Элементы статистики для построения гипотез A/B-тестирования

Доверительный
интервал для оценки параметров и связь с p-value

Доверительный интервал для оценки
параметров и связь с p-value

Цели урока

- Понять, как оценивать выборку доверительным интервалом и при чём здесь p-value
- Узнать, какие есть основные приложения статистики для A/B-тестирования

Доверительный интервал для оценки
параметров и связь с p-value

Задачи урока

- Познакомиться с понятием точечной оценки генеральной совокупности
- Увидеть, что такое доверительный интервал и как им оценивать генеральную совокупность
- Узнать, что такое значение **p-value** и как его использовать
- Познакомиться с примерами использования статистики для тестирования
- Увидеть применение доверительного интервала
- Узнать, как, опираясь на p-value, удалять выбросы из данных

Доверительный интервал для оценки
параметров и связь с p-value

Точечная оценка

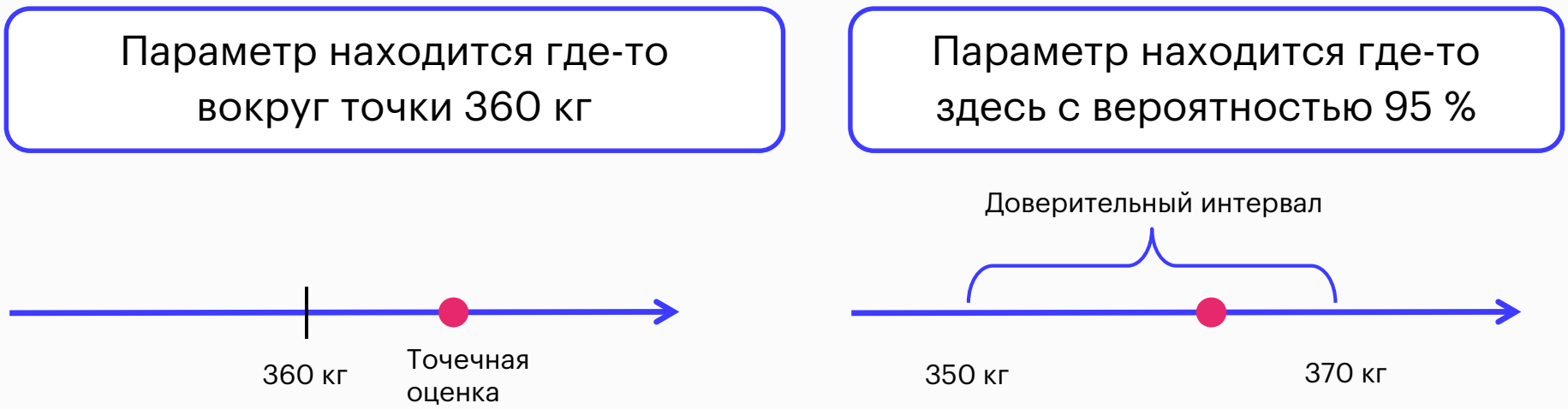
Точечной оценкой называется число, которое используют для оценки параметра генеральной совокупности (среднего).

Пример: мы провели опрос с целью выявить средний вес инопланетян на какой-то планете. Взвешивание 1 000 взрослых инопланетян с этой планеты показало, что средний вес инопланетянина 368 кг. Это только показатель по малой выборке по сравнению со всей планетой, поэтому такая оценка может не всё говорить о реальной картине. Что в таком случае делать?

Доверительный интервал для оценки
параметров и связь с p-value

Что такое
доверительный интервал?

Доверительный интервал — интервал, который покрывает возможные значения параметра с некоторой вероятностью.



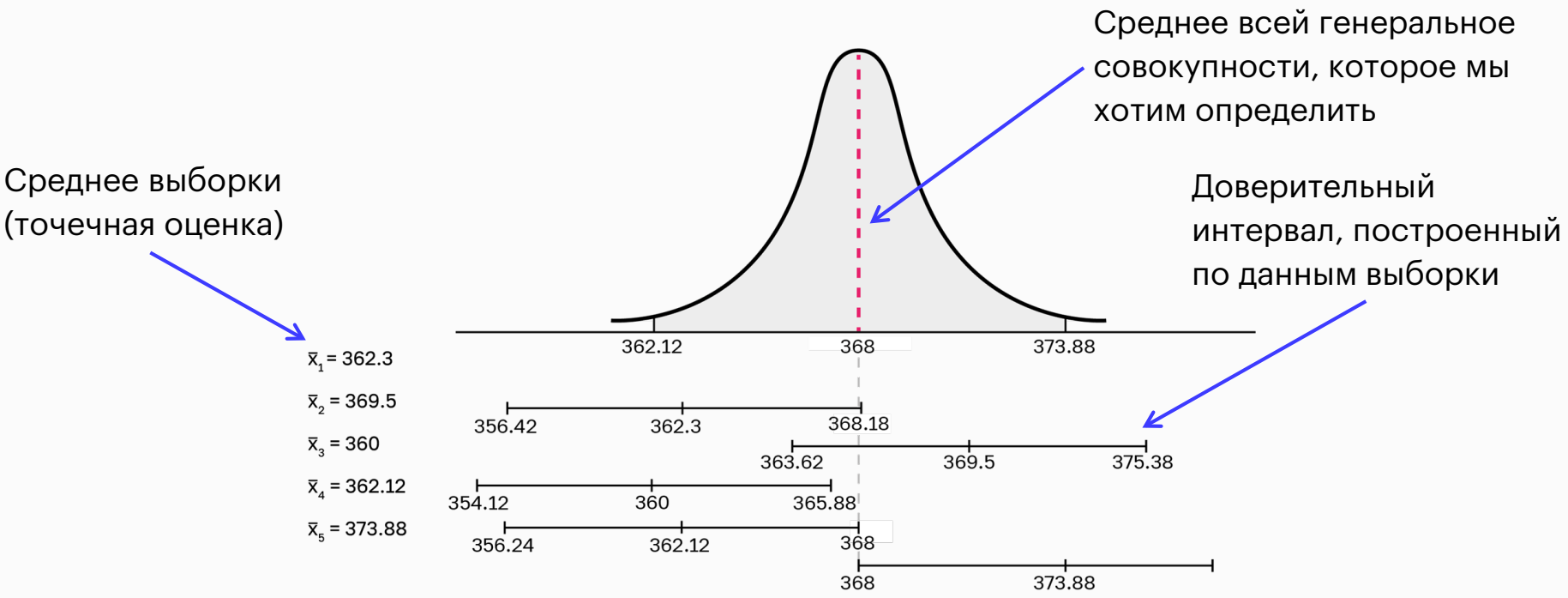
Намного лучше знать не просто оценку 360, а интервал, в котором с большей вероятностью будет находиться реальный средний вес инопланетян. Скажем, с вероятностью 95 % средний вес инопланетян лежит в интервале от 350 кг до 370 кг.

Доверительный интервал для оценки параметров и связь с p-value

Точечная оценка vs доверительный интервал

Доверительный интервал — это интервал, который с заданной вероятностью накрывает оцениваемый нами параметр ГС.

Заметим, что ДИ для разных выборок одной и той же ГС могут отличаться!



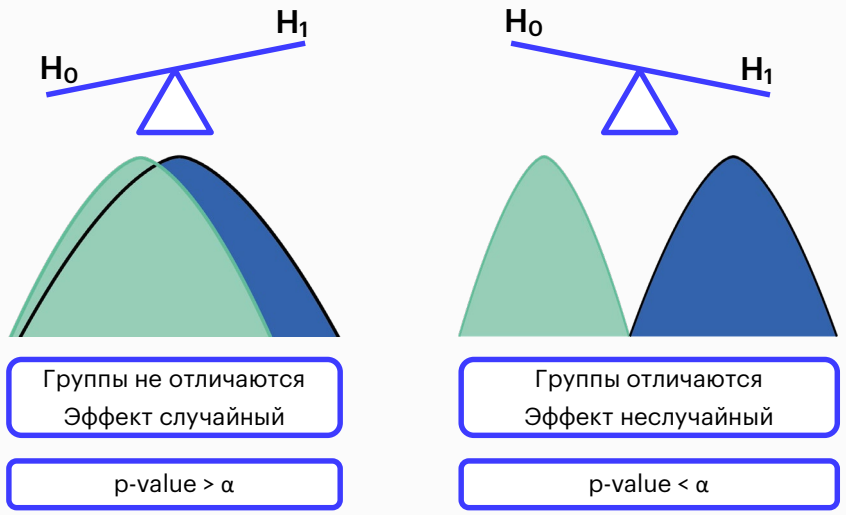
Доверительный интервал для оценки параметров и связь с p-value

Применение p-value

Как по p-value определить, есть ли основания отвергнуть нулевую гипотезу? Тут важно сначала зафиксировать уровень значимости α, а потом уже делать выводы.

p-value — это вероятность отвергнуть нулевую гипотезу при условии, что она верна.

Уровень значимости α — это минимальное значение p-value, на котором нулевая гипотеза может быть отвергнута.



Как правило, за α берётся 5 %

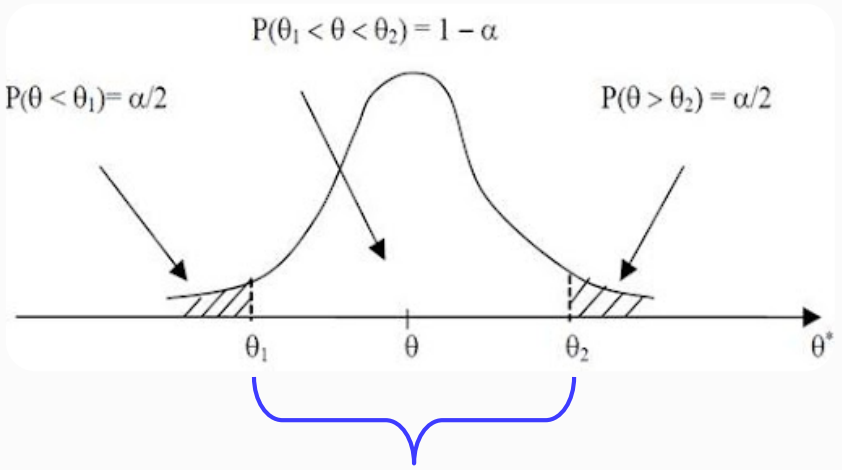
Соответственно, если **p-value** меньше нашего фиксированного **уровня значимости** α, на котором мы проверяем гипотезу, то **нулевую гипотезу** следует отвергнуть, если больше — оснований отвергать нулевую гипотезу нет.

Доверительный интервал для оценки параметров и связь с p-value

Связь p-value
и доверительного интервала

P-value — это вероятность не попасть в доверительный интервал.

Уровень значимости α — это минимальная вероятность не попасть в доверительный интервал.



Доверительный интервал

Доверительный интервал для оценки параметров и связь с p-value

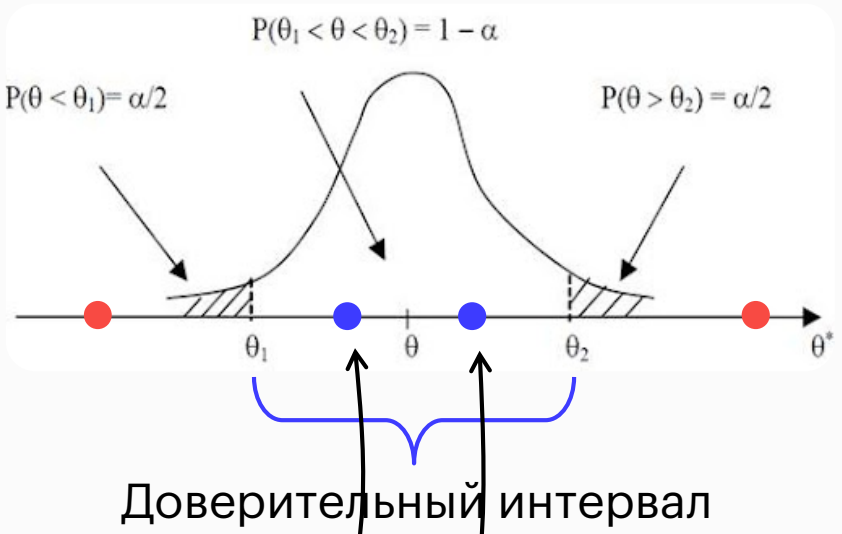
Связь p-value
и доверительного интервала

P-value — это вероятность не попасть в доверительный интервал.

Уровень значимости α — это минимальная вероятность не попасть в доверительный интервал.

Например, красные точки будут иметь **p-value** меньше α, т. к. точки лежат вне доверительного интервала.

Синие точки будут иметь **p-value** больше α, т. к. точки лежат внутри доверительного интервала.



Вычисление p-value

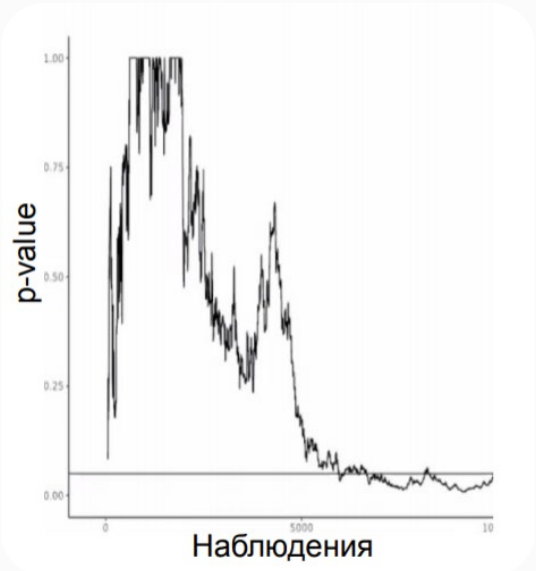
Значение статистики критерия, полученное из выборки, связывают с уже известным распределением, которому она подчиняется, чтобы получить значение p, площадь обоих «хвостов» (или одного «хвоста» в случае односторонней гипотезы) распределения вероятности.

Отличная новость: большинство компьютерных пакетов обеспечивают автоматическое вычисление двустороннего значения p-value (Python). Поэтому вам не придётся считать его самим. Несколько примеров мы разберём на практикуме.

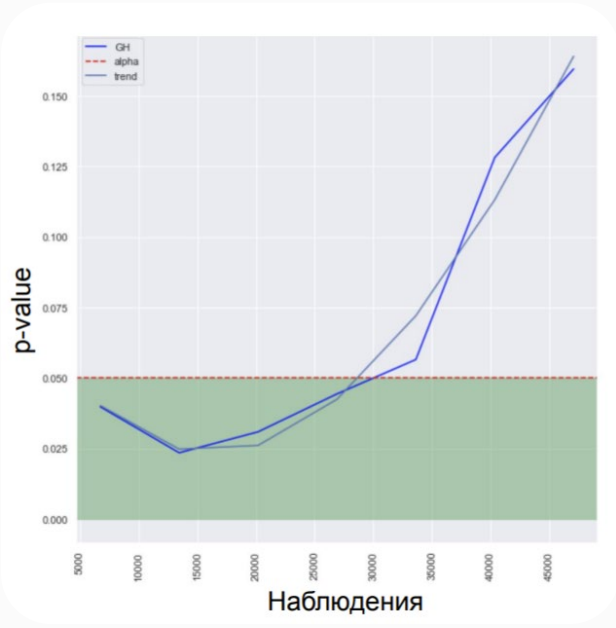
Доверительный интервал для оценки
параметров и связь с p-value

Единое p-value

- Пользователи могут оценить фичу через некоторое время. Сначала значимых различий между группами не будет, но через какое-то время они появятся



- Пользователям сначала не нравилась фича, а потом понравилась. Отсюда различие перестаёт быть значимым



Доверительный интервал для оценки параметров и связь с p-value

Накопительный p-value

Идея: пересчитываем p-value по мере добавления наблюдений.

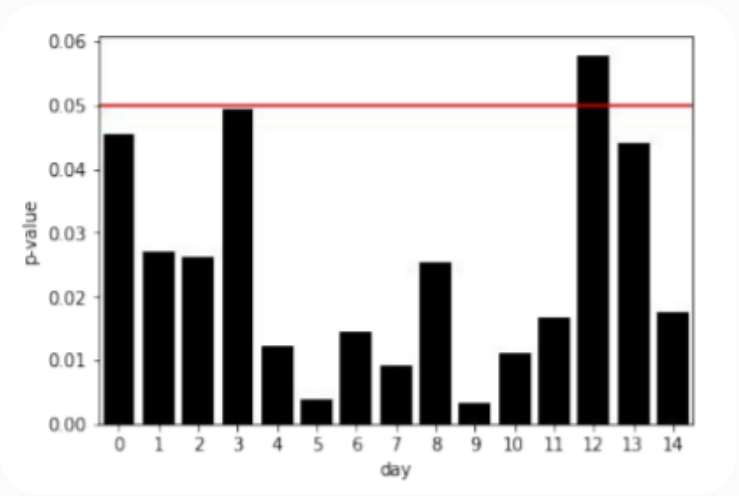
Преимущество: понимаем результаты теста по тренду p-value.

Замечание: на практике общее количество наблюдений бьётся на адекватное количество равных участков, для каждого из которых вычисляем p-value.

Доверительный интервал для оценки параметров и связь с p-value

Накопительный p-value: проблемы

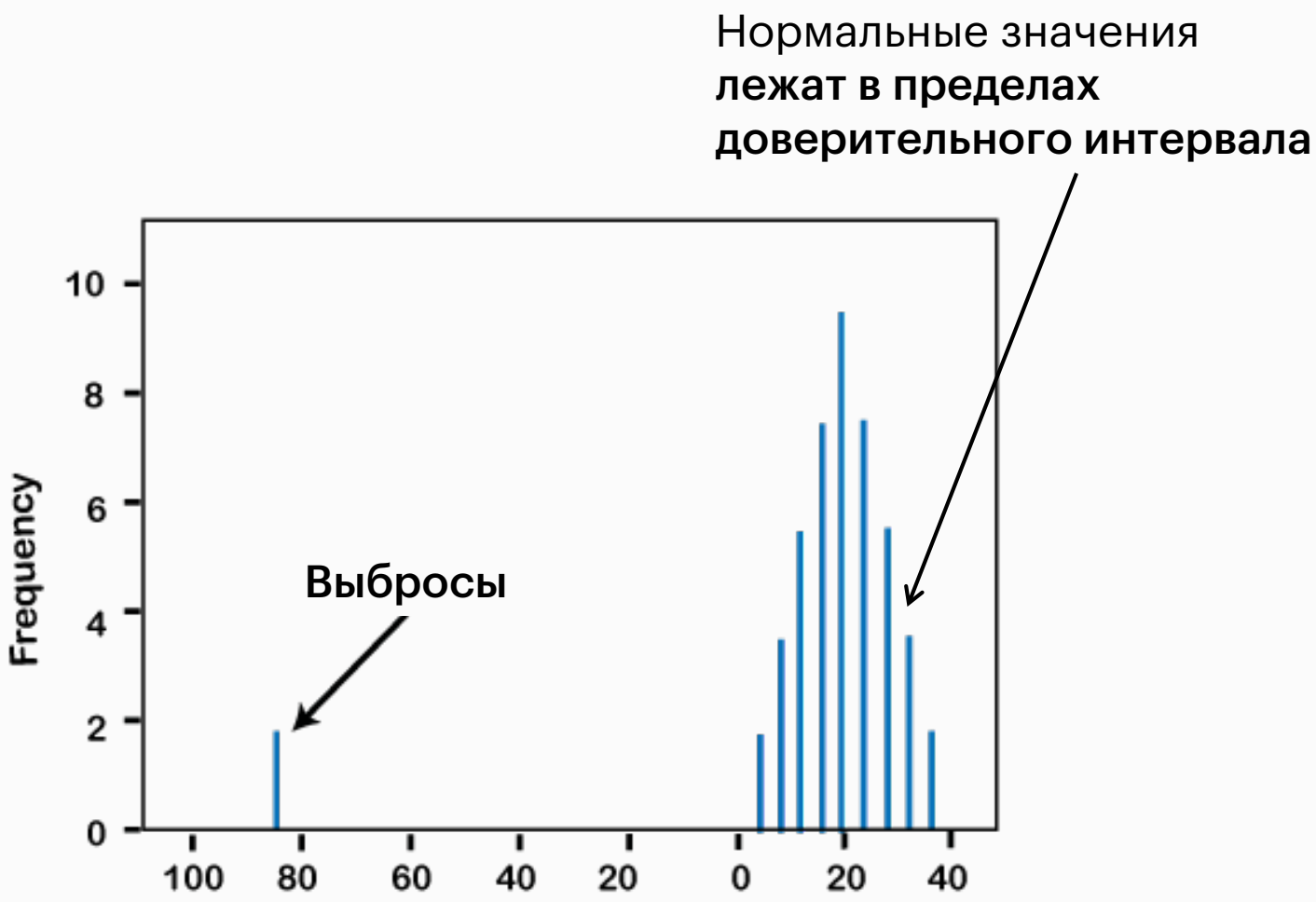
Накопительный p-value может сильно уменьшиться из-за дней с выбросами (дни, в которые поведение пользователей между группами сильно отличалось от остальных дней) и не успеть увеличиться, чтобы не было ошибки I рода. Это может произойти случайно. И это является аргументом в пользу того, что стоит смотреть на тренд.



Можно убрать «дни-выбросы», используя периодический p-value: нужно смотреть на p-value по дням/часам/неделям. Так дни с выбросами отфильтровываются.

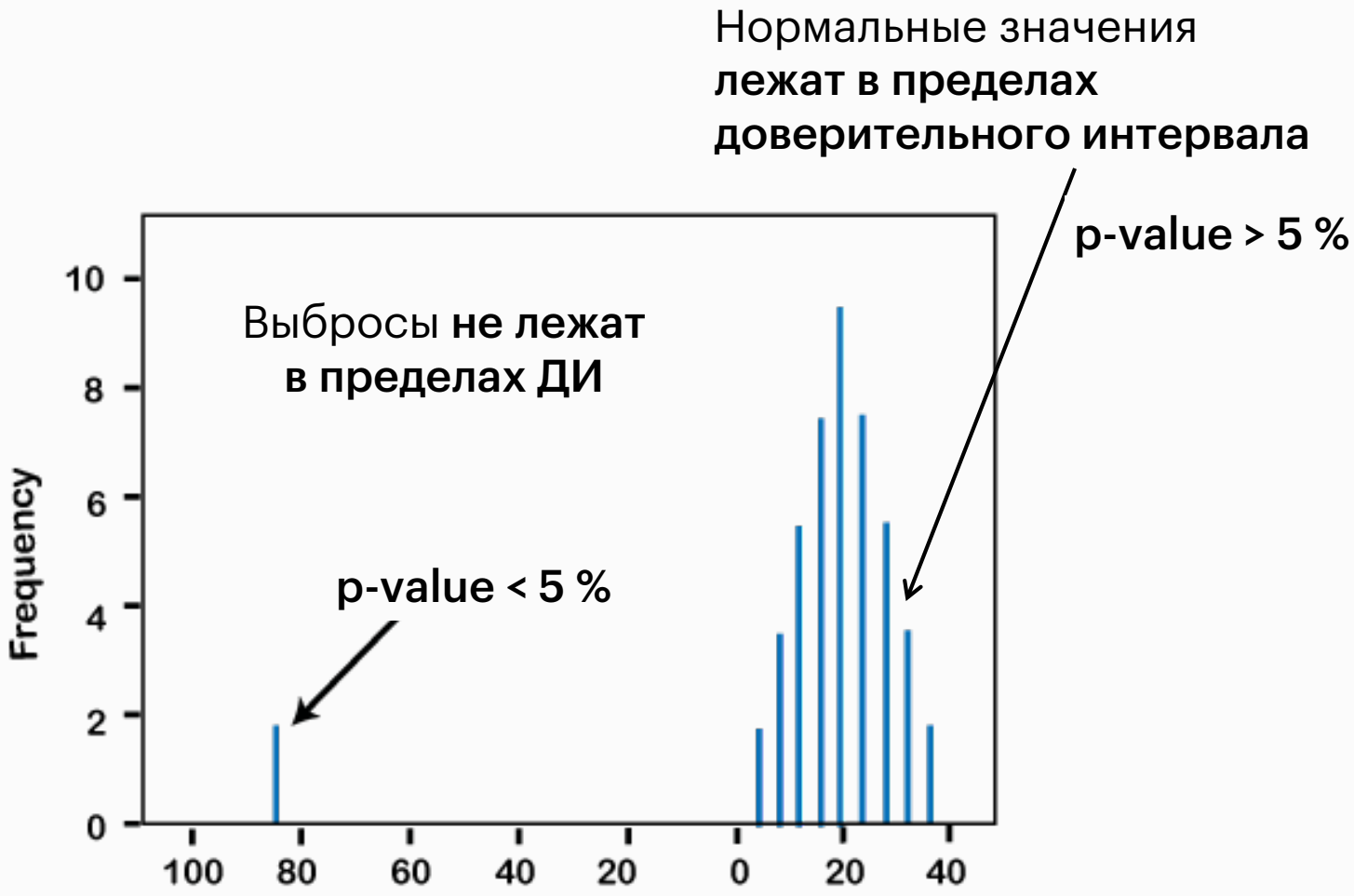
Доверительный интервал для оценки
параметров и связь с p-value

Что если значение
не попало в доверительный интервал?



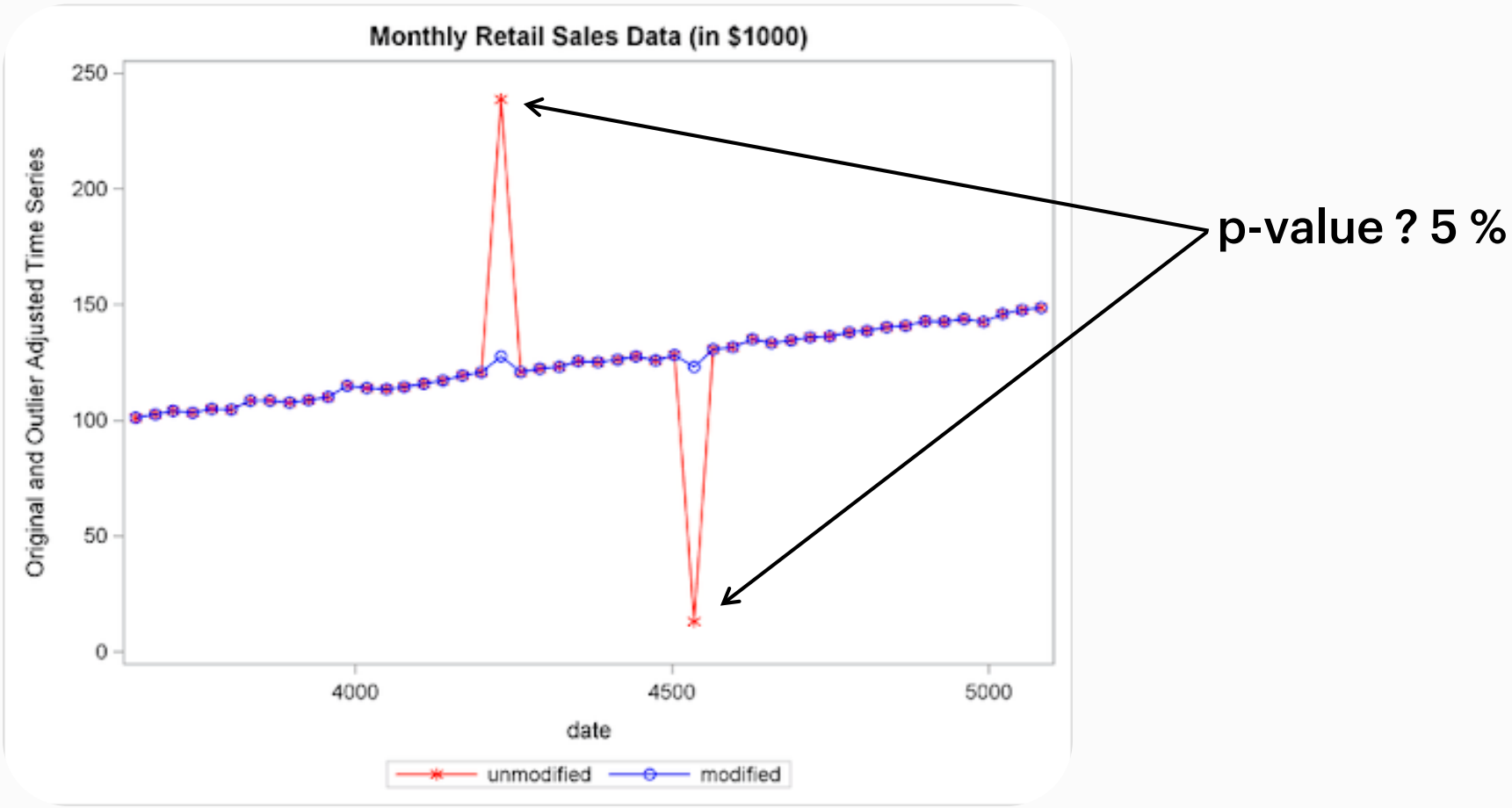
Доверительный интервал для оценки
параметров и связь с p-value

А p-value тут при чём?!
— А как иначе тут определить выброс?



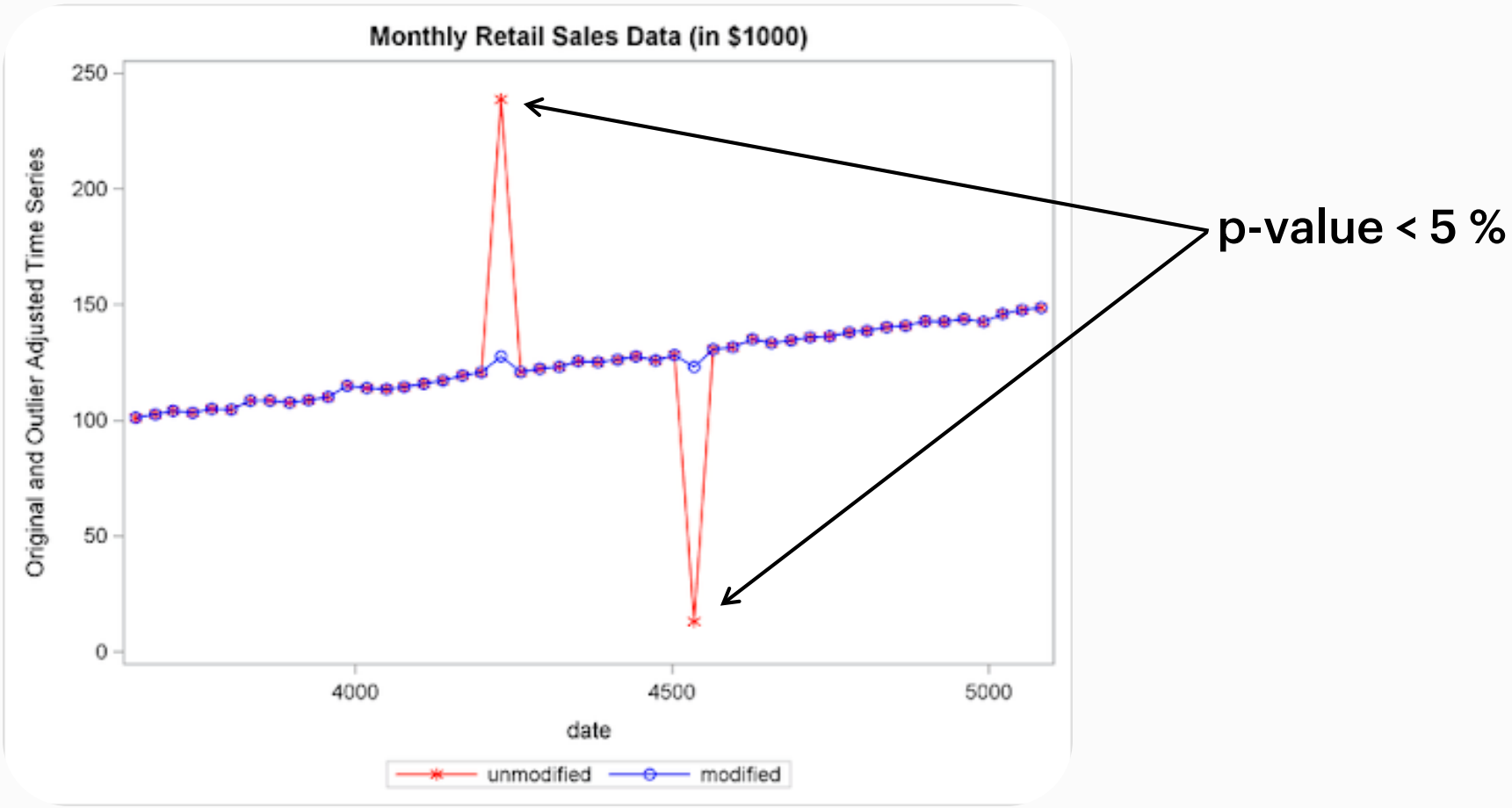
Доверительный интервал для оценки
параметров и связь с p-value

P-value спасает
при работе с временными рядами



Доверительный интервал для оценки
параметров и связь с p-value

P-value спасает
при работе с временными рядами



Доверительный интервал для оценки параметров и связь с p-value

Итоги урока

- Лучше оценивать данные с помощью доверительного интервала
- P-value — это вероятность не попасть в доверительный интервал
- В некоторых случаях стоит использовать накопительный p-value
- Часто **p-value** и ДИ полезны при исключении выбросов из данных
- На практике эти инструменты помогают в работе даже с временными рядами