Статистические инструменты для проведения A/B-теста

Инструменты
для работы с ненормальными распределениями

Цель урока

Изучить базовые инструменты для работы с ненормальными распределениями.

Инструменты для работы
с ненормальными распределениями

Задачи урока

- Рассмотреть приёмы удаления выбросов
- Рассмотреть логарифмирование данных
- Рассмотреть Z-преобразование

Инструменты для работы
с ненормальными распределениями

Где встречаются ненормальные распределения?

Почти всегда это:

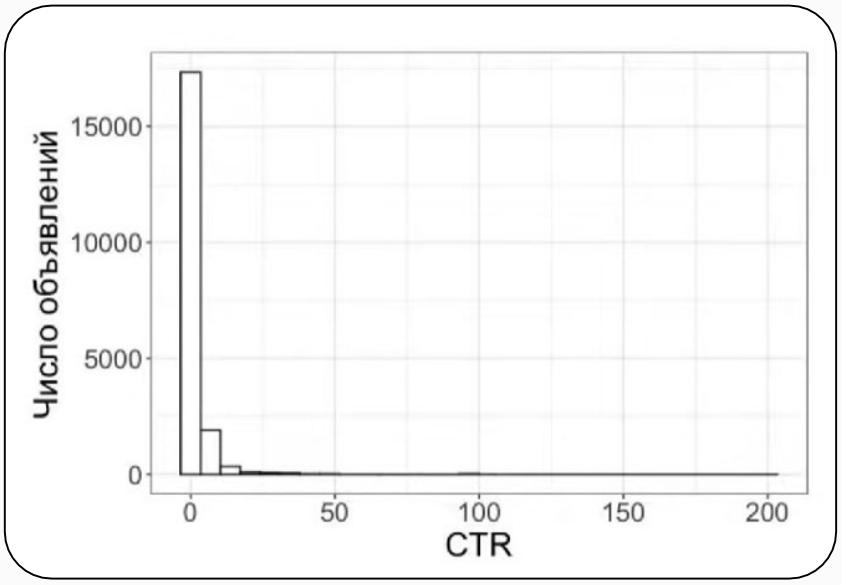
- деньги
- вовлечённость пользователей
- CTR, CPM

Можно попробовать свести к нормальному:

- удалить выбросы
- прологарифмировать

Или применить:

- бутстрап
- бакетинг
- ранговые критерии (Манна — Уитни / Краскела — Уоллиса)

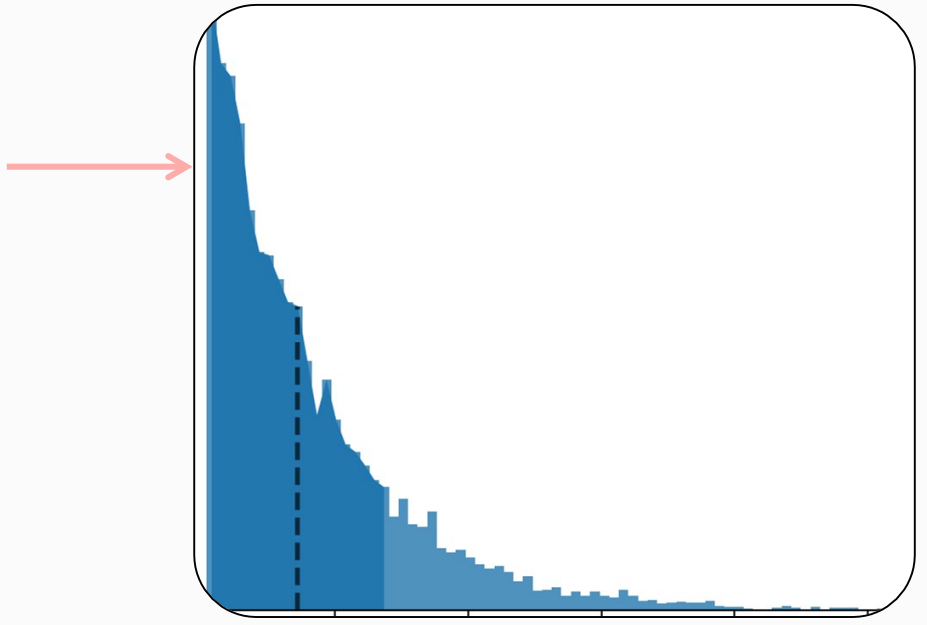


Инструменты для работы
с ненормальными распределениями

Оценка нормальности распределения: Шапиро — Уилка

Как понять, что у вас ненормальное распределение?

1. Можно нарисовать и посмотреть визуально

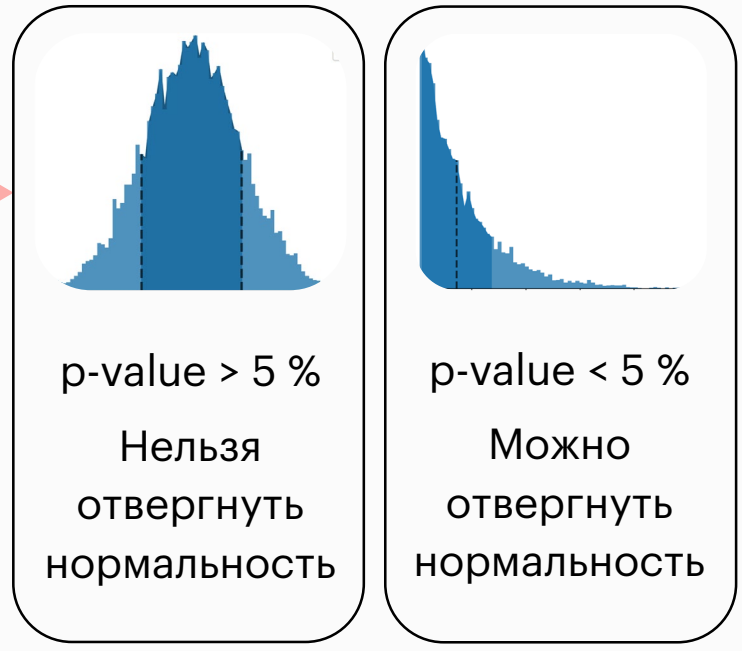


Инструменты для работы
с ненормальными распределениями

Оценка нормальности распределения: Шапиро — Уилка

Как понять, что у вас ненормальное распределение?

2. Воспользоваться специальными критериями. Существуют несколько критериев, которые позволяют ответить на вопрос, насколько ваше распределение похоже на нормальное. Их называют **критериями согласия**



Например: тест Шапиро — Уилка на нормальность.

Нулевая гипотеза: распределение нормальное.

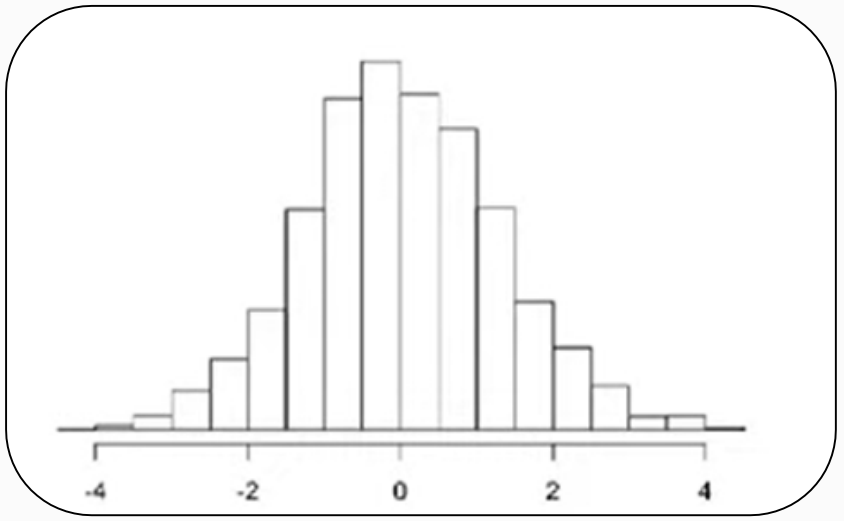
Альтернативная гипотеза: распределение иное.

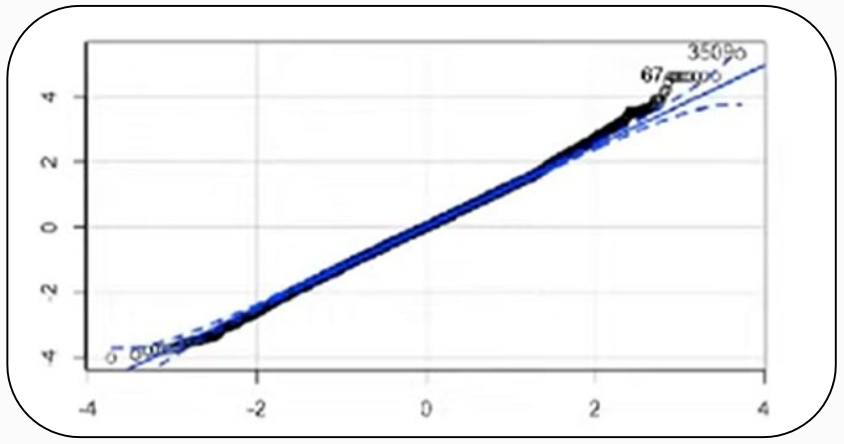
Выдаёт вам **p-value**, которое вы уже умеете интерпретировать.

Инструменты для работы
с ненормальными распределениями

Оценка нормальности распределения

- На больших данных не стоит полагаться только на p-value критериев
- Тест на нормальность выдаёт p-value < 0,001
- Но корреляция между ожидаемыми и предсказанными квантилями — 0,98

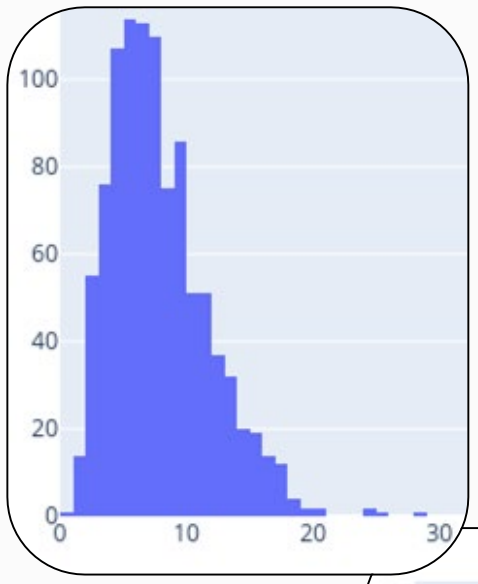


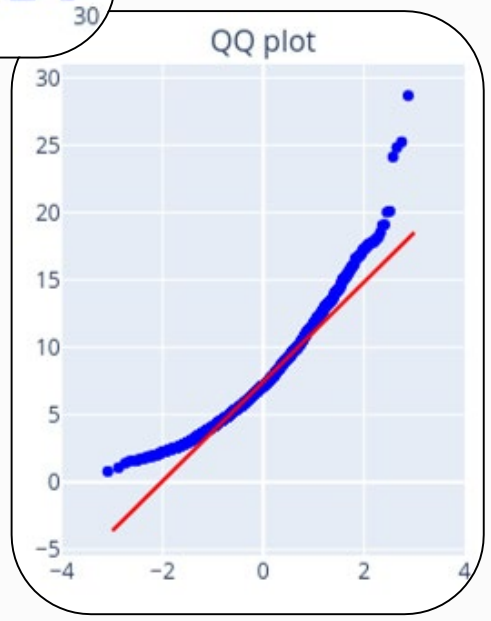


Инструменты для работы
с ненормальными распределениями

Оценка нормальности распределения

- На больших данных не стоит полагаться только на p-value критериев
- Тест на нормальность выдаёт p-value > 0,05
- Но корреляция между ожидаемыми и предсказанными квантилями — 0,68



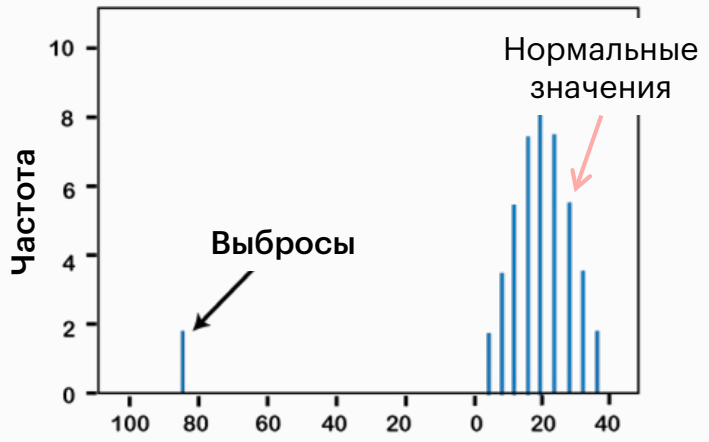


Инструменты для работы
с ненормальными распределениями

Удаление выбросов

При обнаружении выбросов существует несколько стратегий их обработки:

- **игнорировать**: в этом случае чувствительные к выбросам критерии будут давать менее точный результат

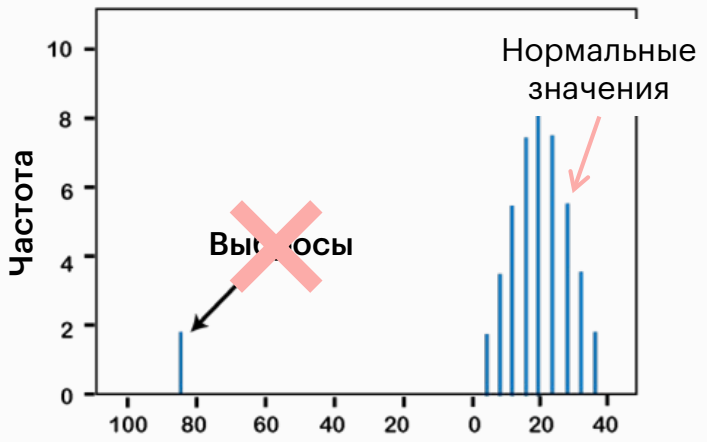


Инструменты для работы
с ненормальными распределениями

Удаление выбросов

При обнаружении выбросов существует несколько стратегий их обработки:

- игнорировать: в этом случае чувствительные к выбросам критерии будут давать менее точный результат
- **удалить совсем**: достаточно жёсткий способ, т. к. теряется часть данных, но самый простой

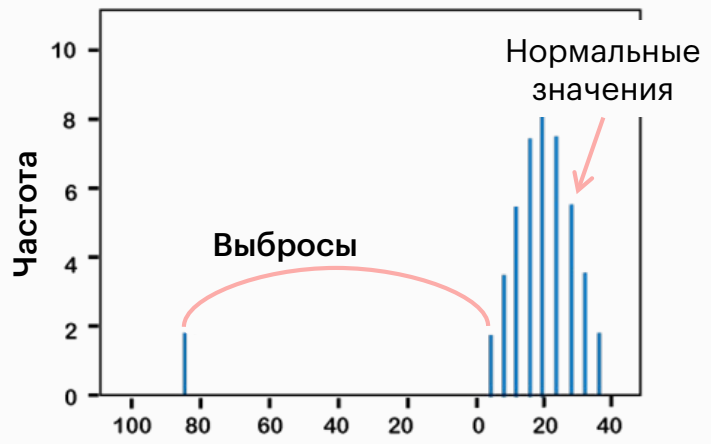


Инструменты для работы
с ненормальными распределениями

Удаление выбросов

При обнаружении выбросов существует несколько стратегий их обработки:

- игнорировать: в этом случае чувствительные к выбросам критерии будут давать менее точный результат
- удалить совсем: достаточно жёсткий способ, т. к. теряется часть данных, но самый простой
- **заменить на какое-то максимальное значение** (нормализация): такой подход сохраняет факт аномального значения, но смягчает его влияние



Инструменты для работы
с ненормальными распределениями

Удаление выбросов

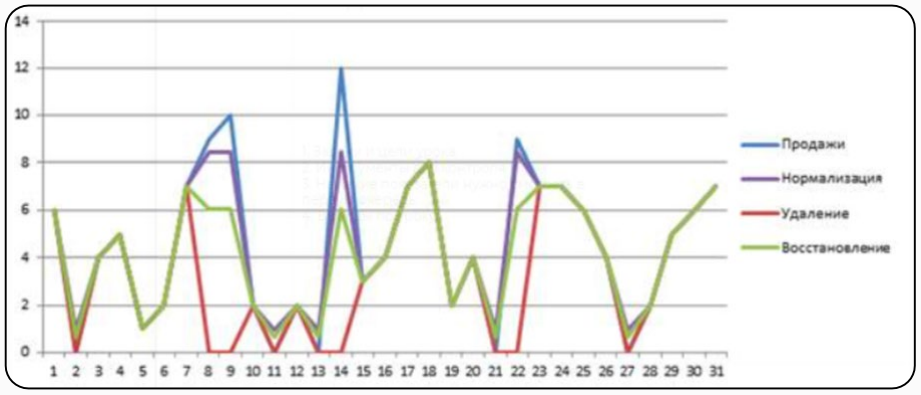
При обнаружении выбросов существует несколько стратегий их обработки:

- игнорировать: в этом случае чувствительные к выбросам критерии будут давать менее точный результат
- удалить совсем: достаточно жёсткий способ, т. к. теряется часть данных, но самый простой
- заменить на какое-то максимальное значение (нормализация): такой подход сохраняет факт аномального значения, но смягчает его влияние
- или заменить на среднее (восстановление/усреднение): ещё более мягкий способ

Инструменты для работы
с ненормальными распределениями

Удаление выбросов

Выбор способа зависит
от ситуации и вида данных.
К примеру, во временных рядах
хорошо подходит усреднение
между двумя соседними
точками. Иногда лучше всего
совсем исключить выбросы
из анализа.



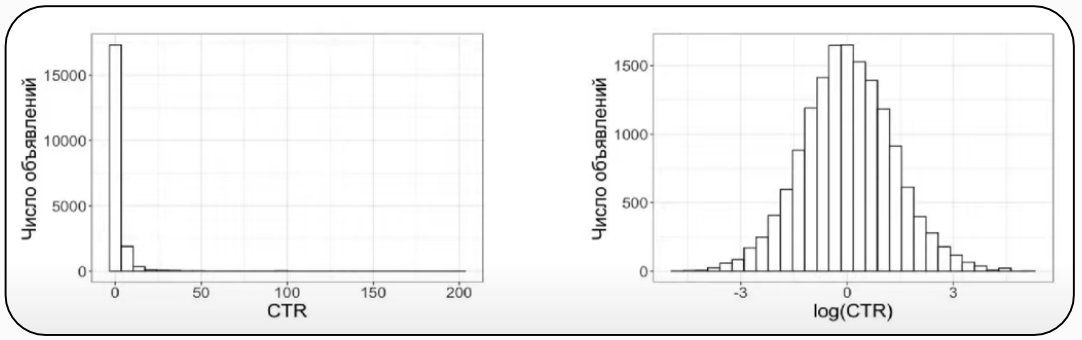
Пример данных с периодом 1 день

Инструменты для работы
с ненормальными распределениями

Логарифмирование

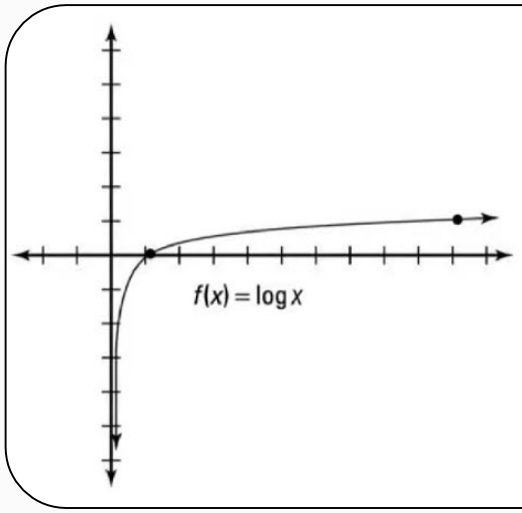
Как это работает?

Просто берём логарифм
от наших данных,
тем самым изменяя
масштаб их шкалы.



Почему
это работает?

Идея в том, что логарифм
медленно сдвигает
малые значения и сильно
сдвигает большие.



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

Инструменты для работы
с ненормальными распределениями

Логарифмирование: итого

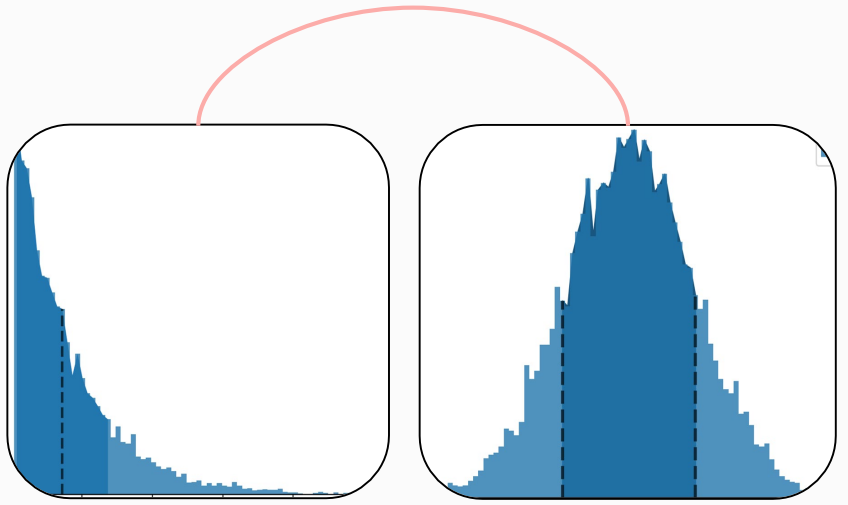
Нормальные распределения — большая редкость, но это не проблема, ведь можно логарифмировать!

Логарифмирование:

+ работает с асимметричными распределениями (картинка)
+ решает проблему выбросов
+ не гарантирует нормальность распределения

Иногда требует предобработки (при наличии отрицательных или нулевых значений в данных).

Логарифмирование



Инструменты для работы
с ненормальными распределениями

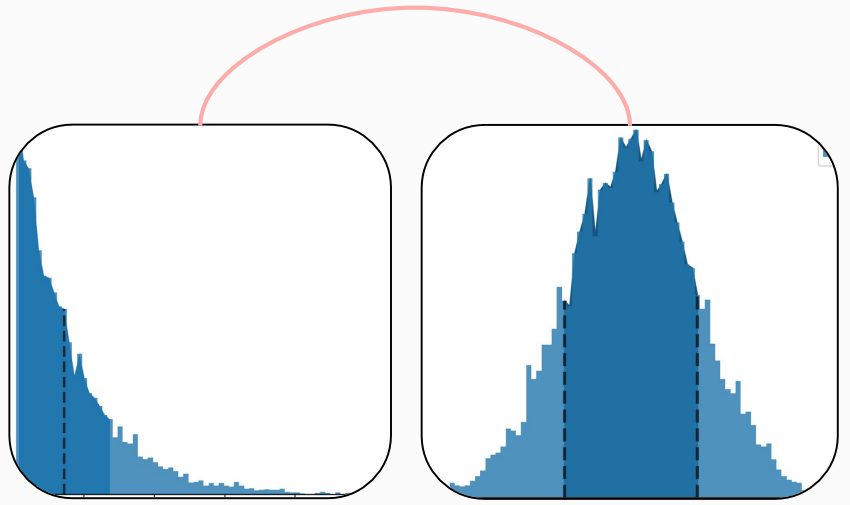
Логарифмирование: итого

Можно использовать другие математические операции, например, возведение в степень или более общий случай — преобразование Бокса-Кокса:

+ можно выправить распределение
+ улучшает визуализацию

Усложняет интерпретацию.

Логарифмирование



Инструменты для работы
с ненормальными распределениями

Z-преобразование

Простое и элегантное преобразование, позволяющее нормировать ваши данные.

Алгоритм такой:

- вычитаем из каждого значения среднее всей выборки
- делим каждое значение на дисперсию всей выборки

$$Z_i = \frac{x_i - \overline{X}}{\sigma}$$

Полученные величины будут иметь среднее 0 и дисперсию 1.

Примечание: такие данные проще использовать в некоторых ситуациях.

Итоги урока

Теперь вы понимаете:

- Какие критерии помогают заподозрить в распределении ненормальность
- Что делать в случае нестандартного распределения в своих данных