Metodologická rubrika

Faktorová analýza kontingenčních tabulek: Možnosti využití standardních programů

MICHAL ČAKRT
CAPA Praha

1. Úvod

Přestože dnes sociologové mají k dispozici mocné prostředky pro práci s daty, reprezentované například moderními multivariačními metodami, zůstává dodnes jakýmsi „evergreenem" empirické sociologické práce analýza kategorizovaných dat, uspořádaných do kontingenčních tabulek. Tabulky nepoužíváme pouze pro potřeby rozborů, ale také – spolu s grafy – jako velmi srozumitelný prostředek prezentace výsledků.

Tradiční přístup k analýze kontingenčních tabulek je zaměstnán především testováním *modelu nezávislosti*. Ačkoli mnoho sociologů dodnes za tento rámec nevychází, pro seriózní práci je pouhé zamítnutí nulové hypotézy málo uspokojivé. Pro interpretaci většinou potřebujeme hlubší analytickou informaci o *struktuře* a *směrech* závislosti zkoumaných veličin. Jedním z pokusů o překonání tohoto omezení je dnes již obecně známá metoda *znaménkového schématu*, opřená o rozbor reziduí, jejíž aplikabilita, přednosti a nedostatky jsou v naší odborné literatuře dostatečně zhodnoceny, především pracemi J. Řeháka a B. Řehákové [1978; 1987].

Variantní, i když dnes již ne zcela nový, leč však dlouho opomíjený [viz Hill, 1974], přístup představuje *kanonická dekompozice reziduálních odchylek očekávaných a skutečných četností* v tabulkách. Kanonická korelační analýza dvojrozměrných tabulek byla do statistické literatury uvedena R. A. Fisherem [1949]. Její souhrnný výklad podali především M. G. Kendall a A. Stuart [1961, ruský překlad 1973]. Tento přístup dal vzniknout skupině metod, z nichž nejznámější je nyní *model korespondenční analýzy*. První zmínky o kanonické analýze kontingenčních tabulek jsou u nás spjaty se jmény K. Laubera a J. Linharta [1978]. Instruktivní a především srozumitelný výklad podávají časopisecky I. Loučková a J. Řehák [1984] (první zmínka o metodě LINDA pochází z roku 1979), knižně pak J. Řehák a B. Řeháková [1986], při formulaci a popisu prvních aplikací metody *LINDA–A*. Zde je uplatněn souběžný *model faktorový*, jehož základním cílem je odhalit případnou existenci společných latentních faktorů, vysvětlujících strukturu reziduálních odchylek v tabulce [srov. Loučková, Řehák 1984 : 174].

Aby čtenář nebyl nucen – jakkoli by to bylo mnohokrát užitečné – listovat několik let starými Metodologickými rubrikami Sociologického časopisu anebo hledat příslušnou pasáž v uvedené knize, připomeňme nyní ve stručnosti základní výchozí teze faktorového modelu a některé jeho analytické aspekty.

Zamítáme-li hypotézu o nezávislosti veličin v n rozměrné kontingenční tabulce, máme důvod předpokládat, že všechny vlivy, působící v tabulce se mohou spojit do poměrně nízkého (vzhledem k dimenzím tabulky a k celkovému množství kategorií) počtu souběžných, avšak navzájem nezávislých příčinných komplexů, analogických společným faktorům, do nichž se spojují vlivy, projevující se ve výchozí kovarianční struktuře (například v korelační matici), a které budou navíc natolik silné, že se v datech statisticky konzistentně projeví. Již na tomto

místě se sluší poznamenat, že tento fakt s sebou nese kromě mnoha předností, o nichž bude řeč dále, i některé nectnosti, dobře známé z „tradiční" faktorové analýzy: příslušné vlivy se především mohou seskupovat k souběžnému působení velmi mnoha způsoby a tato skutečnost často působí mnohé interpretační potíže. Právě zde mj. pramení mnoho výhrad vůči faktorové analýze vůbec.

Přednosti faktorového přístupu lze naopak stručně shrnout do několika bodů:

- na strukturu tabulky se pohlíží jako na **celek,** zatímco tradiční přístupy, včetně znaménkového schématu, hodnotí každé políčko zvlášť proti všem ostatním, tzn. rozkládají celkovou chí-kvadrátovou asociaci na separátní aditivní složky; v tomto smyslu je faktorový přístup simultánní;
- k analýze se přistupuje exploračně, **nezávisle na statistickém testování hypotéz,** to znamená, že metodu lze uplatnit i tam, kde klasické testy nezávislosti nejsou buď vhodné (např. úplná šetření), málo informativní (extrémně velké soubory), popřípadě prakticky vyloučené (výskyt políček s nulovými nebo velmi malými četnostmi);
- jde o realizaci vícerozměrného škálovacího postupu s několika množinami objektů, který má současně přednosti **relačních zobrazení:** jak jeho numerické, tak i grafické výsledky jsou velmi názorné a mohou odhalit zajímavé, například též nelineární vztahy v datech;
- faktorový přístup navíc překračuje omezení daná klasickou kanonickou dekompozicí tím, že umožňuje analyzovat tabulky o **libovolné dimenzi** – v praxi je ovšem počet rozměrů přirozeně z mnoha stran omezen.

Metoda byla dlouho stranou pozornosti, především pro svoji výpočetní náročnost a spolu s jinými, například právě s faktorovou analýzou, čekala na svoje vzkříšení výkonnou výpočetní technikou. I tak byl její vstup do sféry aplikací poněkud opožděn a teprve v několika málo posledních letech jsou metody korespondenční analýzy zahrnuty do novějších verzí velkých statistických výpočetních systémů, například SAS (PROC CORRESP) či BMDP (program CA), u nás však, žel, dosud jen obtížně dostupných.

Výpočty, které vyžaduje jak kanonická analýza kontingenčních tabulek, tak i faktorový model lze, jak dokázal Mickey [1970], naštěstí, při dodržení vcelku nenáročných a vskutku výsledku přiměřených podmínek, s překvapivou snadností realizovat pomocí většiny standardních statistických programů, běžně v praxi používaných pro kanonické korelace nebo faktorovou analýzu. Tyto podmínky se týkají:

1. převedení údajů shrnutých do tabelární formy do takové datové struktury, kterou je příslušný program schopen akceptovat a
2. v předpisu pro výpočet korelační matice, z níž budou faktory extrahovány.

V následujícím textu bude kladen důraz především na *aplikační stránku analýz,* tak aby ji sociolog s pomocí příkladů mohl zvládnout sám s tím, že případní zájemci o hlubší teoretické pozadí jsou odkázáni na uvedenou literaturu.

2. Datová struktura

Základem předkládané techniky je převod dat z tabelární formy do podoby, která je standardnímu programu umožní odpovídajícím způsobem načíst. V zásadě je nutno zachovat dva typy údajů: informaci o **poloze políčka** a informaci o **jeho četnosti.**

Nechť (n_{ij}), $i = 1,\ldots, p$, $j = 1, \ldots, q$ popisuje dvojrozměrnou kontingenční tabulku $p \times q$ a $\Sigma n_{ij} = N$ pak celkovou četnost. Generujme N případů (např. datových řádků, pozorování, štítků apod.) o $p + q$ veličinách $x_i, \ldots, x_p, y_1, \ldots, y_q$, takových, že pro případ n_{ij} platí:

$$
\begin{aligned}
x_i &= y_j = 1 \\
x_{i'} &= 0 \qquad\qquad i' \neq i \\
y_{j'} &= 0 \qquad\qquad j' \neq j
\end{aligned} \tag{1}
$$

Těchto p × q případů lze pak zpracovat standardním programem pro analýzu kanonických korelací, za podmínky, že příslušné korelace v matici spočítáme tzv. kolem počátku, to znamená, že průměry jednotlivých veličin položíme rovny nule. Lze dokázat, že takto získané výsledky jsou totožné s výsledky analýzy popsané Stuartem a Kendallem [1961].

Smyslem použití korelací vypočtených kolem počátku je vyloučení nepravých korelací, například mezi sousedními (často ordinálními) kategoriemi uvnitř jednotlivých dimenzí (sloupců, řádků atd.) tabulky. Jinými slovy jde o to, „přinutit" korelace, aby vyjadřovaly především sílu vztahů mezi různými dimenzemi tabulky a pokud existuje vztah i uvnitř těchto dimenzí, pak aby vycházel najevo pouze „prostřednictvím" této kontigence.

$$
\begin{bmatrix} -R\Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & -R\Sigma_{yy} \end{bmatrix} = 0, \tag{2}
$$

kde Σ_{xx}, Σ_{yy} a Σ_{xy} jsou maticemi součinů p × p mezi veličinami x, resp. q × q veličin y a p × q mezi veličinami x a y, přičemž R označuje koeficient kanonické korelace. V případě, že průměry veličin x a y jsou nulové pro všechny veličiny definované (1), pak se z rovnice (2) stává rovnice:

$$
\begin{bmatrix} -Rn_{i\cdot} & n_{ij} \\ n_{ji} & -Rn_{\cdot j} \end{bmatrix} = 0. \tag{3}
$$

V notaci zavedené Stuartem a Kendallem n, označuje diagonální matici p × p, jejíž i-tý diagonální prvek má hodnotu

$$
n_{i\cdot} = \sum_j n_{ij}
$$

n_j označuje diagonální matici q × q, jejíž j-tý diagonální prvek má hodnotu

$$
n_{\cdot j} = \sum_i n_{ij}
$$

a n_{ij} označuje původní kontingenční tabulku p × q, vyjádřenou ve formě matice. Rovnice (3) je současně úlohou, která řeší problém charakteristických čísel a současně i problém kanonických korelací dvojrozměrné kontingenční tabulky, popsaný Kendallem a Stuartem.

Tento přístup lze **rozšířit a zobecnit na n-rozměrnou kontingenční tabulku** s tím, že není nutno pracovat pouze se dvěma množinami veličin x a y, ale libovolnou n-ticí a problém výpočtu charakteristických čísel není otázkou kanonických korelací mezi dvěma množinami, nýbrž charakteristických čísel **faktorové analýzy**. Láme se tak nejenom bariéra dvojrozměrnosti, ale řešení současně umožňuje *rotaci* faktorů, což podstatně přispívá k hledání optimální interpretace výsledků.

Moderní programy pro faktorovou analýzu umožňují výpočet *vážené korelační matice*, což

Tabulka 1. Symbolické rozložení kategorií

a	b	c
d	e	f
g	h	i
j	k	l

obchází nutnost generovat plný počet N případů. Postačí pouze p × q případů, respektive suma políček s nenulovou četností podle předpisu uvedeného dále.

Předpokládejme, že analyzujeme dvojrozměrnou tabulku o 4 řádcích a 3 sloupcích podle tabulky 1.

Tabulka je tvořena 12 políčky, označenými *a* až *l*, které lze podle (1) z hlediska jejich polohy popsat pomocí 12 případů se 3 + 4, tj. 7 veličinami. Veličiny y_1 až y_3 korespondují se sloupci tabulky, veličiny x_1 až x_4 pak s jejími řádky. Vzniká tak incidenční struktura v podobě dichotomické vzorovací matice, v níž jedničky na pozicích příslušného řádku a sloupce indikují dané políčko. Údaj o četnosti může být buď (méně elegantně) implikován opakovaným výskytem případu v datové struktuře (případů by tudíž nebylo 12, nýbrž N) nebo (úhledněji) může být zahrnut v další veličině n_{ij}, která vyjadřuje váhu případu. Skutečnost, že nejprve udáváme sloupcovou polohu plyne čistě z technických důvodů snazší práce s rychleji se měnícím indexem jako prvním, což odpovídá i intuitivnímu „čtení" tabulky (a jak později uvidíme i zadávání do počítače) po řádcích zleva doprava. Výsledná podoba shora uvedené tabulky je na obrázku 1.

Je-li četnost políčka nulová můžeme buď vyznačit nulovou váhu nebo jednoduše celý řádek vypustit.

Vzorec pro výpočet korelačního koeficientu kolem počátku ze shora uvedené datové matice je následující:

$$n_{ij} = \frac{n_{ij}}{\sqrt{n_{\cdot i} \times n_{j\cdot}}}, \qquad (4)$$

kde n_{ij} je četnost příslušného políčka a $n_{\cdot j}$ a $n_{i\cdot}$ jsou příslušné marginální četnosti. Pokud bychom měli být zcela přesní, nejde v pravém slova smyslu o korelační koeficienty – nemohou totiž nabýt záporných hodnot – nýbrž o *normované součiny*

Aplikace uvedené metody bude nyní ilustrována na několika příkladech.

Obrázek 1. Obecná datová struktura tabulky

y_1	y_2	y_3	x_1	x_2	x_3	x_4	četnost	políčko
1	0	0	1	0	0	0	n_{11}	a
0	1	0	1	0	0	0	n_{12}	b
0	0	1	1	0	0	0	n_{13}	c
1	0	0	0	1	0	0	n_{21}	d
0	1	0	0	1	0	0	n_{22}	e
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
0	0	1	0	0	0	1	n_{43}	l

3. Příklady

Příklad 1: Dvojrozměrná tabulka, všechna pole obsazena

Naší první ilustrací bude tabulka uvedená jako Příklad 81 v učebnici V. Lamsera a L. Růžičky *Základy statistiky pro sociology* [Lamser, Růžička 1970: 270]. Byla zkoumána závislost mezi spokojeností posluchačů určitého rozhlasového programu a stupněm dosaženého vzdělání. Vzdělání je rozděleno do čtyř kategorií (základní – Z, střední všeobecné – SV, střední odborné – SO a vysokoškolské – VŠ) a spokojenost je vyjádřena třemi stupni (spokojen, nerozhodný, nespokojen). Výsledky průzkumu uvádí tabulka 2.

Tabulka 2. Spokojenost s pořadem podle vzdělanostních skupin

Vzdělání	Spokojen	Nerozhodný	Nespokojen	Celkem
Z	22	15	133	170
SV	25	11	74	110
SO	20	8	42	70
VŠ	30	7	13	50
Celkem	97	41	262	400

Klasický test pomocí statistiky χ^2 vedl k zamítnutí nulové hypotézy o nezávislosti projevu spokojenosti se sledovaným pořadem a vzděláním posluchače. K tomu, abychom se dozvěděli více o struktuře závislostí v naší tabulce existuje nepochybně celá řada přístupů. Je možné sledovat rozložení sloupcových, řádkových a celkových procent, diference mezi skutečnými a očekávanými četnostmi či zkoumat rezidua, například pomocí znaménkového schématu. Nám tato učebnicová tabulka poslouží k předložení numericko-grafické prezentace pomocí faktorové analýzy.

Podle (1) vyjádříme tabulku 2 následovně (kategorie spokojen, nerozhodný a nespokojen pro stručnost označíme +, 0 a –); tak jak ukazuje obr. 2.

Výsledná korelační matice, vypočtená podle (4) vypadá následovně, viz obr. 3.

Podle (4) je například korelace mezi kategorií osob se základním vzděláním a kladným hodnocením programu:

$$\frac{22}{(170 \times 97)^{1/2}} = \frac{22}{128.4134} = 0.171321$$

Pozorný čtenář si jistě povšiml specifické vlastnosti korelací (resp. normovaných součinů) vypočítávaných bez odečítání průměrů. Uvnitř dimenzí tabulky, tj. mezi řádkovými kategoriemi a sloupcovými kategoriemi jsou korelace nulové.

Obrázek 2. Datová struktura tabulky spokojenosti s pořadem podle vzdělanostních skupin

Spokojenost			Vzdělání				Četnost
+	0	–	Z	SV	SO	VŠ	
1	0	0	1	0	0	0	22
0	1	0	1	0	0	0	15
0	0	1	1	0	0	0	133
1	0	0	0	1	0	0	25
0	1	0	0	1	0	0	11
0	0	1	0	1	0	0	74
1	0	0	0	0	1	0	20
0	1	0	0	0	1	0	8
0	0	1	0	0	1	0	42
1	0	0	0	0	0	1	30
0	1	0	0	0	0	1	7
0	0	1	0	0	0	1	13

Obrázek 3. Korelační matice spočtená z tabulky 2.

	+	0	–	Z	SV	SO	VŠ
+	1.000						
0	0.000	1.000					
–	0.000	0.000	1.000				
Z	0.171	0.180	0.630	1.000			
SV	0.242	0.164	0.436	0.000	1.000		
SO	0.243	0.149	0.310	0.000	0.000	1.000	
VŠ	0.431	0.155	0.114	0.000	0.000	0.000	1.000

Shora uvedená korelační matice pak byla zpracována programem na faktorovou analýzu – v tomto konkrétním případě to byl dosud jeden z nejlepších v oboru, BMDP4M. Program extrahoval 3 faktory (s použitím standardního kritéria hodnoty charakteristického čísla faktoru většího než 1) a po rotaci metodou varimax vypadaly faktorové zátěže tak, jak popisuje Výstup 1.

Jak numerická, tak i grafická prezentace výsledků zřetelně napovídají, že posluchači s nejvyšším dosaženým vzděláním byli s daným pořadem nejvíce spokojeni, zatímco ti, kteří uvedli pouze základní vzdělání byli spokojeni nejméně. Mezi touto polaritou jsou osoby se středním všeobecným vzděláním, které zaujaly spíše nerozhodný postoj. Respondenti se

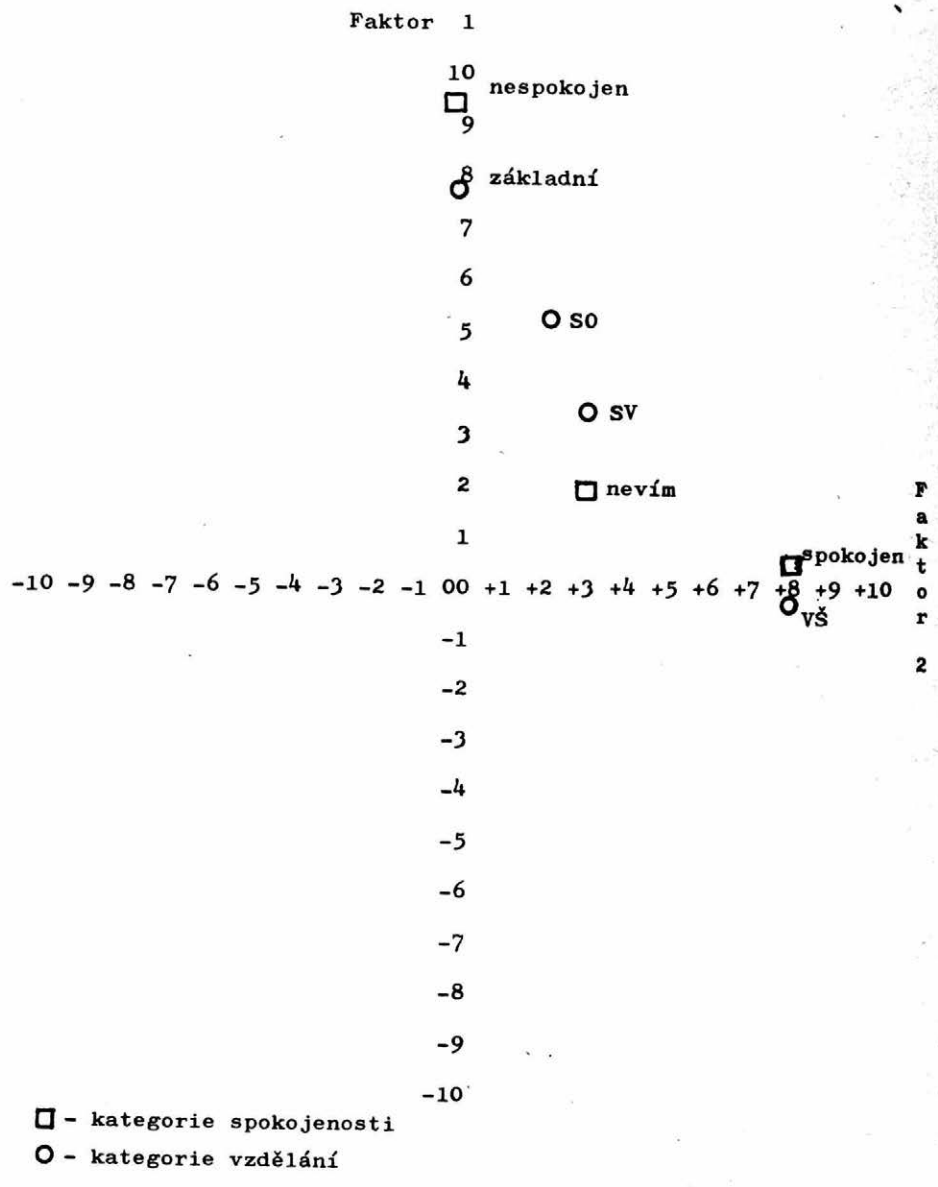
Výstup 1. Setříděné faktorové zátěže tabulky 2[1]

	Faktor 1	Faktor 2	Faktor 3
–	0.933	0.0	0.0
Z	0.801	0.0	0.0
+	0.0	0.849	0.0
VŠ	0.0	0.820	0.0
SV	0.0	0.0	0.753
0	0.0	0.0	0.747
SO	0.480	0.0	0.0

Grafické řešení zobrazuje Obrázek 4.

[1] Faktorové zátěže jsou setříděny tak, že faktory jsou uváděny sestupně podle vyčerpané variance. Řádky jsou uspořádány, tak aby v každém faktoru byly nejvýše ty veličiny, které mají zátěže vyšší než 0.5, přičemž zátěže nižší než 0.25 jsou nahrazeny nulami.

Obr. 4. Faktorové řešení tabulky 2



středním odborným vzděláním nevyjádřili žádné vyhraněné stanovisko, jejich názor je kdesi napůl cesty mezi středoškoláky se všeobecným vzděláním a posluchači, kteří absolvovali pouze základní školu.

Jak již bylo řečeno, faktorová analýza kontigenčních tabulek není omezena pouze na dvojrozměrné tabulky. Lze ji dobře použít i pro vícedimenzionální tabulky a její výsledky jsou, pro tabulky s dobře obsazenými políčky, srovnatelné s logaritmicko lineární analýzou. Výhodná může být aplikace této metody i jako předběžného nástroje pro prohledávání

složitějších tabulek předtím, než se výzkumník rozhodne, jaké logaritmicko lineární modely bude testovat.

V dalším příkladě proto budeme analyzovat trojrozměrnou tabulku s dichotomickými kategoriemi a výsledky faktorové analýzy srovnáme s log-lineárním modelem.

Příklad 2: Trojrozměrná tabulka, všechna pole obsazena

Následující tabulka nebyla vybrána náhodou. Mohou se na ni rozpomenout někdejší účastníci cyklu Vybraných kapitol ze sociologických metod a technik, na nichž nás v červnu 1980 uváděl Tomáš Havránek do tajů analýzy vícerozměrných tabulek [viz Havránek, 1980]. Tím je do značné míry vysvětlena i skutečnost, proč nejde o data sociologická, nýbrž medicínská. Tabeluje se zde výskyt ischemické choroby srdeční ve dvou automobilkách (LIAZ – L a Autoškoda – A) u osob mladších a starších 40 let.

Tabulka 3. Výskyt ischemické choroby srdeční

Věk	Závod	ICHS Ano	ICHS Ne	
> 40	A	26	786	
	L	12	189	
< 40	A	7	145	
	L	7	69	
				1241

Vstupní datovou strukturu naznačuje pro tento příklad Obrázek 5.

Obrázek 5. Datová struktura tabulky výskytu ischemické choroby srdeční

Věk		Závod		ICHS		Četnost
> 40	< 40	A	L	IA	IN	
1	0	1	0	1	0	26
1	0	1	0	0	1	786
1	0	0	1	1	0	12
1	0	0	1	0	1	189
0	1	1	0	1	0	7
0	1	1	0	0	1	145
0	1	0	1	1	0	7
0	1	0	1	0	1	69

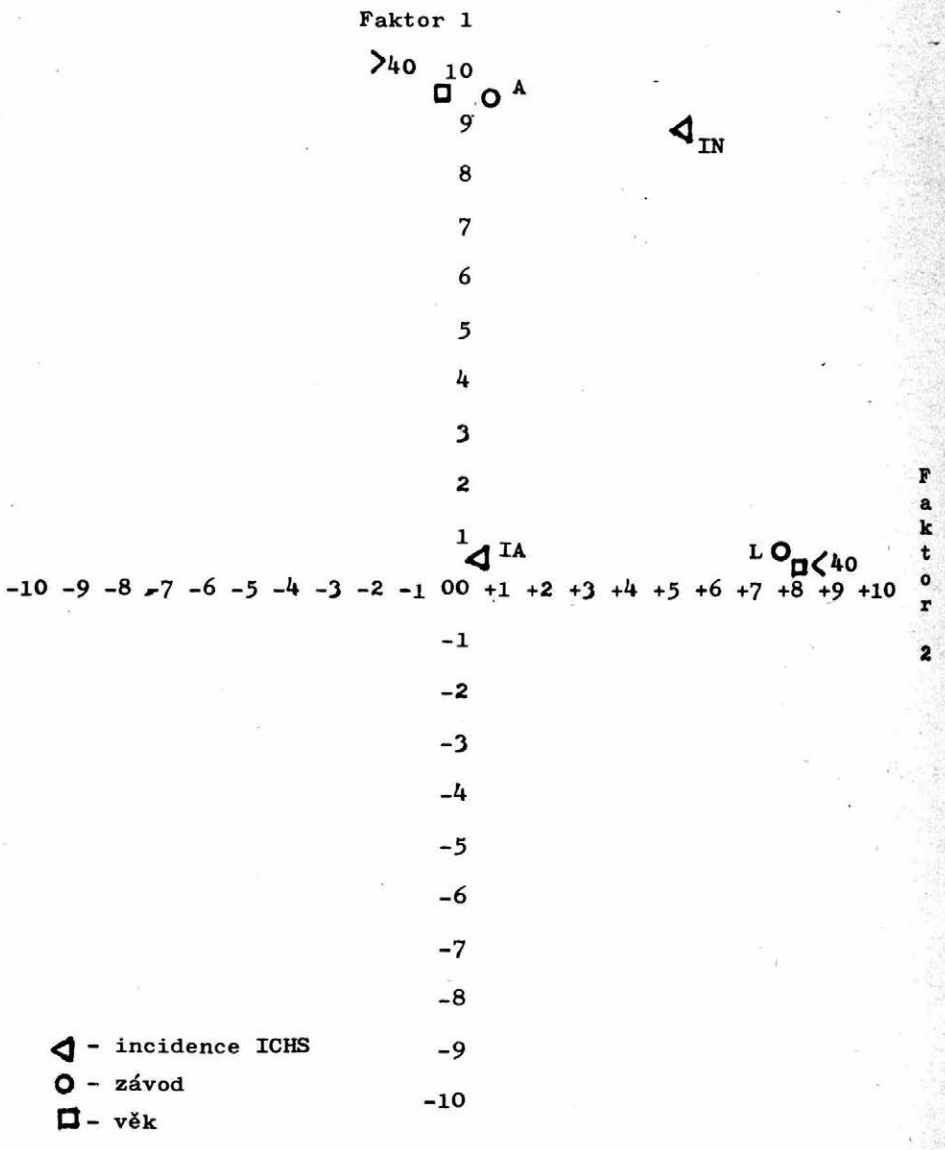
Faktorové řešení ukazuje Výstup 2. Opět byly extrahovány 3 faktory s charakteristickým číslem vyšším než 1.0 a použita rotační metoda varimax.

Výstup 2. Setříděné faktorové zátěže tabulky 3

	Faktor 1	Faktor 2	Faktor 3
> 40	0.951	0.0	0.0
A	0.943	0.0	0.0
IN	0.910	0.405	0.0
< 40	0.0	0.826	0.0
L	0.0	0.761	0.0
IA	0.0	0.0	0.981

Grafické řešení prvních dvou faktorů zobrazuje Obrázek 6.

Obr. 6. Faktorové řešení tabulky 3



Řešení nade vší pochybnost ukazuje, že jediný vztah, který lze v dané tabulce vysledovat je relace mezi závodem a věkovou strukturou, totiž, že v pro závod LIAZ byla v době sběru dat populace typicky mladší než v závodě Autoškoda. Pro vyslovení závěrů týkajících se výskytu ischemické choroby srdeční v závodech nebo věkových skupinách nám tato data neposkytují oporu. Protože naše tabulka má obsazena všechna pole, je možno ji analyzovat i alternativně, pomocí logaritmicko lineárního modelu.

O logaritmicko lineárním modelu měli čtenáři Metodologické rubriky možnost dozvědět se více již dříve a nepokládám proto za nutné podrobněji se o něm nyní na tomto místě rozepisovat. V krátkosti lze jen říci, že strategií je zde vyhledávat taková zjednodušení

vícerozměrné tabulky, vynecháváním některých efektu (tj. dimenzí tabulky), která ještě jsou schopna v rámci přípustné chyby s dostatečnou přesností reprodukovat původní četnosti. Zde půjde o jeho hierachickou variantu, tzn. že vynechání například některého efektu druhého řádu, to je interakce mezi dvěma dimenzami, vypouštíme i efekt třetího řádu, tzn. interakci mezi třemi dimenzemi.

Nejprve se snažíme zjistit, zda lze tabulku vubec zjednodušit, jinými slovy lze-li pracovat s jiným než saturovaným modelem. K tomu poslouží test interakcí. Analýzy byly opět prováděny pomocí programu BMDP, tentokrát BMDP4F.

Řád interakce	χ^2_{LR} [df]	významnost
1	2227.77[3]	0.00000
2	25.37[3]	0.00001
3	0.02[1]	0.89248

χ^2_{LR}[df] je hodnota statistiky chí-kvadrát poměrem věrohodnosti, přičemž v závorkách [ ] jsou uvedeny stupně volnosti

Jak patrno, pro naši tabulku budou postačovat efekty druhého řádu.

Dalším krokem je zjišťování bez kterých efektů se při zjednodušené rekonstrukci tabulky neobejdeme. K tomu nám poslouží testy marginálních a parciálních asociací.

Efekt	Parciální X_{LR}^2	významnost	Marginální X_{LR}^2	významnost
I (ICHS)	1288.67	0.0		
Z (závod)	402.61	0.0		
V (věk)	536.50	0.0		
IZ	4.79	0.0285	5.66	0.0174
IV	1.54	0.2142	2.40	0.1212
ZV	17.31	0.0000	18.17	0.0000

Zjišťujeme jednak, že jak z hlediska závodů, tak z hlediska věku i incidence sledované choroby, nejsou respondenti kategorií rozloženi rovnoměrně. Důležitější je nález, že z interakcí druhého řádu je vysoce významný efekt spolupůsobení věku a závodu a ve výrazně nižší míře i vztah choroby a závodu.

Nyní již můžeme otestovat, nakolik jednotlivé modely reprodukují celkovou strukturu našich výchozích dat.

Model(df)	X_{LR}^2	významnost
IZ(4)	556.23	0.0
IV(4)	425.59	0.0
ZV(4)	1295.88	0.0
I, ZV(3)	**7.23**	**0.0654**
Z, IV(3)	22.98	0.0
V, IZ(3)	19.73	0.0002
IZ, IV (2)	17.33	0.0002
IV, ZV(2)	4.81	0.0901
ZV, IZ(2)	1.56	0,4582

Podle kritérií uváděných Bennedettim a Brownem [1978] vybíráme model I. ZV (uvedený zvýrazněně), tj. marginální nezávislost vztahu věku a závodu na výskytu nemoci. Jinými slovy to znamená, že incidence ICHS není ovlivněna ani věkem a ani závodem, a že skutešně významná souvislost, kterou můžeme z našich dat zjistit, je vztah mezi věkovou strukturou a závodem, což je přesně ona interpretace, k níž docházíme rozborem faktorového modelu. V další analýze, do níž se zde pouštět nebudeme, bychom se patrně zaměřili především na marginální tabulku věk vs. závod. Zcela opominout bychom však ani neměli určitý náznak souvislosti mezi věkem a výskytem onemocnění, kterou lze vyčíst jak z faktorového řešení, tak i z logaritmicko lineární analýzy (významnost efektu IZ a velmi dobrá shoda modelu ZV, IZ s daty). Konkrétní volba modelu v takovéto situaci závisí na výzkumné otázce.

V dalším příkladu se podíváme na možnosti analýz neúplných tabulek, tj. takových, kde nejsou všechna pole obsazena, popřípadě blokována.

Příklad 3: Dvojrozměrná tabulka, některá pole neobsazena

Z následující tabulky 4 by patrně byl leckterý výzkumník značně nešťastný, protože nejenom obsahuje mnoho polí s velmi nízkými empirickými četnostmi, ale i mnoho polí zcela nulových. Situaci komplikuje dále skutečnost, že nejde o tabulku pocházející z náhodného výběru, nýbrž o vyčerpávající údaje, a tudíž sotva o přijetí nebo nebo zamítnutí modelu nezávislosti. Předmětem tabelace je věk ženicha a nevěsty, vyjádřený v pětiletých intervalech, pro všechny sňatky v New Haven, v USA v roce 1948. Pro snazší čitelnost vynechávám v tabulce nulové četnosti.

Tabulka 4. Věk ženicha a nevěsty v pětiletých intervalech pro sňatky v New Haven, 1948

Věk ženicha	Věk nevěsty							
	15–19	20–24	25–29	30–34	35–39	40–44	45–49	nad 50
15–19	42	10	3					
20–24	153	504	51	10	1			
25–29	52	271	184	22	7	2		
30–34	5	52	87	69	13	5		
35–39	1	12	27	29	21	2	3	
40–44		1	9	18	17	8	2	1
45–49	1		3	6	16	16	7	1
nad 50			1	4	11	15	21	43

Je pochopitelné, že i méně zkušený analytik by z tabulky i bez pomoci matematicko-statistického aparátu leccos vyčetl, například skutečnost, že ačkoliv pro novomanžele je typická poměrná věková shoda (silně obsazená diagonála), přesto pánové mají větší možnosti výběru partnerky. Pokud bychom takovýchto tabulek měli z několika po sobě jdoucích let více, bylo by jistě zajímavé sledovat, nedochází-li k určitým posunům věkové struktury snoubenců v čase. To však by byla již jiná úloha, řešitelná jinými prostředky.

Nás na tomto místě ani tolik nezajímají sňatečná specifika amerického města krátce po 2. světové válce, jako spíše otázka, jak se s touto běžné analýze vzdorující tabulkou vyrovná předkládaná technika.

Výstupy 3, 4 a 5 a ilustrují výsledky po extrakci jednoho, dvou a šesti faktoru. Jednofaktorové řešení je převážně informativní a pomáhá zjistit, jak a nakolik jsou jednotlivé sloupcové a řádkové kategorie uspořádány. Podobně jako v předchozích výstupech jde o setříděné faktorové zátěže, přičemž tisk hodnot nižších než 0.25 je pro přehlednost potlačen. Předpona N před označením věkové kategorie značí nevěstu, Ž pak ženicha.

Výstup 3 Setříděné faktorové zátěže tabulky 4 – řešení s jedním faktorem.

	Faktor 1
N20–24	0.680
Ž20–24	0.625
Z25–29	0.541
N25–29	0.446
N15–19	0.371
Ž30–34	0.355
Ž30–34	0.293
Z15–19	0.0
N35–39	0.0
N40–44	0.0
N45–49	0.0
NNAD50	0.0
Ž35–39	0.0
Ž40–44	0.0
Ž45–49	0.0
ŽNAD50	0.0

Sňatečně „nejexponovanější" je skupina osob ve věku 20–24 let, respektive 25–29 let, které se ponejvíce žení a vdávají mezi sebou, případně s přispěním nevěst z řad teenagerů. Poměrně vysoká sňatkovost je ještě v kategoriích do 35 let, kde se může hypoteticky projevovat vliv odložených válečných sňatků. Podíl novomanželů vyššího věku, jakož i nejmladších ženichů je nízký. Kategorie jsou uspořádány způsobem, který by patrně nečinil při interpretaci velké obtíže.

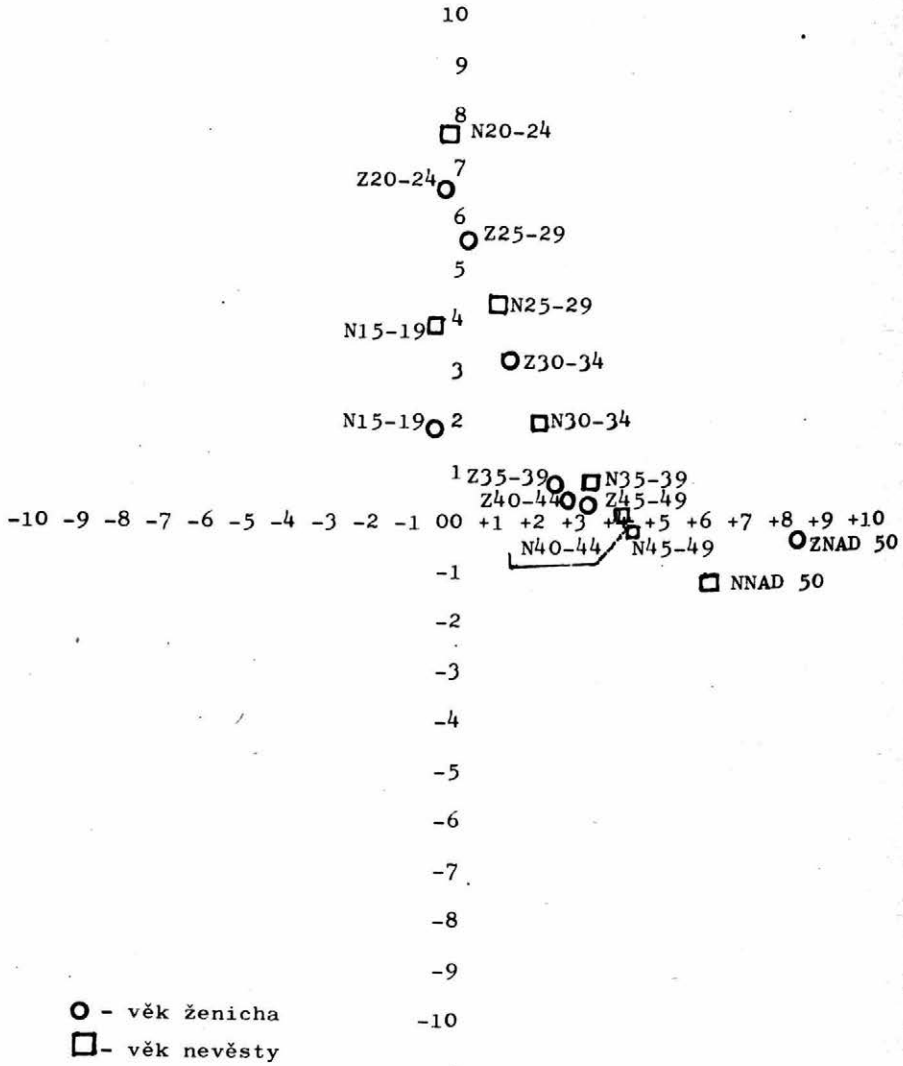
V dalších dvou výstupech si budeme ilustrovat vliv počtu faktorů, na „optiku", s níž na tabulku nazíráme. Nejprve Výstup 4 a Obrázek 7, uvádějící dvojfaktorové řešení.

Výstup 4: Setříděné faktorové zátěže tabulky 4 – řešení se dvěma faktory.

	Faktor 1	Faktor 2
N20–24	0.747	0.0
Ž20–24	0.698	0.0
Ž25–29	0.568	0.0
N25–29	0.435	0.0
N15–19	0.419	0.0
Ž30–34	0.313	0.0
ŽNAD50	0.0	0.815

NNAD50	0.0	0.626
N45–49	0.0	0.446
N40–44	0.0	0.403
Ž45–49	0.0	0.370
N35–39	0.0	0.359
Ž40–44	0.0	0.258
Ž15–19	0.0	0.0
Ž35–39	0.0	0.0
N30–34	0.0	0.0

Obr. 7. Faktorové řešení tabulky 4



Výstup 5. Setříděné faktorové zátěže tabulky 4 – řešení se šesti faktory.

	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5	Faktor 6
NNAD 50	0.938	0.0	0.0	0.0	0.0	0.0
ŽNAD 50	0.830	0.0	0.0	0.0	0.0	0.0
N20–24	0.0	0.894	0.0	0.0	0.0	0.0
Ž20–24	0.0	0.887	0.0	0.0	0.0	0.0
N30–34	0.0	0.0	0.770	0.0	0.0	0.0
Ž30–34	0.0	0.0	0.582	0.333	0.0	0.0
Ž35–39	0.0	0.0	0.518	0.0	0.0	0.0
N25–29	0.0	0.0	0.0	0.829	0.0	0.0
Ž25–29	0.0	0.0	0.0	0,787	0.0	0.0
Ž45–49	0.0	0.0	0.0	0.0	0.780	0.0
N40–44	0.0	0.0	0.0	0.0	0.653	0.0
N15–19	0.0	0.0	0.0	0.0	0.0	0.827
Ž15–19	0.0	0.0	0.393	0.0	0.0	0.809
Ž40–44	0.0	0.0	0.384	0.0	0.308	0.0
N35–39	0.0	0.0	0.387	0.0	0.466	0.0
N45–49	0.401	0.0	0.0	0.0	0.265	0.0

Dvojfaktorové řešení poskytuje pohled celistvější, jakoby z většího odstupu, a zřetelně z něj vyplývají specifika sňatečnosti mladších věkových skupin na jedné straně, a starších na straně druhé. Řešení využívající šesti faktorů je podrobnější a ukazuje například téměř deterministickou sňatečnou vázanost žen nad 50 let na své vrstevníky, přičemž opačný vztah je již mnohem slabší. Další faktory odhalují jemnější vazby uvnitř jednotlivých kohort, kde se v podstatě reprodukuje vysoká míra souladu věku snoubenců s tím, že muži si nezřídka berou ženy i nižších věkových kategorií. Střední věkové skupiny jsou poněkud nevyhraněné a jejich specifika by odkryl detail ve formě dalších jednoho nebo dvou faktorů.

Tento příklad, myslím, velmi dobře ilustruje situaci, kdy lze snadno analyzovat tabulku, v níž je závislost obou veličin zcela evidentní, přičemž striktní statistický test je nedostupný. Na její strukturu je možno se podívat řadou „objektivů" od širokoúhlého, až po teleobjektiv.

Nabízí se nyní otázka v jakém vztahu jsou výsledky získané touto technikou ve srovnání s dobře známou metodou LINDA-A. To bude předmětem následujícího příkladu.

Tabulka 5. Věková struktura podle profesí – počty osob

Profese	Věkové kategorie					
	do 20	do 30	do 40	do 50	do 60	nad 60
			let			
Strojní zámečník	386	2774	1730	1284	699	78
Soustružník	230	1575	518	346	107	4
Frézař	153	693	322	154	57	2
Brusič	55	475	287	222	102	16
Kovář	3	69	65	61	48	2
Slévač	12	125	111	97	49	3
Seřizovač	0	36	17	20	9	1
Údržbář	74	520	480	461	310	54

Příklad 4: Srovnání s metodou LINDA-A

V pracích [Loučková, Řehák 1984 a Řehák, Řeháková 1987] je metodou LINDA-A analyzována předchozí tabulka věkové struktury některých dělnických profesí.

Rotované dvoufaktorové řešení ukazuje Výstup 6:

Výstup 6: Setříděné faktorové zátěže tabulky 5

	Faktor 1	Faktor 2
Strojní		
zámečník	0.613	0.330
Údržbář	0.552	0.0
do 50	0.527	0.0
do 60	0.510	0.0
do 30	0.0	0.785
Soustružník	0.0	0.677
Seřizovač	0.0	0.0
Slévač	0.0	0.0
do 20	0.0	0.346
Kovář	0.0	0.0
do 40	0.464	0.0
Frézař	0.0	0.450
Brusič	0.0	0.0
do 60	0.0	0.0

Interpretace zde předkládaného řešení (které již neuvádím v grafické podobě) by patrně byla poněkud odlišná od té, kterou uvádějí Loučková, Řehák, resp. Řeháková, Řehák. Hlavní diferenciace (která, možná, následně způsobuje další, tzv. zavlečené odlišnosti ve faktorovém řešení) spočívá v nepotvrzené specifičnosti skupiny seřizovačů. LINDA-A v tehdejším stadiu vývoje nebyla zcela adekvátně připravena pracovat s nulovými četnostmi, a proto neobsazená kategorie nejmladších seřizovačů mohla způsobit ono vychýlení. V našem řešení není kategorie seřizovačů nijak mimořádná. Podle sdělení autorů je nový algoritmus LINDY na tuto okolnost již připraven.

3. Závěry

Popisovaná technika byla – prozatím experimentálně – úspěšně vyzkoušena na tabulkách až se šesti dimenzemi. Většina rozumných programů pro faktorovou analýzu umožňuje specifikaci počtu extrahovaných faktorů uživatelem, volbu výpočtu korelační matice bez odečítání průměrů a zejména schopnost pracovat se singulární korelační maticí (jedním z nemnohých populárních softwarových produktů, které poslední dvě možnosti nenabízejí, jsou programy SPSS, a to jak ve verzi X, tak i PC).

Charakteristiku použití faktorové analýzy kontigenčních tabulek lze shrnout:

– lze analyzovat tabulky s neobsazenými poli, včetně blokovaných nebo strukturálně neobsazených;

- není třeba ověřovat prakticky žádné výchozí předpoklady, obvyklé při testování hypotéz či běžné faktorové analýze (normalita rozložení apod.). Je pochopitelně nutné, aby hodnoty uváděné v tabulce byly měřeny na společné stupnici, a tak byla zajištěna srovnatelnost dat. Jinak je technika použitelná především tam, kde by tradiční metody testování hypotéz a výstavby modelů nebyly přiměřené anebo tam, kde by měly jenom malou informační hodnotu;
- je analyzována tabulka jako celek, nikoli jenom jedno pole proti všem ostatním jako při chí-kvadrátové dekompozici celkové asociace do individuálních složek;
- technika je příbuzná s modelem vícerozměrného škálování a její grafické i numerické výstupy nabízejí rychlou a kompaktní informaci o datech i tehdy, nelze-li okamžitě nalézt vyhovující faktorovou interpretaci;
- techniku lze nasadit i tam, kde tabulka není sestavena z dat pocházejících ze striktně nezávislého výběru (např. vícenásobné odpovědi);
- lze analyzovat prakticky jakoukoli ortogonální strukturu, nejenom četností, ale i měr, jako jsou průměry, různé skóry apod.;
- techniku není nutno aplikovat samostatně, lze ji využít i jako předběžný nástroj pro získání informací o struktuře složitější tabulky před testováním variantních logaritmicko lineárních modelů;
- ve srovnání s testy založenými na chí-kvadrátové statistice poměrem věrohodnosti v logaritmicko lineárních modelech, má diskutovaná technika „sklon" jednak zastírat význačné dobře interpretovatelné faktorové řešení i v těch případech, kdy test zamítá nulovou hypotézu o nezávislosti veličin, jednak vést řešení spíše k modelům, které nejlépe vystihují původní data, než k modelům, jež lze charakterizovat jako poslední přípustné zjednodušení, které bývá cílem výstavby explikačních modelů.

V žádném případě tato technika nemá za cíl nahrazovat existující a osvědčené metody analýzy kategoriálních dat. Trpí navíc přirozeně mnoha nectnostmi faktorové analýzy, jež jsou v literatuře jejími oponenty, stejně jako příznivci bohatě diskutovány. Jak již bylo naznačeno výše, matice normovaných součinů, vypočítávaná ze striktně deterministické datové struktury je nutně singulární, tj. její hodnost je nižší než součet počtu všech kategorií, což může některým algoritmům faktorové analýzy působit numerické obtíže. Podobně jako v tradiční faktorové analýze, ani zde doposud není k dispozici spolehlivě vodítko kolik faktorů má být extrahováno; lze proto doporučit volit více řešení a pak se rozhodnout pro interpretačně nejpřiléhavější. Minimum faktorů jsou 2, resp. počet dimenzí tabulky, maximální rozumný počet je omezen počtem kategorií nejpočetnější dimenze. Obvyklý limit počtu faktorů, daný velikostmi charakteristických čísel vektorů větších než 1, nevede vždy k uspokojivému řešení. Pro lepší interpretaci by bylo v budoucnu vhodné dořešit problém centrování dat tak, aby faktorové zátěže nebyly soustředěny kolem prvního kvadrantu.

Pro častá řešení tabulek podobných velikostí si může uživatel napsat jednoduchý program pro generování výchozí datové struktury anebo pro konverzi analyzovaných tabulek do formy přijatelné pro faktorovou analýzu.

Dodatek

Jak již bylo uvedeno v textu, analýzy byly prováděny programy BMDP, instalovanými ve výpočetních střediscích ČSAV v Praze. Při generování datové struktury byla využita skutečnost, že dočasnou součástí firemního programu se může stát i uživatelský kód v jazyce FORTRAN. V následujícím textu uvedu zadání dvou analýz pro dvojrozměrnou a třírozměr-

nou tabulku, včetně FORTRANských transformací, potřebných pro vytvoření vzorovací matice. Jde o verzi BMDP pro počítače IBM řady 370.

Tabulka uvedená v Příkladu 1.

	Příkazy	Komentář
// EXEC BIMEDT, PROG-BMDP4M		vyvolání procedury a programu
//FORT.SYSIN DD *		signál následování FORTRANu
	DO 1=2, NVAR	cyklus vynulování
	X(I)=0	vzorovacího vektoru
1	CONTINUE	
	IC=3	IC udává počet sloupců
	IPS=KASE/IC	přírůstkový faktor
	IF (MOD (KASE, IC). EQ. 0)	rektifikace přechodu na novou
	IPS=IPS–1	dimenzi
	ICP=KASE–IPS*IC+1	definice sloupcové pozice v matici
	IRP=IC+2+IPS	definice řádkové pozice v matici
	X(ICP)=1	dosazení „1" do sloupcové pozice
	X(IRP)=1	dosazení „1" do řádkové pozice
//SYSIN DD *		vstupní proud příkazů BMDP
/PROBLEM	TITLE IS LAMSER.	titulek (nepovinné)
/INPUT	CASES ARE 12.	12 = počet polí tabulky
	VARIABLES ARE 1.	vstupuje jediná veličina – četnost
	FORMAT IS FREE.	ve volném formátu
/VARIABLE	NAMES ARE CETNOST, SPOK, NEROZH, NESPOK, SV, SO, STRED, VS.	názvy veličin
	ADD=NEW.	celkem veličin = vstupující četnosti + vzorovací indikátory 0 a 1
	WEIGHT=CETNOST.	určení veličiny udávající četnost
/FACTOR	FORM=OCORR.	výpočet matice kolem počátku
/PRINT	FSCORE=0	potlačení tisku faktorových skórů (nepovinné)
/PLOT	FSCORE=0	potlačení grafu skórů (nepovinné)
/END		konec příkazů, následují data:
22 15 133 25 11 74 20 8 42 30 7 13		četnosti
/END		ukončení dat

Tabulka uvedená v Příkladu 2.

	Příkazy	Komentář
// EXEC BIMEDT, PROG=BMDP4F		
//TRANSF DD *		alternativní forma
	DO 1 I=2, NVAR	
	X(I)=0	
1	CONTINUE	
	IR=2	počet řádků
	IC=2	počet sloupců
	ICR=IR+IC	rektifikace přechodů na nové
	IMR=IR*IC	dimenze tabulky
	IF1=KASE/IC	přírůstkový faktor sloupců
	IF (MOD (KASE, IC). EQ. 0)	
	IF1=IF1–1	
	IF2=KASE/IMR	přírůstkový faktor „hloubky"
	IF (MOD (KASE, IMR). EQ. 0)	
	IF2=IF2–1	
	ICP=KASE–IC*IF1+1	

```
          IRP=IC+IF1+1-IF2*IR
          IDP=ICR+ICR+IF2+2          definice hloubkové pozice v matici
          X(ICP)=1
          X(IRP)=1
          X(IDP)=1
/SYSIN DD *
PROBLEM TITLE IS 'TABULKA HAVRANEK'.
INPUT     VARIABLES ARE 1.     FORMAT IS FREE.     CASES ARE 8.
VAR       NAMES ARE CETNOST, IA, IN, A, L, ' > 40', ' < 40 '.
          ADD=NEW.             WEIGHT=CETNOST.
FACTOR    FORM=OCORR.
          NUMBER=3.                                 požadovaný počet
                                                    faktorů (nepovinné)
PRINT     FSCORE=0.
PLOT      FSCORE=0.
END
2 6 786 12 189 7 145 7 69                           četnosti
END
```

Poznámky:

Fortranské příkazy využívají automatických veličin BMDP.X(I), kde index I poukazuje na pořadí veličiny v datovém vektoru (proto nulovací cyklus začíná od druhé veličiny, neboť první je načtená četnost).

NVAR je celkový konečný počet veličin v datovém vektoru, použito v ukončení cyklu.

KASE veličina indikující právě zpracovaný případ, pozorování, v našem případě pole tabulky.

Protože jsou i používány celočíselné operace, mají pomocné veličiny předponu „I", indikující v konvenci FORTRANu přirozené číslo.

Tisk a vykreslování faktorových skórů jsou potlačeny, protože zpravidla nemívají pro interpretaci význam.

Příkazy pro generování jsou pouze inspirativní. Potřebnou datovou strukturu lze generovat pomocí příkazů vnitřního jazyka i jiných statistických systémů (např. SAS, SYSTAT aj.) anebo kteréhokoli programovacího jazyka dostupného na tom kterém počítači (BASIC, PASCAL, C apod.).

Literatura

Bennedetti, J. K., Brown, M. B.: *Strategies for the selection of log-linear models.* Biometrics 34 (1978), str. 680–686.

Čakrt, M.: *Factor Analysis of N-Way Contingency Tables,* v Faulbaum, F. u. Ühlinger, H.-M. (Hrgs.): Fortschritte der Statistik Software 1, Gustav Fisher Verlag, Suttgart 1988, str. 27–38

Fisher, R. A.: *The Precision of Discriminant Funstions.* Ann. Eug. 10 (1949), str. 422–429

Hill, M. O.: *Correspondence Analysis: A Neglected Multivariate Method.* Applied Statistics 23 (1974), str. 340–354

Havránek, T.: *Třídění třetího a čtvrtého stupně – úlohy a metody.* V Řehák, J. (red.): Kapitoly ze sociologických metod a technik III. „Statistická analýza dat a kauzální závěry, ČSSS při ČSAV, Živohošť 1980

Lamser, V., Růžička, L.: *Základy statistiky pro sociology.* Praha Svoboda, 1970

Lauber, J., Linhart, J.: *Kanonická analýza kontingenčních tabulek.* Sociologický časopis XIV (1978), str. 610–618

Kendall, M. G., Stuart, A.: *The Advanced Theory of Statistics.* Vol 2, London, Charles Griffin & Co., Ltd., Londýn 1961

Mickey, M. R.: *Novel Use of BMD Programs: Canonical Correlation Analysis of Contingency Tables.* HSCF UCLA BMDP Technical Report No. 2, Los Angeles 1970

Řehák, J., Loučková, I.: *LINDA 1979 – Soubor programů pro analýzu reziduálních odchylek v dvojrozměrných uspořádáních dat.* In: Současný vývoj programů pro potřeby sociologie, Štiřín ČSS, sekce metod a technik s. 24–44.

Řehák, J., Loučková, I.: *Korespondenční analýza.* V Řehák, J. (red.) Kapitoly ze sociologických metod a technik IV: Škálování a kvantifikace, ČSSS při ČSAV, Mikulov 1981

Řehák, J., Loučková, I.: *Faktorová analýza v kontingenčních tabulkách; Metoda LINDA-A – popis a použití.* Sociologický časopis XX (1984), str. 174–192

Řehák, J., Řeháková, B.: *Analýza kontingenčních tabulek: rozlišení dvou základních typů úloh a znaménkové schéma.* Sociologický časopis XIV (1978), str. 619–631

Řehák, J., Řeháková, B.: *Analýza kategorizovaných dat v sociologii.* Praha, Academia 1987