НЕТОЛОГИЯ
групп

РАЗБОР КЕЙСОВ РЕАЛЬНЫХ БИЗНЕСОВ. ПОИСК ИНСАЙДОВ В ДАННЫХ

Максим Чикуров

Data Scientist и руководитель команды аналитики

Работал в компаниях Citibank, BNP Paribas, Barclays Bank, Teradata

maxim.chikurov@gmail.com

О ЧЕМ ПОГОВОРИМ И ЧТО СДЕЛАЕМ

План лекции

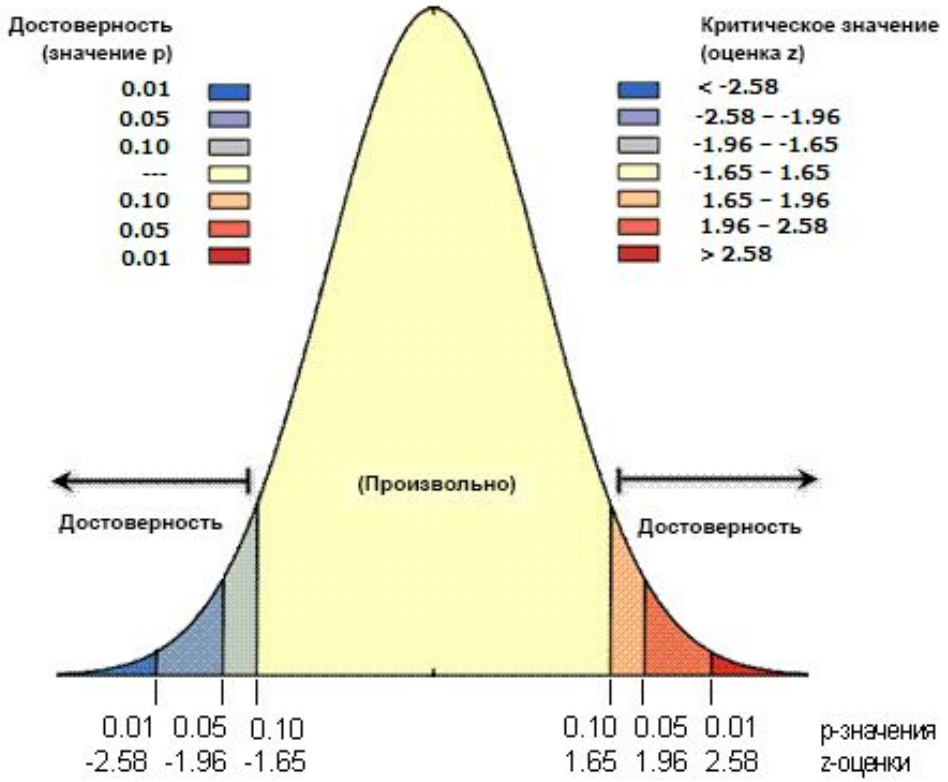
- Статистические тесты
- a/b тестирование
- Корреляция
- Практика в Gretl

Статистические тесты

Z-test

Применяется при проверке нулевой гипотезы о том, что математическое ожидание случайной величины равно некоторому значению m

Z-оценки являются стандартными отклонениями. Если, например, инструмент возвращает z-оценку +2.5, вы сказали бы, что результат – это 2.5 стандартных отклонений. И z-оценки, и p-значения связаны со стандартным нормальным распределением, как показано на картинке.



Z-test

Важно понимать!

Z-test - класс методов статистической проверки гипотез (статистических критериев), основанных на нормальном распределении.

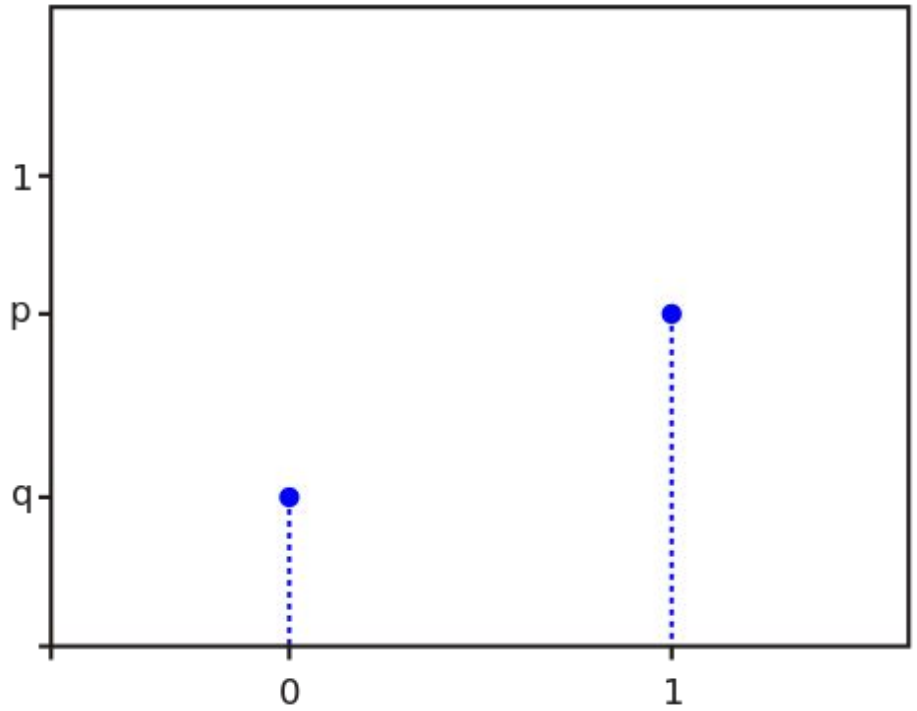
Продолжение кейса грузоперевозок

Известно, что наша компания испытывает существенные проблемы с задержками в доставке грузов. При этом компания конкурент распространила информацию о том, что якобы не более 70% грузов доставляются нашей компанией вовремя. Информация о времени доставки не фиксируется в информационных системах (CRM / ERP). Вы выбрали **100 случайных** накладных на перевозки и определи, что в 82 случаях грузы были доставлены во время.

Вопрос: Можем ли мы говорить о том, что конкурент распространяет ложную информация о нашей компании?

Распределение Бернулли

Случайная величина **X** имеет распределение Бернулли, если она принимает всего два значения: **1 и 0** с вероятностями **p** и **q = 1-p** соответственно.



Математическое ожидание (среднее) = **p**

Дисперсия $\nu = p \times q$

Стандартное отклонение $\sigma = \sqrt{p \times q}$

Решение задачи

Исходя из условий задачи определяем характеристики распределение генеральной совокупности при вероятности успеха 70%:

1. Это распределение Бернулли
2. Среднее (p) = 0,7
3. Стандартное отклонение (sigma) = 0,45

Решение задачи

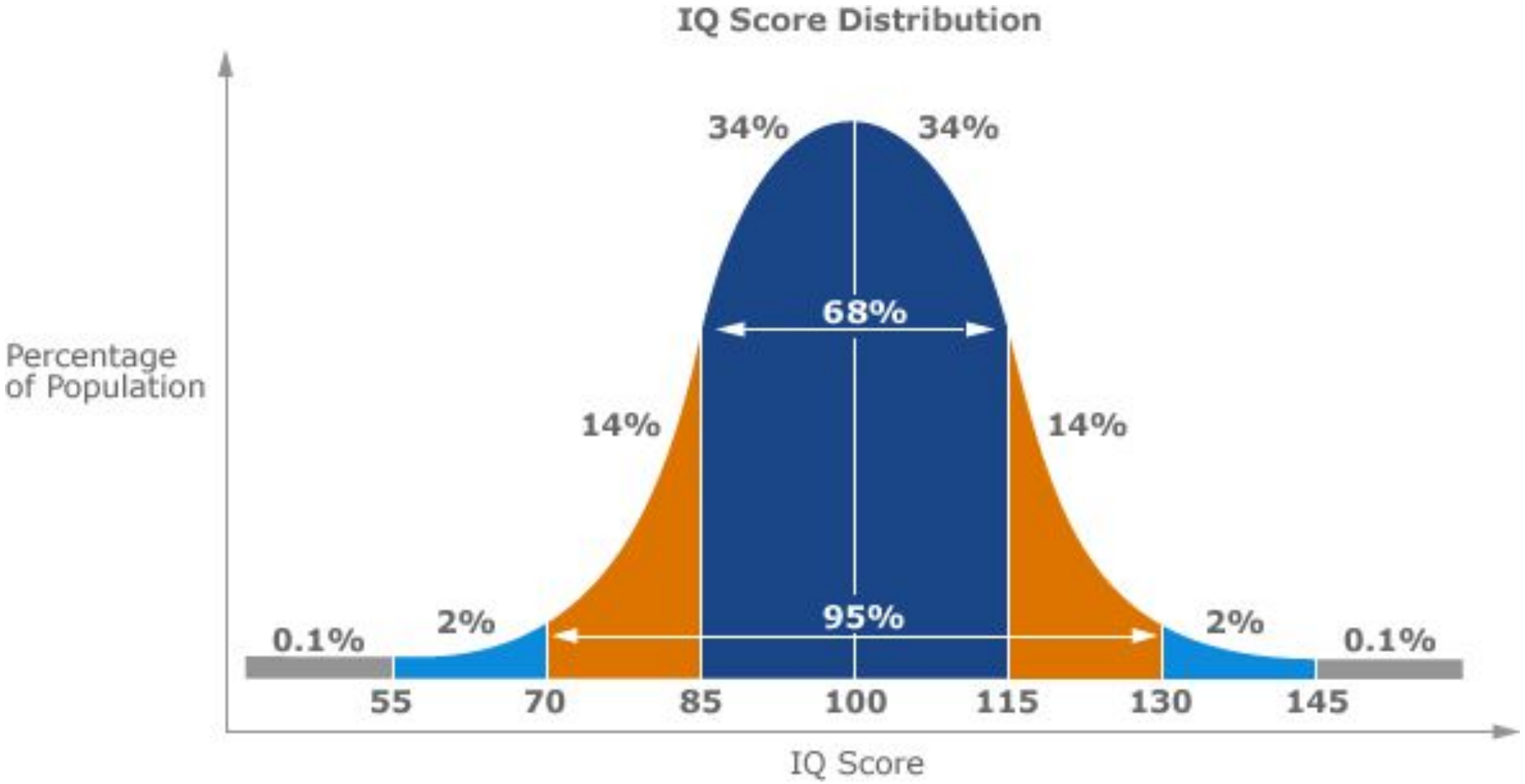
Для решения задачи необходимо определить вероятность получения выборки из 100 значений со средним 0,82. (p = 0.82)

$$SE = \frac{\sigma}{\sqrt{n}} \approx \frac{0,45}{10} \approx 0,05$$

$$Z = \frac{0,82 - 0,7}{0,05} = 2,4$$

Такое Z, соответствует P-value = 0,99. Что следует интерпретировать как вероятность получения выборки из 100 значений со средним меньше 0,82.

ПРАВИЛО ДВУХ СИГМ



A/B ТЕСТИРОВАНИЕ

a/b тестирование

Формально:

A/B тестирование – способ сравнения двух вариантов переменной (выборок A и B), основанный на методах статистики. Результатом A/B тестирования является оценка большей эффективности одного из вариантов. Другими словами это способ ответить на вопрос является ли вариант A лучше варианта B или наоборот.

a/b тестирование

Это мощный маркетинговый инструмент для повышения эффективности работы вашего интернет-ресурса.

С помощью A/B тестов повышают конверсию посадочных страниц, подбирают оптимальные заголовки объявлений в рекламных сетях, улучшают качество поиска.

Зачем?

Представим, наш проект запущен в жизнь, на нем собирается трафик, пользователи активно используют ресурс.
И в один прекрасный день мы решили что-то поменять, например, разместить всплывающий виджет для удобства подписки на новости.

Но

Наши предположения и гипотезы строятся на основе личного опыта и наших взглядов, которые совсем не обязательно совпадают со взглядами аудитории нашего ресурса.

Для чего используют а/б тесты

1

UI/UX

2

Эластичность спроса

3

Микро-запуски

Раздача листовок

Промоутеры раздали
по 3 000 листовок двух видов:

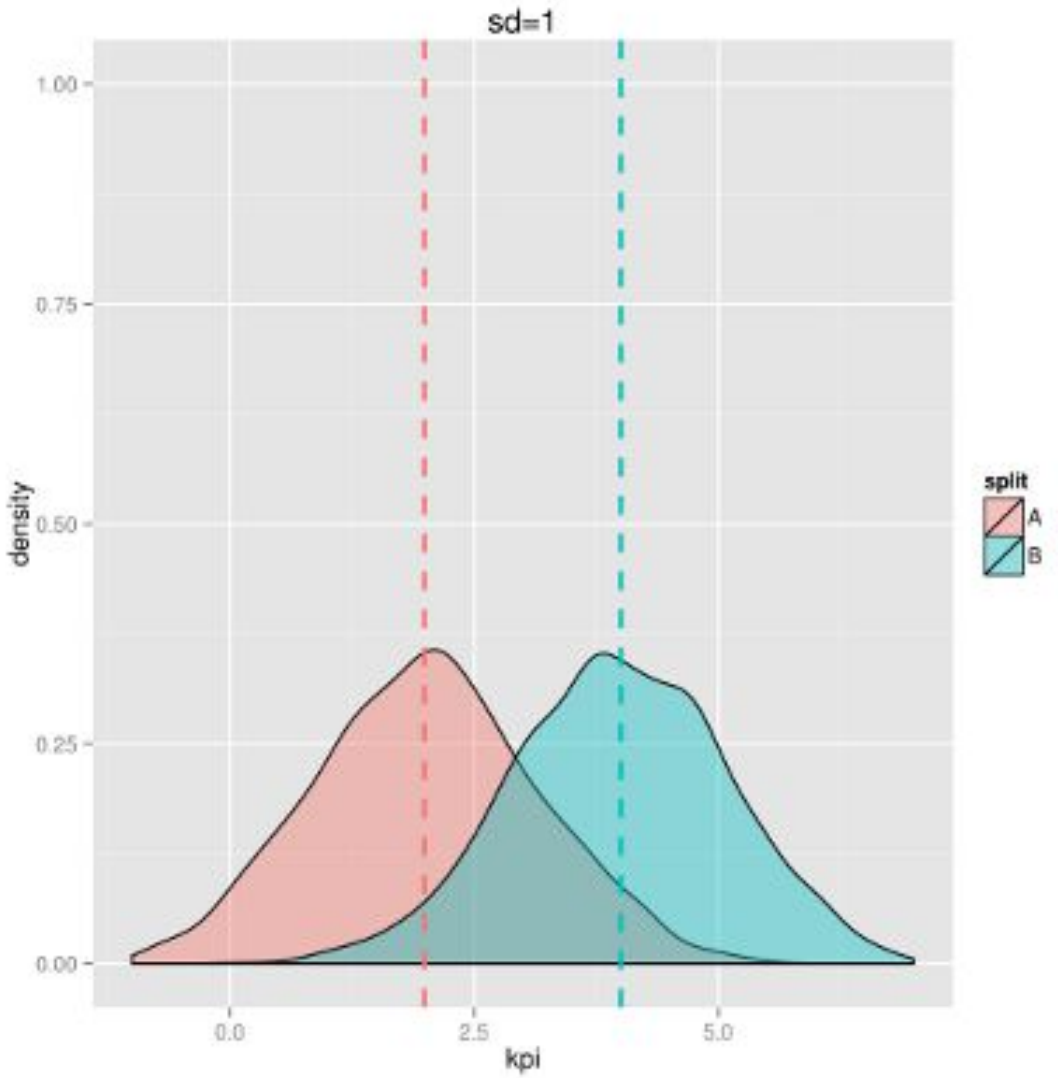
- На листовку А пришли 134 человека
- На листовку Б пришел 121 человек

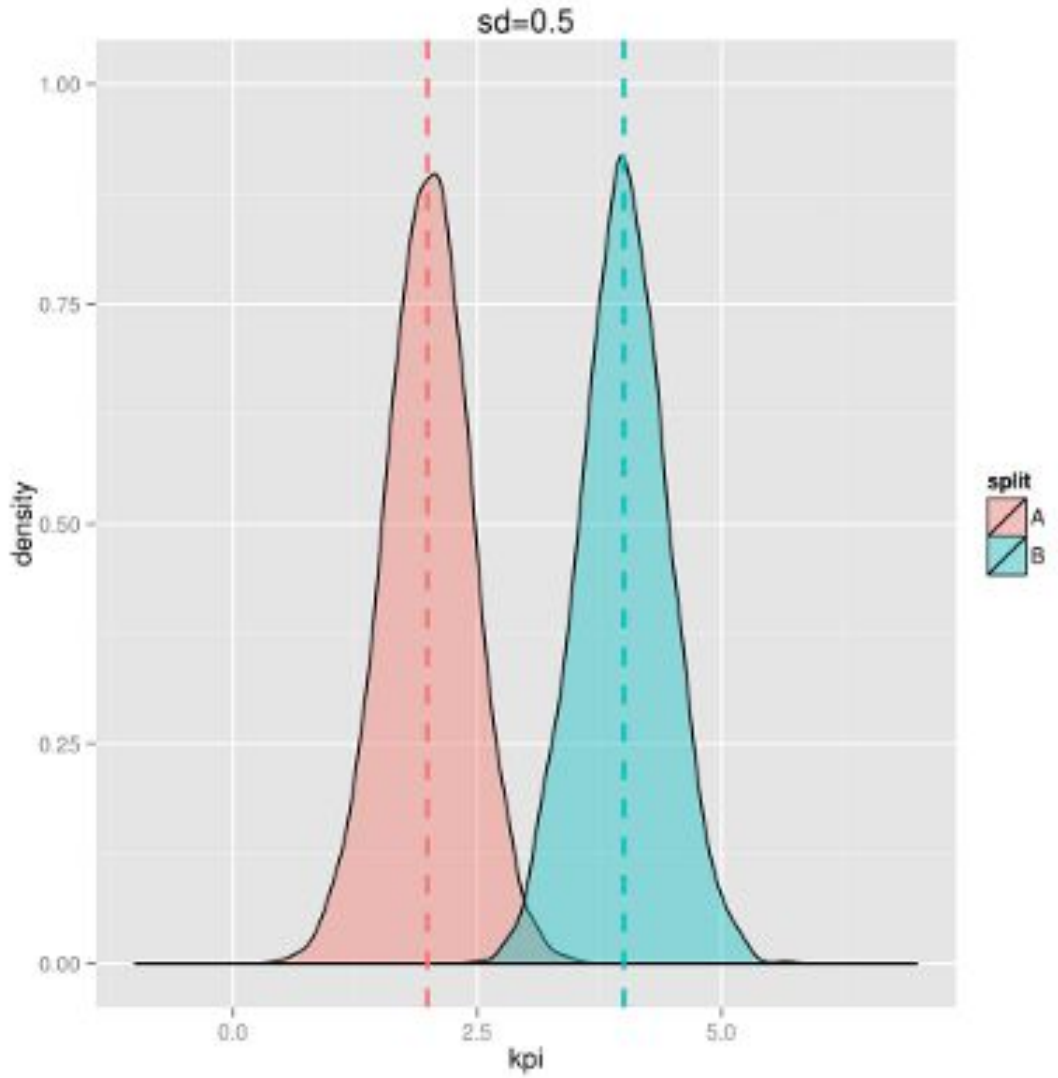
**Можем ли мы сделать вывод,
что листовка А работает лучше?**

Раздача листовок

	Всего	Успех	Конверсия	95% доверительный интервал	
A	3000	134	4,47%	3,73%	5,21%
B	3000	121	4,03%	3,33%	4,74%

A/B ТЕСТИРОВАНИЕ





Формулы

$$Z = \frac{p_1 - p_2}{\sqrt{SE_1{}^2 + SE_2{}^2}} \qquad SE = \sqrt{\frac{p \times (1-p)}{n}}$$

https://vwo.com/ab-split-test-significance-calculator/

https://vwo.com/downloads/ab_testing_significance_calculator.xls

Peeking problem

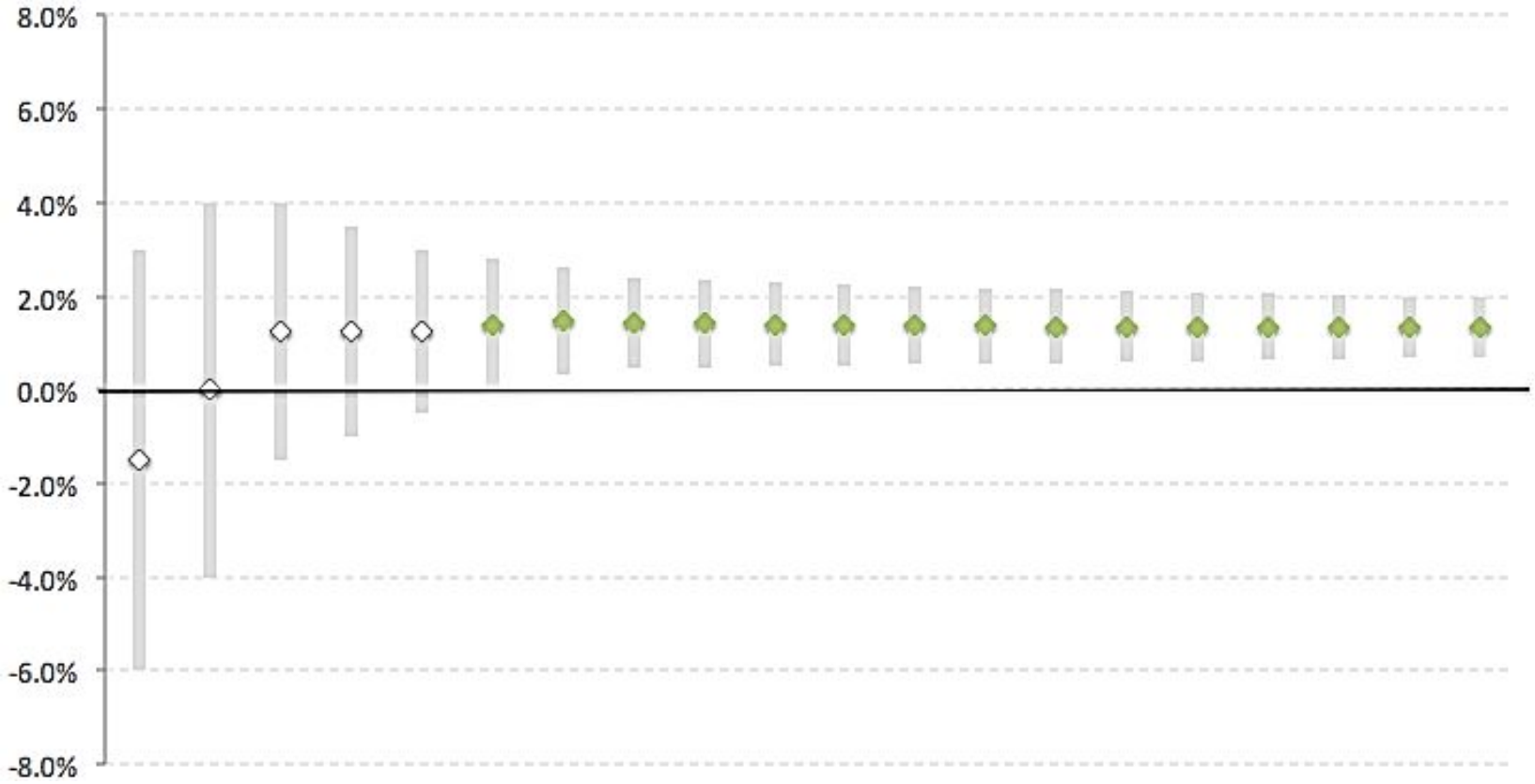
Разработчики проявили инициативу и *написали скрипт,* который каждые несколько часов считал конверсию в первую покупку для тестовой и контрольной версий и проверял, является ли разница значимой.

Спустя несколько дней система выдала сообщение о наличии значимой разницы.

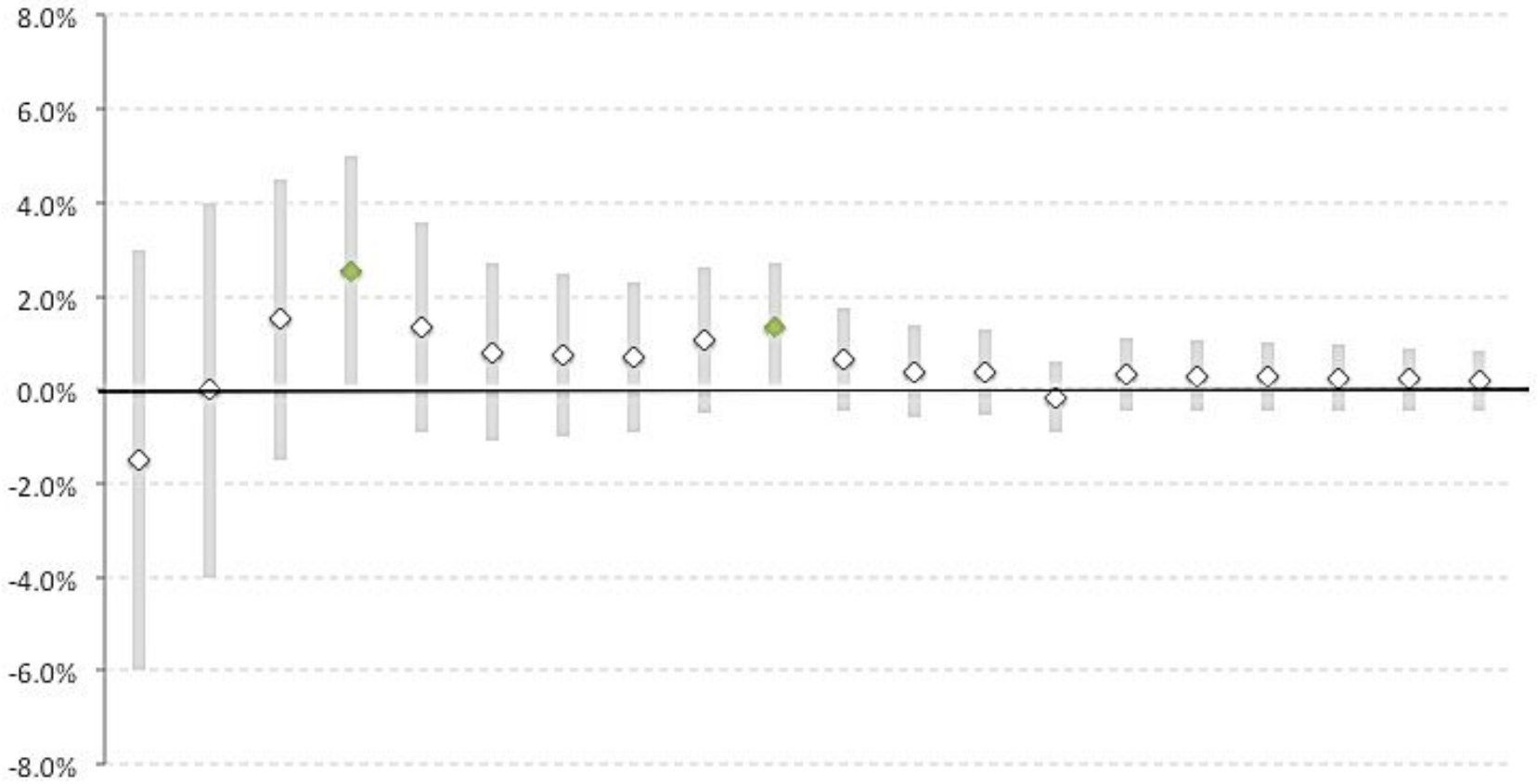
Эксперимент признали успешным *и новую версию раскатили на всех пользователей.*

Вы могли не заметить, но в процесс анализа эксперимента **закралась коварная ошибка.**

Peeking problem



Peeking problem

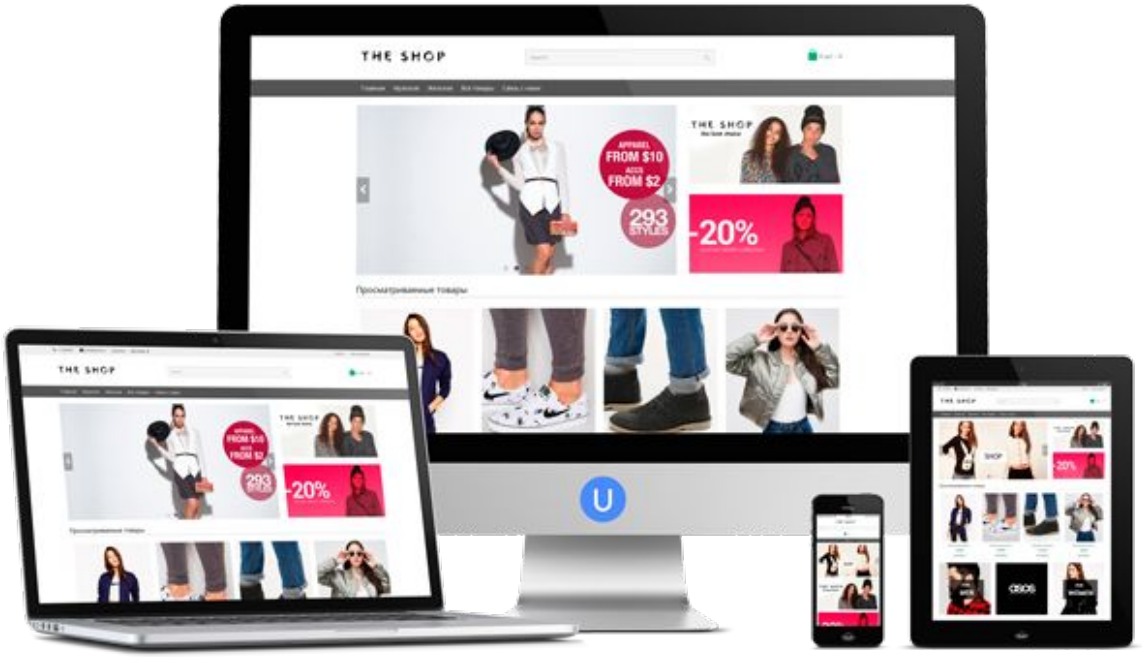


Заметить разницу в 1 %

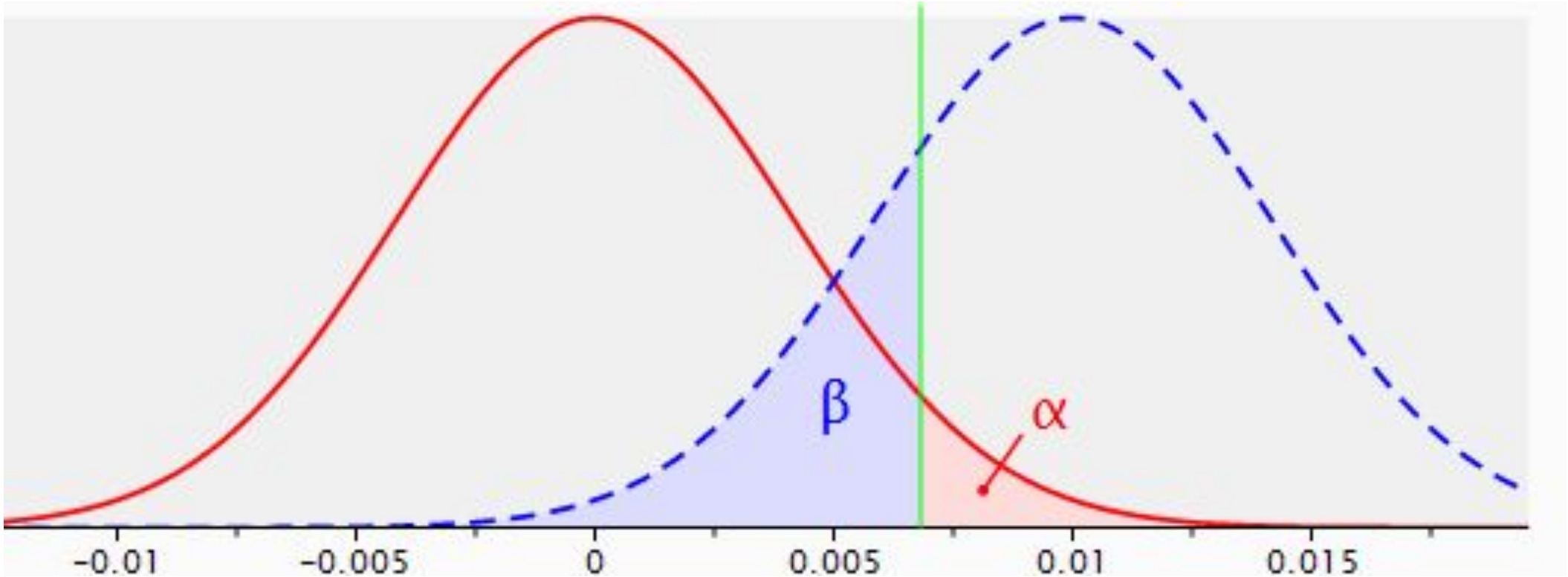
Мы хотим обновить страницу товара
нашего интернет магазин
и ожидаем прирост конверсии
в покупку на 1%
в абсолютном значении

Знаем, что средняя историческая конверсия в покупку — 5 %

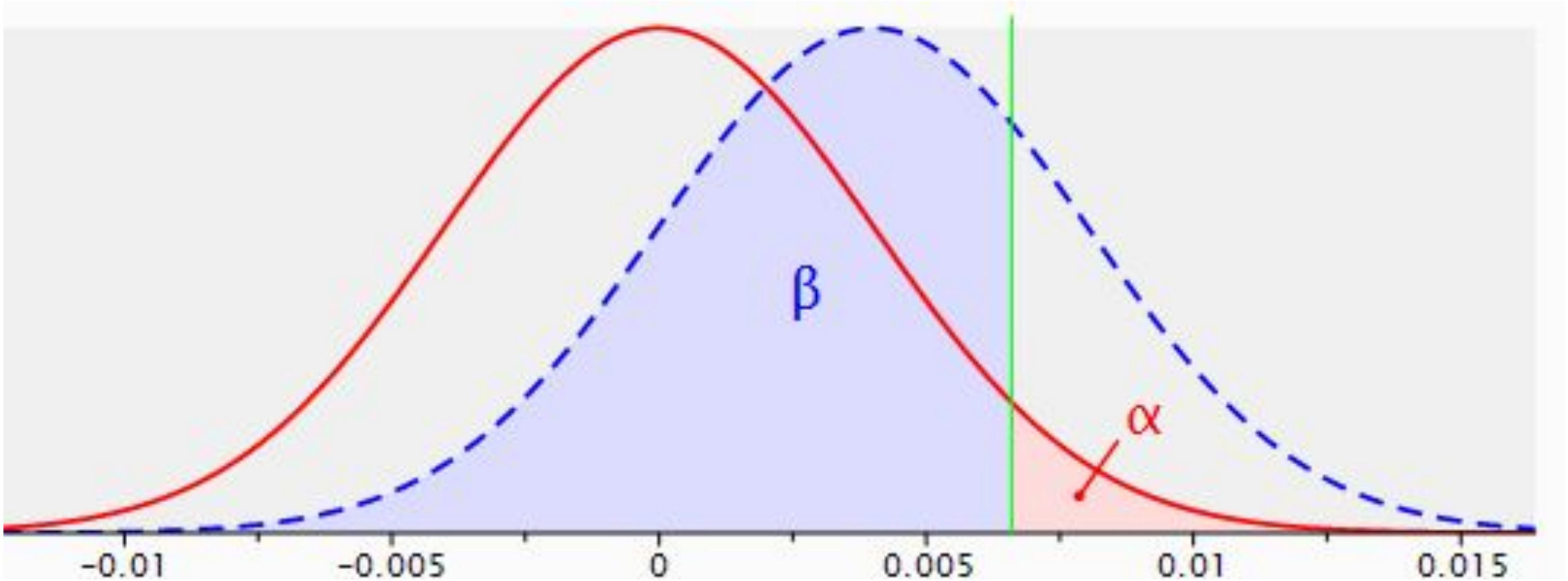
На какое количество пользователей нам нужно раскатить решение, чтобы добиться статистической значимости?



A/B ТЕСТИРОВАНИЕ



A/B ТЕСТИРОВАНИЕ



Вывод:
Перед проведением а/б теста необходимо предварительно **оценить размер выборки**

ПРАКТИКА

ПРАКТИКА

Email рассылка

	Получили	Открыли	Доля, %	Перешли	Доля, %	Заказали	Доля, %
Базовый вариант	35000	3150	9.0	2025	63.3	59	2.90
Зеленая кнопка	2050	189	9.2	134	71.2	4	2.87
Красная кнопка	1950	199	10.2	131	66.1	3	2.60

Что делать дальше?

ПРАКТИКА

Запускаем приложение

	Перешли	Скачали	Конверсия, %	Купили	Конверсия, %
Лендинг 1	2427	457	18.8	14	0.58
Лендинг 2	2144	622	29.0	16	0.75

Что делать дальше?

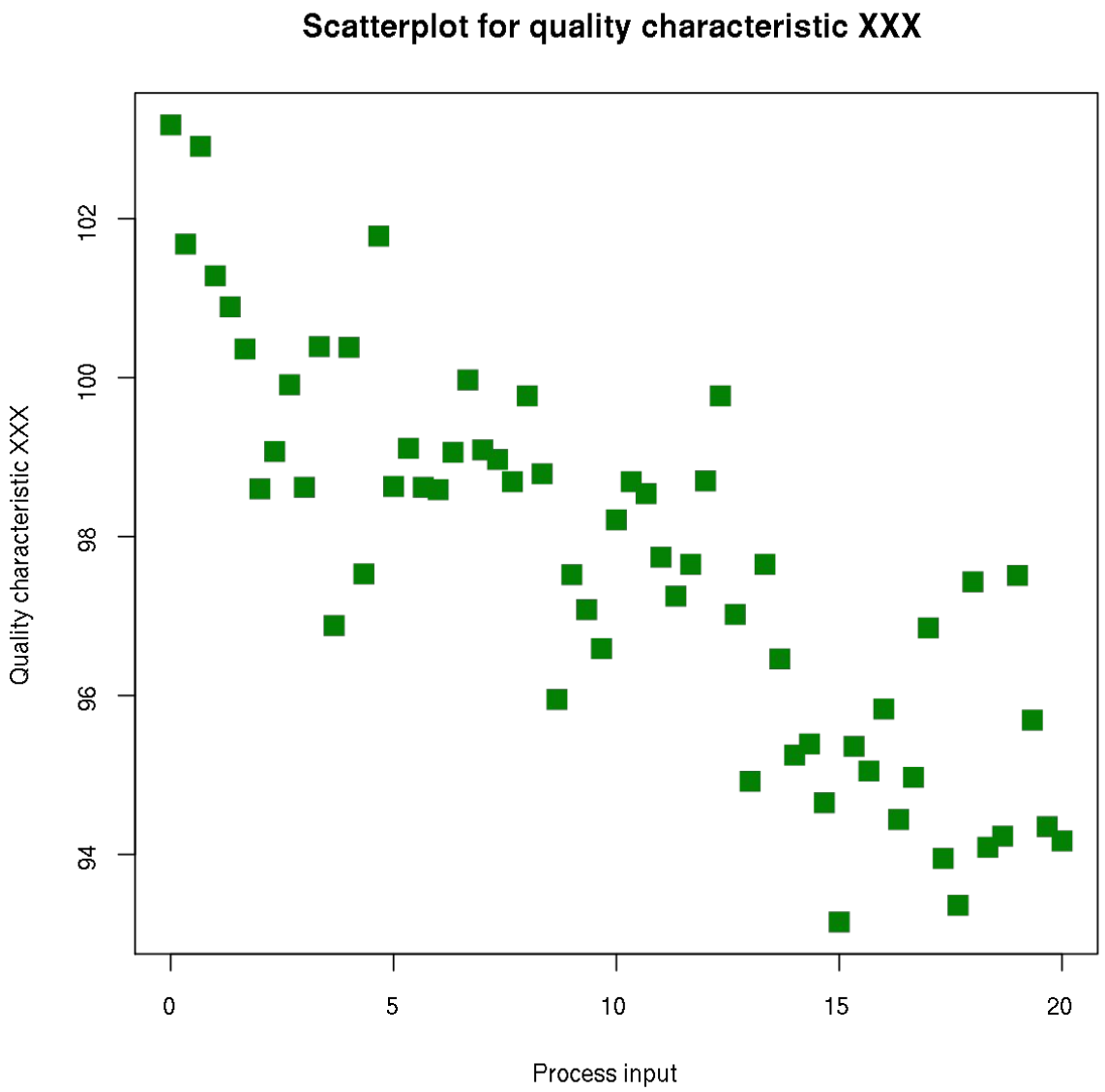
КОРРЕЛЯЦИЯ

Корреляция

Коротко:

Взаимозависимость двух или нескольких случайных величин.

Суть ее заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшению или увеличению) другой(-их) переменной(-ых).



Коэффициент корреляции

параметр, который **характеризует степень линей-ной взаимосвязи между двумя выборками,** рассчитывается по формуле:

$$r_{xy} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

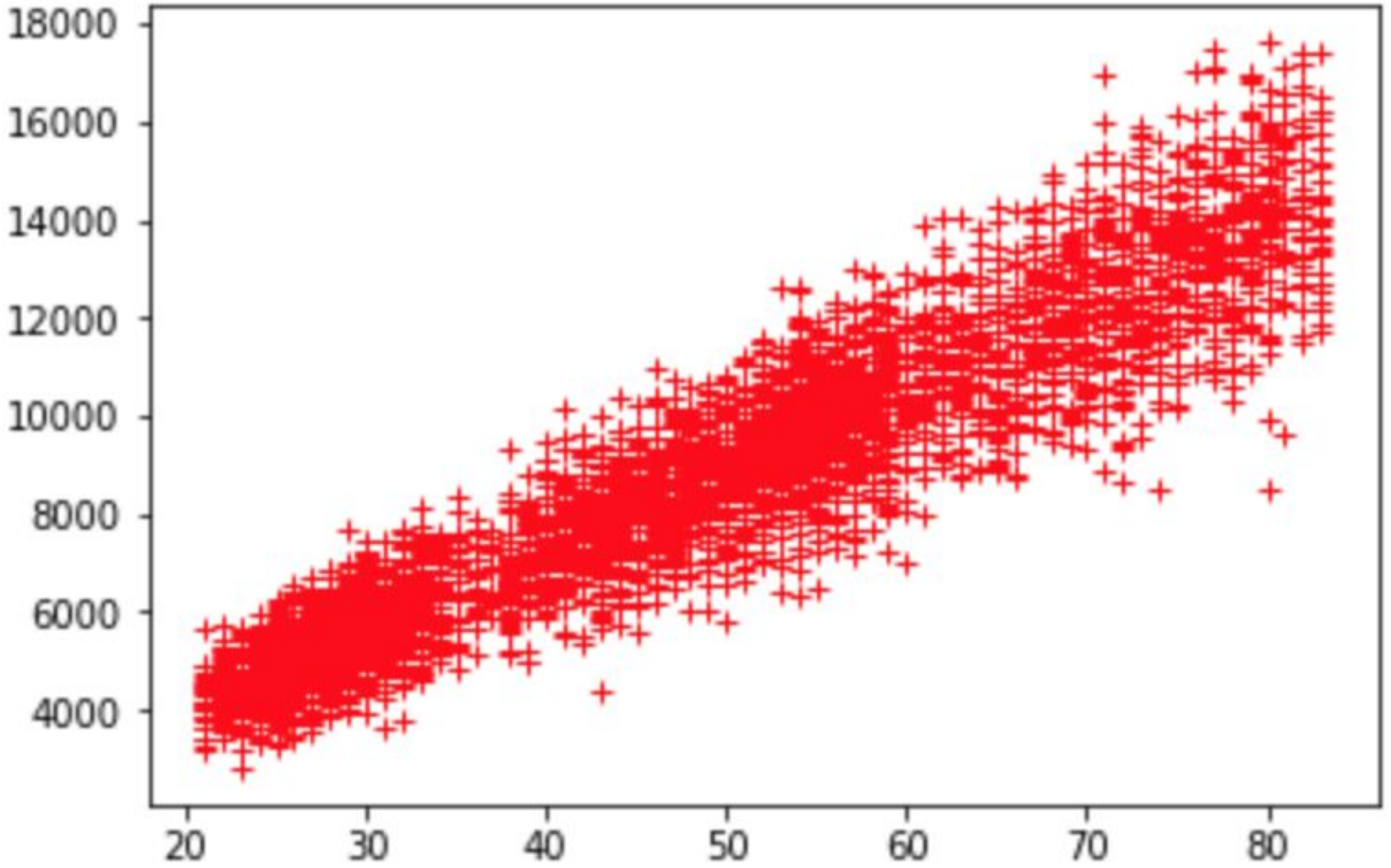
Практика с набором данных “Цены на квартиры”

https://clck.ru/EjWL5

ЗАВИСИМОСТИ

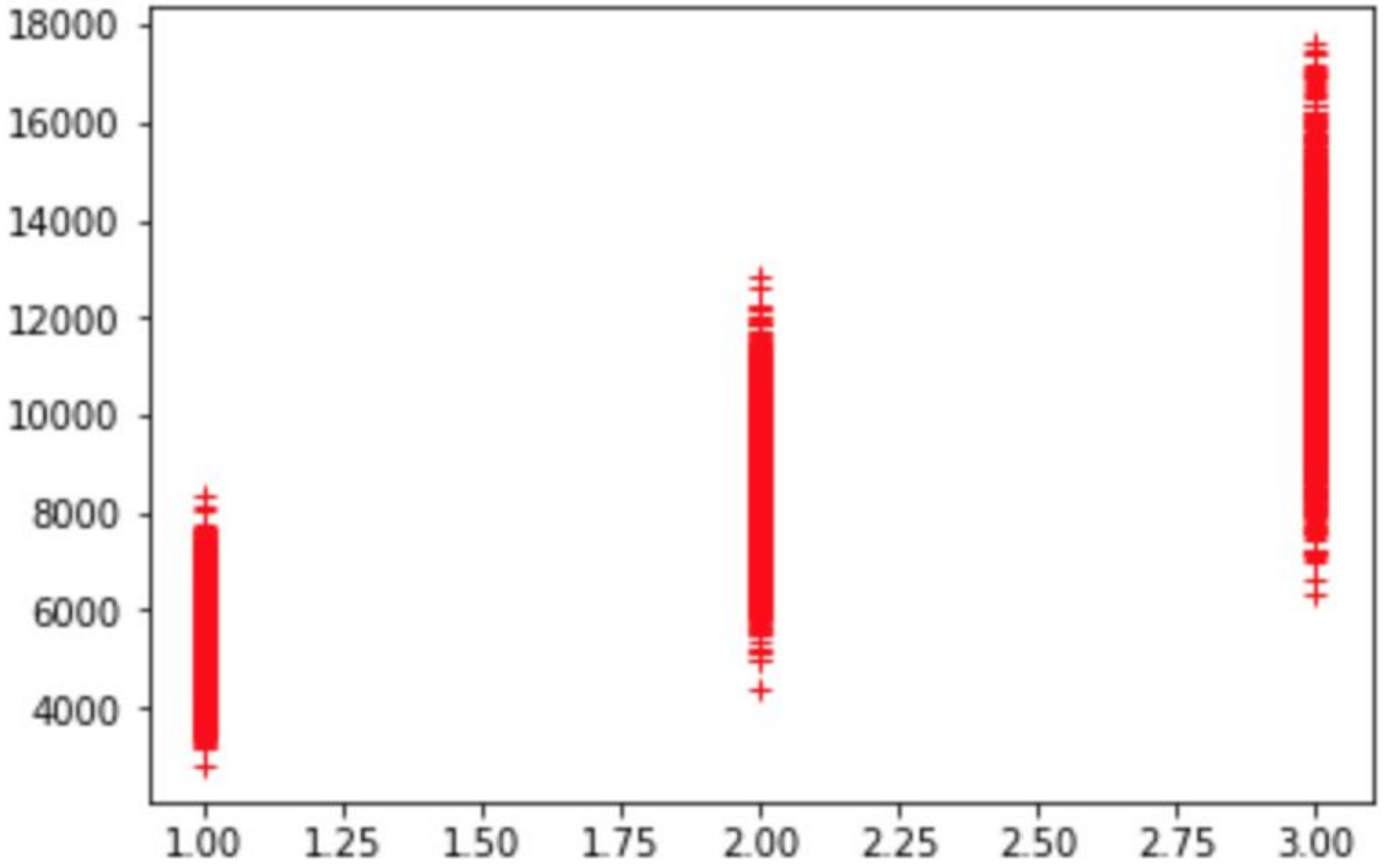
ЗАВИСИМОСТИ

Area и Price



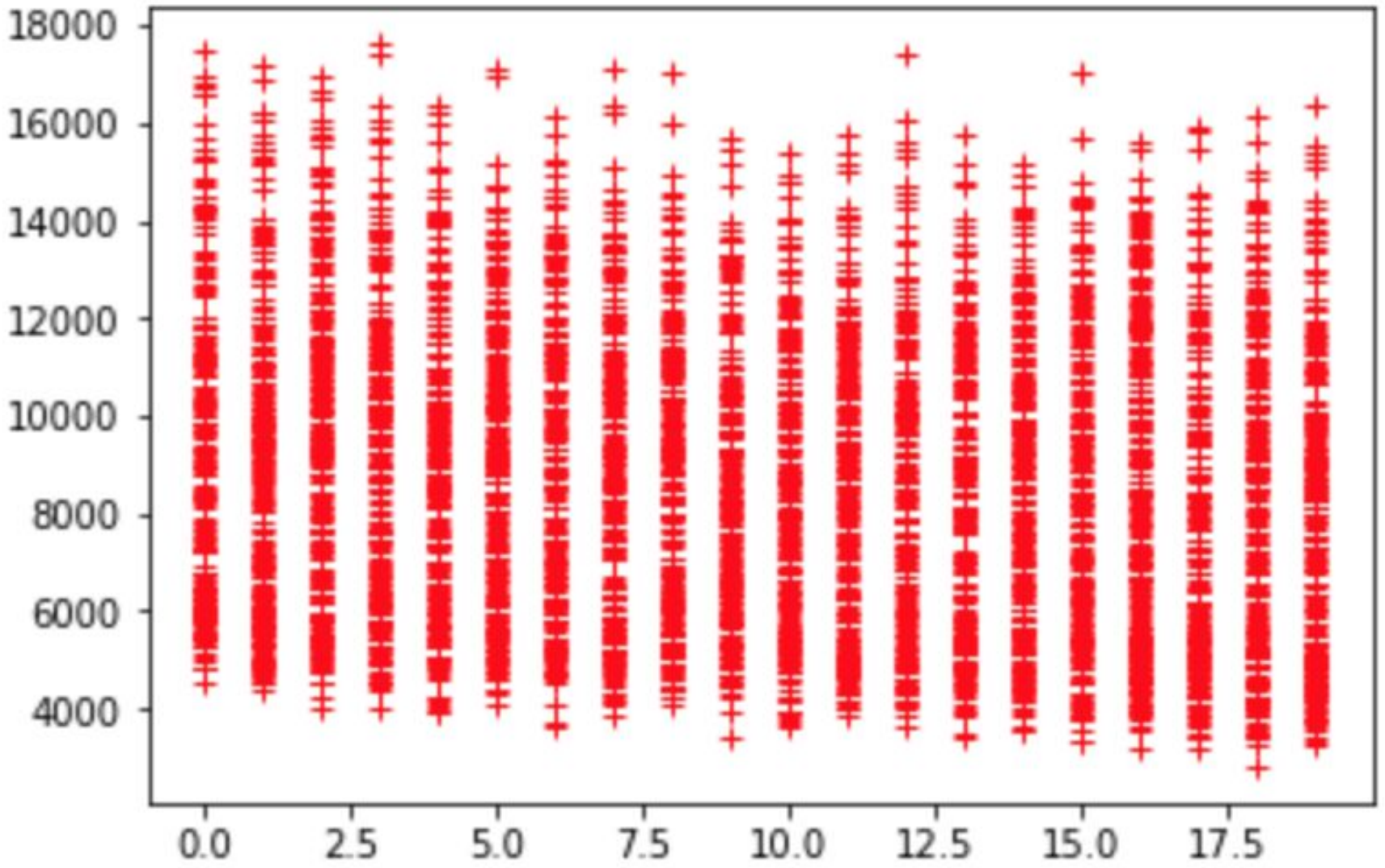
ЗАВИСИМОСТИ

Rooms и Price



ЗАВИСИМОСТИ

DistMetro и Price



ПРАКТИКА

Дополнительные материалы

1. ru.wikihow.com/рассчитать-линейный-коэффициент-корреляции
2. gopractice.ru
3. retailrocket.ru/blog/
4. habr.com/post/233911/
5. www.evanmiller.org/ab-testing/sample-size.html
6. hungrysites.ru/ab
7. www.evanmiller.org/how-not-to-run-an-ab-test.html
8. https://netology.ru/blog/03-2019-statisticheskaya-znachimost

СПАСИБО ЗА ВНИМАНИЕ