

Правительство Российской Федерации Федеральное государственное автономное
образовательное учреждение высшего образования “Национальный исследовательский
институт «Высшая школа экономики»

Кафедра «Компьютерная безопасность»

ОТЧЁТ
К ИНДИВИДУАЛЬНОМУ ЗАДАНИЮ
по дисциплине
“Программирование алгоритмов защиты информации”

Выполнил
Каназирский И.В. СКБ 201

Постановка задачи

Основной задачей проекта являлась реализация статистического теста на однородность хи-квадрат для двух двоичных последовательностей. Тест позволяет определить, имеют ли две двоичные выборки одинаковое распределение значений.

Дополнительно была добавлена возможность проверки одной двоичной последовательности на однородность с равномерным распределением. В данном случае тест оценивает, насколько распределение единиц и нулей в последовательности близко к равномерному распределению.

Таким образом, программа позволяет:

1. Проверять гипотезу об однородности распределений двух двоичных последовательностей.
2. Проверять гипотезу о равномерности распределения значений в одной двоичной последовательности.

Результаты тестирования сравниваются с пороговым значением критерия хи-квадрат для заданного уровня значимости и делается вывод о принятии или отклонении нулевой гипотезы об однородности данных.

Теоретические основы

Проверка однородности двух последовательностей

Для подсчёта статистики хи-квадрат для двух выборок использовалась следующая формула:

$$\hat{\chi}_{n_1, n_2}^2 = n_1 n_2 \sum_{j=1}^N \frac{1}{v_{1j} + v_{2j}} \left(\frac{v_{1j}}{n_1} - \frac{v_{2j}}{n_2} \right)^2$$

где, n_1, n_2 — размеры выборок 1 и 2 соответственно,

v_{ij} — частоты исходов

Проверка одной последовательности на равномерное распределение

Для подсчёта статистики хи-квадрат для одной выборки на однородность с равномерным распределением, использовалась следующая формула:

$$\hat{\chi}_n^2(p) = \sum_{j=1}^N \frac{(v_j - np_j)^2}{np_j}$$

где v_j — частоты значений, np_j — ожидаемые частоты

Работа программы

Программа реализует хи-квадрат тест для проверки распределения бинарных данных. Если на вход передаётся один файл с бинарной последовательностью и уровень

значимости (α), программа проверяет, распределены ли биты в выборке равномерно, то есть содержится ли в последовательности примерно поровну нулей и единиц. Если на вход передаются два файла и уровень значимости, программа сравнивает две выборки на однородность их распределений.

Сначала программа читает бинарные последовательности из указанных файлов. Входные файлы должны содержать только символы 0 и 1, которые могут быть разделены пробелами или новой строкой. Далее программа вычисляет статистику хи-квадрат. Для одной выборки она сравнивает количество нулей и единиц с ожидаемым равномерным распределением. Для двух выборок она анализирует частоты появления нулей и единиц в обеих выборках и определяет, насколько их распределения схожи.

После вычисления хи-квадрат статистики результат сравнивается с пороговым значением, которое зависит от выбранного уровня значимости α : 6.635 для 0.01, 3.841 для 0.05 и 2.706 для 0.1. Если полученное значение хи-квадрат меньше или равно пороговому значению, то гипотеза о равномерности или однородности принимается. В противном случае гипотеза отклоняется.

Запуск программы

Для запуска программы необходимо передать аргументы командной строки в зависимости от задачи, которую нужно выполнить. Программа принимает бинарные последовательности из текстовых файлов (формат ввода обсуждаем и можно по необходимости изменить) и уровень значимости α для проведения хи-квадрат теста.

Если нужно проверить равномерность распределения битов в одной выборке, команда запуска программы выглядит следующим образом:

```
./chi_square_test file1.txt alpha
```

Если необходимо сравнить две выборки на однородность распределений, команда запуска выглядит так:

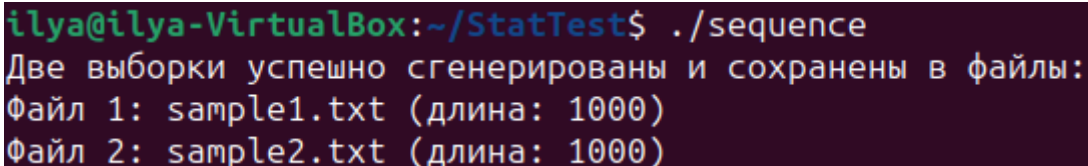
```
./chi_square_test file1.txt file2.txt alpha
```

Файлы, передаваемые в качестве аргументов, должны содержать бинарные последовательности, то есть только символы **0** и **1**, разделённые пробелами или переводами строк.

В результате выдаётся вывод о принятии или отклонении гипотезы об однородности.

Тестирование

Генерация выборок



```
ilya@ilya-VirtualBox:~/StatTest$ ./sequence
Две выборки успешно сгенерированы и сохранены в файлы:
Файл 1: sample1.txt (длина: 1000)
Файл 2: sample2.txt (длина: 1000)
```

Результат работы проверки однородности хи-квадрат

```
ilya@ilya-VirtualBox:~/StatTest/build$ ./ChiSquareTest sample1.txt sample2.txt 0.05
Хи-квадрат для двух выборок: 0.242
Гипотеза об однородности выборок принимается
ilya@ilya-VirtualBox:~/StatTest/build$ ./ChiSquareTest sample1.txt 0.05
Хи-квадрат для одной выборки: 0.1
Гипотеза об однородности выборки принимается
ilya@ilya-VirtualBox:~/StatTest/build$
```