INTERAKTIVNÍ STATISTICKÝ PROGRAMOVÝ SYSTÉM

(příklad, zkušenosti, diskuse)

J. Tvrdík, H. Hájková, KHS Ostrava

Cílem tohoto příspěvku je upozornit na některé technické otázky vývoje programového vybavení pro statistickou analýzu dat. Na příkladu programového systému STATIS jsou ukázány některé užitečné funkce, které by měly statistické soubory programů obsahovat. Na základě zkušeností se systémem STATIS jsou diskutovány další požadavky na programové systémy pro statistickou analýzu dat a vysloveny doporučení pro jejich návrh a programové řešení.

Programový systém STATIS

STATIS je interaktivní programový systém pro jednoduché metody statistické analýzy dat. Je naprogramován v jazyku BASIC pro stolní mikropočítač TRS III s operačním systémem NEWDOS 80 /1/. Tento mikropočítač umožňuje využívat přibližně 38 kB operační paměti, má dvě jednotky pružných disků 5 1/4 palce, klávesnici, alfanumerický displej s možností hrubého grafického zobrazení (128 x 48 bodů). K mikropočítači je připojena mozaiková tiskárna (80 znaků/řádek).

Celý systém STATIS má asi 2 000 řádků zdrojového textu a je i s potřebnými moduly operačního systému uložen na dvou disketách. Anotační karta programového systému STATIS je uvedena v /2/.

Při návrhu systému STATIS byly vysloveny tyto nejdůležitější požadavky na vlastnosti řešeného souboru programů:

- obsluha počítačově nezkušeným uživatelem
- možnost uchování vložených dat
- možnost úpravy dat (opravy, změna uspořádání dat, dodatečné vkládání dat)
- opakované zpracování těchže dat různými statistickými postupy
- dovolit chybějící hodnoty (missing) v datech
- zařadit především jednoduché procedury použitelné nestatistikem
- jednoduché rozšiřování systému o další procedury.

S ohledem na technické charakteristiky mikropočítače TRS III je systém určen pro zpracování nepříliš rozsáhlých dat (počet případů x počet veličin $\leq$ 4 000).

Tyto požadavky v podstatě zformulovaly vlastnosti řešeného systému. Celý soubor programů STATIS se uživateli jeví jako jediný program, jehož běh lze spustit jedním příkazem. Další komunikace uživatele s počítačem probíhá jako dialog inicializovaný počítačem. Uživatel většinou odpovídá volbou z "menu" nabízeného na obrazovce nebo vyplňováním formuláře s možností "default" hodnot. Jedna z voleb úvodní obrazovky je tisk jednostránkového návodu a stručného popisu funkcí systému. V některých situacích lze vyžádat podrobnější návod k využití zvoleného postupu.

Funkce systému STATIS lze rozdělit do dvou skupin (podrobněji - viz 3 manipulace s daty a statistická analýza dat. Do první skupiny patří tyto funkce:

VSTUP DAT	umožňuje vkládání dat z klávesnice (buď po případech nebo po veličinách) a vytvoření datového souboru (přechodného nebo trvalého) na pružném disku
EDIT & LIST	- opravy kterékoliv datové položky, přidávání a vypouštění případů nebo veličin - výpis dat ve zvoleném formátu na obrazovku nebo na tiskárnu
SUBFILES	rozdělení dat do 2 až 12 podsouborů (podsouborem se rozumí nepřerušovaná posloupnost případů z datového souboru)

SORT	setřídění případů podle hodnot zadané veličiny
NAME	přejmenování úlohy, veličin nebo podsouborů, vytvoření kopie souboru
JOIN	spojení dvou dříve vytvořených datových souborů do jednoho buď přidáním objektů nebo přidáním veličin
RECODE	vytvoření kategorií podle zadaných bodů řezu (cutpoints)
TRANSF	vytváření odvozených veličin podle uživatelem zadané sekvence příkazů v jazyku BASIC

Uživatel vytváří základní datové soubory pomocí funkce VSTUP DAT. Ostatní popsané funkce čtou tyto soubory a po provedení interaktivně zadaného zpracování dat vytvářejí nový opravený soubor, který může být dále zpracováván. Výběr funkcí pro manipulaci s daty v systému STATIS odpovídá např. možnostem poskytovaným v HP STATPACK na počítači Hewlett-Packard 9845 /4/.

Do programů statistické analýzy dat vstupují datové soubory vytvořené manipulačními programy. Řídící parametry se vkládají dialogově z klávesnice. Uživatel zadává, zda se má zpracovat celý soubor nebo jen některé podsoubory. Výsledky statistických programů jsou vždy tištěny na tiskárně, aby byly uživateli k dispozici i po ukončení komunikace s počítačem. K vypočítaným statistikám se tiskne odpovídající hodnota pravděpodobnosti. Ve výstupu je využito názvů souboru, veličin, podsouborů a úlohy, takže výsledky jsou přehledné i v případě, kdy je provedeno více výpočtů za sebou. Ke statistické analýze dat jsou zatím implementovány programy:

BASISTAT	výpočty běžných jednorozměrných statistik (výběrový průměr, rozptyl, směrodatná odchylka, minimum, maximum, z-skóre minima a maxima, na vyžádání se počítá šikmost, špičatost, medián, kvartily a tiskne histogram)
STUDENT	jednovýběrový, dvouvýběrový a párový t-test na veličinách zadaného souboru dat, resp. výpočet statistiky T na základě výběrových průměrů a směrodatných odchylek zadaných z klávesnice
ANOVA	jednoduché analýza rozptylu
CHIKVAD	test nezávislosti veličin ve dvourozměrné kontingenční tabulce R x C. Tiknou se i očekávané četnosti a adjustované standardizované residuály, u tabulek 2 x 2 Yuleovo Q, Yatesova korekce a na vyžádání Fisherův exaktní test. Vstupem tohoto programu může být buď datový soubor, ze kterého je kontingenční tabulka programem vytvořena, nebo četnosti zadávané z klávesnice
REGRESE	provádí se volitelně výpočet korelační matice zadaných veličin a vyhodnocení lineární regrese 2 veličin, která je graficky zobrazena na displeji
NEPARAM	jednoduché neparametrické metody /5/: mediánový a Wilcoxonův test, Kruskal-Wallisův a Friedmanův test, pořadové korelace.

Pro další rozšiřování systému STATIS o procedury analýzy dat je možné využívat program PANELY, který obsahuje podprogramy pro čtení dat, opakující se úseky interaktivního řízení chodu programu a podprogramy na výpočet hodnoty distribuční funkce běžných rozdělení podle /6/. Důslednější modularitu (např. knihovnu podprogramů) jazyk BASIC na počítači TRS III nedovoluje, neboť mohou být užívány jen vnitřní procedury bez parametrů. Všechny proměnné v tomto BASICu jsou globální. Tato skutečnost je jedním z nejpodstatnějších omezení dalšího rozšiřování STATISu o složitější statistické procedury. Pomalost výpočtu interpretovaných BASICovských programů téměř vylučuje využití mnohorozměrných metod na TRS III v rámci systému STATIS.

Zkušenosti z užívání STATIS

Programový systém STATIS rok po jeho implementaci využívá příležitostně asi 20 uživatelů. Z nich asi polovina je schopna samostatně racionálně využít nabízené možnosti při manipulaci s daty.

Od začátku provozu bylo nutno provést několik menších úprav v textech zpráv, které jsou součástí komunikace počítače s uživatelem. Šlo především o užití některých pojmů práce se soubory, které byly uživateli neprogramátorovi a nestatistikovi nesrozumitelné, o některá opomenutí nebo naopak nadbytečné dotazy, které vznikly při návrhu nesprávným odhadem psychologie uživatele, nebo o významové posuny způsobené nepřítomností diakritických znamének v českém textu apod.

Osvědčil se dialogový způsob komunikace inicializované počítačem. Tento způsob zadávání řídících parametrů výpočtu považujeme z hlediska uživatele (zejména nepravidelného) za výhodnější, než komunikaci prostřednictvím parametrického řídícího jazyka běžného u dávkových systémů - viz např. /7,8/. Výhodné je pořizování tištěných výstupů se statistických procedur, neboť vytištěné a grafické výstupy umožňují uživateli ve srovnání s výstupem na obrazovku důslednější úvahu nad právě ukončeným krokem analýzy dat a lépe dokumentují postup řešení. Interaktivní systémy by obecně měly mít rychlou odezvu na požadavky uživatele. Tu však nelze někdy z různých důvodů (pomalost technického vybavení, výpočetní složitost) zajistit. V situacích, kdy odezva není téměř okamžitá se v systému STATIS osvědčila odpověď počítače s informací o tom, že požadavek uživatele je zpracováván, ale že je třeba čekat; o průběhu zpracování počítač podává občas uklidňující zprávy. Malý rozsah zpracovávaných dat a dosavadní implementace pouze elementárních statistických procedur nejsou závažnými omezeními současných uživatelů STATISu.

Nejzávažnější nevýhodou STATISu se ukázala datová nekompatibilita systému (a mikropočítače TRS III)s ostatními systémy statistické analýzy dat na jiných počítačích. Domníváme se, že každý systém statistických programů má umožňovat manipulaci s daty. Tudíž každý takový systém by měl umožňovat i vytvoření výstupního souboru z předzpracovaných dat ve formě čitelné jinými statistickými systémy.

Dalším podstatným nedostatkem je malá portabilita systému STATIS. STATIS je snadno přenositelný na mikropočítač TRS III s operačním systémem NEWDOS 80 u česky rozumějících uživatelů.Podle našich informací tedy snadno přenositelný není asi nikam. Převod systému STATIS na jiný mikropočítač je obtížnější; zatím se uskutečnil převod na HP 85 a vyžadoval několik týdnů práce programátora. Tyto nedostatky však nebylo možno ovlivnit řešením systému STATIS.

Obecně požadované vlastnosti statistických systémů

V tomto odstavci se pokusíme shrnout naše zkušenosti a vyslovit návrhy pro programové řešení statistických systémů. Tyto návrhy rozhodně netvoří nějakou ucelenou, nediskutovatelnou a bezrozpornou soustavu.

Systém statistických programů má obsahovat funkce pro předzpracování (manipulaci) dat. Manipulační procedury lze od statistických účinně oddělit. Jejich propojení pak umožňuje vhodně definovaná struktura systémového datového souboru (tzv. savefile). Důležitou vlastností statistického systému je jeho otevřenost, a to ve dvojím smyslu:

a) datovém (oboustranná datová kompatibilita s jinými systémy)
b) funkčním (snadné vytváření nových procedur, které rozšiřují funkční možnosti systému).

Pro návrh a řešení statistického (i jiného) programového systému jsou závažné především tyto faktory:

- technické vybavení
- obecné programové vybavení (podpora komunikace a grafických výstupů, numerické podprogramy apod.)
- třída úloh, která má být pomocí systému řešena
- uživatelé (i potenciální)
- zdroje, které jsou pro řešení k dispozici (počet řešitelů, jejich erudice, čas apod.).

Tyto faktory zdaleka nejsou nezávislé, některé působí proti sobě (např. třída úloh vs uživatelé). Z toho vyplývá, že dobrý (nebo alespoň opakovaně použitelný) statistický programový systém může vzniknout pouze kompromisem kvalifikovaně posuzujícím vlivy těchto faktorů na řešení již od jeho počátečního stadia formulace požadavků na úroveň portability, na dokumentaci a na údržbu a rozšiřování programového systému.

LITERATURA

1. TRS-80 Model III Operation and BASIC Language Reference Manual, Tandy Corp., Fort Worth, Texas 1980
2. Tvrdík, J., Hájková, H.: Interaktivní systém pro jednoduché metody statistické analýzy dat, In: sborník semináře Využití počítačů v lékařství a zdravotnictví, DT ČSVTS Praha 1984
3. Hájková, H.: Programový systém pro analýzu dat, diplomová práce, VŠB Ostrava 1983
4. HP STATPACK-soubor programů pro statistickou analýzu dat na HP 9845
5. Anděl, J.: Matematická statistika, SNTL Praha 1978
6. Olehla, M., Věchet, V., Olehla, J.: Řešení úloh matematické statistiky ve FORTRANU, NADAS, Praha 1982
7. Dixon, W.J. (red.): BMDP - Biomedical Computer Programs, University of California Press, Los Angeles 1975
8. Havránek, T.: The present state of the GUHA software, Int. J. Man-Machine Studies 15, 253-264, 1981
9. Myers, G.J.: Composite/Structured Design, Van Nostrand Reinhold Comp. 1978
10. Čimbura, V., Tvrdík, J.: Problémy návrhu modulárních programů, In: sborník semináře Programování 80, DT ČSVTS Ostrava 1980