

38. Критерий хи-квадрат.



Status

Completed

Теорема Пирсона

$X \sim F$, проверяются гипотезы как обычно.

Разобьем область возможных значений X_1 на k непересекающихся промежутков Δ_j .

$$p_j = P_{H_0}(X_1 \in \Delta_j),$$

ν_j - число наблюдений, попавших j -ый промежуток,

$$\Psi_n = \sum_{j=1}^k \frac{(\nu_j - np_j)^2}{np_j}$$

Если $0 < p_i < 1$ при всех $i = 1, \dots, k$, то для любого $y > 0$, при верной H_0

$$\Psi_n \Rightarrow \eta \sim \chi_{k-1}^2.$$

Докажем теорему Пирсона при $k = 2$.

В этом случае $\nu_2 = n - \nu_1$, $p_2 = 1 - p_1$. Посмотрим на ρ и вспомним ЦПТ:

$$\begin{aligned} \rho(\mathbf{X}) &= \frac{(\nu_1 - np_1)^2}{np_1} + \frac{(\nu_2 - np_2)^2}{np_2} = \frac{(\nu_1 - np_1)^2}{np_1} + \frac{(n - \nu_1 - n(1 - p_1))^2}{n(1 - p_1)} = \\ &= \frac{(\nu_1 - np_1)^2}{np_1} + \frac{(-\nu_1 + np_1)^2}{n(1 - p_1)} = \frac{(\nu_1 - np_1)^2}{np_1(1 - p_1)} = \left(\frac{\nu_1 - np_1}{\sqrt{np_1(1 - p_1)}} \right)^2 \end{aligned}$$

Но величина ν_1 есть сумма n независимых случайных величин с распределением Бернулли B_{p_1} , и по ЦПТ

$$\frac{\nu_1 - np_1}{\sqrt{np_1(1 - p_1)}} \Rightarrow \xi,$$

где ξ имеет стандартное нормальное распределение. Поэтому

$$\rho(\mathbf{X}) = \left(\frac{\nu_1 - np_1}{\sqrt{np_1(1 - p_1)}} \right)^2 \Rightarrow \xi^2.$$

Величина ξ^2 имеет χ^2 -распределение H_1 с одной степенью свободы. □

Критерий Хи-квадрат

Проверяет, что распределение выборки равно некоторому известному.

Пусть $\vec{X} \sim F$

$$H_0 = \{F = F_0\}$$

$$H_a = \{F \neq F_0\}$$

$$\Psi_n = \sum_{j=1}^k \frac{(\nu_j - np_j)^2}{np_j}$$

$$\delta = \begin{cases} 0, & \Psi_n < c \\ 1, & \Psi_n \geq c \end{cases}$$

где $\chi_{k-1}^2(c) = 1 - \epsilon$.



Метод разбиения на промежутки тут не указан, но он важен! Если сделать 1 промежуток, то, понятно, критерий будет всегда срабатывать, а если слишком много — никогда. Постройте аналогию с гистограммой. Также, если распределение имеет артефакты внутри промежутков, вы это никак не узнаете, так что, по сути, проверяется немного другая гипотеза.

Свойства

1. Критерий состоятельный

Упражнение. Вспомнить закон больших чисел и доказать, что если H_1' неверна, то найдется $j \in \{1, \dots, k\}$ такое, что

$$\frac{(\nu_j - np_j)^2}{np_j} = \frac{n}{p_j} \left(\frac{\nu_j}{n} - p_j \right)^2 \xrightarrow{p} \infty.$$

Осталось построить критерий в соответствии с [K2](#).

Понимаем, что $\frac{\nu_j}{n} = (I(X_1 \in \Delta_j) + \dots + I(X_n \in \Delta_j))/n \stackrel{\text{ЗБЧ}}{=} EI(X_1 \in \Delta_j) = p_j$ (при H_0). Иначе — разность будет ненулевой, и с увеличением n как минимум одно слагаемое будет стремиться к бесконечности.

2. Критерий имеет асимптотический размер $1 - \epsilon$.
3. Реально допустимый уровень значимости: $\epsilon^* = P(\eta \geq \Psi_n) = 1 - \chi_{k-1}^2(\Psi_n)$.