Programování třídění a statistických výpočtů pomocí samočinných počítačů

JIŘÍ LINHART — ZDENĚK ŠAFÁŘ
Fakulta všeobecného lékařství UK

Již téměř před třemi lety byly na katedře DHM FVL UK učiněny první kroky ve využití samočinného počítače pro zpracování dat ze sociologického průzkumu.[1] Zkušenosti se během doby vytříbily a proto považujeme za užitečné o nich opět čtenáře informovat.

Předkládáme zde zadání programu pro zpracování zatím u nás nejzastoupenějších typů sociologických průzkumů a to ve formě, která může sloužit téměř kterémukoliv zručnějšímu programátorovi, aby uvedený postup adaptoval pro konkrétní typ počítače i pro značně široké spektrum sociologických dat. Naše řešení vychází z čistě praktických manipulačních zkušeností se sociologickými daty — chce neuspořádaná vstupní data (např. odpovědi na otázky v dotazníku) dostat do formy, ze které by se tato data dala poměrně jednoduše interpretovat. Domníváme se, že námi uváděné zadání je sice přibližným, ale nicméně praktickým řešením problémů, které má řada pracovníků-sociologů se shromážděným materiálem. Doufáme, že uvedením programu vyvoláme diskusi a výměnu zkušeností na tomto poli, a zároveň poskytneme praktickou pomůcku pro současnou sociologickou praxi.

Zadání programu

Vstupní data: cca „n" dotazníků (záznamových archů apod.) s „x" otázkami (typem znaků), které mohou nabývat jen jedné z hodnot 1, 2, 3 . . . j.

Počet dotazníků a počet otázek, které je možno bez obtíží zpracovat, závisí na kapacitě paměti počítače — např. 5000 dotazníků o 150 otázkách zpravidla nečiní obtíží. Počet možných alternativ v jednotlivých otázkách je nejlépe omezit na devět, desátá alternativa je nula.

1. úloha — třídění I. stupně

Roztřídit zpracovávaný soubor o rozsahu „n" (počet dotazníků apod.) podle otázky (znaku) A-té na „r" tříd (hodnot znaku) a vytisknout absolutní a relativní četnosti. Zjistit počet hodnot 0. Tisknout jen tolik hodnot, kolik má A-tá otázka možných alternativ odpovědí. Úlohu opakovat pro všech „x" otázek. Na formě tisku nezáleží.

Tedy vytisknout: otázka A-tá:

třída (hodnota) znaku 1 . . .	a dotazníků	tj.	a'%
třída (hodnota) znaku 2 . . .	b dotazníků	tj.	b'%
třída (hodnota) znaku . . .			
třída (hodnota) znaku r . . .	p dotazníků	tj.	p'%
třída (hodnota) znaku 0 . . .	q dotazníků	tj.	q'%

Relativní četnosti vypočítat podle vzorce $\frac{w_j \cdot 100}{n}$, ve kterém „n" je počet všech dotazníků (rozsah souboru), „w_j" je počet prvků zahrnutých do j-té třídy znaku A-tého.

Zřejmě dále a+b+ . . .p+q=n, a'+b'+ + . . . +p'+q'=100 %.

2. úloha — třídění II. stupně

Soubor o rozsahu „n" roztřídit podle otázky (znaku) A-té na „r" tříd $A_1, A_2, \ldots A_r$, a zároveň podle otázky (znaku) B-té na „s" tříd $B_1, B_2, \ldots B_s$, tedy celkem na „r" . „s" skupin, tak, aby každý prvek souboru byl zařazen do jedné skupiny. Nepracovat s nulovými třídami znaku A-tého a B-tého.

Třídění provést u otázek: viz příloha 1 (může být doplňováno postupně, programátorovi je vhodné uvést předběžný předpokládaný počet zadávaných tabulek).

[1] Lauber J.–Šafář Z.: *Možnosti zpracování sociologických dat na samočinných počítačích*, Filosofický časopis 2, 1965.

Výsledky třídění sestavit do tabulky (matice) o dvojím vchodu, která se často nazývá kontigenční tabulkou. (Viz tab. 1.)

Zavedeme přílohu 2, do které pracovník, který provádí průzkum, vypíše programátorovi zadání třídění III. stupně.

Tabulka 1

Otázka A-tá / Otázka B-tá	$A_1, A_2 \ldots A_r$	Součty
B_1	$n_{11}, n_{12} \ldots n_{1r}$	b_1
B_2	$n_{21}, n_{22} \ldots n_{2r}$	b_2
.	. . . .	.
B_s	$n_{s1}, n_{s2} \ldots n_{sr}$	b_s
Součty	$a_1 \quad a_2 \ldots a_r$	n

V tabulce značí n_{11}, n_{12}, ...n_{rs} (obecně n_{ij}) počet prvků, u kterých byla zjištěna i-tá třída znaku A-tého a j-tá třída znaku B-tého.

Nepracovat s nulovými třídami znaků A-tého a B-tého, pouze pod tabulkou uvést jejich počet.

3. úloha — třídění III. stupně

Soubor o rozsahu „n" roztřídit podle otázky A-té na „r" tříd, podle otázky B-té na „s" tříd a podle otázky C-té na „k" tříd, tedy celkem na „r" . „s" . „k" skupin tak, aby každý prvek souboru byl zařazen do jedné skupiny. Opět nepracovat s nulovými třídami všech znaků, pouze vytisknout počet všech jedinců, kteří na některou otázku neodpověděli.

Výsledky třídění sestavit do „k" kontingenčních tabulek (v případě, že změníme postup třídění, tak do „r" nebo do „s" tabulek).

Vyhneme se složitému formálnímu vyjádření obsahu třídění III. stupně, neboť jeho podstata je shodná se strukturou třídění II. stupně. Nejprve předtřídíme celý soubor „n" jedinců do souborů c_1, c_2, ...c_k , a tyto soubory potom třídíme stejně, jak je naznačeno v tabulce 1 — forma výstupních tabulek bude totiž obdobná, každý prvek tabulky bychom označili pouze c_k n_{ij}, původní prvky n_{ij} by se objevily v součtových marginálních jednotlivých tabulek, celková suma „n" v každé tabulce je nahrazena počtem prvků v jednotlivých třídách znaku C-tého.

4. úloha — třídění IV. stupně a vyšších stupňů

Soubor o rozsahu „n" roztřídit podle otázky A-té na „r" tříd, podle otázky B-té na „s" tříd, podle otázky C-té na „k" tříd, podle otázky D-té na „g" tříd, tedy celkem na „r" . „s" . „k" . . „g" skupin tak, aby každý prvek souboru byl zařazen. Opět nepracovat s těmi jedinci, u kterých se v některé otázce vyskytla nulová odpověď.

Výsledky jsme tentokrát již nuceni sestavit do „k" . „g" kontingenčních tabulek (popř. do počtu tabulek odpovídajícího jinému násobku počtu tříd dvou ze čtyř tříděných znaků). Při třídění vyšších stupňů opakujeme týž postup pro více než čtyři otázky, logika třídění zůstává stejná.

Programátorovi oznámíme, do kterého stupně třídění chceme materiál zpracovávat a zavedeme přílohu 2, ve které vypíšeme zadání pro vyšší stupně třídění.[2]

Statistické výpočty

1. Výpočet procent pro třídění I. stupně viz úloha 1.

2. Výpočet tzv. očekávaných četností pro kontingenční tabulku 1.

Vypočítej pro skupinu prvků n_{ij} z tabulky 1 očekávanou četnost o_{ij} ze vzorce

$$[1] \qquad o_{ij} = \frac{a_i \cdot b_j}{n}$$

ve kterém o_{ij} je očekávaná četnost odpovídající příslušné skupině, n_{ij}, a_i je počet prvků zařazených do i-té třídy zna-

[2] V případě, že chceme pouze některé z tabulek třídění vyšších stupňů, je lépe vytisknout všechny tabulky příslušného třídění a z vytištěného si vybírat až při interpretaci.

[3] Matice očekávaných četností představuje model nulové hypotézy: mezi znaky A a B není žádný vztah — koeficient kontingence této matice je roven 0.

ku A-tého, b_j je počet prvků zařazených do j-té třídy znaku B-tého, „n" je počet prvků zařazených do všech tříd znaku A-tého a B-tého mimo třídu 0. Očekávané četnosti vypočítej pro všechny skupiny n_{ij} a výsledky sestav do tabulky identické s tabulkou 1.[3]

Výpočet testovacího kritéria $\varkappa^2$ pro kontingenční tabulky. Vypočítej pro každou skupinu n_{ij} hodnotu testovacího kritéria $\varkappa^2$ ze vzorce [2]

$$\chi^2 = \frac{(n_{ij} - o_{ij})^2}{o_{ij}}$$

ve kterém n_{ij} je příslušná skupina prvků kontingenční tabulky, o_{ij} je vypočítaná očekávaná hodnota pro tuto skupinu (viz vzorec [1] a $\varkappa^2$ je hodnota testovacího kritéria. Vypočítej tuto hodnotu pro všechny n_{ij} a sečti, výsledek označ $\Sigma\varkappa^2$. V případě, že některá hodnota $o_{ij} = 0$, pak platí, že $\varkappa^2 = 0 . \Sigma\varkappa^2$ vytiskni pod každou kontingenční tabulku.

Určení statistické významnosti kontingenční tabulky

Zjisti pro každou vytříděnou kontingenční tabulku tzv. počet stupňů volnosti podle vzorce [3]

$$df = (r-1) . (s-1)$$

ve kterém „df" je počet stupňů volnosti, „r" je počet sloupců a „s" počet řádků kontingenční tabulky.

Pro příslušné df vyhledej v tabulce (viz příloha č. 3) hodnotu veličiny $\Sigma\varkappa^2$, která odpovídá 5 %, popř. 1 % hladině významnosti. Označujeme dále $\varkappa^2$ alfa.

Porovnej vypočtenou hodnotu $\Sigma\varkappa^2$ příslušné kontingenční tabulky s tabelovanou hodnotou $\varkappa^2$ pro zvolené alfa, a jestliže

$$\Sigma\chi^2 > \chi^2 \text{ alfa}$$

vytiskni tabulku jako *významnou*, jestliže

$$\Sigma\chi^2 \leq \chi^2 \text{ alfa}$$

pracuj s tabulkou jako *nevýznamnou* (viz dále).

Výpočet koeficientů kontingence C_{norm}

Vypočti normalizovaný koeficient kontingence odvozený K. Pearsonem ze vzorce [5]

$$C_{norm} = \frac{\sqrt{\frac{\Sigma\chi^2}{n + \Sigma\chi^2}}}{\sqrt{\frac{r-1}{r}}}$$

ve kterém „n" je celkový počet jedinců tříděných v příslušné kontingenční tabulce a „r" označuje počet sloupců tabulky. V případě, že počet sloupců „r" je větší než počet řádků „s", nahradíme ve vzorci [5] „r"=„s".

Vypočtený koeficient vytisknout na čtyři desetinná místa pod příslušnou kontingenční tabulku.

Určení znaménkového schématu tabulky

Urči pro každý prvek n_{ij} základní tabulky znaménko pomocí výpočtu náhodné veličiny „z" ze vzorce [6],

$$z = \frac{(\%\, n_{ij} - \%\, o_{ij})}{\sqrt{\%\, o_{ij}\,(100 - \%\, o_{ij})}} \sqrt{n}$$

ve kterém „n" je celkový počet prvků v kontingenční tabulce

$$\%\, n_{ij} = \frac{100 . n_{ij}}{n}$$

$$\%\, o_{ij} = \frac{100 . o_{ij}}{n}$$

Vypočtenou hodnotu veličiny „z" ohodnoť podle přílohy 4 a vytiskni jí odpovídající znaménkový symbol.

Opakuj pro všechna n_{ij} a výsledky v symbolické podobě (tj. ve znaménkách) vytiskni ve formě tabulky identické s tabulkou 1.

Výstup z počítače

Tisk výsledků by měl být uspořádán do co nejpřehlednějších tabulek (viz dále ukázky konkrétní formy výstupu); a to:

1. *Tabulka je statisticky významná a v matici tabulky není více než 20 % všech „o_{ij}" menších než 5.*

Vytiskni tabulku absolutních četností (viz tab. 1), tabulku očekávaných četností (stejná forma), hodnotu $\Sigma\varkappa^2$ a významnost alfa, hodnotu koeficientů C_{norm} a znaménkové schéma. Hodnoty $\Sigma\varkappa^2$, C_{norm} vytiskni na 4 desetinná místa.

2. *Tabulka je statisticky významná, ale více než 20 % o_{ij} je menší než 5, a zároveň více než 50 % „o_{ij}" je větší než 5.*

Vytiskni totéž s poznámkou „Statisticky neprůkazné".

3. *Tabulka statisticky významná, ale 50% o_{ij} je menší než 5.*

Netisknout vůbec — k natištěné hlavičce tabulky vytisknout poznámku „Malé četnosti".

4. *Tabulky statisticky nevýznamné, méně než 20% o_{ij} menší než 5.*

Vytiskni znaménkové schéma $\Sigma\chi^2$ a hodnotu koeficientů C_{norm}, a poznámku „Statisticky nevýznamné".

5. *Tabulky statisticky nevýznamné, více než 20% o_{ij} menší než 5.*

Vytiskni pouze poznámku „Statisticky nevýznamné". Každá tabulka musí být opatřena hlavičkou udávající číslo, event. zkrácené znění tříděných otázek.[4]

Popis programu

1. Vstup do počítače

Tím jsou řádně zakódované odpovědi na otázky v dotazníku, záznamy z rozhovorů, z pozorování apod. Tyto údaje se děrují na děrnou pásku nebo na děrné štítky.

Děrování je dlouhou a pracnou záležitostí, která je závislá na počtu údajů a na jejich formě. Je nutno dbát na to, aby materiál určený k děrování byl co nejpřehledněji uspořádán. Grafickou formu je nutno dohodnout se střediskem, které bude provádět děrování.

Děrování a přezkoušení je možno provést na děrnoštítkové soupravě, je ale nutno předem zjistit, zda vstup do počítače je upraven pro ten typ štítků, na kterých budeme mít údaje naděrovány. Kódované údaje (znaky) pro náš program mají jednotnou formu: Každý typ znaku může nabývat jen jednu z hodnot, zpravidla 1, 2, ... 9 nebo 0 (neodpověděl). V případě, že znak tuto formu nemá, je nutno ho upravit (např. spojitý znak kategorizovat apod.). Se znakem 0 zpravidla nepočítáme pro další operace.

2. Třídění

1. úloha — pro všechny znaky (např. otázky v dotazníku) zjistit frekvence výskytu jednotlivých hodnot znaku (včetně počtu nulových tříd znaku).
2. úloha — pro vybrané znaky (otázky) podle zadání zjistit četnosti třídění II. stupně (popř. zadat, kdy pracovat s údaji 0).

Zadání pro třídění je možno podat (na základě dohody s programátorem) v určité dohodnuté formě, ale nejčastěji:

— v seznamu typu 1) 2, 4, 5, 11, 13 4) 3 až 9, 13, 15 apod. To znamená např. třídění II. stupně otázky 1) s otázkou 2), dále 1) s otázkou 4) atd.;
— ve vyčerpávajícím zadání typu 1) se všemi, 3) se všemi, popř. každé s každou,
— v jiné dohodnuté podobě (tzv. šachovnicová tabulka aj.).

3., 4. úloha — zadání pro třídění by mělo být maximálně přehledné a jednoduché — např. na základě dohody otázky 1) 3), 5, 3) 7)9 nebo pouze 2), 4, 6, 3), 5) 2, 14. kde pořadí čísel znaků udává postup třídění.

3. Statistické výpočty

a) Použití výpočtu relativních četností (procent) jako běžně používaného postupu.

b) Použití χ^2 testu[5] je jakýmsi východiskem z nouze — lepší statistický nástroj pro tento typ dat není k dispozici. Malou účinnost tohoto testu v případě malých četností v kontingenční tabulce se snažíme poněkud paralyzovat instrukcemi, které se váží na velikost očekávaných četností (popř. je možno brát v úvahu pravidlo, že žádná očekávaná četnost nesmí být menší než 1,0). Problematičnost použití χ^2 testu proto stoupá s prací s malými vzorky (do 200—300), vyššími stupni třídění a zvyšováním počtu alternativ v otázkách.

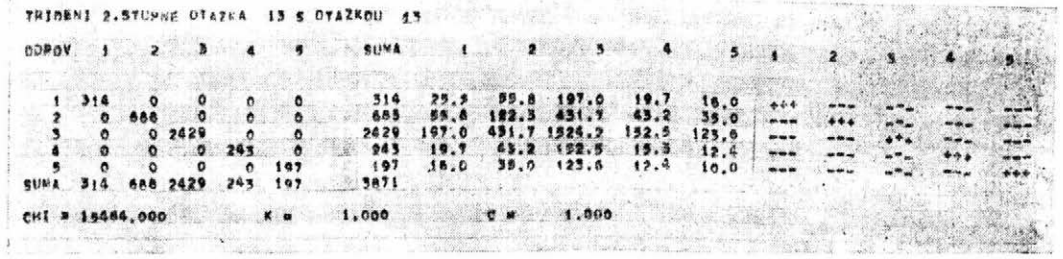
c) Programovaný normalizovaný koeficient kontingence chápeme jako přibližnou míru síly vztahu mezi dvěma znaky (odpověďmi na otázky). C_{norm} nabývá hodnot od 0 do 1,0, a to vždy kladných. Jeho značnou nevýhodou je to, že silně závisí na hodnotě „n", tj. na celkovém počtu prvků v kontingenční tabulce — srovnávání C_{norm} v tabulkách se značně rozdílnými „n" je proto sporné. Vcelku však hodnoty $C_{norm} > 0{,}5$ můžeme považovat za značně vysoké. Jako příklad maximálního vztahu mezi dvěma otázkami uvádíme tabulku 2.

Velikost koeficientu nám umožňuje hrubé srovnání intezity vztahu mezi

[4] Předkládáme jenom jednu z možných forem výstupu z počítače. V určitých případech bude zřejmě vhodnější vytisknout tabulku, ve které by byly vytištěné procentuální četnosti počítané z marginálií tabulky, tedy z marginálií otázky A a otázky B.

[5] Součet hodnot χ^2 nám charakterizuje odchylku empirického modelu (rozložení četností v tabulce 1) od teoretického (tabulka očekávaných četností). Gnoseologický význam má pouze sumace všech příspěvků χ^2. Z příspěvků k celkové hodnotě $\Sigma\chi^2$ v jednotlivých políčkách kontingenční tabulky je problematické dělat závěry.

Tabulka 2



TRIDENI 2.STUPNE OTAZKA 13 S OTAZKOU 15

ODPOV	1	2	3	4	5	SUMA	1	2	3	4	5	1	2	3	4	5
1	314	0	0	0	0	314	25.9	55.8	197.0	19.7	16.0	+++	---	---	---	---
2	0	688	0	0	0	688	56.8	122.3	431.7	43.2	35.0	---	+++	---	---	---
3	0	0	2429	0	0	2429	197.0	431.7	1525.2	152.5	123.6	---	---	+++	---	---
4	0	0	0	243	0	943	19.7	43.2	152.5	15.3	12.4	---	---	---	+++	---
5	0	0	0	0	197	197	16.0	35.0	123.6	12.4	10.0	---	---	---	---	+++
SUMA	314	688	2429	243	197	3871										

CHI = 15466.000 K = 1.000 C = 1.000

dvěma znaky (odpověďmi na dvě otázky) v různých kontingenčních tabulkách, a to ještě značně podmíněně — pouze tam, kde znaménkové schéma ukazuje zřetelné uspořádání odchylek ve směru diagonály tabulky. V případě, že odchylky jsou uspořádány různě po celé ploše tabulky, potom i míra vztahu v podstatě ztrácí svůj smysl.

Další statistické operace s touto mírou vztahu nepovažujeme za spolehlivé, to znamená, že s koeficientem kontingence nelze v žádném případě počítat parciální a množné korelace a že jej nelze dosazovat do matic zpracovávaných faktorovou analýzou.[6]

Významnost koeficientu je dána mírou významnosti tabulky, ze které jsou počítány.

d) Znaménkové schéma považujeme za velmi vhodnou optickou pomůcku pro interpretaci získaných tabulek — je to test významnosti rozdílu mezi absolutní četností a jí odpovídající teoretickou četností očekávanou (za předpokladu nezávislosti). Za testovací kritérium byla zvolena tzv. veličina „z", meze hodnot pro přiřazení znamének (viz příloha) — např.:

+++ = empirická hodnota příslušného políčka je o mnoho větší než hodnota teoretická;

— = empirická hodnota je slabě, ale významně menší než vypočítaná teoretická hodnota;

0 = rozdíl mezi empirickou a očekávanou hodnotou, je nevýznamný,

byly určeny odhadem. Jelikož frekvence výskytu vysokých odchylek závisí bezesporu na charakteru zkoumaných znaků, je možno meze hodnot veličiny „z" přizpůsobovat podle charakteru výzkumu. Veličina „z" nedostatečně rozlišuje odchylky velmi malých empirických a očekávaných hodnot (v tomto případě je její použití dokonce nesprávné). Jelikož však slouží pouze jako optické vodítko k interpretaci tabulky, domníváme se, že tento nedostatek lze pouze vzít na vědomí, popřípadě zabudovat další „obrannou" instrukci do programu, například: „pokud rozdíl mezi očekávanou a empirickou četností je menší než 5,0, nechť tato hodnota je hodnocena ve schématu jako 0 bez ohledu na velikost »z«" — nebo lze použít jiného testovacího kritéria.[7]

[6] Paralelně s výpočtem C_{norm} jsme používali výpočtu koeficientu K (odvozeného A. A. Čuprovem, viz V. J. Urbach), který je uváděn v ukázkách tabulek v textu. Tento koeficient má zhruba stejnou gnoseologickou hodnotu jako C_{norm}; jeho vzorec je následující:

$$K = \sqrt{\frac{\chi^2}{n(r-1)(s-1)}}$$

Pokouší se korigovat nedostatek C_{norm}, který je nespolehlivý v tabulkách, kde „r" a „s" (tj. počet sloupců a řádků) se od sebe příliš liší (např. tabulka 2×9). Jeho praktické použití však ukázalo, že jeho vypočtené hodnoty jsou zhruba poloviční proti hodnotám C_{norm} při velikosti hodnot C_{norm} do 0.7 — tj. diferencuje hůře síly vztahů, se kterými se v sociologii setkáváme nejčastěji, proto od jeho používání upouštíme a domníváme se, že ho plně nahradí C_{norm} Podobnou variantou C_{norm}, která se podobně jako K pokouší odstranit uvedený nedostatek, je následující vzorec uváděný Hoffstätterem:

$$C_{norm} = \sqrt{\frac{\chi^2}{n+\chi^2}} : \left[\left(\sqrt{\frac{r-1}{r}} + \sqrt{\frac{s-1}{s}}\right) \cdot \frac{1}{2}\right]$$

S jeho distribuční charakteristikou a vhodností použití zatím nemáme přímé zkušenosti.

Zdá se, že vhodnou mírou vztahu pro tabulky, kde jak v řádcích, tak ve sloupcích kontingenční tabulky jsou data (otázky) ordinální stupnice, jsou koeficienty pořadové korelace, z nichž za vhodné pokládáme koeficient pořadové korelace Spearmana a koeficient pořadové korelace odvozený Kendallem a označovaný jako τ.

Pro jejich výpočet je třeba použít upravených vzorců pro případy stejného pořadí, hodnoty těchto koeficientů jsou v rozmezí od -1 do $+1$. Vzorce pro jejich výpočet jsou poměrně složité, proto je zde nebudeme uvádět a odkazujeme zájemce na příslušnou literaturu (N. Siegel).

Vzorce pořadové korelace jsou velmi praktické, protože umožňují počítat množnou a parciální korelaci a lze je použít k faktorové analýze.

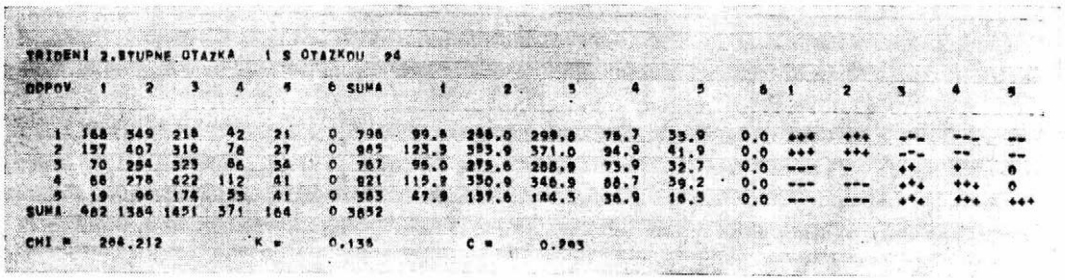
První pokusy s programováním koeficientu pořadové korelace se ukázaly být velmi slibné.

[7] Místo testu významnosti dvou procent (viz Malý), který zde uvádíme, lze použít jakéhokoliv jiného testu (např. lze testovat rozdíl procent pomocí sínové transformace). V úvahu by také přicházela adaptace McNamara testu významnosti změn. (Viz N. Siegel).

Příklady tabulek

Uvádíme dvě typické tabulky ve formě, v níž byly získány na základě nepodstatné modifikace našeho programu pomocí počítače ICT.

Tabulka 3



TRIDENI 2.STUPNE OTAZKA 1 S OTAZKOU 24

ODPOV	1	2	3	4	5	6	SUMA	1	2	3	4	5	6	1	2	3	4	5
1	168	349	218	42	21	0	798	99.6	288.0	299.8	78.7	33.9	0.0	+++	+++	---	---	---
2	157	407	316	78	27	0	985	123.3	383.9	371.0	94.9	41.9	0.0	+++	+++	---	--	--
3	70	284	323	86	54	0	767	95.0	275.6	288.9	73.9	32.7	0.0	---	-	++	+	0
4	88	278	422	112	41	0	921	115.2	330.9	346.9	88.7	39.2	0.0	---	---	+++	+++	0
5	19	96	174	53	41	0	383	47.9	137.6	144.5	36.9	16.3	0.0	---	---	+++	+++	+++
SUMA	482	1384	1451	371	164	0	3852											

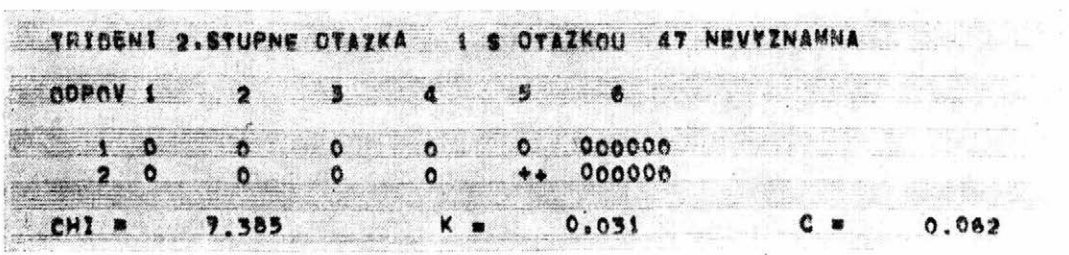
CHI = 284.212 K = 0.136 C = 0.293

V průsečíku sloupce „1" a řádku „1" je četnost 168, tj. počet osob, které odpověděly na otázku 1) „alternativou 1" a na otázku 24) také „alternativou 1". Celkem potom např. na otázku 1) odpovědělo „alternativou 4" 371 osob (suma v řádku). Podobně celkem 985 osob odpovědělo na otázku 24) „alternativou 2" (suma ve sloupci). Celkem na obě otázky odpovědělo 3 852 osob.

Další sloupce čísel (na 1 desetinné místo) jsou očekávané četnosti. Četnost v průsečíku sloupce 1 a řádku 1 (99,6) je očekávaná četnost k empirické hodnotě 168. Z porovnání vidíme, že absolutní četnost je o d o s t větší než četnost očekávaná — významnost pro tento rozdíl je oklasifikována opět v průsečíku sloupce 1 a řádku 1 ve znaménkovém schématu jako +++. Obdobně pro každé políčko tabulky absolutních četností nalézáme odpovídající četnost očekávanou a znaménko ve znaménkové matici.

Suma χ^2=284,212, což odpovídá hladině významnosti $\alpha < 0{,}1\ \%$; to znamená, že mezi odpověďmi na obě otázky je vysoce významný vztah. Síla tohoto vztahu je oklasifikována velikostmi koeficientu Cnorm. Směr tohoto vztahu je zřejmý ze znaménkové matice. Je vždy významně více těch, kteří odpovídají na otázku 1) a na otázku 24) shodnými alternativami — v získaném materiálu jde o vztah mezi odpověďmi na otázku „Zajímá Vás práce, kterou vykonáváte?", o alternativách 1 — velmi mne zajímá, 2 — více zajímá než nezajímá, 3 — zajímá i nezajímá (tak 50 : 50), 4 — spíše nezajímá, 5 — vůbec nezajímá, a odpověďmi na otázku „Domníváte se, že Vám nadřízení dávají možnost projevit iniciativu a vyslovit svůj názor na věci týkající se celého pracoviště?" o alternativách 1 — vždy, 2 — jen někdy, 3 — nemohu posoudit, 4 — velmi zřídka, 5 — nikdy.

Tabulka 4



TRIDENI 2.STUPNE OTAZKA 1 S OTAZKOU 47 NEVYZNAMNA

ODPOV	1	2	3	4	5	6	
1	0	0	0	0	0	000000	
2	0	0	0	0	++	000000	

CHI = 7.385 K = 0.031 C = 0.062

Další tabulka ukazuje nevýznamný vztah mezi odpověďmi na otázku 1) a otázku 47, neboť $\alpha > 10\ \%$. (Viz tab. 4.)

Významnost výkyvu (++) v jednom políčku tabulky lze interpretovat velmi podmíněně — spíše jako určitou tendenci.

Shrnutí

1. Z použitých statistických postupů vyplývá, že program je vhodný k testování

hypotéz za předpokladu nezávislosti. — V praxi to znamená, že můžeme tímto programem zpracovat nejběžněji dostupný materiál. Program je nevhodný pro vyhodnocování panelových průzkumů, ku testování rozdílů např. mezi pretestem a posttestem a jako test významnosti změn.

2. Použijeme-li tohoto programu k vyhodnocování dat, která jsou svou podstatou ordinální nebo intervalová, musíme mít na paměti, že použité statistické operace nám v těchto případech poskytnou pouze aproximativní výsledky, protože χ^2 test je určen pro data nominální stupnice.

3. Použití programu proto není vhodné tehdy, máme-li proměnné vyjádřené většinou znaky intervalové povahy. Rozhodneme-li se ho přesto použít, je nutno znaky vhodně kategorizovat.

4. Otázky, jejichž forma umožňuje respondentovi uvést více alternativ odpovědí, popř. tyto alternativy nějak seřadit (podle důležitosti apod.), není zpravidla možno zpracovat přímo a je nutno je vhodně upravit.

5. Výhodné je pracovat s otázkami o přibližně stejném počtu alternativ (např. 4, 5), nevhodné je používat otázky s vysokým počtem alternativ (např. 9).

6. Vyhodnocením předvýzkumu je vhodné zajistit, aby všechny alternativy odpovědí byly zastoupeny dostatečně vysokým počtem jedinců.

7. Přestože zpracování dat samočinným počítačem je při použití tohoto programu poměrně levné, je třeba vždy pečlivě zvážit i jiné možnosti (např. děrné štítky, popřípadě analyzátorové karty). Použití programu je nejvhodnější tam, kde potřebujeme velké množství tabulek složitého třídění a řadu statistických výpočtů.

Perspektivy výstavby vyhodnocovacích programů

Jak již bylo řečeno, princip tohoto programu byl vyzkoušen na počítači Minsk 22 v roce 1964-65 při vyhodnocování průzkumu studentů FVL UK. Později byly na modifikacích programu zpracovány ještě dva další průzkumy téhož pracoviště.

Přibližně v téže době, jak je nám známo, začala výstavba univerzálního programu PROSA, organizovaná manželi Perglerovými, na němž se stále pracuje. Tento program počítá s daleko širší paletou statistických operací (např. pro intervalové znaky) a pokouší se rozpracovat i prvky samoprogramování.

Princip zde rozebíraného programu byl upraven pro počítače Eliot 503 pro výpočty a třídění výzkumu ÚML Praha (výzkum v AZNP Mladá Boleslav), nejpraktičtější upravená varianta programu je zatím odladěna ve Výzkumném ústavu výpočetní techniky Praha 2, Nekázanka 5 (vedoucí ing. Pešan, programátor prom. mat. Ctiborský) na zakázku sociologického pracoviště ÚV ČSM. Středisko v současné době tímto programem zpracovává několik různých sociologických výzkumů a má zájem o další zakázky podobného typu.

Zatím nejúspěšnějším nasazením programu zůstává průzkum MV KSČ „Politická aktivita a poměr k práci u mládeže na pražských strojírenských závodech", který zpracovávalo na počítači ICT výpočetní středisko ČKD Praha (vedoucí ing. Brand) a z něhož jsou i ukázky tabulek v textu.

Tyto zkušenosti ukazují některé nesporné přednosti práce na samočinných počítačích, které ještě jednou krátce shrnujeme:

— Třídění na počítači je velmi rychlé a úsporné.
— Programování statistických výpočtů není zpravidla složité, rychlost a přesnost výpočtů srovnávat s prací pomocí kalkulačky nemá smysl.
— Je možno uvažovat o programování i interpretačních postupů. Pro informaci uvádíme, že cca 2000 tabulek námi uvedeného typu bylo získáno na počítači ICT za cca 6 hodin chodu stroje. Tatáž tabulka (výstup byl ovšem horší kvality) na počítači Minsk 22 spotřebovala přibližně 17 vteřin chodu stroje. (Pro srovnání odhad pomocí zatím běžné techniky: cca 7 minut třídění na třidiči + nejméně 3 hodiny pro statistické výpočty na kalkulačce bez záruky přesnosti.)

Finanční náklady závisí na počtu zpracovávaných dat a rozsahu třídění. Při výzkumech středního typu se pohybují kolem 5000 Kčs.

Domníváme se, že některé z ústředních sociologických pracovišť by mělo věnovat pozornost (a hlavně finanční prostředky) rozvíjení základního výzkumu v ob-

Měření souvislostí společenských jevů

PŘEMYSL PERGLER

Souhrn

V předložené práci se navrhuje míra souvislosti, založená na některých pojmech teorie informace, a to s využitím výsledků prací některých našich i zahraničních autorů. Je vhodná pro měření vzájemných souvislostí většího počtu společenských jevů a lze ji gnoseologicky jednoznačně interpretovat pro znaky nominální i kvantitativní. Ukazuje se, že v některých případech je odůvodněné zavést koeficient souvislosti r. Pearsonův korelační koeficient je pak speciálním případem koeficientu souvislosti. Postup při měření souvislosti je poměrně jednoduchý; pro rozsáhlé výzkumné akce lze vyčíslení výhodně provést na samočinném číslicovém počítači. Numerické výsledky mohou být podkladem interpretace výsledků výzkumné akce (např. pomocí kauzálního nebo jiného modelu), výběru otázek pro dotazník apod. V textu se vysvětluje na příkladech princip a použití, dodatky obsahují odvození a připomínky.

1

Pro vysvětlení pojmu souvislosti vyjděme z fiktivního příkladu, v němž nejprve předpokládáme vyčerpávající šetření (tj. byli vyšetřováni všichni jedinci v populaci).

Příklad 1

Sledujeme nominální alternativní znak u_1 . . . Postoj k práci . . . dobrý . . . špatný. (Pro jednoduchost budeme sledovat malý počet alternativních znaků; hlavní výhody popisované techniky se však jeví u většího počtu množných znaků. Otázkou vztahu znaků samotných ke sledovaným jevům se v této práci nezabýváme). Veškeré podklady pro příklad 1 jsou obsaženy v prvotní tabulce T 1.

Tabulka 1

Relativní četnost p		
Postoj k práci u_1	dobrý	0,50
	špatný	0,50
	součet	1,00

Pokusme se předpovědět, jaká je pravděpodobnost, že náhodně vybraný jedinec z populace má dobrý postoj k práci. Podle prvotní tabulky je to pravděpodobnost $p = 0{,}50$.

Naše předpověď není zcela určitá; lze říci, že zahrnuje jistou neurčitost. Znaky, jejichž předpověď neobsahuje žádnou neurčitost, není nutno se zabývat (byl by to např. případ, kdy všichni jedinci mají dobrý postoj k práci).

Příklad 2

Kromě znaku u_1 sledujeme i nominální alternativní znak

u_2 . . . Kvalifikace jedince . . . vyšší
nižší

Prvotní četnosti jsou uvedeny v tabulce T 2.

Tabulka T 2

Relativní četnost p		Kvalifikace u_2		
		vyšší	nižší	součet
Postoj k práci u_1	dobrý	0,00	0,50	0,50
	špatný	0,50	0,00	0,50
	součet	0,50	0,50	1,00

Předpokládejme nyní, že u jedince náhodně vybraného z populace známe hodnotu znaku u_2 (kvalifikace) a pokusme se předpovědět postoj k práci téhož jedince. Existuje-li nějaká souvislost mezi oběma znaky, je jiná pravděpodobnost, že kvalifikovaný jedinec má dobrý postoj k práci, než pravděpodobnost, že nekvalifikovaný jedinec má dobrý postoj k práci: kdyby tomu tak nebylo, nemohli bychom o souvislosti obou jevů mluvit. (V našem extrémně zjednodušeném příkladě je bezprostředně patrné, že existuje silná souvislost.) Naše předpověď o pracovním postoji náhodně vybraného jedince bude při znalosti jeho kvalifikace dokonalejší než bez znalosti jeho kvalifikace, pokud oba jevy spolu souvisí. Znalost kvalifikace jedince nám přináší jistou informaci o jeho pracovním postoji. V našem extrémním případě dokonce úplně

zmizí neurčitost předpovědi pracovního postoje jedince, jehož kvalifikaci známe. Informace o postoji, kterou s sebou nese znalost kvalifikace, je dokonalá; v našem případě je to dokonce nejdokonalejší informace, jakou si dovedeme představit.

Příklad 3

Prvotní četnosti by však mohly být rozděleny i jinak, například podle tabulky T 3.

Tabulka T 3

Relativní četnost p		Kvalifikace u_2		
		vyšší	nižší	součet
Postoj k práci u_1	dobrý	0,25	0,25	0,50
	špatný	0,25	0,25	0,50
	součet	0,50	0,50	1,00

Tabulka T 1, která popisuje rozložení četnosti pouze podle postoje k práci, zůstává nadále v platnosti. Možnosti odhadu postoje náhodně vybraného pracovníka na podkladě znalosti jeho kvalifikace jsou však ve srovnání s příkladem 2 podstatně odlišné. Znalost kvalifikace nám zde nepřináší žádnou informaci pro předpověď postoje. Neurčitost předpovědi se znalostí kvalifikace nesnížila. Oba jevy spolu nesouvisí.

Úvaha, ilustrovaná uvedenými příklady, vede k tomuto myšlenkovému pokusu: Mějme předpovědět, s jakou pravděpodobností bude u náhodně vybraného jedince souboru mít znak u_1 jistou (předem určenou) hodnotu a proveďme předpověď pro všechny možné hodnoty znaku u_1. Tato předpověď nebude zcela určitá, tj. bude zatížena jistou neurčitostí. Sledujme, jaký význam má pro tuto předpověď znalost hodnoty znaku u_2 téhož jedince, tj. kolik informace o znaku u_1 nese sebou znalost znaku u_2. Pak platí zároveň výroky 1° a 2°.

1° Čím více souvisí znak u_1 se znakem u_2, tím více informace o znaku u_1 nese s sebou znak u_2.

2° Čím více souvisí znak u_1 se znakem u_2, tím více se na podkladě znalosti znaku u_2 sníží neurčitost předpovědi o hodnotě znaku u_1.

(V textu předpokládáme, že platí ekvivalence dvou dvojic výroků:

Předpověď je méně určitá — Předpověď nese s sebou větší neurčitost.

Znalost hodnoty u_2 má větší význam pro předpověď hodnoty znaku u_1. — Znak u_2 nese s sebou větší informaci o znaku u_1).

Termín „předpověď" se vztahuje k myšlenkovému pokusu; není to tedy například předpověď o vývoji znaku v čase. Výroky 1° a 2° mají smysl, existuje-li užitelná kvantitativní míra neurčitosti a informace a neodporují-li si výsledky dosazení těchto měr do obou výroků. Je-li tomu tak (a jsou-li splněny ostatní požadavky zadání úlohy), lze z pojmů neurčitosti a informace odvodit hledanou míru souvislosti. Jak plyne z formulace obou výroků, půjde o proměnnou ordinální. Teorie informace, jak známo [11], [12], míru informace a neurčitosti definuje. Na jejím podkladě se pak v dodatku 2 odvozuje míra souvislosti znaků, která se nazývá koeficientem souvislosti. Platí potom vztahy podle tabulky T 4.

Tabulka T 4

Jev	Míra jevu	
	Slovní název	Označení
Neurčitost znaku u_1	Entropie znaku u_1	H (1)
Informace o znaku u_1 nesená znakem u_2	Informační obsah u_1 obsažený v u_2	I (2, 1)
Souvislost znaku u_1 se znakem u_2	Koeficient souvislosti znaku u_1 se znakem u_2	r (2, 1)

Dosavadní úvahu shrnují výroky 3° a 4°.

3° Intenzita souvislosti znaku u_1 se znakem u_2 je dána informací o znaku u_1, obsaženou ve znaku u_2.

4° Intenzita souvislosti znaku u_1 se znakem u_2 je dána poklesem neurčitosti předpovědi hodnoty znaku u_1 na podkladě znalosti znaku u_2.

Oba tyto výroky lze považovat za definici intenzity souvislosti.

V dodatku 2 se dále ukazuje, že navržený postup je aplikovatelný i na množné znaky nominální, i na znaky kvantitativní. Pearsonův korelační počet lze považovat za zvláštní případ měření souvislosti a korelační koeficient je identicky rovný koeficientu souvislosti.

O výpočetním postupu je stručná zmínka v závěru práce. (Je zajímavé si všimnout, že při popisovaném postupu nevadí, jsou-li ně-