O ANALÝZE MNOHOROZMĚRNÝCH KONTINGENČNÍCH TABULEK

Tomáš Havránek

Podívejme se na následující frekvenční tabulku:

E	D	C	B	A		%
				0	1	
3.0	140	0	0	95	1	1.04
			1	201	8	3.83
		1	0	295	5	1.67
			1	38	2	5.00
	140	0	0	49	5	11.36
			1	117	15	11.36
		1	0	192	15	7.25
			1	20	3	13.04
3.0	140	0	0	58	7	12.73
			1	158	6	3.89
		1	0	145	10	6.45
			1	22	3	12.00
	140	0	0	47	6	12.77
			1	137	17	9.06
		1	0	117	16	6.25
			1	22	9	29.30

kde A je výskyt ischemické choroby srdeční v průběhu pěti let po vyšetření (0 ne, 1 ano), B psychická náročnost práce (subjektivně, 0 ne, 1 ano), C fyzická náročnost práce (subjektivně, 0 ne, 1 ano), D systolický krevní tlak (pod 140, nad 140 včetně), E index alfa a beta lipoproteinů. Hodnoty B až E jsou zjišťovány při vstupním vyštření. Hodnoty veličin D a E nejsou ve zdrojových datech kategorizovány, ale kategorizace se provádí zejména pro tabulační účely. Data pocházejí z výzkumu prováděného Angiologickou laboratoří fakulty všeobecného lékařství UK pod vedením prof.MUDr.Z.Reiniše,DrSc.

Čeho si na tabulce všimneme? Především neobsahuje žádné nulové frekvence. To nám při zpracování ušetří mnohé starosti. Dále je pravděpodobně zřejmé, že A závisí na čtveřici B,C,D,E. Otázkou je však struktura této závislosti i závislosti B,C,D a E mezi sebou. Dále je asi vhodné konstatovat, že žádná z veličin není veličinou řízenou.

1. Stojíme nejprve před otázkou, jak vyjadřovat hypotézy o nezávislosti resp. závislosti náhodných veličin, jejichž pozorováním (nezávislým homogenním náhodným výběrem) vznikne tabulka tohoto typu. Jedna z možností, která se v posledních letech čím dál tím více používá je logaritmicko-lineární vyjádření. O co jde si nejprve ukážeme na případě tří náhodných veličin. Pro jednoduchost budeme předpokládat, že všechny nabývají pouze hodnot 0 a 1.

Veličiny si označíme A,B,C . Nechť nyní p^{ABC}_{ijk} je pravděpodobnost, že trojice $\langle A,B,C\rangle$ nabývá hodnoty $\langle i,j,k\rangle \in \{0,1\}^3$. Tyto pravděpodobnosti můžeme rozepsat jako

$$\log p^{ABC}_{ijk} = \theta + \lambda^A_i + \lambda^B_j + \lambda^C_k + \lambda^{AB}_{ij} + \lambda^{AC}_{ik} + \lambda^{BC}_{jk} + \lambda^{ABC}_{ijk}, \qquad (1)$$

kde požadujeme, aby $\sum_i \lambda^A_i = \sum_j \lambda^B_j = \sum_k \lambda^C_k = \sum_i \lambda^{AB}_{ij} = \sum_j \lambda^{AB}_{ij} = \sum_i \lambda^{AC}_{ik} = \sum_k \lambda^{AC}_{ik} = \sum_j \lambda^{BC}_{jk} = \sum_k \lambda^{BC}_{jk} = \sum_i \lambda^{ABC}_{ijk} = \sum_j \lambda^{ABC}_{ijk} = \sum_k \lambda^{ABC}_{ijk} = 0$

Tento pohled je podobný pohledu při obvyklém modelu analýzy rozptylu s více faktory. Čísla λ^A_i, λ^B_j, λ^C_k nazýváme hlavními efekty (nebo efekty prvního řádu), λ^{AB}_{ij}, λ^{AC}_{ik}, λ^{BC}_{jk} efekty druhého řádu a λ^{ABC}_{ijk} efektem třetího řádu (často se též říká dvoufaktorový efekt) atd. Zpravidla, mluvíme-li o efektech, pak efektem myslíme vždy např. λ^{AB}_{ij} pro každé $i \in \{0,1\}$ a $j \in \{0,1\}$. Model (1), resp. struktura závislosti jemu odpovídající, "vysvětluje" jakoukoliv trojrozměrnou tabulku. Problém je, zda tabulku nelze "vysvětlit" jednodušším způsobem, jednodušší strukturou závislosti vzniklou vynecháním některých efektů, např. zda nestačí předpokládat, že

$$\log\left(p^{ABC}_{ijk}\right) = \theta + \lambda^A_i + \lambda^B_j + \lambda^C_k + \lambda^{BC}_{jk}, \qquad (2)$$

což odpovídá nezávislosti páru C,B na A.

Jednodušší struktury závislosti dostáváme vynecháváním efektů; budeme předpokládat, že vynecháme-li některý efekt druhého řádu, vynecháme i efekt třetího řádu, t.j. λ^{ABC}_{ijk}. Takto získané modely nazýváme hierarchické logaritmicko-linární modely a můžeme je zapisovat zkráceným zápisem. Zapisujeme vždy jen horní indexy nejvyšších obsažených efektů (efekt druhého řádu je vyšší než efekt prvního řádu atd.). Tedy pravděpodobnostní model (1) má zápis (ABC), model (2) má zápis (A,BC) a model úplné nezávislosti $\log\left(p^{ABC}_{ijk}\right) = \theta + \lambda^A_i + \lambda^B_j + \lambda^C_k$ má zápis (A,B,C). Tento způsob zápisu je dostatečně přehledný a mnohdy snadno interpretovatelný: (A,B,C) znamená, že A,B a C jsou navzájem nezávislé atd. Model ABC se nazývá saturovaný model, nebo úplný model třetího řádu, model (AB,AC,BC), který obsahuje všechny efekty prvního a druhého řádu, se nazývá úplný model druhého řádu, podobně (A,B,C) je úplný model prvního řádu.

Důležité je si uvědomit, že o modelech můžeme mluvit pouze na základě jejich velmi stručného zápisu, který je tvořen vlastně výrazy slovy určitého velmi jednoduchého formálního jazyka. S těmito výrazy lze provádět určité operace, které mohou být popsány syntaktickými pravidly a které mohou mít určitý sémantický význam v prostoru odpovídajících pravděpodbností (pravděpodobnostních modelů). Jednoduchý formální zápis modelů je zároveň velice důležitý pro jejich interpretaci a komunikování nestatistikům. Zároveň je vhodný i po stránce počítačové.

Podívejme se nyní na seznam logaritmicko-lineárních hierarchických modelů, které připadají rozumně v úvahu u trojrozměrné tabulky (pomíjíme modely nižší dimenze a modely vzniklé vynecháním efektů prvního řádu):

model	vyjádření v pravděpodobnostech (horní indexy vynecháváme)	stupně volnosti
(A,B,C)	$p_{ijk} = p_{i..} p_{.j.} p_{..k}$	4
(AB,C)	$p_{ijk} = p_{ij.} p_{..k}$	3
(AC,B)	$p_{ijk} = p_{i.k} p_{.j.}$	3
(A,BC)	$p_{ijk} = p_{i..} p_{.jk}$	3
(AB,AC)	$p_{ijk} = p_{ij.} p_{i.k} / p_{i..}$	2
(AB,BC)	$p_{ijk} = p_{ij.} p_{.jk} / p_{.j.}$	2
(AC,BC)	$p_{ijk} = p_{ij.} p_{.jk} / p_{..k}$	2
(AB,AC,BC)	$(p_{111}p_{001})/(p_{101}p_{011}) = (p_{110}p_{000})/(p_{100}p_{010})$	1

Všimněme si, že prvních 7 modelů lze snadno vyjádřit v řeči nezávislosti a podmíněné nezávislosti. Navíc, maximálně věrohodné odhady pravděpodobností p_{ijk} při platnosti každého z těchto modelů dostaneme snadno součinem (či podílem) přímých odhadů marginálních pravděpodobností. Tyto modely se nazývají přímé nebo multiplikativní. Poslední model (AB,AC,BC) má jiný charakter: jednak jeho interpretace je jiná a jednak odhady metodou maximální věrohodnosti musíme získávat iteračním postupem (např. "iterative proportional fitting, viz BISHOP(ová),FIENBERG,HOLLAND(1975)).

2. Logaritmicko lineární modely lze použít i pro tabulky vyšší dimenze, kdy vyjadřování v pravděpodobnostech je velmi složité a nepřehledné. Ve vyšších dimenzích je zřetelně vidět jednoduchost formálního zápisu modelů. Logaritmicko-lineární model odpovídající (A,BCDE) si jistě již každý umí sestavit. Uvažujme nyní obecně tabulky dimenze n a nechť veličiny jsou očíslovány $1,\ldots,n$. Pak každý zápis $(A_1,\ldots,A_k)$, kde $A_i \subseteq \{1,\ldots,n\}$ a pro každé $i \neq j$ není ani $A_i \subseteq A_j$, ani $A_i \supseteq A_j$ definuje jednoznačně logaritmicko-linární hierarchický model. $(A_1,\ldots,A_k)$ se pak nazývá generující sentencí (množinou) modelu. Obvykle se ovšem prakticky používají místo čísel písmena, neboť konkrétně zatím nikdo pravděpodobně nepracuje s tabulkou dimenze větší než $n = 22$. Navíc místo pedantického zápisu $\{\{A,B\},\{B,C,D,E\}\}$ atp. se používá zjednodušený zápis (AB,BCDE), který jsme také používali výše.

Ve vyšších dimenzích již drtivě převažují hierarchické logaritmicko-lineární modely, které nejsou multiplikativní. Viz následující tabulku:

dimenze: n	2	3	4	5	výraz
hierarchické	5	19	167	7580	2^{2^n-1}
z toho					
ZPA, grafové	5	18	113	1450	$\sum_{i=0}^{n}\binom{n}{i}2^{\binom{i}{2}}$
multiplikativní	5	18	110	1233	není znám

(3)

S třídou ZPA logaritmicko-lineárních hierarchických modelů se seznámíme později.

Jak rozlišíme multiplikativní a nemultiplikativní modely? Jednoduchý algoritmus využívající generující sentence lze nalézt v knize BISHOP,FIENBERG a HOLLAND (1975). Zde ho uvádíme v nepatrně pozměněné formulaci:

§ Měj $(A_1,\ldots,A_k)$. Je-li $k = 2$ zastav.
(1) Vyhoď písmeno, které se vyskytuje ve všech A_i.
(2) Vyhoď písmeno, které se vyskytuje v jediném A_i. Nebylo-li možné použít ani (1) ani (2), zastav.
Tím vznikne $(A'_1,\ldots,A'_k)$ (některé A'_i může být prázdné).
(3) Vyhoď každé A_i, které je vlastní částí jiného. Je-li $A_i = A_j$ $i < j$, vyhoď A_j (pro všechna i,j)
Tím vznikne $(A''_1,\ldots,A''_\ell)$, $\ell \leq k$.
Polož $k = \ell$ a $(A_1,\ldots,A_k) = (A''_1,\ldots,A''_\ell)$ a jdi na §.

$(A_1,\ldots,A_k)$ vstupující do prvního cyklu je multiplikativní právě když je výpočet zastaven v některém cyklu pro podmínku $k = 2$.

Uveďme si příklady pro $n = 5$:
(ABC,ACDE) je triviálně multiplikativní.
$(ABE,BCE,ADE,CDE) \xrightarrow{(1)} (AB,BC,AD,CD)$ a již nelze aplikovat (1) a (2); model není multiplikativní.
$(ABE,ADE,BC) \xrightarrow{(1)} (ABE,ADE,B) \xrightarrow{(3)} (ABE,ADE)$ - model je multiplikativní.
Pro n=7: $(ABD,BCE,ACF,EG,ABC) \xrightarrow{(2)} (ABD,BCE,AC,EG,ABC) \xrightarrow{(3)} (ABD,BCE,EG,ABC) \xrightarrow{(2)} (ABD,BCE,E,ABC) \xrightarrow{(3)} (ABD,BCE,ABC)$
$\xrightarrow{(1)} (AD,CE,AC) \xrightarrow{(2)} (A,CE,AC) \xrightarrow{(3)} (CE,AC)$ - model je multiplikativní.
Zde je opět na místě zdůraznit, že jde o čistě syntaktický postup, používající pouze formální

vyjádření modelů.

3. Máme-li dva logaritmicko-lineární hierarchické modely, můžeme uvažovat jejich konjunkci, t.j. model spočívající v platnosti omezení danných oběma modely současně. Píšeme formálně pro tento případ např. $(ABC,ACDE)\&(ABE,ADE,BC)$. Platí, že výsledný model je opět hierarchický logaritmicko-lineární model a že existuje konstruktivní postup, jak nalézt jeho generující sentenci ze sentencí modelů vstupujících do konjunkce:

Budiž $(A_1,\ldots,A_k)$, $(B_1,\ldots,B_\ell)$ dvě generující sentence. Definujme jejich spojení $\sqcap$ jako $(A_1,\ldots,A_k)\sqcap(B_1,\ldots,B_\ell) = (A_1\cap B_1, A_1\cap B_2, \ldots, A_k\cap B_\ell)$ (přitom každé $A_i\cap B_j$, které bude vlastní částí jiného průniku, vynecháme a v případě rovnosti dvou průniků ponecháme ve výrazu vpravo jen jeden z nich).

Platí, že model, který je konjunkcí dvou hierarchických logaritmicko lineárních modelů definovaných generujícími sentencemi $(A_1,\ldots,A_k)$ a $(B_1,\ldots,B)$ je definován generující sentencí $(A_1,\ldots,A_k)\sqcap(B_1,\ldots,B_\ell)$.

Tedy například $(ABC,ACDE)\&(ABE,ADE,BC)$ vyjádříme jako $(ABC,ACDE)\sqcap(ABE,ADE,BC) =$ $= (AB,A,BC,AE,ADE,C) = (AB,BC,ADE)$.

4. Z hlediska intepretačního rozumná třída modelů by měla být uzavřena vzhledem ke konjunkci. Naneštěstí třída multiplikativních modelů uzavřená na konjunkci není. Viz následující příklad:

$(ABC,BCD)\&(ABD,ACD) = (AB,AC,BD,CD)$ (ENKE, 1980) .

Je tedy vhodné hledat nejmenší třídu modelů, uzavřenou na konjunkci a obsahující třídu multiplikativních modelů (samozřejmě jako podtřídu hierarchických logaritmicko-lineárních modelů). Cesta k hledání takové třídy vede přes standardní representaci hierarchických logaritmicko-lineárních modelů.

Uvažujme modely dané pevné dimenze n. Generující sentenci sestávající se z množin kardinality právě n-1 nazveme elementární. Příklad pro n = 5 : $(ABCD,ACDE,ABCE)$. Počet množin v generující sentenci nazveme velikostí sentence.

Platí, že každou generující sentenci lze jednoznačně vyjádřit konjunkcí $\varphi_1\&\ldots\&\varphi_r$ kde $\varphi_1,\ldots,\varphi_r$ jsou elementární sentence. Tuto konjukci nazýváme <u>standardní representací</u>.

Maximální velikostí sentence v $\varphi_1\&\ldots\&\varphi_r$ můžeme měřit složitost původní sentence jejímž standardním vyjádřením daná konjunkce je. Modely, jejichž standarní vyjádření obsahuje sentence velikosti maximálně 2 , nazýváme <u>ZPA modely</u>. Z praktických důvodů se většinou omezujeme na případ, kdy standardní vyjádření obsahuje právě sentence velikosti dvě (jinak by chyběl některý efekt prvního řádu).

Příklady standardních vyjádření: n = 4

$(A,B,C,D) = (ABC,ACD)\&(ABC,BCD)\&(ABC,ABD)\&(ABD,ACD)\&(ABD,BCD)\&(ACD,BCD)$ je ZPA i multiplikativní.

$(AB,AC,AD,BC) = (ABC,ACD)\&(ABC,ABD)\&(ABD,ACD,BCD)$ není ZPA ani multiplikativní.

$(AB,AC,BD,CD) = (ABC,BCD)\&(ABD,ACD)$ je ZPA a není multiplikativní.

(AC,AD,BD) $(ABC,ABD)\&(ABD,ACD)\&(ACD,BCD)$ je ZPA i multiplikativní.

Co znamená ZPA? "Zero partial association". Model ABC,ACD odpovídá nulové parciální asociaci mezi B a C , t.j. podmíněné nezávislosti B a C podmíněno AD (jde o parciální asociaci v Birchově smyslu) .

Platí následující skutečnosti:

(1) Třída ZPA modelů je uzavřena vzhledem ke konjunkci,

(2) třída ZPA modelů obsahuje třídu multiplikativních modelů a

(3) třída ZPA modelů je nejmenší třída hierarchických logaritmicko-lineárních modelů obsahující třídu multiplikativních modelů.

Vidíme, že třída ZPA modelů má vlastnosti naší hledané třídy. ZPA modely, které nejsou multiplikativní, neumožňují ovšem přímé odhady pravděpodobností a mají obtížnější interpretaci. Rozdíl v počtu ZPA modelů a multiplikativních modelů je vidět z tabulky (3). Třída ZPA modelů odpovídá třídě grafových modelů definovaných v práci DARROCH, LAURITZEN a SPEED, (1980). V této práci je také navržen grafický způsob reprezentace těchto modelů, který může být vhodný pro komunikaci modelů zadavatelům a zákazníkům.

5. Máme-li klasickou úlohu testovat danou hypotézu popsanou generující sentencí $\varphi = (A_1, \ldots, A_k)$, víme jak postupovat (za určitých podmínek např. na frekvence v tabulce apod.). Můžeme například použít χ^2 test poměrem věrohodnosti. Ukažme si jej opět pro speciální případ dimenze tabulky, abychom se vyhnuli obtížným formálním zápisům. Nechť tedy $n = 4$. Jsou-li $f_{ijk\ell}$ napozorované frekvence v tabulce, označíme $m = \sum_{i,j,k,\ell} f_{ijk\ell}$ a očekávané frekvence můžeme vyjádřit jako $F_{ijk\ell}(\varphi) = m\, p_{ijk\ell}(\varphi)$, kde $p_{ijk\ell}(\varphi)$ jsou pravděpodobnosti odhadnuté metodou maximální věrohodnosti za předpokladu platnosti hypotézy popsané pomocí φ. K testování pak použijeme statistiku

$$\chi^2(\varphi) = 2 \sum_{ijk\ell} f_{ijk\ell} \log \left(f_{ijk\ell} / F_{ijk\ell}(\varphi) \right)$$

s příslušným počtem stupňů volnosti (7 - počet odhadovaných parametrů).

Někdy je vhodné zkoumat vztahy mezi hypotézami. Řekneme, že $\varphi \models \psi$ (φ logicky vyplývá z ψ) plyne-li z pravdivosti modelu definovaného φ pravdivost modelu definovaného φ (je-li φ splněno, je splněno i ψ). Tuto relaci opět snadno rozpoznáme na syntaktické úrovni :

platí totiž, že $(A_1, \ldots, A_k) \models (B_1, \ldots, B_\ell)$ právě když $(A_1, \ldots, A_k) \leq (B_1, \ldots, B_\ell)$,

kde relace $\leq$ je definována takto: $(A_1, \ldots, A_k) \leq (B_1, \ldots, B_\ell)$, jestliže pro každé A_i existuje B_j tak, že $A_i \subseteq B_j$.

Chceme-li nyní testovat "významnost rozdílu" mezi φ a ψ, $\varphi \models \psi$, použijeme statistiky

$$\chi^2(\varphi / \psi) = \chi^2(\varphi) - \chi^2(\psi) \quad ;$$

stupně volnosti dostaneme rovněž odečtením. Je dobré si všimnout, že φ může být považována za jednodušší hypotézu.

6. Velmi často se ocitáme v situaci popsané na začátku tohoto článku: nemáme žádnou specifikovanou hypotézu, nebo několik málo specifikovaných hypotéz, ale musíme naopak nějaké hypotézy "podporované daty" nalézt. To pak znamená zkoušet celou řadu hypotéz, které by mohly připadat v úvahu a pozorovat jejich shodu s daty. Úmyslně zde je použito slovo "zkoušet" místo testovat, aby bylo jasné odlišení této situace od klasické situace testování. V našem případě, i když používáme testových statistik, dosažených hladin významnosti atd., chápeme tyto spíše jako míry shody a neshody hypotéz s daty, než jako testy v klasickém smyslu (neuvažujeme zde metody simultánní inference). Metody vyhledávání hypotéz by měly být z teoretického hlediska vlastně hodnoceny jinak, než klasické metody testování. Obrana proti chybě prvního druhu (resp. globální chybě prvního druhu) zde možná není vůbec tou nejdůležitější věcí. Spíše zde záleží na něčem, co by mohlo být nazváno silou - například na tom, s jakou pravděpodobností metoda odhalí danou (známou) strukturu závislosti. Zde je tedy mnoho otevřených otázek a my zatím používáme řekněme empiricko-intuitivní přístup.

Vraťme se nyní k frekvenčním tabulkám. Víme,že hierarchických logaritmicko-lineárních hypotéz, které by mohly teoreticky připadat v úvahu je velmi velké množství. Musíme se tedy omezit na nějakou rozumnou podtřídu. Z řady výše uvedných důvodů je prvním kandidátem na takou třídu právě třída ZPA modelů:

(1) je uzavřená vůči konjunkci,(2) obsahuje třídu multiplikativních modelů a je to nejmenší taková třída a (3) všechny ZPA modely jsou popsatelné konjukcemi elementárních sentencí,které popisují jednoduché ZPA modely vyjadřující nulovou parciální asociaci párů veličin.

Na druhé straně je vidět,že i ZPA hypotéz může být poměrně mnoho.Je proto nutné využít vztahy mezi hypotézami k tomu,abychom memuseli zkoumat celou třídu ZPA hypotéz. Zvolíme si pevnou hladinu významnosti; hypotézy s dosaženou hladinou významnosti nižší než zvolená zamítáme, ostatní "akceptujeme".

Zdá se vhodné využít toto pravidlo:

Je-li $\varphi \vDash \psi$ (a tedy $\varphi \leq \psi$) pak při zamítnutí ψ v daných datech zamítneme i φ (aniž bychom ho testovali).

Mohlo by se zdát,že je vhodné použít i opačného vztahu: při "akceptování" φ "akceptovat" i ψ. Tedy například,kdybychom akceptovali některé elementární ZPA modely ,akceptovali bychom i jejich konjunkci. Víme však,že role zamítnutí a nezamítnutí není v tomto smyslu symetrická a že zamítnutí můžeme považovat za "jistější".

(Předběhneme nyní k dokončení našeho příkladu. Nezamítnutí hypotéz označených 2,3,5,6,0 při dané hladině významnosti by tedy znamenalo akceptování hypotézy 23560 = (AB,BC,ADE),která má však hodnotu $\chi^2 = 41.56$ při 20 stupních volnosti a tedy dosaženou hladinu významnosti 0.0032!)

Samozřejmě,že při použití testu χ^2 poměrem věrohodnosti (i Pearsonova χ^2) se nám může stát, že sice ψ zamítneme,ale φ při testování nikoliv.Při použití výše zmíněného pravidla dojde tedy k chybnému rozhodnutí (jde ovšem o něco jiného než chyba I. a II. druhu) vzhledem k daným datům. Praxe však ukazuje,že tyto případy nejsou příliš časté a takto zamítnuté hypotézy bez testování nemívají dosaženou hladinu významnosti příliš vyšší než je zvolená mez a nepatřily by k "nejlepším" hypotézám podle kriterí naznačených v bodě 7 .Celá věc by si však zasloužila ještě hlubší zkoumání.Jde o zjištění jak časté či pravděpodobné jsou chyby zmíněné výše a nebo o konstrukci testu,který by neměl tuto nepříjemnou (i z teoretického hlediska) vlastnost χ^2 testů.

Celý postup vyhledávání "akceptovatelných" hypotéz vypadá s použitím námi navrženého pravidla takto:

ZPA hypotézy generujeme a testujem v pořadí opačném k uspořádání $\leq$. To znamená,že v prvním kroce testujem elementární ZPA hypotézy o parciální asociaci. V dalším kroce testujeme hypotézy odpovídající dvoučlenným konjunkcím elementárních sentencí,v třetím kroce hypotézy odpovídající trojčlenným konjunkcím atd. Tak máme zajištěno,že pro dané φ $(= \varphi_1 \& \cdots \& \varphi_k)$ budou všechny hypotézy ψ takové,že $\psi \vDash \varphi$ (a tedy takové,které v případě zamítnutí už nemusíme generovat a testovat) uvažovány později než φ a můžeme tedy informaci o zamítnutí φ využít. Výsledkem je množina "akceptovaných" hypotéz,t.j. takových hypotéz,které byly testovány a nezamítnuty (hypotézy ,které jsou zamítnuty na základě našeho pravidla nejsou testovány).

Takto získaná množina kandidátů může být ještě dosti velká. Je proto vhodné z ní vybírat "nesilnější","nejzajímavější", "nejlepší" apod. hypotézy. Výše uvedená slova je nutné nějakým zpsobem specifikovat. Jednou z čistě logických možností je hledat minimální prvky množiny akceptovaných hypotéz vzhledem k uspořádání $\leq$ .Tím dostáváme vlastně i nejjednodušší akceptované hypotézy pokud měříme složitost počtem obsažených logaritmicko-lineárních parametrů (efektů) .To je ve shodě se strategií propagovanou BENEDETTI ovou a BROWnem .

7. Dokončeme nyní náš příklad:

	Testováno:	zamítnuto (0.1)	parciální a marginální asociace dle Browna:	
1.	ABCD,ABCE	+	+	+
2.	ABCD,ABDE		+	+
3.	ABCD,ACDE		+	+
4.	ABCD,BCDE	+	+	+
5.	ABCE,ABDE			
6.	ABCE,ACDE			
7.	ABCE,BCDE	+	+	+
8.	ABDE,ACDE	+	+	+(nejsilnější vazba)
9.	ABDE,BCDE	+	+	
10.	ABDE,BCDE			

23	ABCD,ADE	+ multiplikativní	23 znamená konjunkci elementárních sentencí 2 a 3.Testovány jsou pouze dvouclenné konjunkce elementárních sentencí nezamítnutých v prvním kroce.
25	ABC,ABDE	multiplikativní	
26	ABC,ACD,ABE,ADE	ne	
20	BCD,ABDE	multiplikativní	
35	ABC,ABD,ACE,ADE	ne	
36	ABC,ACDE	multiplikativní	Je tedy testováno pouze 10 hypotéz místo $\binom{10}{2}$ 45 možných.
30	ABD,BCD,ADE,CDE	+ ne	
56	ABCE,ADE	multiplikativní	
60	ABE,BCE,ADE,CDE	ne	
256	ABC,ABE,ADE	multiplikativní	V třetím kroce jsou testovány pouze ty konjunkce,které nejsou zamítnuty již na na základě zamítnutí konjunkcí v předchozích krocích.Tedy netestujeme např 125,nebo 350. Testované 5 hypotéz místo 105 možných.
250	BC,ABDE	multiplikativní	
260	BC,CD,ABE,ADE	ne	
356	ABC,ACE,ADE	multiplikativní	
560	ABE,BCE,ADE	multiplikativní	
2560	ABE,ADE,BC	multiplikativní	Ve čtvrtém kroce máme již pouze jedinou konjunkci,které není zamítnuta na základě předchozích výsledků.

Množina"akceptovatelných" hypotéz je tedy tvořena všemi hypotézami (modely),sentencemi které byly testovány a nezamítnuty. Zkráceně jde tedy o modely zapsané výše kódy 2,3,5,6,0,25,26,35,30,60,256,250,260,356,560,2560.

Které z uvedených hypotéz jsou ty "nejlepší" a měly by být prezentovány jako možné hypotézy pro další zkoumání? Jeden z možných přístupů je hledat minimální prvky množiny "akceptovaných" sentencí vzhledem k uspořádání $\leq$. Platí totiž,že na teoretické úrovni jsou všechny ostatní sentence v množině "akceptovaných" sentencí jejich logickými důsledky. Zde jsou dva takové prvky a to 356 a 2560, naštěstí obě multiplikativní .

Pro vyhledávání těchto prvků je vhodné si všimnou vztahu mezi relací $\leq$ mezi sentencemi a inkluzí jejich kódů . $\varphi \geq \psi$ je ekvivalentní kód(φ) $\subseteq$ kód(ψ). Např.:(ABE,BCE,ADE) $\geq$ (ABE,ADE,BC) a $560 \subseteq 2560$. Hlednání minimálních prvků v $\leq$ se tedy převádí na hledání maximálních kódů vzhledem k inkluzi.

Někdy může být užitečná i celá množina"akceptovaných" hypotéz. Navržené kriterium pro výběr "nejzajímavějších" hypotéz není jediné možné. Při výběru mohou hrát roli i apriorní znalosti zadavatele o tom,co s čím může souviset, nebo jiná kriteria výběru "nejlepších" hypotéz, například podle nevyšší dosažené hladiny významnosti chápané jako kr terium shody modelu s dat, či podle různých modifikací χ^2 jako míry shody, např. AIC (Akaike information criterion), viz SAKAMOTO a AKAIKE, (1978) ; AIC = χ^2 - 2 d.f. .V našem příkladě jsou dvě hypotézy s nižším AIC (a tedy lepším) než hypotézy vybrané a to 256 a 56.

Při větších příkladech to znamená skladovat množinu "akceptovaných" hypotéz spolu s některými numerickými údaji a pomocí dalších programů z ní pak vybírat podle různých návrhů a kriterií výše zmíněnných i podle dalších kriterií,která mohou být z různých důvodů uplatněna.

Příklad byl spočítán pomocí programu BMDP3F ze systému BMDP (viz DIXON and BROWN,1977). Tento program umožňuje snadno testovat zadaný model.Pro každý testovaný model bylo nutné napsat příkaz MODEL* ABC,ABE,ADE. atp. ,což je při větších příkladech,kdy by bylo testováno větší množtví hypotéz již dosti pracné. Napsání programu,který by prováděl celý výše popsaný postup automaticky máme v plánu.

8. Kromě přímo citovaných prací se tex opírá o dosud nepublikované preprinty DARROCH a SPEED(1979) a HAVRÁNEK(1981a,b) Seznam literatury obsahuje další související práce.

LITERATURA:

AITKIN M.(1980). A note on the selection of log-linear models,Biometrics 36,173-178.

BENEDETTI J.K.,BROWN M.B. (1978). Strategies for the selection of log-linear models. Biometrics 34,680-686.

BISHOP Y.M.M.,FIENBERG S.E.,HOLLAND P.W. (1975) .Discrete multivariate analysis: Theory and practice, MIT Press, Cambridge,Mass.

DARROCH J.N.,LAURITZEN S.L.,SPEED T.P. (1980).Markov fields and log-linear interaction models for contingency tables. The Annals of Statistics 8,522-539.

DARROCH J.N.,SPEED T.P. (1979). Multiplicative and additive models and interactions. Dept,of theoretical statistics,res.rep.49,University of Aarhus.

DIXON W.J.,BROWN M.B. (1977) .Biomedical computer programs, BMDP. University of California Press, Los Angeles.

ENKE H. (1980) To some reasonable test procedures in multiple contingency tables to investigate certain epidemiological or medicin-sociological relationships. Biometrical Journal 22, 779-793.

GOODMAN L.A. (1970) The multivariate analysis of qualitative data: interactions among classifications. JASA 65,226-256.

GOODMAN L.A. (1971) Partitioning of chi-square ,analysis of contingency tables,and estimation of expected frequencies in multidimensional contingency tables. JASA 66, 339-344.

HAVRÁNEK, T. (1981a) On some possibilities of logical analysis of the dependency structure in multidimensional contingency tables, EMS Wroclaw.

HAVRÁNEK, T. (1981b) A procedure for model search in multidimensional contingency tables, Statistical semetster of the Banach Center,Warszawa.

WERMUTH, N. (1976). Model search among multiplicative models, Biometrics 32, 253-263.

WERMUTH, N. (1980). Linear recursive equations,covariance selection and path analysis, JASA 75, 963-972.