JACK-KNIFEOVÁNÍ KOEFICIENTŮ ASOCIACE

Jaromír Běláček, MFF UK

Výsledky presentované v tomto příspěvku jsou aplikovatelné ve všech situacích, kde máme co činit s omezenými dostatečně hladkými funkcemi (normovaného) multinomického vektoru, pořízeného nezávislým výběrem reprezentantů populace klasifikované do m-tříd. Takovýmito funkcemi jsou například rozličné koeficienty asociace vyjadřující míru jisté specifické statistické závislosti mezi zkoumanými znaky kontingenční tabulky, četné míry vnitřní variability (diversity) souboru apod. Příspěvek pojednává o teoretickém srovnání základních statistických vlastností výše zmíněných empirických veličin s odhady, které lze z těchto získat užitím elementárních metod typu jack-knife, zejména s ohledem na redukci vychýlení původního odhadu a na jack-knifeový odhad asymptotického rozptylu. V závěru je zdůrazněna především jednoduchost a praktická aplikovatelnost obou základních jack-knifeových odhadů představených v textu.

1. MATEMATICKÝ APARÁT

V celém článku budeme vycházet z multinomického rozdělení $M(n,\mathbf{p})$ m-složkového vektoru $n\cdot\mathbf{f}_n$ ($\mathbf{p} \equiv (p_1,\dots,p_m)^T$ pro $p_1,\dots,p_m \geq 0$, $\sum p_j = 1$ a $\mathbf{f}_n \equiv (f_{n1},\dots,f_{nm})^T$ je vektor relativních četností - frekvenční vektor). Předmětem našeho zkoumání je náhodná veličina $\delta(\mathbf{f}_n)$, kde $\delta(\cdot): \mathbb{R}^m \to \mathbb{R}$ je funkce m-proměnných omezená na svém definičním oboru.

Je-li $\delta(\cdot)$ spojitá v bodě $\mathbf{p}$, je $\delta(\mathbf{f}_n)$ zřejmě konzistentním odhadem (při $n\to\infty$) parametru $\delta(\mathbf{p})$ ($\delta(\mathbf{p})$ je teoretická hodnota koeficientu na celé populaci). Za silnějších předpokladů na diferencovatelnost funkce $\delta(\cdot)$ lze hlubší statistické vlastnosti koeficientu $\delta(\mathbf{f}_n)$ indikovat z Taylorova rozvoje funkce $\delta(\cdot)$ v blízkosti bodu $\mathbf{p}$. Z hlediska výpočtů i teoretických tvrzení je výhodné vyjádření v tvaru (viz. [10], str. 173):

$$(1)\qquad \delta(\mathbf{t}) = \delta(\mathbf{p}) + \sum_{k=1}^{s} \frac{1}{k!}\, \mathbf{D}_k\cdot(d\mathbf{t})^k + \mathbf{Z}_s(\mathbf{t})\;(d\mathbf{t})^s, \qquad \mathbf{t}\in\mathbb{R}^m, \quad \mathbf{t}\to\mathbf{p},$$

kde $\mathbf{D}_k$ je k-indexová matice všech k-tých parciálních derivací funkce $\delta(\cdot)$ v bodě $\mathbf{p}$ (každý z k indexů probíhá od 1 do m), $(\mathbf{y})^k \equiv (\mathbf{y}\times\mathbf{y}\times\dots\times\mathbf{y})$ pro $\mathbf{y}\in\mathbb{R}^m$ označuje tenzorový součin k vektorů $\mathbf{y}$ (což je opět k-indexová matice obsahující všechny smíšené součiny délky k všech m souřadnic vektoru $\mathbf{y}$), $d\mathbf{t}$ zastupuje rozdíl (diferenci) $(\mathbf{t}-\mathbf{p})$ a symbol $\mathbf{D}_k\cdot(d\mathbf{t})^k$ je skalární součin dvou k-indexových matic (definovaný jako součet součinů všech navzájem si odpovídajících indexových složek). Zbytkový člen v (1) je vyjádřen jako skalární součin dvou s-indexových matic, $\mathbf{Z}_s(\mathbf{t})$ a $(d\mathbf{t})^s$, kde $\mathbf{Z}_s(\cdot): \mathbb{R}^m \to (\mathbb{R}^m\times\dots\times\mathbb{R}^m)$ je m-složková vektorová funkce mající na okolí bodu $\mathbf{p}$ vlastnost

$$(2)\qquad \lim_{\mathbf{t}\to\mathbf{p}} \mathbf{Z}_s(\mathbf{t}) = \mathbf{Z}_s(\mathbf{p}) = \mathbf{0}_s$$

($\mathbf{0}_s$ symbolisuje s-indexovou matici sestavenou ze samých nul).

Spojitost zbytkového členu v bodě $\mathbf{p}$ umožňuje využít rozvoje (1) pro stanovení (nedegenerovaného) asymptotického rozdělení n.v. $n^{s/2}(\delta(\mathbf{f}_n) - \delta(\mathbf{p}))$ (pro vhodně volené $s\in\mathbb{N}$) bez zbytečně silných předpokladů na stupeň diferencovatelnosti funkce $\delta(\cdot)$.

Obdobně lze vlastnost (2) využít při aproximaci centrovaných momentů veličiny $\delta(\mathbf{f}_n)$ se zbytkem ve tvaru $o(\cdot)$. Na úrovni tohoto článku se omezíme na asymptotická rozdělení při $s = 1$ a aproximace vychýlení a středních čtvercových odchylek veličiny $\delta(\mathbf{f}_n)$ a některých dalších odvozených statistik.

2. ASYMPTOTICKÁ NORMALITA A APROXIMACE ZÁKLADNÍCH MOMENTOVÝCH CHARAKTERISTIK KLASICKÝCH ODHADŮ

Za předpokladu existence 1. totálního diferenciálu funkce $\delta(\cdot)$ v bodě $\mathbf{p}$ vyplývá asymptotické rozdělení veličiny $n^{1/2}(\delta(\mathbf{f}_n) - \delta(\mathbf{p}))$ z asymptotické normality rozdělení $M(n,\mathbf{p})$. Platí (viz např. [9], str. 430 v kombinaci s [1], str. 198), že

$$n^{1/2}(\delta(\mathbf{f}_n) - \delta(\mathbf{p})) \xrightarrow{\mathcal{D}} N(0, \sigma^2(\mathbf{p})), \tag{3}$$

kde

$$\sigma^2(\mathbf{p}) = \mathbf{D}_1^T \;[\;\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T]\mathbf{D}_1 \tag{4}$$

pro diagonální matici $\mathrm{diag}(\mathbf{p})$ s diagonálními prvky $p_1,\dots,p_m$. (Asymptotické rozdělení je nedegenerované, pokud $\sigma^2(\mathbf{p}) \neq 0$.) Předpokládáme-li existenci vyšších diferenciálů funkce $\delta(\cdot)$, vzniká možnost explicitních aproximací prvních dvou centrovaných momentů n.v. $\delta(\mathbf{f}_n)$ prostřednictvím vzorců (viz [5], str. 233-238, resp. [4], str. 452-4):

$$[E\delta(\mathbf{f}_n) - \delta(\mathbf{p})] = \sum_{k=1}^{s} \frac{1}{k!}\, \mathbf{D}_k \cdot [E(d\mathbf{f}_n)^k] + o(n^{-s/2}), \qquad n \to \infty \tag{5}$$

$$E[\delta(\mathbf{f}_n) - \delta(\mathbf{p})]^2 = \sum_{\substack{j=1 \\ j+k\leq s}}^{s-1} \sum_{k=1}^{s-1} \frac{1}{j!}\frac{1}{k!}\, \mathbf{D}_{j\times k} \cdot [E(d\mathbf{f}_n)^{j+k}] + o(n^{-s/2}), \quad n \to \infty, \tag{6}$$

kde $E(d\mathbf{f}_n)^k$ značí k-indexové matice sestavené ze středních hodnot prvků matic $(d\mathbf{f}_n)^k$ a $\mathbf{D}_{j\times k}$ jsou $(j+k)$-indexové tensorové součiny matic $(\mathbf{D}_j \times \mathbf{D}_k)$ (obsahující všechny vzájemné součiny prvků matic $\mathbf{D}_j$ a $\mathbf{D}_k$).

Pro frekvenční (multinomický) vektor $\mathbf{f}_n$ lze explicitně vyjádřit libovolný člen stojící na pravé straně v (5) a (6). Zavedeme-li označení $\mathbf{D}_2^2$ pro jednoindexovou matici diagonálních prvků matice D_2, D_3^3 pro jednoindexovou matici sestavenou z prvků matice D_3, jejíž všechny tři indexy jsou si rovny, D_3^{21} pro dvouindexovou matici sestavenou z prvků matice D_3, jejíž první dva indexy jsou si rovny atd., pak pro aproximace ad(5) do řádu $O(n^{-2})$ můžeme využít vzorců

$$\begin{aligned}
&\text{i)} && \mathbf{D}_1 \cdot [E(d\mathbf{f}_n)] = 0 \\
&\text{ii)} && \mathbf{D}_2 \cdot [E(d\mathbf{f}_n)^2] = n^{-1}(\mathbf{D}_2^2 \cdot (\mathbf{p}) - \mathbf{D}_2 \cdot (\mathbf{p})^2) \\
&\text{iii)} && \mathbf{D}_3 \cdot [E(d\mathbf{f}_n)^3] = n^{-2}(\mathbf{D}_3^3 \cdot (\mathbf{p}) - 3\mathbf{D}_3^{21} \cdot (\mathbf{p})^2 + 2\mathbf{D}_3 \cdot (\mathbf{p})^3) \\
&\text{iv)} && \mathbf{D}_4 \cdot [E(d\mathbf{f}_n)^4] = 3[(n-1)/n^3](\mathbf{D}_4^{22} \cdot (\mathbf{p})^2 - 2\mathbf{D}_4^{211} \cdot (\mathbf{p})^3 + \mathbf{D}_4 \cdot (\mathbf{p})^4) \; + \\
&&& \qquad + n^{-3}(\mathbf{D}_4^4 \cdot (\mathbf{p}) - 4\mathbf{D}_4^{31} \cdot (\mathbf{p})^2 + 6\mathbf{D}_4^{211} \cdot (\mathbf{p})^3 - 3\mathbf{D}_4 \cdot (\mathbf{p})^4 \;,
\end{aligned} \tag{7}$$

které lze odvodit například na základě znalosti smíšených momentů multinomického rozdělení $M(n,\underset{\sim}{p})$. Obdobně lze ověřit platnost identit

$$\text{i)}\quad \underset{\sim}{D}_{1\times 1}\cdot\left[E(d\underset{\sim}{f}_n)^2\right] = n^{-1}(\underset{\sim}{D}^2_{1\times 1}\cdot(\underset{\sim}{p}) - \underset{\sim}{D}_{1\times 1}\cdot(\underset{\sim}{p})^2 = n^{-1}\sigma^2(\underset{\sim}{p})$$

$$(8)\qquad \text{ii)}\quad \underset{\sim}{D}_{1\times 2}\cdot\left[E(d\underset{\sim}{f}_n)^3\right] = n^{-2}(\underset{\sim}{D}^3_{1\times 2}\cdot(\underset{\sim}{p}) - \underset{\sim}{D}^{12}_{1\times 2}\cdot(\underset{\sim}{p})^2 - 2\underset{\sim}{D}^{21}_{1\times 2}\cdot(\underset{\sim}{p})^2 + 2\underset{\sim}{D}_{1\times 2}\cdot(\underset{\sim}{p})^3)$$

$$\begin{aligned}\text{iii)}\quad \underset{\sim}{D}_{1\times 3}\cdot\left[E(d\underset{\sim}{f}_n)^4\right] &= 3\left[(n-1)/n^3\right](\underset{\sim}{D}^{22}_{1\times 3}\cdot(\underset{\sim}{p})^2 - \underset{\sim}{D}^{211}_{1\times 3}\cdot(\underset{\sim}{p})^3 - \underset{\sim}{D}^{112}_{1\times 3}\cdot(\underset{\sim}{p})^3 + \underset{\sim}{D}_{1\times 3}\cdot(\underset{\sim}{p})^4) + \\ &+ n^{-3}(\underset{\sim}{D}^4_{1\times 3}\cdot(\underset{\sim}{p}) - 3\underset{\sim}{D}^{31}_{1\times 3}\cdot(\underset{\sim}{p})^2 - \underset{\sim}{D}^{13}_{1\times 3}\cdot(\underset{\sim}{p})^2 + 3\underset{\sim}{D}^{211}_{1\times 3}\cdot(\underset{\sim}{p})^3 + \\ &+ 3\underset{\sim}{D}^{121}_{1\times 3}\cdot(\underset{\sim}{p})^3 - 3\underset{\sim}{D}_{1\times 3}\cdot(\underset{\sim}{p})^4)\end{aligned}$$

použitelných v aproximacích ad(6). (Za předpokladu existence s-tého diferenciálu $\delta(\cdot)$ na otevřeném okolí bodu $\underset{\sim}{p}$ jsou matice $\underset{\sim}{D}_k$, $k=1,\dots,s$, symetrické vzhledem k záměně svých indexů. Vzorce 7iii)-iv) jsou speciální případy 8ii)-iii) pro symetrické matice.)

V praktických případech bývá vzorec (4) využíván pro odhad $\sigma^2(\underset{\sim}{f}_n)$ as.rozptylu $\sigma^2(\underset{\sim}{p})$ (do předpisu (4) dosazujeme $\underset{\sim}{f}_n$ na místo $\underset{\sim}{p}$), na němž jsou často založeny intervaly spolehlivosti pro parametr $\delta(\underset{\sim}{p})$ využívající kritických hodnot normovaného normálního rozdělení. Užití aproximace ad(3) pro testy hypotéz o skutečné hodnotě parametru $\delta(\underset{\sim}{p})$ však nebere na zřetel skutečnost, že empirické koeficienty $\delta(\underset{\sim}{f}_n)$ mohou být často odhady vychýlené. V takové situaci můžeme původní koeficient $\delta(\underset{\sim}{f}_n)$ opravit tak, že od něj odečteme odhad vychýlení, získaný dosazením $\underset{\sim}{f}_n$ na místo $\underset{\sim}{p}$ v předpisu (5) (upraveném podle (7)).

Nejjednodušším odhadem konstruovaným takto pro redukci vychýlení řádu $O(n^{-1})$ bude zřejmě statistika

$$(9)\qquad \bar{\delta}(\underset{\sim}{f}_n) \overset{\text{df}}{=} \delta(\underset{\sim}{f}_n) - (2n^{-1})(\bar{\underset{\sim}{D}}^2_2\cdot(\underset{\sim}{f}_n) - \bar{\underset{\sim}{D}}_2\cdot(\underset{\sim}{f}_n)^2)\ ,$$

kde $\bar{\underset{\sim}{D}}_2$ je matice parciálních derivací funkce $\delta(\cdot)$ v bodě $\underset{\sim}{f}_n$. S použitím vzorců (5) a (6) aplikovaných na koeficient $\bar{\delta}(\underset{\sim}{f}_n)$ lze ověřit platnost vztahů

$$(10)\qquad \left[E\,\bar{\delta}(\underset{\sim}{f}_n) - \delta(\underset{\sim}{p})\right] = (2n)^{-1}\,\underset{\sim}{D}_2\cdot\left[E(d\underset{\sim}{f}_n)^2\right] - \tfrac{1}{3}\,\underset{\sim}{D}_3\cdot\left[E(d\underset{\sim}{f}_n)^3\right] - \tfrac{1}{24}\,\underset{\sim}{D}_4\cdot\left[E(d\underset{\sim}{f}_n)^4\right] + o(n^{-2}),\ n\to\infty.$$

$$\begin{aligned}(11)\qquad E\left[\bar{\delta}(\underset{\sim}{f}_n) - \delta(\underset{\sim}{p})\right]^2 &= \underset{\sim}{D}_{1\times 1}\cdot\left[E(d\underset{\sim}{f}_n)^2\right] + \tfrac{1}{4}\,\underset{\sim}{D}_{2\times 2}\cdot\left[E(d\underset{\sim}{f}_n)^4\right] - \tfrac{1}{4}\,\underset{\sim}{D}_{2\times 2}\cdot\left[E(d\underset{\sim}{f}_n)^2\right]^2 + o(n^{-2}) = \\ &= n^{-1}\sigma^2(\underset{\sim}{p}) + \operatorname{Var}\left[\tfrac{1}{2}\,\underset{\sim}{D}_2\cdot(d\underset{\sim}{f}_n)^2\right] + o(n^{-2}), \qquad n\to\infty\ .\end{aligned}$$

Protože vychýlení koeficientu (9) je zřejmě řádu $O(n^{-2})$, je aproximace (11) vhodná taktéž pro aproximaci rozptylu $\operatorname{Var}\left[\bar{\delta}(\underset{\sim}{f}_n) - \delta(\underset{\sim}{p})\right] = \operatorname{Var}\left[\bar{\delta}(\underset{\sim}{f}_n)\right]$. Pro účely statistické inference však obvykle vystačíme s členem $n^{-1}\sigma^2(\underset{\sim}{p})$ (resp. s jeho konzistentním odhadem), neboť vzhledem k předpisu (9) je asymptotické rozdělení veličiny $n^{1/2}(\bar{\delta}(\underset{\sim}{f}_n) - \delta(\underset{\sim}{p}))$ zřejmě opět $N(0,\sigma^2(\underset{\sim}{p}))$.

Teoretických aproximací (5), (6), (10) a (11) lze s výhodou použít pro srovnání s koeficienty typu jack-knife, u nichž je docíleno analogických efektů jako u koeficientů $\bar{\delta}(\underset{\sim}{f}_n)$ a $\sigma^2(\underset{\sim}{f}_n)$ při použití zcela odlišných výpočtových prostředků.

3. ODHADY TYPU JACK-KNIFE A JEJICH ZÁKLADNÍ STATISTICKÉ VLASTNOSTI

O metodách jack-knife je přehledně pojednáno např. v článku [8], kde lze nalézt i značné množství odkazů na literaturu. Z hlediska našeho použití jsou nejvýznamnější reference na [6], [7] a [2].

Za základní jack-knifeový odhad $\delta(\underset{\sim}{p})$ založený na konzistentním koeficientu $\delta(\underset{\sim}{f}_n)$ lze považovat odhad $\hat{\delta}(\underset{\sim}{f}_n)$ definovaný předpisem (viz [8], str. 3)

$$\hat{\delta}(\underset{\sim}{f}_n) = n\cdot\delta(\underset{\sim}{f}_n) - [(n-1)/n] \sum_{i=1}^{n} \delta(\underset{\sim}{f}_{-i}) , \tag{12}$$

kde $\delta(\underset{\sim}{f}_{-i})$ je odhad koeficientu $\delta(\underset{\sim}{p})$ počítaný z vektoru relativních četností $\underset{\sim}{f}_{-i} = \frac{1}{n-1}(n\cdot\underset{\sim}{f}_n)_{-i}$ pro vektor $(n\cdot\underset{\sim}{f}_n)_{-i}$ vzniklý z $(n\cdot\underset{\sim}{f}_n)$ vyřazením (vynecháním, vyškrtnutím) i-tého pozorování. (Je-li $n\cdot\underset{\sim}{f}_n = \sum_{i=1}^{n} \underset{\sim}{e}_i$ pro nezávislé n.v. $\underset{\sim}{e}_i$ rozdělené multinomicky $M(1,\underset{\sim}{p})$, pak $\underset{\sim}{f}_{-i} = \frac{1}{n-1}(n\cdot\underset{\sim}{f}_n - \underset{\sim}{e}_i)$. (Za vhodných předpokladů na hladkost funkce $\delta(\cdot)$ lze dokázat asymptotickou normalitu statistiky $n^{1/2}(\hat{\delta}(\underset{\sim}{f}_n) - \delta(\underset{\sim}{p}))$. K témuž výsledku je možno dojít buď přímočarým zobecněním výsledků článku [6] na hladké funkce výběrového průměru mnohorozměrných statistik nebo správnou aplikací mnohem obecnějšího výsledku pro funkce několika U-statistik: Za předpokladu omezenosti 2. parciálních derivací $\delta(\cdot)$ na okolí bodu $\underset{\sim}{p}$ dostáváme (viz [2], str. 2084), že

$$n^{1/2}(\hat{\delta}(\underset{\sim}{f}_n) - \delta(\underset{\sim}{p})) \xrightarrow{\mathcal{D}} N(0,\sigma^2(\underset{\sim}{p})), \tag{13}$$

kde $\sigma^2(\underset{\sim}{p})$ bylo definováno vzorcem (4). Lze tedy s pomocí koeficientu $\hat{\delta}(\underset{\sim}{f}_n)$ provádět analogické testy hypotéz jako s $\delta(\underset{\sim}{f}_n)$, resp. $\tilde{\delta}(\underset{\sim}{f}_n)$.

Teoreticky je konstrukce odhadu (12) (pořízeného na základě nezávislého náhodného výběru) motivována možností redukce vychýlení řádu $O(n^{-1})$ původního odhadu $\delta(\underset{\sim}{f}_n)$ (jak se snadno formálně ověří aplikací symbolu E na předpisu (12)). Přesnější aproximaci vychýlení (řádu $O(n^{-2})$) je možné získat za pomoci rozvoje funkce $\hat{\delta}(\cdot)$ jakožto lineární kombinace Taylorových rozvojů veličin $\delta(\underset{\sim}{f}_n)$, $\delta(\underset{\sim}{f}_{-1}),\ldots,\delta(\underset{\sim}{f}_{-n})$ s následnou aplikací symbolu E.

Výsledek

$$[E\,\hat{\delta}(\underset{\sim}{f}_n) - \delta(\underset{\sim}{p})] = -\frac{1}{6}\underset{\sim}{D}_3\cdot[E(d\underset{\sim}{f}_n)^3] - \frac{1}{24}\underset{\sim}{D}_4\cdot[E(d\underset{\sim}{f}_n)^4] + O(n^{-3}), \qquad n\to\infty \tag{14}$$

se nepatrně liší od srovnání aproximace (10) pro $[E\,\tilde{\delta}(\underset{\sim}{f}_n) - \delta(\underset{\sim}{p})]$. Je zajímavé, že obdobně získaná aproximace střední odchylky

$$E[\hat{\delta}(\underset{\sim}{f}_n) - \delta(\underset{\sim}{p})]^2 = n^{-1}\sigma^2(\underset{\sim}{p}) + \mathrm{Var}\left[\frac{1}{2}\underset{\sim}{D}_2\cdot(d\underset{\sim}{f}_n)^2\right] + O(n^{-3}), \qquad n\to\infty \tag{15}$$

do řádu $O(n^{-2})$ vede k témuž vyjádření jako pro $E[\tilde{\delta}(\underset{\sim}{f}_n) - \delta(\underset{\sim}{p})]^2$ ve vzorci (11). Tento výsledek evokuje možnost vzájemné zaměnitelnosti koeficientů $\hat{\delta}(\underset{\sim}{f}_n)$ a $\tilde{\delta}(\underset{\sim}{f}_n)$. (Poznámka: Zatímco shoda asymptotických rozdělení ad(3) a (13) je dána skutečností, že 1. členy Taylorových rozvojů koeficientů $\tilde{\delta}(\underset{\sim}{f}_n)$ - resp. $\delta(\underset{\sim}{f}_n)$ - a $\hat{\delta}(\underset{\sim}{f}_n)$ jsou stejné, je eliminace vychýlení řádu $O(n^{-1})$ u $\tilde{\delta}(\underset{\sim}{f}_n)$ a $\hat{\delta}(\underset{\sim}{f}_n)$ dána specifickou konstrukcí těchto koeficientů. Shoda členů řádu $O(n^{-2})$ v aproximacích středních čtvercových chyb - resp. rozptylů - je však již netriviální vlastností ovlivněnou (skrze $\tilde{\delta}(\cdot)$) speciálním tvarem varianční matice multinomického rozdělení.)

Jack-knifeovému odhadu (12) přísluší dále odhad

$$(16) \qquad \widehat{\sigma^2}(\underset{\sim}{f}_n) = (n-1) \sum_{i=1}^{n} \left(\delta(\underset{\sim}{f}_{-i}) - n^{-1} \sum_{j=1}^{n} \delta(\underset{\sim}{f}_{-j}) \right)^2 ,$$

který je analogií výběrové statistiky S^2 pro odhad rozptylu v nezávislém náhodném výběru. Za předpokladu existence 1. totálního diferenciálu funkce $\delta(\cdot)$ na otevřeném okolí bodu $\underset{\sim}{p}$ je statistika $\widehat{\sigma^2}(\underset{\sim}{f}_n)$ konzistentním odhadem asymptotického rozptylu $\sigma^2(\underset{\sim}{p})$. (Uvedené tvrzení opět vyplývá jako speciální případ z článku [2], str. 2085). Jelikož vyslovený předpoklad zaručuje konzistenci i klasickému odhadu $\sigma^2(\underset{\sim}{f}_n)$, je zajímavé teoreticky srovnávat i momentové charakteristiky obou těchto odhadů. Aproximace vychýlení a středních čtvercových odchylek lze pro oba koeficienty explicitně vyjádřit do členů řádu $O(n^{-1})$ jakožto funkce tenzorových součinů prvních až třetích diferenciálů funkce $\delta(\cdot)$:

$$(17) \quad E\,\sigma^2(\underset{\sim}{f}_n) = \frac{n-1}{n}\,\sigma^2(\underset{\sim}{p}) + n^{-1}\Big\{2\underset{\sim}{D}_{1\times 2}\cdot\big[E(d\underset{\sim}{e})^3\big] + \underset{\sim}{D}_{1\times 3}\cdot\big[E(d\underset{\sim}{e})^2 \times E(d\underset{\sim}{e})^2\big] + \underset{\sim}{D}_{2\times 2}\cdot E\big[(d\underset{\sim}{e}_I)\times(d\underset{\sim}{e}_{II})^2\times(d\underset{\sim}{e}_I)\big]\Big\} + o(n^{-1}), \qquad n\to\infty ,$$

$$(18) \quad E\big[\sigma^2(\underset{\sim}{f}_n) - \sigma^2(\underset{\sim}{p})\big]^2 = n^{-1}\Big\{\underset{\sim}{D}_{1\times 1\times 1}\cdot\big[E(d\underset{\sim}{e})^4 - E(d\underset{\sim}{e})^2\times E(d\underset{\sim}{e})^2\big]\Big\} + 4n^{-1}\Big\{\underset{\sim}{D}_{1\times 2\times 1\times 1}\cdot\big[E(d\underset{\sim}{e})^2\times E(d\underset{\sim}{e})^3\big] + \underset{\sim}{D}_{1\times 2\times 2\times 1}\cdot\big[E(d\underset{\sim}{e})^2\times E(d\underset{\sim}{e})^2\times E(d\underset{\sim}{e})^2\big]\Big\} + o(n^{-1}), \qquad n\to\infty ;$$

$$(19) \quad E\,\widehat{\sigma^2}(\underset{\sim}{f}_n) = \sigma^2(\underset{\sim}{p}) + n^{-1}\Big\{\underset{\sim}{D}_{1\times 2}\cdot\big[E(d\underset{\sim}{e})^3\big] + \underset{\sim}{D}_{1\times 3}\big[E(d\underset{\sim}{e})^2\times E(d\underset{\sim}{e})^2\big] + \underset{\sim}{D}_{2\times 2}\,E\big[(d\underset{\sim}{e}_I)\times(d\underset{\sim}{e}_{II})^2\times(d\underset{\sim}{e}_I)\big]\Big\} + o(n^{-1}), \qquad n\to\infty ,$$

$$(20) \quad E\big[\widehat{\sigma^2}(\underset{\sim}{f}_n) - \sigma^2(\underset{\sim}{p})\big]^2 = E\big[\sigma^2(\underset{\sim}{f}_n) - \sigma^2(\underset{\sim}{p})\big]^2 + o(n^{-1}), \qquad n\to\infty ;$$

kde $\underset{\sim}{e}, \underset{\sim}{e}_I$ a $\underset{\sim}{e}_{II}$ jsou nezávislé multinomické vektory s rozdělením $M(1,\underset{\sim}{p})$. (Hodnoty $E(d\underset{\sim}{e})^2$, $E(d\underset{\sim}{e})^3$ a $E(d\underset{\sim}{e})^4$ lze spočítat dosazením do vzorců (7) a (8) pro $\underset{\sim}{e}=\underset{\sim}{f}_1$ a $n=1$.)

Shodu aproximací (18) a (20) (která vzniká z analogických příčin jako při aproximacích středních čtvercových chyb koeficientů $\delta(\underset{\sim}{f}_n)$ a $\hat{\delta}(\underset{\sim}{f}_n)$) lze interpretovat jako shodu asymptotických rozptylů n.v. $n^{1/2}\big[\sigma^2(\underset{\sim}{f}_n) - \sigma^2(\underset{\sim}{p})\big]$ a $n^{1/2}\big[\widehat{\sigma^2}(\underset{\sim}{f}_n) - \sigma^2(\underset{\sim}{p})\big]$ nazývaných opět jako hladké funkce normovaného multinomického rozdělení. Z hlediska této shody poskytují aproximace (17) a (19) snadné kriterium, na jehož základě je možné rozhodovat, který z obou koeficientů má (jakožto bodový odhad parametru $\sigma^2(\underset{\sim}{p})$) menší (resp. větší) vychýlení: Pokud pro konkrétní koeficient $\delta(\cdot)$ platí nerovnost

$$(21) \qquad \underset{\sim}{D}_{1\times 2}\cdot\big[E(d\underset{\sim}{e})^3\big] - \sigma^2(\underset{\sim}{p}) > 0 ,$$

je výhodnější užití jack-knifeového odhadu $\widehat{\sigma^2}(\underset{\sim}{f}_n)$, v případě opačné nerovnosti má menší vychýlení (řádu $O(n^{-1})$) koeficient $\sigma^2(\underset{\sim}{f}_n)$.

4. ZÁVĚRY PRO PRAKTICKÉ APLIKACE

Praktická použitelnost prezentovaných jack-knifeových odhadů (a s ní spojených teoretických aproximací předchozí stati) pro funkce frekvenčního (normovaného multinomického) vektoru $\underset{\sim}{f}_n$ vyplývá z povšimnutí, že výpočetní vzorce předpisů (12) a (16) lze ve skutečnosti přepsat jako

$$(*) \qquad \hat{\delta}(\underset{\sim}{f}_n) = n\cdot\delta(\underset{\sim}{f}_n) - [(n-1)/n] \sum_{j=1}^{m} n_j\cdot\delta(\underset{\sim}{f}_{-j})$$

a

$$(**) \qquad \widehat{\sigma^2}(\underset{\sim}{f}_n) = (n-1)\left[\sum_{j=1}^{m} n_j(\delta(\underset{\sim}{f}_{-j}))^2 - n^{-1}\left(\sum_{j=1}^{m} n_j\cdot\delta(\underset{\sim}{f}_{-j})\right)^2\right]$$

(n_j značí absolutní četnost v j-té třídě multinomického vektoru $n\cdot\underset{\sim}{f}_n$), neboť vynechání i-tého pozorování (respondenta) se v každé j-té třídě opakuje n_j-krát s týmž výsledkem $\delta(\underset{\sim}{f}_{-j})$. To znamená, že stačí počítat "navíc" (proti původnímu odhadu $\delta(\underset{\sim}{f}_n)$) pouze m hodnot $\delta(\underset{\sim}{f}_{-j})$, což je při operační rychlosti soudobé výpočetní techniky záležitost zcela zanedbatelná.

Druhou výhodu metod jack-knife lze spatřovat v tom, že její užití není závislé na explicitním výpočtu diferenciálů funkce $\delta(\cdot)$, přestože výsledný efekt (v případě $\hat{\delta}(\underset{\sim}{f}_n)$ - -redukce vychýlení řádu $O(n^{-1})$ a vliv na střední čtvercovou odchylku do řádu $O(n^{-2})$; v případě $\sigma^2(\underset{\sim}{f}_n)$ ovlivnění asymptotického rozptylu) odpovídá použití klasických metod s diferenciály. Tato vlastnost přímo vybízí k tomu, aby "jack-knifeování" bylo zavedeno jako univerzální podprogram do prakticky užívaných statistických procedur, kde např. pro danou kontingenční tabulku (získanou multinomickým způsobem výběru) bývá počítáno několik různých koeficientů asociace současně.

Nevýhodou výše popsané metody jacknife při užití na konkrétních koeficientech asociace je vlastnost, že odhad $\hat{\delta}(\underset{\sim}{f}_n)$ ze vzorce (12) (resp.(*)) nemusí padnout do oboru hodnot původního (omezeného) koeficientu $\delta(\underset{\sim}{f}_n)$ (resp. $\delta(\underset{\sim}{p})$), což by vedlo ke sporu s jeho interpretací. Podobně jako u koeficientu $\delta(\underset{\sim}{f}_n)$ k tomuto dochází, když n je ~~malé~~ tak, že vychýlení odhadu $\delta(\underset{\sim}{f}_n)$ podstatným způsobem ovlivňují i členy vyšších řádů $(O(n^{-2}))$. V těchto situacích bývají ovšem problémy i s použitím normální aproximace (3), takže odhady $\delta(\underset{\sim}{f}_n)$, $\hat{\delta}(\underset{\sim}{f}_n)$ i $\widehat{\sigma^2}(\underset{\sim}{f}_n)$ a $\sigma^2(\underset{\sim}{f}_n)$ jsou příliš hrubé. Odstranit tuto nevýhodu by bylo možné přechodem k analogickým odhadům akceptujícím i členy řádu $O(n^{-2})$, což (v případě klasických odhadů) umožňuje jemnější užití aproximací ad (5) a (6) nebo (v případě jack-knifeových metod) použití tzv. jack-knifeování 2. řádu (viz [8], str. 2).

Pozn.: Při formulaci aproximačních vzorců pro vychýlení a střední čtvercové odchylky v textu uvažovaných koeficientů nebyly z úsporných důvodů stanoveny předpoklady na stupeň diferencovatelnosti funkce $\delta(\cdot)$, jejíž diferenciály se v těchto aproximacích vyskytují. Obecně lze říci, že ve vzorcích s vyjádřením zbytku ve tvaru $o(\cdot)$ vystačíme s předpokladem existence nejvyššího diferenciálu $\underset{\sim}{D}_s$ v daném vzorci se vyskytujícím. Podobně jako v případě aproximace (13) lze však ověřit <u>nutnost</u> nepatrně silnějších předpokladů (typu existence omezených (s+1)-ních parciálních derivací funkce $\delta(\cdot)$ v bodě $\underset{\sim}{p}$) při ověřování platnosti srovnatelných vyjádření pro koeficient $\hat{\delta}(f_n)$. Tento zesílený předpoklad (v aproximacích (14) a (15) se indikuje řádem zbytku $O(\cdot)$) však jistě nebude bránit možnostem praktického používání tohoto koeficientu.

LITERATURA:

[1] Anděl J. (1978): Matematická statistika, SNTL, Praha

[2] Arvesen J.N. (1969): Jack-knifing U-statistics, AMS 40, 2076-100

[3] Běláček J. (1981): Asymptotické vlastnosti korelačního poměru η^2 odvozeného z kontingenční tabulky, Diplomová práce, MFF UK

[4] Hurt J. (1976): Asymptotický rozvoj funkcí statistik, Aplikace matematiky 6, 21/1986

[5] Hurt J. (1978): Asymptotic expansions for moments of functions of statistics, Proceedings of the 2-nd Prague Symp. on As. Statistics, 21-25 August, 233-238

[6] Miller R. (1964): A trustworthy jack-knife, AMS 35, 1594-605

[7] Miller R. (1968): Jackknifing variances, AMS 39, 567-82

[8] Miller R. (1974): The jackknife-review, Biometrika 61

[9] Rao C.R. (1978): Lineární metody statistické indukce a jejich aplikace, Academia, Praha

[10] Sikorski R. (1973): Diferenciální a integrální počet, 52-59, 163-174

- - - -