STATISTICKÉ PROGRAMY A JEJICH UŽIVATELÉ

J. Tvrdík, VVUÚ Ostrava-Radvanice

Abstract:

The paper deals with the analysis of real data. An artificial example is referred in order to illustrate the difference between textbook tasks and the real data analysis. The properties of the practical tasks of data analysis are discussed. Those properties could be used for classiffication of the tasks. According to author's experience four kinds of the tasks can be distinguished. The quality of user seems to be the most significant variable of classiffication. The role of taylored statistical software is briefly mentioned. The purpose of the paper is to open the topic to the more profound discussion.

1. Učebnicový příklad na úvod

Osecký /1/ uvádí jeden příklad analýzy dat, který ve stručnosti lze shrnout:

Úloha: rozpoznat, zda proutkař umí najít vodu

Pokus: 10 dvojic zakrytých věder, voda vždy jen v jedné nádobě z každé dvojice

Data: počet správně určených plných věder

Model: binomické rozdělení

$$P(X \leq x \mid n, \pi) = \sum_{t=0}^{x} \pi^t (1 - \pi)^{n-t},$$

$n = 10, \quad \pi = 0{,}5$

Výsledek:

t	P	1-P
0	0.001	0.999
1	0.010	0.989
8	0.044	0.055
9	0.010	0.011
10	0.001	0.001

Interpretace: Je-li $1-P \leq \alpha$, pak zamítneme hypotézu "náhody". Při obvyklé hladině významnosti $\alpha = 0.05$ to znamená, že jestliže počet správně určených plných věder je 9 nebo 10, pak můžeme věřit na umění proutkaře.

Po přečtení tohoto příkladu se mnou zalomcovala závist, vztek i lítost současně - jiní mají úlohy, které lze bez pochybnosti vyřešit, zatímco já mám tu smůlu, že se potkávám jen s úlohami, které pochybnosti přímo generují. Po podrobnějším přečtení okolního textu negativní pocity vyprchaly: uvedený příklad je fiktivní, jen krásně vymyšlený.

2. Analýza reálných dat a co k tomu patří

Obraťme pozornost k úvaze o úlohách praktických, kdy data jsou obrazem části reálného světa a výsledky analýzy dat neslouží jen jako učebnicové

příklady. Použijeme-li k tomuto rozboru terminologie analýzy dat, pak objekty (případy) jsou úlohy. Určit veličiny je už složitější, za první pokus považujeme tento seznam: data, uživatel, cíl úlohy, řešitel, metody analýzy dat, statistický software, čas na řešení, interpretace výsledků.

Pokusme se uvedené veličiny zhruba popsat a případně i kategorizovat:

DATA - měla by být obrazem výseku reálného světa, o kterém mají výsledky úlohy vypovídat, případně z výsledků inferovat obecnější tvrzení. Zdaleka ne všechno umíme změřit. Struktura dat by měla odpovídat struktuře zobrazovaného kusu světa. Metody analýzy dat však většinou pracují s datovou strukturou ve tvaru matice. Není už v tomto zobrazení hlavní příčina všech dalších pochyb o výsledku řešení?
Domnívám se, že následující kategorizace dat není přehnaně skeptická:
- průhledně jednoduchá
- zjevně defektní
- skrytě defektní

UŽIVATEL - člověk nebo skupina, zajímající se o část světa zobrazenou v datech (jeden malý, ale významný klan čs.analýzy dat pro uživatele používal pracovní označení "GUHA MUDr."). Uživatel se chce dopracovat nějakých výsledků analýzou svých dat. Slušný uživatel hledá pravdivá nová tvrzení. Uživatel je prapříčina i hlavní zdroj obtíží řešitele úloh analýzy reálných dat. Pro hrubou klasifikaci uživatelů použijeme dvou dichotomických veličin (nebo chcete-li, párových kategorií) s těmito ohodnoceními:
soudný - hochštapler
úporný - laxní

CÍL ÚLOHY - Veličina "cíl úlohy" silně souvisí s veličinou "uživatel", neboť uživatel cíle formuluje (byť někdy dost nezřetelně).
Pohlédneme-li nad ty cíle, které formuluje uživatel, pak můžeme rozeznat tyto kategorie:
- splnit formální požadavky časopisu
- splnit požadavky uživatelova šéfa
- cokoliv (uživatel umí učinit efektní závěry z libovolných výsledků)

ŘEŠITEL - (GUHA Ing.") je další osobou v řešení úlohy, tedy i nutný zdroj konfliktů. Pro řešitele je důležitá celá řada vlastností a dovedností jako
- statistické vzdělání
- chuť k aplikacím
- zručnost v počítačovém zpracování dat
- porozumění a trpělivost v komunikaci s uživatelem

Uvedené požadované vlastnosti vysvětlují, proč mezi řešiteli je dosti nestatistiků. Statistik s chutí k aplikacím je bytost u nás velice vzácná, skoro pohádková. Obecně je to škoda, ale díky tomu i my ostatní máme obživu.

Dramatická situace nastává, když uživatel je současně i řešitelem:

pokud nejde o opravdu jednoduchou úlohu, pak to bývá něco mezi bloudením v labyrintu a táborákem u benzinové pumpy.

METODY AD - připomenu jen některé vyčnívající kategorie:
- módní
- uživatelem vyžadované
- řešitelem preferované

Podle mé zkušenosti je podstatné, zda jsou metody podporované statistickým softwarem dostupným řešiteli, neboť co nelze použít snadno a pohodlně, to se většinou nepoužívá.

SOFTWARE - o klasifikaci statistických programů existuje mnoho článků a několik tlustých knih. Zde jen připomeňme kategorie software podle způsobu jeho nabytí (je to důležitá vlastnost software ve vztahu k množství pochybností spojených s analýzou reálných dat):
- legálně získaný
- kradený
- doma udělaný

S přáním, aby druhá kategorie vymizela a statistické programy dělané "doma" vznikaly pouze tehdy, kdy jsou k tomu důvody opravdu rozumné.

ČAS NA ŘEŠENÍ - vyřešení mnoha úloh je časově omezeno. Po určitém termínu ztrácí sebelepší řešení praktický smysl. Formulace cílů a postup řešení musí odpovídat časovým možnostem. Je to záležitost kvalifikovaného kompromisu. Přívlastek "kvalifikovaný" znamená mimo jiné i to, že řešitel se nenechá dohnat na neřemeslnou úroveň řešení.

INTERPRETACE VÝSLEDKŮ - veličina silně závislá na uživateli, řešiteli a jejich vzájemném vztahu. Interpretací výsledků se od zpracování datového modelu obracíme zpátky k reálnému světu. Výstupem ze statistických programů jsou tabulky, čísla a grafy o datovém modelu. Uživatele však zajímají tvrzení o reálném světě. Chce vyjádřit své závěry i jinými prostředky. Zde číhá nebezpečí neoprávněné generalizace, zbytečné vágnosti nebo dokonce úplně nesmyslného zkomolení výsledků.

3. Trocha obecných prohlášení

Popis veličin rozpoznatelných na úlohách analýzy reálných dat naznačuje, že téma příspěvku "statistické programy a jejich uživatelé" souvisí s řadou dalších nesnadno uchopitelných veličin a dějů. Domnívám se, že statistický software by se měl stále více přibližovat potřebám řešení různých typů praktických úloh. Měl by svým uživatelům nejen usnadňovat práci, ale i přiměřeným způsobem rozvíjet jejich konceptuální představy o možnostech analýzy dat. Cestu k opuštění stavu "mnoho dat - málo jejich analýzy" spatřuji nejen ve vývoji obecných statistických programových systémů, ale především ve vytváření na míru střižených programů pro určité typy úloh a pro určité třídy uživatelů-nestatistiků-viz na př. /2/.

4. Trocha empirie

S použitím veličin z odst. 2 uvedu čtyři typy úloh, se kterými jsem se v posledních létech opakovaně setkával:

a) řešitelova noční můra

uživatel:	úporný hochštapler
data:	rozsáhlá, defektní
cíl:	cokoliv
průběh:	nádenická práce řešitele, přerušovaná hlasitými hádkami s uživatelem
výsledek:	hromada potištěného papíru do šuplíku, posléze do sběru

b) klidnější varianta

uživatel:	laxní soudný
data:	jiná než průhledně jednoduchá
cíl:	požadavek uživatelova šéfa
průběh:	nádenická práce řešitele, přerušovaná občasnými rozhovory s uživatelem týkajícími se řešené úlohy jen vzdáleně
výsledek:	hromada potištěného papíru do šuplíku, posléze do sběru

c) lahůdka líného řešitele

uživatel:	laxní hochštapler
data:	nejsou potřeba
cíl:	zvýšení uživatelovy důležitosti
průběh:	krátké rozhovory uživatele a řešitele při náhodných setkáních na chodbě , WC (jsou-li oba stejného pohlaví) a pod.
výsledek:	neodůvodněné zvýšení respektu obou zúčastěných, naštěstí jen v jejich nejbližším okolí

d) dobré řemeslo

uživatel:	úporný soudný
data:	průhledně jednoduchá, rozsáhlá
cíl:	zformulovaný uživatelem i řešitelem společně
čas:	bez tvrdých omezení
průběh:	normální, převážně příjemný
výsledek:	většinou užitečný

Zkušenost ukazuje, že pro klasifikaci úloh analýzy reálných dat je důležitější veličina "uživatel". Uvedené čtyři typy úloh odpovídají čtyřem kategoriím veličiny "uživatel" z odst.2. Ostatní charakteristiky typy úloh spíše jen zvýrazňují. Samozřejmě, že uvedená klasifikace je svázána se subjektem autora a proto ji považujeme za první pracovní a orientační verzi.

5. Trocha optimismu na konec

Letmý pohled na některé problémy analýzy reálných dat mnoho nevyřešil, ale snad aspoň otevřel tuto záležitost k důkladnější diskusi. Možná, že

stav analýzy dat, statistického software a jeho uživatelů je ukazován zbytečně černě. Zkušenost z posledních týdnů však dává naději, že mnohé věci dopadnou lépe, než očekáváme. Proto na závěr jen citát (jméno autora tohoto výroku neznám):

There are NO ROUTINE STATISTICAL QUESTIONS,
there are ONLY QUESTIONABLE STATISTICAL ROUTINES.

Literatura:

1. Osecký P., Principy, přehled a počítačové využití statistických metod, In: Sborník SOFSEM' 88, str.255-274,Beskydy, 1988
2. Tvrdík J., Statistická analýza dat v hygienické službě, Čs. Hyg. 34/4/, str. 145-150, 1989