

Testy nezávislosti v kontingenčních tabulkách

Slovní formulace problému

Při sensorické analýze byly porovnávány tři výroby salámů z různých měst - Prahy, Šumperka a Klatov. Všichni hodnotící mohli každému vzorku přiřadit známku A, B nebo C . Přičemž známka A značí vynikající chuť, B průměrnou chuť a C chuť špatnou. Data z této sensorické analýzy jsou uvedena v následující tabulce.

Město	A	B	C
Praha	7	7	6
Šumperk	4	4	10
Klatovy	6	11	3

My bychom rádi zjistili, zda je chuť salámu závislá na výrobě, ze které pochází.

Označme si X veličinu, která značí, ze kterého města pochází salám a Y veličinu, která odpovídá známce. Závislost veličin X a Y nelze zkoumat pomocí korelačního koeficientu, neboť se jedná o tzv. **nominální veličiny**. Testování závislosti nominálních veličin provedeme pomocí tzv. **kontingenční tabulky**. Jedná se o tabulku četností, jejíž řádky udávají četnosti veličiny X a sloupce četnosti veličiny Y .

Uvažujme obecně kontingenční tabulku (n_{ij}) o r řádcích a c sloupcích. Označme $n = \sum_{i=1}^r \sum_{j=1}^c n_{ij}$ a dále

$$n_{i\cdot} = \sum_{j=1}^c n_{ij},$$
$$n_{\cdot j} = \sum_{i=1}^r n_{ij}.$$

Definujme

$$\hat{N}_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}. \quad (1)$$

Potom testová statistika pro nulovou hypotézu

$$H_0 : \text{veličiny } X \text{ a } Y \text{ jsou nezávislé,}$$

oproti alternativě, že tomu tak není, má tvar

$$R = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - \hat{N}_{ij})^2}{\hat{N}_{ij}}, \quad 1$$

a za platnosti H_0 má asymptoticky χ^2 -rozdělení o $(r-1)(c-1)$ stupních volnosti.

Poznámka: Test nezávislosti v kontingenčních tabulkách je úzce spojený s χ^2 testem dobré shody. Kontingenční tabulky nejsou ve skutečnosti nic jiného než test shody rozdělení. Pro bližší informace nahlédněte poznámky k předmětu Aplikovaná statistika, 11. téma.

¹Viz poznámku pod čarou 1,

$$R = n \left(\sum_{i=1}^r \sum_{j=1}^c \frac{n_{ij}^2}{n_{i\cdot} \cdot n_{\cdot j}} - 1 \right).$$