Program pro statistické vyhodnocení sociologických dat

JOSEF LAUBER
Vysoká škola ekonomická, Praha

S používáním moderní výpočetní techniky pro zpracování sociologických výzkumů se u nás začalo před několika málo lety. Za tak krátkou dobu byla na různých pracovištích vytvořena řada programů, které se od sebe liší jak celkovou koncepcí, tak i tím, pro jaký druh počítače jsou určeny. Přes některé snahy o výměnu zkušeností z této oblasti hromadného zpracování dat nemá dosud sociologická veřejnost dostatečný přehled o obsahu těchto programů, o jejich ekonomických parametrech, ani o dostupnosti programů pro široký okruh zájemců.

Z této situace jsme vycházeli, když jsme se rozhodli publikovat informace o programu, který byl vytvořen na sociologickém pracovišti Fakulty všeobecného lékařství KU. Práce na první verzi tohoto programu byly zahájeny v roce 1964 (viz článek [1]). V následujících letech program po několikerém přepracování dosáhl podoby, kterou popisujeme v tomto článku. Rozhodně by bylo užitečné, kdyby i ostatní pracoviště publikovala základní informace o svých programech.

Před započetím strojového zpracovávání jsme stáli před problémem, na jaký počítač se máme zaměřit. Volba padla nakonec na MINSK 22, který byl — a zatím ještě je — u nás nejrozšířenějším počítačem, a proto je možno zde poměrně snadno získat volnou strojovou kapacitu.

Operace, které může program provádět

1. Třídění souboru až do 50. stupně včetně.
2. Statistické vyhodnocení vzniklých kontingenčních tabulek (ve smyslu článku [2]).
3. Výpočet skóru (sumovaného odhadu) ze zadané množiny znaků a další operace s tímto skórem.
4. Slučování a vypouštění hodnot znaků a tříd (vypouštěním řádků, resp. sloupců u výsledných kontingenčních tabulek nebo jejich slučováním).

Předpoklady užití programu

Program předpokládá data naděrovaná v kódu MINSK nebo v mezinárodním dálnopisném kódu (podle speciálních pokynů), případně na 90sloupcových děrných štítcích.

Struktura programu předpokládá užití znaků, které jsou kódovány přirozenými čísly, tj. 1, 2, ... Nejvýše přípustný počet tříd (hodnot znaků) byl z praktických důvodů stanoven na 30. Číslem 0 kódujeme zpravidla údaj „neodpověděl", a to tehdy, jestliže nás údaj nezajímá, tj. například když jej nepotřebujeme třídit s ostatními znaky. V opačném případě kódujeme i údaj „neodpověděl" číslem jiným než 0. Dalším přirozeným omezením je nutnost, aby se data vešla do vyhrazené části paměti počítače. Program využívá tří magnetických pásek, prakticky použitelná kapacita paměti je tudíž asi 180 000 slov (záleží na kvalitě pásek). — Paměťová buňka (slovo) má u počítače MINSK nosnost informace 36 bitů. Informace o hodnotě znaku, který má maximálně devět tříd, se ve zvoleném systému ukládání zobrazí na čtyřech bitech. Hodnoty znaku, který má 10—30 tříd, se ukládají do osmi bitů paměti. Na tomto základě lze snadno spočítat, kolik bitů paměti zaberou informace o každém jedinci studovaného souboru, a konečně, kolik slov paměti je nutno rezervovat pro údaje o jednom respondentovi. Potřebný počet slov paměti na uchování údajů z jednoho dotazníku označme p, počet respondentů označme N. Potom programem lze zpracovávat výzkumy, u nichž

$$N \cdot p < 180\,000$$

Povelový kód

Povelovým kódem rozumíme jazyk, pomocí kterého předkládá badatel počítači (přesněji řečeno programu) své požadavky na zpracování dat. Tento povelový kód byl vytvořen tak, aby se mu badatel mohl snadno naučit, a zároveň aby dešifrování povelů nezabralo počítači příliš mnoho ča-

su. Povelový kód je tedy jakýmsi kompromisem mezi těmito těžko slučitelnými požadavky. Nebudeme se jím detailně zabývat, uvedeme pouze jeho hlavní rysy, které objasníme na typických a nejvíce používaných příkladech.

A. *Třídění prvního stupně*

Požadavek na třídění prvního stupně, tj. získání rozložení četností jednotlivých hodnot znaků spolu s výpočtem procent, se zadá jednoduše tak, že se napíší pořadová čísla znaků dotazníku, pro která se má tato operace provést.

B. *Třídění druhého stupně*

Třídění druhého stupně (výsledkem je kontingenční tabulka v % s řádky a sloupci spolu se statistickými údaji χ^2 test na 1% a 5% hladině statistické významnosti, informace o korektnosti tohoto testu, znaménkové schéma, normalizovaný koeficient kontingence, Spearmanův koeficient pořadové korelace, atd., viz [2]) je možno zadávat v několika modifikacích:

α) pro každou jednotlivou tabulku zápisem čísel znaků $T_a * T_b$.

β) třídění jednoho znaku T_a s posloupností znaků T_i ($i = 1,2 \ldots n$) se zadá takto: $T_a * (T_1, T_2, \ldots T_n)$.

γ) třídění systémem „každý s každým" ze znaků $T_1, T_2, \ldots T_n$ se zadá takto: $(T_1, T_2, \ldots T_n)$
(počítač sám vytvoří všechny netriviální kombinace dvou znaků ze zadané množiny znaků).

C. *Třídění vyšších stupňů*

V tomto případě rozumíme tříděním vyšších stupňů třídění prvního nebo druhého stupně prováděné na podsouboru, který je specifikován jakousi „hierarchií" vztahů, které mají jedinci tohoto podsouboru splňovat. Co tím míníme, bude nejlépe patrno na příkladě. Chceme například provést třídění druhého stupně $(T_a * T_b)$ pro jedince, kteří na otázku T_x odpověděli kategorií 2, a zároveň na otázku T_y kategorií 6. Zadání pro tento případ:

PRO T_x 2 a T_y/6
$T_a * T_b$

Požadavky na předtřídění, tj. povely tvaru T_x/i, potom platí pro libovolná třídění prvního a druhého stupně, která následují. Posloupnost povelů pro předtřídění lze aditivně rozšiřovat a tím dosáhnout sledování stále speciálnějšího souboru. Chceme-li potom pracovat opět s celým souborem nebo budovat od základu novou „hierarchii" předtřídění, napíšeme nejprve povel „předtřídění zruš" a postup třídění opakujeme na novém zadání. Posloupnost povelů pro předtřídění, které jsou v platnosti, tj. nebyly zrušeny, může obsahovat maximálně padesát prvků (stanoveno opět z praktických důvodů).

D. *Skór z otázek* $T_a \ldots T_z$ (maximálně 35 otázek).

Slovy „skór z otázek $T_a \ldots T_z$" označujeme součet hodnot znaků $T_a \ldots T_z$. Tuto operaci zadáme takto:

SKÓR $(T_a, \ldots Z_z)$

Tento povel počítač interpretuje tak, že u každého jedince spočte příslušný skór a tiskne rozložení souboru podle hodnot tohoto skóru.

Program umožňuje badateli na základě znalosti tohoto rozložení skórů kategorizovat (maximálně do devíti tříd) a pracovat s takto nově vytvořeným znakem jako s kterýmkoliv jiným, to znamená, že tento umělý znak nejen lze třídit s ostatními, ale může vstupovat například jako jeden znak do další operace SKÓR $(T_a, \ldots T_z)$, apod. Jestliže některou otázku napíšeme do povelu SKÓR $(T_a, \ldots T_z)$ vícekrát, potom se v součtu uplatní s příslušnou „váhou". Lze tedy dělat vážený skór, u kterého jsou váhy přirozená čísla. Pomocí kategorizovaného skóru lze, jak je možno ukázat, provádět i jeden speciální způsob překategorizování znaků.

E. *Slučování kategorií u kontingenčních tabulek*

V mnoha případech mají výsledné kontingenční tabulky některé nepříjemné vlastnosti, pro které nelze korektně provádět některé druhy statistických vyhodnocení. Máme tím na mysli hlavně malé četnosti v některých řádcích či sloupcích získaných tabulek. V tomto případě je někdy účelné (je-li to i významově možné) provést sloučení některých řádků a sloupců tabulky tak, aby byla statisticky vyhodnotitelná.

Skutečný povelový kód, kterého užíváme ve styku s počítačem, se ovšem od uvedeného poněkud liší. Odchylky jsou však pouze formálního charakteru, takže pro badatele, který by s programem pracoval

standardně, nepředstavuje jejich dodržování žádné principiální potíže. Požadavky výzkumníka vyjádřené ve zmíněném kódu se naděrují na děrnou pásku, která se spolu s daty předloží počítači. Program počítá také s tím, že některé požadavky je možno uplatnit bezprostředně při práci stroje pomocí klávesnice ovládacího pultu. Výzkumník také není nikterak vázán požadavkem na zadání veškerých operací, které chce s daty provádět. To umožňuje rozdělit celé zpracování na kratší etapy, které mohou po sobě následovat v intervalech nutných pro interpretaci získaných výsledků a pro volbu dalších hypotéz, jejichž ověření se ukazuje účelné.

Ekonomické parametry programu

Při sestavování programu byl kladen velký důraz na to, aby strojové zpracování sociologických dat bylo co možná nejlevnější. Proto, ve snaze po maximální úspoře času při vlastním chodu počítače, byl program napsán přímo ve strojovém kódu. Do jaké míry byla tato snaha úspěšná, bude nejlépe patrno z několika příkladů, z nichž je možno učinit si představu o rychlosti a nákladech na strojovou část zpracování výzkumů:

Při souboru 1000 respondentů a 35 otázkách v jednom dotazníku trvá výpočet jedné tabulky (i s tiskem) asi 8—10 vteřin. Při souboru 10 000 respondentů a 90 otázkách trvá výpočet asi 2—3 minuty. V současné době stojí 1 hodina strojového času na počítači MINSK cca 600—900 Kčs. Při této ceně je náklad na jednu tabulku v prvním případě asi 2 Kčs, v druhém případě asi 30 Kčs.

Abychom zjistili celkové náklady na strojové zpracování, je nutno k nákladům na strojový čas ještě připočítat náklady na děrování údajů a práci programátora.

Veškeré bližší informace o tomto programu, případně o možnostech jeho použití, podá autor nebo pracovníci sociologického pracoviště při katedře filosofie a sociologie Fakulty všeobecného lékařství University Karlovy.

Závěrem je mou povinností poděkovat tomuto pracovišti za pomoc při vytváření koncepce programu a poskytnutí možnosti tento program realizovat.

Literatura:

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

[2] J. Linhart—Z. Šafář, *Programování třídění a statistických výpočtů pomocí samočinných počítačů,* Sociologický časopis 4/1967.

Резюме

Иосиф Лаубэр: Программа для статистической оценки социологических дат

Статья информирует о проверенной программе, применяемой при оценке социологических исследований. При помощи этой программы можно проводить как повседневные операции — распределение самого различного вида, статистическое тестирование гипотез и так далее, так и некоторые операции, служащие для шкаловых процессов, переоценку категоризации знаков и создание новых знаков. Далее в статье описан «язык», при помощи которого работник исследования задает свои требования вместе с экономическими параметрами программы при её применении на счетчике МИНСК-22.

Summary

Josef Lauber: Program for a Statistical Evaluation of Sociological Data

The article presents information on a tested program used in evaluating sociological researches. This program makes it possible to perform both routine operations (classifications of most various kinds, statistical testing of hypotheses, etc.) and some further operations serving the purposes of the scaling procedure, the re-categorization of the values of symbols and the formation of new symbols. The article describes the "language" with the help of which the research worker places his requirements, together with the economic parameters of the program applied in the computer MINSK 22.