

Statistické programy a chybějící hodnoty: varování před nečekanými výsledky

MICHAL ČÁKRŤ

Výzkumný ústav strojírenské
technologie a ekonomiky,
Praha

Soudobý sociologický výzkum se neobejde bez kvalitních prostředků pro analýzu dat. A protože sociologie, snad více než kterákoli jiná společenská věda, velmi často pracuje s jevy hromadného výskytu, tzn. se značně rozsáhlými výběrovými soubory – jak z hlediska počtu respondentů, tak i měřených veličin – je takřka vždy odkázána na zpracování dat na počítačích. Doby, kdy pionýři empirických výzkumů vyhodnocovali sebraný datový materiál ručně nebo za pomoci primitivní techniky, patří již nenávratně minulosti a do říše legend naší vědy. Přestože u nás bylo v minulosti vytvořeno množství jednotlivých programů pro různé úlohy analýzy dat (jenom regresních programů existují stovky), dodnes na trhu chybí kvalitní, prakticky použitelný, uživatelsky orientovaný a cenově dostupný softwarový produkt, pokrývající alespoň nejobvyklejší spektrum požadavků pracovníků výzkumu i praxe, který by si sociolog, resp. jeho zaměstnavatel, mohl bez komplikací opatřit. Co horšího, chybí u nás i takový trh, pro nějž by komukoli „stálo“ za to nějaký náročnější programový systém vyvíjet. Toto skeptické stanovisko je ještě dále podpořeno tím, že na většině modernějších tzv. střediskových počítačů řady EC lze provozovat několik starších, licenčně již nechráněných, avšak dosud vyhovujících systémů západní provenience. Navíc s nástupem osobních počítačů nelze očekávat, že by se vedle nakoupených či pronajatých programů naší zemi do budoucna zcela vyhnuly některé ze stovek „divoce“ obíhajících nebo tzv. „public domain“ (tzv. veřejně dostupných prakticky za cenu nosného media) programů ze zahraničí. Tím bude bezesporu tlak na vývoj domácích produktů ještě dále oslaben.

Tváří tvář této skutečnosti jsou sociologové ve stále větší míře odkázáni buď na jednotlivé statistické programy anebo na programové systémy nejrozličnějšího původu a rozličné kvality, jimiž jsou to či ono výpočetní středisko nebo osobní počítač vybaveny, popřípadě na služby více či méně seriózních externích spolupracovníků, kteří příslušnými programy disponují.

Není účelem této stati jednotlivé programové produkty srovnávat, hodnotit či doporučovat, jakkoli by takový počín byl nanejvýš žádoucí a potřebný. Nakonec sociolog sám jen zřídka určuje, který z nich použije. Na tomto místě chci zatím pouze poukázat na jistá možná nemilá překvapení, která mohou výzkumníka potkat, přechází-li – dobrovolně nebo z nutnosti – na jiný systém anebo, jak uvidíme, na novou generaci téhož systému. Nepůjde nyní o zjevná úskalí, která vyplývají z různých forem zadávání příkazů, rozsahů nabízených metod a možností či odlišného uživatelského komfortu. Ta musí každý uživatel zvládnout z příslušné dokumentace a vyzkoušet v praxi sám. Pozornost se soustředí na do velké míry skryté a maskované rozdíly v činnostech vyvolaných syntakticky mnohdy totožnými či velmi podobnými příkazy. Zaměřím se na několik programových systémů, domácích i zahraničních, o nichž lze předpokládat, že se s nimi náš uživatel může nyní nebo v budoucnu setkat.

V každém typu empirického výzkumu se nevyhnutelně setkáváme s problematikou chybějících hodnot (1). O specifice této otázky konkrétně v sociologii bylo popsáno již mnoho papírů a jeho objem nemá smysl na tomto místě dále zvyšovat. Pro sociologii – na rozdíl například od některých přírodních věd – jsou typické dílčí analýzy na různé vybíraných a členěných podmnožinách výběrového souboru. Při zpracování dat na počítači používá výzkumník v takovýchto případech často možnosti podmíněných příkazů, prováděných na podkladě vyhodnocení logických podmínek týkajících se přítomnosti či nepřítomnosti určitých vlastností, resp. hodnot veličin daného pozorování (respondenta, objektu ap.). Tyto podmínky bývají nejčastěji formulovány jako implicitní nebo explicitní logické implikace, zpravidla ve formě IF (výraz), IF (výraz) THEN (příkaz), ev. ve spojení s ELSE (příkaz), či SELECT IF (výraz) a podobně. Výrazem je v tomto případě relační operace (tzn. že se zjišťuje zda hodnota veličiny je rovna, nerovna, větší, menší atd. hodnotě jiné veličiny či konstanty), anebo (velmi často) několik takových výrazů spojených operátory logické konjunkce nebo disjunkce (AND nebo OR). Na výsledku podmínky závisí, jakou provede program další akci, tzn. že v případě výběru rozhodne o tom, zda pozorování zařadí nebo nezařadí do analýzy, popřípadě do archivního souboru.

Do okamžiku, kdy ve veličinách použitých ve výrazech nejsou vynechány žádné hodnoty, pracuje většina programových systémů očekávaným způsobem. Potíže nastanou ve chvíli, kdy tomu tak není. V této situaci se akce jednotlivých programů mohou začít značně lišit a uvádět výzkumníka v úžas i zoufalství svou těžko pochopitelnou nevypočitatelností.

Podívejme se nyní blíže na to, jaké situace mohou konkrétně v takovýchto případech nastat. Ve většině programovacích jazyků, jimiž může být ten který program napsán, je výsledkem relační operace (či jejich posloupnosti) buď hodnota pravda (TRUE či numericky zpravidla 1), pokud hodnoty operandů splňují relaci, jinak je hodnota nepravda (FALSE či 0). Třetí možnost zpravidla neexistuje. Tento rámec je však pro systémy analýzy dat příliš úzký, protože neposkytuje prostředek k ošetření takových případů, kdy hodnota jednoho operandu není přítomna, nebo přesahuje povolený obor. Některé jazyky na takovou okolnost nejsou vůbec vybaveny a program ukončí abnormálně svoji činnost. To jistě není z hlediska situací, které mohou v analýze dat často nastat, velmi praktické.

Z těchto důvodů má každý rozumný systém pro analýzu dat opatření, jak se s chybějícími nebo extrémními hodnotami vyrovnat. Jejich konkrétní implementační podoba se systém od systému přirozeně liší. Některé mají opatření již na vstupu dat, například takové, že chybějící hodnoty jsou již předem kódovány speciálním symbolem nebo symboly a výsledkem zpravidla bývají tzv. systémové chybějící hodnoty, spočívající v takové vnitřní reprezentaci, že je systém rozpozná jako nepřipustné a s jako takovými s nimi zachází. Jiné systémy pracují s tzv. uživatelsky definovanými chybějícími hodnotami, kdy se programu sděluje, které hodnoty (či intervaly ap.) má pokládat za chybějící. Podobně lze definovat i obor přípustných hodnot pro každou veličinu a stanovit tak dodatečné podmínky pro vlastnosti dat. Uvedené způsoby jsou nezdídka navzájem kombinovány a to vše má bezprostřední vliv na to, jak systém s těmito hodnotami pracuje.

Původní dvojhodnotová logika, rozeznávající pouze TRUE a FALSE, neboli 0 a 1, bývá proto často rozšířena na trojhodnotovou, obohacenou o možnost nabýt chybějící hodnoty, zpravidla označovanou jako MISSING a nemající zpravidla smysluplnou numerickou hodnotu (ve vnitřní reprezentaci často extrémně vysoká negativní hodnota, např. v programech BMDP „minus nekonečno“). V numerických operacích panuje mezi většinou systémů shoda v tom, že pokus o použití chybějící hodnoty vede k výsledku MISSING, v další analýze nepoužitelnému.

Mnohem menší jednomyslnost však nacházíme v zacházení s chybějícími hodnotami v logických operacích. Prozkoumejme nyní, jak si v těchto případech vedou následující systémy pro analýzu dat: ze zahraničních vybíráme SPSS (verzi 9), SPSS-X (Release 2.2), SPSS/PC+ (Update březen 1987), BMDP (Release 1987), BMDPC (1987 Update), SAS (všechny verze od 79.0) a SAS/PC (Release 6.02), z domácích pak SASD (2), PANDA (3) a DAISY (4).

Projdeme nyní v krátkosti způsoby, jimiž některé z programových systémů zacházejí s chybějícími hodnotami v logických výrazech:

SAS, SAS/PC, SPSS 9, PANDA a SASD přistupují k chybějícím hodnotám v logických výrazech jako k platným, bez ohledu na to, jsou-li systémové (PANDA, SASD), uživatelské (SPSS 9), nebo smíšené (obě verze SAS).

SPSS-X se pokouší navzdory výskytu chybějících hodnot dospět k použitelnému výsledku. Logika systému je taková, že při logické konjunkci (operátor AND) nemůže nikdy nastat hodnota TRUE, je-li jedna z podmínek nesplněna, tzn. i tehdy, obsahuje-li veličina chybějící hodnotu. Výsledek však může nabýt hodnoty FALSE, vyskytne-li se taková hodnota ve druhém výrazu. Při disjunkci (OR) nemůže být výsledek FALSE, objeví-li se v kterémkoli výrazu chybějící hodnota. Pouze v případě, že je chybějící hodnota v takovéto operaci kombinována s pravdivým výrazem, je výsledek TRUE.

SPSS/PC, stejně jako DAISY, konvertuje výsledek všech operací s chybějícími hodnotami na příslušnou hodnotu MISSING (systém DAISY totiž rozlišuje typy veličin – kardinální, ordinální a nominální – a jejich reprezentace chybějících hodnot).

BMDP a BMDPC pracují prakticky na téže principu jako výše uvedený systém SPSS-X, s tím rozdílem, že týmž způsobem, jakým se v logických výrazech přistupuje k hodnotám chybějícím, se zachází i s hodnotami mimo povolený obor, i když v transformacích je možno k chybějícím hodnotám i k hodnotám jak pod, tak nad stanoveným oborem poukazovat speciálními klíčovými slovy.

V následujících tabulkách jsou přehledně shrnuty výsledky logických operací v diskutovaných programových systémech, ve dvojicích výrazů spojených operátory konjunkce a disjunkce.

Tabulka 1. Chybějící hodnoty v logických výrazech spojených operátorem „AND“

Výsledek		Výsledek „Výraz 1 AND výraz 2“			
Výraz 1	Výraz 2	SAS, SAS/PC, SPSS 9, PANDA, SASD	SPSS-X	SPSS/PC	BMDP, BMDPC
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	MISSING	*	MISSING	MISSING	MISSING
MISSING	MISSING	*	MISSING	MISSING	MISSING
FALSE	MISSING	FALSE	FALSE	MISSING	FALSE

* v tabulkách znamená, že kódy chybějících hodnot jsou v takovýchto výrazech pokládány za validní. Výsledek je tudíž TRUE ve všech případech, kdy je podmínka vzdor přítomnosti chybějící hodnoty splněna. Má-li, například, složený výraz formu IF(VARA EQ 2 AND VARB NE 3), a hodnota veličiny VARA byla 2, přičemž VARB by měla pro dané pozorování chybějící hodnotu, byl by výsledek TRUE. Totéž platí analogicky i pro operace „OR“. V případě systému SASD bude výsledek pro poslední tři řádky tabulek FALSE, bude-li hodnota kterékoli použité veličiny uživatelem definovaná jako chybějící.

Tabulka 2. Chybějící hodnoty v logických výrazech spojených operátorem „OR“

Výsledek		Výsledek „Výraz 1 OR výraz 2“ . . .			
Výraz 1	Výraz 2	SAS, SAS/PC, SPSS 9, PANDA, SASD	SPSS-X	SPSS/PC+	BMDP, BMDPC
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
TRUE	MISSING	TRUE	TRUE	MISSING	TRUE
MISSING	MISSING	*	MISSING	MISSING	MISSING
FALSE	MISSING	*	MISSING	MISSING	MISSING

Praktické důsledky

Jaké praktické důsledky může různé vyhodnocování chybějících hodnot v logických výrazech mít v analýze dat? Právě proto, že sociolog velmi často provádí separátní či paralelní analýzy na různých podmnožinách výběrového souboru, musí mnohdy spoléhat na výběr na základě provádění podmíněných příkazů. Takovéto příkazy mají zpravidla podobu složených relačních operací, vyhodnocených jako celek logickým operátorem typu „IF“.

Na reálných datech z výzkumu, kde v jedné z otázek byl častý výskyt chybějící odpovědi, lze ilustrovat, jak rozdílným výsledkům je možno dospět. Tabelujeme dvě veličiny, PDETI a RODINA pro respondenty vybrané na základě vyhodnocení disjunkce dvou dalších veličin, SEX a VEK, která byla zadána ve formě:

SELECT IF (SEX EQ 2 AND VEK NE 4)	programy SPSS
IF (SEX EQ 2 AND VEK NE 4);	programy SAS
USE=IF (SEX EQ 2 AND VEK NE 4).	programy BMDP

Uvedené logické podmínky měly do analýzy vybrat ženy (hodnota veličiny SEX byla v tomto případě 2) s vyloučením určité věkové kategorie. Úkolem bylo na této podmnožině výzkumného souboru tabelovat vztah mezi životem v úplné rodině (v tabulkách veličina RODINA s hodnotami 1 = úplná, 2 = neúplná) a skutečností pečuje-li respondentka o děti v předškolním věku (v tabulkách veličina PDETI s hodnotami 1 = ano a 2 = ne). Údaj o věku byl v tomto případě zjišťován poněkud nevhodným způsobem, což – vzhledem ke zvýšené citlivosti určité části ženské populace na tuto otázku – vedlo k časté absenci odpovědi. Počet neuvedených údajů u znaku SEX byl nesrovnatelně nižší. Tato výzkumnická neomalenost asi příliš nepotěšila zadavatele, nám však tak poskytla vhodný datový materiál pro ilustraci chování některých popisovaných systémů. Ve shora uvedených logických podmínkách pro výběr se proto velmi často vyskytly okolnosti popsané v tabulce 1 na čtvrtém a posledním řádku – tedy kombinace TRUE-MISSING a FALSE-MISSING.

Tabulku 3 vytvořily programy SAS a SPSS 9, tabulku 4 SPSS-X a BMDP a tabulku 5 SPSS/PC+. Obě veličiny byly dvojhodnotové, odpovědi jsou uvedeny pouze číselnými kódy. Údaje v závorkách jsou celková procenta.

Tabulky jsou samy o sobě dostatečně výmluvné. Přestože výchozí datový materiál i analytické strategie byly ve všech případech totožné, získali jsme v závislosti na použitém programu dost odlišné výsledky. Dramatické změny v hodnotách statistiky chí-kvadrát i výrazné přesuny v relativních četnostech jsou bezpochyby varující. Zatímco například

Tabulka 3. Kontingenční tabulka PDETI a RODINA z programů SAS a SPSS 9

		RODINA					
		1		2		celkem	
PDETI	1	1508	[52.3]	96	[3.3]	1604	[55.9]
	2	117	[4.1]	1148	[40.0]	1265	[44.1]
Celkem		625	[56.6]	1244	[43.4]	2869	[100.0]

Chi-kvadrát = 2065.712 Stupeň volnosti = 1

Tabulka 4. Kontingenční tabulka PDETI a RODINA z programů SPSS-X a BMDP

		RODINA					
		1		2		celkem	
PDETI	1	828	[44.1]	55	[2.9]	883	[47.0]
	2	99	[5.3]	895	[47.7]	994	[53.0]
Celkem		927	[49.4]	950	[50.6]	1877	[100.0]

Chi-kvadrát = 1310.709 Stupeň volnosti = 1

Tabulka 5. Kontingenční tabulka PDETI a RODINA z programu SPSS/PC+

		RODINA					
		1		2		celkem	
PDETI	1	430	[38.2]	28	[2.5]	458	[40.6]
	2	63	[5.6]	606	[53.8]	669	[59.4]
Celkem		493	[43.7]	634	[56.3]	1127	[100.0]

Chi-kvadrát = 784.847 Stupeň volnosti = 1

v úplných rodinách s předškolními dětmi ve skutečnosti žilo téměř 53 % respondentek z našeho podsouboru, z tabulky 5 bychom nesprávně soudili na necelých 40 %. Ačkoli na první pohled patrný vztah mezi oběma znaky bychom mohli pokládat za nenáhodný ve všech třech případech, v pokusech o věcnou interpretaci, bychom se asi tabulku od tabulky značně rozházeli.

Uživatel kteréhokoli programového systému musí být vždy mimořádně opatrný a předem si ověřit, jak jeho software pracuje v takovýchto, ne zcela řídkých případech analýzy dat. Pochopitelně, že rozumné programy poskytují uživateli prostředky k tomu, aby docílil výsledky uvedené v tabulce 3, tzn. takové, které do analýzy zahrnují maximum respondentů s požadovanými vlastnostmi – v programech SPSS-X a SPSS/PC+ je to například funkce VALUE ve spojení s uživatelsky definovanými chybějícími hodnotami a v programech BMDP zase speciální systémová veličina XMIS atd.

Závěrem

Programy, o nichž je zde zmínka, pochopitelně nepředstavují úplnou softwarovou varietu, s níž se náš sociolog může nyní nebo v nejbližší budoucnosti setkat. Vyčerpávají nicméně patrně většinu možností přístupů k chybějícím hodnotám užitých v nejrozšířenějších statisticko-analytických systémech. Výběr mimo to ilustruje nejen úskalí, která na výzkumníka číhají při příliš bezstarostných přechodech na programové produkty různé provenience, ale i při změně softwaru od jedné a téže firmy. Právě málo koncepční přístup renomované a u nás známé společnosti SPSS Inc. (5), která poukaz na uvedené rozpory zasunula až kamsi do dodatků svých jinak velmi kvalitních manuálů ("Help for Old Friends" pro SPSS-X a "Help for SPSS-X Users" pro PC), by měl být výstrahou před přílišnou programovou promiskuitou, k níž je sociolog vinou neutěšené situace na softwarovém trhu u nás občas sváděn.

K ověření toho, jak pracuje programový systém, který má analytik právě k dispozici, může posloužit jednoduchý test, jaký si každý uživatel může sestavit sám. Jedna z možností je uvedena v Dodatku. Jeho princip spočívá v tom, že simuluje okolnosti popsané v posledních třech „problémových“ řádcích tabulek 1 a 2.

Dodatek

Vytvoříme jednoduchý vstupní soubor o čtyřech pozorováních a obsahující dvě veličiny, označené například A a B s hodnotami (tečky symbolizují kódy chybějících hodnot):

A	B		A	B
2	.		FALSE	MISSING
1	.	čili	TRUE	MISSING
.	.		MISSING	MISSING
.	1		MISSING	TRUE

Příkazy pro vytvoření nových veličin ověříme způsobem vyhodnocení logických podmínek. Výraz

Z = B NE 3 – zjistí jaké numerické vyjádření mají hodnoty TRUE a MISSING, příkazy

X = (A EQ 1 AND B NE 3) a

Y = (A EQ 1 OR B NE 3) – ověří činnost logické konjunkce a disjunkce, pokud se ve výrazu vyskytne hodnota MISSING. Výpis dat, tj. hodnot veličin A, B, Z, X a Y by měl názorně demonstrovat fungování daného systému a poskytnout tak představu, co od něj může analytik očekávat.

(1) V sociologii, zejména v okruhu pravidelnějších účastníků akcí Sekce metod a technik sociologického výzkumu Čs. sociologické společnosti při ČSAV, se příležitostně souběžně razí termín „vynechané hodnoty“. Z hlediska námětu tohoto článku, totiž fungování počítačových programů, není rozdíl mezi tzv. vynechanou a chybějící hodnotou natolik významný, aby bylo nutno čtenáře zatěžovat dosud neustálenými terminologickými nuancemi. Navíc pojem chybějící hodnota je v souladu se vžitou terminologií počítačové statistiky, dikcí většiny programových manuálů a koresponduje s cizojazyčnými výrazy, např. anglickým "missing" či německým „fehlend“ atd.

(2) Programový systém SASD (Statistická Analýza Sociologických Dat) byl vytvořen péčí Domu techniky ČSVTS pro střediskové počítače řady EC (IBM). Autory systému jsou R. Jiroušek, O. Kříž a J. Řehák, kolektiv řídil F. Hollmann.

(3) Programový systém PANDA (Programy pro ANalýzu DAT) vznikl v Ústavu pro výzkum veřejného mínění při FSÚ pro počítač typu CDC Cyber, v oddělení vedeném P. Dvořákem.

(4) Systém programů DAISY (Data Analysis Interactive SYstem) vytvořil J. Trunda. Vznikl původně pro stolní kalkulátory Hewlett-Packard řady 98×5 a je v současné době převáděn na 16ti bitové osobní počítače.

(5) Vůbec firma SPSS, přestože byla původně orientována na společenské vědy, má tradičně dost potíží při práci s chybějícími daty. Další potíže v této oblasti na výzkumníky čekají při pokusech přenést z programů SPSS-X do programů SPSS/PC a PC+ pomocí příkazů EXPORT a IMPORT.

Literature

- BMDP Statistical Software, Inc.: BMDP User's Guide, 1985. Revised Reprint, University of California Press, Los Angeles 1985 (kapitola 6).
- SAS Institute Inc.: SAS User's Guide: Basics, Version 5. Edition, Cary, NC, SAS Institute Inc., 1985 (kapitola 16: Missing Values in Logical operations).
- SAS Institute Inc.: SAS Language Guide for Personal Computers, Version 6 Edition, Cary, NC, SAS Institute 1985, (Kapitola 12: Missing Values in Logical Operations).
- SASD – uživatelský manuál. Pardubice, DT ČSVTS Pardubice 1984.
- SPSS Inc.: SPSS-X User's Guide, 2nd Edition, Chicago, McGraw-Hill 1986 (8. 19 Missing Values and Logical Operators).
- SPSS inc.: SPSS/PC + for IBM PC/XT/AT, Chicago, SPSS inc. 1986 (B 36, 4.22 Missing Values).
- Zull, C. - Ritter, H.: Die Behandlung fehlender Werte in logischen Ausdrucken bei SPSS and SAS: Eine Warnung vor unerwarteten Ergebnissen, In: ZUMA Nachrichten Nr. 21 (November 1987).

Резюме

M. Čákrт: Statistické programy a chybějící hodnoty: varování před neočekávanými výsledky

Programové systémy pro statistickou analýzu dat dnes jsou již nezbytným nástrojem empirických sociologických výzkumů. V datech, získaných sociologem na základě výběrových šetření (sample surveys) v terénních podmínkách, téměř vždy mají místo chybějící hodnoty (missing values). Pro sociologii jsou typické částečné srovnávací analýzy na vybraných a klasifikovaných podskupinách. Při provádění této práce využívá výzkumník velmi často podmíněných výrazů (conditional expression), jež slouží k posouzení logických podmínek, určujících přítomnost určitých vlastností, nebo hodnot proměnných daného pozorování, respondenta a t. p. V případě, že v takovém podmíněném výrazu se objeví chybějící hodnota, různé programové systémy pracují různě.

Článek rozebírá a rozvíjí na srozumitelný způsob rozdíly v činnosti některých domácích a nejznámějších zahraničních programových produktů (SPSS, SAS, BMDP) a upozorňuje na možnosti neočekávaných způsobů fungování těchto systémů a nepředvídaných výsledků.

Summary

M. Čákrт: Statistical Programmes and Missing Values: Warning against Unexpected Results

Programme systems for statistical data analysis nowadays represent an indispensable tool of empirical sociological research. In the data obtained by the sociologist from sample surveys in the field there are almost always some missing values. Typical of sociology are partial comparative analyses carried out on variously selected and classified subsamples. In this work the researcher very often makes use of conditional expressions applied on the basis of the evaluation of logical conditions ascertaining the presence of certain properties or values of the variables of the given observation, of the respondent, etc. In case a missing value appears in such a conditional expression, various programme systems operate in different ways.

The paper is concerned with and lucidly elaborates the differences in the activities of some home and foreign programme products which are best known here (SPSS, SAS, BMDP), and indicates the possibilities of unexpected ways of functioning of these systems as well as of unforeseen results.