PSY117/454
Statistická analýza dat v psychologii
Přednáška 11

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 preference politické strany 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 disjunktní – validita
  - Kategorie mohou vzniknout i transformací z proměnné vyššího řádu – kategorizace pořadí, známek ve škole, „nižší úzkost x vyšší úzkost“ atd.
- Klíčová slova
  - Četnost, relativní četnost, očekávaná četnost, rezidua, χ^2 (Chi-kvadrát)

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

χ^2 – test dobré shody

- Liší se empirické četnosti nějakých jevů od teoreticky očekávaných četností?
  - Házení kostkou – kolikrát padne 1,2,…
  - Preference stran
- H_0: $F(x) = F_0(x)$ vs. H_1: $F(x) \neq F_0(x)$

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

  - kde k je počet kategorií, n_i pozorovaná četnost v kat. i, p_i je očekávaná četnost
- 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 teoretické (10:24:32…)
- ! N empirických = N očekávaných

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

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*
- 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.

	B_1	B_2	…	B_x	Řádkové součty
A_1	n_{11}	n_{12}	…	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	…	n_{2s}	$n_{2.}$
…	…	…	…	…	…
A_x	n_{r1}	n_{r2}	…	n_{rs}	$n_{r.}$
Sloupcové součty	$n_{.1}$	$n_{.2}$	…	$n_{.s}$	n

AJ: contingency table (crosstabulation, ctosstab)

Závislost kategoriálních proměnných

- Postup analogický, jako u jednorozměrné verze testu χ^2
- Očekávané četnosti: m_{ij}
- χ^2
- Stupně volnosti: df = (r-1)*(s-1)

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

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

	B_1	B_2	...	B_x	Řádkové součty
A_1	n_{11}	n_{12}	...	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2s}	$n_{2.}$
...	...	...	...	...	...
A_x	n_{r1}	n_{r2}	...	n_{rs}	$n_{r.}$
Sloupcové součty	$n_{.1}$	$n_{.2}$	...	$n._s$	n

Síla vztahu v kontingenční tabulce

- Koeficient kontingence (Pearson) C_{kor}
- Cramerovo V
- C_{kor} se interpretuje jako Pearsonova korelace, V jako R^2. Tedy $C_{kor}^2 \approx V$.
- 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.
  - Nelineární vztahy
  - Možnost výpočtu reziduí: $n_{ij} - m_{ij} = res_i$
  - Součet residuí v tabulce vždy nula
  - Umožňují zjistit, kde jsou lokalizovány největší odchylky od náhodného rozložení četností v tabulce….
  - V SPSS: Standardizovaná rezidua (Pearsonova): rozdělení reziduí je normální s odchylkou 1; tedy standardizovaná rezidua s hodnotou +- 1,96 jsou „zajímavá".

- 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.
  - Pro jeden výběr: znaménkový, ...
  - Pro párové srovnání: Wilcoxon, ...
  - 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

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

Příklad I
Jeden výběr, znaménkový test

- Liší se hodnota medianu od stanovené?
  - H_0: $Md = Md_0$; H_1: $Md \neq Md_0$ ... =>
  - H_0: $\sigma^2 = \sigma^2_0$; H1: $\sigma^2 \neq \sigma^2_0$
  - Asymptotický test pomocí normálního rozdělení:
    - $d_i = x_i - Md_0$; Z_+ je počet kladných rozdílů, analogicky Z_-; $d_i = 0$ ignorujeme.
    - Platí-li H_0, $Z_+ = Z_-$. $Z_+ + Z_- = n$.
    - $z = (2Z_+ - n)/\sqrt{n}$
    - Padne-li statistika z do intervalu $\pm z_{\alpha/2}$, H_0 nezamítáme.
  - Přesný test by využil binomického rozdělení.
- Silnější alternativou je Wilcoxonův test pro jeden výběr; zohledňuje absolutní velikost rozdílů od Md_0.
- Pro závislé výběry $d_i = x_i - y_i$; znaménkovým nebo W-testem zkoumáme, zda pro H_0 střední hodnota $d = 0$.

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

Neparametrické testy pro nezávislé výběry

- Mediánový test
  - Je–li medián dvou výběrů shodný, leží na jedné straně *Md* 50% každého výběru.
  - Určíme *Md* pro celý soubor; četnosti hodnot ležících nad i pod *Md* by měly být stejné pro *x* i *y*.
  - V asymptotické verzi testu je možné použít kvantily normálního rozložení pro:

	x	y	Σ
<Md	*a*	*b*	*a+b*
>Md	*c*	*d*	*c+d*
Σ	*a+c*	*b+d*	*n*

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

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