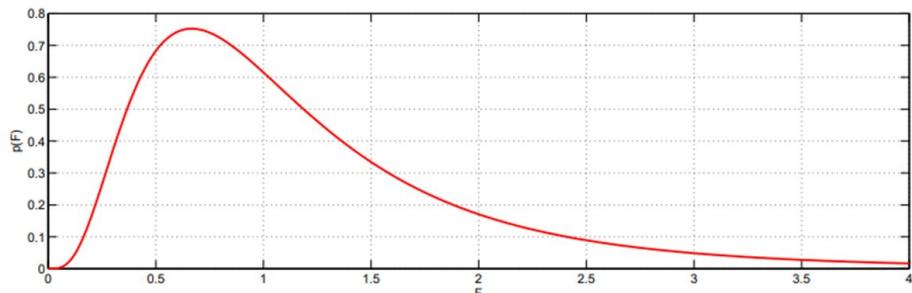
выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2),$
 $X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma_2^2);$

нулевая гипотеза: $H_0: \sigma_1 = \sigma_2;$

альтернатива: $H_1: \sigma_1 < \neq > \sigma_2;$

статистика: $F(X_1^{n_1}, X_2^{n_2}) = \frac{S_1^2}{S_2^2};$

$F(X_1^{n_1}, X_2^{n_2}) \sim F(n_1 - 1, n_2 - 1)$ при $H_0.$



Пример расчета критерия

U-критерий Манна-Уитни (независимые выборки)

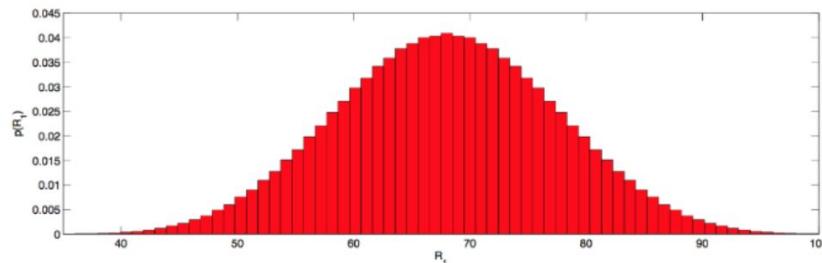
Условия применения:

- Группы однородны по независимым переменным
- Группы могут быть небольшими, но хотя бы по 20 наблюдений в каждой
- Повторяющихся значений практически нет
- Форма распределения в группах должна быть схожей
- Наблюдения независимы
(для зависимых – критерий Уилкоксона,
[пример расчета](#))

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

выборки независимые
 нулевая гипотеза: $H_0: F_{X_1}(x) = F_{X_2}(x)$
 альтернатива: $H_1: F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta < \neq > 0$
 статистика: $X_{(1)} \leq \dots \leq X_{(n_1+n_2)}$ — вариационный ряд
 объединённой выборки $X = X_1^{n_1} \cup X_2^{n_2}$
 $R_1(X_1^{n_1}, X_2^{n_2}) = \sum_{i=1}^{n_1} \text{rank}(X_{1i})$

нулевое распределение: табличное



[Пример расчета критерия](#)

U-критерий Манна-Уитни (независимые выборки)

- В выборочных данных не должно быть совпадающих значений или таких совпадений должно быть очень мало (до 10)
- Подходит для любого распределения

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - R_2$$



$$U = \min\{U_1, U_2\}$$

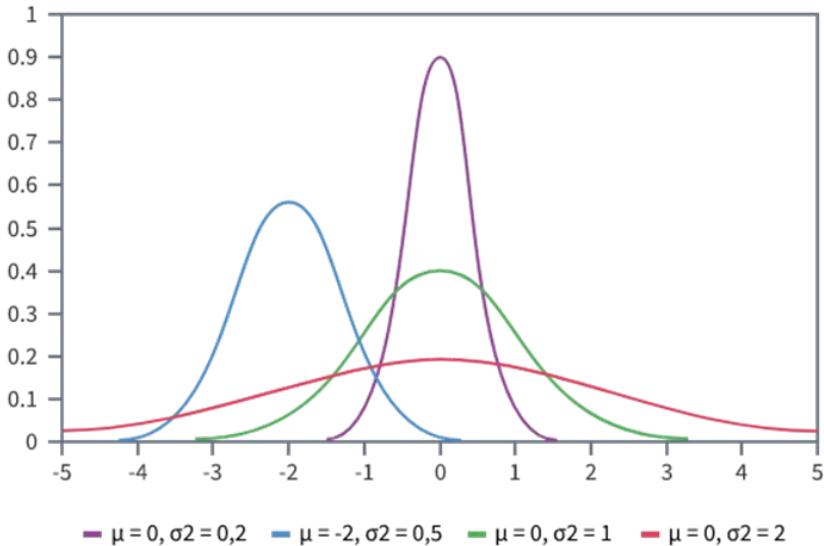
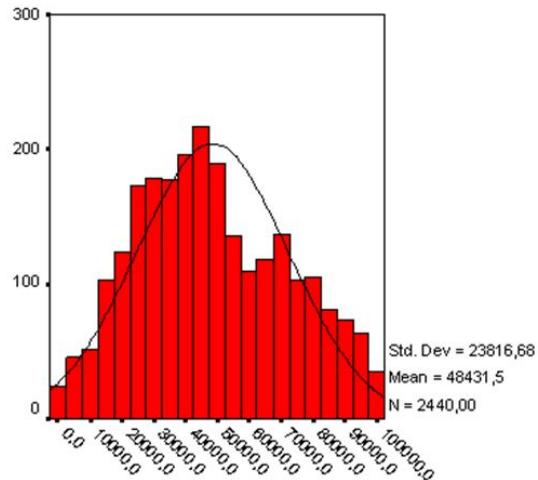
Если $U > U_{кр} \Rightarrow H_0$ НЕ отклоняется

Critical Values for the Mann-Whitney U-Test																														
Level of significance: 5% (P = 0.05)																														
Size of the largest sample (n ₂)																														
3	0	1	1	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	11	11	12	13	13	13	13	13		
4	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	17	18	19	20	21	22	23				
5	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20	22	23	24	25	27	28	29	30	32	33				
6	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27	29	30	32	33	35	37	38	40	42	43					
7		8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48	50	52	54					
8		13	15	17	19	22	24	26	29	31	34	36	38	41	43	45	48	50	53	55	57	60	62	65						
9		17	20	23	26	28	31	34	37	39	42	45	48	50	53	56	59	62	64	67	70	73	76							
10		23	26	29	33	36	39	42	45	48	52	55	58	61	64	67	71	74	77	80	83	87								
11		30	33	37	40	44	47	51	55	58	62	65	69	73	76	80	83	87	90	94	98									
12		37	41	45	49	53	57	61	65	69	73	77	81	85	89	93	97	101	105	109										
13			45	50	54	59	63	67	72	76	80	85	89	94	98	102	107	111	116	120										
14			55	59	64	67	74	78	83	88	93	98	102	107	112	118	122	127	131											
15			64	70	75	80	85	90	96	101	106	111	117	122	125	132	138	143												
16			75	81	86	92	98	103	109	115	120	126	132	138	143	149	154													
17			87	93	99	105	111	117	123	129	135	141	147	154	160	166														
18			99	106	112	119	125	132	138	145	151	158	164	171	177															
19			113	119	126	133	140	147	154	161	168	175	182	189																
20			127	134	141	149	156	163	171	178	186	193	200																	
21			142	150	157	165	173	181	188	196	204	212																		
22			158	166	174	182	191	199	207	215	223																			
23			175	183	192	200	209	218	226																					
24			192	201	210	219	228	238	247																					
25			211	220	230	239	249	258																						
26			230	240	250	260	270																							
27			250	261	271	282																								
28			272	282	293																									
29			294	305																										
30			317																											

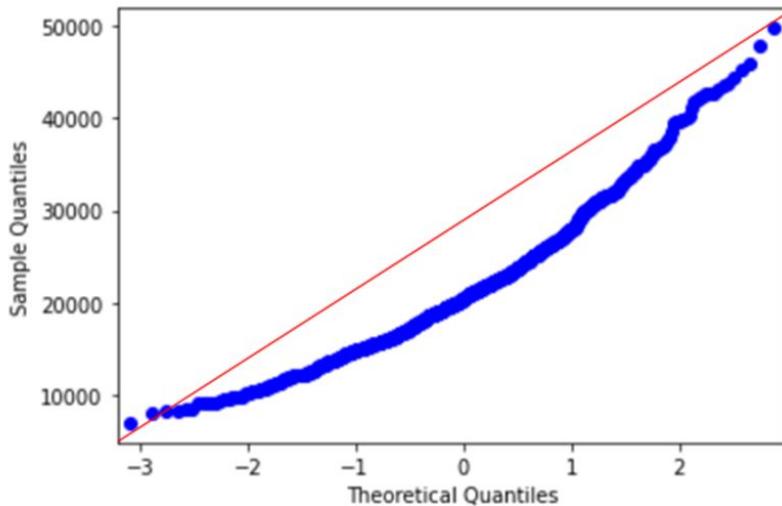
Как проверить нормальность распределения?

Критерий согласия - проверка на согласие подразумевает проверку предположения о том, что исследуемая случайная величина подчиняется предполагаемому закону.

- Q-Q график
- Тест Колмогорова-Смирнова
- Тест Шапиро-Уилка



Q-Q график (quantile-quantile plot)



Как построить:

- Упорядочиваем нашу выборку
- Считаем экспериментальные квантили
- Считаем теоретические квантили
- Строим график с координатами: (теоретическое значение, экспериментальное значение)

Как трактовать:

- Точки совпали с прямой => идеальный вариант нормального распределения
- Точки кривой выше прямой => в выборке завышаем значения относительно нормального распределения
- Точки ниже прямой => значения в выборке ниже, чем должны быть при нормальном распределении
- Справа от нуля – область от середины до хвоста с максимальными значениями, слева – от середины до минимальных

Критерий Колмогорова-Смирнова

Критерий согласия Колмогорова используется для проверки простых гипотез о принадлежности анализируемой выборки некоторому полностью известному закону распределения

Критерий однородности Смирнова используется для проверки гипотезы о принадлежности двух независимых выборок одному закону распределения, то есть о том, что два эмпирических распределения соответствуют одному и тому же закону.

В основе критерия - **статистика Колмогорова-Смирнова**, которая является оценкой расстояния между эмпирической выборочной функцией распределения и кумулятивной функцией теоретического распределения, либо между эмпирическими функциями распределения двух выборок

В случае двух выборок распределение, рассматриваемое в рамках нулевой гипотезы, должно быть непрерывным

Критерий Колмогорова-Смирнова

$$X_1^{n1} = (X_{11}, \dots, X_{1n1})$$

$$X_2^{n2} = (X_{21}, \dots, X_{2n2})$$

$$H_0: F_{X1}(x) = F_{X2}(x)$$

$$H_A: F_{X1}(x) \neq F_{X2}(x)$$

$F_{n1X1}(x)$, $F_{n2X2}(x)$ - эмпирические функции распределения, построенные по выборкам X_1^{n1} и X_2^{n2}

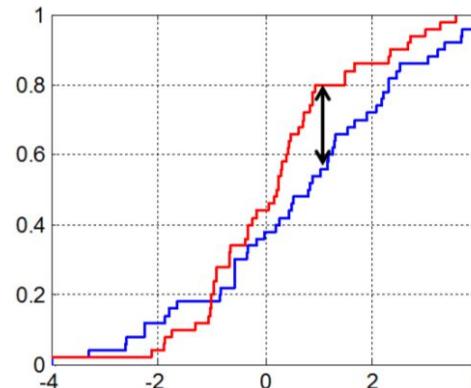
$$\lambda_{\text{наблюдаемое}} = D(X_1^{n1}, X_2^{n2}) * \sqrt{\frac{n_x \cdot n_y}{n_x + n_y}}$$

Пример расчета критерия

Статистика Колмогорова

$$D(X_1^{n1}, X_2^{n2}) = \sup_{-\infty < x < +\infty} |F_{n1X1}(x) - F_{n2X2}(x)|$$

При справедливости H_0 распределение статистики D_n будет одинаковым для любых непрерывных распределений



Критерий Колмогорова-Смирнова

Критические числа Колмогорова-Смирнова

Степень свободы <i>N</i>	Проверка единичной выборки *			Проверка двух выборок **	
	<i>D_{0,10}</i>	<i>D_{0,05}</i>	<i>D_{0,01}</i>	<i>D_{0,05}</i>	<i>D_{0,01}</i>
1	0,950	0,975	0,995	—	—
2	0,776	0,842	0,929	—	—
3	0,642	0,708	0,828	—	—
4	0,564	0,624	0,733	1,000	1,000
5	0,510	0,565	0,669	1,000	1,000
6	0,470	0,521	0,618	0,833	1,000
7	0,438	0,486	0,577	0,857	0,857
8	0,411	0,457	0,543	0,750	0,875
9	0,388	0,432	0,514	0,668	0,778
10	0,368	0,410	0,490	0,700	0,800
11	0,352	0,391	0,468	0,636	0,727
12	0,338	0,375	0,450	0,583	0,667
13	0,325	0,361	0,433	0,538	0,692
14	0,314	0,349	0,418	0,571	0,643

Наблюдаемое значение сравнивается с критическими из таблицы:

- Если наблюдаемое значение БОЛЬШЕ критического, то различия считаются ЗНАЧИМЫМИ
- Если наблюдаемое МЕНЬШЕ критического, то нет оснований отвергать нулевую гипотезу => оба эмпирических распределения соответствуют одному закону распределения

Свойства

- Не требует ни равенства дисперсии, ни одинакового размера групп
- Хорошо подходит для проверки двух групп на однородность по статическим признакам
- Не подходит для оценки эффекта в А/В тестах

План лекции

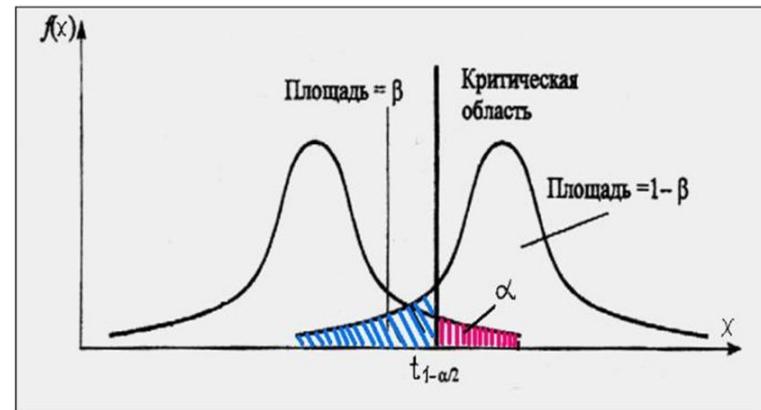
- Что такое А/Б-тестирование
- Основные понятия и предпосылки
- Общая схема подготовки А/Б-теста
- Статистические критерии
- Мощность и корректность

Мощность и корректность

- **Корректность** – вероятность отклонить нулевую гипотезу, когда она верна => вероятность не допустить ошибку 1 рода ($1 - \alpha$)
 - **Мощность** - вероятность отклонить нулевую гипотезу, когда нулевая гипотеза ложна => вероятность не допустить ошибку 2 рода ($1 - \beta$)
 -
- Как правило, берут значения $\alpha = 0,05$ и $\beta = 0,2$

Задача проверки гипотез несимметрична:

- Ограничиваем вероятность ошибки первого рода α
- Минимизируем ошибку второго рода
- Снижая ошибку 1-го рода, мы повышаем ошибку 2-го



Общая схема подготовки к А/Б



Свойства метрики

- Чувствительность метрики — метрика должна реагировать на изменение
- Надежность — если не было никаких изменений, то метрика не меняется

Как проверить?

A/A-тест - тест с разбивкой на группы, но без изменений функционала. Таким образом проверяем, как ведет себя метрика в «естественной» среде, насколько сильно она колеблется и каковы различия между группами.

- В случае надежности метрики ожидается, что A/A-тест не покажет статзначимых различий
- На похожей на A/A-тест механике также проверяется корректность работы выбранного критерия

Оценка корректности

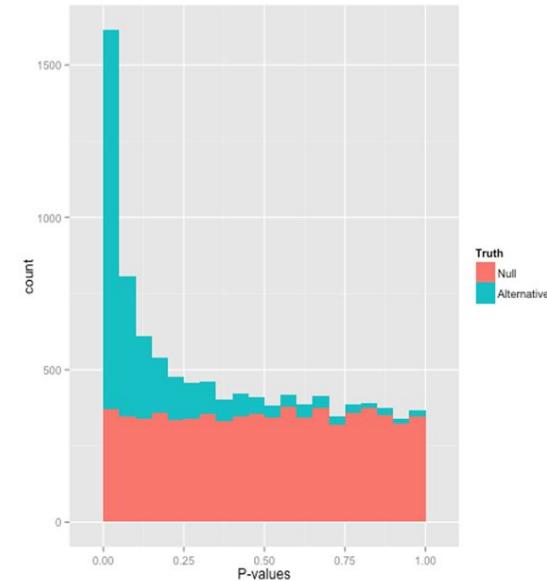
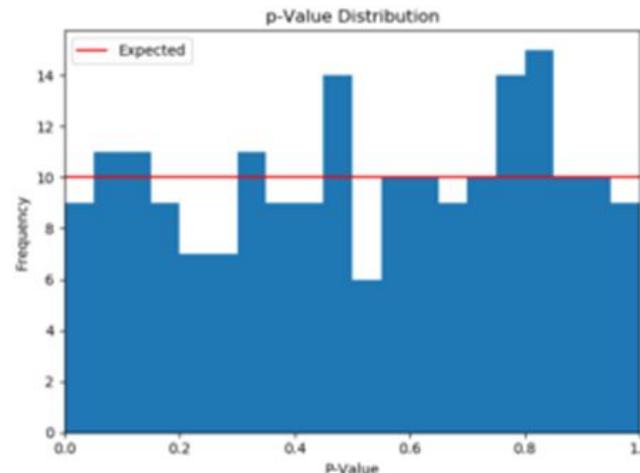
- На исторических данных выбираем N пар похожих групп (либо рандомно разбиваем выборку на группы несколько раз)
- Сравниваем выбранным стат-критерием наши группы в каждой паре и запоминаем p-value
- Считаем процент пар, в которых $p\text{-value} \leq \alpha$: т.е считаем в скольких случаях критерий (сработал) нашел различия, когда его на самом деле нет.

При проверке разбиения на группы перед A/B-тестом в ходе истинного A/A-теста используется похожая логика, но тест проводят не итеративно

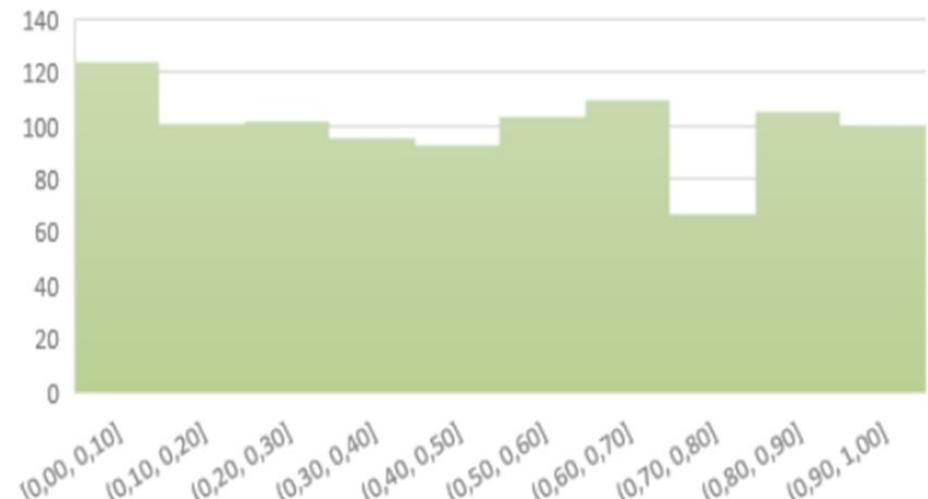
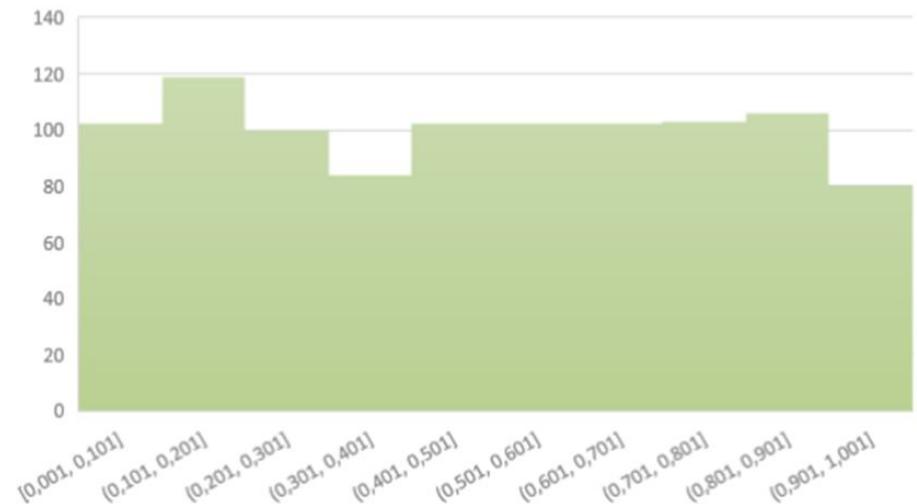
Оценка корректности: распределение p-value

Распределение p-value, которое мы получили после оценки корректности на исторических данных, должно быть распределено равномерно

Иначе выводы будут смещены, критерий и А/В тест будут некорректными



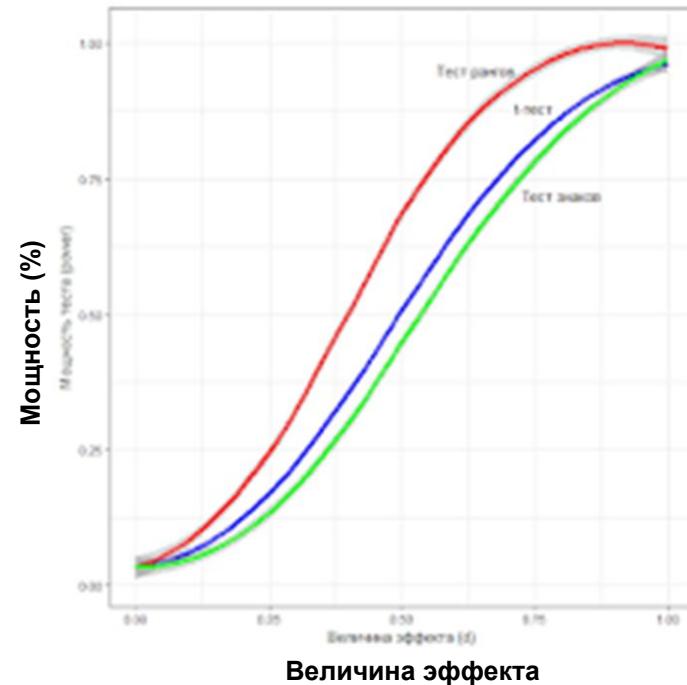
Оценка корректности: распределение p-value



Оценка мощности

Что влияет:

- Размер выборки (с увеличением выборки уменьшается стандартная ошибка, а следовательно, увеличивается мощность)
- Величина ожидаемого эффекта (больше эффект – выше мощность)
- Чувствительность статистики критерия
- Уровень α



Оценка мощности

Алгоритм:

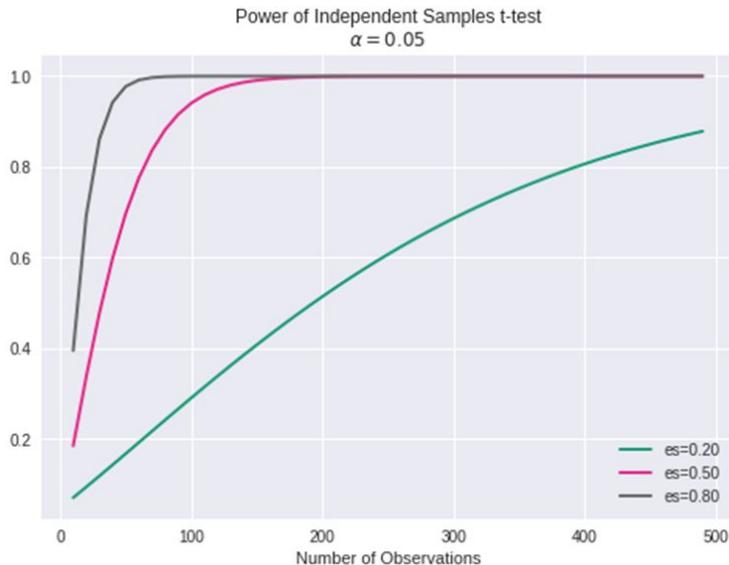
- Для разных сочетаний размера выборки, ожидаемого эффекта и стат.критерия повторяем следующую процедуру:
- Выделяем N пар похожих групп из выборки (или же N раз перемешиваем выборку и разбиваем на количество групп)
- Добавляем разные эффекты размером $x\%$ в каждую пару в одну из групп
- Сравниваем стат. критерием наши группы в каждой паре и запоминаем p-value
- Считаем процент пар, в которых $p\text{-value} \leq \alpha$: т.е считаем в скольких случаях критерий сработал или нашел различия, когда оно на самом деле есть

Как добавить эффект:

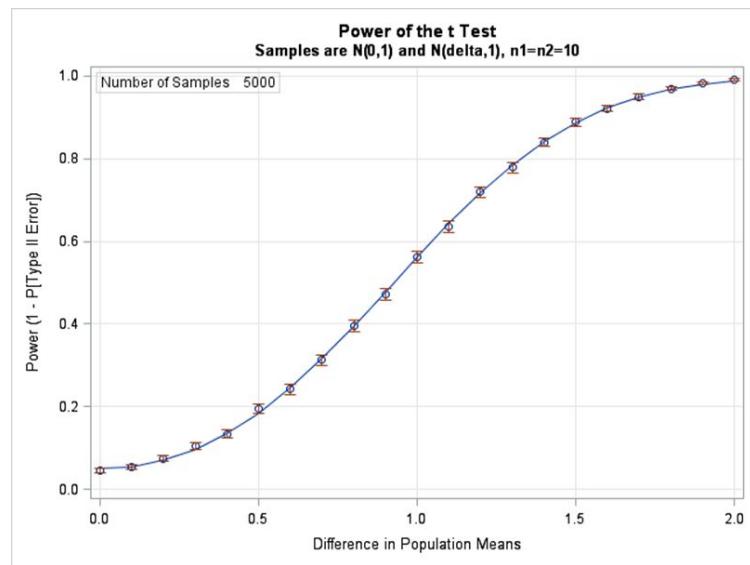
- Сдвинуть наше распределение на определенную константу
- Сдвинуть наше распределение на определенную константу со случайным шумом
- Сдвинуть распределение, смешав его с другим распределением

Оценка мощности

Разные размеры выборок и разные эффекты



Однаковые размеры выборок и разные эффекты



Выводы

- А/В-тест всегда должен основываться на гипотезе, которая нуждается в проверке.
- Не стоит объединять несколько идей в одну гипотезу, а также ограничивайте переменные, вводимые в тест, чтобы вы могли понять их индивидуальное влияние.
- Выбор метрики зависит от лежащей в основе гипотезы, которая проверяется с помощью этого А/В-теста. Метрика и гипотеза определяют, как будет разработан тест, а также насколько хорошо работают предложенные идеи
- Для корректного выбора критерия важно учесть характер данных, эксперимента, а также условия применения критерия
- При выборе критерия необходимо оценить его мощность и корректность, оценить распределение p -value
- После оценки мощности становится возможным выбор наиболее подходящего сочетания параметров: критерий, эффект, размер выборки, продолжительность теста

Спасибо за внимание!