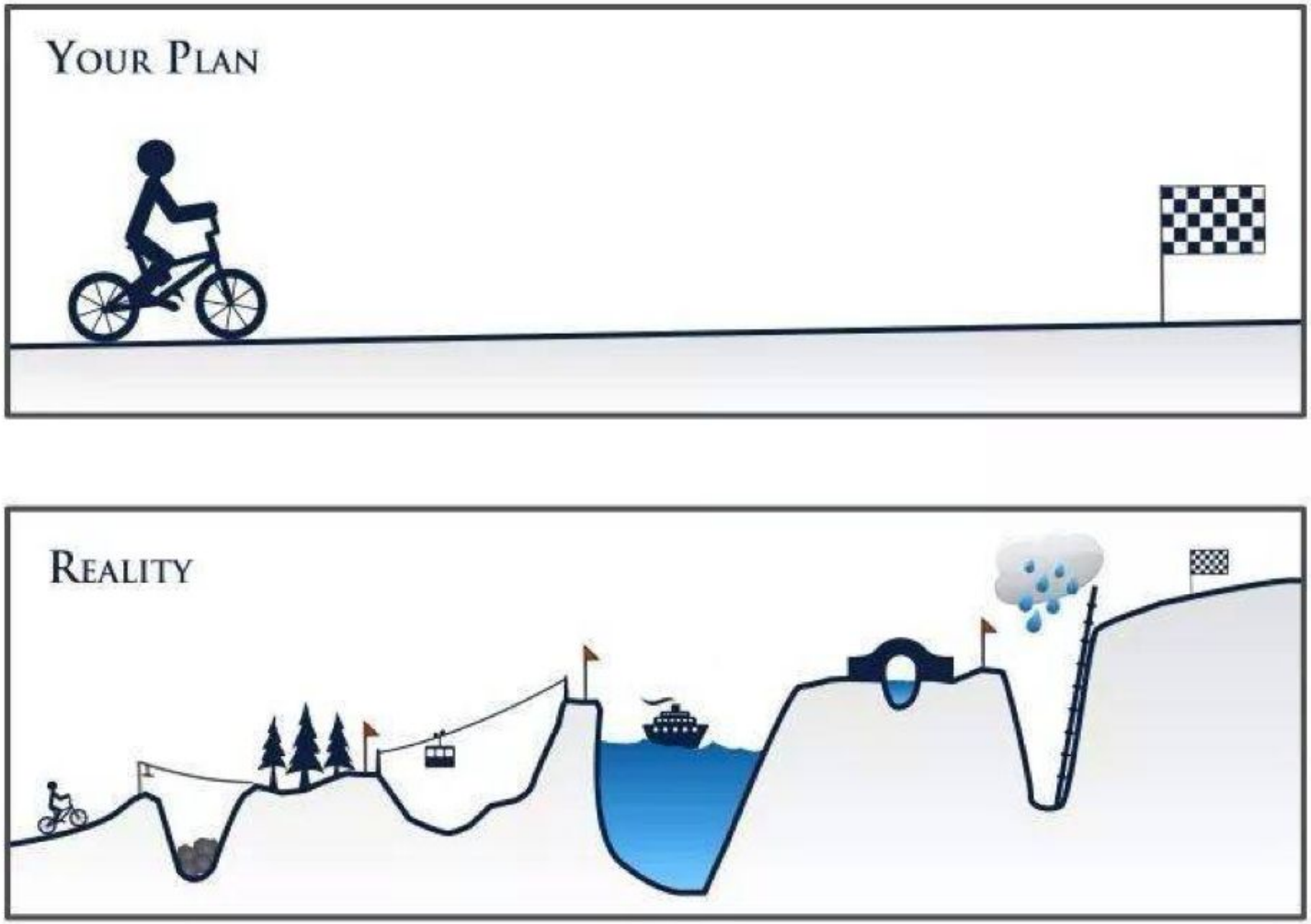
Статистика и реальные данные, как быть и что делать?

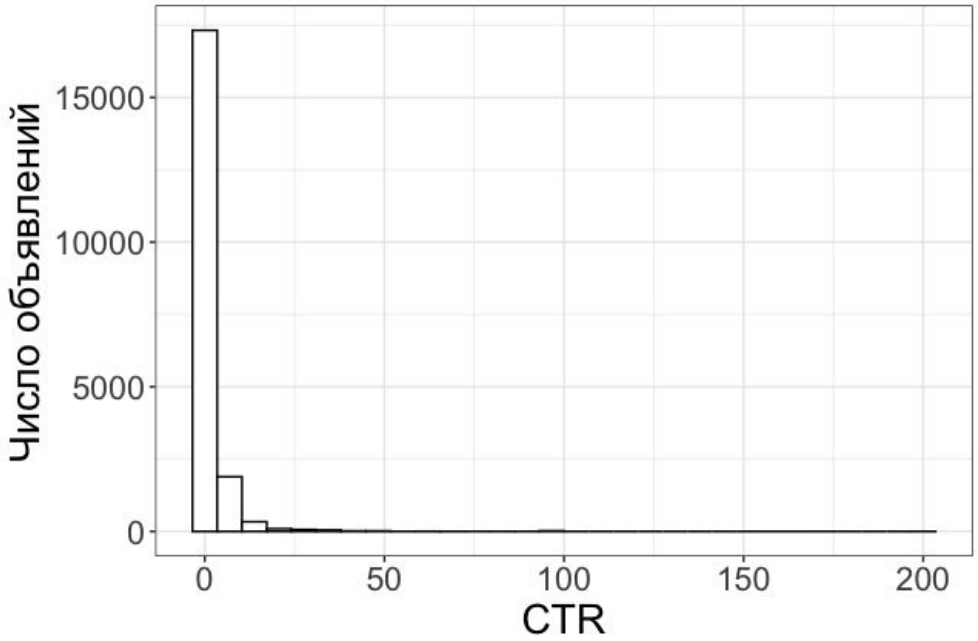
Анатолий Карпов


YOUR PLAN
REALITY

Ненормальные распределения + выбросы

Почти всегда это:

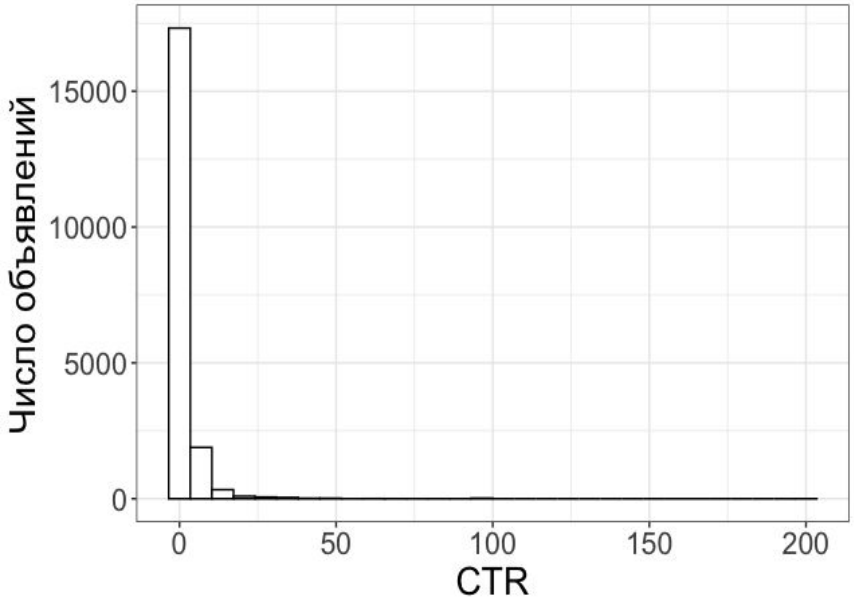
- Деньги
- CTR, CPM
- Вовлеченность юзеров
- ...

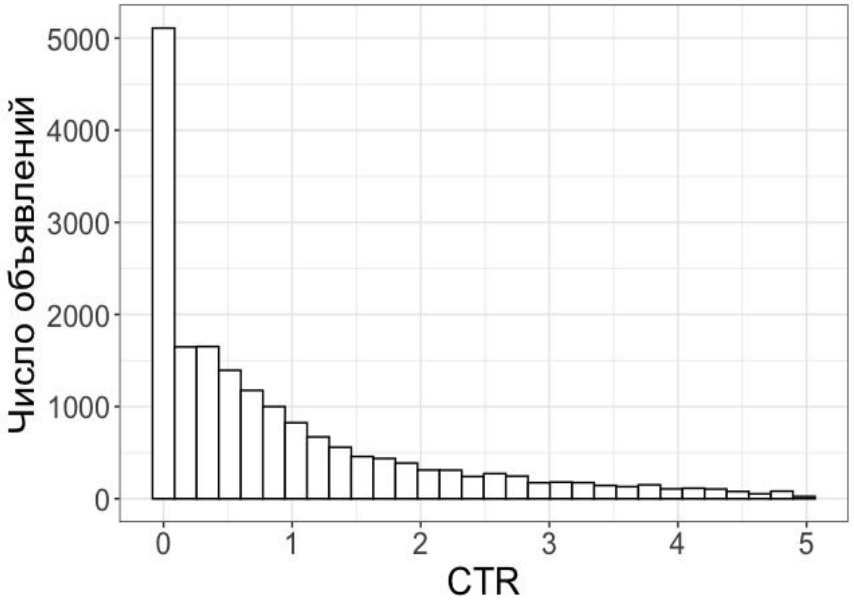


Ненормальные распределения

- Удаление выбросов
- Логарифмирование
- Непараметрика
- Bootstrap

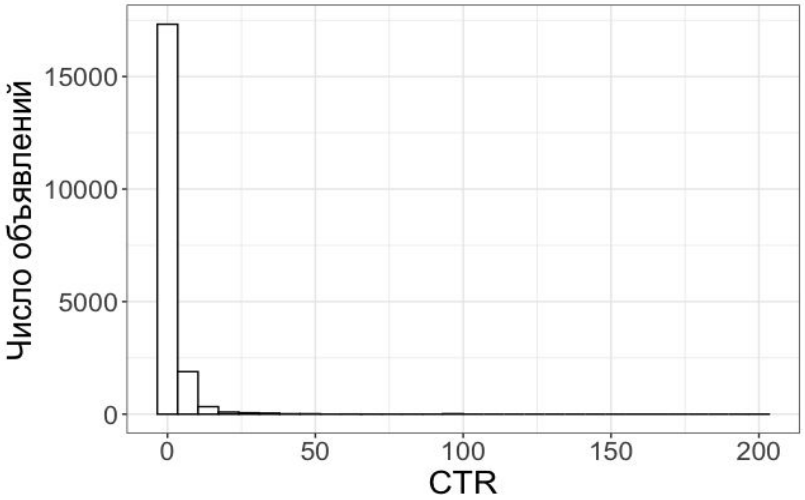
Ненормальные распределения

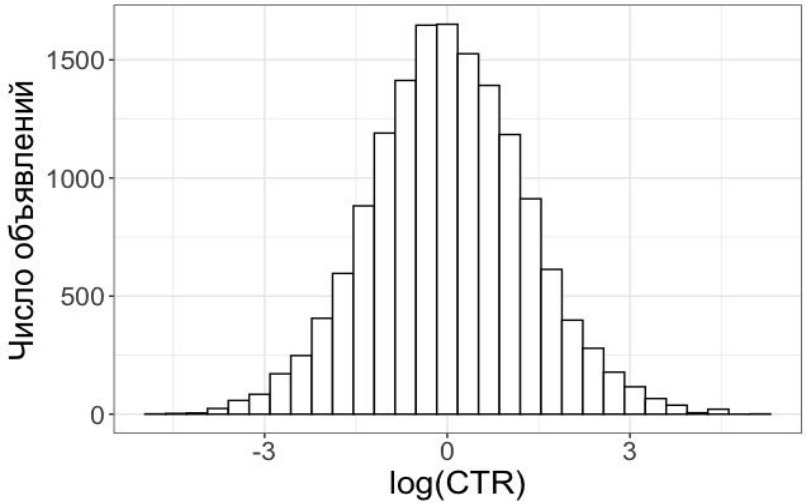




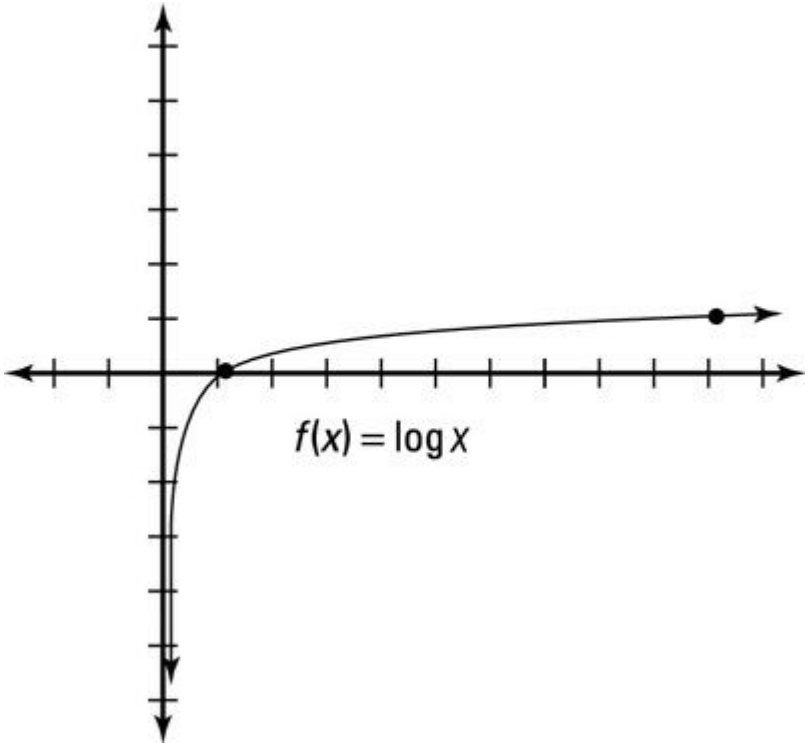
Оставим только объявления с CTR < 5 и посмотрим на реальное распределение.

Логарифмирование переменных



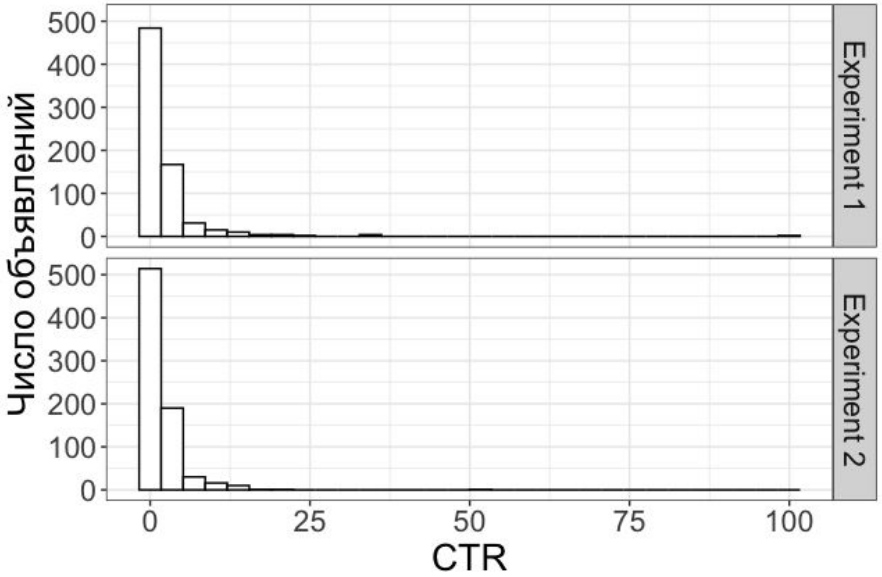


Почему это вообще работает?



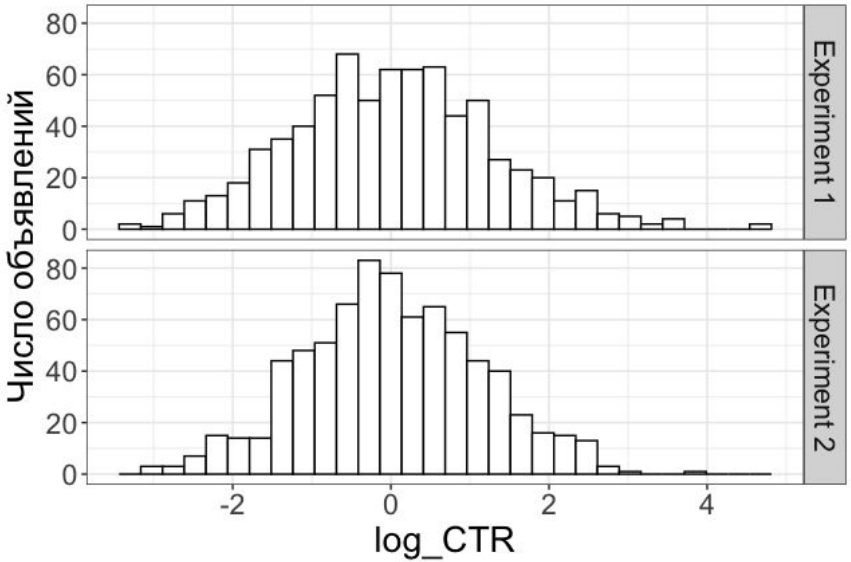
X	log(x)
1	0
2	0.69
3	1.10
100	4.61
1000	6.91

Параметрические тесты - не самая лучшая идея



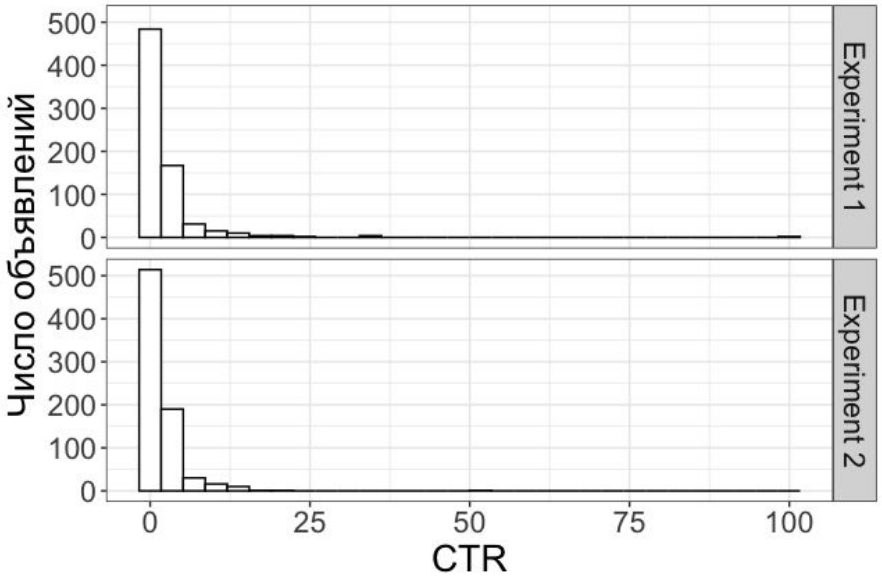

```
        Welch Two Sample t-test

data:  CTR by group
t = 2.2992, df = 1013.3, p-value = 0.0217
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.08942826 1.13131908
sample estimates:
mean in group Experiment 1 mean in group Experiment 2
                  2.556309                   1.945935
```

После логарифмирования сильно лучше




```
        Welch Two Sample t-test

data:  log_CTR by group
t = 0.12812, df = 1439.1, p-value = 0.8981
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1151319  0.1312223
sample estimates:
mean in group Experiment 1 mean in group Experiment 2
                0.03044029                 0.02239509
```

Не забываем про непараметрику



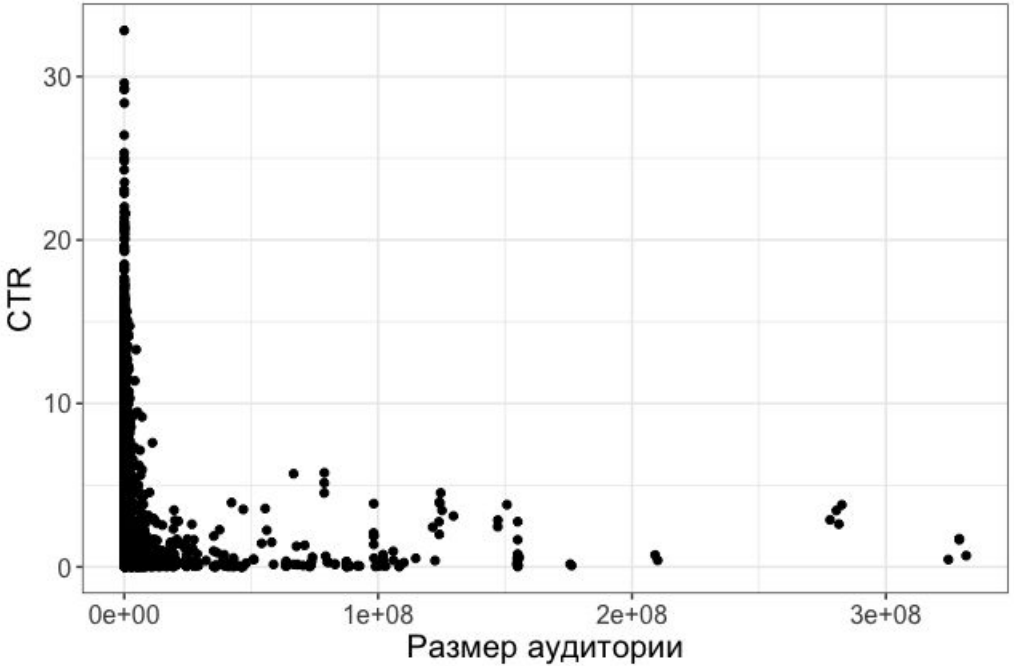

```
        Wilcoxon rank sum test with continuity correction

data:  CTR by group
W = 274294, p-value = 0.8531
alternative hypothesis: true location shift is not equal to 0
```

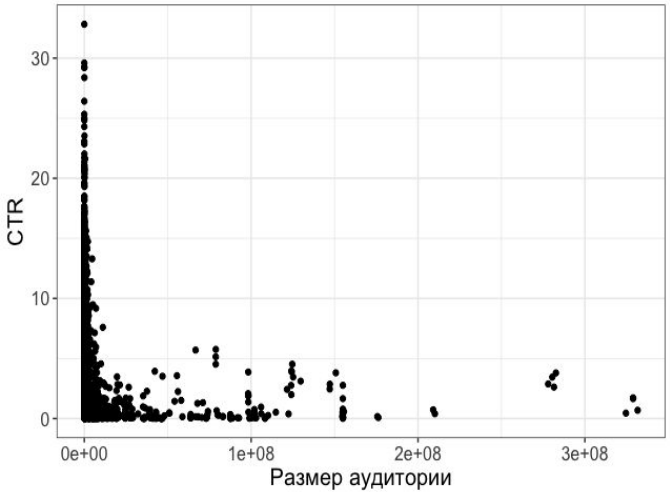
Логарифмирование переменных

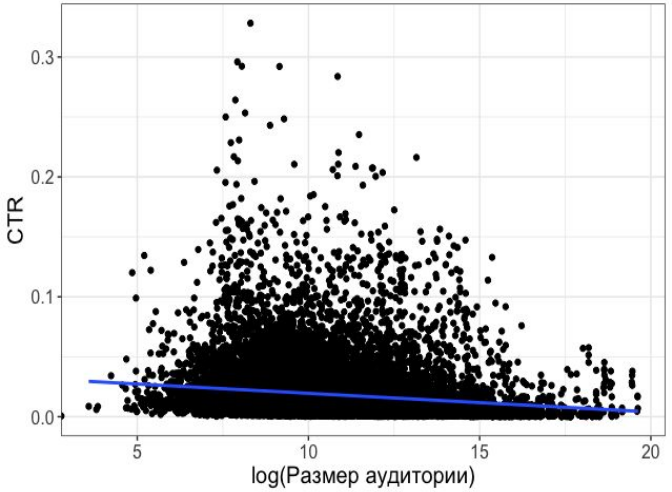
- На удивление хорошо работает на реальных данных
- Не забываем про ноль в исходной переменной (однажды я все сломал)
- Теперь можно применять параметрические тесты

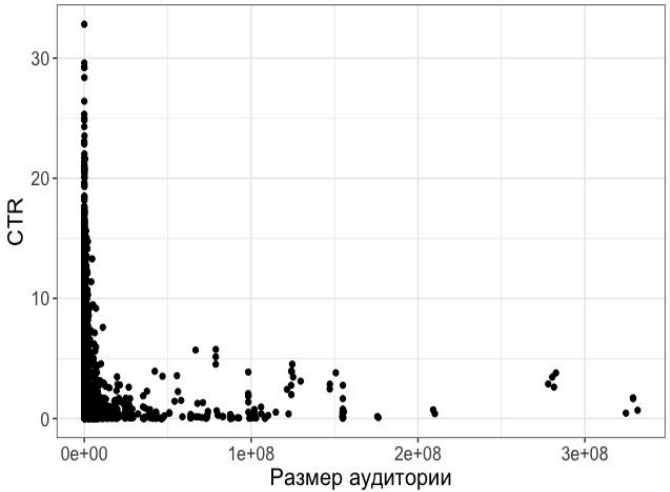
Логарифмирование переменных

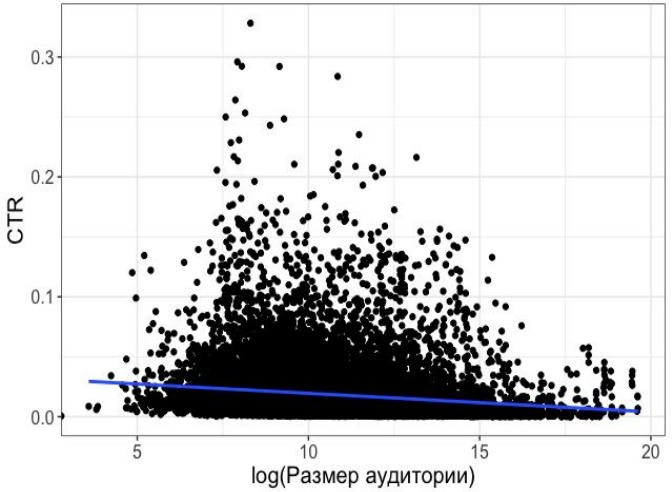
Помогает понять природу взаимосвязи между переменными, как минимум на графиках. Как связан CTR и размер аудитории?

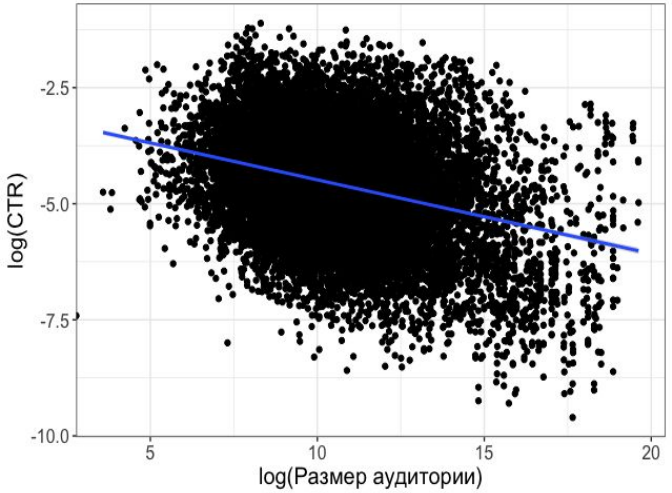



CTR
30
20
10
0
0e+00
1e+08
2e+08
3e+08
Размер аудитории


CTR
0.3
0.2
0.1
0.0
5
10
15
20
log(Размер аудитории)


CTR
30
20
10
0
0e+00
1e+08
2e+08
3e+08
Размер аудитории


CTR
0.3
0.2
0.1
0.0
5
10
15
20
log(Размер аудитории)

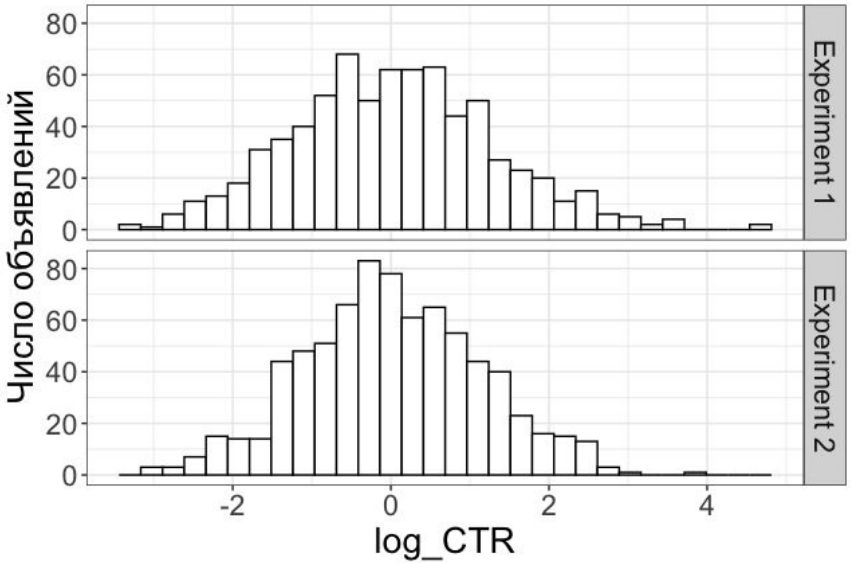

log(CTR)
-2.5
-5.0
-7.5
-10.0
5
10
15
20
log(Размер аудитории)

Логарифмирование переменных

Можно рассматривать как частный случай метода трансформации **Box–Cox transformation** - использовать не только логарифмирование, но и другие математические операции, например, возведение в степень.

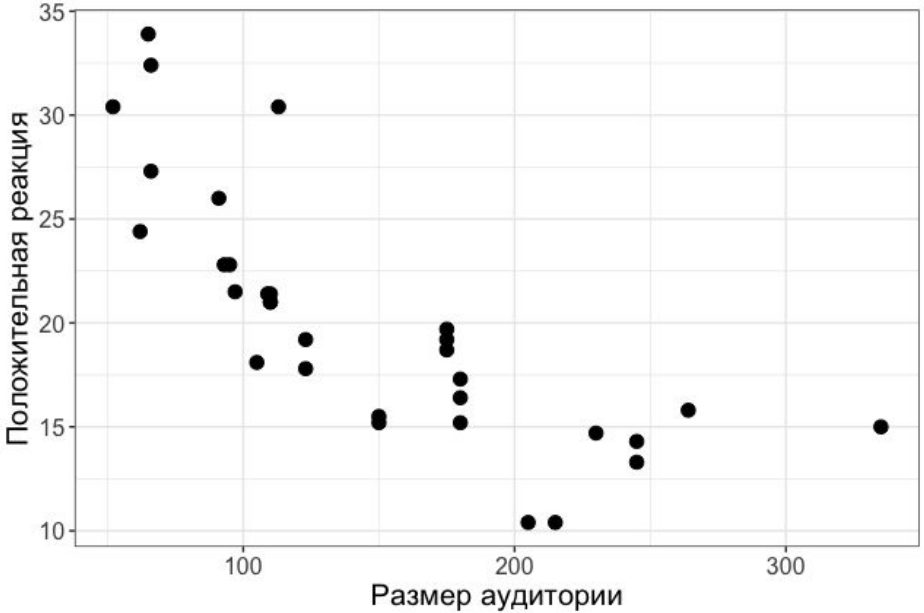
- В большинстве случаев можно выправить распределение
- Помогает визуализировать зависимости
- Усложняет интерпретацию результатов

Интерпретация результатов



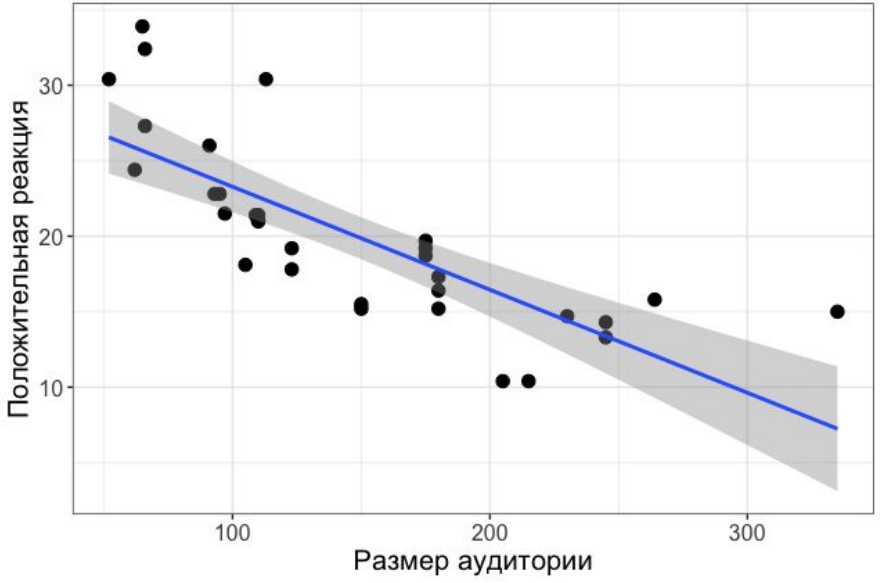
Сравнение групп - в целом, можно описать результаты, все что делает логарифм “схлопывает” экстремальные наблюдения

Интерпретация результатов



С анализом зависимостей ситуация сложнее, но есть статистические лайфхаки!

Интерпретация результатов

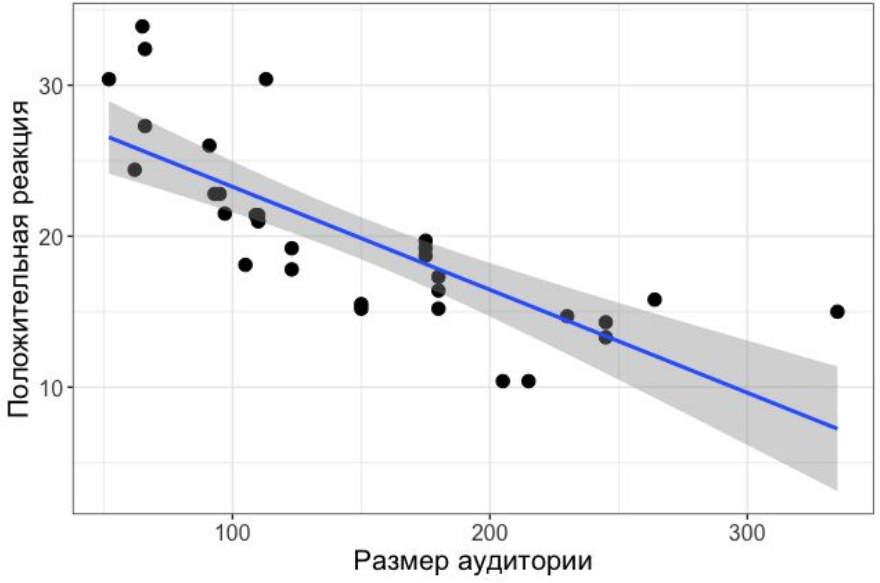


Рейтинг = -0.06 * Аудитория + 30

При единичном изменении размера аудитории рейтинг в среднем снижается на 0.06

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
Target_size -0.06823    0.01012  -6.742 1.79e-07 ***
```

Интерпретация результатов

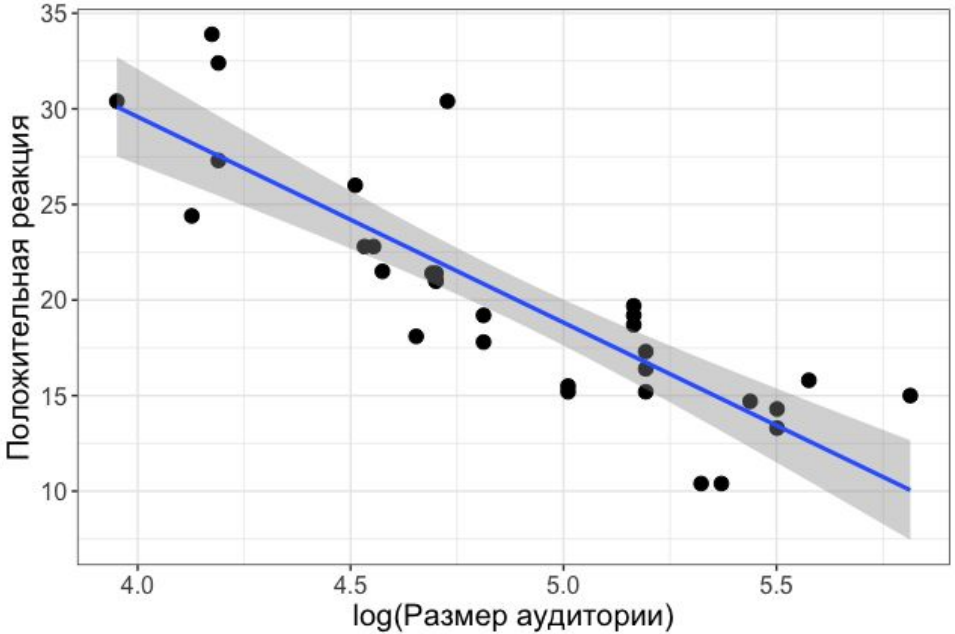



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
Target_size -0.06823    0.01012  -6.742 1.79e-07 ***
```

Очевидно, что нелинейный характер связи никак не учитывается в модели.

1. Использовать более сложные методы.
2. Трансформация переменных

Интерпретация результатов




```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       72.640      6.004  12.098 4.55e-13 ***
log(Target_size) -10.764      1.224  -8.792 8.39e-10 ***
---
```

Корреляция качества рекламных объявлений с логарифмом размера аудитории - звучит не очень понятно(

Интерпретация результатов

Рейтинг = -10 * log(Аудитория) + 72

Изменение на 10% по размеру аудитории в среднем приводит к понижению рейтинга на 1.

- Если мы увеличим аудиторию с 100 до 110, рейтинг упадет на 1
- Если мы увеличим аудиторию с 1000 до 1010, рейтинг упадет на 0.1

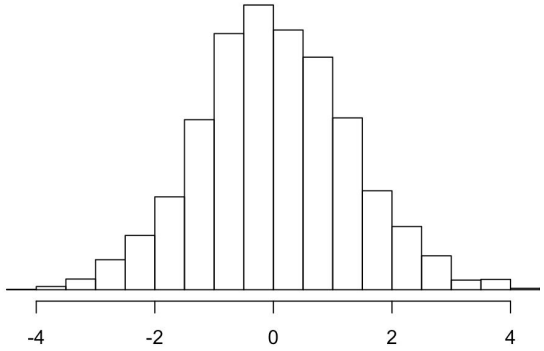
Это и есть нелинейная природа взаимосвязи, чем больше аудитория, с увеличением размера аудитории влияние на рейтинг снижается.

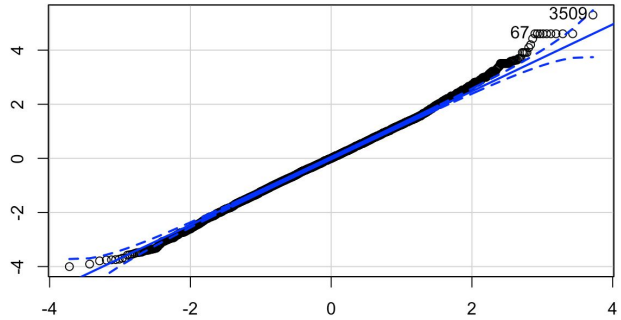
Оценка нормальности распределения

Нетривиальная задача для статистики, особенно на больших данных не стоит полагаться только на p - value соответствующих критериев.

Строго говоря, симметричное распределение - не значит нормальное распределение.

Реальные данные по log(CTR) объявлений





Согласно критерию Shapiro test не можем говорить о нормальности ($p < 0.001$)

При этом корреляция между ожидаемыми и предсказанными квантилями распределения $r = 0.98$

Итого

Нормальное распределение - большая редкость

Логарифмирование - отлично работает с асимметричными распределениями

- Решает проблему выбросов
- Не гарантирует нормальности распределения

Непараметрика - это важно!

Трансформация данных это хорошо, но...

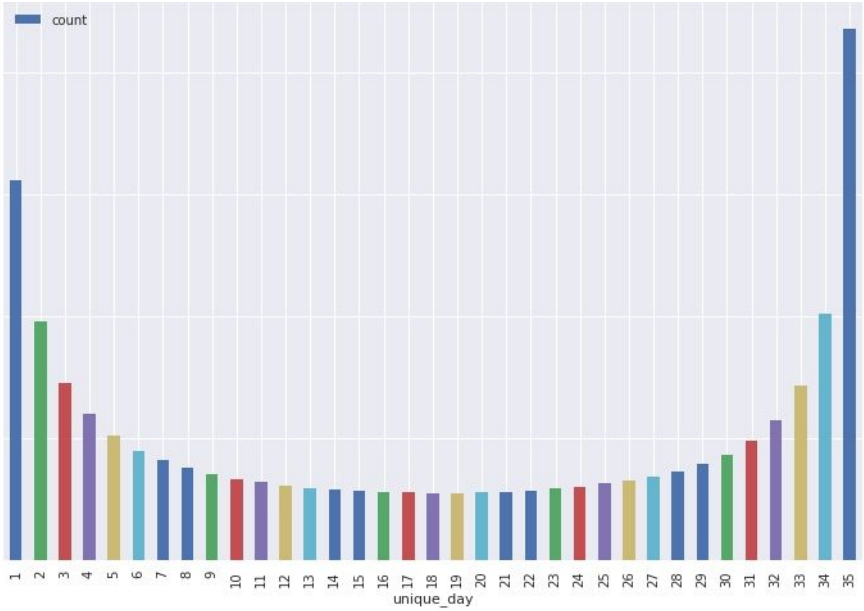
Мы уходим от исходного распределения:

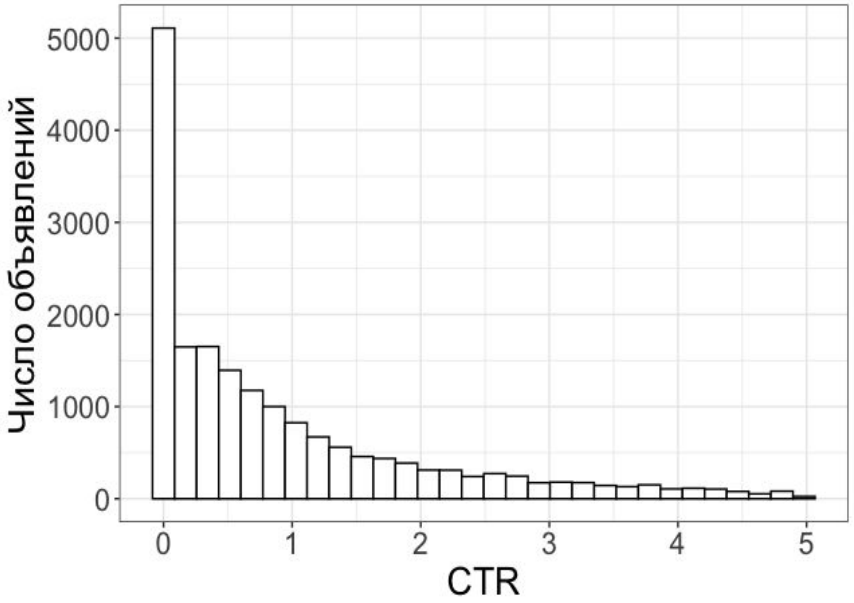
- Зато можем сравнить средние значения

А если хотим работать с исходным распределением?

- Не всегда все так плохо с выбросами
- Не всегда поможет трансформация

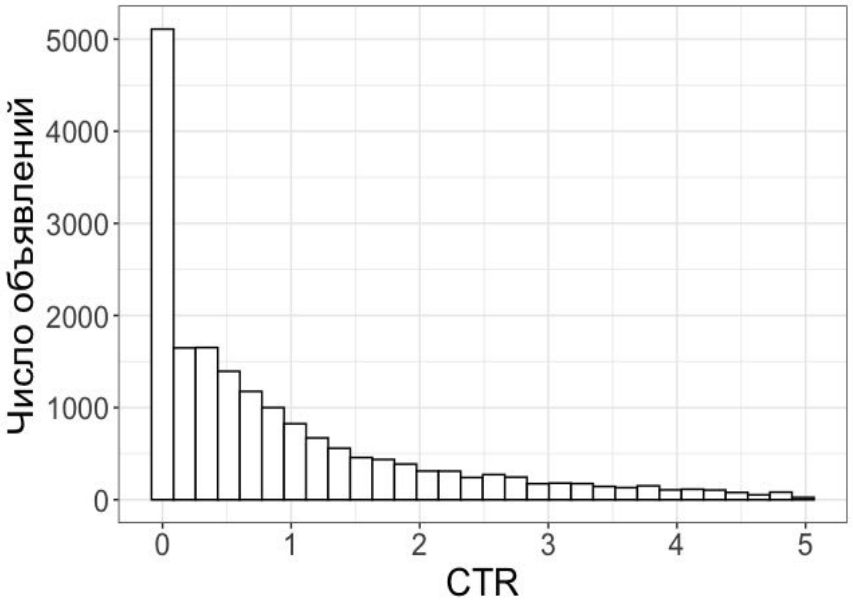
Всякие хитрые распределения





Часто перед нами стоит вопрос не про среднее значение и уж тем более не про сумму рангов.

Всякие хитрые распределения



Изменение формы распределения - переход из одного бакета в другой, может многое значить для продукта.

Сравнение средних или рангов может пропустить такой результат.

Bootstrap и Метод Монте-Карло

Давайте сравнивать действительно то, что нам интересно:

- Медиана
- Максимум
- Минимум
- 13 процентиль

Средние значения тоже можем сравнить

Все самое важное в деталях

Есть новая супер фича, мега усилитель CTR. Выкатили на часть объявлений. Получили следующие результаты:

Группа	**Объявления**	**Показы**	**Клики**	**CTR**
Супер фича	140	190 488	1 727	0.9
Обычные	860	2 798 818	16 218	0.58

Все самое важное в деталях

А если сначала для каждого объявления посчитать CTR, а затем усреднить CTR объявлений в каждой группе?

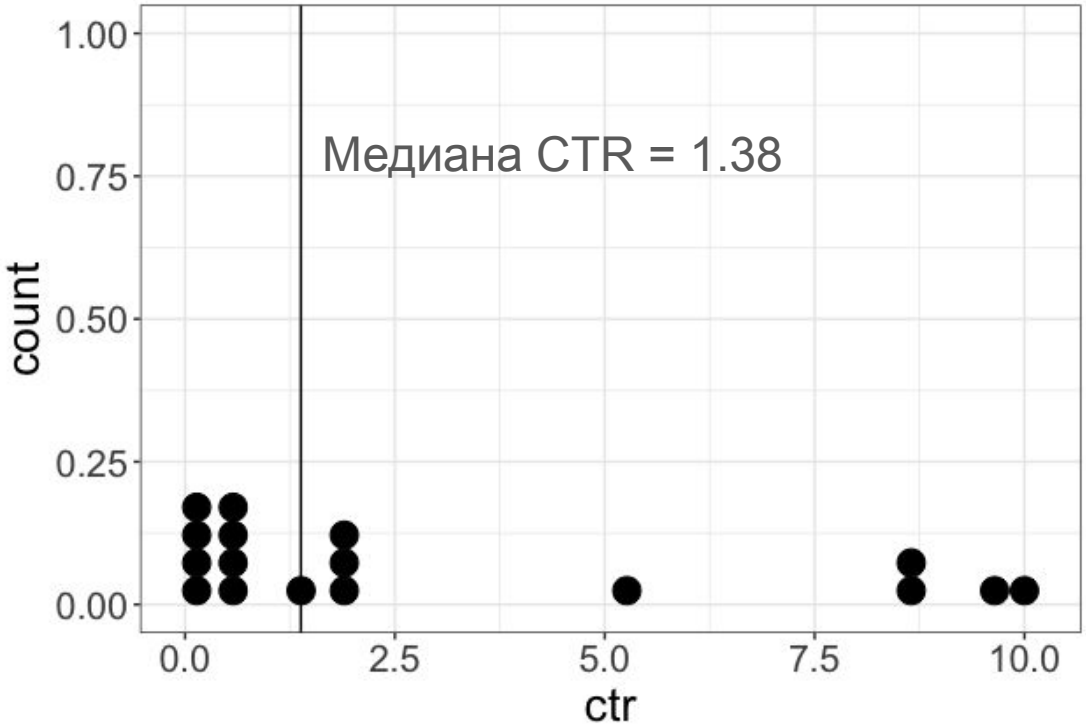
Группа	**Объявления**	**Медиана медиан CTR**
Супер фича	140	1
Обычные	860	1

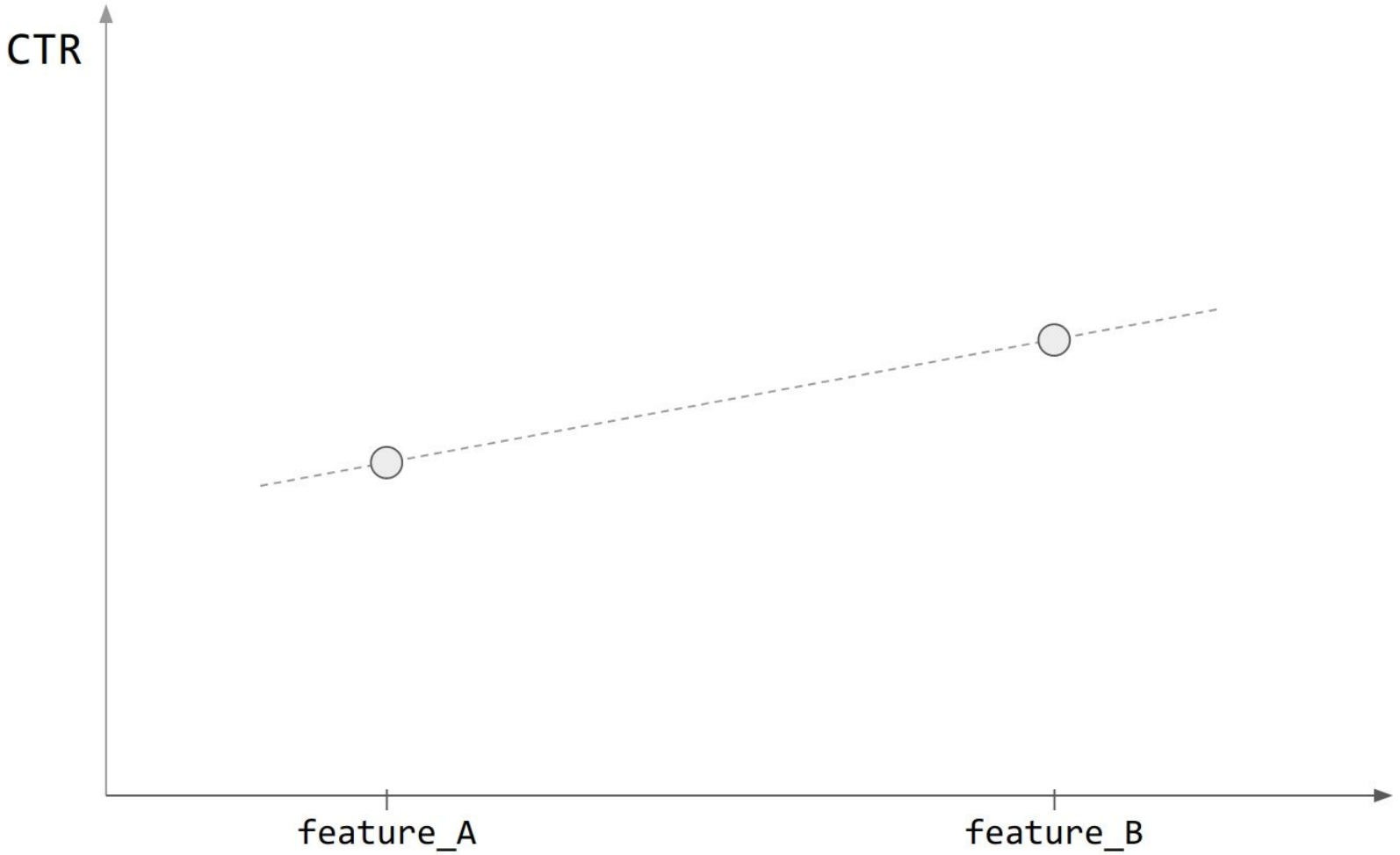
Как так?

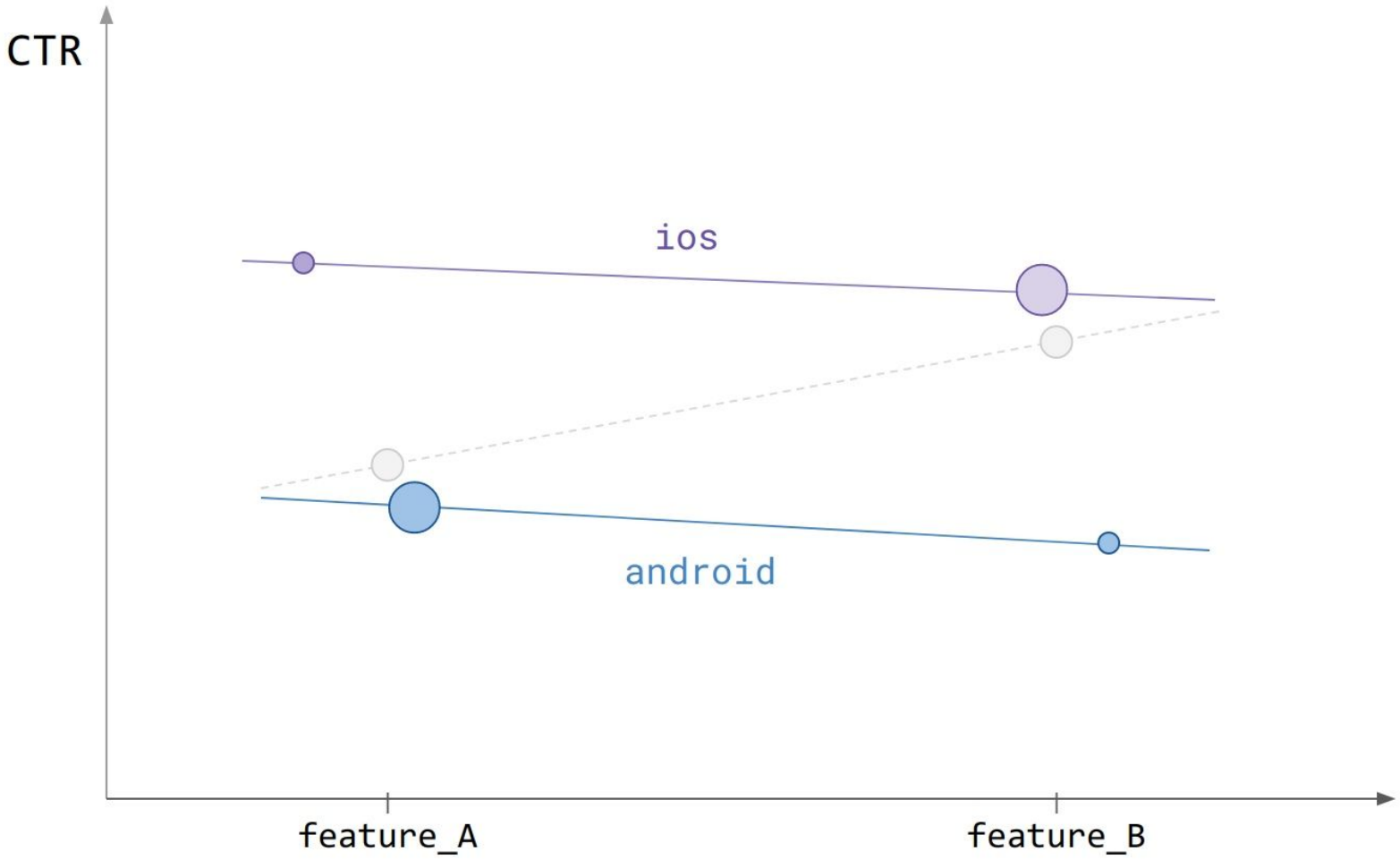
Группа	Объявления	Показы	Клики	CTR в группе
Супер фича	140	190 488	1 727	0.9
Обычные	860	2 798 818	16 218	0.58

Группа	Объявления	Медиана медиан CTR
Супер фича	140	1
Обычные	860	1

Объявления	Показы	Клики	CTR в группе
17	36 417	209	0.57




CTR
feature_A
feature_B


CTR
ios
android
feature_A
feature_B

Больше статистики на stepik.org

100/116 Продолжить

Основы статистики

Bioinformatics Institute

★★★★★ 4.9 92.7K

52/112 Продолжить

Введение в Data Science и машинное обучение

Bioinformatics Institute

★★★★★ 4.9 22.7K