JACK-KNIFEOVÁNÍ KOEFICIENTŮ ASOCIACE II

Jaromír Běláček, GGÚ ČSAV

Stejnojmenný příspěvek ze sborníku ROBUST 86 poukázal na možnost aplikace nejjednodušších transformací typu jackknife na koeficientech asociace (matematicky hladkých funkcích frekvenčního multinomického vektoru), které poskytují méně vychýlené odhady empirických koeficientů a alternativní (neparametrické) odhady rozptylu. Smyslem tohoto příspěvku je poskytnout informaci o univerzálních zobecněních těchto metod konstruovaných pro eliminaci vychylujících členů vyšších řádů nežli $O(n^{-1})$, a to i pro situaci několika nezávislých náhodných výběrů. Na příkladech vybraných koeficientů asociace jsou referovány některé zobecnitelné praktické závěry.

V návaznosti na terminologickou diskusi odehrávající se po odeznění prvotního příspěvku na ZŠ ROBUST 86 si dovoluji upozornit, že termín "jackknife" znamená podle doslovného překladu (viz.[5],str.1174) "velký kapesní zavírací nůž, zavírák" . V tomto smyslu interpretoval "jackknife" i profesor P.K.Sen na své loňské předvánoční přednášce na KPMS přiřazujíce jednotlivým čepelím svého "nože" rozličné speciální interpretace. Jak mne ovšem upozornil můj pozorný kolega dr.P.Hrala ([6]), lze na základě věrných historických pramenů přisuzovat anglické terminologické přiřazení dané metodě pracovním aktivitám českých národnostních menšin před druhou světovou válkou. Jak známo Valaši (jinak též profesionální klestiči dobytka) se proslavili právě výrobou a aktivním používáním kapesních nožů, jejichž prostřednictvím účinně potlačovali nežádoucí aktivity domácího skotu po celé Evropě. Slušno podotknout, že kategorická informace o tom, že adekvátní český ekvivalent anglického "jackknife" je tudíž "okleštování", způsobila pozitivní nezájem mého nejbližšího nadřízeného dr.J.Řeháka o další terminologické nuance.

1. JEDNOVÝBĚROVÁ JACKKNIFEOVÁ TRANSFORMACE K-TÉHO ŘÁDU

Pro libovolný odhad Θ_n pořízený na základě náhodného výběru o rozsahu n a $k<n$ zaveďme veličiny

$$(1)\qquad \bar{\Theta}_{n-v} = \begin{cases} \binom{n}{v}^{-1} \sum\limits_{1\le i_1<\dots<i_v\le n} \Theta_{-i_1,\dots,-i_v} & \text{pro } v=1,\dots,k \\ \Theta_n & \text{pro } v=0 \end{cases}$$

jako průměry všech $\binom{n}{v}$ odhadů získaných z původního výběru při postupném vynechávání v-tic pozorování. Jednovýběrový jackknifeový odhad k-tého řádu byl definován v [10] předpisem

$$(2)\qquad J_k(\Theta_n) := \frac{\begin{vmatrix} \bar{\Theta}_{n-0} & \bar{\Theta}_{n-1} & \dots & \bar{\Theta}_{n-k} \\ n^{-1} & (n-1)^{-1} & \dots & (n-k)^{-1} \\ \vdots & \vdots & & \vdots \\ n^{-k} & (n-1)^{-k} & \dots & (n-k)^{-k} \end{vmatrix}}{\begin{vmatrix} 1 & 1 & \dots & 1 \\ n^{-1} & (n-1)^{-1} & \dots & (n-k)^{-1} \\ \vdots & \vdots & & \vdots \\ n^{-k} & (n-1)^{-k} & \dots & (n-k)^{-k} \end{vmatrix}}$$

jakožto řešení pro $\hat{\Theta}$ soustavy lineárních rovnic

$$\bar{\Theta}_{n-\hat{v}} = \hat{\Theta} + \sum_{v=1}^{k} \hat{\alpha}_v \cdot (n-\hat{v})^{-v} \quad , \quad \hat{v}=0,1,\dots,k \tag{3}$$

v proměnných $\hat{\Theta}, \hat{\alpha}_1, \dots, \hat{\alpha}_k$. Vzhledem k tomu, že Vandermondův determinant matice soustavy (3) -tj.jmenovatel zlomku v (2)- je různý od nuly, je $J_k(\Theta_n)$ definován jednoznačně jako lineární kombinace $\bar{\Theta}_{n-0}, \dots, \bar{\Theta}_{n-k}$ s pevně určenými koeficienty. Rozvojem čitatele v (2) podle prvního řádku lze získat explicitní vyjádření

$$J_k(\Theta_n) = \frac{1}{k!}\left[\sum_{v=0}^{k}(-1)^v \binom{k}{v}(n-v)^k \bar{\Theta}_{n-v}\right] \tag{4}$$

snad poprvé uvedené v [12].

Základní vlastnost transformace $J_k(.)$ vyplývá z níže uvedeného tvrzení -

-<u>Lemma 1</u>: Nechť odhad Θ_n má konečný rozvoj střední hodnoty typu

$$E\,\Theta_n = \Theta + \sum_{v=1}^{\infty} \alpha_v \cdot n^{-v} \quad , \tag{5}$$

kde parametry $\Theta, \alpha_1, \alpha_2, \dots$ nezávisí na n. Pak platí

$$E\,J_k(\Theta_n) = \Theta + O(n^{-k-1}) \quad , \quad n \longrightarrow \infty . \tag{6}$$

<u>Důkaz</u> viz [10],str.527 nebo v [3] na str.54-5 .

Je důležité poznamenat, že význam předpokladu (5) nespočívá v explicitní znalosti parametrů $\Theta, \alpha_1, \alpha_2, \dots$, nýbrž v pouhé specifikaci struktury vychylujícího rozvoje odhadu Θ_n . Platnost (5) se zbytkem řady nahrazeným symbolem $O(n^{-v})$ pro některé $v>k$ lze ověřit pro většinu standardních odhadů Θ_n založených na posloupnosti nezávislých stejně rozdělených náhodných veličin (viz také [11], §6).

Zajímavá je interpretace vlastnosti (6) z hlediska explicitního vyjádření (4). Koeficienty $(-1)^v\binom{k}{v}$ pro $v=0,1,\dots,k$ charakterizují <u>k-tou zpětnou diferenci</u> běžně zaváděnou pro libovolnou posloupnost $\{x_n\}$ reálných čísel rekurentními vztahy

$$\begin{aligned}
\nabla x_n &:= x_n - x_{n-1} \\
\nabla^2 x_n &:= \nabla x_n - \nabla x_{n-1} = x_n - 2x_{n-1} + x_{n-2} \\
&\;\;\vdots \\
\nabla^k x_n &:= \nabla^{k-1} x_n - \nabla^{k-1} x_{n-1} = \sum_{v=0}^{k}(-1)^v\binom{k}{v} x_{n-v} \quad .
\end{aligned} \tag{7}$$

Není těžké ukázat, že je-li x_n nezáporná mocnina n řádu $j \geq 0$, pak $\nabla^k x_n$ je polynom stupně $j-k$ při $j<k$ identicky rovný nule; dokonce každý pól v nule typu n^{-j} pro kladné j převádí k-tá diference na pól typu $O(n^{-j-k})$. V situaci odpovídající splnění (5) definujeme

$$x_n := n^k \cdot E\,\Theta_n = \Theta \cdot n^k + \sum_{v=1}^{\infty} \alpha_v \cdot n^{k-v} \quad , \; n>k \; , \tag{8}$$

takže k-tá diference aplikovaná na x_n (tj. $\nabla^k x_n = (k!)E\left[J_k(\Theta_n)\right]$) musí být nutně výraz typu (6).

Explicitní znalost transformačního vzorce (4) umožňuje vyšetřovat otázku asymptotického rozdělení $J_k(\Theta_n)$ ve vztahu k původnímu rozdělení odhadu Θ_n . Platí-li asymptotická normalita pro n.v. $n^{1/2}(\Theta_n - \Theta)$, lze stejné normální aproximace použít obvykle i v případě n.v. $n^{1/2}(J_k(\Theta_n) - \Theta)$. Odhadem společného asymptotického rozptylu je statistika

$$S_1^2(\Theta_n) := (n-1)\sum_{i=1}^{n}(\Theta_{-i} - \bar{\Theta}_{n-1})^2 \quad . \tag{9}$$

2. S-VÝBĚROVÁ JACKKNIFEOVÁ TRANSFORMACE K-TÉHO ŘÁDU

Definici odhadu (2) lze analogicky rozšířit na situace, kdy parametr Θ je odhadován na základě $S(\geq 1)$ nezávislých náhodných výběrů s rozsahy $n_1,\dots,n_s$ ($n_1+\dots+n_s=n$). Pokud zanedbáme členy řádů $O(n^{-k-1})$, lze typickou strukturu vychylujícího rozvoje odpovídajícího odhadu Θ_n^s charakterizovat vyjádřením

$$E\,\Theta_n^s = \Theta + \sum_{r=1}^{s\wedge k}\ \sum_{1\le a_1<\dots<a_r\le s}\ \sum_{r\le v_1+\dots+v_r\le k} \alpha_{v_1\dots v_r}^{a_1\dots a_r}\, n_{a_1}^{-v_1}\dots n_{a_r}^{-v_r}\ , \tag{10}$$

kde Θ a všechna $\alpha_{\dots}^{\dots}$ jsou parametry nezávislé na $n_1,\dots,n_s$ a poslední sumace sčítá přes systém $\{[v_1,\dots,v_r];\ 1\le v_1,\dots,v_r\le k\ \text{et}\ v_1+\dots+v_r\le k\}$. V nejjednodušším případě $k=1$ lze jackknifeovou transformaci vyjádřit jako řešení soustavy lineárních rovnic definované vzorcem (viz [7], str.1013)

$$J_1^s(\Theta_n^s) := \frac{\begin{vmatrix} \Theta_n^s & \bar\Theta_1^1 & \dots & \bar\Theta_1^s \\ n_1^{-1} & (n_1-1)^{-1} & \dots & n_1^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ n_s^{-1} & n_s^{-1} & \dots & (n_s-1)^{-1} \end{vmatrix}}{\begin{vmatrix} 1 & 1 & \dots & 1 \\ n_1^{-1} & (n_1-1)^{-1} & \dots & n_1^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ n_s^{-1} & n_s^{-1} & \dots & (n_s-1)^{-1} \end{vmatrix}} = (n-s+1)\Theta_n^s - \sum_{a=1}^{s}(n_a-1)\bar\Theta_1^a\ , \tag{11}$$

kde $\bar\Theta_1^a$ pro $a=1,\dots,s$ je průměr všech n_a odhadů získaných z Θ_n^s vynecháním právě jednoho pozorování v rámci stratifikační proměnné a.

Pro obecné $k<\min\{n_1,\dots,n_s\}$ máme k dispozici systém pomocných odhadů

$$\bar\Theta_{\hat v_1\dots\hat v_{\hat r}}^{\hat a_1\dots\hat a_{\hat r}} := \binom{n_{\hat a_1}}{\hat v_1}^{-1}\dots\binom{n_{\hat a_{\hat r}}}{\hat v_{\hat r}}^{-1} \sum_{\substack{1\le i_1^{\hat a_1}<\dots<i_{\hat v_1}^{\hat a_1}\le n_{\hat a_1}\\ \vdots \\ 1\le i_1^{\hat a_{\hat r}}<\dots<i_{\hat v_{\hat r}}^{\hat a_{\hat r}}\le n_{\hat a_{\hat r}}}} \Theta_{-i_1^{\hat a_1},\dots,-i_{\hat v_1}^{\hat a_1},\dots\dots,-i_1^{\hat a_{\hat r}},\dots,-i_{\hat v_{\hat r}}^{\hat a_{\hat r}}} \tag{12}$$

při

$$1\le\hat r\le s\wedge k\ ,\quad 1\le\hat a_1<\dots<\hat a_{\hat r}\le s\ ,\quad \hat r\le\hat v_1+\dots+\hat v_{\hat r}\le k\ . \tag{13}$$

Pro libovolnou sestavu indexů z (13) zřejmě platí

$$E\,\bar\Theta_{\hat v_1\dots\hat v_{\hat r}}^{\hat a_1\dots\hat a_{\hat r}} = \Theta + \sum_{r=1}^{s\wedge k}\ \sum_{1\le a_1<\dots<a_r\le s}\ \sum_{r\le v_1+\dots+v_r\le k} \alpha_{v_1\dots v_r}^{a_1\dots a_r}\, \tilde n_{a_1}^{-v_1}\dots\tilde n_{a_r}^{-v_r}\ , \tag{14}$$

když

$$\tilde n_a \begin{cases} n_a & \text{pro } a\notin\{\hat a_1,\dots,\hat a_{\hat r}\} \\ (n_{\hat a_o}-\hat v_o) & \text{kdykoli } a \text{ je rovno některému indexu } \hat a_o \text{ z množiny } \{\hat a_1,\dots,\hat a_{\hat r}\}\ . \end{cases} \tag{15}$$

Nahradíme-li střední hodnoty na levých stranách rovnic (10) a (14) jejich přirozenými odhady Θ_n^s a $\bar\Theta_{\hat v_1\dots\hat v_{\hat r}}^{\hat a_1\dots\hat a_{\hat r}}$ a všechny teoretické parametry doplníme symbolem ^ , definuje <u>řešení pro $\hat\Theta$</u> takto vzniklé soustavy $\binom{s+k}{s\wedge k}$ lineárních rovnic o $\binom{s+k}{s\wedge k}$ neznámých odhad $J_k^s(\hat{\theta}_n^s)$ zobecňující jednovýběrový odhad ze stati 1 (speciálně $J_k(\hat{\theta}_n) = J_k^1(\hat{\theta}_n^1)$).

Jednoznačnost definice i přesnější představu o struktuře odhadu $J_k^s(\hat{\theta}_n^s)$ poskytuje následující technické tvrzení -

-<u>Lemma 2</u>: Zaveďme symbol $\underset{\sim}{\nabla}^{a_1\dots a_r}$ pro $\binom{k}{r}$ složkový vektor sestavený z prvků

$$(16)\quad \nabla^{a_1\dots a_r}_{v_1\dots v_r} := \bar{\theta}^{a_1\dots a_r}_{v_1\dots v_r} - \left[\bar{\theta}^{a_1\dots a_{r-1}}_{v_1\dots v_{r-1}} + \dots + \bar{\theta}^{a_2\dots a_r}_{v_2\dots v_r}\right] + \dots + (-1)^{r-1}\left[\bar{\theta}^{a_1}_{v_1} + \dots + \bar{\theta}^{a_r}_{v_r}\right] + (-1)^r \hat{\theta}_n^s$$

pro systém indexů korespondující s (13).

Platí

$$(17)\quad J_k^s(\hat{\theta}_n^s) = \theta_n^s + \sum_{r=1}^{s\wedge k} (-1)^r \left\{ \sum_{1\le a_1<\dots<a_r\le s} \hat{\alpha}^*_{(r)}(\underset{\sim}{\nabla}^{a_1\dots a_r}) \right\} ,$$

kde $\hat{\alpha}^*_{(r)}(\underset{\sim}{\nabla}^{1\dots r})$ je jediné řešení pro $\hat{\alpha}^{1\dots r}_{1\dots 1}$ soustavy lineárních rovnic

$$(18)\quad \frac{\nabla^{1\dots r}_{\hat{v}_1,\dots,\hat{v}_r}}{\hat{v}_1\dots\hat{v}_r} = \sum_{r\le v_1+\dots+v_r\le k} \hat{\alpha}^{1\dots r}_{v_1\dots v_r}\cdot(n_1-\hat{v}_1)^{-v_1}\dots(n_r-\hat{v}_r)^{-v_r} , \quad \begin{matrix}1\le r\le s\wedge k\\ r\le\hat{v}_1+\dots+\hat{v}_r\le k\end{matrix} .$$

<u>Důkaz</u> je podrobně rozepsán v [3] .

Při $k<\min\{n_1,\dots,n_s\}$ lze ověřit jednoznačnost řešení každé soustavy typu (18) a za předpokladu (10) i požadovanou vlastnost

$$(19)\quad E\, J_k^s(\theta_n^s) = \theta .$$

Explicitní vyjádření (17) pro k=1,2,3 udávají vzorce (11) a

$$(20)\quad J_2^s(\theta_n^s) = \theta_n^s - \left\{\sum_{a=1}^{s}\left[(n_a-1)^2\nabla_1^a - \tfrac{1}{2}(n_a-2)^2\nabla_2^a\right]\right\} + \left\{\sum_{1\le a_1<a_2\le s}\left[(n_{a_1}-1)(n_{a_2}-2)\nabla_{11}^{a_1a_2}\right]\right\} ,$$

$$(21)\quad J_3^s(\theta_n^s) = \theta_n^s - \left\{\tfrac{1}{6}\sum_{a=1}^{s}\left[3(n_a-1)^3\nabla_1^a - 3(n_a-2)^3\nabla_2^a + (n_a-3)^3\nabla_3^a\right]\right\} +$$
$$+\left\{\tfrac{1}{2}\sum_{1\le a_1<a_2\le s}\left[(n_{a_1}-2)^2(n_{a_2}-1)\nabla_{21}^{a_1a_2} - 2(n_{a_1}-1)(n_{a_2}-1)\begin{vmatrix}(n_{a_1}-1) & (n_{a_2}-2)\\ (n_{a_2}-2) & (n_{a_1}-1)\end{vmatrix}\nabla_{11}^{a_1a_2} +\right.\right.$$
$$\left.\left. + (n_{a_1}-1)(n_{a_2}-2)^2\nabla_{12}^{a_1a_2}\right]\right\} - \left\{\sum_{1\le a_1<a_2<a_3\le s}(n_{a_1}-1)(n_{a_2}-1)(n_{a_3}-1)\nabla_{111}^{a_1a_2a_3}\right\} ,$$

kde každá sumace přes systém $1\le a_1<\dots<a_r\le s$ pro $r>s$ je definována jako nula.

<u>3. APLIKACE NA KOEFICIENTY ASOCIACE</u>

Uvažujme koeficienty asociace typu $\delta(\underset{\sim}{f}_n^s)$, kde $\delta(\underset{\sim}{.}): \mathbb{R}^m \longrightarrow \mathbb{R}$ je funkce m proměnných omezená na svém definičním oboru a $\underset{\sim}{f}_n^s := (\underset{\sim}{f}_{n_1}^1,\dots,\underset{\sim}{f}_{n_s}^s)$ sestává z s nezávislých vektorů odpovídajících po řadě c-rozměrným multinomickým rozdělením $M(n_1,\underset{\sim}{p}^1), \dots ,M(n_s,\underset{\sim}{p}^s)$. Případ s=1 pokrývá jednovýběrovou situaci z příspěvku [2] , při s>1 odpovídají nezávislé výběry přirozené stratifikaci vzhledem k hodnotám některé proměnné kontingenční tabulky. Korektní aplikace jackknifeové transformace $J_k^s(.)$ ze stati 2 požaduje pro $\theta_n^s = \delta(\underset{\sim}{f}_n^s)$ strukturu vychylujícího rozvoje typu (10). Za předpokladu dostatečné hladkosti funkce $\delta(.)$ na okolí teoretického bodu $\underset{\sim}{p} := (\underset{\sim}{p}^1,\dots,\underset{\sim}{p}^s)$ lze tuto vlastnost odvodit ze standardní aproximace Taylorovým rozvojem vyjádřené ve tvaru

$$(22)\quad \left[E\,\delta(\underset{\sim}{f}_n^s) - \delta(\underset{\sim}{p})\right] = \sum_{j=1}^{2k}\frac{1}{j!}\underset{\sim}{D}_j\cdot\left[E(d\underset{\sim}{f}_n^s)^j\right] + o(n^{-k}) , \quad n\longrightarrow\infty ,$$

kde $\underset{\sim}{D}_j$ je matice j-tých parciálních derivací $\delta(\underset{\sim}{.})$ v bodě $\underset{\sim}{p}$ blokově strukturovaná jakožto $\underset{\sim}{D}_j := \left[\underset{\sim}{D}_j^{a_1\dots a_j}\right]_{1\le a_1,\dots,a_j\le s}$, $\left[E(d\underset{\sim}{f}_n^s)^j\right]$ symbolizuje matici j-tých centrálních momentů m=(s.c)-složkového vektoru $(d\underset{\sim}{f}_n^s) := (\underset{\sim}{f}_n^s - \underset{\sim}{p})$ a tečka značí skalární součin.

Elementární rozpis každého součinu z (22) pro matice $\underset{\sim}{L}_j$ symetrické vzhledem k permutacím indexů

(23) $$\underset{\sim}{D}_j \cdot [E(d\underset{\sim}{f}_n^s)^j] = \sum_{r=1}^{s} \sum_{1 \le a_1 < \ldots < a_r \le s} \sum_{\substack{1 \le j_1, \ldots, j_r \le j \\ j_1 + \ldots + j_r = j}} \frac{j!}{j_1! \ldots j_r!} \underset{\sim}{D}_j^{a_1 \ldots a_r} \cdot [E(d\underset{\sim}{f}_{n_{a_1}}^{a_1})^{j_1} \times \ldots \times E(d\underset{\sim}{f}_{n_{a_r}}^{a_r})^{j_r}]$$

umožňuje využít explicitních vyjádření (7) resp.(8) z článku [2] pro výpočet každého dílčího součinu z (23) typu $\underset{\sim}{D}_j^{\cdot a \cdot} \cdot [\cdot \times E(d\underset{\sim}{f}_{n_a}^a)^j \times \cdot]$. S ohledem na (22) a (23) lze snadno nahlédnout explicitní závislost každého koeficientu $\alpha_{v_1 \ldots v_r}^{a_1 \ldots a_r}$ z (10) pouze na parametrech $\underset{\sim}{p}^{a_1}, \ldots, \underset{\sim}{p}^{a_r}$ a dané sestavě $v_1, \ldots, v_r$. Speciálně při vynechání všech nulových sčítanců v (23) -odpovídajících situaci $j_a = 1$ pro některé $a = 1, \ldots, r$ - dostaneme vyjádření koeficientů u řádů n_a^{-1}, n_a^{-2} a $n_{a_1}^{-1} n_{a_2}^{-1}$ tvaru

(24) $$\begin{aligned} \alpha_1^a &= \tfrac{1}{2} \underset{\sim}{D}_2^a \cdot [E(d\underset{\sim}{e}^a)^2] , && \\ \alpha_2^a &= \tfrac{1}{6} \underset{\sim}{L}_3^a \cdot [E(d\underset{\sim}{e}^a)^3] + \tfrac{1}{8} \underset{\sim}{D}_4^a \cdot [E(d\underset{\sim}{e}^a)^2 \times E(d\underset{\sim}{e}^a)^2] , && a = 1, \ldots, s , \\ \alpha_{11}^{a_1 a_2} &= \tfrac{1}{4} \underset{\sim}{D}_4^{a_1 a_2} \cdot [E(d\underset{\sim}{e}^{a_1})^2 \times E(d\underset{\sim}{e}^{a_2})^2] , && 1 \le a_1 < a_2 \le s , \end{aligned}$$

kde $\underset{\sim}{e}^a, \underset{\sim}{e}^{a_1}, \underset{\sim}{e}^{a_2}$ jsou nezávislé multinomické vektory s rozdělením $M(1, \underset{\sim}{p}^a), M(1, \underset{\sim}{p}^{a_1})$ a $M(1, \underset{\sim}{p}^{a_2})$. Analogické výpočtové vzorce se komplikují s přechodem k vyšším řádům $O(n^{-k})$ a k nesymetrickým maticím $\underset{\sim}{L}_j$.

Naznačené možnosti teoretického výpočtu vychylujících parametrů pro každý zvolený koeficient $\delta(\underset{\sim}{f}_n^s)$ a jejich následné využití pro účely eliminace vychýlení jsou ovšem nivelizovány při aplikaci transformací jackknife, které mají univerzální schopnost eliminace těchto členů bez ohledu na analytický předpis $\delta(\underset{\sim}{.})$. Výpočtové vzorce (12) pro $\hat{\theta}_{v_1 \ldots v_r}^{a_1 \ldots a_r} := \delta(\underset{\sim}{f}_{v_1 \ldots v_r}^{a_1 \ldots a_r})$ budou mít sice složitější strukturu typu

(25) $$\delta(\underset{\sim}{f}_{v_1 \ldots v_r}^{1 \ldots r}) =$$

$$= \binom{n_1}{v_1}^{-1} \cdots \binom{n_r}{v_r}^{-1} \sum_{\substack{0 \le m_1^1, \ldots, m_c^1 \le v_1 \\ m_1^1 + \ldots + m_c^1 = v_1 \\ \vdots \\ 0 \le m_1^r, \ldots, m_c^r \le v_r \\ m_1^r + \ldots + m_c^r = v_r}} \left\{ \binom{n_1^1}{m_1^1} \cdots \binom{n_c^1}{m_c^1} \cdots \binom{n_1^r}{m_1^r} \cdots \binom{n_c^r}{m_c^r} \delta(\underset{\sim}{f}_{-(m_1^1 \ldots m_c^1), \ldots, -(m_1^r \ldots m_c^r)}^{1 \cdot \ldots \cdot r}) \right\} ,$$

kde sumace v (25) sčítá přes počty $m_1^1, \ldots, m_c^1, \ldots, m_1^r, \ldots, m_c^r$ všech sestav indexů vyhazovaných po řadě z 1. až c-té třídy 1. až r-té stratifikační proměnné a $n_1^1, \ldots, n_c^1, \ldots, n_1^r, \ldots, n_c^r$ jsou odpovídající absolutní četnosti (pokud $n_b^a < m_b^a$ pro některou dvojici $[a,b] \in \{1, \ldots, r\} \times \{1, \ldots, c\}$, je celý výraz ve složené závorce definován jako nula) , ale jsou formálně stejné pro libovolný koeficient $\delta(\underset{\sim}{f}_n^s)$.

4. ZÁVĚRY ZE SIMULAČNÍHO EXPERIMENTU

Statistické chování jackknifeových transformací bylo zkoumáno na jednovýběrové $J_k(.)$ pro $k = 1, \ldots, 10$ aplikované na korelačním poměru η^2 (viz [9],str.251) počítaném z kontingenční tabulky 4x5 a na logaritmu součinového poměru v tabulce 2x2 , který lze chápat jako speciální případ logaritmické interakce z [4]. V obou případech lze teoreticky zdůvodnit volbu řádu k jackknifeové transformace odpovídající faktické numerické významnosti několika prvních členů v (5) pro dostatečně velké n (kupř. $n \ge 100$).

V případě korelačního poměru η^2 je možno ověřit nárůst významnosti teoretických koeficientů z (24) při snižující se hodnotě některé z marginálních řádkových (resp. stratifikačních) pravděpodobností - z teoretických i simulačních výsledků však vyplývá pro všechny běžné volby teoretických parametrů (a rozsahů n) potřeba použití $J_k(.)$ (resp. $J_k^s(.)$) pro k nejvýše rovné dvěma. Méně jednoznačná je volba "správného řádu k" v situaci logaritmické interakce, jejíž teoretickou hodnotu pro jednovýběrový případ lze definovat vztahem (viz [1], str.101)

$$\delta_{\gamma}(\underset{\sim}{p}) := \sum_{m=1}^{c} \gamma_m \cdot \log p_m \quad , \tag{26}$$

kde $\underset{\sim}{p} := (p_1, \ldots, p_c)$ je pravděpodobnostní rozdělení obecně vícerozměrné kontingenční tabulky s nenulovými složkami a $\underset{\sim}{\gamma} := (\gamma_1, \ldots, \gamma_c)$ nenulový vektor konstant vyhovující podmínce $\gamma_1 + \ldots + \gamma_c = 0$. Problémy s definicí (26) na hranici jednotkového simplexu se v případě výběrového koeficientu $\delta_{\gamma}(\underset{\sim}{f}_n^1)$ kompenzují připočtením jisté korekční konstanty ke všem absolutním četnostem analyzované tabulky. Dá se ukázat, že tato úprava nic nezmění na platnosti asymptotického vztahu

$$\begin{aligned} E\,\delta_{\gamma}(\underset{\sim}{f}_n^1) = \delta_{\gamma}(\underset{\sim}{p}) &- n^{-1}\left\{\tfrac{1}{2}\Gamma_{-1}\right\} + n^{-2}\left\{\tfrac{1}{2}\Gamma_{-1} - \tfrac{5}{12}\Gamma_{-2}\right\} - n^{-3}\left\{\tfrac{1}{2}\Gamma_{-1} - 1\tfrac{1}{4}\Gamma_{-2} + \tfrac{3}{4}\Gamma_{-3}\right\} + \\ &+ n^{-4}\left\{\tfrac{1}{2}\Gamma_{-1} - 2\tfrac{11}{12}\Gamma_{-2} + 4\tfrac{1}{2}\Gamma_{-3} - 2\tfrac{11}{120}\Gamma_{-4}\right\} - \\ &- n^{-5}\left\{\tfrac{1}{2}\Gamma_{-1} - 6\tfrac{1}{4}\Gamma_{-2} + 18\tfrac{3}{4}\Gamma_{-3} - 20\tfrac{11}{12}\Gamma_{-4} + 7\tfrac{11}{12}\Gamma_{-5}\right\} + \\ &+ n^{-6}\left\{\tfrac{1}{2}\Gamma_{-1} - 12\tfrac{11}{12}\Gamma_{-2} + 67\tfrac{1}{2}\Gamma_{-3} - 135\tfrac{23}{24}\Gamma_{-4} + 118\tfrac{3}{4}\Gamma_{-5} - 37\tfrac{439}{504}\Gamma_{-6}\right\} + o(n^{-6}) \,, n \to \infty , \end{aligned} \tag{27}$$

pro

$$\Gamma_{-v} := \sum_{m=1}^{c} (\gamma_m)(p_m)^{-v} \quad , \quad v = 1, \ldots, 6 \quad , \tag{28}$$

který lze odvodit z aproximací typu (22) při $k=6$ s ohledem na jednoduchou (diagonální) strukturu všech matic parciálních derivací a na znalost centrálních momentů elementárního multinomického rozdělení (viz [3], str.35-7). Paktická významnost koeficientů stojících v (27) u jednotlivých řádů n^{-v} tedy bezprostředně závisí na rychlosti konvergence hodnoty $1/(n \cdot \min\{p_1, \ldots, p_c\})$ k nule. Pro dosti malá n ($n \leq 50$) a konkrétní volby vektorů $\underset{\sim}{\gamma}$ a $\underset{\sim}{p}$ je tudíž oprávněno použití jackknifeových transformací vyššího řádu k.

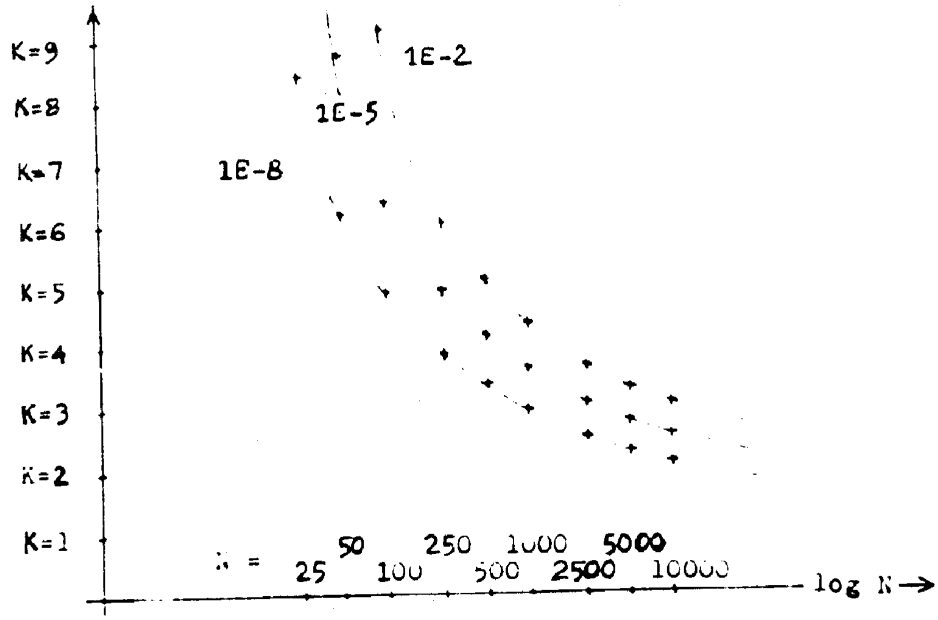
Formální simulační aplikace $J_k(.)$ na oba uvažované koeficienty přinesla tyto zobecněné závěry:

i/ Obecný vzorec (4) nelze použít pro neomezeně velká n a k, neboť sčítání a odčítání příslušných členů typu $O(n^k)$ klade zvýšené nároky na numerickou přesnost použité výpočetní techniky. Z tohoto důvodu je nutné počítat hodnoty koeficientů asociace, na něž se aplikuje transformace $J_k(.)$ při $k>2$, s vysokou přesností (double precision). Volba maximálního řádu k v závislosti na rozsahu n a zvolené přesnosti zaokrouhlovací chyby vyplývá z empirického grafu na obr.1 .

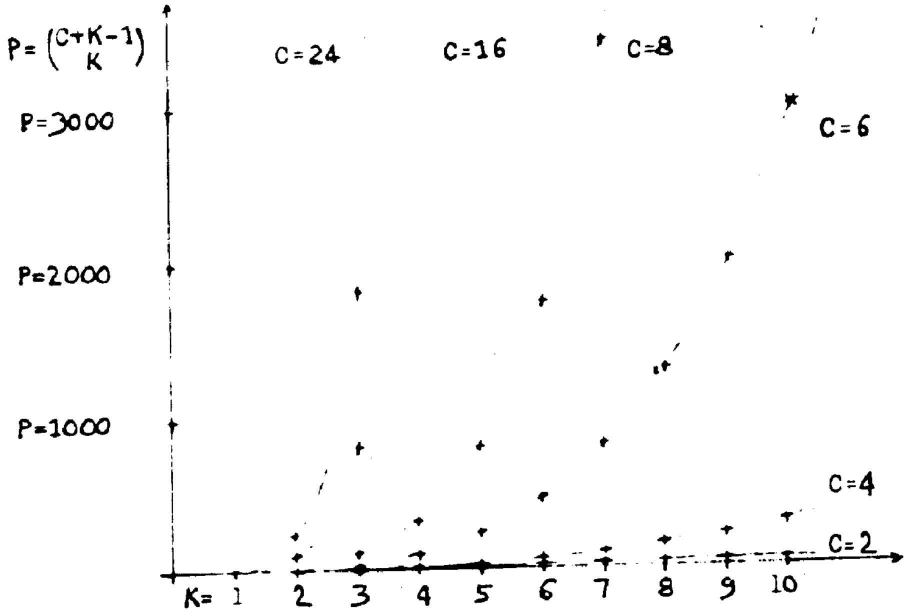
ii/ Počet $\binom{n}{k}$ vynechávaných pozorování pro účely následného přepočítávání hodnot původního odhadu podle (1) lze v případě koeficientů asociace úspěšně redukovat na počet $\binom{c+k-1}{k}$ pro c kategoriálních tříd. Jak ovšem vyplývá z grafu na obr.2, vzrůstá při zvyšujícím se c a k i tento počet nad časově zvládnutelnou hranici.

iii/ Pokud zvolený koeficient $\delta(\underset{\sim}{p})$ není definován na celém jednotkovém simplexu, může dojít k situaci, že od jistého k není (na rozdíl od empirické hodnoty $\delta(\underset{\sim}{f}_n^1)$) definován odhad $J_k(\delta(\underset{\sim}{f}_n^1))$. Bodové vlastnosti jackknifeových odhadů jsou potom při malých n velmi citlivé na případná předefinování-korekce jako za vzorcem (26).

iv/ Respektujeme-li omezení vyplývající z i/-iii/, vykazují všechny numericky dostupné odhady značnou stabilitu projevující se vysokou párovou korelací mezi odhady $J_1(\delta(\underset{\sim}{f}_n^1)), \ldots, J_k(\delta(\underset{\sim}{f}_n^1))$, z nichž všechny se statisticky významně liší od původní-vychýlené hodnoty $\delta(\underset{\sim}{f}_n^1)$. Pro malé rozsahy n je ovšem třeba mít na paměti, že rozptyl jackknifeových odhadů je obvykle téhož řádu jako původní odhad, takže jednotlivé pozorované hodnoty -přestože jsou fakticky nevychýlené- mohou být značně variabilní!!



Obr.1: Závislost maximálního výpočetního řádu K jackknifeové transformace $J_K(.)$ na rozsahu N a zvolené přesnosti E zaokrouhlovací chyby.



Obr.2: Závislost počtu P snovu přepočítávaných hodnot koeficientu asociace na řádu K jackknifeové transformace $J_K(.)$ a počtu C kategoriálních tříd.

LITERATURA:

[1] Anděl J.(1973): Interaction in contingency taoles. Aplikace matematiky 18, str.99-109

[2] Běláček J.(1986): Jackknifeování koeficientů asociace. Sborník ROBUST 86, str.21-6

[3] Běláček J.(1988): Eliminace vychýlení obecných měr asociace. Kand.práce MFF UK, Praha

[4] Goodman L.(1964): Interactions in multidimensional contingency tables. Ann.Math.Statist.35, str.632-46

[5] Hais K.-Hodek B.(1984): Velký anglicko-český slovník. Academia, Praha

[6] Hrala P.(1987) - osobní sdělení

[7] Krewski L.-Rao J.N.(1981): Inference from stratified samples - properties of the linearization, jackknife and balanced repeated replication methods. Ann.Statist.5, str.1010-19

[8] Parr W.C.-Tolley H.D.(1982): Jackknifing in categorial data analysis. Austral.J.Statist.24, str.67-79

[9] Řehák J.-Řeháková B.(1986): Analýza kategorizovaných dat v sociologii. Academia, Praha

[10] Schucany W.R.-Gray H.L.-Owen L.B.(1971): On bias reduction in estimation. JASA 66, str.524-33

[11] Serfling R.J.(1980): Approximation theorems of mathematical statistics. J.Willey and sons, New York-Chichester-Brisbane-Toronto, str.210-42

[12] Smith E.P.-van Belle G.(1984): Nonparametric estimation of specied richnes. Biometrics 40, str.119-29