PSY117
Statistická analýza dat v psychologii
Přednáška 11 2018

TESTY PRO KATEGORICKÉ PROMĚNNÉ – NEPARAMETRICKÉ METODY

... a to mělo, jak sám vidíte, nedozírné následky.

Smrť

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.

AJ: frequency, relative frequency, expected frequency, residuals, Chi-square

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* rozdílu 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

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

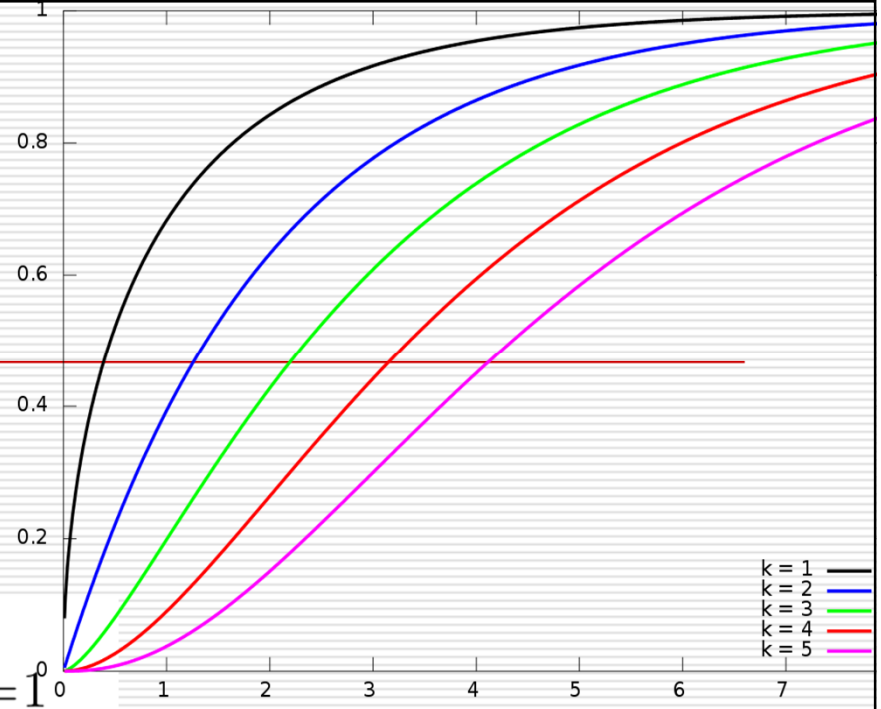
- H_0: $\chi^2 = \nu$ vs. H_1: $\chi^2 > \nu$

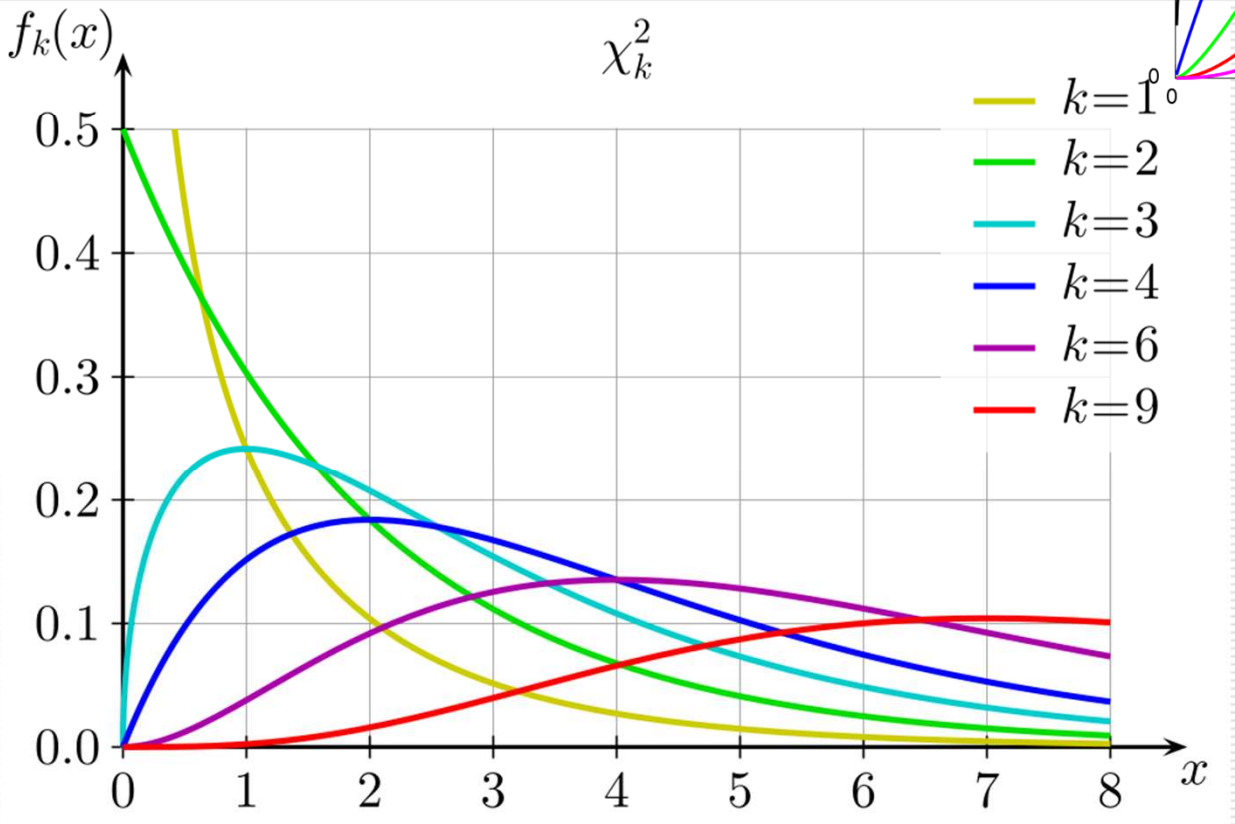
k je počet kategorií, *n* velikost vzorku, n_i četnost kat. i,
p_i teoretická p-nost jevu v kategorii i;

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

AJ: Chi-square goodness-of-fit test, observed (empirical) frequency vs. expected frequency

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
Chi²				**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**

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

$P(\chi^2 > 11{,}64 \mid \chi^2 = 4) = 1\text{-CHISQ.DIST}(11{,}64; 4; 1) = 0{,}02$

Závislost kategorických proměnných

- Jaká je souvislost preference politické strany a úrovně hrubého příjmu voliče?
- Jaká je pravděpodobnost společného výskytu dvou jevů z *x* a *y* možných?
- Kontingenční tabulka … řádky x sloupce = *r* x *s*; *i* x *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**

AJ: contingency table (crosstabulation, ctosstab), marginal frequencies

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)

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

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(homogenity) 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**

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

Velikost účinku 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* x *s* **Cramerovo** $V = \sqrt{\frac{\chi^2}{n(k-1)}}$ kde *k* ke menší z *r* a *s*
- Všechny koeficienty v intervalu <0;1>.
- 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): $\boldsymbol{R = (n_{ij} - m_{ij})/\sqrt{m_{ij}}}$
    - rozdělení standardizovaných reziduí je normální s průměrem 0 a sm. odchylkou 1; tedy R ≥ +- 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.

AJ: strength of association, contingency coefficient, standardized residuals

kontingenční koeficient C = √(17,9/(17,9+100))=0,4
Cramérovo V =√(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**

Testy středních hodnot pro ordinální proměnné – neparametrické metody

- Metody užívající *parametrů* normálního rozložení (*m*, *s*) mají svá omezení, když…
  - data pochází z rozložení, které se od normálního výrazně liší (tvar, či odlehlé hodnoty)
  - data mají spíše ordinální charakter; nebo se jedná o krátké intervalové škály
- *Neparametrické* metody
  - jsou *robustní* vůči rozložení dat…
  - mají nižší sílu testu (tj. vyšší požadavky na velikost vzorku)
- Testy pro mediány
  - Pro jeden výběr: znaménkový test, Wilcoxonův test
  - Pro párové srovnání: Wilcoxonův test
  - Pro 2 nezávislé výběry: Mann-Whitney U, Kolmogorov-Smirnov Z

Non-parametric, robust, data assumptions, sign test

Jeden výběr, znaménkový test

- *H*: Je medián roven *k*? $H_0: Md = k$; $H_1: Md \neq k$
- Platí-li H_0, mělo by nad i pod hypotetizovaným mediánem být stejné množství případů
- Asymptotický test pomocí normálního rozložení:
  - Z^+ (Z^-) je počet hodnot vyšších (nižších) než hypotetizovaný medián
  - Hodnoty rovné mediánu ignorujeme a odečítáme z *n*
  - Platí-li H_0, $Z^+ = Z^-$ *a* $Z^+ + Z^- = n$.
  - Testová statistika $z = (2Z_+ - n)/\sqrt{n}$ má asymptoticky normální rozložení. (přesně má binomické rozložení).
  - *P*=2*(1–NORM.S.DIST(*z*;1)) (nebo přímo 2*BINOM.DIST(MIN(Z^+;Z^-); n; 0,5; 1)
- Jedná se tedy o alternativu *t*-testu pro jeden výběr;
- Pro závislé výběry *(=párové srovnání)* spočítáme $d_i = x_i - y_i$ a znaménkovým testem testujeme $H_0: Md_d = 0$.

- Silnější a používanější variantou je Wilcoxonův signed-ranks test

Neparametrické testy pro nezávislé výběry

Mediánový test

- Je–li společný medián dvou výběrů shodný, leží na jedné straně *Md* 50% každého výběru.
- Určíme *Md* pro celý soubor; pokud platí H_0, četnosti hodnot ležících nad i pod *Md* by měly být stejné pro *x* i *y*.
- Pokud H_0 neplatí, budou četnosti výrazně asymetrické, v „diagonále".
- Při $n>30$ lze užít asymptoticky normálně rozloženou testovou statistiku *z*:

$$z = \frac{(ad - bc)\sqrt{n}}{\sqrt{(a+b)(b+d)(a+c)(c+d)}}$$

	Sk A	Sk B	Σ
<Md	*a*	*b*	*a+b*
>Md	*c*	*d*	*c+d*
Σ	*a+c*	*b+d*	*n*

Silnější alternativou je Wilcoxonův test pro nezávislé výběry nebo Mann-Whitney U, popřípadě další.

Shrnutí

- Pro nominální data máme testy založené na chí-kvadrátu
  - Test dobré shody
  - Test nezávislosti/homogenity
- Pro ordinální data a výrazně nenormálně rozložená intervalová máme „neparametrické“ testy
  - Jejich primitivní verze jsem si ukázali
  - „Pojmenované“ testy je zpřesňují