Kontingenční tabulka

T-test i pořadové testy počítají s číselnými proměnnými, ale v praxi se často vyskytují i nominální (faktorové) proměnné.

Proměnná Z má I úrovní.

Proměnná Y má J úrovní.

Dohromady máme IJ kombinací úrovní faktorů Z a Y.

Kontingenční tabulka: sečteme výskyty jednotlivých kombinací (Z, Y) ve výběru a výsledek zaneseme do tabulky s I řádky a J sloupci.

V každém *políčku* je uveden počet měření s odpovídající kombinací hodnot Z a Y.

Příklad

$$\mathcal{X} = \left(\begin{array}{ccc|c} 4 & 0 & 2 & 6 \\ 0 & 1 & 1 & 2 \\ 1 & 1 & 4 & 6 \\ \hline 5 & 2 & 7 & 14 \end{array}\right)$$

$\leftarrow$ Daňové úniky

$\leftarrow$ Hospodářská kriminalita

$\leftarrow$ Násilné trestné činy

$\uparrow$ Praha

$\uparrow$ Brno

$\uparrow$ Ostrava

Sdružené rozdělení: $\pi_{ij} = P(Z = i, Y = j)$ je pravděpodobnost, že Z se rovná i a zároveň Y se rovná j.

Marginální rozdělení Z: $\pi_{i\cdot}$ je pravděpodobnost, že Z se rovná i.

Marginální rozdělení Y: $\pi_{\cdot j}$ je pravděpodobnost, že Y se rovná j.

Nezávislost

Vztah mezi diskrétními náhodnými veličinami Z a Y může být popsán jejich sdruženým rozdělením, podmíněným rozdělením Z za podmínky Y nebo podmíněným rozdělením Y za podmínky Z.

Pokud bychom znali sdružené pravděpodobnosti π_{ij}, pak bychom snadno mohli rozhodnout, jestli jsou Z a Y nezávislé, tj. jestli platí:
$\pi_{i|j} = \pi_{ij}/\pi_{\cdot j} = \pi_{i\cdot}$,
$\pi_{j|i} = \pi_{ji}/\pi_{i\cdot} = \pi_{\cdot j}$
nebo $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$.

V praxi se ale rozhodujeme na základě nějakého náhodného výběru.

Pozorované relativní četnosti budeme značit $p_{ij} = x_{ij}/x_{\bullet\bullet}$, kde x_{ij} jsou absolutní četnosti a $x_{\bullet\bullet}$ je rozsah výběru.

Pro odvození testu i pro správnou interpretaci odhadnutých pravděpodobností může být důležité, jakým způsobem kontingenční tabulka vznikala.

- Poisson sampling
  (vše je náhodné)
- Multinomial sampling
  (celkový počet pozorování, tj. součet x_{ij} je pevně daný)
- Independent multinomial sampling
  (celkový počet pozorování v každém řádku nebo sloupci je pevně daný)

Odvození a interpretace odhadu sdružených pravděpodobností π_{ij} sice závisí na tom, jakým způsobem kontingenční tabulka vznikla, ale odhad naštěstí vyjde pokaždé stejně $\widehat{\pi}_{ij} = p_{ij} = x_{ij}/x_{\bullet\bullet}$.

Za předpokladu nezávislosti stačí odhadnout marginální pravděpodobnosti a z nich už snadno získáme odhady:

$$\tilde{\pi}_{ij} = \hat{\pi}_{i.}\hat{\pi}_{.j} = (x_{i\bullet}x_{\bullet j})/x_{\bullet\bullet}^2.$$

Testování nezávislosti je tedy zřejmě možné založit na vhodné míře vzdálenosti mezi $x_{ij}/x_{\bullet\bullet}$ a $(x_{i\bullet}x_{\bullet j})/x_{\bullet\bullet}^2$ nebo na nějaké míře vzdálenosti mezi pozorovanými četnostmi x_{ij} a četnostmi očekávanými za předpokladu nezávislosti

$$E_{ij} = \tilde{\pi}_{ij}x_{\bullet\bullet} = (x_{i\bullet}x_{\bullet j})/x_{\bullet\bullet}.$$

Test poměrem věrohodností (G-test)

Standardním matematicko-statistickým postupem lze snadno odvodit tzv. test nezávislosti poměrem věrohodností v kontingenční tabulce (Likelihood-Ratio Test of Independence).

Testová statistika je:

$$G^2 = -2 \log \frac{\prod_i \prod_j (x_{i\bullet} x_{\bullet j})^{x_{ij}}}{x_{\bullet\bullet}^{x_{\bullet\bullet}} \prod_i \prod_j x_{ij}^{x_{ij}}} = 2 \sum \sum x_{ij} \log(x_{ij}/E_{ij}),$$

kde E_{ij} je odhad očekávaných četností za předpokladu nezávislosti. Za platnosti nulové hypotézy (nezávislost) má testová statistika G^2 rozdělení $\chi^2_{(I-1)(J-1)}$ (nezávislost tedy zamítáme na hladině α, pokud nám vyjde testová statistika větší než kvantil $\chi^2_{(I-1)(J-1)}(1-\alpha)$).

Pearsonův χ^2-test nezávislosti

Nejčastěji používaný test nezávislosti v kontingenční tabulce je založen na testové statistice:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(x_{ij} - E_{ij})^2}{E_{ij}}.$$

Za platnosti nulové hypotézy (nezávislost) má testová statistika X^2 rozdělení $\chi^2_{(I-1)(J-1)}$.

Jako předpoklad pro použití tohoto testu se obvykle požaduje, aby všechny očekávané četnosti byly alespoň 5. Pokud jsou četnosti nižší, lze použít např. Fisherův přesný test.

```
EYE/HAIR      black    brown      red      blond
d.brown          68      119       26          7
l.brown          15       54       14         10
green             5       29       14         16
blue             20       84       17         94

> chisq.test(eyehair)

Pearson's Chi-squared test

data:  eyehair
X-squared = 138.2898, df = 9, p-value < 2.2e-16
```

Zde tedy zamítáme nezávislost a jako další krok bychom se mohli podívat, které pozorované četnosti jsou příliš malé nebo příliš velké.

Drcení zubů v lisu

Příklad

Údaje o zubech (Anděl, Statistické metody: příklad 11.9):

	zlomen	nezlomen	celkem
tepelný šok	21	29	50
bez šoku	11	39	50
celkem	32	68	100

Má tepelný šok vliv na pevnost zubu?

```
> chisq.test(matrix(c(21,11,29,39),2))
```

Rozumové schopnosti zločinců

Příklad

Údaje o zločincích (Anděl, Statistické metody: příklad 11.11):

	hmotnost		
	do 150 liber	nad 150 liber	celkem
normální schopnosti	272	124	396
snížené schopnosti	82	15	97
celkem	354	139	493

Souvisí hmotnost zločinců s jejich rozumovými schopnostmi?

```
> chisq.test(matrix(c(272,82,124,15),2))
```