

# 10 Testování nezávislosti v kontingenčních tabulkách

## 10.1 Kontingenční tabulky

- jeden výběr ... dva nominální znaky  $X$  a  $Y$
- znak  $X$  ...  $r$  variant; znak  $Y$  ...  $s$  variant

### Kontingenční tabulka (KT)

- $n_{jk}$  ... absolutní simultánní četnosti  $j$ -té varianty znaku  $X$  a  $k$ -té varianty znaku  $Y$
- $n_{j.} = n_{j1} + \dots + n_{js}$  ... absolutní marginální četnosti  $j$ -té varianty znaku  $X$
- $n_{.k} = n_{1k} + \dots + n_{rk}$  ... absolutní marginální četnosti  $k$ -té varianty znaku  $Y$

### Pearsonův $\chi^2$ test

- **asymptotický test**
  - musíme ověřit podmínku dobré aproximace
  - `chisq.test(data, correct=F)$expected`
  - alespoň 80% případů musí být  $\geq 5$  a zbylých 20% nesmí klesnout pod 2.
- $H_0$  :  $X, Y$  jsou stochasticky nezávislé.
- $H_1$  :  $X, Y$  nejsou stochasticky nezávislé.
- porovnáváme pozorované četnosti  $n_{jk}$  a teoretické četnosti  $\frac{n_{j.}n_{.k}}{n}$  dvojice variant  $(x_{[j]}, y_{[k]})$
- za platnosti  $H_0$  si jsou  $n_{jk}$  a  $\frac{n_{j.}n_{.k}}{n}$  podobné
- Testovací statistika:

$$K = \sum_{j=1}^r \sum_{k=1}^s \frac{(n_{jk} - \frac{n_{j.}n_{.k}}{n})^2}{\frac{n_{j.}n_{.k}}{n}}$$

- Kritický obor:  $W = \langle \chi_{1-\alpha}^2((r-1)(s-1)), \infty \rangle$
- `chisq.test(data, correct=F)`

## Měření závislosti, Cramérův koeficient

- Cramérův koeficient

$$V = \sqrt{\frac{K}{n(m-1)'}}$$

kde  $m = \min\{r, s\}$ .

Cramérův koeficient	interpretace
0 – 0.1	zanedbatelná závislost
0.1 – 0.3	slabá závislost
0.3 – 0.7	střední závislost
0.7 – 1	silná závislost

- `cramersV(data)` z knihovny `lsr`

## 10.2 Čtyřpolní kontingenční tabulky

- náhodné veličiny  $X, Y$  mají pouze 2 varianty  $\rightarrow$  *čtyřpolní kontingenční tabulka*
- značení:  $n_{11} = a, n_{12} = b, n_{21} = c, n_{22} = d$

### 10.2.1 Pearsonův $\chi^2$ test

- asymptotický test; viz výše
- kritický obor:  $W = \langle \chi_{1-\alpha}^2(1), \infty \rangle$

### 10.2.2 Fisherův faktoriálový test

- přesný test
- `fisher.test(data)`

## Podíl šancí ve čtyřpolní KT

- pokus se provádí za dvojích různých okolností a může skončit buď úspěchem nebo neúspěchem

- 1.okolnost: podíl počtu úspěchů ku počtu neúspěchů:  $\frac{a}{c}$
- 2.okolnost: podíl počtu úspěchů ku počtu neúspěchů:  $\frac{b}{d}$
- $o\rho$  ... teoretický podíl šancí
  - $X, Y$  nezávislé  $\rightarrow$  potom  $o\rho = 1$

- $OR$  ... výběrový podíl šancí

$$OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}$$

- Závislost  $X, Y$  je tím silnější, čím více se  $OR$  ( $o\rho$ ) liší od 1.
- $OR$  resp.  $o\rho \in \langle 0; \infty \rangle$  (nesymetrický interval)  $\rightarrow$  preferujeme logaritmus podílu šancí
- $\ln(OR)$  resp.  $\ln(o\rho) \in \langle -\infty; \infty \rangle$

## Test podílem šancí

- $H_0 : X, Y$  jsou stochasticky nezávislé ...  $o\rho = 1 \rightarrow \ln o\rho = 0$
- $H_1 : X, Y$  nejsou stochasticky nezávislé ...  $o\rho \neq 1 \rightarrow \ln o\rho \neq 0$ .
- Testová statistika

$$T_0 = \frac{\ln OR}{\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$$

- Kritický obor:  $W = (-\infty; -u_{1-\alpha/2}) \cup \langle u_{1-\alpha/2}; \infty \rangle$
- $100(1 - \alpha)\%$  **asymptotický** interval spolehlivosti

$$(d, h) = \left( \ln OR - \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} u_{1-\alpha/2}; \ln OR + \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} u_{\alpha/2} \right).$$