12 Критерии хи-квадрат и отношения правдоподобий для дискретных данных

Статистика критерия хи-квадрат для простой гипотезы выглядит как

$$\sum_{i=1}^{k} \frac{(\nu_i - np_i)^2}{np_i}.$$

При верной гипотезе она сходится к χ^2_{k-1}. Статистика критерия отношения правдоподобий предлагает взамен брать

$$\sum_{i=1}^{k} \nu_i \ln \frac{\nu_i}{np_i}.$$

Наконец, более общий подход Кресси-Рида предлагает рассматривать

$$\frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} \nu_i \left(\left(\frac{\nu_i}{np_i} \right)^{\lambda} - 1 \right),$$

где при $\lambda \in \{-1, 0\}$ выражение доопределяется из соображений по непрерывности.

1. Начнем с проверки простой гипотезы.

   Найдите первые 1000 цифр числа π после запятой. С помощью критерия хи-квадрат проверьте, можно ли при уровне значимости 0.05 считать эти цифры случайными равномерными?

2. Проверим однородность и независимость. Использовать данные из файла Priem.csv и встроенный критерий.

   (a) Ответить на вопрос - отличаются ли мальчики и девочки в плане успешности сдачи ЕГЭ? Для этого попарно проверьте на однородность суммарные баллы, баллы по русскому, баллы по математике.

   (b) Правда ли, что оценки по математике и русскому независимы?

3. Построим критерий Кресси-Рида для проверки простой гипотезы о полиномиальном распределении. Давайте сравним наши критерии для различных λ. Рассмотрите λ, равные $-1, 0, 0.5, 1, 2$. Постройте график p-value для каждого из них и выберите наиболее удачный критерий. Используйте исходное равномерное распределение (все p_i равны) и неравномерное на свой вкус.

4. * Переходим к параметрической гипотезе.

   Среди 2020 семей, имеющих 2 детей, 527 семей, в которых 2 мальчика, и 476 - две девочки. Можно ли при уровне значимости 0.05 считать, что количество мальчиков – биномиальная случайная величина?

   В этой задаче нужно сначала найти ОМП для параметрической гипотезы (формулу для нее) на листочке, затем вычислить ее значение для данных из условия задачи и воспользоваться встроенным критерием хи-квадрат.

   Теперь исследуем работу получившегося критерия на модельных данных (нужно использовать ту же формулу для ОМП, что и раньше). Рассмотреть следующие распределения:

   1) $Binom(2, 1/2)$, $Binom(2, 1/8)$,

   2) равномерное распределение $\mathcal{R}\{0, 1, 2\}$,

   3) $\mathbf{P}(X = 0) = \mathbf{P}(X = 2) = 3/8, \mathbf{P}(X = 1) = 1/4$ (X – число мальчиков).

   Для каждого распределения сгенерировать по 100 выборок. К каждой выборке применить построенный критерий, получить p-value. Построить графики p-value для каждого распределения, сравнить их.