Взаимосвязи порядковых признаков

А. С. Пушняков
pushnyakovalex@mail.ru

31 марта 2014 г.

Порядковые признаки

- Даны выборка $\{X^\ell\}_{\ell=1}^N$ и два признака K_1 и K_2: $X_1^\ell \in K_1$, $X_2^\ell \in K_2$.
- Если множества K_1 и K_2 конечные и на них задано отношение порядка, то соответствующие признаки будем называть порядковыми. Можно считать, что $K_1 = \{1, 2 \dots k_1\}$, $K_2 = \{1, 2 \dots k_2\}$.
- Вопрос: есть ли монотонная зависимость между признаками K_1 и K_2?

Таблица сопряженности

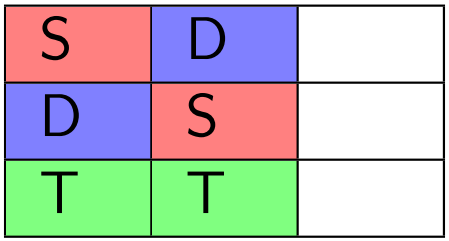
K_1 \ K_2	1	…	j	…	k_2	$\sum$
1						
⋮						
i			R_{ij}			$R_{i\bullet}$
⋮						
k_1						
$\sum$			$R_{\bullet j}$			N

$R_{ij} = |\{X^\ell |\ X_1^\ell = i,\ X_2^\ell = j\}|$ — число элементов выборки, для которых первый признак равен i, а второй — j.

Согласованные пары

Пара клеток (i_1, j_1) , (i_2, j_2) называется

- *согласованной*, если $i_1 > i_2$ и $j_1 > j_2$ или $i_1 < i_2$ и $j_1 < j_2$.
- *несогласованной*, если $i_1 > i_2$ и $j_1 < j_2$ или $i_1 < i_2$ и $j_1 > j_2$.
- *связанной*, если $i_1 = i_2$ или $j_1 = j_2$.



- $P_s = 2\sum\limits_i \sum\limits_j R_{ij} \sum\limits_{i'>i} \sum\limits_{j'>j} R_{i'j'}$ — число согласованных пар.
- $P_d = 2\sum\limits_i \sum\limits_j R_{ij} \sum\limits_{i'>i} \sum\limits_{j'<j} R_{i'j'}$ — число несогласованных пар.
- $P_t = N^2 - P_s - P_d = \sum\limits_i R_{i\bullet} + \sum\limits_j R_{\bullet j} - \sum\limits_i \sum\limits_j R_{ij}$ — число связанных пар.

- Вместо таблицы сопряженности R_{ij} рассмотрим таблицу вероятностей ρ_{ij}.
- Обозначим вероятности согласованных, несогласованных и связанных пар как Π_s, Π_d и Π_t соответственно.
- γ-коэффициент — это разность условных вероятностей согласованных и несогласованных пар при условии, что пара не является связанной:

$$\gamma = \frac{\Pi_s - \Pi_d}{1 - \Pi_t}.$$

- МНП оценка γ-коэффициента:

$$G = \frac{P_s - P_d}{N^2 - P_t}.$$

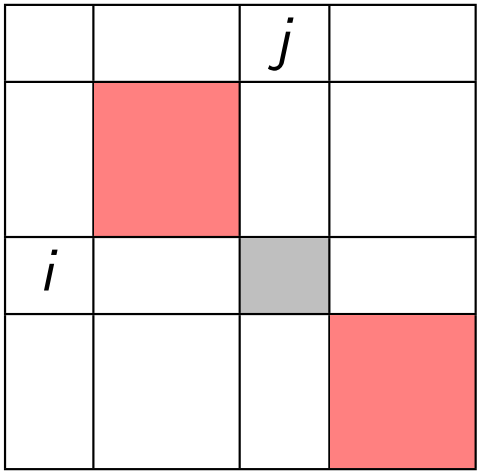
Свойства коэффициента γ

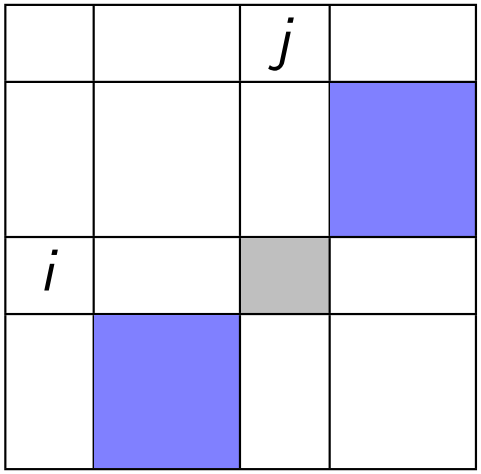
- $|\gamma| \leqslant 1$.
- γ не определен, если $\rho_{ij} = 0$ для всех i (или j) кроме одного.
- $\gamma = \begin{cases} 1, & \text{если } \Pi_d = 0, \\ -1, & \text{если } \Pi_s = 0. \end{cases}$
- $\gamma = 0$, если признаки K_1 и K_2 независимы, но обратное не верно.

0.2	0	0.2
0	0.2	0
0.2	0	0.2

Свойства коэффициента γ

$$S_{ij} = \sum_{i'>i}\sum_{j'>j} \rho_{i'j'} + \sum_{i'<i}\sum_{j'<j} \rho_{i'j'}, \quad D_{ij} = \sum_{i'<i}\sum_{j'>j} \rho_{i'j'} + \sum_{i'<i}\sum_{j'>j} \rho_{i'j'}.$$





$$\Pi_s = \sum_i \sum_j S_{ij}, \quad \Pi_d = \sum_i \sum_j D_{ij}.$$

$$\sqrt{N}(G-\gamma) \xrightarrow{\mathrm{d}} \mathcal{N}(0,\sigma^2),$$

$$\sigma^2 = \frac{16}{(1-\Pi_t)^4}\{\Pi_s^2\Pi_{dd} - 2\Pi_s\Pi_d\Pi_{sd} + \Pi_d^2\Pi_{ss}\},$$

$$\Pi_{ss} = \sum_i\sum_j \rho_{ij}S_{ij}^2, \quad \Pi_{dd} = \sum_i\sum_j \rho_{ij}D_{ij}^2, \quad \Pi_{sd} = \sum_i\sum_j \rho_{ij}S_{ij}D_{ij}.$$

$$\sigma^2 \leqslant \frac{2(1-\gamma^2)}{1-\Pi_t}.$$

В силу последнего неравенства на практике можно полагать

$$(G-\gamma)\sqrt{\frac{N^2-P_t}{2N(1-G^2)}} \approx \mathcal{N}(0,1).$$

С вероятностью $1-\alpha$

$$-K_{\alpha/2} \leqslant (G-\gamma)\sqrt{\frac{N^2-P_t}{2N(1-G^2)}} \leqslant K_{\alpha/2},$$

$$(G-\gamma)^2\left(\frac{N^2-P_t}{2N(1-\gamma^2)}\right) \leqslant K_{\alpha/2}^2.$$

Решив квадратное неравенство, можно получить доверительный интервал для γ.

Пример

R Table			
8	5	3	3
0	8	1	0
0	4	14	4

S Table			
31	19	4	0
22	26	17	16
0	8	21	25

D Table			
0	0	12	27
11	6	7	18
20	7	3	0

Обозначим за $A \times B$ поэлементную свертку матриц A и B.

$$P_s = R \times S = 1006,\ P_d = R \times D = 242,$$

$$P_{ss} = R \times S \times S = 24168,\ P_{sd} = R \times S \times D = 2617,$$

$$P_{dd} = R \times D \times D = 3278.$$

Пример

Вначале для оценки дисперсии воспользуемся точной формулой:

$$G = 0.612, \quad \hat{\sigma} = 0.151.$$

Если взять $\alpha = 0.05$, то получим доверительный интервал $(0.316, 0.908)$.
Если использовать верхнюю оценку на дисперсию, то получим $\hat{\sigma} = 0.224$ и доверительный интервал $(0.174, 1)$.
Если использовать вышеописанное квадратное неравенство, то получим интервал $(0.058, 0.878)$.