Kdybychom se pokusili krátce zhodnotit dojmy, jež si ze semináře odnášíme, byl by na prvním místě obraz o výrazném rozvoji sociologické výzkumné práce v SSSR, a to v nejrůznějších institucích pod nejrůznějšími názvy a pochopitelně i se značně rozdílnou úrovní. Malá výstavka sociologické literatury vycházející v současné době v SSSR ukázala značné mezery v našich znalostech sovětské literatury. Obrovský zájem účastníků semináře o jakékoli sociologické texty však ukázal, že nebude snadné literaturu dostávat a vůbec bude těžké — zvláště do budoucna — si udržet přehled o vývoji sovětské sociologie jako celku.

Na druhém místě je vědomí o plodné spolupráci matematiků se sociology, i když i zde byla pociťována tradiční propast mezi čistě matematicky orientovanými modely a „aplikovanou matematikou", značná samorostlost a originálnost v řešení často tradičních metodologických úloh, provázená ovšem i nebezpečím neekonomických přístupů — zvláště při zatímní neexistenci základních metodologických příruček pro sociology v SSSR (mimo snad masově rozšířené leningradské publikace *Člověk a jeho práce*).

Výrazný dojem zanechala již zmíněná plánovitost práce Novosibirské skupiny, jež přímo provokuje ke srovnání s našimi Sociologickými ústavy (odděleními), v nichž celková náplň práce se podřizovala privátním zájmům jednotlivých badatelů, tedy byla řízena přesně opačným pojetím, než jaké jsme zaznamenali tam.

V závěrečném komuniké účastníci semináře vyjádřili přesvědčení, že použití matematiky je vnitřním rysem marxistické sociologie a že výměna mezinárodních zkušeností socialistických zemí na tomto poli pomůže vytvořit jediný marxisticko-leninský přístup k řešení základních problémů současné sociologie. Vyjádřili potřebu pořádat podobná shromáždění častěji (jednou za dva roky), vždy v jiné socialistické zemi, a za tím účelem vytvořit stálý metodologický komitét pověřený přípravou seminářů. Doporučili, aby byly vydány materiály semináře a aby byla organizována výměna materiálů z této oblasti mezi socialistickými zeměmi.

K období zvládnutí optimální výměny zkušeností v metodologii marxistické sociologie máme pravděpodobně ještě daleko. Je však již jisté, že novosibirským seminářem byly položeny solidní základy této spolupráce, a je potěšitelné, že k tomu přispěli i českoslovenští sociologové.

Zdeněk Šafář

Programy pro statistické vyhodnocení sociologických dat v Laboratoři počítacích strojů v Brně

V současné době existuje v Československu už poměrně značné množství programů pro statistické vyhodnocení sociologických dat. Sociologický časopis přinesl už informace o některých z nich[1] a tento článek je chce rozšířit zprávou o sadě programů, které jsou k dispozici v Laboratoři počítacích strojů VUT v Brně.

Sada programů vznikala postupně v letech 1966—1970, byla několikrát přepracována a její možnosti byly rozšiřovány. Při jejím vytváření bylo nutno volit mezi dvěma možnostmi: buď usilovat o vytvoření jednoho víceúčelového programu zdokonalováním výchozího základu vzhledem k určité žádoucí cílové vlastnosti (např. vzhledem k maximální rychlosti, a tedy ekonomičnosti prováděných výpočtů); anebo rozšiřovat původní program dalšími pomocnými programy tak, aby vznikla sada speciálních programů, jež ve variabilní sestavě odpovídají na určité, v sociologii často i velmi různorodé požadavky. Byla zvolena druhá varianta; pokusili jsme se vytvořit programovou stavebnici, ve které si původní díly zachovávají jednou již získané výhody a další díly umožňují dosáhnout např. vyšší rychlosti výpočtů v těch případech rutinních výzkumů, které pracují stereotypně s několika základními operacemi, anebo naopak umožňují speciální, jen zřídka vyžadované operace.

Celá sada programů byla vyvinuta pro počítač SAAB D21. Volba počítače byla dána skutečností, že je to nejvýkonnější počítač v Brně a patří v současné době k nejlepším počítačům střední velikosti v ČSSR. Kapacita vnitřní rychlé paměti, ve které se provádějí všechny výpočty, je 32 000 slov (slovo je např. jedno celé číslo). Maximální rychlost aritmetických operací ve vnitřní paměti je 100 000 operací/vt. K počítači je připojeno 8 mechanismů s magnetickými páskami jako vnější paměť počítače. Jedna magnetická páska je dlouhá 1 km. Při hustotě záznamu 4 slova na 1 mm pojme celkem 4 milióny slov, což představuje např. 40 000 dotazníků o 100 znacích. Magnetické pásky jsou vyměnitelné, takže kapacita vnější paměti je prakticky neomezená. Tisk výsledků se provádí na širokořádkové tiskárně, která má volitelný formát papíru, s maximem 132 typů na řádek. Maximální rychlost tisku je 20 řádků/vt. Výsledky je možno též děrovat maximální rychlostí 150 znaků/vt.

1. Snímání, uchování a kumulace vstupních informací

Vstup do počítače je z děrných štítků. Počítač je přečte a uloží na magnetickou pásku (= vnější paměť počítače), kde je má po celou dobu zpracování výzkumu uschovány. Rychlost snímání je max. 780 štítků za minutu. Data je možno snímat též z pětistopé nebo osmistopé pásky rychlostí max. 500 nebo 1000 znaků/vt.

Součástí stavebnice na snímání dat je několik pomocných programů, jež umožňují tyto operace:

[1] Srovnej: J. Lauber — Z. Šafář, *Možnosti zpracování sociologických dat na samočinných počítačích*, Filosofický časopis 2/1965, J. Linhart — Z. Šafář, *Programování a třídění statistických výpočtů pomocí samočinných počítačů*, Sociologický časopis 4/1967, J. Lauber, *Program pro statistické vyhodnocení sociologických dat*, Sociologický časopis 3/1969.

1. *Setřídění* několika děrných štítků pro případ, že informace z jednoho dotazníku přesahují kapacitu jednoho děrného štítku. Maximální možný počet znaků u jednoho respondenta je tisíc znaků, sejmutých nejvýše z dvaceti štítků. Každý znak může mít hodnotu od 0 do 2^{22}.
2. Ve vnější paměti počítače uložený *soubor dotazníků může být dodatečně doplňován* přidáním dalších dotazníků, což je značná výhoda např. tehdy, je-li výzkum prováděn v několika etapách v delším časovém období.
3. *Soubor dotazníků může být také redukován*, a to:
   a) náhodným výběrem — používáme zejména k redukci neekonomicky velkého souboru tam, kde je účelné pro některé operace použít jen náhodně vybrané části souboru;
   b) záměrným výběrem — používáme k vyloučení „chybných" dotazníků, tj. takových, které nesplňují určitou podmínku; programem ad 2 můžeme tyto dotazníky nahradit jinými.
4. *Doplněna* může být také *sada informací u kteréhokoli dotazníku* o libovolný počet znaků, což lze s výhodou použít zejména tam, kde teprve po zpracování některých znaků dotazníku kategorizujeme a kódujeme otevřené otázky a získáváme nové znaky, ale také tam, kde jsme se v druhé etapě šetření dotazovali téhož respondenta a chceme mít možnost třídit informace u téhož respondenta z první a druhé etapy dotazování navzájem. Dodatečné informace mohou být zapsány nejenom přidáním existujících informací, ale i tak, že některou položku vymažeme a nahradíme opraveným novým údajem (v případě, že jsme zjistili chybu v předchozím řetězu zpracování dat — při kódování, děrování apod.).

II. Třídicí programy

1. Třídění podle logických formulí — program MF 2 B

Tento program pracuje s logickými formulemi *rovností, nerovností* (větší než, menší než), *konjunkcí, disjunkcí a negací.* Uvedené jednoduché logické formule mohou tvořit složité logické formule. Každá logická formule nabývá hodnot *ano* nebo *ne* podle toho, jakou hodnotu pro tento dotazník mají odpovědi na otázky (znaky) vyskytující se v logické formuli. Jednoduchá logická formule rovnosti na příklad

$$5 = 1$$

vybere ze souboru všechny dotazníky, kde znak 5 má hodnotu (rovná se) 1. Logické formule oddělujeme čárkami, konec každého stupně lomítkem.
Zadání

$$5 = 1,\ 2,\ 5V2/$$

jsou tři logické formule. Pro splnění první formule musí mít znak s číslem 5 hodnotu 1, pro splnění druhé hodnotu 2 a pro splnění třetí musí mít hodnotu větší než 2.

Toto zadání je zadáním třídění prvního stupně — soubor se v tomto případě rozdělí do tří částí.
Třídění druhého stupně se zadáním

$$5 = 1,2,5V2/7 = 1,2,3,4/$$

vytvoří tuto tabulku:

	5 = 1	5 = 2	5 V2	Σ
7 = 1				
7 = 2				
7 = 3				
7 = 4				
Σ				

Zadání třetího stupně s nejjednoduššími logickými formulemi, které nekombinují ani znaky navzájem, ani hodnoty uvnitř znaků, má tento tvar:

$$5=1,2,5V2/7=1,2,3,4/12=1,2,3,4,5,6,7/$$

V tomto zadání se vytvářejí stejné tabulky jako nahoře, ale pro podsoubory 12=1, 12=2, . . . atd.
Logická formule však může být i velmi složitá a vybírat dotazníky podle určité kombinace hodnot několika znaků. Tak na příklad zápis

$$5=2.1V7.(2=1+3).N(3=3)$$

je formule, která nabývá hodnoty *ano*, jestliže platí, že
— znak 5 (např. pohlaví) má hodnotu 2
— a současně
— znak 1 (např. příjem) má hodnotu větší než 7
— a současně
— znak 2 (např. rodinný stav) má hodnotu 1 nebo 3
— znak 3 (např. srážky z platu, splátky) však současně nemá hodnotu 3.

Zdánlivě složitou logickou formuli lehce přečteme, nahradíme-li v našem případě číselné označení znaků a hodnot slovním popisem: u zvoleného příkladu nám program vybere ze souboru ty jedince, kteří:
— *jsou* muži (zápis 5=2)
— *a současně* mají příjem *větší než* 2300 (zápis .1V7)
— *a současně* jsou svobodní *nebo* rozvedení (zápis .(2=1+3)
— *současně však nemají* žádné srážky z platu (zápis .N(3=3)).

Při kombinování více znaků musí být vždy celý výraz, z něhož se provádí negace, uzavřen v závorkách a logický součin má přednost (prioritu) před logickým součtem. Chce-

me-li dosáhnout jiné priority, musíme výraz, který se má zhodnotit dříve, uzavřít do závorek. Např.

N(5=3.12V2+15=8+20M6),
(5=3+12V2).(15=8+20M6),

první formule není splněna, jestliže

(5=3 a současně 12V2)

nebo je 15=8

nebo je 20M6,

druhá formule je splněna, jestliže

5=3 nebo 12V2

a současně 15=2 nebo 20 M6.

Program vytváří tabulky tak, že vybírá z magnetické pásky jeden dotazník za druhým, vyhodnocuje pro zadanou tabulku logické formule a ve své vnitřní rychlé paměti zaplňuje políčka tabulky. Může pochopitelně nejen podle logických formulí přetřiďovat hodnoty jednoho nebo kombinace více znaků, ale může měnit pořadí hodnot, resp. uspořádat zvolené kombinace v libovolném pořadí, což lze výhodně použít, hledáme-li např. určitou originalitu v získané typologii. Program je velice pružný a je možné jeho pomocí vytřídit prakticky libovolnou tabulku neomezeného počtu logických kombinací všech variant všech znaků dotazníku. Jeho nevýhoda spočívá v tom, že pro jeho složitost není možno vytvářet na jeden průchod magnetické pásky vnější paměti více než jednu tabulku druhého stupně, i když v libovolném počtu hloubek třetího a dalších stupňů. Korespondence s vnější pamětí je totiž oproti ostatním operacím velmi pomalá. Druhá nevýhoda spočívá v tom, že vytváříme-li více tabulek, v nichž se některé logické formule opakují, vyhodnocují se tyto formule zbytečně vícekrát.

Jako každý jiný, tak i tento program odpovídá určitému pojetí sociologické analýzy. Jeho přednosti vyniknou zejména v pokročilejší fázi výzkumné práce, kdy už zkoumaný soubor známe natolik, že jsme schopni zadat třídění k testování složitějších, multideterminačních hypotéz. Takovéto počínání může být jistě kritizováno pro nutnou arbitrárnost a velkou míru předběžné teoretické úvahy, která bývá někdy označována jako sociologická spekulace. Zastánci tohoto způsobu analýzy však mohou — kromě meritorní obrany funkce teoretické úvahy a neredukovatelnosti osobnosti sociologa ve výzkumu — poukázat i na to, že tato část stavebnice záměrně v maximální míře využívá určitých schopností počítače, kdežto jiných, protikladných, využívá v sadě programů program MB (o něm viz dále), po kterém sáhneme, chceme-li přenechat i výběr hypotéz do jisté míry počítači.

2. Vytváření umělých proměnných — programy STAV a DOT 1

Popsané technické nevýhody programu MF 2B vedly ke snaze po jejich odstranění. Uvažovali jsme takto: je nutné vypracovat jednodušší program, aby bylo možno vytvářet více tabulek při jednom průchodu magnetické pásky dotazníků (vnější paměti) a odstranit zbytečné vyhodnocování opakujících se formulí. Tyto úvahy vedly v první řadě k programu na vytváření tzv. umělých proměnných (umělých znaků). Umělé znaky jsou nové znaky vytvořené z původních znaků dotazníku anebo z dříve vytvořených umělých znaků. Lze je vytvořit pomocí popsaných logických formulí programem STAV anebo pomocí aritmetických operací programem DOT 1. Logickými formulemi můžeme kombinovat znaky jednoho dotazníku, aritmetické operace nám dovolují hodnoty znaků v jednom dotazníku mezi sebou sečítat, odčítat, násobit a dělit. Výsledek se trvale přiřazuje jako hodnota nového znaku (tzv. „umělé proměnné") každému dotazníku ve zkoumaném souboru, podle toho, jak vyhoví podmínkám zadání.

Tak např. u zvoleného příkladu chceme rozdělení souboru na potenciální ženichy na jedné straně a na ostatní na druhé straně zachovat jako trvale užívaný třídící znak. Zadáme tedy vytvoření umělé proměnné na nejnižším neobsazeném čísle znaku — dejme tomu 91:

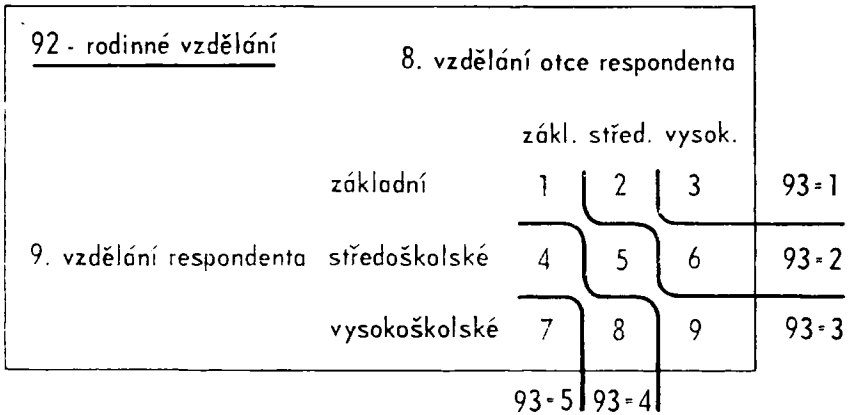
:91, 1:5=2.1V7. (2=1+3).N(3=3),

:2:5=1+1M8+(2=2)+(3=1+2+4+5),

což čteme: umělá proměnná 91 *nechť nabývá hodnoty 1*, když znak 5 je roven dvěma a současně znak 1 má hodnotu větší než sedm a současně znak dvě ... atd.; *nechť nabývá hodnoty 2*, když znak 1 má hodnotu menší než osm nebo ... atd. Před každou logickou formulí je mezi dvojtečkami hodnota umělého znaku, před první i číslo tohoto znaku.

Častěji než k trvalému zapsání výsledku třídění podle tak složitých logických formulí používáme vytváření umělých proměnných ke kategorizaci hodnot jednoho znaku: umožňuje nám to nejen sloučit málo četné hodnoty jednoho znaku, ale v praxi nás možnost vytvoření umělých proměnných vede zejména k tomu, že nám dovolí vyhnout se arbitrární kategorizaci otázek, které mají pro zpracování nepřijatelně velký počet alternativ. Tak např. věkové skupiny, podle kterých budeme třídit soubor, nemusíme určovat už v dotazníku, ale věk, zjištěný podle ročníku narození (99 variant), kategorizujeme teprve při zpracování do „přirozených" skupin podle toho, jak se nám chová jako nezávislá proměnná vzhledem k vybrané meritorní proměnné. Zejména při dichotomizaci výhodně stanovíme hranice kategorií na místě, kde se mění charakteristika korelace alternativ. Možnost kombinace hodnot několika znaků používáme nejenom k vytváření tzv. skóre (součtů hodnot sady znaků), ale umělá proměnná může být i průměrem hodnot několika znaků, resp. jakoukoli jinou hodnotou vzniklou aritmerickými operacemi s hodnotami znaků jednoho dotazníku. Násobení a dělení nám umožňuje zejména vážení hodnot přirozeného znaku zvoleným koeficientem a získání průměru jako hodnoty umělého znaku.

Umělá proměnná může také zachycovat určitou typologii v nominálních hodnotách umělého znaku. Tak např. kombinací znaků „vzdělání respondenta" (znak 8) a vzdělání otce respondenta (znak 9) získáme při třech hodnotách každého znaku umělou proměnnou 92 o devíti hodnotách pro všechny možné kombinace:



Z tohoto umělého znaku můžeme pak dál disjunkcí polí oddělených silnějšími přepážkami vytvořit další umělý znak *„(93) mobilita rodinného vzdělání"* s pěti hodnotami: *1. silně sestupná:* syn základní, otec vysokoškolské; *2. sestupná:* syn základní, otec středoškolské. nebo syn středoškolské, otec vysokoškolské; *3. žádná:* oba stejné vzdělání; *4. mírně vzestupná:* syn ... atd.

Zadání takové umělé proměnné (znak 93) zní:

:93, 1 :92=3, :2: 92=2+6, :3: 92=1+5+9 ... atd.

3. Třídění jedné hodnoty ze sady znaků do jedné tabulky

Kromě možnosti tvorby umělých znaků získal program MF 2B ještě jedno zlepšení, řešící triviální, ale v praxi dosti nepříjemný problém. V dotaznících se často setkáváme s výčtovou otázkou s možností výběru libovolného počtu alternativ. Běžný způsob zpracování, kdy každou alternativu třídíme zvlášť jako dichotomický znak, je velmi nevýhodný, neboť už vytřídění jedné takové výčtové otázky s deseti alternativami desetkrát zvyšuje počet tabulek u každé nezávisle proměnné, takže při jejich studiu se utápíme v hromadě papíru. Při další úpravě programu jsme vycházeli z předpokladu, že alternativy jedné otázky tvoří logickou sadu a při analýze obvykle potřebujeme srovnat relativní četnost kladných odpovědí v podsouborech podle nezávisle proměnné. Tak například nás zajímá, které nemoci z povolání jsou relativně četnější u veterinárních pracovníků v terénu a které jsou relativně četnější u veterinárních pracovníků v laboratoři. Program MF 2B nám umožňuje vybrat u každé nemoci pouze jedince s kladnou odpovědí a vytřídit výčtovou otázku s prakticky libovolným počtem variant do jedné přehledné tabulky.

Musíme si tu být vědomi, že počet dotazníků a počet voleb (onemocnění) není totožný. Řádkový součet v tabulce neudává celkový počet dotazníků v řádku, ale celkový počet udaných onemocnění (voleb); říkáme, že přináší *součtové četnosti podle tabulky.* Máme sice odpovědi na alternativní otázku v jedné tabulce, ale má to tu nevýhodu, že tabulka relativních četností řádků nám — stejně jako tabulka absolutních četností — ukazuje pouze, že u obou zaměstnaneckých kategorií je relativně nejčastější nemocí brucelóza, neumožňuje nám však zjistit, zda určitá nemoc se relativně častěji vyskytuje u terénních nebo u laboratorních pracovníků. (Viz tab. 1.)

Takovou analýzu můžeme udělat až po převedení absolutních četností na četnosti relativní, nikoli ovšem k celkovému počtu voleb

Tabulka 1

	Znak 25 = 1 ornitóza	Znak 26 = 1 toxoplasmóza	Znak 27 = 1 brucelóza	Prodělaných onemocnění celkem
Terénní pracovník (n = 300)	50 14 %	100 29 %	200 57 %	350 100 %
Laboratorní pracovník (n = 200)	40 14 %	100 33 %	160 53 %	300 100 %

(onemocnění), nýbrž relativní k celkovému počtu pracovníků dané kategorie. Potřebujeme tabulku:

	Znak 25 = 1 ornitóza	Znak 26 = 1 toxoplasmóza	Znak 27 = 1 brucelóza	Pracovníků celkem
Terénní pracovník	50 16 %	100 33 %	200 66 %	300 100 %
Laboratorní pracovník	40 20 %	100 50 %	160 80 %	200 100 %

z níž teprve vidíme, že ačkoli v absolutních hodnotách prodělalo brucelózu více terénních pracovníků a ačkoli je to nemoc v obou zaměstnaneckých kategoriích nejčastější, relativně častěji jsou touto nemocí postihováni laboratorní pracovníci. Takovou tabulku dostaneme, zadáme-li u programu MF 2B *součtové četnosti podle podmínky:* řádkové součty tu nejsou součty voleb, nýbrž součty jedinců v podsouboru (v podmínce), ke kterým jsou také počítány relativní četnosti.

4. Rychlý třídicí program MB

Programy dosud popsané nám umožňují upravit každý znak či kombinaci znaků tak, aby ve výsledku byly hodnoty uspořádány v žádoucí posloupnosti a do žádoucích tříd. Za těchto podmínek je možno uvažovat o programu, který počítá co nevíce tabulek na jeden průchod magnetické pásky a má jednoduché zadání. Rutinní sociologický výzkum nás často staví před problém, která třídění druhého stupně zadat ze všech možných kombinací znaků dotazníku. Víme, že už z dotazníku o padesáti znacích můžeme dostat

$$\frac{50^2 - 50}{2} = 1225$$

neopakujících se smysluplných tabulek druhého stupně. Z tohoto velkého množství tabulek je ovšem většina bezvýznamných; které to jsou, rozhodujeme na základě znalosti souboru a podle sledovaných hypotéz arbitrárně. K výběru jsme nuceni nejenom ohledy na ekonomičnost provozu, ale také proto, že se musíme vyhnout abundantnímu množství informací: je-li program dostatečně rychlý a naše ekonomické možnosti přiměřené, staví nás počítač při podrobnější analýze souboru nikoli před problém *jak získat* co nejvíce tabulek, ale naopak před problém *jak se vyhnout* tomu, abychom byli zahlceni informacemi a jak si vybrat co nejefektivněji z velkého množství možných tabulek právě ty, které jsou hodny naší pozornosti. SAAB D21 je výkonný počítač; další vývoj budované sady programů směřoval tedy nutně nikoli k rozšiřování možností při vytváření jedné tabulky, ale naopak usiloval o získání předběžné, redukované informace tak, abychom začínali pracovat nikoli s tabulkou jako určitým rozložením souboru do kombinací dvou znaků, nýbrž nejprve znali vztah mezi znaky, jenž se v tomto rozložení projevuje, vztah vyjádřený co nejúsporněji — jediným číslem — např. korelačním koeficientem.

Dalším logickým krokem ve vývoji sady programů byl tedy krok od tabulek četností k tabulkám koeficientů. Na úrovni třídění druhého a třetího stupně řeší problémy v tomto směru *program MB*. Počítá s rychlým vytvářením velkého množství tabulek a má velmi jednoduché zadání:

1=1:2, 2=2:5, 3=1:3, 4=0:4/ 1=1:2,
2=2:5, 3=1:3, 4=0:4, 5=1:2/'

Program vytváří tabulky ze všech kombinací dvojic znaků před prvním lomítkem a před druhým lomítkem o řádcích a sloupcích, které jsou vymezeny hodnotami před dvojtečkou a za dvojtečkou. Např. 2=2:5 zadává vytvoření 4 sloupců tabulky pro hodnoty 2, 3, 4, 5 znaku. 2.

Na uvedené zadání se vytvoří 10 tabulek

o sloupcích ze znaků	2	3	4	3	4	4	1	2	3	4
o řádcích ze znaků	1	1	1	2	2	3	5	5	5	5

Tabulky vytvořené jen v absolutních četnostech se ukládají dovnitř paměti počítače tak, že za sebou následují vždy tabulky z jednoho zadání a (bylo-li zadáno třídění třetího stupně) pro jeden podsoubor. Vznikne tak *matice tabulek*, jejíž sloupce tvoří znaky před prvním lomítkem a řádky znaky mezi prvním a druhým lomítkem. *Tabulky se netisknou ani se nepočítají tabulky relativních četností a koeficienty.* Jestliže se vyskytnou stejné znaky zadání pro řádky a v zadání pro sloupce, opakující se tabulky a tabulky na diagonále matice se nevytvářejí. Někdy samozřejmě není možné zapsat všechny požadované tabulky do jednoho zadání. Je-li zadání více a počet tabulek ze všech zadání je tak velký, že se nevejde do vnitřní paměti počítače pro jeden průchod magnetické pásky dotazníků, počítač sám rozdělí tabulky do skupin, které se mohou vytvořit na jeden průchod. Rozdělení zadání do skupin je optimální. Poněvadž program nepočítá relativní četnosti, koeficienty a netiskne tabulky, využívá maximálně vnitřní paměti a je velmi rychlý: na jeden průchod magnetické pásky dotazníků může vytvořit až 2500 tabulek 2×2.

V další práci s nimi postupujeme tak, že nejprve dáme povel k tisku *matice koeficien-*

tů, která nám slouží sama o sobě, nebo nám poskytuje orientaci při výběru tabulek, se kterými budeme dále pracovat. Matice koeficientů má tvar tzv. praporu:

SPEARMANŮV KOEFICIENT POŘADOVÉ KORELACE

	000 : 000							
	019	022	111	155	130	131	118	175
019 =		+0.159	−0.227	+0.005	−0.142	−0.068	+0.196	+0.191
022 =	+0.159		−0.090	−0.122	−0.122	−0.020	+0.126	+0.164
111 =	−0.227	−0.090		−0.037	+0.096	−0.002	−0.140	−0.008
155 =	+0.005	−0.122	−0.037		+0.472	−0.285	−0.145	−0.044
130 =	−0.142	−0.122	+0.096	+0.472		+0.231	−0.070	−0.115
131 =	−0.068	−0.020	−0.002	−0.285	+0.231		+0.105	−0.101
118 =	+0.196	+0.126	−0.140	−0.145	−0.070	+0.105		+0.391
175 =	+0.191	+0.164	−0.008	−0.044	−0.115	−0.101	+0.391	

Zadání pro třídění třetího stupně zapisujeme opět mezi druhé a třetí lomítko

...../...../14=1 : 7, 20=1 : 2/

a program vytváří matici koeficientů (prapor) pro podsoubor každé hodnoty třídícího znaku třetí hloubky od hodnoty před a po hodnotu za dvojtečkou — u zvoleného příkladu sedm praporů pro podsoubory podle hodnot znaku 14 a dva pro podsoubory podle hodnot znaku 20.

Hlavní výhodou tohoto programu je, že vytváří rychle velké množství tabulek, jež netiskne, ale dává nám o nich informaci maticí koeficientů. U jednoduššího výzkumu nám to umožňuje třídit na úrovni druhého stupně až všechny znaky navzájem, takže se můžeme vyhnout arbitrární volbě sledovaných vztahů a máme jistotu, že naší pozornosti neunikne nepředpokládaný vztah. V sadě ordinálních či kardinálních proměnných je při náročnějším výzkumu matice korelací podkladem pro řadu dalších operací, od hledání zástupných indikátorů a vytváření stupnic až po faktorovou analýzu. Za vysokou produktivitu při vytváření tabulek platíme dvěma omezeními: první spočívá v tom, že můžeme pracovat jen se znaky o omezeném rozsahu hodnot (u znaků pro sloupce do dvaceti, u znaků pro řádky do padesáti). Hodnoty přitom musí být uspořádány za sebou, nelze je přehazovat, slučovat ani vypouštět (kromě krajních hodnot, které vypouštět lze). Této nevýhodě čelíme tím, že si znaky nejdříve upravíme programem pro vytváření umělých znaků STAV a DOT 1. Druhou základní nevýhodou je, že program MB třídí soubor pouze do třetího stupně. I tomu lze odpomoci do jisté míry vytvářením umělých znaků z kombinací znaků pro třetí a čtvrtý, resp. vyšší stupně. Daleko lépe však odstraňuje tuto nevýhodu další program z popisované sady.

5. Hloubkový průzkum podle podsouborů — program HP

Víme, že třídění druhého stupně nás často klame: neupozorní nás na falešné korelace a existenci intervenujících proměnných. Určitým příspěvkem k řešení tohoto problému je program HP. Vychází ze skutečnosti, že dva znaky, které vykazují v třídění druhého stupně nízkou korelaci, mohou už v některém podsouboru vyššího stupně souviset velmi těsně a míra souvislosti se může měnit vlivem intervenující proměnné. Sledovat tyto změny tak, že budeme srovnávat hodnotu určitého korelačního koeficientu v matici koeficientů druhého stupně a v sadě matic třetího stupně, vytvořených podle intervenující proměnné programem MB, je velmi nepřehledné a pracné. Program HP pracuje na stejném principu jako program MB: i on vytváří tabulky jen v absolutních četnostech, vynechává opakující se tabulky, ukládá je nejprve do paměti a tiskne až na povel. I on vychází z jednoduchého zadání o těchže limitech. Na rozdíl od programu MB však může vytvářet podsoubory až do šestého stupně a neuspořádává tabulky do matice, ale vytváří za sebou tytéž tabulky pro všechny zadané podsoubory. Z těchto tabulek se programem HL TISK na řádky pod sebe tisknou vybrané koeficienty. Tak např. zadání

12 = 0 : 3, 15 = 1 : 2/
1 = 2 : 6, 4 = 1 : 4, 5 = 1 : 8/
120 = 1 : 3, 122 = 0 : 1/
60 = 1 : 2/
82 = 1 : 4

vytváří a na povel tiskne do řádků zvolené koeficienty tabulky 12 = 0 : 3 / 1 = 2 : 6 pro celý soubor

pro 120 = 1
pro (120 = 1) . (60 = 1)
pro (120 = 1) . (60 = 1) . (82 = 1)
pro (120 = 1) . (60 = 1) . (82 = 2)
pro (120 = 1) . (60 = 1) . (82 = 3)
pro (120 = 1) . (60 = 1) . (82 = 4)
pro (120 = 1) . (60 = 2)
pro (120 = 1) . (60 = 2) . (82 = 1)
.
.
. atd.

Tisk tabulek a koeficientů u třídících programů a jejich grafická úprava

Program MB tabulky utváří a uchovává v paměti. Program PRAP z nich počítá a tiskne matice statistických koeficientů,[2] a to matici:

a) chí-kvadrátů,
b) hladin významnosti chí-kvadrátů,
c) Čuprovova koeficientu kontingence,
d) Pearsonova normalizovaného koeficientu kontingence,
e) Spearmanova koeficientu pořadové korelace,
f) testu Spearmanova koeficientu

podle výběru určeného zadáním. Nekorektní[3] chí-kvadrát a koeficienty z něho odvozené označuje hvězdičkou. Program TISK TAB tiskne

a) všechny vytvořené tabulky,
b) jen ty tabulky z matice, které byly zadány pořadovým číslem.

Každá tabulka může mít tento obsah:

1. tabulku absolutních četností
2. tabulku relativních četnosti
3. tabulku relativních četností řádkových
4. tabulku relativních četností sloupcových
5. tabulku korelace alternativ
6. tabulku znaménkového testu
7. chí-kvadrát
8. hladinu významnosti pro chí-kvadrát ($<0{,}1\ \%$, $<1\ \%$, $<5\ \%$, $<10\ \%$, $>10\ \% = *$)
9. Čuprovův koeficient kontingence
10. Pearsonův koeficient pořadové korelace
11. Spearmanův koeficient pořadové korelace
12. test Spearmanova koeficientu

Abychom se vyhnuli zahlcení daty, program obsahuje možnost vypustit při tisku z obsahu tabulky kteroukoli z výše uvedených položek. Také grafické úpravě tabulky byla proto věnována velká péče. Aby bylo možno tabulky z pásu papíru roztrhat na naznačené listy a stáhnout je rychlovazačem v knihu, začíná každá tabulka na začátku listu. Tabulka o pěti řádcích je při zadání celého obsahu rozvržena na celý list počítačového papíru; do deseti sloupců se tiskne na úzký papír, od deseti do dvaceti na široký a při větším počtu sloupců se musí dělit. Pokud má tabulka více než pět hodnot v řádcích a nechceme vypouštět nic z obsahu, přesáhne na další list na pásu, nová tabulka však opět začíná na začátku listu (viz Tabulka 2).

Program MF 2B tiskne tabulky hned když je vytváří. Je-li v zadání, tiskne i slovní označení tabulky (hlavičku až do 62 písmen). Jinak pro obsah a grafickou úpravu tabulky platí totéž, co jsme uvedli u programu MB. Jak jsme už vysvětlili, relativní četnosti řádků mohou být počítány podle tabulky nebo podle podmínky.

Program HP tiskne u zadaného souboru a podsouboru do 6. hloubky do řádků všechny koeficienty, jež jsme uváděli u programů MB i MF 2B. Podle naší volby lze kterýkoli koeficient vypustit.
Celé tabulky lze vytisknout tiskovým programem TISK TAB uvedeným u MB.
Jako *kontrolní tisk* je možno zadat tisk všech dotazníků nebo tisk kterékoli položky ze všech dotazníků.

Ekonomika provozu

Doba všech výpočtů je ovlivněna vždy počtem dotazníků a počtem znaků v dotazníku, v některých případech i jinými okolnostmi (např. složitostí logické formule, rozměry tabulek atd.).

---

[2] Dále uváděné statistické koeficienty byly převzaty z těchto pramenů:
Jaroslav Janko, *Statistické tabulky*, Praha 1958 (chí-kvadrát, hladina významnosti).
J. Linhart – Z. Šafář: citovaný článek (Pearsonův normalizovaný koeficient kontingence C_{norm}; znaménkový test),
K. Szaniawski, *Metody statystyczne w sociologii*, Warszava 1968 (Čuprovův koeficient kontingence; viz též Linhart, Šafář, cit. článek),
R. Reisenauer, *Metody matematické statistiky*, Praha 1965 (test Spearmanova koeficientu).
Spearmanův koeficient pořadové korelace je počítán podle vzorce:

$$RS = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma d^2}{2 \cdot \sqrt{\Sigma x^2 \cdot \Sigma y^2}},$$

kde

$$\Sigma x^2 = \frac{N^3 - \sum_j n_{ij}^3}{12} \qquad \Sigma y^2 = \frac{N^3 - \sum_i n_{ij}^3}{12}$$

$$\Sigma d^2 = \sum_i \sum_j (P_i - P_j)^2 \cdot n_{ij},$$

přičemž průměrné pořadí řádků

$$P_i = \sum_{k=1}^{i-1} \sum_j n_{kj} + \frac{\sum_j n_{ij} + 1}{2}$$

průměrné pořadí sloupců

$$P_j = \sum_{l=1}^{j-1} \sum_i n_{il} + \frac{\sum_i n_{ij} + 1}{2}$$

(upraveno podle K. Ctiborský, *Typový program pro matematickostatistickou analýzu výsledků sociologického výzkumu*, rozmnožený rukopis).

Korelace alternativ je P. Oseckým upravený asociační koeficient φ, který se vypočítává pro čtyřpolní tabulky vzniklé redukcemi tabulky o m řádcích a n sloupcích:

$$S_{ij} = \frac{n_{ij} \cdot N - \sum_i n_{ij} \cdot \sum_j n_{ij}}{\sqrt{\sum_i n_{ij} \cdot \sum_j n_{ij} \, (N - \sum_i n_{ij}) \, (N - \sum_j n_{ij})}}$$

kde

$$N = \sum_i \sum_j n_{ij}$$

[3] Za nekorektní je označen Chí-kvadrát a koeficienty z něho odvozené, jestliže a) více než 20 % všech očekávaných četností je menší než 5; b) jestliže v jednom políčku kontingenční tabulky je očekávaná četnost menší než 1. (Podle R. Reisenauer, citované dílo.)

Tabulka 2

118*174
(111=002)

ABS. CETNOST

	01	02	03	04	05	06	
001	16	54	25	14	42	5	156
002	5	37	20	11	30	1	104
003	7	22	14	13	28	5	89
004	1	23	13	4	23	6	70
	29	136	72	42	123	17	419

REL. CETNOST

	01	02	03	04	05	06	
001	3.8	12.9	6.0	3.3	10.0	1.2	37.2
002	1.2	8.8	4.8	2.6	7.2	0.2	24.8
003	1.7	3.3	3.3	3.1	6.7	1.2	21.2
004	0.2	5.5	3.1	1.0	5.5	1.4	16.7
	6.9	32.5	17.2	10.0	29.4	4.1	100.0

REL. CETN. R.

	01	02	03	04	05	06	
001	10.3	34.6	16.0	9.0	26.9	3.2	100.0
002	4.8	35.6	19.2	10.6	28.8	1.0	100.0
003	7.9	24.7	15.7	14.6	31.5	5.6	100.0
004	1.4	32.9	18.6	5.7	32.9	8.6	100.0
	6.9	32.5	17.2	10.0	29.4	4.1	100.0

REL. CETN. S.

	01	02	03	04	05	06	
001	55-3	39.7	34.7	33.3	34.1	29.4	37.2
002	17.2	27.2	27.8	26.2	24.4	5.9	24.8
003	24.1	16.2	19.4	31.0	22.8	29.4	21.2
004	3.4	16.9	18.1	9.5	18.7	35.3	16.7
	100.0	100.0	100.0	100.0	100.0	100.0	100.0

KOR. ALTERNATIV

	01	02	03	04	05	06
001	0.10	0.04	−0.02	−0.03	−0.04	−0.03
002	−0.05	0.04	0.03	0.01	−0.01	−0.09
003	0.02	−0.09	−0.02	0.08	0.02	0.04
004	−0.10	0.00	0.02	−0.06	0.03	0.10

ZNAMENKOVY TEST

	01	03	04	05	06
001	+ +				
002					−
003			+		
004	− −				+ +

CHI = 19.990 HV = * SP = +0.121 CU = 0.111 PE = 0.246

Zvolme si tedy soubor o 1000 dotaznících a 100 znacích a cenu strojového času 1950,— Kčs/hod (užívanou v Brně).

Pak 1 tabulka o rozměrech 5×5 vytvořená programem MF 2B obsahující pouze jednoduché logické formule trvá 28 vt. a stojí 15,18 Kčs pro jeden podsoubor (63 vt.; 34,15 Kčs pro 10 podsouborů). Vyskytuje-li se v tabulce složitá logická formule, čas na její výpočet se prodlužuje průměrně čtyřikrát.

Vytvoření umělých znaků pomocí 50 jednoduchých logických formulí trvá 80 vt. a stojí 43,36 Kčs. Vytvoření umělých znaků pomocí 50 aritmetických operací trvá 37 vt. a stojí 20,05 Kčs.

Vytváření umělých znaků závisí velice silně na počtu průchodů magnetické pásky. Při jednom průchodu magnetické pásky nezávisí prakticky na počtu aritmetických operací, ale závisí na počtu logických formulí. (Vytvoření 1 umělého znaku pomocí 1 jednoduché logické formule trvá 80 vt. a stojí tedy 43,36 Kčs.)

V programu MB při plně využité kapacitě paměti připadne na vytvoření 1 tabulky o rozměrech 2×2 0,3 vt., tj. 0,11 Kčs. Její vytištění trvá asi 3 vt. a stojí tedy 1.63 Kčs. (Pro tabulku 5×5 je to: 0,9 vt. — 0,49 Kčs, 4,3 vt. — 2,33 Kčs.)

Výpočet tabulek pomocí programu HP při max. využití paměti:
Tab. 2×2 — vytvoření 0.6 vt., cena 0,33 Kčs.
Tab. 5×5 — vytvoření 1,3 vt.. cena 0,71 Kčs.

Tisk matice koeficientů o rozměrech 10×10 je hotov za 17 vt. a stojí tedy 9,21 Kčs.

U rychlých programů se výkon počítače dotkl nepředvídaného limitu: Program ve speciálních případech i pro velký počet dotazníků počítá a tiskne tabulku pod cenu papíru se třemi kopiemi. V praxi se cena tabulky poněkud zvyšuje o podíl na nákladech se snímáním souboru.

III. Programy pro statistické operace s kardinálními znaky

Mezi sociology existují autoři, kteří zastávají názor, že v sociologii prakticky nemáme pravé kardinální proměnné. Mít plat Kčs 3500.— měsíčně není prostě více než mít plat Kčs 980,—; je to něco jiného. Nanejvýš lze předpokládat ordinalitu; předpokládat kardinalitu kteréhokoli sociologického znaku je věc konvence. Nicméně je to užitečná konvence, která umožňuje četné statistické operace. Proto také byla vypracována sada programů pro operace s kardinálními znaky.

1. *Program PAV*
   provádí analýzu tabulky, v níž aspoň jeden znak je kardinální. Vychází z tabulky absolutních četností, spočítané programy MB nebo HP, přičemž počítá se skutečnými hodnotami znaků ve sloupcích a řádcích. Pro každý řádek a sloupec tabulky počítá aritmetický průměr, jeho přípustnou chybu, rozptyl jako statistickou veličinu, relativní odchylku, variační koeficient, počet objektů sloupce nebo řádku a teoretický počet dotazníků, který by byl nutný pro odhad průměru s 95% spolehlivostí a s přesností na 5 %.
   Dále se provádí jednoduchá analýza rozptylu tabulky. Počítají se korelační poměry, rozptyly uvnitř řádků a sloupců a mezi nimi a pro celou tabulku s příslušnými stupni volností, průměrné „součty čtverců", testovací hodnota pro poměr rozptylů a variační koeficienty uvnitř skupin, mezi sloupci a celkem. Nakonec se provádí výpočet součinového korelačního koeficientu a jeho test.
2. *Program CHAR 1*
   počítá a tiskne ze souboru nebo podsouboru daného variantou třídícího znaku aritmetický průměr, jeho interval spolehlivosti, modus, medián, směrodatnou odchylku výběrovou, odhad směrodatné odchylky základního souboru, její interval spolehlivosti, tytéž hodnoty pro rozptyl, variační koeficient, teoretický nutný počet objektů po dosažení zadané přesnosti aritmetického průměru, strmost rozložení a jeho asymetrii. Mezi zadanými znaky provádí program po dvojicích F-test shody rozptylů a t-test shody aritmetických průměrů. Pravděpodobnostní hladina se volí 95% nebo 99%.
3. *Program CHAR 2*
   porovnává skutečné rozložení jevu se zadaným teoretickým rozložením. Pracuje za stejných podmínek jako CHAR 1 a navíc umožňuje u každého znaku před testem shody provést libovolnou transformaci znaku pomocí aritmetického výrazu zadaného z děrné pásky. Test shody je možno provést s těmito teoretickými rozloženími: normální, logaritmicko-normální, Poisonovo, binomické.
4. *Program KOREL*
   zkoumá lineární závislost jednoho znaku na jiných zadaných znacích ve zvoleném podsouboru. Počítá a tiskne: aritmetické průměry, směrodatné odchylky, matici regresních a korelačních koeficientů a lineární závislost ve formě rovnice.

*

Jsme si dobře vědomi toho, že všechny zde uvedené programy mají výrazné limity. Při jejich výstavbě jsme sledovali také myšlenku úspornosti. Usilovali jsme o to vyhnout se abundantním kvalitám, abychom ušetřenou kapacitu investovali tam, kde se jí skutečně používá. V tom smyslu také chtěl náš program být nikoli tak „dokonalý", jak jen možno, ale spíše tak „nedokonalý", jak jen možno. V žádném případě není tato sada programů uzavřeným celkem. Dále se vyvíjí a je doplňována. V současné době už pracují také programy pro analýzu časových řad a pro faktorovou analýzu. Informace o nich by však neúnosně rozšířila tento článek; posečkáme s ní, až bude i s těmito programy více zkušeností.

František Krejčí - Ivo Možný

Doprava a společnost

Koncem minulého roku uspořádal výbor Československé vědeckotechnické společnosti pro dopravu a spoje ve spolupráci s Výzkumným ústavem dopravním v Praze sympozium s me-