Секция «19.14 Теория вероятностей и математическая статистика»

Адаптивный критерий хи-квадрат для простой гипотезы

Научный руководитель – Хиль Елена Викторовна

Андреев Данила Евгеньевич
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра математической статистики и случайных процессов, Москва, Россия
E-mail: danila.andreev@math.msu.ru

Предположим, что у нас есть выборка независимых одинаково распределённых случайных величин $X_1, \ldots, X_n$ с неизвестной функцией распределения F_1. Рассмотрим задачу проверки согласия:

$$H_0 : F_1 = F_0 \quad \text{против} \quad H_1 : F_1 \neq F_0.$$

Пусть P_i обозначает вероятностную меру, соответствующую F_i, $i = 0, 1$. Разобьём прямую на N интервалов Δ_i, $i = 1, \ldots, N$, равновероятных относительно P_0. Пусть ν_i — число наблюдений, попавших в i-й интервал, $i = 1, \ldots, N$. Классический критерий согласия χ^2 впервые был введён Карлом Пирсоном в 1900 году [1]. Статистика критерия в нашей модели имеет следующий вид:

$$\chi_n = \sum_{i=1}^{N} \frac{(\nu_i - np)^2}{np}.$$

Предположим, что $F_1 \neq F_0$, то есть нулевая гипотеза не верна. Пусть $p_i^{(1)}$ — вероятность интервала Δ_i при P_1, $i \in 1, \ldots, N$. Даже если F_1 и F_0 различаются, их вероятности внутри каждого интервала могут быть почти одинаковыми; то есть $p_i^{(1)} \approx 1/N$ для всех i. Поскольку χ^2-критерий сравнивает частоты только по этим интервалам, он может не обнаружить различие между F_1 и F_0. Чтобы снизить влияние этот эффекта, мы используем разбиения на N интервалов. Для каждого разбиения мы группируем интервалы в k ячеек, причём каждая ячейка является объединением одного или нескольких соседних интервалов. Для этого разбиения мы вычисляем χ^2-статистику. Итоговая статистика есть сумма этих малых статистик. Такой подход позволяет учитывать различия, проявляющиеся на нескольких интервалах, и помогает различать F_1 и F_0. Пусть S_n обозначает статистику критерия. Представим S_n в виде

$$S_n = X_n A X_n^T,$$

где

$$X_n = \left(\frac{\nu_1 - np}{\sqrt{np}}, \ldots, \frac{\nu_N - np}{\sqrt{np}} \right),$$

а $A = \{\alpha_{ij}\}$ — матрица размера $N \times N$. Элементы α_{ij} зависят от параметров N и k. Наш основной результат следующий. Теорема. При выполнении нулевой гипотезы имеет место:

$$S_n \xrightarrow{d} Z^T M Z, \quad n \to \infty,$$

где Z — столбец размерности $N-1$ со стандартным многомерным нормальным распределением, а M — симметричная положительно определённая матрица размера $(N-1)\times(N-1)$. Явный вид матрицы M будет приведён в докладе.

Источники и литература

1) Pearson K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling // Philosophical Magazine. Series 5. Vol. 50(302). P. 157–175. DOI: 10.1080/14786440009463897.