Pandas: продолжение. Статистика вывода

Андрей Куртасов
Системный аналитик

Проверка связи

Отправьте «✚», если меня видно и слышно

Если у вас нет звука или изображения:

- перезагрузите страницу
- попробуйте зайти заново
- откройте трансляцию в другом браузере (используйте Google Chrome или Microsoft Edge)
- с осторожностью используйте VPN, при подключении через VPN видеопотоки могут тормозить

План семинара

1. Продолжим практику работы с датасетами в Pandas:
    a. Задача с группировкой и агрегацией
    b. Работа с XML-датасетом
2. Обсудим задания из LMS
3. Обсудим тему тестирования статистических гипотез на примере тестов для сравнения двух выборок

Практика: группировка и агрегация

Упражнение: где больше всего потребляется тот или иной продукт?

Дан набор данных:
https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-02-18/food_consumption.csv

Описание: https://www.nu3.de/blogs/nutrition/food-carbon-footprint-index-2018

Задача: вывести страны, которые потребляют максимум того или иного продукта относительно других стран, с указанием продукта, максимума потребления и страны.

МФТИ

Практика: работа с XML-датасетом

Извлекаем информацию из публичного датасета

Задача: исследовать датасет https://proverki.gov.ru/portal/public-open-data

МФТИ

Разбор заданий из LMS

Неделя 4. Задание 2: расчет доверительного интервала

У вас есть данные возрастов репрезентативной выборки сотрудников компании. Рассчитайте промежуток, в котором будет находиться среднее генеральной совокупности с 99% надежностью?

Тестирование гипотез

Статистические тесты двух выборок

МФТИ

- Цель тестирования – сравнение двух выборок.
- Виды тестов:
  - Параметрические (t-тест Стьюдента, ANOVA)
  - Непараметрические (критерий Манна-Уитни, t-критерий Вилкоксона)
- Инструменты:
  - Электронные таблицы
  - Python
  - R и множество других

Параметрические и непараметрические тесты

МФТИ

- Если исходить из предположения о нормальном распределении, можно использовать t-тесты:
  - `stats.ttest_ind()`
  - `stats.ttest_rel()`
- В случае сомнений о нормальности можно использовать критерий Уилкоксона или Манна — Уитни:
  - `stats.wilcoxon()`
  - `stats.mannwhitneyu()`

Упражнение 1

Тестируется эффективность нового препарата для контроля артериального давления. Массив drug содержит средние значения систолического давления испытуемой группы пациентов за 6 мес. наблюдений, а массив placebo – контрольной группы. Сделать вывод об эффективности препарата на основе этих данных:

```
placebo = [128, 127, 118, 115, 144, 142, 133, 140, 132, 131, 111, 132, 149, 122, 139, 119, 136, 129, 126, 128]
drug = [118, 115, 112, 120, 124, 130, 123, 110, 120, 121, 123, 125, 129, 130, 112, 117, 119, 120, 123, 128]
```

1. Вычислить средние значения по каждой группе.
2. Какой будет нулевая гипотеза?
3. Применить t-тест для сравнения этих двух выборок.

Упражнение 2

Исследователи хотят подтвердить эффективность обезболивающего.
Для этого у 10 пациентов измеряется уровень боли по количественной шкале до (before) и после (after) курса лечения.

```
before = [5, 6, 7, 4, 8, 7, 6, 5, 6, 4]
after  = [3, 4, 6, 2, 5, 4, 3, 3, 5, 2]
```

1. В чем ключевое отличие от предыдущей задачи?
2. Какой будет нулевая гипотеза?
3. Применить t-тест и сделать вывод.

МФТИ

Свободная дискуссия

Ваши вопросы? Пожелания?

До встречи!