**Разработка и реализация категориальных методов анализа данных типа времени жизни с применением к изучению использования суперкомпьютера**

**Теоретическая часть.**

1. Изучение непараметрических методов анализа данных типа времени жизни: оценки Нельсона—Аалена и Каплана—Мейера

<https://drive.google.com/file/d/1R1RhBbmPUAAzAH_gC18mix7GNkEi1lCU/view>

2. Использование асимптотической нормальности оценок Нельсона—Аалена и Каплана—Мейера для построения категориальных критериев типа хи-квадрат проверки однородности *d* выборок.

(<https://www.amsa.conf.nstu.ru/amsa2013/about-workshop/AMSA2013-proceedings.pdf> (стр. 173-177))

**Реализация.**

Предлагается реализовать методы категориального анализа данных типа времени жизни с использованием среды программирования R (на базе имеющихся предварительных разработок).

**Изучение мощности и эффективности построенных критериев**

1. Создание плана и реализация инструментов моделирования данных типа времени жизни с независимым цензурированием и различными распределениями времен отказов и цензурирования.

2. Исследование мощности построенных критериев с использованием методов Монте-Карло (моделирования).

3. Проведение сравнительного анализ эффективности реализованных критериев с критериями проверки однородности Gρ для 2 выборок.

**Создание R-пакета**

<https://bdemeshev.github.io/r_cycle/cycle_files/20_package.html>

**Анализ результатов работы суперкомпьютера.**

Данные, полученные с суперкомпьютера, организованы в виде двух файлов, один из которых содержит данные заявки пользователя, а второй – результаты работы суперкомпьютера.

*Предварительный этап.* Исследовать структуру имеющихся данных, очистить данные от неинформативных записей. Сформировать модель статистического эксперимента. Удалить записи, неприменимые в рамках построенной модели. Объединить файлы заявок пользователя и результатов работы суперкомпьютера, включая только запуски задач, присутствующие в обоих файлах. Установить зависимость между кодом завершения задачи и переменной, определяющей успешное завершение задачи или снятие незавершенной задачи по той или иной причине. Установить зависимость ID пользователя и области знаний, к которому относится запуск задачи. Дальнейшее исследование провести для исходных данных результатов работы суперкомпьютера и для объединенного файла данных. Установить процентное соотношение суммарных времен исполнения задач и компьютерных времен исполнения задач в различных областях знаний.

*Разведочный анализ данных.* Построить непараматрические оценки распределения усредненного времени выполнения задач (Каплана-Мейера) и времен выполнения задач с разделением по областям знаний. Построить графики и таблицы квартилей. Сравнить визуально результаты, полученные с использованием исходного файла результатов работы суперкомпьютера и отфильтрованного файла запусков, для которых имелись данные заявки пользователя. Выбрать факторы,

*Применение методов категориального анализа.*

Провести проверку значимости влияния выбранных факторов на время исполнения задачи.