Прикладная статистика и анализ данных

Задание 8

Правила:

- Дедлайн **04 мая 16:30**. После дедлайна работы не принимаются кроме случаев наличия уважительной причины.
- Выполненную работу нужно отправить на почту `mipt.stats@yandex.ru` , указав тему письма `"[asda] Фамилия Имя - задание 8"` . Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- Прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `8.N.ipynb` и `8.N.pdf` , где `N` - ваш номер из таблицы с оценками.
- Решения, размещенные на каких-либо интернет-ресурсах не принимаются. Кроме того, публикация решения в открытом доступе может быть приравнена к предоставлении возможности списать.
- Для выполнения задания используйте этот ноутбук в качестве основы, ничего не удаляя из него.
- Никакой код из данного задания при проверке запускаться не будет.
- В каждой задаче не забывайте делать **пояснения и выводы**.

Баллы за задание:

- Задача 1 - 3 балла
- Задача 2 - 3 балла
- Задача 3 - 6 баллов
- Задача 4 - 7 баллов

Задача 1.

Предположим, что вы разработали лекарство от коронавируса. Перед применением оно обязательно должно пройти клинические испытания. Для начала было разрешено проверить лекарство на двух независимых группах по 10 человек. Одна группа принимает плацебо, другая -- ваш препарат. Большее количество пациентов на первом этапе брать не разрешают -- слишком велики риски отрицательного результата.

Для каждого пациента измерялось количество дней от приема препарата до выздоровления. Получились следуюшие результаты:

In [ ]:

```
x = [6, 16, 8, 13, 9, 4, 7, 10, 3, 14]  # плацебо
y = [5, 10, 3, 1, 5, 3, 19, 2, 2, 5]  # лекарство
```

Что вы можете сказать на основе этих результатов?

- Лекарство эффективнее, подтверждается статистическими методами;
- Наверное, лекарство эффективнее, но статистическими методами это пока не подтверждено, нужно продолжить эксперимент. Подумайте, как обосновать необходимость продолжение эксперимента;
- По результатам эксперимента нельзя сделать какой-либо вывод. Стоит ли продолжать эксперименты? Если да, то четко это обоснуйте;
- Лекарство неэффективно, нужно немедленно прекращать эксперимент.

Задача 2.

В задании 6 по машинному обучению вы предсказывали цену жилья по его характеристикам, в процессе чего принимали участие на Kaggle. Пришло время внедрять разработки в продакшн. На первый взгляд может показаться, что внедрять нужно решение победителя, однако оно может повлечь множество технических сложностей при работе в продакшне, поэтому предлагается выбрать оптимум между качеством и технической сложностью модели. В данной задаче вам предлагается исследовать модели на наличие статистически значимой разницы по их качеству. Не забывайте про практическую значимость результата.

Сравните три модели по качеству предсказания на тестовой выборке: свою и два решения, рассказанных на семинарах. В решении данного задания не нужно приводить код каждой из моделей. Достаточно прочитать 4 файла: истинные ответы на тесте и предсказания трех моделей.

Задача 3.

Сессией в интернете называется промежуток времени, охватывающий работу пользователя с момента открытия первой страницы и до закрытия последней. В каждой сессии пользователь может кликнуть на целевой объект. Предположим, базовая конверсия сессии в клик равна 0.08 (т.е. вероятность клика за сессию). С целью проведения AB-тестирования для 5 процентов аудитории выкатывается новый дизайн, от которого ожидается улучшение конверсии на 0.01, т.е. вероятность клика станет равна 0.09. Оцените двумя способами срок такого AB-теста, считая, что все сессии независимы, а на сайте происходит 1000 сессий в день.

Задача 4.

Имеются две задачи:

1. Пусть $X_i = (X_{i1}, \ldots, X_{id}), i = 1, \ldots, n$ -- выборка из d-мерных объектов. Для всех пар признаков требуется проверить гипотезу о независимости этих признаков, т.е. $\mathsf{H}_{jk}\colon X_{ij}, X_{ik}$ независимы. Для проверки используется критерий на основе коэффициента корреляции Спирмена.
2. Пусть решается задача регрессии и $X_i = (X_{i1}, \ldots, X_{id}), i = 1, \ldots, n$ -- выборка из d-мерных регрессоров, а $Y_1, \ldots, Y_n$ -- соответствующие значения отклика. Требуется отобрать значимые признаки, для чего проверяются гипотезы $\mathsf{H}_j\colon X_{ij}, Y_i$ независимы. Для проверки используется критерий на основе коэффициента корреляции Спирмена.

В обоих задачах предполагается использовать перестановочный критерий, в котором статистикой является коэффициент корреляции Спирмена.

Задание:

- Предложите группу перестановок для реализации данного критерия, а так же метод генерации бутстрепных выборок.
- Можно ли в данных задачах использовать метод множественной проверки гипотезы на основе перестановок (maxT-статистика)? Если нет, то нужно привести пример, для которого нарушается какое-либо требование. Если да, то нужно привести пару примеров, для которых свойство выполняется. Примеры можно привести с помощью семплирования.