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

**(НИР магистратура 1 курс)**

***Этап 2.2:*** **Реализация категориальных критериев типа хи-квадрат проверки однородности d выборок.**

Оформление документации к реализованному на R набору инструментов для проверки гипотез и построения доверительных интервалов:

* использовать документацию к R-функциям в качестве образца
* рассмотреть возможность реализации методов объектно-ориентированного программирования для результатов выполнения функций пакета (S3-объект) с прицелом на создание R-пакета.

***Этап 3.1:*** **Изучение статистических свойств построенных критериев**

Основной мотив: контроль вероятности ошибки I рода

Цели: убедится в работоспособности построенных функций, убедиться в корректности построенных методов и рассмотреть возможность использования методов на малых объемах данных.

**Аргументы планирования эксперимента:** (i) порядок модели (число выборок d); (ii) тип теста; (ii) размеры выборок (n1=n/d); (iii) распределение времен отказов; (iv) распределения времен цензурирования (базовый показатель – процент цензурированных данных q); (v) элементы группировки (brk, dim(brk)=s); (vii) граница критической области

**Выбор аргументов:**

(i). выбор порядка модели: d=2, d=3, d=5

(ii). Рассмотреть 4 типа тестов: на базе оценки Каплана-Мейера (2 теста - прямой и дифференцированный (типа хи-квадрат) test=с("cumulative", "histogram"),type="km"); на базе оценки Нельсона-Аалена (2 теста - прямой и дифференцированный (типа хи-квадрат) test=с("cumulative", "histogram"),type="na")

(iii). выборки одинакового объема (сбалансированный план): n1=10 (очень малые выборки), n1=50 (малые выборки), n1=100 (малые выборки), n1=300 (относительно большие выборки), n1=1000 (большие выборки), n1=10000 (очень большие выборки)

(iv). выбор распределений времен отказа: использовать стандартное показательное распределение

(v).1 формирование независимой выборки из показательного распределения: подбор параметра q=0.1 (малый процент цензурированных данных), q=0.4 (небольшой процент цензурированных данных), q=0.6 (большой процент цензурированных данных); q=0.8 (очень большой процент цензурированных данных) /в дальнейшем, можно будет попробовать использовать распределение Вейбулла или гамма с 1-2 фиксированными параметрами формы для проверки влияния самого распределения цензурирования, а также разные распределения со случайным параметром/

* подбор параметра осуществляется приближенно на экстемально больших выборках (10^6, 10\*10^5)
* можно разделить выборку на 10 или более подвыборок для оценивания СКО числа цензурированных наблюдений

(v).2 использование цензурирования II типа (U\_i=T\_{(1-q)})

(vi). выбор группировки

* (!) м.б. дописать действие при отсутствии наблюдений в группе: ничего не делать (warning), останавливать анализ или объединять с соседней группой.
* число групп s=2, s=3, s=5, s=10.
* элементы группировки выбирать как квантили распределения T: brk[i]=Z\_{T}(i/s) (например, при s=2 dim(brk)=1, brk – медиана стандартного показательного распределения)

(vii). выбор границы критической области: p=0.05, p=0.01, p=0.001

(viii). использование различных сравнений

* по идее выбор сравнения не должен (существенно) влиять на результат

(!) Выбор аргументов может быть скорректирован в процессе проведения эксперимента для получения более информативных результатов.

Моделирование

(i). При фиксированном наборе параметров

* реализуется N циклов моделирования: в идеале, N=100000, но начать надо с 10000, или даже с 1000 (в конце концов, выберем максимальное доступное N)
* проводится группировка (10 групп, каждая подгруппа содержит N/10 циклов) для оценивания дисперсии
* по результатам формируется эмпирическое распределение (10 групп эмпирических распределений) p-значений
* выводится набор квантилей порядков p=0.05,0.01,0.001
* в каждой группе вычисляются квантили порядков p=0.05,0.01,0.001; вычисляются средние значения и среднеквадратические отклонения для каждого значения p

(ii). В каждом цикле моделирования

* моделируется выборка из распределения T размера n
* моделируется выборка из распределения U размера n (за исключением цензурирования II типа)
* формируются цензурированные данные (X,\delta);
* проводится разделение на подгруппы размера n1 путем введения дополнительной переменной group
* формируется объект Surv()
* запускается стандартная функция survfit()

(!) в дальнейшем надо будет написать запуск chisq.surv() на базе объекта Surv()

* запускается chisq.surv() и выводится p-значение

(iii). Вывод результатов

* при каждом наборе параметров выводится строка таблицы

Q5% | QM5% | MSE5% || Q1% | QM1% | MSE1% || Q0.1% |QM0.1% | MSE0.1%

* базовая иерархия признаков

1. тип теста – 4 (2x2) уровня (test=с("cumulative", "histogram"),type=с("km","na"))
2. порядок модели – 3 уровня
3. выбор сравнения –3 уровня для контроля отсутствия влияния сравнений на результат: contr.helmert(); contr.poly(); contr.SAS()-contr.treatment() (или NA)
4. тип цензурирования (случайное, II типа) – 2 уровня
5. размер выборки – 6 уровней
6. процент цензурированных данных – 4 уровня
7. группировка – 4 уровня
8. граница критической области – 1 уровень (в одну строку)

(!) При формировании результатов можно выводить уровни (a-b) в разные таблицы.

(iv). Изучение полученных результатов, написание отчета

* полученные квантили распределений p-значений должны быть близки к соответствующим значениям границ критической области
* при малом числе наблюдений может наблюдаться смещение или большой разброс квантилей распределений p-значений (наиболее нежелательно смещение в меньшую сторону)
* наличие смещения для больших выборок скорее всего сигнализирует о наличии ошибки в коде