Python задание 3: бутстрап

1 Что такое доверительный интервал (ДИ) и зачем он нужен?

Повсюду будем предполагать, что у нас есть некоторая выборка $X_1, \ldots, X_n$. Т.е. некоторое подмножество данных, извлечённое из большего набора (генеральной совокупности), используемое для анализа и деления выводов о всей совокупности.

Доверительный интервал (ДИ) — это диапазон значений, который с определенной степенью уверенности включает истинное значение параметра генеральной совокупности на основе выборочных данных. Проще говоря, это интервал, в котором, как мы предполагаем, находится истинное значение параметра, и мы можем указать, насколько уверены в этом предположении (обычно в процентах, например, 95%).

Зачем нужны доверительные интервалы?

1. *Оценка неопределенности.* Когда мы проводим исследования или эксперименты, мы часто работаем с выборкой данных, а не со всей генеральной совокупностью. Из-за этого наши оценки параметров могут быть неточными. Доверительный интервал помогает нам понять степень этой неопределенности.

2. *Интерпретация результатов.* Доверительные интервалы предоставляют более полную информацию, чем просто точечные оценки. Вместо того чтобы сказать, что среднее значение равно 50, мы можем сказать, что среднее значение с 95% уверенностью находится между 47 и 53.

3. Сравнение групп. При сравнении различных групп или условий доверительные интервалы позволяют оценить, значимы ли наблюдаемые различия.

2 Медианное абсолютное отклонение

Медианное абсолютное отклонение (MAD) — это мера разброса данных $X_1, \ldots, X_n$, которая менее чувствительна к выбросам по сравнению со стандартным отклонением. Она определяется как медиана абсолютных отклонений наблюдений от общей медианы выборки:

$$\text{MAD} = \text{median}(\{|X_i - \text{median}(X_1, \ldots, X_n)|\}_{i=1}^n) = \text{median}(\{|X_i - \hat{\mu}_n|\}_{i=1}^n)$$

Однако аналитическое получение доверительных интервалов для MAD может быть сложной задачей. В таких случаях на помощь приходит бутстрап — метод, позволяющий оценивать характеристики статистических распределений без сложных выкладок и предположений об изначальной выборке.

3 Метод бутстрапа

Бутстрап — это вычислительный метод статистики, основанный на принципе повторной выборки с возвращением из исходного набора данных для оценки распределения статистики интереса.

Основные шаги бутстрапа:

1. *Выборка с возвращением.* Из исходного набора данных $X_1, \ldots, X_n$ многократно (M раз) генерируются новые выборки такого же размера путем случайного выбора элементов с возвращением:

$$X^*_{1;1}, X^*_{2;1}, \ldots, X^*_{n;1};$$

$$X^*_{1;2}, X^*_{2;2}, \ldots, X^*_{n;2};$$

$$\ldots$$

$$X^*_{1;M}, X^*_{2;M}, \ldots, X^*_{n;M}$$

2. *Вычисление статистики.* Для каждой бутстрап-выборки вычисляется интересующая нас статистика (в данном случае MAD).

3. *Оценка распределения.* Полученные значения статистики образуют новую выборку, которое используется для оценки доверительных интервалов и других характеристик. Для получания ДИ уровня 95% можно выбросить по 2.5% самых маленьких и самых больших элементов этой выборки, а в качестве ДИ взять наименьший интервал, содержащий все оставшиеся элементы.

   Аналогично можно получить ДИ любого уровня $1 - \varepsilon$.

Этот алгоритм не требует предположений о форме распределения данных/структуре генеральной совокупности, применим к широкому спектру статистик и моделей, а также легко реализуется с помощью современных вычислительных средств.

4 Формулировка задания

Используя метод бутстрапа, оцените 95% доверительный интервал для MAD по заданной выборке. Саму выборку вы можете получить, например, с помощью такого кода:

```
import numpy as np

X = np.random.exponential(scale=1.0, size=10000)
```

Хотя за вами остаётся право использовать и какой-нибудь другой метод.

Кроме этого также используйте метод бустрапа для оценки 95% доверительный интервал для выборочной дисперсии по той же самой выборке. Сравние полученные доверительные интервалы и сделайте выводы.