DVA INTERAKTIVNÍ STATISTICKÉ SYSTÉMY PRO SMEP Z POHLEDU UŽIVATELE

Josef Tvrdík

Úvod

Na začátek tři provokativní tvrzení podepřené (dosti silně) zkušenostmi:

a) V analýze empirických dat se užívají především ty metody, jejichž aplikace není příliš pracná a časově náročná. Tzn. ty metody, které jsou podporovány dostupnými programy. Hledisko přiměřenosti metody, robustnosti, efektivnosti algoritmů a splnění statistických předpokladů je většinou až druhořadé.

b) S dobrým programovým vybavením (tj. dobrým statistickým programovým systémem) může podstatnou část úloh analýzy dat vyřešit uživatel sám.

c) Užívání dobrého programového systému pro analýzu dat má významný vedlejší pedagogický efekt. Umožňuje uživateli prakticky vyzkoušet a relativně rychle poznat různé metody předzpracování a popisu dat, vybrat z možností v rámci dostupného statistického programového systému nejvhodnější procedury pro danou úlohu a vybavuje ho znalostmi i slovníkem užitečnými pro konzultace s odborníkem - statistikem při aplikaci náročnějších statistických metod.

Příspěvek se zabývá uživatelskými zkušenostmi se dvěma interaktivními statistickými systémy implementovanými na počítačích SM4/20 (a ekvivalentních) pod operačním systémem DOS RV (RSX-11). Jsou to programový systém MSTAT vyvíjený na UK Bratislava /1/ a program ISP vyvinutý ve VÚLB Bratislava /2/. Systém MSTAT je (nebo bude) dodáván Datasystémem, ISP je distribuován řešitelským pracovištěm, zdravotnickým organizacím zdarma. Příspěvek vznikl až po dlouhém váhání, neboť řada okolností nabádala spíše řídit se příslovím "Darovanému koni na zuby nehleď" Pro presentování příspěvku nakonec rozhodly tyto důvody:

- stále není u nás běžně dostupný obecný statistický systém programů;
- jsou (snad) v dohlednu výsledky výzkumného úkolu zabývajícího se vývojem statistického software pro počítače řady SMEP (tzv. SMEPPACK);
- jedním z anotovaných témat ROBUSTu 86 je statistický software;
- oživení systémů MSTAT a ISP si vyžádalo řadu hodin práce a přesto nevedlo k očekávanému výsledku;
- neexistují (nebo autorovi příspěvku nejsou známé) univerzální a jednoznačné návody pro návrh interaktivních statistických (ani jiných) programů.

Zkušenost ukazuje, že kritické posouzení vlastností existujících programových produktů je pro návrh software velmi důležité. Argumentace příkladem je velmi závažná a výmluvná, neboť pro návrh software nejsou k dispozici obecné a dokazatelné pravdy jako např. v matematice. Návrhem software se jako červená nit táhne kompromis. Pokud navrhovaný software má být dobrý (tj. vyhovovat vysloveným i nevysloveným požadavkům uživatelů), je nutné, aby kompromisy přijímané v návrhu byly kompromisy kvalifikované. A tato kvalifikovanost se opírá především o zkušenosti s návrhem a využitím podobných programových produktů.

2. Systém MSTAT

MSTAT je systém programů, spojených povelovým souborem v jeden funkční celek. Komunikace uživatele se systémem tedy probíhá ve dvou úrovních: jednak s povelovým souborem, jednak s programy. Způsob komunikace je však podobný a přechody mezi úrovněmi nepůsobí rušivě.

Systém MSTAT se skládá z programu HELPST pro předzpracování dat a sedmi programů pro některé úkoly statistické analýzy dat. Program HELPST instruuje uživatele o způsobu další komunikace se systémem MSTAT a umožňuje vytvoření typizovaného souboru dat (SAVEFILE), který je pak vstupním souborem programů statistické analýzy. Zpracovávaná data do programu HELPST (a tedy i do celého systému MSTAT) vstupují ze sekvenčního znakového souboru s větami pevné délky. Tento soubor musí být připraven předem např. programem EDI nebo GOLEM. Max. počet veličin je 256, popis formátu vstupních dat (obvyklé fortranské konverze F a A) nesmí přesáhnout 80 znaků. Data nesmí obsahovat chybějící údaje (MISSING). Program HELPST dovoluje zadat věty a typ veličin. Jsou rozlišovány 4 typy veličin - dichotomické, nominální, kardinální a identifikační (analogie LABEL v BMDP).

Dále má program některé funkce předzpracování dat (editování, transformace veličin), a výpočet základních statistik (průměr, směrodatná odchylka, minimum, maximum u kardinálních veličin a frekvence jednotlivých hodnot u nominálních a dichotomických veličin).

Výsledky vystupují na obrazovku, program nabízí možnost tzv. protokolování, tj. zápisu některé části dialogu a výsledků do dočasného souboru na disku. Tento soubor lze před skončením běhu systému MSTAT vypsat na tiskárnu.

Vstupem pro statistické programy je typizovaný soubor dat, vytvořený programem HELPST. Některé statistické programy mají další omezení na tvar dat a počet proměnných. Součástí systému MSTAT jsou tyto statistické programy:
(názvy odpovídající nabídce systému)

Základní statistiky a testy - obsahuje různé t-testy. Prakticky využitelné jsou však jen jednovýběrový a párový t-test, neboť dvouvýběrovou úlohu nelze programu prostředky systému popsat.

Regresní analýza - obvyklá mnohorozměrná lineární regrese, max. počet nezávislých proměnných je 20. Výsledky jsou přehledně uspořádány v tabulce, neobsahují však hodnoty pravděpodobnosti odpovídající t-statistikám jednotlivých proměnných. Na vyžádání je dostupná analýza rozptylu a koeficient mnohonásobné korelace.

Diskriminační analýza - max. 20 veličin. Počty ve skupinách musí zadávat uživatel. Výsledky obsahují jen koeficienty lineární diskriminační funkce a Mahalanobisovu vzdálenost skupin.

Faktorová analýza - pozorování musí být nezávislá. Pouze extrakce faktorů, lze zadat jejich počet a změnit odhad komunalit.

Analýza rozptylu - shodný počet pozorování ve skupinách, komplikované zadávání popisu modelu.

Neparametrické metody - obsahuje Kolmogorov-Smirnovův test shody dvou rozdělení pro různé případy, Kendallův a Spearmanův pořadový korelační koeficient, znaménkové testy. Mezi neparametrické metody je zařazen i x^2-test nezávislosti v dvourozměrné kontingenční tabulce; tento program však pouze stanoví četnosti v jednotlivých políčkách, zřejmým nedopatřením chybí výstup hodnoty statistiky.

Grafické průběhy - výstup histogramů nebo grafu závislosti dvou proměnných na alfanumerickém výstupním zařízení.

Všech sedm statistických procedur vypisuje výsledky na obrazovku a zapisuje je i do dočasného souboru, který lze před ukončením běhu systému vytisknout na tiskárnu. Komunikaci se systémem ztěžuje to, že názvy těchto dočasných souborů nebyly vytvořeny podle jednotných pravidel (patrně tento nedostatek bude

odstraněn v další verzi MSTAT) a zejména odlišný způsob komunikace v programu Grafické průběhy od všech ostatních programů MSTATu. V tomto programu jsou data z klávesnice snímána okamžitě (způsob INKEY). Přechod do a z tohoto režimu je pro uživatele obtížný. Obecně pro komunikaci ve statistických programech režim INKEY považuji za nevhodný.

3. ISP

Systém ISP je tvořen několika úlohami (TASK), které jsou volány přímo z programu ISP. Vstupními daty je znakový soubor zapsaný volným formátem. Oddělovačem položek je mezera, čárka nebo tabulátor. Vstup dat do systému ISP zajišťuje modul INPUT, který umožňuje interaktivně specifikovat vstupní data a název problému, ze kterého se odvodí jména tří systémových souborů s daty a doplňujícími informacemi k problému (analogie SAVEFILE).
Modul INPUT umožňuje i vstup dat z výstupních tiskových souborů dotazovaného programu DATATRIEVE (DTS, DTR). Toto je z mnoha hledisek výhodná vlastnost systému ISP, neboť pomocí DATATRIEVE lze pohodlně provést řadu funkcí předzpracování dat (vstup dat, editování, některé transformace, výběr veličin a případů k dalšímu statistickému zpracování).

Vstupní data ISP musí vyhovovat omezením, z nichž některá jsou při praktickém použití ISP velmi nepříjemná. Počet veličin nesmí přesáhnout 20, maximální délka vstupní věty je 132 B. Systém ISP neumí pracovat s daty obsahujícími chybějící hodnoty (MISSING).
Dále systém ISP neobvyklým způsobem ošetřuje nulové hodnoty v datech. Program INPUT nabízí jejich náhradu průměrem nebo mediánem, některé výpočetní programy (např. lineární regrese) vůbec nulové hodnoty datových položek nedovolují.

Komunikace se systémem ISP probíhá v angličtině. Volbu angličtiny zdůvodňují autoři systému nejednoznačností slovenské a české statistické terminologie. Bohužel právě příklad ISP ukazuje, že jednoznačná statistická terminologie není postačující podmínkou přehledného dialogu. Stejně důležitá je i volba vhodných textů (a jejich uspořádání na obrazovce) s informací o tom, v které úrovni a na které větvi dialogového stromu právě jsme. Tuto podmínku ISP často nesplňuje.

Ze statistických procedur obsahuje ISP výpočty základních statistických charakteristik (průměr, medián, kvantily, rozptyl, šikmost, špičatost, histogram), různé t-testy, znaménkový test, Kolmogorov-Smirnovovy testy dobré shody, analýzu rozptylu, faktorovou analýzu, různé programy pro lineární regresi (pro 2 proměnné, mnohorozměrnou lineární regresi, výběr nejlepšího regresního modelu) a regresi polynomem.
Seznam statistických procedur je dosti široký a pokrývá řadu běžných uživatelských požadavků. Použitelných je však podstatně méně programů systému ISP, neboť v mnoha programech jsou různé chyby od metodických (např. ošetření nulových hodnot) přes komunikační (dialog neodpovídá skutečnému průběhu programu) až po výpočetní (např. chybné frekvence v histogramech, numericky zcela chybné hodnoty vypočtených statistik a pod.). Dalším pro praktické používání dost podstatným nedostatkem ISP jsou nepřehledné a neuspořádané výsledkové sestavy. Výsledky jsou rozházeny po velké ploše, často chybí tabulková či jiné pro interpretaci vhodné hutné uspořádání. Někdy dokonce slovní označení hodnoty ve výsledcích neodpovídá skutečnosti.

Tento výčet nedostatků a chyb ukazuje, že současnou verzi systému ISP nelze doporučit k rutinnímu používání.

4. Závěr

Ze stručně popsaných uživatelských zkušeností je zřejmé, že systémy MSTAT a ISP se odchylují od požadavků na dobrý statistický programový systém vágně a trochu skrytě obsažených v provokativních tvrzeních v úvodu. Pomineme přirozené obecné požadavky na vlastnosti programového vybavení jako numerická správnost vypočítaných hodnot a vyřešení všech větví programu odpovídajících dovoleným volbám uživatele, kterým zejména systém ISP až příliš často nevyhovuje. Pro návrh interaktivních statistických systémů se ve světle uvedených zkušeností ukazují důležité i další vlastnosti programů:

- jednotný styl komunikace v rámci celého systému
- stručnost, přehlednost a jednoznačnost dialogu
- přehlednost a obsažnost výsledků pro potřeby interpretace (hodnoty pravděpodobnosti pro vypočtené statistiky, tabulkové a grafické uspořádání výsledků apod.)
- zpracování dat s chybějícími položkami (MISSING)
- možnosti jednoduchého zadávání úloh dvou a více výběrů
- a další .

U statistických systémů se interakce může týkat několika oblastí:

- interaktivního zadávání vstupních parametrů
- interaktivní interpretace výsledků
- interaktivního řízení výpočtu

Domnívám se, že pro programy určené širšímu použití je postačující interaktivní podpora zadávání vstupních parametrů. To je oblast, ve které se občasný uživatel při formulaci své úlohy počítači dopouští nejčastěji drobných chyb, které by mu interaktivní podpora měla umožnit rychle opravit, případně mu poradit ve výběru možností ("cesty" programem). Požadavek podpory interpretace výsledků lze většinou uspokojit přehlednou grafickou úpravou a obsažností tištěných výstupů. Interakce s vlastní výpočetní procedurou je určena pouze kvalifikovanému uživateli, který je schopen podle dílčích výsledků rozhodnout o dalším postupu výpočtu. Pro většinu běžných statistických metod tedy není nutná a zřejmě ani žádoucí.

Literatura:

1. Manuál systému MSTAT, Katedra aplikované matematiky UK Bratislava
2. Manuál ISP, Výzkumný ústav lékařské bioniky, Bratislava