

12. Analýza kategoriálních dat

χ^2 test dobré shody

- Srovnání s teoretickou distribucí
- Nominální, ordinální, diskrétní data
- Spojitá data kategorizujeme
- Předpoklady: kategorie vzájemně nezávislé
očekávané frekvence > 5

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n\pi_i)^2}{n\pi_i}$$

- n_i – pozorované četnosti pro jednotlivé kategorie
- $n\pi_i$ – očekávané četnosti při daných pravděpodobnostech π_i

χ^2 test dobré shody

- χ^2 má pro výběry velkých rozsahů přibližně χ^2 rozdělení s $k-1$ stupni volnosti
- X má předpokládané rozložení dané psmi $\pi_i \Rightarrow$ hodnota testové statistiky χ^2 bude malá, pro velké hodnoty χ^2 zamítáme H_0
- Kritická hodnota = kvantil $\chi^2_{1-\alpha}(df)$ o $df = k-1$ stupních volnosti
- Pokud ověřujeme pouze typ, ale ne hodnoty parametrů rozložení, musí být parametry z výběru předem odhadnuty \Rightarrow snížení počtu df

χ^2 test dobré shody - příklad

10 000 lidí hází mincí -> rub: 4 000 případů (R)

-> líc: 6 000 případů (L)

Lze výsledek považovat za statisticky významně odlišný od očekávaného poměru R:L = 1:1?

$$\chi^2 = \frac{(4000 - 5000)^2}{5000} + \frac{(6000 - 5000)^2}{5000} = 400$$

- Kritická hodnota: $\chi^2_{(0,95)}(df = 1) = 3,84$
- $3,84 \ll 400 \Rightarrow$ zamítáme H_0 o shodnosti očekávaného a pozorovaného poměru

χ^2 test dobré shody - příklad

Studujeme rozdělení počtu pacientů, kteří přijdou na zubní pohotovost ve všední den. Ordinační dobu rozdělíme do půlhodinových intervalů a v každé půlhodině zjistíme počet pacientů, kteří se během ní na zubní pohotovost dostavili. Ověřte na 5% hladině významnosti, zda je přijatelný předpoklad o Poissonově rozdělení počtu pacientů.

H_0 : Počet příchodů pacientů během 30 min. má Poissonovo rozlož.
 H_1 : Počet příchodů pacientů během 30 min. nemá Poissonovo rozlož.

Za platnosti H_0 pst příchodu určitého počtu pacientů x :

$$P(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

λ neznáme => odhadneme jako vážený průměr:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i = \frac{1}{1200} (79 \cdot 0 + 188 \cdot 1 + \dots + 0 \cdot 11) = \frac{3364}{1200} = 2,80$$

Pro $\lambda = 2,80$ počítáme psti : $P(x_1=0)=\pi_1, P(x_2=1)=\pi_2, \dots$

$P(x_{12}=11 \text{ a více})=\pi_{12}$

Očekávané četnosti: $n \pi_i$

Okrajové třídy spojíme ($n \pi_i < 5$)

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n\pi_i)^2}{n\pi_i} = 8,50$$

$$df = k - m - 1 = 9 - 1 - 1 = 7 \Rightarrow \chi^2_{(0,95)}(7) = 14,07$$

$8,50 < 14,07 \Rightarrow$ nezamítáme H_0 o Poissonově rozložení

| Číslo kategorie | Počet pacientů | Pozorovaná četnost | Očekávaná četnost |
|--------------------|-------------------|-----------------------|----------------------|
| i | x_i | n_i | $n\pi_i$ |
| 1 | 0 | 79 | 72,97 |
| 2 | 1 | 188 | 204,32 |
| 3 | 2 | 282 | 286,05 |
| 4 | 3 | 275 | 266,98 |
| 5 | 4 | 196 | 186,89 |
| 6 | 5 | 114 | 104,66 |
| 7 | 6 | 45 | 48,84 |
| 8 | 7 | 10 | 19,54 |
| 9 | 8 | 7 | 6,84 |
| 10 | 9 | 3 | 2,13 |
| 11 | 10 | 1 | 0,78 |
| 12 | 11 a více | 0 | 0,00 |
| Celkem | - | 1200 | 1200,00 |

| | | | |
|---|----------|----|------|
| 9 | 8 a více | 11 | 9,75 |
|---|----------|----|------|

Kontingenční tabulky

- Data kvalitativní, diskrétní kvantitativní, spojité kvantitativní, ale s hodnotami sloučenými do skupin
- Dva znaky tohoto typu – kategorie jednoho znaku = řádky, kategorie druhého znaku = sloupce => kontingenční tabulka
- Jeden znak r kategorií, druhý znak s kategorií => kontingenční tabulka typu $r \times s$
- Kontingenční tabulka typu 2×2 = čtyřpolní tabulka

Testy v kontingenčních tabulkách

- Hypotéza o shodnosti struktury (1 znaku ve dvou a více výběrech)
- Hypotéza o nezávislosti (2 znaků v jednom výběru)
- Hypotéza o symetrii (2 znaků či opakovaných měřeních v jednom výběru)

Příklady testů

Příklad č.1: Byl studován výskyt mihulí v tocích České republiky. Předběžné výsledky ukázaly, že jejich přítomnost/nepřítomnost v toku není určena současným stupněm znečištění ani znečištěním v minulosti (nelze ale vyloučit jednorázovou intoxikaci). Byly tedy studovány další vlastnosti jednotlivých toků, zvl. mechanické zábrany, které mohou limitovat pohyb kruhoústých a ryb v toku. Toky byly klasifikovány do 2 typů: a) s přítomností jezů a splavů zabraňujících zpětnému návratu vodních obratlovců a b) bez přítomnosti jezů a splavů. Bylo celkem vyšetřeno 100 toků. Z nich bylo 50 s jezy a 50 bez jezů. Z toků typu a) byly mihule nalezeny v 10 případech, v tocích typů b) ve 40 případech. Je poměr toků s výskytem/absencí mihulí shodný v obou typech toků (tj. v tocích s bariérami/bez bariér)?

Příklad č. 2: Zkoumáme vzájemný výskyt dvou druhů na skalní stepi. Celkem jsme na plochu rozmístili náhodně 100 plošek o rozměru 1x1 m. Na každé ploše jsme zaznamenali přítomnost/nepřítomnost druhu A a druhu B. Oba druhy se vyskytovaly v 36 čtvercích, ani jeden ve 20 čtvercích, pouze druh A se vyskytoval ve 30 čtvercích. Vyskytují se druhy vzájemně nezávisle?

Příklad č. 3: Sledujeme skupinu 20 pacientů, kteří byli léčeni dvěma různými hypertenzivy A a B. Každý pacient dostával po dobu 1 měsíce lék A a po odeznění případných účinků po dobu 1 měsíce lék B. Výsledek byl klasifikován jako úspěch (tlak snížen o více než 15 mm Hg) či neúspěch. Liší se léky v účinku?

Test hypotézy o shodnosti struktury

- Shodnost struktury jednoho ze sledovaných znaků za různých podmínek, které vyjadřují kategorie druhého znaku
- Očekávaná četnost = (součet v řádku x součet ve sloupci)/celkový počet pozorování

$$X^2 = \sum \frac{(\text{pozorovaná četnost} - \text{očekávaná četnost})^2}{\text{očekávaná četnost}}$$

- Sčítáme přes všechna políčka v tabulce
- Kritická hodnota: kvantil $X^2_{1-\alpha}(df)$
- $df = (\text{počet řádků} - 1)(\text{počet sloupců} - 1)$

Čtyřpolní tabulka

| | | |
|-----|-----|-----|
| a | b | a+b |
| c | d | c+d |
| a+c | b+d | n |

$$F(A) = \frac{(a+b)(a+c)}{n}$$

$$F(B) = \frac{(a+b)(b+d)}{n}$$

$$F(C) = \frac{(a+c)(c+d)}{n}$$

$$F(D) = \frac{(b+d)(c+d)}{n}$$

$$X^2 = n \frac{(ad - bc)^2}{(a+b)(a+c)(c+d)(b+d)}$$

$$df = 1$$

Test o shodnosti struktury - příklad

| kouření/vzdělání | základní | odborné | střední | VŠ | Suma |
|------------------|----------|---------|---------|-----|------|
| Nekuřák | 14 | 22 | 55 | 73 | 197 |
| Bývalý kuřák | 11 | 28 | 44 | 42 | 125 |
| Kuřák | 14 | 24 | 24 | 17 | 79 |
| Silný kuřák | 78 | 189 | 175 | 106 | 548 |
| Suma | 117 | 296 | 298 | 238 | 949 |

$$\begin{aligned} X^2 &= (14 - 197 \cdot 117 / 949)^2 / (197 \cdot 117 / 949) \\ &+ (22 - 296 \cdot 197 / 949)^2 / (296 \cdot 197 / 949) + \dots \\ &= 4,36 + 0,68 + \dots = 38,68 \end{aligned}$$

$$df = (r-1)(s-1) = 9$$

$$X^2_{1-0,001,9} = 27,88$$

$38,68 > 27,88 \Rightarrow$ zamítáme H_0
na hladině významnosti 0,001

Test hypotézy o nezávislosti - příklad

Při studiu vztahu mezi barvou vlasů a očí v populaci Němců antropolog pozoroval náhodný výběr 6800 lidí s těmito výsledky:

| | | Barva vlasů | | Celkem |
|-----------|--------|-------------|--------|--------|
| | | Tmavá | Světlá | |
| Barva očí | Tmavá | 726 | 131 | 857 |
| | Světlá | 3129 | 2814 | 5943 |
| Celkem | | 3855 | 2945 | 6800 |

H_0 : Barva očí je nezávislá na barvě vlasů
 H_0 : Barva vlasů je nezávislá na barvě očí
 H_0 : Barva očí a barva vlasů jsou vzájemně nezávislé

$$X^2 = 6800 \cdot \frac{(726 \cdot 2814 - 131 \cdot 3129)^2}{(726 + 131)(726 + 3129)(3129 + 2814)(131 + 2814)} =$$
$$6800 \cdot \frac{2,67 \cdot 10^{12}}{857 \cdot 3855 \cdot 5943 \cdot 2945} = 341,5$$

$$X^2_{(1-0,05,1)} = 3,84$$

$341,5 > 3,84 \Rightarrow$ zamítáme H_0 o nezávislosti barvy očí a barvy vlasů

Fisherův exaktní test

- Analyzuje všechny možné 2x2 tabulky, které dávají stejnou sumu řádků a sloupců jako zdrojová tabulka
- Každé tabulce se přiřazuje p st, že taková situace nastane, je-li H_0 pravdivá

Fisherův exaktní test – ilustrační příklad

| | | Delikventi | Nedelikventi | Celkem |
|--------------|--------|------------|--------------|--------|
| Nošení brýlí | Ano | 1 | 5 | 6 |
| | Ne | 8 | 2 | 10 |
| | Celkem | 9 | 7 | 16 |

Všechny možné varianty tabulky s danou sumou řádků a sloupců

| | | | | | | | | | | | |
|-------|--|---|---|---|---|-------|--|---|---|---|---|
| (I) | <table border="1"><tr><td>0</td><td>6</td></tr><tr><td>9</td><td>1</td></tr></table> | 0 | 6 | 9 | 1 | (V) | <table border="1"><tr><td>4</td><td>2</td></tr><tr><td>5</td><td>5</td></tr></table> | 4 | 2 | 5 | 5 |
| 0 | 6 | | | | | | | | | | |
| 9 | 1 | | | | | | | | | | |
| 4 | 2 | | | | | | | | | | |
| 5 | 5 | | | | | | | | | | |
| (II) | <table border="1"><tr><td>1</td><td>5</td></tr><tr><td>8</td><td>2</td></tr></table> | 1 | 5 | 8 | 2 | (VI) | <table border="1"><tr><td>5</td><td>1</td></tr><tr><td>4</td><td>6</td></tr></table> | 5 | 1 | 4 | 6 |
| 1 | 5 | | | | | | | | | | |
| 8 | 2 | | | | | | | | | | |
| 5 | 1 | | | | | | | | | | |
| 4 | 6 | | | | | | | | | | |
| (III) | <table border="1"><tr><td>2</td><td>4</td></tr><tr><td>7</td><td>3</td></tr></table> | 2 | 4 | 7 | 3 | (VII) | <table border="1"><tr><td>6</td><td>0</td></tr><tr><td>3</td><td>7</td></tr></table> | 6 | 0 | 3 | 7 |
| 2 | 4 | | | | | | | | | | |
| 7 | 3 | | | | | | | | | | |
| 6 | 0 | | | | | | | | | | |
| 3 | 7 | | | | | | | | | | |
| (IV) | <table border="1"><tr><td>3</td><td>3</td></tr><tr><td>6</td><td>4</td></tr></table> | 3 | 3 | 6 | 4 | | | | | | |
| 3 | 3 | | | | | | | | | | |
| 6 | 4 | | | | | | | | | | |

Pravděpodobnost náhodného vzniku variant tabulky

| | a | b | c | d | P |
|--------------|---|---|---|---|----------------|
| (I) | 0 | 6 | 9 | 1 | 0,00087 |
| (II) | 1 | 5 | 8 | 2 | 0,02360 |
| (III) | 2 | 4 | 7 | 3 | 0,15734 |
| (IV) | 3 | 3 | 6 | 4 | 0,36713 |
| (V) | 4 | 2 | 5 | 5 | 0,33042 |
| (VI) | 5 | 1 | 4 | 6 | 0,11014 |
| (VII) | 6 | 0 | 3 | 7 | 0,01049 |
| Total | | | | | 0,99999 |

Test hypotézy o symetrii

- Pro 2x2 tabulku => McNemarův test

| | Léčba II | | |
|---------|----------|-----|--------|
| Léčba I | + | - | Celkem |
| + | a | b | a+b |
| - | c | d | c+d |
| Celkem | a+c | b+d | n |

$$X^2 = \frac{(b - c)^2}{b + c}$$

- Kritická hodnota: kvantil $X^2_{1-\alpha}(df)$, kde $df=1$
- Lze testovat i časový vývoj

McNemar test - příklad

Srovnání dvou metod stanovení antigenu v krvi (antigen vždy přítomen)

H_0 : metoda I = metoda II

| Metoda I | Metoda II | Frekvence |
|----------|-----------|-----------|
| úspěch | úspěch | 202 |
| úspěch | neúspěch | 60 |
| neúspěch | úspěch | 42 |
| neúspěch | neúspěch | 10 |

$$\Sigma = 102$$

$$X^2 = \frac{(b - c)^2}{b + c} = \frac{(60 - 42)^2}{102} = \frac{324}{102} = 3,18$$

$$X^2_{(1-0,05,1)} = 3,84$$

Nelze zamítnout H_0 o ekvivalentnosti metod