Parciální asociační koeficienty v kontingenčních tabulkách

Jan Řehák, Ústav pro filosofii a sociologii ČSAV, Praha
Blanka Řeháková, Ústav pro výzkum veřejného mínění při FSÚ, Praha

Parciální asociační koeficienty jsou zaváděny analogicky k parciální korelaci a to pro řešení obdobných úloh v kontingenčních tabulkách. Pro nominální případ jsou známy parciální asociační koeficienty u Guttmanova λ (Goodman, Kruskal [1963]) a u Wallisova τ (Gray, Williams [1981]).

V tomto příspěvku je vyvinut obecný tvar parciálního asociačního koeficientu pro případ zobecněné závisle proměnné A, na niž působí komplex nezávislých nominálních proměnných $(B_1, B_2, \ldots, B_M)$, tj. pro model $(B_1, B_2, \ldots, B_M) \to A$. Vycházíme z výsledků D-modelu a z koeficientu explanační síly rozkladu (viz Řehák [1976], Řehák, Řeháková [1979, 1982]).

1. Základní definice a pojmy

Zobecněnou proměnnou $A = \{a_1, \ldots, a_K ; D\}$ nazveme rozkladové pravidlo $\{a_1, \ldots, a_K\}$ a matici $D = \|d_{ij}\|$ o rozměru KxK, jejíž prvky splňují podmínky:

a) $d_{ii} = 0 \quad (i = 1, \ldots, K)$
b) $d_{ij} = d_{ji} \quad (i, j = 1, \ldots, K)$
c) $d_{ij} \geq 0 \quad (i, j = 1, \ldots, K)$
d) d_{ij} je skórem nepodobnosti mezi kategoriemi a_i, a_j, tj. $d_{ij} = d(a_i, a_j)$ nabývá tím vyšší hodnoty, čím nepodobnější jsou kategorie a_i, a_j.

Označíme $f' = (f_1, f_2, \ldots, f_K)$ rozložení na A, $\sum f_i = 1$, $f_i \geq 0$ $(i = 1, \ldots, K)$. Zobecněnou varianci distribuce f na A definujeme jako

$$(1) \qquad \mathrm{Gvar}\, f = \sum_i \sum_j f_i f_j d_{ij} = f' D f$$

Pro kontingenční tabulku RxK označme $f_{(r)}' = (f_{1|r}, \ldots, f_{K|r})$ podmíněné rozložení r-tého řádku, $w' = (w_1, \ldots, w_R)$ marginální sloupcové rozložení a $f = \sum_r w_r f_{(r)}$ marginální řádkové rozložení. Platí věta o rozkladu zobecněné variance

$$(2) \qquad \mathrm{Gvar}\, f = \sum_r w_r \mathrm{Gvar}\, f_{(r)} + \frac{1}{2} \sum_r \sum_s w_r w_s D^2(f_{(r)}, f_{(s)}) = \sum_r w_r \mathrm{Gvar}\, f_{(r)} + \sum_r w_r D^2(f_{(r)}, f),$$

kde $D(.,.)$ je vzdálenost mezi distribucemi:

$$(3) \qquad D^2(f_{(r)}, f_{(s)}) = (f_{(r)} - f_{(s)})' D (f_{(s)} - f_{(r)}).$$

Podle analogie k definici korelačního poměru η^2 definujeme koeficient explanační síly rozkladu B (řádková proměnná kontingenční tabulky) jako

$$(4) \qquad \delta_{A|B} = 1 - \frac{\sum_r w_r \mathrm{Gvar}\, f_{(r)}}{\mathrm{Gvar}\, f}$$

2. Vícenásobná a parciální asociace

Vícenásobný asociační koeficient pro měření síly vztahu, který se projevuje mezi souborem proměnných $B_1, B_2, \ldots, B_M$ a závislou proměnnou A, tj. pro model $(B_1, B_2, \ldots, B_M) \rightarrow A$ definujeme jednoduše jako asociační koeficient δ mezi kombinací všech B_m a A:

$$(5) \qquad \delta_{(B_1, \ldots, B_M) \rightarrow A} = \delta_{A \mid (B_1, \ldots, B_M)} = \delta_{A \mid B_1 \times \ldots \times B_M}$$

Parciální koeficient vyjadřuje přínos proměnné B_M k variabilitě proměnné A poté, co byl již zjištěn vliv $B_1, B_2, \ldots, B_{M-1}$. Je to čistý vliv B_M po odečtení vlivů ostatních proměnných, intenzita vztahu $B_M \rightarrow A$, jsou-li $B_1, B_2, \ldots, B_{M-1}$ konstantní.

$$(6) \qquad \delta_{A \mid B_M}(B_1, \ldots, B_{M-1}) = \frac{\delta_{A \mid B_1 \times \ldots \times B_M} - \delta_{A \mid B_1 \times \ldots \times B_{M-1}}}{1 - \delta_{A \mid B_1 \times \ldots \times B_{M-1}}}$$

je to tedy relativní přírůstek asociace komplexu $B_1 \times \ldots \times B_M$ oproti $B_1 \times \ldots \times B_{M-1}$.

Pomocí zobecněných varianci lze vzorec (6) zapsat jako (zde pro M=2, pro vyšší M se nahradí B_1 kombinací proměnných)

$$(7) \qquad \delta_{A \mid B_2}(B_1) = 1 - \frac{\sum_r^R \sum_s^S w_{rs}\ \mathrm{Gvar}\ f_{(rs)}}{\sum_r^R w_r\ \mathrm{Gvar}\ f_{(r)}},$$

kde $f_{(rs)}$ jsou podmíněná rozložení A, jsou-li podmínkami jednotlivé kombinace hodnot proměnných B_1, B_2 (r-tá hodnota B_1 a s-tá hodnota B_2) a w_{rs} je marginální rozložení kombinace $B_1 \times B_2$ v tabulce $B_1 \times B_2 \times A$.

Parciální koeficient má vlastnosti (formulováno pro M=2):

a) δ je definován kdykoliv je definován $\delta_{A \mid B_1}$ a $\delta_{A \mid B_1} < 1$ (tj. existuje ještě prostor pro vysvětlení zbytkové variability A, která nebyla vlivem B_1 zcela vysvětlena).

b) $0 \leq \delta \leq 1$.

c) $\delta = 0$ právě když $\delta_{A \mid B_1 \times B_2} = \delta_{A \mid B_1}$, tj. přidáním nové proměnné se celková asociace neměni, informace o A uložená v B_2 se již plně projevila v rozkladu B_1.

d) $\delta = 1$ právě když $\delta_{A \mid B_1 \times B_2} = 1$, tj. B_2 vysvětluje plně zbytkovou variancí, kterou u A zanechává B_1.

Mezi koeficienty parciální a vícenásobné asociace platí analogické vztahy jako u korelační analýzy:

(8) $$1 - \delta_{A \mid B_1 \times B_2} = (1 - \delta_{A \mid B_1})(1 - \delta_{A \mid B_2(B_1)})$$

$$1 - \delta_{A \mid B_1 \times B_2 \times B_3} = (1 - \delta_{A \mid B_1})(1 - \delta_{A \mid B_2(B_1)})(1 - \delta_{A \mid B_3(B_1, B_2)})$$

atd.

3. Asymptotický rozptyl koeficientů

Pro maximálně věrohodné odhady $\hat{\delta}$ koeficientů δ platí, že jsou konzistentní a asymptoticky normální. Dále uvádíme asymptotický rozptyl koeficientů pro případ multinomického výběru přes celou tabulku (výsledky jsou vzaty z prací: Statistická dokumentace k programu DIANA (B. Řeháková [1979]) pro (4) a z tezí k disertační práci (B. Řeháková [1982]) pro (7) .

Věta. Je-li $\hat{\delta}$ maximálně věrohodný odhad δ při multinomickém výběru o velikosti n přes celou kontingenční tabulku, pak platí:

a) $\sqrt{n}(\hat{\delta} - \delta) \underset{n \to \infty}{\sim} N(0, \sigma^2(p))$ za předpokladu, že $\sigma^2(p) > 0$ a všechny marginální četnosti jsou nenulové; odhad rozptylu je $\hat{\sigma}^2(f) = \hat{\sigma}^2$.

b) pro $\delta_{A \mid B_1}$ je při $w_r = f_{r.}$

$$\hat{\sigma}^2 = \frac{1}{\xi^4} \sum_{b=1}^{R} \sum_{a=1}^{K} f_{ba} \left\{ \nu \left(2 \sum_{k=1}^{K} f_{bk} d_{ka} - \xi \right) - \xi \left(2 \sum_{a=1}^{K} \frac{f_{bk}}{f_{b.}} d_{ka} - \text{Gvar } f_{(b)} \right) \right\}^2,$$

(9) $$\nu = \sum_{r=1}^{R} f_{r.} \quad \text{Gvar } f_{(r)} \quad , \quad \xi = \text{Gvar } f_{\cdot}$$

c) pro $\delta_{A \mid B_2(B_1)}$ je při $w_{rs} = f_{rs.}$, $w_r = f_{r..}$

$$\hat{\sigma}^2 = \frac{1}{\xi^4} \sum_{b=1}^{R} \sum_{c=1}^{S} \sum_{a=1}^{K} f_{bca} \left\{ \nu \left(2 \sum_{k=1}^{K} d_{ka} \frac{f_{b.k}}{f_{b..}} - \text{Gvar } f_{(b)} \right) - \right.$$

(10) $$\left. - \xi \left(2 \sum_{k=1}^{K} d_{ka} \frac{f_{bck}}{f_{bc.}} - \text{Gvar } f_{(bc)} \right) \right\}^2 ,$$

$$\nu = \sum_{r=1}^{R} \sum_{s=1}^{S} f_{rs.} \text{ Gvar } f_{(rs)} , \quad \xi = \sum_{r=1}^{R} f_{r..} \text{ Gvar } f_{(r)} .$$

Při aplikaci určujeme 100 γ %-ní intervaly spolehlivosti jako $\delta = \hat{\delta} \pm \frac{1}{\sqrt{n}} z_\gamma \hat{\sigma}$. Pro nominální případ (parciální Wallisovo τ) byl asymptotický rozptyl uveden (v jiném tvaru) v práci Gray, Williams [1981]. Obdobné výsledky je možno obdržet pro jiné typy multinomického výběru určené stratifikací podle některých proměnných $B_1, \ldots, B_M$.

Závěr - aplikační poznámky

Použití parciální asociace je výhodné tam, kde je konstatována existence nebo neexistence vztahu a určení závislostního modelu nedostačující

a kde požadujeme navíc měření síly vztahů a určení vlivu jednotlivých proměnných. Výsledky tohoto příspěvku jsou podstatné pro analytické situace, v nichž závislá proměnná je jiného typu než běžná prostá klasifikace. V praxi autorů se vyskytlo již několik typů zobecněných proměnných, zejména:

a) typ geografických vztahů mezi lokalitami $a_{\bullet}$; zde d_{ij} je čtverec vzdálenosti, nebo vzdálenost,

b) typ proměnné, u níž vztahy mezi lokalitami jsou dány sousedstvím nebo zprostředkovaným sousedstvím ,

c) kombinované typologie, u nichž d_{ij} je určeno jako počet vlastností, kterým se kategorie odlišují, např. u kombinace tří vlastností A, B, C : $d(AB\bar{C}, A\bar{B}C) = 2$, $d(ABC, AB\bar{C}) = 1$, $d(ABC, \bar{A}\bar{B}C) = d(AB\bar{C}, \bar{A}\bar{B}C) = 3$ atd.

Důležitost modelu spočívá v tom, že poskytuje jednotný modelový základ pro nejrůznější typy závisle proměnné A a tím i srovnatelnost analýz pro různé případy.

LITERATURA:

Goodman L. A., Kruskal W. H. [1963] : Measures of Association for Cross Classifications III: Approximate Sampling Theory. JASA 58, str. 310 - 364

Gray L. N., Williams J. S. [1981] : Goodman and Kruskal's Tau, Multiple and partial Analogs. Sociological Methods and Research 10, str. 50 - 62

Řehák J. [1976] : Základní deskriptivní míry pro rozložení ordinálních dat. Sociologický časopis XII, str. 416 - 431

Řehák J., Řeháková B. [1979] : Základní charakteristiky proměnných s konečným počtem hodnot a distanční analýza jejich rozložení. Sociologický časopis XV, str. 214 - 233

Řehák J., Řeháková B. [1982] : Distanční přístup k analýze kategorizovaných dat a jeho aplikace na problém shody. Robust, Podkost, str. 76 - 80

Řeháková B. [1979] : Statistická dokumentace k programu DIANA

Řeháková B. [1982] : Téze k disertační práci .