УДК 004.21

МЕТОД ОЦЕНКИ МАЛЫХ ВЕРОЯТНОСТЕЙ В ПЕРМУТАЦИОННЫХ ТЕСТАХ С ИСПОЛЬЗОВАНИЕМ МНОГОУРОВНЕВОГО РАСЩЕПЛЕНИЯ

Голиков Н.Р. (ИТМО)

Научный руководитель – кандидат технических наук, доцент Сергушичев А.А. (ИТМО)

Введение. Пермутационные тесты часто используются для проверки статистических гипотез. Примерами таких тестов являются тесты Колмогорова-Смирнова и Манна-Уитни, проверяющие что две статистические выборки сделаны из одного распределения. Классический метод Монте-Карло не подходит для ситуации, когда вычисляемое *p*-значение очень мало, так как число перестановок, на которых вычисляется значение, становится слишком большим. В работе [1] был разработан метод для оценки малых вероятностей в пермутационных тестах с использованием многоуровневого расщепления [2]. Как было показано, у данного подхода возникали проблемы при работе с сильно дискретными пространствами. В работе [3] нами был разработан метод для оценки низких *p*-значений в дискретных пространствах на примере теста Колмогорова-Смирнова. Метод заключался в решении проблемы с совпадением уровней в многоуровневом расщеплении путем использования хеширования для расширения пространства статистик. Для расширения области применимости метода необходимо было доработать его для случая вещественных значений статистики, как например в задаче анализа представленности генов [4], а также обобщить метод на произвольные пермутационные тесты.

Основная часть. Разработанный алгоритм для оценки низких вероятностей был обобщен для случая вещественных значений статистики. В качестве примера такого случая был использован метод FGSEA [4], для которого тест Колмогорова-Смирнова является частным случаем. Полученное решение с использованием хеширования было внедрено в основную ветку программного пакета.

Был разработан программный пакет на языке Python для оценки малых *p*-значений в произвольных пермутационных тестах. Интерфейс пакета позволяет пользователю определить произвольный пермутационный тест путем реализации метода подсчета статистики. В результате пользователь получает эффективный метод для оценки *p*-значения произвольной величины по значению статистики. Для ускорения работы метода была добавлена возможность реализовать метод на языке C++.

В качестве примера использования пакета были использованы тесты Колмогорова-Смирнова и Манна-Уитни. Для них была продемонстрирована сходимость метода путем сравнения оцененных *p*-значений и точных *p*-значений.

Выводы. Алгоритм для оценки малых *p*-значений в пермутационных тестах был обобщен для случая вещественного значения статистики. Полученный метод был интегрирован в программный пакет FGSEA. Был разработан программный пакет для оценки малых *p*-значений в произвольных пермутационных тестах. Сходимость метода была показана на примере тестов Колмогорова-Смирнова и Манна-Уитни.

Список использованных источников:

1. Сухов В.Д., Короткевич Г.В., Сергушичев А.А. Многоуровневое расщепление в методе Монте-Карло для оценки вероятностей редких событий в пермутационных тестах // Научно-технический вестник информационных технологий, механики и оптики. – 2024.
2. Multilevel Splitting for Estimating Rare Event Probabilities / P. Glasserman [et al.] // Operations Research. — 1999. — Vol. 47, no. 4. — P. 585–600. — DOI: 10.1287/opre.47.4.585.
3. Голиков Н.Р., Сухов В.Д. (науч. рук. Сергушичев А.А.) Исследование многоуровневого расщепления для оценки малых вероятностей в дискретных

пространствах // Сборник тезисов докладов конгресса молодых ученых. Электронное издание. – СПб: Университет ИТМО, [2025]. URL: https://kmu.itmo.ru/digests/article/14011
4. Fast gene set enrichment analysis / G. Korotkevich [et al.] // bioRxiv. — 2021. — Feb. — P. 060012. — eprint: 060012. — URL: https://doi.org/10.1101/060012