Asymptotické testy hypotéz pro rozdělení D-proměnné

Blanka Řeháková, Ústav pro výzkum veřejného mínění při FSÚ, Praha

Model pro obecné kategorizované proměnné, u jejichž hodnot (kategorií) uvažujeme číselně vyjádřitelné relace nepodobnosti, a jeho teorie byly popsány v pracech Řeháka [1976], Řeháka a Řehákové [1979], [1984], [1986a], [1986b], Řehákové [1982], [1985]. Příspěvek předkládá asymptotické testy dobré shody a homogenity pro kontingenční tabulky, v nichž se obecná proměnná (dále D-proměnná) vyskytuje v roli závisle proměnné.

Testy jsou založeny na předpokladu multinomického výběru vytvářejícího tabulku. Odtud pak plyne asymptotická normalita rozdělení četností a z ní použitelnost výsledků teorie rozdělení kvadratických forem normálně rozdělených veličin. Testové statistiky navržené v práci Řehákové [1985] vyplývají přirozeně z modelu, mají tvar kvadratických forem a vycházejí ze (semi)metriky.

<u>Základní definice a vlastnosti</u>

D-proměnnou $A = \{A_1, A_2, \ldots, A_K; \underset{\sim}{D}\}$ nazveme úplnou soustavu jevů $A_1, A_2, \ldots, A_K$ spolu s maticí $\underset{\sim}{D} = \| d_{jk} \|$ typu KxK, pro jejíž prvky $d_{ij} = d(A_i, A_j)$ platí

identita: $d_{ii} = 0$ pro $i = 1, \ldots, K$,
symetrie: $d_{ij} = d_{ji}$ pro $i, j = 1, \ldots, K$,
nezápornost: $d_{ij} \geq 0$ pro $i, j = 1, \ldots, K$ a
$d_{ij} > 0$ pro alespoň jednu dvojici i, j,
interpretabilita: d_{ij} je tím větší, čím nepodobnější jsou si kategorie A_i, A_j.

Prvky d_{ij} matice D nazveme skóry vzdálenosti kategorií A_i, A_j, $\underset{\sim}{D}$ nazveme maticí skórů vytvářejících typ proměnné.

Nominální, ordinální a kardinální kategorizované proměnné jako speciální případy D-proměnné jsou vytvářeny po řadě maticemi s prvky $d_{ij} = 1 - \delta_{ij}$ ($\delta_{ij} = 1$ pro $i = j$, $\delta_{ij} = 0$ pro $i \neq j$), $d_{ij} = | i - j |$, $d_{ij} = (x_i - x_j)^2$ (x_i je číselná hodnota přiřazená kategorii A_i).

Označíme $Q_K = \{\underset{\sim}{f} : \underset{\sim}{f} = (f_1, f_2, \ldots, f_K)', f_k \geq 0,\ k = 1, 2, \ldots, K,\ \sum_1^K f_k = 1\}$ K-rozměrný simplex rozdělení. Pro volbu testových statistik jsou důležité následující vlastnosti:

1. Funkce $D(\underset{\sim}{f}, \underset{\sim}{g}) = \sqrt{(\underset{\sim}{f} - \underset{\sim}{g})' \underset{\sim}{D} (\underset{\sim}{g} - \underset{\sim}{f})}$ je (semi)metrika, právě když matice $\underset{\sim}{D}^* = \| d^*_{ij} \|$ typu (K-1)x(K-1), pro niž $d^*_{ij} = d_{iK} + d_{Kj} - d_{ij}$ je pozitivně (semi)definitní matice.

2. Pro zobecněnou varianci $\text{Gvar}\ \underset{\sim}{f} = \underset{\sim}{f}' \underset{\sim}{D} \underset{\sim}{f}$, $\underset{\sim}{f} \in Q_K$ platí za předpokladu existence semimetriky D(.,.) rozklad: je-li $\underset{\sim}{f} = \sum w_r \underset{\sim}{f}_{(r)}$, $\underset{\sim}{f}_{(r)} \in Q_K$, $r = 1, \ldots, R$, $\underset{\sim}{w} \in Q_R$, pak

$$\text{Gvar}\ \underset{\sim}{f} = \sum_r w_r\, \text{Gvar}\ \underset{\sim}{f}_{(r)} + \frac{1}{2} \sum_r \sum_s w_r w_s D^2(\underset{\sim}{f}_{(r)}, \underset{\sim}{f}_{(s)}) = \sum_r w_r \text{Gvar}\ \underset{\sim}{f}_{(r)} + \sum_r w_r D^2(\underset{\sim}{f}_{(r)}, \underset{\sim}{f}).$$

Testy jsou založeny na tomto výsledku (viz např. Ruben [1962]). Nechť $\underset{\sim}{X} = (X_1, \ldots, X_K)'$ má rozdělení $N(\underset{\sim}{0}, \underset{\sim}{\Sigma})$, $h(\underset{\sim}{\Sigma}) = K$ a $Y = \underset{\sim}{X}' \underset{\sim}{A} \underset{\sim}{X}$, kde $\underset{\sim}{A}$ je symetrická pozitivně definitní nebo pozitivně semidefinitní matice. Pak $L[Y] = L[\sum_{i=1}^{K} \lambda_i Z_i^2]$, kde $Z_1^2, \ldots, Z_K^2$ jsou nezávislé náhodné veličiny, z nichž každá má rozdělení chí-kvadrát o jednom stupni volnosti a $\lambda_1, \ldots, \lambda_K$ jsou charakteristická čísla matice $\underset{\sim}{\Sigma} \underset{\sim}{A}$.

Dosažené hladiny významnosti $1 - P(\underset{\sim}{X}' \underset{\sim}{A} \underset{\sim}{X} \leq t)$ lze počítat pomocí algoritmu AS 106 (viz Sheil, O'Muircheartaigh [1977], který kombinuje výsledky Rubena [1962] a Kotze, Johnsona, Boyda [1967]:

$$P(\underset{\sim}{X}' \underset{\sim}{A} \underset{\sim}{X} \leq t) = \sum_{k=0}^{\infty} c_k\, F(K' + 2k, t/\beta), \quad t > 0$$
$$= 0, \quad t \leq 0$$

kde $F(K', y)$ je distribuční funkce centrálního rozdělení chí-kvadrát o K' stupních volnosti, $K' = h(\sum \underset{\sim}{A}) = h(\underset{\sim}{A})$. (Volba konstatny β, výpočet c_k a další podrobnosti viz výše uvedená literatura.)

Následující testy A–C zavedla Řeháková [1985], test D Řehák (nepublikováno).

A) Test dobré shody s předpokladem

Nechť $\underset{\sim}{f} \in Q_K$ je vektor relativních četností pro $A = \{A_1, \dots, A_K; \underset{\sim}{B}\}$, $n\underset{\sim}{f} \sim M_K(n, \underset{\sim}{p})$, nechť $\underset{\sim}{\pi} \in Q_K$ je daný vektor. Uvažujme hypotézu $H_0: \underset{\sim}{p} = \underset{\sim}{\pi}$ a alternativu $H_A: D(\underset{\sim}{p}, \underset{\sim}{\pi}) \neq 0$. Testová statistika je $D^2(\underset{\sim}{f}, \underset{\sim}{\pi}) = (\bar{\underset{\sim}{f}} - \bar{\underset{\sim}{\pi}})' \underset{\sim}{D}^{*} (\bar{\underset{\sim}{f}} - \bar{\underset{\sim}{\pi}})$, kde $\bar{\underset{\sim}{f}}$ je vektor, který má za složky prvních $K-1$ složek vektoru $\underset{\sim}{f}$, obdobně $\bar{\underset{\sim}{\pi}}$. Za hypotézy H_0 platí, že při $n \to \infty$

$$L[\, n D^2(\underset{\sim}{f}, \underset{\sim}{\pi})\,] \to L\Big[\sum_{i=1}^{K-1} \lambda_i Z_i^2\Big],$$

kde Z_i^2 $(i=1,\dots,K-1)$ jsou nezávislé náhodné veličiny s rozdělením chí-kvadrát o jednom stupni volnosti a λ_i $(i=1,\dots,K-1)$ jsou charakteristická čísla matice $(\underset{\sim}{D}_{\pi} - \bar{\underset{\sim}{\pi}}\bar{\underset{\sim}{\pi}}')\underset{\sim}{D}^{*}$, $\underset{\sim}{D}_{\pi} = \operatorname{diag}\{\pi_1, \dots, \pi_{K-1}\}$. (Matice $\underset{\sim}{D}_{\pi} - \bar{\underset{\sim}{\pi}}\bar{\underset{\sim}{\pi}}'$ je regulární právě když $\pi_k > 0$ pro $k=1,\dots,K$).

B) Test shody dvou rozdělení

Dvouvýběrový problém je u D-proměnných řešitelný pomocí statistiky $D^2(\underset{\sim}{f}, \underset{\sim}{g}) = (\bar{\underset{\sim}{f}} - \bar{\underset{\sim}{g}})' \underset{\sim}{D}^{*} (\bar{\underset{\sim}{f}} - \bar{\underset{\sim}{g}})$, kde $\underset{\sim}{f}, \underset{\sim}{g} \in Q_K$ jsou dvě nezávislá rozdělení proměnné $A = \{A_1, \dots, A_K; \underset{\sim}{B}\}$, $n\underset{\sim}{f} \sim M_K(n, \underset{\sim}{p})$, $m\underset{\sim}{g} \sim M_K(m, \underset{\sim}{r})$.

$H_0: \underset{\sim}{p} = \underset{\sim}{r}$, $H_A: D(\underset{\sim}{p}, \underset{\sim}{r}) \neq 0$.

Za hypotézy H_0 platí, že když n a m jdou do nekonečna tak, že $\lim m/(m+n) = \omega$, $0 < \omega < 1$, pak

$$L[\,(n+m) D^2(\underset{\sim}{f}, \underset{\sim}{g})\,] \to L\Big[\sum_{i=1}^{K-1} \lambda_i Z_i^2\Big],$$

kde Z_i^2 jsou nezávislé náhodné veličiny s rozdělením chí kvadrát o jednom stupni volnosti a λ_i $(i=1,\dots,K-1)$ jsou charakteristická čísla matice

$$\frac{1}{\omega(1-\omega)} (\underset{\sim}{D}_p - \bar{\underset{\sim}{p}}\bar{\underset{\sim}{p}}')\underset{\sim}{D}^{*}, \quad \underset{\sim}{D}_p = \operatorname{diag}\{p_1, \dots, p_{K-1}\}.$$

V praxi vezmeme za ω číslo $m/(m+n)$ a místo $\underset{\sim}{p}$ dosadíme vektor se složkami $\hat{p}_i = (n_i + m_i)/(n+m)$ za předpokladu, že $n_i + m_i > 0$ $(i=1,\dots,K)$.

Test je speciálním případem obecnější situace, v níž se porovnává $R \geq 2$ rozdělení (viz C). Jeho důležitost je dána nejen častým výskytem dvouvýběrového problému, ale též proto, že porovnání $R > 2$ rozdělení lze rozložit do $R(R-1)/2$ porovnání dvojic při použití simultánní inference.

C) Test shody R rozdělení

Nechť $\underset{\sim}{f}_{(1)}, \dots, \underset{\sim}{f}_{(R)}$ jsou nezávislé náhodné vektory (řádková rozdělení četností proměnné $A = \{A_1, \dots, A_K; \underset{\sim}{B}\}$ v tabulce $R \times K$), nechť $\underset{\sim}{w} = (w_1, \dots, w_R)' \in Q_R$, $w_r > 0$ $(r=1,\dots,R)$, speciálně $w_r = n_r/n$, kde n_r je rozsah r-tého výběru, $n = \sum n_r$. Rozdělení $n_r \underset{\sim}{f}_{(r)}$ je $M_K(n_r, \underset{\sim}{p}_{(r)})$, $r = 1,\dots,R$. Označme $\underset{\sim}{f} = \sum w_r \underset{\sim}{f}_{(r)}$.

Testovou statistikou je druhý člen rozkladové formule pro Gvar $\underset{\sim}{f}$ a to

$$\sum w_r D^2(\underset{\sim}{f}_{(r)}, \underset{\sim}{f}) = (\bar{\underset{\sim}{f}}_{(1)} - \bar{\underset{\sim}{f}}, \dots, \bar{\underset{\sim}{f}}_{(R-1)} - \bar{\underset{\sim}{f}})' (\underset{\sim}{W}_{R-1} \otimes \underset{\sim}{D}^{*}) (\bar{\underset{\sim}{f}}_{(1)} - \bar{\underset{\sim}{f}}, \dots, \bar{\underset{\sim}{f}}_{(R-1)} - \bar{\underset{\sim}{f}}),$$

kde $\underset{\sim}{W}_{R-1}$ je symetrická pozitivně definitní matice s prvky

$$W_{rr} = \frac{w_r(w_R + w_r)}{w_R}, \quad W_{rs} = \frac{w_r w_s}{w_R}, \quad r \neq s, \quad r,s = 1,\dots,R-1,$$

$\otimes$ je Kroneckerův součin matic.

H_o : $\underset{\sim}{p}_{(1)} = \dots = \underset{\sim}{p}_{(R)}$ $(= \underset{\sim}{p})$,

H_A : alespoň pro jednu dvojici $r \neq s$ je $D(\underset{\sim}{p}_{(r)}, \underset{\sim}{p}_{(s)}) \neq 0$. Nechť n_r jdou do nekonečna tak, že $n_r/n \to \omega_r$, $0 < \omega_r < 1$, $(r=1,\dots,R)$, $\sum \omega_r = 1$. Pak za hypotézy H_o

$$L\left[n \sum_{r=1}^{R} w_r D^2 (\underset{\sim}{f}_{(r)}, \underset{\sim}{f}) \right] \longrightarrow L\left[\sum_{i=1}^{(R-1)(K-1)} \lambda_i Z_i^2 \right],$$

kde Z_i^2 jsou nezávislé náhodné veličiny, z nichž každá má rozdělení chí-kvadrát o jednom stupni volnosti a λ_i $(i=1,\dots,(R-1)(K-1))$ jsou charakteristická čísla matice $(\underset{\sim}{B}_{R-1} \otimes \underset{\sim}{\Sigma}_{K-1})(\underset{\sim}{E}_{R-1} \otimes \underset{\sim}{E}^{m})$, kde $\underset{\sim}{\Sigma}_{K-1} = \underset{\sim}{D}_{\bar{p}} - \bar{\underset{\sim}{p}}\,\bar{\underset{\sim}{p}}'$, $\underset{\sim}{D}_{\bar{p}} = \mathrm{diag}\{p_1,\dots,p_{K-1}\}$ a $\underset{\sim}{B}_{R-1}$ je symetrická positivně definitní matice typu $(R-1) \times (R-1)$ s prvky

$$b_{rr} = \frac{1}{\omega_r} - \frac{2w_r}{\omega_r} + \sum_{i=1}^{R} \frac{w_i^2}{\omega_i}, \quad b_{rs} = -\frac{w_r}{\omega_r} - \frac{w_s}{\omega_s} + \sum_{i=1}^{R} \frac{w_i^2}{\omega_i}, \quad r \neq s.$$

V praxi vezmeme za ω_r číslo n_r/n a za $\underset{\sim}{p}$ dosadíme všude $\hat{\underset{\sim}{p}}$ se složkami $\hat{p}_k = \sum_r n_{rk}/n$ za předpokladu, že $\hat{p}_k > 0$ pro $k=1,\dots,K$.

D) Test shody marginálních rozdělení ve čtvercové tabulce

Nechť $A = \{A_1,\dots,A_K; \underset{\sim}{D}\}$, $B = \{B_1,\dots,B_K; \underset{\sim}{D}\}$ jsou dvě D-proměnné o stejném počtu kategorií a stejné matici škál $\underset{\sim}{D}$. Čtvercová tabulka $A \times B$ vznikla jako výběr o rozsahu n z multinomického rozdělení $M(n; p_{11},\dots,p_{KK})$. Označme n_{ij} četnost v poli (i,j), $n_{+j} = \sum_i n_{ij}$, $n_{i+} = \sum_j n_{ij}$, $f_{i+} = n_{i+}/n$, $f_{+j} = n_{+j}/n$.

H_o: $\underset{\sim}{p}_A = \underset{\sim}{p}_B$, H_A : $D(\underset{\sim}{p}_A, \underset{\sim}{p}_B) \neq 0$, $\underset{\sim}{p}_A = (p_{1+},\dots,p_{K+})'$, $\underset{\sim}{p}_B = (p_{+1},\dots,p_{+K})'$.

Testová statistika je $D^2(\underset{\sim}{f}_A, \underset{\sim}{f}_B) = (\bar{\underset{\sim}{f}}_A - \bar{\underset{\sim}{f}}_B)' \underset{\sim}{E}^{m} (\bar{\underset{\sim}{f}}_A - \bar{\underset{\sim}{f}}_B)$, $\bar{\underset{\sim}{f}}_A = (f_{1+},\dots,f_{K-1,+})'$, $\bar{\underset{\sim}{f}}_B = (f_{+1},\dots,f_{+,K-1})'$. Za platnosti H_o

$$L\left[n D^2 (\underset{\sim}{f}_A, \underset{\sim}{f}_B) \right] \longrightarrow L\left[\sum_{i=1}^{K-1} \lambda_i Z_i^2 \right],$$

kde Z_i^2 jsou nezávislé náhodné veličiny, z nichž každá má rozdělení chí-kvadrát s jedním stupněm volnosti a λ_i $(i=1,\dots,K-1)$ jsou charakteristická čísla matice $\underset{\sim}{\Sigma}\,\underset{\sim}{E}^{m}$, kde $\underset{\sim}{\Sigma}$ je symetrická positivně definitní matice typu $(K-1) \times (K-1)$ s prvky

$$\sigma_{ii} = p_{i+} + p_{+i} - 2p_{ii}, \quad \sigma_{ij} = -(p_{ij} + p_{ji}), \quad i \neq j$$

předpokládáme $p_{ij} > 0$ pro $i,j = 1,\dots,K$).
V praxi použijeme odhadů $\hat{p}_{ij} = n_{ij}/n$, $\hat{p}_{i+} = n_{i+}/n$, $\hat{p}_{+i} = n_{+i}/n$.

LITERATURA

Kotz S., Johnson N.L., Boyd D.W. [1967] Series Representations of Distributions of Quadratic Forms in Normal Variables I. Central Case. AMS 38, 838-848.

Ruben H. [1962] Probability Content of Regions under Spherical Normal Distributions, IV: The Distribution of Homogeneous and Non-homogeneous Quadratic Functions of Normal Variables. AMS 33, 542-570.

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

Řehák J., Řeháková B. [1979] Základní charakteristiky proměnných s konečným počtem hodnot a distanční analýza jejich rozložení. Sociologický časopis XV, 214-231

Řehák J., Řeháková B. [1984] Parciální asociační koeficienty v kontingenčních tabulkách. In: Antoch J., Jurečková J. (ed.) Robust 84, 105-108.

Řehák J., Řeháková B. [1986a] Classifications with relations: a model for the description of distributions and their distances. Kybernetika (v tisku).

Řehák J., Řeháková B. [1986b] Vícenásobná a parciální asociace v kontingenčních tabulkách. Sociologický časopis (v tisku).

Řeháková B. [1982] Koeficienty parciální asociace pro obecný typ kategorizované proměnné. Práce k aspirantskému minimu.

Řeháková B. [1985] Model a metoda pro analýzu kategorizovaných dat s relacemi. Kandidátská disertační práce.

Sheil J., O'Muircheartaigh I. [1977] The Distribution of Non-negative Quadratic Forms in Normal Variables. Applied Statistics 26, 92-98.