Модифицированный Критерий Хеллер-Хеллера-Горфина для проверки однородности

Алексей П. Бузин

Московский государственный университет, Москва, Россия.
E-mail: aleksei.buzin@math.msu.ru

В работе [1] Кресси и Рид ввели общий критерий, обобщающий критерий хи-квадрат. Однако, при применении подобного подхода к данным общего вида существенным аспектом, влияющим на качество критерия, является выбор отрезков разбиения. Эта проблема остается актуальной для всех видов модификаций критериев Кресси-Рида, однако, нас она будет интересовать применительно к задаче однородности.

Один из вариантов решения данной проблемы, значительно повышающих мощность, для критериев проверки однородности и независимости был предложен в работе [2] (будем называть данный подход HHG). В этом подходе было предложено рассматривать всевозможные разбиения и суммировать или брать максимум у соответствующих разбиениям статистик критериев Кресси-Рида (в оригинальной работе рассматривались лишь два частных случая: критерий отношения отношения правдоподобий и хи-квадрат).

Однако, в подходе HHG не удается описать предельное распределение статистики, что приводит к большим вычислительным затратам. В настоящей работе приводится модификация критерия, в которой удается найти предельное распределение статистики для гипотезы однородности.

Рассмотрим задачу проверки гипотезы однородности двух независимых выборок из независимых наблюдений: $X_1, \dots X_{n_1} \sim F_1$ н.о.р., $Y_1, \dots Y_{n_1} \sim F_2$ н.о.р. с непрерывными функциями распределения.

Пусть $n = n_1 + n_2$,

$$\lim_{n\to\infty} n_1/n =: \alpha_1, \ \lim_{n\to\infty} n_2/n =: \alpha_2.$$

Рассмотрим совместное распределение $R(\cdot)$, соответствующее функции распределения $\alpha_1 F_1(\cdot) + \alpha_2 F_2(\cdot)$. Введем статистики

$$D = \sup_{T:\widehat{R}_n(\Delta_i(T))>\varepsilon} \chi^2(T), \quad D' = \frac{1}{\{\{\Delta\} : \widehat{R}_n(\Delta_i(T)) > \varepsilon\}} \sum_{T:\widehat{R}_n(\Delta_i(T))>\varepsilon} \chi^2(T),$$

где $\widehat{R}_n$ — совместное эмпирическое распределение выборок, $\chi^2(T)$ статистика хи-квадрат от разбиения T, ε — фиксированное неотрицательное число.

Теорема *При гипотезе D имеет невырожденное предельное распределение, для которого получено явное выражение.*

В докладе будет показана стостоятельность критерия, построенного по статистике D и рассмотрено предельное распределение статистики D при гипотезе и альтернативе.

Также в докладе будет обсуждаться свойства статистики D' и поведение статистик в случае $\varepsilon = \varepsilon(n) \to 0$ (при $n \to \infty$).

Литература

1. Cressie N., Read T. R. C. *Multinomial goodness-of-fit tests* //Journal of the Royal Statistical Society Series B: Statistical Methodology. – 1984. – Т. 46. – №. 3. – С. 440-464.

2. Heller R. et al. *Consistent distribution-free K-sample and independence tests for univariate random variables* //Journal of Machine Learning Research. – 2016. – Т. 17. – №. 29. – С. 1-54.