И долго мне ждать?

Команда «даже для нас это большая лужа»

Датасет hh.ru

500 000 строк

Каждая строка - информация про взаимодействие работодателя и претендента на вакансию

hh.ru — один из самых крупных сайтов по поиску работы и сотрудников в мире (по данным рейтинга Similarweb)

Датасет

Структура данных

Количественные данные	Качественные данные
• год рождения соискателя • желаемый уровень зарплаты в рублях • кол-во месяцев рабочего опыта • предлагаемая зарплата вакансии в рублях, от • предлагаемая зарплата вакансии в рублях, до	• дата создания коммуникации • первоначальный статус взаимодействия • финальный статус взаимодействия • дата создания резюме • профессия • пол • город размещения резюме • уровень образования соискателя • отношение к переезду по работе • готовность к командировкам • предпочитаемый график работы • предпочитаемый тип занятости • набор навыков из резюме • дата создания вакансии • город размещения вакансии • предлагаемый график работы • тип занятости из вакансии • набор требуемых навыков из вакансии

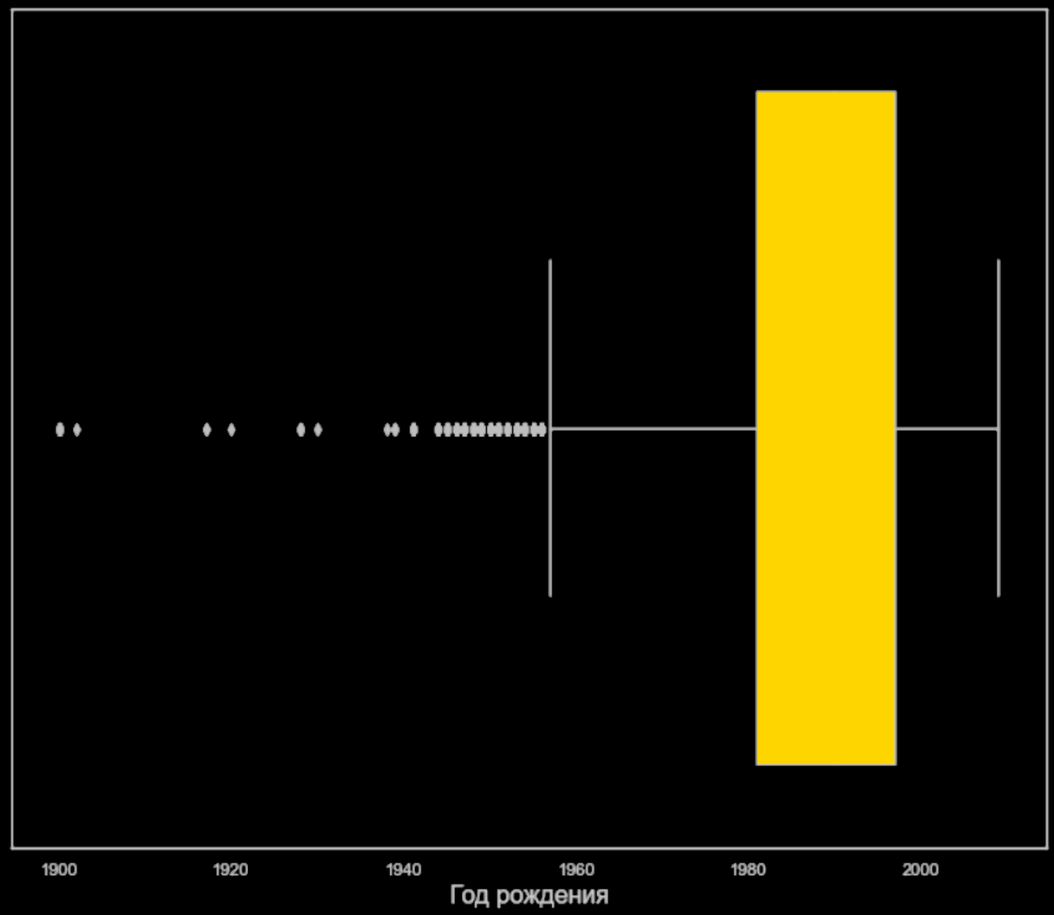
Предварительный анализ

	year_of_birth	expected_salary	work_experience_months	compensation_from	compensation_to
count	500 000	489 214	500 000	470 618	361 775
mean	1988.33	97 074.31	107.90	69 245.91	81 808.27
std	11.12	4 725 634	96.12	47193.40	197 702.10
min	1900	1	0	0	0
25 %	1981	40 000	32	40 000	35 000
50 %	1990	60 000	85	60 000	70 000
75 %	1997	80 000	160	85 000	110 000
max	2009	996 482 600	1273	1 600 000	100 000 000

Предварительный анализ

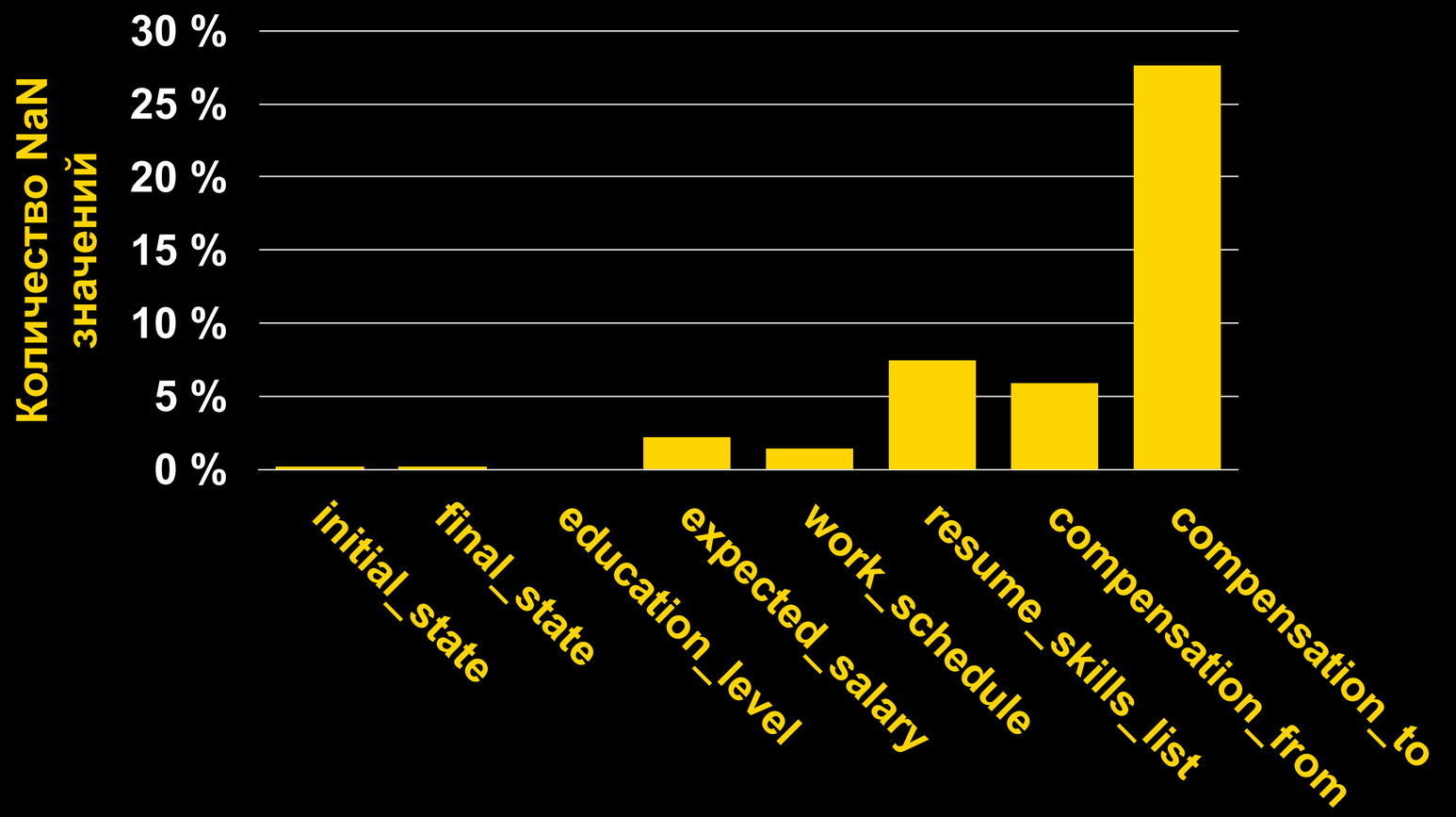
Проверка и удаление выбросов с помощью boxplot

- Год рождения
- Ожидаемая зарплата



Предварительный анализ

Удаление **NaN** значений в тех столбах, где их **меньше 5%**



Предварительный анализ

Дубликаты удалены

NaN-значения почищены

Выбросы удалены

Отрицательные значения удалены

Итого

451 361

Строки

Исследовательский вопрос и гипотеза

Как неопределенность в предполагаемой зарплате влияет на время ожидания отклика на вакансию?

Работодатель зачастую не указывает конкретную зарплату в вакансии, надеясь сэкономить на ней

Соискатель работы не доверяет вакансии и ищет варианты получше

Компаниям, у которых не указана конкретная зарплата на вакансии, приходится дольше ждать отклика соискателей

Исследуемые показатели

Наличие/отсутствие конкретной зарплаты на вакансии

Указал ли работодатель обе границы предполагаемой зарплаты или нет

Влияние

Время ожидания отклика на вакансию

Время прошедшие с момента создания вакансии и первого взаимодействия по этой вакансией

Метод исследования

1

Разбитие на 2 выборки

A - есть **NaN** значения в полях **compensation_to** и/или **compensation_from**

B - нет **NaN** значений в полях **compensation_to** и **compensation_from**

2

Создание необходимых переменных

time_waiting_for_response

3

Тесты данных

Тест Колмогорова-Смирнова

U-критерий Манна-Уитни

Очистка данных

Очистка от выбросов по boxplot

В выборке B

compensation_to и
compensation_from

В выборках A и B

time_waiting_for_response

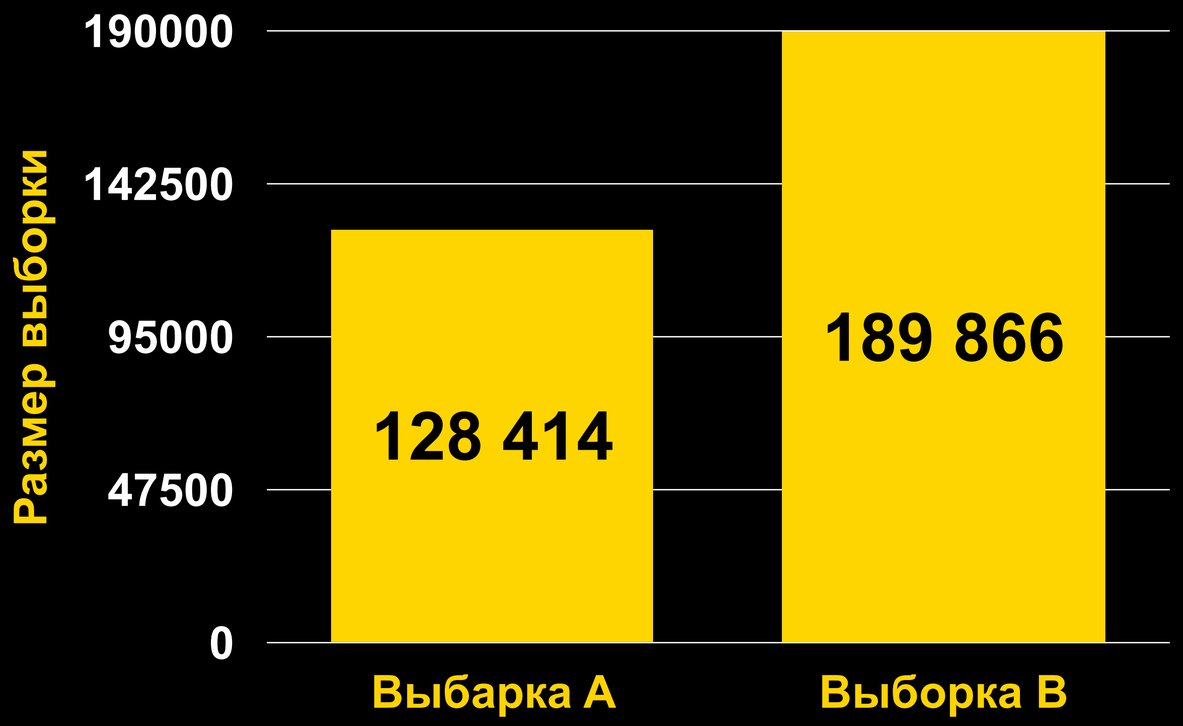
Проверка данных на нормальность

Одновыборочный критерий согласия Колмогорова

	p_value
Выборка A	0.0
Выборка B	0.0

Данные в обеих выборках не подчиняются нормальному распределению

Исследуемые показатели

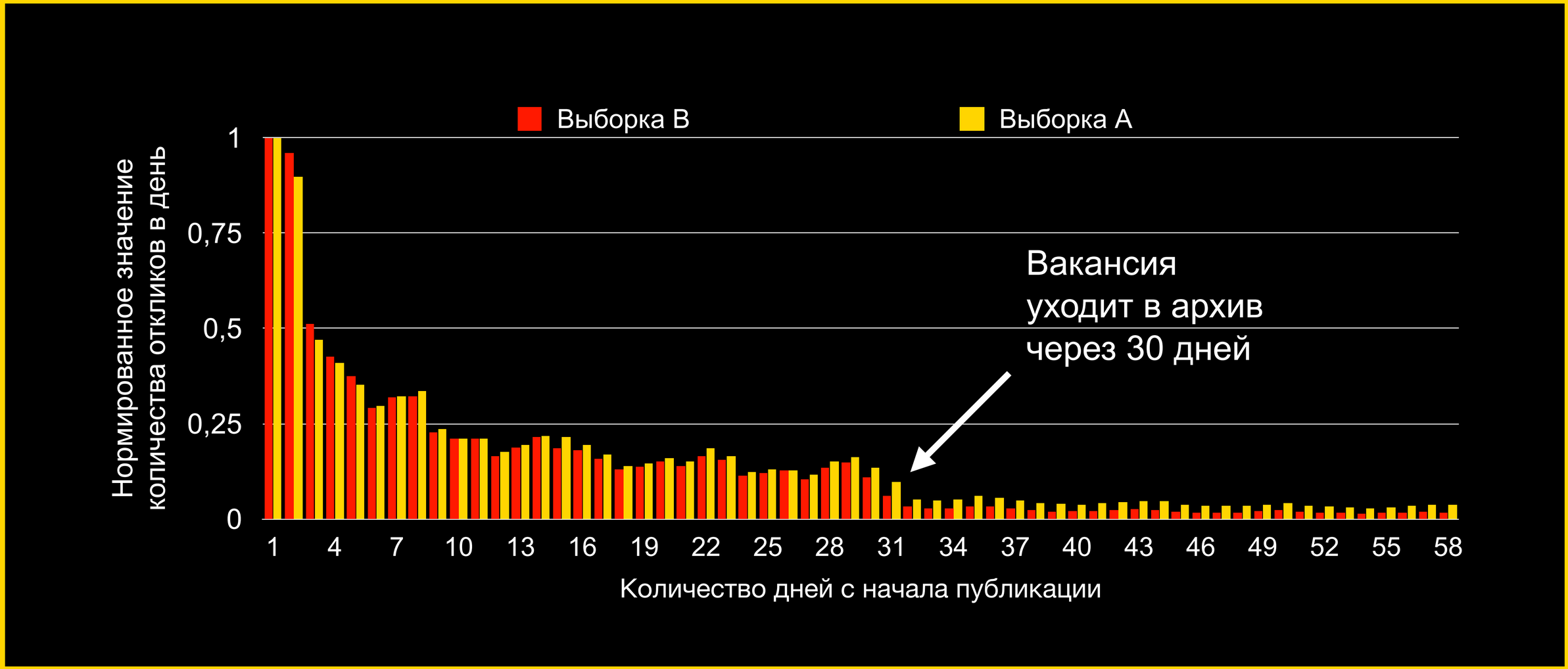


Итого

318 280

Строк для исследования

Исследуемые показатели



Анализ данных

1

Для каждой из выборок подсчитывалось количество наблюдений с i-тым днем

2

Для каждого из массивов обсчитывался критерий Колмогорова-Смирнова и Манна-Уитни и выводился p-value

scipy.stats.mannwhitney(a, b)

scipy.stats.ks_2samp(a, b)

Пороговым значением **p-value**, при котором отвергалась нулевая гипотеза о принадлежности данных одному распределению, было принято общепринятое значение **0.05**

Результаты анализа

	1 месяц	2 месяца
U-критерий Манна-Уитни p_value	0.522	0.0276
Тест Колмогорова-Смирнова p_value	0.963	0.000184

Устойчивость модели

Разделение выборки **B** на три равные выборки по упорядоченному уровню зарплаты

+

Перекрестная проверка при помощи теста Колмогорова-Смирнова и критерия Манна-Уитни

Итог:

P-value не отличается от первоначальной выборки

Вывод

Гипотеза о том, что компаниям, у которых не указана минимальная и/или максимальная предполагаемая зарплата, приходится дольше ждать отклика

Отвергается за временной отрезок, равный **одному месяцу**

Принимается на более широком диапазоне времени

Практическая польза

1

Создание уникальной формулы привлекательной вакансии

2

Уточнение исследований проведенных hh.ru

3

Прогнозирование времени ожидания отклика на вакансию

Policy implication

Справочная информация и помощь в создании привлекательной вакансии

1. Государство заинтересовано в уменьшении количества безработных людей
2. Компании/работодатели заинтересованы в быстром отклике на вакансию
3. hh.ru заинтересован в том, чтобы поддерживать репутацию платформы для быстрого поиска профессии

Ограничения исследования

Масштаб

Исследование нельзя масштабировать на мир

Временное ограничение

Исследование нельзя масштабировать на весь год

Неполный датасет

Отсутствие всех видов профессий для более достоверного анализа

Вакансии без отклика

Отсутствие данных о вакансиях без отклика

Перспективы развития

Совместное исследование

Сотрудничество с **hh.ru** и публикация статьи по теме

Проверка устойчивости

Проверить устойчивость на подгруппах других параметров

Ссылки на исследования и статьи hh.ru

- “Почему работодатели не указывают зарплату?” *Hh.ru*, hh.ru/article/2071
- “Почему в вакансиях не всегда указывают зарплату.” *Hh.ru*, hh.ru/article/23913
- “Работайте с вакансиями в архиве.” *Hh.ru*, hh.ru/article/26128

Наша команда

Глеб Филимонов

Программист, аналитик

Марк Калинин

Аналитик, тимлид, программист

Арсен Рябуха

Программист, дизайнер

Михаил Личманов

Мат. модель, аналитик

Иван Караулов

Мат. модель

Спасибо за внимание

Приложение 1: Результаты проверки на устойчивость

	1 месяц			2 месяца		
	Низкий уровень зарплаты	Средний уровень зарплаты	Высокий уровень зарплаты	Низкий уровень зарплаты	Средний уровень зарплаты	Высокий уровень зарплаты
U-критерий Манна-Уитни p_value	0.59	0.49	0.49	0.038	0.027	0.024
Тест Колмогорова-Смирнова p_value	0.99	0.82	0.82	0.0004	0.0002	3.23e-5

Приложение 2: Тест Колмогорова-Смирнова

1. Ищется максимальное расстояние между графиками, по оси абсцисс: для всех координат по X-у берётся модуль разности значений функций в этом X-е, и из всех таких значений берётся супремум
2. Эта величина умножается на корень из произведения размера выборок, делённого на сумму размеров выборок, и получается Kn - распределение Колмогорова
3. Далее по этому распределению строится график и считается p-value
4. P-value отвечает за схожесть данных в выборках

Приложение 3: U-критерий Манна-Уитни

Тест можно проводить не только с нормальными распределениями, в отличие, например, от t-теста

1. Тест собирает все значения обеих выборок
2. Располагает их в порядке возрастания
3. Привязывает к ним «ранги»
4. Считает сумму для каждой выборки
5. Сложные вычисления приводят к z-value, через которое потом считается p-value