PSY117/454
Statistická analýza dat v psychologii
Přednáška 10

TESTY PRO NOMINÁLNÍ A ORDINÁLNÍ PROMĚNNÉ – NEPARAMETRICKÉ METODY

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

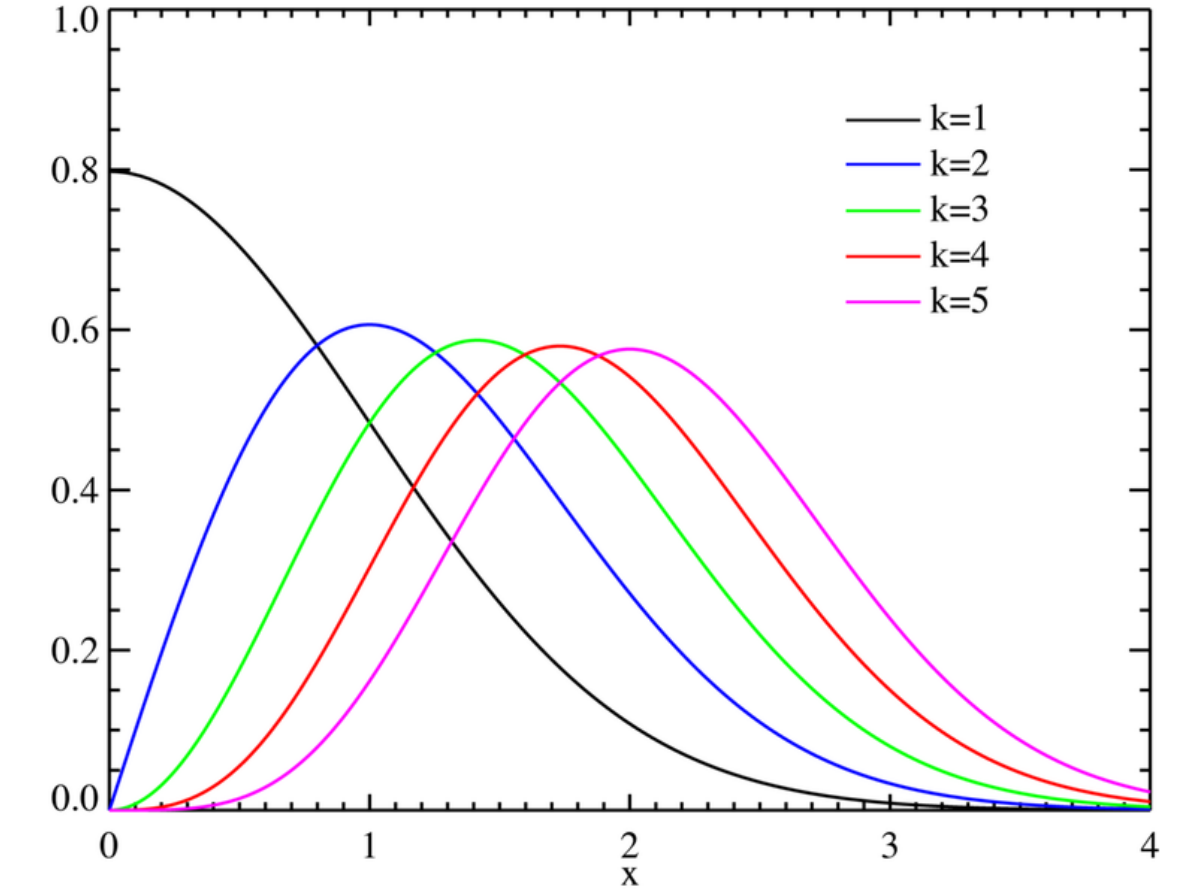
Smrť

Analýza četností hodnot nominální proměnné

- Výzkumné otázky…
  - Liší se významně preference nějakých 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 jedné populaci),
    - k rozdílu četností jevu mezi různými proměnnými (četnost jevu v různých populacích),
    - Nebo k pravděpodobnosti výskytu dvou (či více) jevů současně.
- Nominální proměnná
  - Též kategoriální, alternativní
  - Zařazení jevu do určité kategorie
  - Jednotlivé kategorie musí být vzájemně disjunktní – metodologie & logika věci
  - Kategorie mohou vzniknout i transformací z proměnné vyššího řádu – kategorizace pořadí, známek ve škole, „nižší úzkost x vyšší úzkost" atd.
  - Ordinální proměnné o málo opakujících se pořadích ($k<10$) mohou být analyzovány jako nominální
- Klíčová slova
  - Četnost, relativní četnost, očekávaná četnost, rezidua, χ^2 (Chi-kvadrát)

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

Rozdělení Chi^2



χ^2 – test dobré shody

- Liší se empirické četnosti nějakých jevů od teoreticky očekávaných četností?
  - Házení kostkou – kolikrát padne 1,2,…
  - Preference politických stran ve volbách…
  - Tedy jedna nominální proměnná, jeden výběr
- Testujeme pravděpodobnost daného rozdílu mezi empirickými a očekávanými hodnotami v rámci jednoho výběru
- H_0: $F(x) = F_0(x)$ vs. H_1: $F(x) \neq F_0(x)$

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

  - k je počet kategorií, n velikost vzorku, n_i pozorovaná četnost v kat. i, p_i teoretická pravděpodnost jevu v kategorii (0 až 1); $\sum n_i = \sum np_i$
- Rozdělení χ^2; stupně volnosti df = k-1
- Překoná-li hodnota χ^2 kritickou mez, H_0 zamítáme.
- Pro získání pravděpodobnosti χ^2 CHIDIST(x,volnost); CHIINV(prst, volnost)
- Očekávané četnosti… při uniformním rozložení 1:1:1…; nebo libovolně teoreticky odvozené (10:24:32…)
- N_i i NP_i vždy jako četnosti; nikdy ne procenta = relativní četnosti (ztráta informace o velikosti vzorku.

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

Ve kterém městě by jste žili nejraději?

Uniformní/náhodné rozdělení

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**		**p**	**1,000**

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

Ve kterém městě by jste žili nejraději?

Empirické rozdělení

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**		**p**	**0,02**

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

Závislost kategoriálních proměnných

- Jaká je souvislost preference politické strany a úrovně hrubého příjmu voliče?
- Jaká je pravděpodobnost společného výskytu dvou jevů z *x* a *y* možných? Podmínka disjunkce!
- 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 kategoriálních proměnných

- Postup analogický, jako u jednorozměrné verze testu dobré shody χ^2
- Očekávané četnosti: 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)

$$m_{ij} = \frac{n_{i.} n_{.j}}{n}$$

$$\chi^2 = \sum_{r=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

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**

Síla vztahu v kontingenční tabulce

- Koeficient kontingence (Pearson) C_{kor}
- Cramerovo *V*
  - Oba koeficienty v intervalu (0;1). Neindikují ovšem žádným způsobem „směr" vztahu. Směrů je v kontingenční tabulce mnoho :-)
- A proto… jsou kontingenční tabulky mnohdy účelné i tehdy, máme-li k dispozici data na vyšší úrovni měření.
- Možnost odhalení nelineárních vztahů
  - 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 reziduí 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* ☺
- Hendl str. 297 – 313.

AJ: strength of association, contingency coefficient, standardized residuals

Testy středních hodnot pro ordinální proměnné – neparametrické metody

- Metody užívající *parametrů* normálního rozložení nejsou dobře použitelné v případech, kdy
  - Data nepochází z normálního rozložení
  - Data mají ordinální charakter; nebo se jedná o krátké intervalové škály
  - Jsou malé výběry
  - Obecně parametry *m, s* nedávají dobrou informaci

- *Neparametrické* metody problém překonávají, jsou *robustní* vůči rozložení dat… *(nezávisí na parametrech norm. rozl.)*
  - Pro jeden výběr: znaménkový, ...
  - Pro párové srovnání: Marginal Homogeneity, ...
  - Pro 2 nezávislé výběry: Mann-Whitney U, Kolmogorov-Smirnov Z
  - a mnoho dalších...
  - na velkém vzorku je ale koneckonců robustní i *t*-test – platnost centrální limitní věty; ovšem pozor na bimodalitu a další „zvláštní jevy".

Non-parametric, robust, data assumptions, sign test, sample distribution etc.

Jeden výběr, znaménkový test

- Je „průměrná“ známka z matematiky v nějaké třídě „2“?
- Liší se empirická hodnota medianu od stanovené?
  - H_0: $Md = Md_0$; H_1: $Md \neq Md_0$ ... =>
  - H_0: $\sigma^2 = \sigma^2_0$; H1: $\sigma^2 \neq \sigma^2_0$
  - Pokud se hodnoty mediánů shodují, mělo by nad i pod teoretickým medianem být stejné množství případů
  - Asymptotický test pomocí normálního rozdělení:
    - rozdíly $d_i = x_i - Md_0$; Z_+ je počet kladných rozdílů, analogicky Z_-; $d_i = 0$ ignorujeme.
    - Platí-li H_0, $Z_+ = Z_-$. $Z_+ + Z_- = n$.
    - Testovací statistika:

      $z = (2Z_+ - n)/\sqrt{n}$
    - Padne-li statistika z do intervalu $\pm z_{\alpha/2}$, H_0 nezamítáme.
  - z má tvar asymptoticky normálního rozdělení, přesný test by využil binomického rozdělení.
- Jedná se tedy o alternativu t-testu pro jediný výběr;
- Pro závislé výběry *(=párové srovnání)* $d_i = x_i - y_i$; znaménkovým testem zkoumáme, zda pro H_0 střední hodnota $d = 0$.

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 H0, 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".
  - V asymptotické verzi testu je možné použít kvantily normálního rozložení pro:

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

	x	y	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.