

Технологический практикум (практикум на ЭВМ)

316 группа, 2022–2023 уч.гг.

Все задания ниже реализовать для **собственных** данных с использованием языков программирования **R** и **Python**. **Снабдить** программный код необходимыми **комментариями**, включая описание характеристик используемых данных, их источник и т.п.

1. Реализовать аппроксимацию распределений данных с помощью ядерных оценок.
2. Реализовать анализ данных с помощью `cdplot`, `dotchart`, `boxplot` и `stripchart`.
3. Проверить, являются ли наблюдения выбросами с точки зрения формальных статистических критериев Граббса и Q-теста Диксона. Визуализировать результаты.
4. Воспользоваться инструментами для заполнения пропусков в данных. Пропуски внести вручную и сравнить результаты заполнения с истинными значениями.
5. Сгенерировать данные из нормального распределения с различными параметрами и провести анализ с помощью графиков эмпирических функций распределений, квантилей, метода огибающих, а также стандартных процедур проверки гипотез о нормальности (критерии Колмогорова-Смирнова, Шапиро-Уилка, Андерсона-Дарлинга, Крамера фон Мизеса, Колмогорова-Смирнова в модификации Лиллиефорса и Шапиро-Франсия). Рассмотреть выборки малого (не более 50-100 элементов) и умеренного (1000-5000 наблюдений) объемов.
6. Продемонстрировать пример анализа данных с помощью графиков квантилей, метода огибающих, а также стандартных процедур проверки гипотез о нормальности. Рассмотреть выборки малого и умеренного объемов.
7. Продемонстрировать применение для проверки различных гипотез и различных доверительных уровней (0.9, 0.95, 0.99) следующих критериев:

а. Стьюдента, включая односторонние варианты, когда проверяемая нулевая гипотеза заключается в том, что одно из сравниваемых средних значений больше (или меньше) другого. Реализовать оценку мощности критериев при заданном объеме выборки или определения объема выборки для достижения заданной мощности;

б. Уилкоксона-Манна-Уитни (ранговые);

с. Фишера, Левене, Бартлетта, Флигнера-Килина (проверка гипотез об однородности дисперсий).

8. Исследовать корреляционные взаимосвязи в данных с помощью коэффициентов корреляции Пирсона, Спирмена и Кендалла.

9. Продемонстрировать использование методов хи-квадрат, точного теста Фишера, теста МакНемара, Кохрана-Мантеля-Хензеля.

10. Проверить наличие мультиколлинеарности в данных с помощью корреляционной матрицы и фактора инфляции дисперсии.

11. Исследовать зависимости в данных с помощью дисперсионного анализа.

12. Подогнать регрессионные модели (в том числе, нелинейные) к данным, а также оценить качество подобной аппроксимации.

Отчет должен быть выполнен в виде единого файла, желательно в форматах PDF и HTML. Результаты вывода (анализа данных) для R и Python должны аккуратно сравниваться между собой, включая получаемые графики. В случае невозможности объединения допустимо подготовить отдельный файл с отчетом для каждого из языков. Исходный код должен быть полностью представлен либо непосредственно в отчете, либо отдельным приложением. Задание сдается путем отправки материалов на почту a.k.gorshenin@gse.cs.msu.ru до **20 ноября 2022 года** (включительно).