VÝUKA STATISTIKY NA POČÍTAČOVÉ UČEBNĚ

Martina Litschmannová[*]

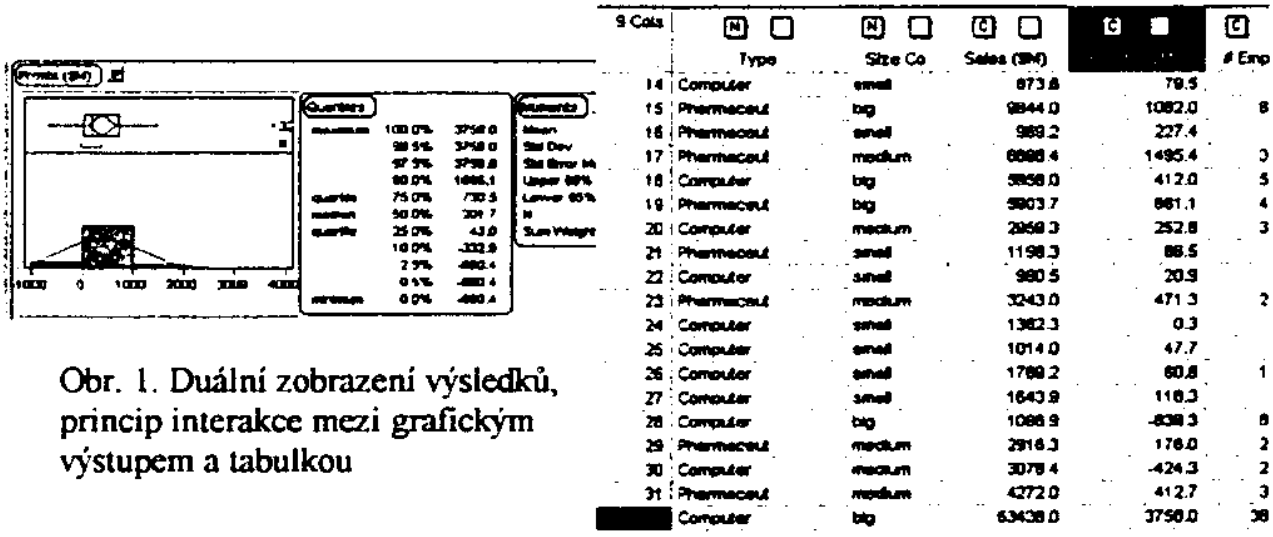
Již nenávratně pryč je doba, kdy těžiště výuky statistiky bylo zavaleno množstvím „ručních" výpočetních operací. Vznik mikropočítače a rozvoj uživatelského software učinily rozsáhlé metody složité statistické analýzy široce dostupnými i pro nepříliš zkušené uživatele. Vlivem nedostatku statistických dovedností se tak často stává, že návrh, analýza a interpretace těchto dat jsou neefektivní, neadekvátní a dokonce zcela vadné.

V současné době probíhá výuka statistiky na FEI TU Ostrava na počítačové učebně, kde máme k dispozici jak specializovaný statistický software (JMP-In, Statgraphics, popř. pro studenty vyšších ročníků jednouživatelské verze S-plus a Fiabex), tak i velmi rozšířený „obyčejný" tabulkový procesor Excel. Výpočty tedy ponecháváme na systému a můžeme se věnovat analýze výsledků.

Vzhledem k tomu, že série přednášek a cvičení ze statistiky (úvodní kurz) je koncipována tak, že klade velký důraz na individuální práci studenta, snažíme se využívat software vyznačující se snadným ovládáním. Pro úvodní kurz statistiky, nezahrnující časové řady, můžeme doporučit programový systém JMP-In, s nímž máme několikaleté zkušenosti. JMP-IN se vyznačuje především velmi snadným ovládáním, minimálními nároky na paměť počítače (7,6 MB) a příznivou cenou (http://www.jmpdiscovery.com/).

JMP-IN

Jedná se o program vyvinutý institutem SAS. JMP-In pracuje pod systémem Windows a je kompatabilní s produkty firmy Microsoft (Word, Excel), což student ocení zejména při tvorbě semestrální práce. Hlavní důraz se v JMP-Inu klade na interaktivní práci s daty. Program je uzpůsoben tak, že čím podrobnější analýzu dat požadujeme, tím více voleb v kontextu s analýzou se nám nabízí. Výsledky každé analýzy jsou prezentovány ve dvou formách – textové a grafické. Výhodou pak je zejména možnost vyhledávání souvislosti mezi grafickým výstupem a vstupními daty (využití např. při identifikaci odlehlých pozorování).



9 Cols	Type	Size Co	Sales ($M)		# Emp
14	Computer	small	873.8	79.5	
15	Pharmaceut	big	9844.0	1082.0	8
16	Pharmaceut	small	989.2	227.4	
17	Pharmaceut	medium	6888.4	1495.4	3
18	Computer	big	5858.0	412.0	5
19	Pharmaceut	big	5803.7	661.1	4
20	Computer	medium	2958.3	252.8	3
21	Pharmaceut	small	1198.3	88.5	
22	Computer	small	980.5	20.9	
23	Pharmaceut	medium	3243.0	471.3	2
24	Computer	small	1382.3	0.3	
25	Computer	small	1014.0	47.7	
26	Computer	small	1788.2	60.8	1
27	Computer	small	1643.9	118.3	
28	Computer	big	1086.9	-838.3	8
29	Pharmaceut	medium	2916.3	176.0	2
30	Computer	medium	3078.4	-424.3	2
31	Pharmaceut	medium	4272.0	412.7	3
	Computer	big	63438.0	3756.0	38

Obr. 1. Duální zobrazení výsledků, princip interakce mezi grafickým výstupem a tabulkou

[*] *Katedra apl. matematiky, VŠB–TU Ostrava, 17. listopadu 15, 708 33 Ostrava-Poruba, Martina.Litschmannova@vsb.cz*

Softwarové požadavky

JMP-IN je možno používat buď na platformě Apple Macintosh nebo na PC. Vzhledem k tomu, že školní učebny jsou v drtivé většině vybaveny osobními počítači, zmíníme se nyní blíže o softwarových požadavcích JMP-INu. JMP-IN je aplikace spustitelná pod operačními systémy Windows firmy Microsoft. V případě Windows '95 či Windows NT není pro činnost JMP-INu nutné instalovat žádný podpůrný software, v případě starších Windows 3.1x je nutné mít nainstalovanou 32-bitovou knihovnu WIN32S.

Výhody systému JMP-IN

Velkou předností programu JMP-IN je to, že plně využívá vlastnosti hostitelského operačního systému. Práce s JMP-INem je tak velmi pohodlná, zvláště proto, že jeho ovládání je v mnoha rysech stejné jako u jiných aplikací Windows. Tuto skutečnost pak oceníme především tehdy, pracujeme-li se studenty, kteří daný operační systém znají a často v něm pracují.

Za druhý největší přínos JMP-INu považujeme možnost interakce při současném zobrazení grafického a textového výstupu, což oceníme zejména při explorační analýze dat. Výsledky každé analýzy se nám zobrazí, pokud je to možné, ve dvou formách – textové a grafické. Student tak může s využitím interakce okamžitě podrobit výsledky kontrole. Aktivujeme-li myší libovolnou část (bod) grafu, dojde k jeho zvýraznění a zároveň se zvýrazní odpovídající data v datovém souboru (využití najdeme např. při identifikaci odlehlých pozorování). Toto pochopitelně lze provést i obráceně: vybereme-li určitá data z datového souboru, v grafu se nám zvýrazní příslušná plocha. Výhody duálního zobrazení dat demonstruje Obr. 1.

Nevýhody systému JMP-IN

JMP-IN skrývá obrovské možnosti především v oblasti explorační analýzy dat. Dokáže zpracovávat milióny dat, přičemž rozsah datového souboru je omezen pouze vlastní pamětí počítače. V čem tedy spočívají zásadní nedostatky?

- ✓ *Rozsáhlé soubory* – JMP-IN vytváří grafický výstup ve všech úlohách, avšak při více než tisíci bodech se grafy stávají nepřehlednými a tím i neužitečnými.
- ✓ *Programování* – JMP-IN je program určený pouze pro analýzu dat prostřednictvím standardních metod; lze jej tedy doporučit jako výukový program, případně jako systém pro základní analýzu dat, nikoliv jako program vhodný pro vědeckou práci.
- ✓ *Speciální oblasti statistiky* – JMP-IN pokrývá rozsáhlou část statistické analýzy, avšak je zde několik oblastí, například časové řady, které v programu nejsou implementovány.

Zadávání dat, JMP-IN jako tabulkový procesor

JMP-IN umožňuje jak přímé zadávání dat, tak i import tabulek v textovém (*.txt) nebo SAS-import (*.xpt) formátu. Při ručním zadávání dat máme dvě možnosti. Buď data přímo zapisujeme, nebo využíváme funkce *Calculator*, která nám umožňuje data zadávat jako funkce dříve definovaných (zadaných) proměnných.

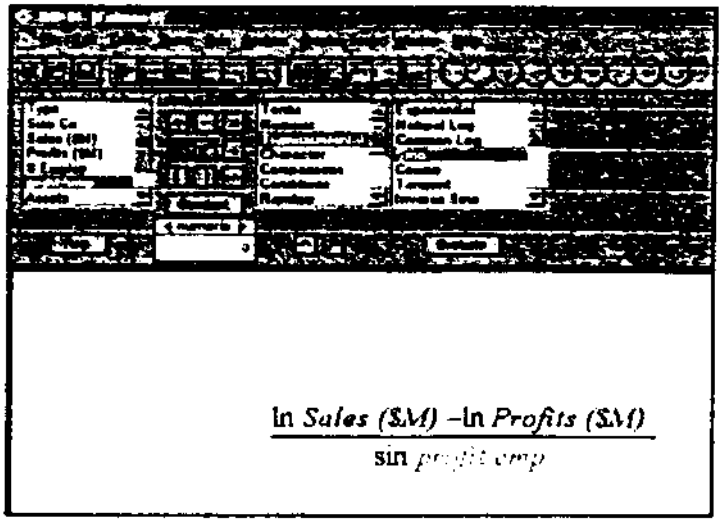
Přímé zadávání – vypisováním

JMP-IN nám nabízí předem možnost vybrat si typ dat zadávaných do příslušného sloupce. Volíme mezi numerickými a kategoriálními datovými typy a u obou pak specifikujeme další vlastnosti. U kategoriálních dat je pak zvláště příjemná možnost vytvoření seznamu povolených hodnot. Toho využíváme v případě proměnných, jejichž hodnoty se opakují. Například: zavedeme-li v souboru osobních dat studentů VŠB - TUO proměnnou Fakulta,

její hodnoty budou tvořit názvy pěti fakult TUO a ty se budou neustále opakovat. Využití seznamu povolených hodnot tak povede k velké úspoře času při zadávání těchto údajů.

Přímé zadávání – prostřednictvím Calculatoru

Dalším způsobem zadávání dat je využití možnosti definovat hodnoty nové proměnné jako funkce dříve definované proměnné. V případě, že zvolíme tuto možnost definování nové veličiny, zobrazí se nám na obrazovce pracovní panel (Obr. 2) obsahující mimo jiné seznam dříve definovaných veličin, panel matematických operátorů, seznam funkcí, pole pro zadávání konstant, a konečně pole, v němž se zobrazuje námi definovaný vztah pro novou proměnnou. Na tomto místě bychom snad měli podotknout, že seznam funkcí je dostatečně rozsáhlý a umožňuje práci jak s numerickou, tak i s kategoriální proměnnou.



Obr. 2 Calculator

JMP-IN jako tabulkový procesor

Nyní se zmíníme o možnostech JMP-INu ve funkci tabulkového procesoru. V nabídce *Tables* se skrývá několik funkcí umožňujících úpravu stávajících tabulek, případně tvorbu tabulek nových.

- *Group /Summary* - vytváří nové tabulky obsahující statistické ukazatele (střední hodnoty, mediány, četnosti, ...) původní (zdrojové) tabulky.
- *Subset* - vytváří nové tabulky obsahující pouze námi vybrané hodnoty z tabulky původní.
- *Sort* - modifikuje původní tabulku na základě vícenásobného třídění.
- *Stack Columns* - převádí kontingenční tabulku do standardního datového formátu.
- *Split Columns* - převádí data ze standardního datového formátu do tvaru kontingenční tabulky.
- *Transpose* – transponuje původní tabulku (převádí řádky na sloupce)
- *Concatenate* – umožňuje horizontální slučování dvou vybraných tabulek, případně horizontální slučování pouze určitých proměnných obsažených v těchto dvou tabulkách
- *Join* – je obdobou funkce *Concatenate* pro vertikální slučování

Tyto funkce nám ulehčují práci s rozsáhlými datovými soubory a bývají často používány, zejména při vytváření výstupu statistické analýzy. Poněkud jiná je situace v případě prvotního zadávání dat. Podle našich zkušenosti volí uživatelé v tomto případě raději, zejména u rozsáhlejších datových souborů, jiný tabulkový procesor, převážně Excel, z něhož pak data importují přes textový formát do JMP-INu.

Analýza dat

Postup při zpracování datového souboru oceníme zvláště z hlediska výuky. Analýzu zahajujeme vždy explorační analýzou dat, přesněji řečeno zobrazením veličiny, případně zobrazením závislosti dvou veličin.

Jednorozměrná veličina

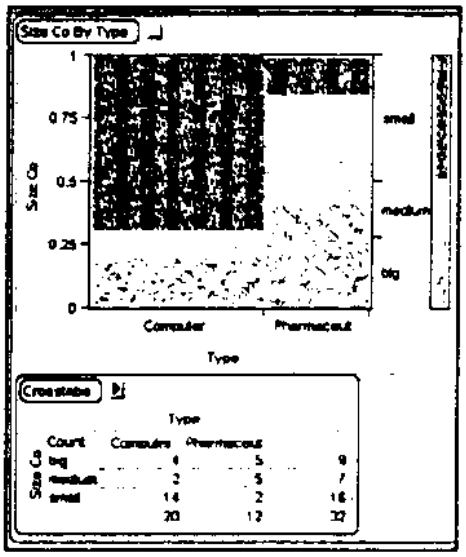
Grafická část výstupu se v případě **kategoriální veličiny** skládá z histogramu a z mozaikového grafu, což je doplněno textovým výstupem ve formě tabulky obsahující absolutní, relativní a kumulativní četnosti. (Pokud kategoriální proměnná není ordinární, lze kumulativní četnosti vypustit.) Jedinou možností další analýzy je test výběrového podílu jednotlivých hodnot této veličiny, který je nám nabízen jako rozšíření explorační analýzy.

Jedná-li se o **numerickou veličinu**, grafický výstup (Obr. 1) je tvořen histogramem a krabicovým grafem, textová část obsahuje hodnoty význačných kvantilů a momentů. V této chvíli je velmi efektní využití interakce mezi grafickým výstupem a datovým souborem. Například chceme-li identifikovat odlehlá pozorování v datovém souboru, stačí je označit v krabicovém grafu a toto označení se zpětně promítne do tabulky (příslušné hodnoty se „vysvítí"). Pak již zbývá pouze rozhodnout o ponechání či vyřazení těchto hodnot z dalšího zpracování. Samozřejmostí je dále možnost změny měřítka na jednotlivých osách, změna velikosti jednotlivých grafických výstupů a odstraňování, popř. přidávání dalších grafických výstupů. V rámci další analýzy se nám nabízí možnost testování střední hodnoty, testování normality výběru a zobrazení grafu „stem & leaf".

Dvourozměrná veličina

Chceme-li testovat závislost veličiny Y na veličině X, liší se výsledky podle typu jednotlivých veličin. Pro tyto testy nám slouží funkce *Fit Y by X*.

Grafickým výstupem **testu závislosti dvou kategoriálních veličin** je mozaikový graf (je-li tento graf tvořen vodorovnými pásy, neexistuje závislost mezi jednotlivými veličinami, čím větší je členitost grafu, tím větší je závislost mezi veličinami), který je doplněn výsledky testu Chi-kvadrát a kontingenční tabulkou (Obr. 3), kterou lze postupně upravovat tak, aby byla zřejmá konstrukce Chi-kvadrát testu (marginální pravděpodobnosti, předpokládané hodnoty, odchylky od skutečných hodnot, čtverce těchto odchylek).



Obr. 3. Mozaikový graf

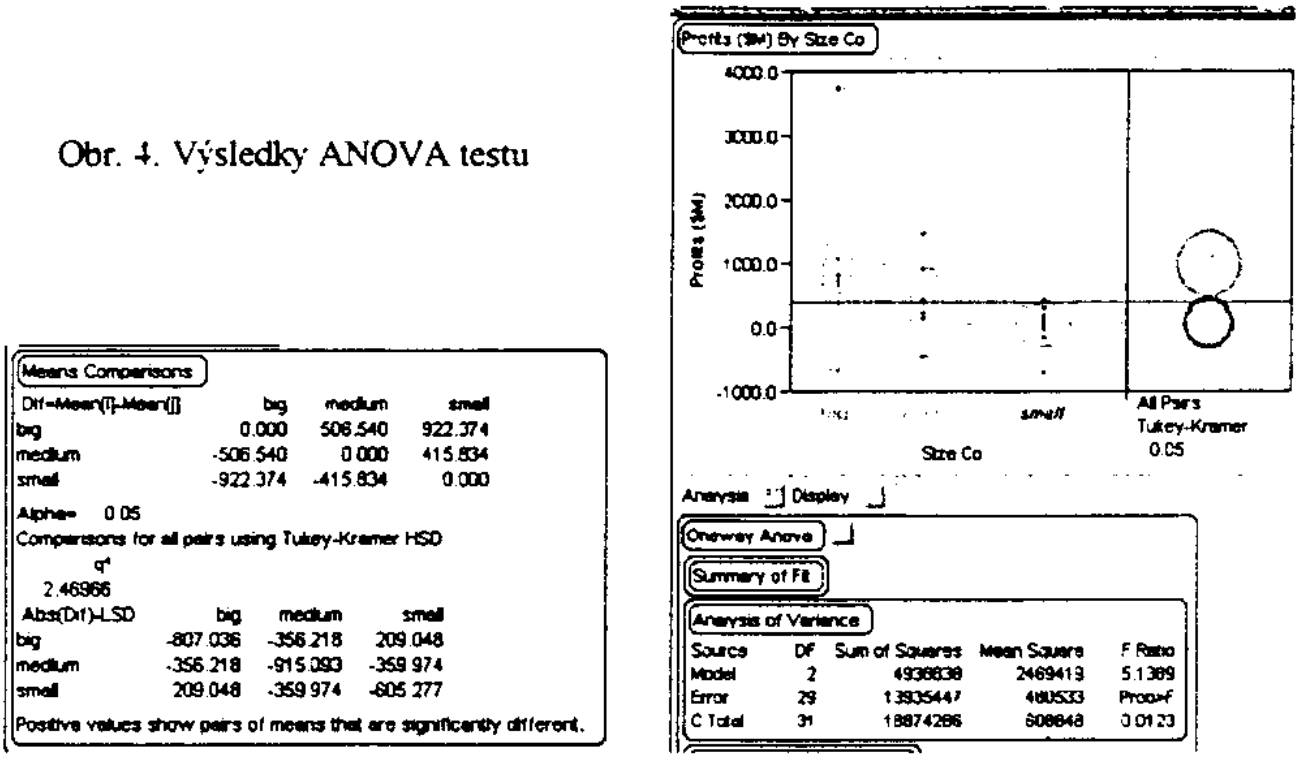
Sledujeme-li **závislost numerické veličiny na veličině kategoriální**, získáme jako grafický výstup vícenásobný krabicový (nebo diamantový) graf, v němž můžeme navíc zobrazit např. střední hodnoty se směrodatnými odchylkami. Jako textový výstup můžeme volit hodnoty středních hodnot a směrodatných odchylek jednotlivých kategorií, hodnoty jednotlivých kvantilů, výsledky Studentova testu, neparametrických testů mediánu, Wilcoxonova testu. V případě více než dvou kategorií to pak jsou výsledky ANOVY a výsledky komparačních testů, a to jak v textové podobě, tak i ve formě grafické (Obr. 4)

Vzhledem k rozsahu referátu se zmíníme pouze o možnostech JMP-INu v oblasti **jednofaktorové ANOVY**. Máme zde k dispozici jak klasický test, tak i neparametrickou Kruskalovou – Wallisovu analýzu, která na rozdíl od klasického testu nepředpokládá normalitu rozdělení pravděpodobnosti v základních souborech. Datovým výstupem testu je tabulka ANOVA v klasické podobě. V případě zamítnutí hypotézy o rovnosti středních hodnot máme možnost vybrat si ze dvou typů komparačních testů založených buď na Studentově nebo na Tukeyově – Kramerově testu. Výsledek těchto testů je dán opět jak v textové, tak i v grafické podobě. Textový výstup sestává z matice (symetrické podle hlavní diagonály), v níž kladné hodnoty na příslušných pozicích signalizují podstatný rozdíl mezi středními hodnotami daných souborů. Grafický výstup je prezentován

skupinou kružnic. V tomto případě (Obr. 4) je podstatný rozdíl mezi středními hodnotami signalizován tím, že se příslušné kružnice vůbec neprotínají.

Nakonec nám zbývá zmínit se o **testování závislosti dvou numerických veličin**. Grafickým výstupem je tentokrát tzv. bodový graf, jenž je doplněn nabídkou pro výpočet a zakreslení různých regresních modelů (lineární, polynomický, ...), nabídkou transformace pro regresi (logaritmická, exponenciální, párového t-testu).

Obr. 4. Výsledky ANOVA testu



Závěr

Rozsah tohoto referátu neumožňuje zmínit se o všech možnostech JMP-INu. Cílem bylo ukázat, že JMP-IN je statistický nástroj, který předčí starší programy především v tom, že je plně grafický, snadno ovladatelný a interaktivní. Přínos všech těchto výhod se projevuje jak na zvýšení zájmu studentů o statistiku (zvýšená aktivita na cvičeních, vyšší kvalita samostatných semestrálních prací), tak i na jejich konečných znalostech.

Literatura:

1. SAS Institute Inc., JMP Statistics and Graphics Guide, SAS Institute, USA, 1995
2. SAS Institute Inc., JMP Introductory Guide, SAS Institute, USA, 1995
3. SAS Institute Inc., JMP User's Guide, SAS Institute, USA, 1995
4. Dummer R. M., Klimková M., Statistika I. (Cvičení), VŠB – TU Ostrava, 1997