
TESTY PRO KATEGORICKÉ PROMĚNNÉ – NEPARAMETRICKÉ METODY

Analýza četností hodnot kategorických (=O, N) proměnných

Výzkumné otázky...

- Liší se preference politických stran?
- Liší se poměrné zastoupení kuřáků mezi ženami a muži?
- Souvisí nějak individuální volební preference s odhadem měsíčního příjmu respondenta?
- Otázky směřují
 - buď k rozdílu četností různých jevů v rámci jedné proměnné (četnost různých jevů v populaci),
 - k rozdílu četností jevu mezi různými proměnnými (četnost jevu v různých populacích),
 - Nebo k pravděpodobnosti výskytu dvou (či více) jevů současně.

χ^2 test dobré shody

- Liší se empirické četnosti nějakých jevů od teoreticky očekávaných četností?
 - Preference politických stran ve volbách...
 - Tedy jedna nominální proměnná, jeden výběr

- Testujeme p rozdílů mezi empirickými-pozorovanými (f_o) a očekávanými (f_e) četnostmi

- Mírou rozdílu je hodnota χ^2 , která má rozložení χ^2 s $\nu=k-1$ stupni volnosti a průměrem ν

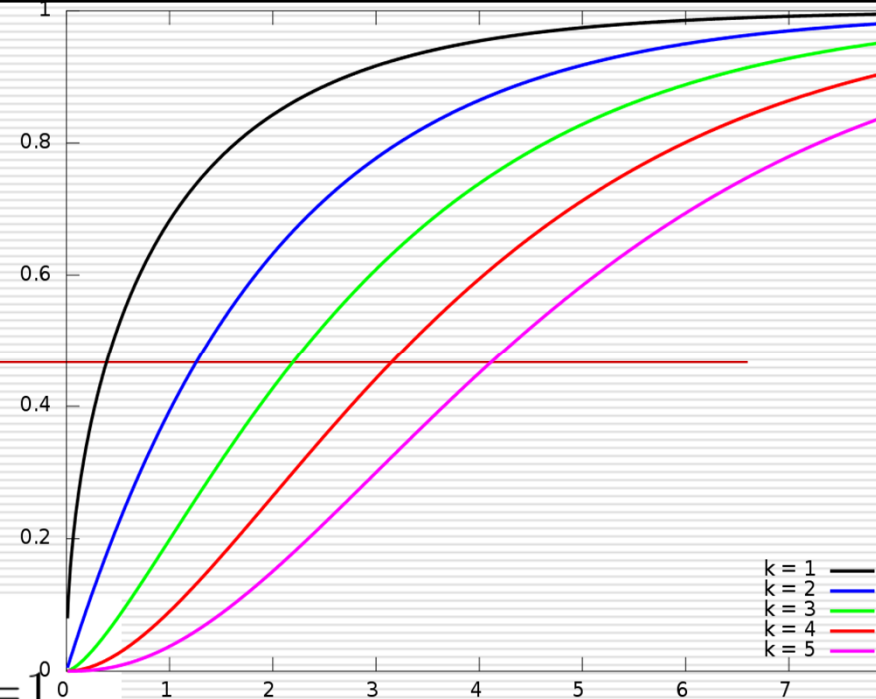
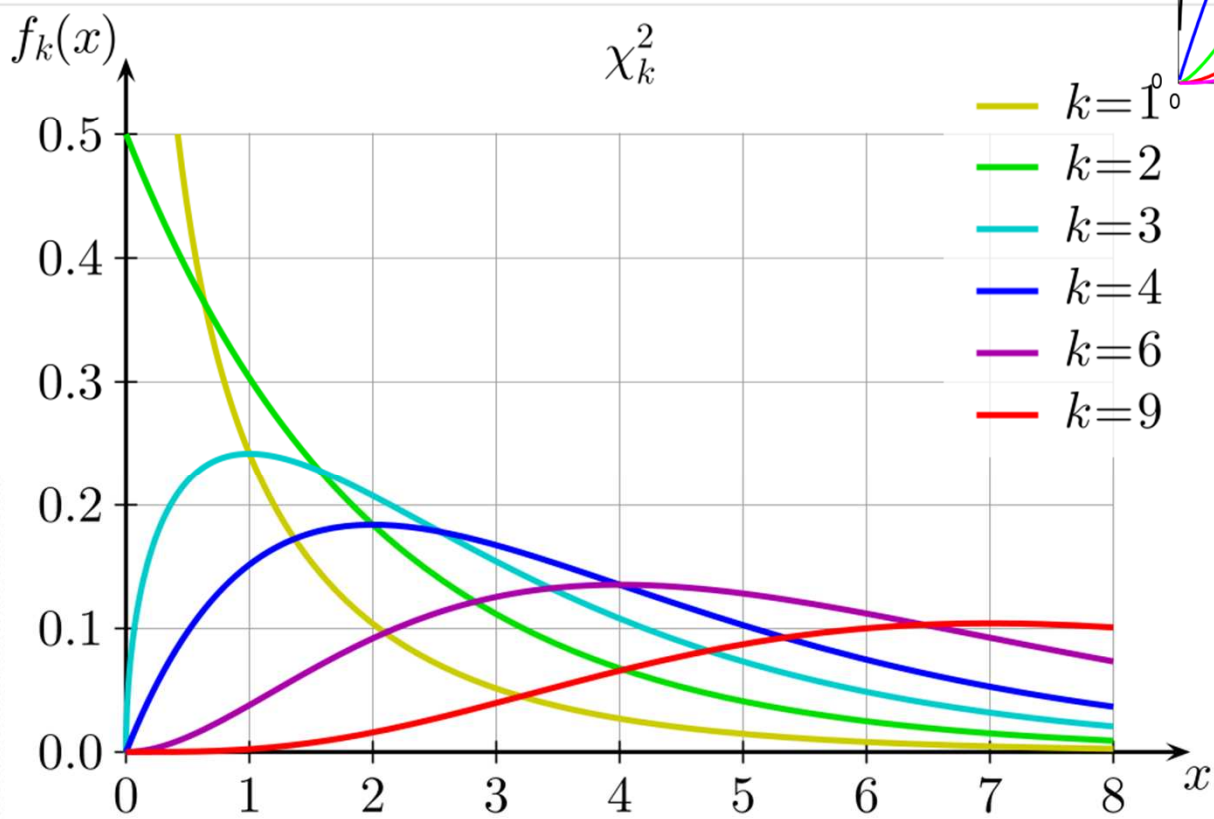
$$\chi^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^k \frac{(fo_i - fe_i)^2}{fe_i}$$

- $H_0: \chi^2 = \nu$ vs. $H_1: \chi^2 > \nu$

k je počet kategorií, n velikost vzorku, n_i četnost kat. i ,
 p_i teoretická p -nost jevu v kategorii i ;

- Pro získání pravděpodobnosti χ^2 CHISQ.DIST(χ^2 ; df; 1); CHISQ.INV(p ; df)
- Očekávané četnosti stanovujeme na základě teoretického předpokladu.
- n_i a np_i vždy jako **četnosti**; nikdy ne procenta (ztráta informace o velikosti vzorku)

Rozdělení χ^2



Ve kterém městě byste žili nejraději?

Kategorie	n	p	np	(n-np)^2/np
Paříž	28	0,2	28	0
New York	28	0,2	28	0
Londýn	28	0,2	28	0
L.A.	28	0,2	28	0
Tokio	28	0,2	28	0
Celkem	140	1	140	0
Chi ²				0

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

$$P(\chi^2 > 0 \mid \chi^2 = 4) \approx 1$$

Ve kterém městě byste žili nejraději?

Kategorie	n	p	np	(n-np)^2/np
Paříž	38	0,2	28	3,57
New York	37	0,2	28	2,89
Londýn	22	0,2	28	1,29
L.A.	25	0,2	28	0,32
Tokio	18	0,2	28	3,57
Celkem	140	1	140	11,64
Chi2				11,64

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

Závislost kategorických proměnných

- ❑ Jaká je souvislost preference politické strany a úrovně hrubého příjmu voliče?
- ❑ Jaká je pravděpodobnost společného výskytu dvou jevů x a y možných?
- ❑ Kontingenční tabulka ... řádky \times sloupce = $r \times s$; $i \times j$
- ❑ Ve těle tabulky jsou četnosti jednotlivých kombinací, v okrajích tzv. **marginální četnosti** – sumy sloupců nebo řádků. Tedy n_{12} znamená počet osob ve druhém sloupci prvního řádku; počet osob, u nichž nastal jev A_1 a současně B_2 .

Kategorie	B_1	B_2	...	B_s	Řádkové součty
A_1	n_{11}	n_{12}	...	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2s}	$n_{2.}$
...
A_r	n_{i1}	n_{i2}	...	n_{ij}	$n_{i.}$
Sloupcové součty	$n_{.1}$	$n_{.2}$...	$n_{.j}$	n

Závislost kategorických proměnných

- ❑ χ^2 test nezávislosti(homogenity)
- ❑ Očekávané četnosti $f_e: m_{ij}$ (očekávaná četnost v i - j -té buňce)(i – řádky, j –sloupce)
- ❑ Testová statistika je χ^2
- ❑ Stupně volnosti: $df = (i-1)*(j-1)$

$$f_e = m_{ij} = \frac{n_{i.} \cdot n_{.j}}{n} \quad \chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - m_{ij})^2}{m_{ij}} = \sum_{i=1}^r \sum_{j=1}^s \frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$$

Kategorie	B ₁	B ₂	...	B _s	Řádkové součty
A ₁	n ₁₁	n ₁₂	...	n _{1s}	n _{1.}
A ₂	n ₂₁	n ₂₂	...	n _{2s}	n _{2.}
...
A _r	n _{i1}	n _{i2}	...	n _{ij}	n _{i.}
Sloupcové součty	n_{.1}	n_{.2}	...	n_{.j}	n

Př. χ^2 test nezávislosti(homogeneity) Vztah bydliště a počtu holínek

Pozorované Řádková %	0	1	>2	Řádkové součty
Velkoměsto	10 67%	1 7%	4 27%	15
Maloměsto	15 43%	19 54%	1 3%	35
Vesnice	15 30%	20 40%	15 30%	50
Sloupcové součty	40	40	20	100

Očekávané/ dílčí χ^2	0	1	>2	Řádkové součty
Velkoměsto	6/ 2,7	6/ 4,2	3/ 0,3	15
Maloměsto	14/ 0,1	14/ 1,8	7/ 5,1	35
Vesnice	20/ 1,3	20/ 0	10/ 2,5	50
Sloupcové součty	40	40	20	100

$$\chi^2 = 17,9 \quad df = (3-1) * (3-1) = 4 \quad P(\chi^2 > 17,9 \mid \chi^2 = 4) = 0,001$$

Síla vztahu v kontingenční tabulce

- ❑ Pro tabulky 2x2 **Phi** $\phi = \sqrt{\frac{\chi^2}{n}}$
- ❑ Pro tabulky 3x3 a více **koeficient kontingence** (Pearson) $C = \sqrt{\frac{\chi^2}{\chi^2+n}}$
- ❑ Pro tabulky $r \times s$ **Cramerovo V** $V = \sqrt{\frac{\chi^2}{n(k-1)}}$ kde k je menší z r a s
- ❑ Všechny koeficienty v intervalu $\langle 0;1 \rangle$.
- ❑ Pro tabulky větší než 2x2 je často třeba identifikovat buňku(y), kde jsou největší odchylky od očekávaných četností
 - Skrze výpočet **reziduí**, tj. rozdílů mezi pozorovanou a očekávanou četností: $n_{ij} - m_{ij} = res_i$
 - ❑ tyto „zbytkové“ hodnoty lokalizují odchylky od pravděpodobnostního rozdělení
 - ❑ Součet residuí v tabulce je vždy nula
 - **Standardizovaná rezidua** (Pearsonova): $R = (n_{ij} - m_{ij})/\sqrt{m_{ij}}$
 - ❑ rozdělení standardizovaných residuí je normální s průměrem 0 a sm. odchylkou 1; tedy $R \geq \pm 1,96$ jsou „zajímavá“ pro interpretaci, významně přispívají k signifikanci χ^2 .
- ❑ Analýza tabulky skrze χ^2 je nespolehlivá, je-li $\min(m_{ij}) < 5$. *I řídké jevy musí mít šanci 😊*
 - Pro tuto situaci máme tzv. permutační testy (v SPSS „exact“)
- ❑ Hendl str. 297 – 313.

kontingenční koeficient $C = \sqrt{(17,9/(17,9+100))}=0,4$
 Cramérovo $V = \sqrt{(17,9/(100*2))}=0,3$

Pozorované Řádková % St. rezidua	0	1	>2	Řádkové součty
Velkoměsto	10 67% 1,6	1 7% -2,0	4 27% 0,6	15
Maloměsto	15 43% 0,3	19 54% 1,3	1 3% -2,3	35
Vesnice	15 30% -1,1	20 40% 0	15 30% 1,6	50
Sloupcové součty	40	40	20	100

Shrnutí

- Pro nominální data máme testy založené na chí-kvadrátu
 - Test dobré shody
 - Test nezávislosti/homogeneity
-