Практическое задание 1

Файл `conversion.csv` содержит данные для AB-тестирования, проводимого с целью понять, приводит ли переход к новому типу меню к увеличению конверсии сайта.

Столбцы содержат значения 0 и 1, где 1 соответствует пользователю, который кликнул на меню и продолжил знакомство с сайтом, 0 – пользователю, который не стал продолжать знакомство с сайтом. Столбец `OLD` соответствует сайту со старым меню, а столбец `NEW` – сайту с новым меню.

Часть 1

1. Используя R, реализуйте бутстрэп с числом итераций $N = 1500$ для проверки гипотезы об отсутствии различий в значениях конверсии. Посчитайте p-value.
2. Для визуализации полученного p-value постройте, используя `ggplot2`, гистограмму распределения разностей долей, отметив красной вертикальной линией значение разности, равное 0.

Подсказка. Создайте датафрейм `diff_df` с одним столбцом, в котором хранятся значения разностей. Для добавления вертикальной линии используйте слой `geom_vline()`.

3. Проверьте, согласуются ли полученные результаты с интуицией. Другими словами, правда ли разница `p_new - p_old` кажется такой маленькой, как показывает значение p-value и график?

Важно! В уроке у нас результаты совпадали с интуицией: разница невелика, p-value, наоборот, велико, поэтому не отвергаем гипотезу об отсутствии различий. Здесь мы получили что-то странное: разница в долях велика, но и p-value велико!

На самом деле, когда мы используем бутстрэп, мы получаем распределение долей на основе наших данных, которое отличается от распределения долей, ожидаемого в случае, если нулевая гипотеза верна, в случае, если различий действительно нет. Если различий в долях действительно нет, то разница между ними равна нулю. А тогда и значения долей на гистограмме должны располагаться симметрично относительно нуля!

На нашей гистограмме значения симметричны относительно значения, сильно отличного от нуля. Как быть? Для вычисления p-value и построения графика *центрировать разности долей*, то есть из каждого значения в векторе разностей вычесть среднее значение по этому вектору.

4. Используя уточнения из предыдущего пункта, пересчитайте p-value и постройте новый график.
5. Сделайте вывод о том, есть ли различия в значениях конверсии старой и новой версии сайта.

Часть 2

1. Используя Python, реализуйте бутстрэп с числом итераций `N=1500` для проверки гипотезы об отсутствии различий в значениях конверсии. При подсчете p-value не забудьте про центрирование, которое было описано в части 1.

Подсказка: схема реализации алгоритма на Python такая же, как и в R, только вместо вектора перед циклом нужно создать одномерный массив NumPy из пропущенных значений или нулей. Для случайного выбора значений используйте функцию `choices()` из модуля `random`. Пример:

```
import random
random.choices([1, 3, 5, 6], k=2)  # выбор 2х элементов из списка
```

2. Сравните результаты, полученные в R и Python (они могут отличаться – работает случайность выбора). Сделайте выводы о том, правда ли, что новый дизайн сайта лучше старого.