Tabulkové procesory a statistika

Hana Řezanková, Jiří Žváček

Důvodem pro sestavení tohoto příspěvku byl dotaz teoretiků: „K čemu jsou statistikovi tabulkové procesory, když máme tak skvělé pakety?“. Pokusíme se na něj odpovědět (a ukázat, že může být položen i opačně).

1. Pomocí čeho se dělá statistika?

Dovolíme si vyslovit provokativní hypotézu, že statistiku většinou dělají nestatistikové pomocí nestatistických systémů.

Běžná statistická práce je práce s daty. Většina analýz vychází z požadavků praxe a má tudíž zcela elementární charakter a standardní formu. Chce se rozbor trhu, vývoj akcií, cenový index, odhad z výběru. Tedy v podstatě přehledná tabulka výsledků, popisná statistika, jednoduchý výpočet a především ohromující graf. Vše krásně vytištěno se zvýrazněnými výsledky a bez gramatických chyb. Tuto práci samozřejmě nemusí dělat statistikové. Často je dokonce lepší, když ji dělají příslušně vzdělaní odborníci, což samozřejmě neznamená, že už se nejedná o statistiku.

Možná, že někdo namítne, že je to pohled ekonoma – ale těch je nejvíc (a začínají mít peníze). A nechť se každý zamyslí nad tím, kolikrát použil pokročilé metody ve srovnání s triviálními a nad čím strávil nejvíce času (navíc znalci tvrdí, že celá matematická statistika funguje tak od 20 do 50 pozorování – pro méně je všechno vidět na první pohled a nedá se to prokázat, pro více je zase významný každý nesmysl).

Odmyslíme-li si na chvíli fakt, že většina práce a času se na počítačích realizuje v textových procesorech (optimisté tvrdí, že pouze 70 %) a že data máme v databankách, tak většina toho, co bychom mohli nazvat vlastní statistikou, se „dělá“ v tabulkových procesorech.

Důvody jsou v podstatě tři:

a) *Většina statistické práce je vykonávána lidmi, kterým je dostupnější tabulkový procesor než statistický paket.*

Příčin je více. Tyto programové systémy jsou řádově levnější (protože se jich prodávají milióny – např. zaváděcí cena produktu Quattro Pro for Windows byla 50 USD), uživatelsky přátelské a obsahují velmi mocné a flexibilní nástroje pro analýzu dat. Pokud tedy lze provést analýzu tím, co mám po ruce a s čím umím zacházet, nestrávím zbytek života tím, že budu splácet a učit se statistický paket. Jak ukážeme, překvapivě mnoho statistiky už v tabulkových procesorech je.

b) *Většinu statistické práce lze pohodlněji a lépe vykonat v tabulkových procesorech než ve statistických paketech.*

Lidstvo touží především po přehledných, řádně popsaných a komentovaných tabulkách a krásných grafech. Tyto dvě základní činnosti jsou ve statistických paketech obvykle na velmi nízké úrovni. Statistické pakety umějí vytvářet pouze velmi jednoduché souhrnné statistické tabulky, neexistuje možnost vhodně je graficky upravit – různě orámovat, zvýraznit významná políčka. Většinou nezahrnují grafy vhodné k publikování (chybí zejména třírozměrné varianty základních grafů) a nedokáží spojit textový a grafický výstup do jednoho celku.

c) *Ještě nevymřela touha si něco sám prohlédnout, vyzkoušet, přepočítat, experimentovat.*

Data jsou ve statistických paketech „odosobněna". Člověka až zamrazí, když si uvědomí, že teprve nejnovější verze slavného SPSS umožňuje pohodlně si prohlédnout data v tabulce a vyhledat konkrétní pozorování. Pokud se však jediný údaj opraví, musí uživatel znovu přepočítat všechny odvozené proměnné.

I použití výsledků jako dat pro další analýzu je ve statistických paketech těžkopádné, zatímco v tabulkových procesorech zůstáváme stále ve stejném prostředí se všemi možnostmi systému.

Statistické pakety obvykle nejsou dostatečně flexibilní, aby bylo možno v jejich rámci na současné úrovni programovat nebo si alespoň něco nového zkusit. Možnosti jsou vymezeny nabídkovým (resp. příkazovým) režimem, vytvoření uživatelských aplikací je prakticky nemožné.

Na druhé straně se současné programovací jazyky natolik vzdálily chápání neprofesionálů, že v nich už skoro žádný neprofesionál neprogramuje. Týká se to i u nás velmi zřídka používaného statistického jazyka S-Plus, takže vývoj nových metod a ilustrace mizí.

Tabulkové procesory naopak něco na způsob programování umožnují i laikům a pro pokročilejší mají mocné programovací prostředky. Nejnovější systémy (Lotus, Excel, Quattro Pro) obsahují i editory dialogů.

Domníváme se, že pro vytváření jednodušších specializovaných statistických paketů je prostředí současných tabulkových procesorů vysloveně vhodné.

2. Programové systémy a (aplikovaná) statistika

Světu programových systémů pro PC dominují tři typy produktů. Jsou to

- *textové procesory* (MS Word, Lotus AmiPro, WordPerfect),
- *databankové systémy* (MS FoxPro, MS Access, Lotus Approach, dBASE, Paradox),
- *tabulkové procesory* (MS Excel, Lotus 1-2-3, Lotus Improv, Quattro Pro).

A tím také tři firmy, které jediné tento sortiment plně pokrývají, a tím prakticky monopolizují – *Microsoft*, *Lotus* a společenství *Borland/WordPerfect* (u nás z povzdálí přihlíží Software602).

Zde se vydělává nejvíce peněz, zde je největší konkurence a nesporně se jedná o vrcholy současného programování na PC. Bez nadsázky je možno říci, že výše uvedené programové systémy určují cestu vývoje dalšího softwaru a každá nová verze přináší mnoho zásadních novinek.

V té či oné podobě je má na svém počítači každý a tráví nad nimi většinu času (i nestatistického). Je tedy přirozené, že je přece jenom zná trochu lépe než sporadicky používané specializované systémy.

I statistická práce má v podstatě 3 fáze

- pořízení správných dat,
- vlastní výpočet,
- předání komentovaných výsledků.

Je vcelku jasné, že na získávání a údržbu rozsáhlejších souborů dat bude nejlepší použít rozšířený databankový systém. Většinou už v něm data jsou (nestatistik určitě nepoužije pro vstup dat statistický paket), nebo je možné pomocí něho data pohodlně připravit (kdo trochu zná dBASE nebo jakýkoliv dnešní databankový systém, tak mu dá jistě přednost, především z důvodu možnosti kontroly dat).

Stejně tak, jako dáme přednost textovému procesoru při úpravě výstupů. (Zkuste dnes publikovat neupravený výstup z SPSS s nalepovanými obrázky!) V nejnovějších statistických paketech pro Windows sice můžeme připravovat primitivní výstupní ASCII soubory (bez zarovnání, dělení slov, gramatické kontroly - o vzorcích ani nemluvě) s nabubřelými dělenými tabulkami bez pořádných rámečků (a když tam jsou, tak zas nejdou do textového procesoru přenést), ale možnost integrovat text a grafy do výstupu je vzdálený sen. Navíc výsledky je stále častěji nutno předávat k publikování jako data (na disketách nebo přímo sítí). Bez textového procesoru se dnes prostě nikdo neobejde.

Ještě, že v prostředí Windows existují určité prostředky (Clipboard, OLE), které znalcům umožňují problémy s výstupy obejít (ale to není zásluha paketů).

Zdánlivě tedy ještě zbývá pro statistické pakety trochu prostoru v oblasti statistických výpočtů. Rádi bychom upozornili na fakt, že už to dávno není tak jisté. Spoustu výpočtů lze provést již v databankových systémech, na závěrečné dopočty a vymýšlení výsledků stačí i textové procesory a v záloze číhají tabulkové procesory, které jsou dokonce na výpočty specializovány.

3. Charakteristika tabulkových procesorů

Základem tabulkových procesorů (dále TP) jsou veškeré operace s tabulkami, jako jsou zejména

- *výpočty v tabulkách* (včetně spousty funkcí a maticových operací),
- *grafická úprava tabulek*,
- a nejnověji i *možnost* nejrůznějším způsobem *přestavovat tabulky*.

Kromě toho obsahují

- základní operace s databázemi, včetně možnosti vytvářet z nich analytické tabulky (např. kontingenční),
- obchodní grafiku, včetně úprav grafu,
- základní ekonomicko-matematické (tedy i statistické) metody,
- psaní textů a jejich gramatickou kontrolu,
- integraci textu, tabulek a grafů do jednoho výstupu.

Ovládání TP je vysoce propracované a intuitivní, založené zejména na nabídkách a ikonách. Můžeme v nich však i programovat, vytvářet posloupnosti příkazů využívajících všech možností systému. Nejjednodušší postup jsou tzv. makra, která můžeme vytvořit zaznamenáním vykonávaných činností.

Činnost TP budeme ilustrovat na čtyřech základních úlohách, pro které jsou podle našeho názoru tyto programové systémy vhodnější než současné statistické pakety:

I. Kontrola a oprava dat
II. Vytváření a úprava souhrnných tabulek
III. Publikování grafů
IV. Programování jednoduchých ilustračních úloh

3.1 Klasické tabulkové procesory. Patří sem zejména Excel 4.0, Quattro Pro 5.0 a Lotus 1-2-3 verze 4 (nemá smysl hovořit o jiných systémech než pro Windows).

Základním pojmem tabulkových procesorů je pracovní plocha (sheet, zřejmě list), kterých může být v moderních TP více. Pracovní plocha má strukturu mřížky s políčky, které mají pevnou adresu složenou z kombinace písmene (písmen), které označuje sloupec, a čísla (řádku) – první políčko má tedy adresu A1. Toto označování se nazývá *poziční*.

Do políček lze zapisovat

- *číslo*,
- *text*,
- *formuli*, která se může skládat z adres políček, operátorů, čísel a funkcí,
- *prvky makra*, určitého druhu programu, pomocí kterého můžeme vytvářet složitější aplikace s využitím všech možností systému.

Popularita TP je dána zejména tím, že v nich můžeme vytvářet složité tabulky, které obsahují *popisy, data* a zejména *výpočty*.

Tabulky jsou přitom velmi flexibilní, protože

- při změně dat se automaticky přepočtou všechny formule (jen tam, kde je to nutné),
- lze měnit způsob zobrazení tabulky (šířku sloupců, formáty zobrazení obsahu políček atd.),
- tabulky lze různě orámovat, zvýrazňovat políčka.

Práce s tabulkovými procesory se provádí ve dvou režimech:

a) *vstupu údajů*, kdy se pohybujeme v tabulce,
b) *příkazů*, v jehož rámci zadáváme složitější operace.

Základem je práce s oblastmi, což jsou čtyřúhelníkové skupiny políček, vymezené dvěma adresami v protilehlých rozích.

Nejdůležitějšími příkazy jsou *přesunování a kopírování oblastí*. Zvláště užitečné je kopírování formulí s využitím automatických změn adres.

Dalšími důležitými operacemi s oblastmi jsou zejména

- *změna formátu oblasti* (způsobu zobrazení obsahu políček),
- *výpočet nad oblastmi*, např. regresní analýza, maticové operace, optimalizace atd.
- *grafické znázornění oblasti*,
- *zobrazení grafu do oblasti*.

Výsledkem práce s TP je tedy formulář, skládající se z textů, tabulek a grafů, který se automaticky přizpůsobuje změnám dat – tedy program.

3.2 Dynamické tabulkové procesory

Názvy těchto systémů ještě nejsou ustálené, záleží na tom, na který aspekt těchto nových systémů se zaměříme. V podstatě se jedná o tři nové myšlenky (viz dále), z nichž nejvíce bije do očí možnost rozšiřování tabulek. Alternativním názvem je např. *systémy pro vícerozměrnou analýzu dat*, protože podstatnou se zdá myšlenka přestavby tabulek pomocí kategorií.

Revoluci začal a nejznámější je Improv. Je lehčí jej používat, než popsat. Na rozdíl od klasických tabulkových procesorů zobecňuje typickou strukturu ¡písmeno,číslo¿, takže sloupce i řádky mohou mít zcela libovolné názvy, tzv. *items* (položky sloupců a řádků). Položkám lze přiřadit *categories*, typy informace, kterou obsahují.

Změny jsou poučné přinejmenším ve třech směrech:

- *políčka se generují* pouze při zavedení nové položky (klávesa Enter) – příklad vytvoření tabulky:

pohlaví

			muži	ženy	celkem
fakulta 1	bakalářské	semestr 1			
		semestr 2			
		...			
	inženýrské	semestr 1			
		semestr 2			
		...			
fakulta 2	dtto				
...					

fakulta | studium | semestr

- *formule* se u některých systémů (Improv) zadávají *mimo tabulku* ve formě: *výsledná oblast = formule*, kde ve formuli mohou uvedeny celé skupiny položek – zadává se tedy zároveň, kam se má výsledek uložit, a velmi se redukuje počet zadávaných funkcí; například pomocí formule *celkem=muži+ženy* se definují součty pro všechny řádky najednou, používá se tedy *nepoziční označování*,

- tabulku lze velmi jednoduše přebudovat pomocí *kategorií* (ve výše uvedeném příkladu jsou kategoriemi *pohlavi, fakulta, studium a semestr*), čímž se dají generovat různé varianty tabulek (jednodušeji než u specializovaných generátorů tabulek známých ve statistickém softwaru – např. Tables v SPSS).

fakulta | pohlaví

		fakulta 1		fakulta 2		...
		muži	ženy	muži	ženy	
bakalářské	semestr 1					
	semestr 2					
	...					
inženýrské	semestr 1					
	semestr 2					
	...					

studium | semestr

Kategorie jsou zaváděny i do nejnovějších verzí „starých“ tabulkových procesorů, práci s nimi obsahuje např. Quattro Pro for Windows.

3.3 Statistické funkce

Tabulkové procesory obsahují stovky matematických, časových, textových, finančních a přirozeně i statistických funkcí, takže základní výpočty můžeme provádět velmi jednoduše (viz též [5]).

Statistické funkce lze rozdělit např. do následujících skupin (konkrétní možnosti jsou uvedeny pro Quattro Pro for W.):

a) *jednorozměrné míry*, které zahrnují všechny popisné charakteristiky (včetně kvantilů a dokonce useknutého průměru),
b) *charakteristiky lineární regrese* (včetně extrapolací a směrodatných odchylek),
c) *transformace* (pořadí, normování, Fisherova),
d) *výsledky statistických testů* (chí-kvadrát test dobré shody, t-test shody průměrů, F-test shody rozptylů, normální test o střední hodnotě – zde z-test),
e) *užitečné funkce* (neúplná gama a beta funkce, počty permutací a kombinací),
f) *pravděpodobnostní funkce* (hustoty či pravděpodobnosti, distribuční funkce a kvantily) pro 15 nejznámějších rozdělení,
g) *nejrůznější druhy součtů*.

Pro konstrukci souhrnných tabulek jsou podstatné statistické *databázové funkce*.

3.4 Statistické metody. I když většinu metod lze realizovat pomocí funkcí, jsou základní metody „předprogramovány" do standardní podoby. Metody zavedl Excel 4, ale prakticky totožné má nové Quattro Pro for W. (QPW).

Metoda	Název metody Excel 4.0	QPW 5.0
Jednorozměrné statistické charakteristiky	Descriptive Statistics	
Jednorozměrná tabulka četností	Histogram	
Jednofaktorová analýza rozptylu	Anova: Single-Factor	One-Way
Dvoufaktorová analýza rozptylu s opakováním	Anova: Two-Factor with Replication	Two-Way
Dvoufaktorová analýza rozptylu bez opakování	Anova: Two-Factor without Replication	Two-Way
Regresní analýza pro lineární model	–	Advanced Regression
	Regression	
Matice korelačních koeficientů	Correlation	
Matice kovariancí a rozptylů	Covariance	
Generování náhodných čísel (7 typů rozdělení)	Random Number Generation	
Uspořádání hodnot (od největší k nejmenší), jejich pořadí a údaj, kolik % hodnot je menších nebo rovno příslušné hodnotě.	Rank and Percentile	
Výběr hodnot (buď každá k-tá hodnota nebo náhodný výběr k-hodnot)	Sampling	
Exponenciální vyrovnání	Exponential smoothing	
Klouzavé průměry	Moving Average	
Fourierova analýza	Fourier Analysis	Fourier

Párový t-test o shodě průměrů	t-Test: Paired Two-Sample for Means
t-test o shodě průměrů pro dva výběry se stejnými rozptyly	t-Test: Two-Sample Assuming with Equal Variances
t-test o shodě průměrů pro dva výběry s nestejnými rozptyly	t-Test: Two-Sample Assuming with Unequal Variances
z-test o shodě průměrů pro dva výběry	z-Test: Two-Sample z-Test for Means
F-test o shodě rozptylů pro dva výběry	F-Test: Two-Sample F-Test for Variances

4. Příklad kontroly a opravy dat

Zdrojová statistická data mají zpravidla formu datové matice, která odpovídá tabulce databáze. Jedná se tedy o strukturu

pozorování	pozorované proměnné	odvozené proměnné
1		
2		
...		

Je třeba si uvědomit, že správná data v reálném životě neexistují a že první činností každé analýzy musí být odstranění či oprava chyb. Ty zpravidla nezjistíme v pozorovaných proměnných, ale až analýzou odvozených proměnných. Typickým příkladem mohou být data z oblasti medicíny, kde na řádově 100 pozorovaných znaků připadají stovky znaků odvozených, které mají určitý věcný význam.

Zpravidla tedy

- nalezneme chybu v odvozené proměnné,
- opravíme pozorovanou proměnnou,
- znovu prozkoumáme odvozené proměnné.

Tento postup je velmi zdlouhavý ve statistických paketech, kde je nutno znovu vypočítat odvozené proměnné, ale velmi přirozený v tabulkových procesorech, kde se při změně údaje provedou všechny nutné přepočty automaticky.

Zcela elementární kontrola je tedy např.

pozorování	pozorované proměnné	odvozené proměnné
1		
2	oprava ⟵	vyhledání
...		
minimum		
maximum	⟶ změna ...	nesmysl

Extrémní pozorování pak můžeme snadno vyhledat a případně opravit či vypustit.

Závažným problémem ovšem je, že tabulkové procesory neumějí pracovat s chybějícími pozorováními (přesněji řečeno nezahrnují algebru chybějících pozorování, některé speciální funkce s nimi pracovat umějí). Výsledkem operace mezi textovou a číselnou hodnotou je nula.

5. Vytváření souhrnných tabulek

Jak je dnes již všeobecně známo, individuální data prakticky nelze publikovat (podle zákona o ochraně individuálních údajů). Navíc pro vytvoření základní představy o datech jsou souhrnné tabulky nezbytné.

Další oblastí využití je publikování výsledků testů. Je jistě mnohem přehlednější publikovat jednu tabulku výsledků týkající se několika proměnných, než samostatné výstupy pro jednotlivé proměnné:

	tlak	tep	atd.
nový lék	@AVG(X1)	@AVG(Y1)	
placebo	@AVG(X2)	@AVG(Y2)	
t-test	@TTEST(X1,X2)	@TTEST(Y1,Y2)	

Základním typem dvourozměrných tabulek jsou kontingenční tabulky, které obsahují četnosti výskytů jednotlivých kombinací hodnot dvou kategoriálních proměnných. Tyto četnosti mohou být buď absolutní nebo relativní, a to v rámci sloupců, řádků, či celé tabulky. Tvorba kontingenčních tabulek bývá jednou ze základních procedur statistických paketů.

Tabulkové procesory umožňují zobrazovat v políčkách tabulky kromě četností též statistické charakteristiky třetí (obvykle spojité) proměnné – aritmetický průměr, minimum, maximum, součet, rozptyl, směrodatnou odchylku. Nabízená činnost se obvykle jmenuje Crosstabs (Excel, Quattro Pro) a vyžaduje data ve tvaru databáze (sloupce jsou označeny názvy). U statistických paketů jsou tyto možnosti obsaženy ve speciálních modulech (Tables v SPSS).

6. Grafika

Ze širšího hlediska můžeme pod grafikou chápat veškeré grafické ztvárnění výstupů. V současné době se jedná zejména o grafickou úpravu

- *textů*, která se již díky Windows ve všech systémech koncepčně neliší (lze si přinejmenším zvolit fonty – formu a velikost zobrazení znaků),
- *tabulek*, které lze různě rámovat a zvýrazňovat,
- *grafů*.

Pro TP platí jednoduché pravidlo – nejlepší je novější z produktů Quattro Pro a Excel. Až do Excelu 5 to tedy bude QPW.

6.1 Tabulky. Grafická úprava tabulek je ve statistických paketech naprosto nesrovnatelná s tabulkovými procesory, zejména díky dvěma prvkům TP, kterými jsou

- rámování a
- zvýrazňování políček.

Nejnovější textové procesory (zatím WordPerfect 6.0, ale další budou jistě brzy následovat) obsahují kromě klasických objektů text, graf a vzorec i objekt tabulka. Dokáží importovat tabulky z lepších tabulkových procesorů, včetně jejich vlastností. Zachovávají tedy formáty, vzorce, rámování, zvýrazňování atd.

Zásadním nedostatkem statistických paketů je to, že zatímco vstupní data jsou tabulkou, s níž lze provádět podobné operace jako v tabulkových procesorech, výstupy jsou textové soubory. Dochází tak k situaci, že nemůžeme přímo přenést vstupní data do výstupu a výstupy do dat. Pokud tedy chceme získat pěknou výstupní tabulku, musíme to provést v jiném systému.

Exportovat textovou tabulku výsledků ze statistického paketu do tabulkového procesoru je poměrně pracné, protože je třeba

- upravit ji v textovém procesoru (zejména odstranit nadbytečné řádky a sloupce a spojit rozdělené tabulky),

- vložit ji jako text do tabulkového procesoru a pomocí formátovacího řádku rozdělit souvislý text do sloupců tabulky (operace „parse“).

6.2 Grafy. Situace již došla tak daleko, že některé původně statistické grafy jsou známější pod názvem „obchodní“ grafika a jejich publikování v původní dvourozměrné podobě není známkou kvality. Na druhé straně je možno konstatovat, že nové statistické pakety v prostředí Windows již mají integrovány základní operace s grafy (které měly tabulkové procesory vždy).

Srovnání tabulkových procesorů (TP) a statistických paketů (SP) provedeme na následujících systémech:

SG – Statgraphics, v. 6.1 Plus,
SPSS – for Windows, v. 5.0.2,
SYSTAT – for Windows, v. 5.0 (demoverze),
STATIS – Statistica for Windows, v. 4.0,
QPW – QuattroPro for Windows, v. 5.00,
Excel – Microsoft Excel, v. 4.0,
Improv – Lotus Improv for Windows, v. 2.0,
1234 – Lotus 1-2-3 for Windows, v. 4.

V níže uvedených tabulkách je použita následující symbolika:

*	možnost je součástí programového systému,
–	možnost není součástí programového systému,
číslo	počet různých způsobů, max. stupeň polynomu apod.,
.	daný graf lze vytvořit zprostředkovaně (úpravou dat),
nic	nemá smysl uvažovat.

Pokud není určitý produkt zařazen do některého srovnání, nepodařilo se zjistit, zda jsou sledované možnosti přítomny.

Základní operace s grafy

	SG	SPSS	SYSTAT	STATIS	QPW	Excel	Improv	1234
šrafování	*	*	*	*	*	*	*	*
změna barev	*	*	*	*	*	*	*	*
změna typů čar	*	*	*[1]	*	*[2]	*	*	*
titulek	*	*	*	*	*	*	*	*
podtitulek	*	*	–	*	*	–	*	*
poznámka	*	*	–	*	–	–	–	*
legenda	*	*	*	*	*	*	*	*
změna měřítka	*	*	*	*	*	*	*	*
rotace 3D grafů	*	*	*	*	*	*	*	–

Důležité jsou též možnosti kreslení a provádění dodatečných úprav do vytvořených grafů. Přidávanými prvky, používanými pro zvýrazňování objektů grafu, jsou zejména přímka, obdélník, elipsa a text umisťovaný na libovolné místo (ekvivalentní způsob pro zadávání titulků a poznámek).

Dodatečné úpravy grafů

	SG	SPSS	SYSTAT	STATIS	QPW	Excel	Improv	1234
přímka	–	–	*	*	*	–	–	*
obdélník	–	–	*	*	*	–	–	*
elipsa	–	–	*	*	*	–	–	*
text	–	–	*	*	*	–	–	–

Graf lze obvykle buď uchovat ve vlastním formátu programového systému nebo v některém rozšířeném grafickém formátu – pro převod do textových procesorů, příp. grafických systémů (s přenosem ve Windows prostřednictvím programu Clipboard mohou být spojeny určité problémy). Možnosti exportu jsou však (kromě QPW) velmi omezené.

Nejdůležitější grafické formáty jsou zejména

.BMP (Bitmap), bitmapový formát používaný systémy MS-Windows (3.x) a OS/2 Presentation Manager,

.CGM (Computer Graphics Metafile), nejrozšířenější vektorový formát,

.WMF (Windows Metafile), vektorový formát,

.EPS (Encapsulated PostScript), formát využívaný dražšími laserovými tiskárnami,

.GIF (Graphics Interchange Format), bitmapový formát, který používá CompuServe Information Service (CIS) a další BBS,

.PCX standard pro přenos obrázků v bitmapovém formátu, podporovaný např. systémy Paintbrush a Quattro Pro for DOS,

.TIF (Tag Image File Format – TIFF) je bitmapový formát, který má dvě varianty – s kompresí a bez komprese.

Export grafů

formát	SG	SPSS	SYSTAT	STATIS	QPW	Excel
.BMP	–	–	*	*	*	*
.CGM	*	–	–	–	*	*
.WMF	–	–	*	*	–	–
.EPS	–	–	–	–	*	–
.GIF	–	–	–	–	*	–
.PCX	–	–	–	–	*	–
.TIF	–	–	–	–	*	–

Základní tzv. obchodní statistické grafy jsou: *výsečový, sloupkový* a *spojnicový*. Vyskytují se v mnoha variantách a kombinacích, podle vžitých publikačních zvyklostí. Ze statistického hlediska se vesměs jedná o grafy dvourozměrné pro jednu proměnnou. Vzhledem k tomu, že jde o zobrazování posloupnosti hodnot, používá se též termín řada.

K zobrazení více řad se v případě sloupkového grafu používají následující prostředky:

- *shlukování* (odpovídající hodnoty jsou znázorněny vedle sebe),
- *kumulace* (odpovídající hodnoty jsou znázorněny nad sebou),
- *segmentace* (odpovídající hodnoty jsou znázorněny za sebou – tím se zavede třetí rozměr).

Základní rozdíl mezi TP a SP spočívá v tom, že v TP se zobrazují hodnoty z tabulky, zatímco v SP se zobrazují spíše četnosti (resp. průměry či jiné charakteristiky), které systém automaticky počítá na základě původních hodnot. Pokud chceme zobrazit četnosti z původních dat v TP, musíme nejprve zadat jejich výpočet.

V zásadě je možno říci, že dnes nelze v obchodní praxi publikovat klasický dvourozměrný graf, ale pouze jeho třírozměrnou (3-D) variantu. Třírozměrného efektu se dosahuje kromě segmentace zavedením hloubky (výšky výseče, tloušťky sloupce, šířky čáry). Kromě výše uvedených tří základních typů se používají další specializované grafy.

Obchodní grafy se používají v těchto variantách (anglická terminologie se v různých systémech někdy liší, v dalším textu jsou většinou použity názvy z QPW):

a) Výsečový (Pie) se používá pro znázornění podílu na celku.
Esteticky má velký význam zejména

- *hloubka* (pro jednotnost, protože zde by bylo lepší použít termín výška), která způsobuje třírozměrný efekt (proto se používá označení 3-D Pie),
- vysouvání výsečí (*exploze*), aby se zvýraznily některé aspekty.

Statisticky důležité je

- *kolapsování* (sdružení malých podílů do položky „ostatní"), které lze ovšem v tabulkových procesorech řešit úpravou výchozí tabulky, a
- více grafů v jednom obrázku pro porovnání.

Variantou výsečového grafu je *prstencový graf (Doughnut)*. Jediný rozdíl od výsečového grafu je „otvor" uprostřed, do kterého lze zakreslit či napsat potřebné texty.

Terčový graf (Ring), který je obsažen v SYSTATu jako varianta výsečového grafu, je koncepčně chybný (a proto se nevyskytuje v nestatistických systémech). Zobrazuje totiž data ve formě kumulativních soustředných kruhů, jejichž poloměr je úměrný velikosti hodnoty (např. tedy četnosti výskytu varianty). Například dvakrát větší hodnota tedy bude mít dvakrát větší poloměr, zdánlivě tedy dvakrát větší kruh. To je ovšem optický klam, protože velikost je zobrazena dvourozměrným obrazcem (prstencem), jehož plocha roste se čtvercem poloměru a je ve skutečnosti 3 krát větší než plocha vnitřního kruhu. Podobné triky jsou popsány ve většině prací o statistickém podvodnictví (např. i v knize Swoboda: Moderní statistika, která je u nás velmi populární).

Varianty výsečového grafu

	SG	SPSS	SYSTAT	STATIS	QPW	Excel	Improv	1234
s hloubkou	–	–	–	*	*	*	*	*
s vysouváním	*	*	–	*	*	*	*	*
s kolapsováním	–	*	–	–	.	.	.	.
prstencový	–	–	–	–	*	–	–	–
terčový	–	–	*	–	–	–	–	–

b) Sloupkový graf (Bar) se používá pro znázornění posloupnosti hodnot (resp. jejich četností atd.) jedné či více proměnných.
Základní varianty jsou tři:

- jednoduchý graf pro jednu proměnnou,
- jednoduchý graf pro více proměnných a
- sekvenční graf pro více proměnných (tento graf je již trojrozměrný).

ba) Speciálními typy jednoduchého grafu *pro jednu proměnnou* jsou

- **odchylkový graf (Variance)**, který zobrazuje o dchylky od zvolené úrovně (obsažen v SG, SPSS, STATIS, QPW a Excelu) a
- **histogram**, specializovaný graf pro znázornění četností v intervalech (vzdálenost mezi sloupky je nulová).

U histogramu je užitečná možnost zakreslit do grafu hustotu, resp. pravděpodobnostní funkci statistických rozdělení (zejména normálního), případně jej vyhladit spojitou hladkou křivkou (např. splinem). Histogramy jsou již jednoznačně doménou SP (normální křivku lze zakreslit ve všech čtyřech sledovaných paketech), v některých TP je lze vytvořit, ale pouze úpravou šířky sloupků nebo nastavením nulové vzdálenosti sloupků. Hustotu

nebo pravděpodobnostní funkci jiného než normálního rozdělení umožňuje kreslit SG (18 typů rozdělení) a STATIS (10 typů přímo v histogramu a další v jiných modulech).

bb) *Pro více proměnných* existují dvě základní varianty jednoduchého grafu:

- **shlukový (Clustered)**, ve kterém se kreslí těsně vedle sebe sloupky pro různé řady nebo různé hodnoty klasifikační proměnné (třeba pohlaví),
- **kumulativní (Stacked)**, ve kterém se hodnoty (které má smysl sčítat) kreslí nad sebe.

Pro kumulativní graf existují varianty

- **porovnávací (Comparison)**, se spojením jednotlivých řad mezi sloupci čarou (velmi názorné pro srovnávání)
- **100 % kumulativní (100 % Stacked)**, přepočet na 100 % (pro srovnávání struktur).

Esteticky významné jsou varianty

- *s hloubkou*, ve které těmto grafům přidáváme z estetických důvodů třetí rozměr (2.5 Bar, 3–D Stacked a 100 % 3–D Stacked),
- změna šířky a vzdálenosti sloupců,
- *záměna os* (rotace dvourozměrného grafu).

Většinu těchto grafů si lze představit jak v klasické vertikální podobě, tak v podobě horizontální. Často se příslušná operace nazývá záměna os (X–Y). Opravdovou rotaci má Statistica, která má všechny čtyři možnosti (včetně „na stropě" a doprava).

Varianty jednoduchého sloupkového grafu pro více proměnných

	SG	SPSS	SYSTAT	STATIS	QPW	Excel	Improv	1234
Typy								
shlukový	*	*	–	–	*	*	*	*
kumulativní	*	*	*	*	*	*	*	*
porovnávací	–	–	–	–	*	*	–	*
100 %	*	*	*	–	*	.	.	.
Vlastnosti								
hloubka	–	*	–	*	*	*	*	*
šířka sloupce	–	–	*	*	*	*	*	*
vzdál.sloupců	–	*	–	–	*	–	*	–
záměna os	XY	XY	XY	4	XY	XY	XY	XY

bc) Pravé *třírozměrné grafy* jsou sekvenční:

- **sloupkový (3–D Bar)**, ve kterém se kreslí řady sloupků (s hloubkou) za sebou ve třírozměrném prostoru,
- **schodovitý (3–D Step)**, ve kterých je vzdálenost mezi sloupky nulová (jedná se vlastně o sekvenci histogramů).

Z vlastností mají největší význam

- *záměna os*,
- *rotace*, ve které se „díváme" na třírozměrný graf z různých míst.

Varianty sekvenčního sloupkového grafu

	SG	SPSS	SYSTAT	STATIS	QPW	Excel	Improv	1234
Typy								
sloupkový	–	–	–	*	*	*	*	*
schodovitý	–	–	–	–	*	–	–	–
Vlastnosti								
záměna os				–	*	–	–	*
rotace				*	*	*	*	–

c) **Spojnicový graf (Line)** je primárně určen pro časové řady a spojuje jednotlivé údaje (znázorněné jako body) čarou (spojení je obvykle možné též potlačit, takže mohou zůstat zobrazené pouze body).

ca) *Variantou pro jednu řadu* je např. graf **plošný (Area)**. Rozdíl spočívá v tom, že plocha pod čarou je vyplněna barvou (příp. je vyšrafována).

Jednoduchým způsobem pro vyjádření třírozměrného efektu je u obou typů grafů *hloubka*. Speciálně spojnicový graf s hloubkou je **stužkový (Line with Depth** nebo **Ribbon)**, který odpovídá zavedení hloubky u předchozích grafů a vytváří třírozměrný efekt. Pro plošný se používá obdobného názvu (Area with Depth).

Pro více proměnných jsou určeny kumulativní a sekvenční varianty.

cb) *Kumulativní variantou pro více řad* (jejichž hodnoty má smysl sčítat, např. podíly na celku v čase) je graf **plošný kumulativní** (chybí pouze v systému STATIS), který může být zobrazen buď bez hloubky **(Stacked Area)** nebo s hloubkou **(3–D Area)**.

cc) *Sekvenční varianty pro více řad* existují pro grafy

- **spojnicový (3–D Line)**, čáry pro více proměnných,
- **stužkový (3–D Ribbon)**,
- **plošný (3–D Unstacked Area)**.

Speciální modifikací spojnicového grafu je graf **povrchový** (úseky mezi spojnicemi odpovídajících bodů u dvou čar jsou vyplněny plochou) s několika variantami

- prostý **(3–D Surface)**, kdy jsou plochy mezi řadami vybarveny stejnou barvou (příp. stejně vyšrafovány),
- kontury **(3–D Contour)**, kde jsou plochy mezi řadami barevně rozlišeny,
- stínovaný **(3–D Shaded Surface)**, kde je vytvořen efekt stínů (představa je, že „slunce" svítí shora a nejjasnější jsou rovné plochy).

Varianty spojnicového grafu

	SG	SPSS	SYSTAT	STATIS	QPW	Excel	Improv	1234
Typy								
spojnicový 3–D	–	–	–	*	–	–	–	–
stužkový 3–D	–	–	–	*	*	*	*	*
plošný kum. 3–D	–	–	–	–	*	*	*	*
plošný sekv.3–D	–	–	–	*	*	*	*	*
povrchový								
prostý	–	–	–	–	*	–	–	–
kontury	–	–	–	–	*	*	–	–
stínovaný	–	–	–	*	*	–	–	–

U jednoduchých spojnicových grafů bývá v případě časových řad užitečné znázornit zároveň vyrovnávací křivku (spojení bodů pomocí čáry může být v tomto případě potlačeno). SP používají zpravidla širokou škálu křivek, k nimž patří zejména přímka, parabola, exponenciála, vyhlazení (nejčastěji splinovými funkcemi) a čára klouzavých průměrů.

Varianty vyrovnávací křivky

	SG	SPSS	SYSTAT	STATIS	QPW	Excel	Improv	1234
přímka	*	*	–	*	*	.	.	.
parabola	*	3	–	5	–	–	–	–
exponenciála	*	–	–	*	*	–	–	–
vyhlazení	*	*	–	*	–	–	–	–
klouz. průměry	*	.	–	–	*	.	.	.

d) Další grafy již jsou různým způsobem specializovány, zmiňujeme se o nich jenom pro úplnost:

da) Minimaxový graf (High-Low) Jedná se o specializovaný finanční graf, používaný k popisu denního vývoje cen a objemu prodeje akcií.

V grafu lze zobrazit

(1) nejvyšší dosaženou cenu,
(2) nejnižší dosaženou cenu,
(3) cenu při uzavření burzy (znak),
(4) cenu při otevření burzy (znak),

rozdíl (1) a (2) je spojen úsečkou, další řady jsou kresleny jako spojnicové grafy.

Zpravidla se jako pátá řada zobrazuje čára objemu prodeje. Zde bývá často názornější použít pro tuto řadu sloupkový graf (viz kombinované grafy).

db) Radarový graf (Radar) Tento graf je navrhován pro zkoumání přítomnosti trendu a zakresluje hodnoty řady postupně do kruhu. Hodnoty jsou interpolovány úsečkou (Excel) nebo spojitou křivkou (QPW) a poslední hodnota je spojena s první. Vznikají tím pěkné obrázky (v QPW dostaneme pro hodnoty 1,3, 4,3 srdce) s pramalou vypovídací schopností. Radarový graf by měl určitý smysl spíše pro analýzu periodických časových řad, pokud by kruh odpovídal délce periody.

Specializované grafy

	SG	SPSS	SYSTAT	STATIS	QPW	Excel	Improv	1234
Hi–Lo	–	–	*	*	*	*	*	*
radar	–	–	*	–	*	*	–	*

dc) Korelační graf (Scatterplot) je dvourozměrným (resp. třírozměrným grafem) a znázorňuje body odpovídající dvojici (resp. trojici) hodnot. Podobně jako histogram je již spíše statistickým grafem a možnosti TP nejsou na úrovni SP (u TP obvykle existuje pouze základní dvourozměrná varianta). Stejně jako u spojnicového grafu a histogramu i zde je velmi užitečné vyrovnání. Vhodné je též zobrazení několika korelačních grafů v jednom obrázku, přičemž uspořádání těchto grafů odpovídá matici korelačních koeficientů.

Možnosti u korelačního grafu

	SG	SPSS	SYSTAT	STATIS	QPW	Excel	Improv	1234
Typy								
maticový	*	*	*	*	–	–	–	–
Vyrovnání								
přímka	*	*	*	*	*	.	.	.
parabola	–	3	2	5	–	–	–	–
exponenciála	*	–	–	*	*	–	–	–
vyhlazení	–	*	*	*	–	–	–	–

dd) *Kombinované grafy* (Combo) jsou vesměs kombinace se sloupcovým grafem. QPW obsahuje

Line–Bar, Area–Bar, High Low–Bar.

de) *Vícenásobné grafy* umožňují prezentovat více výsledků v jediném grafu. Tato mmožnost bývala donedávna vyhrazena pouze specializovaným grafickým systémům a objevovala se ve formě matice grafů ve statistických paketech.

QPW umožňuje zakreslit do jednoho obrázku až 4 grafy: Multiple Columns, Multiple 3–D Columns, Multiple Pies, Multiple Bar, Multiple 3–D.

7. Jednoduché programy

Tabulkové procesory jsou v podstatě prostředí pro vývoj aplikací (viz [7]). Dokonce nejjednodušší vzorec je vlastně program. Pro složitější aplikace je určen makrojazyk a specializované prostředky pro tvorbu uživatelských rozhraní (např. dialogových rámečků).

7.1 Tabulky. Především je třeba si uvědomit, že každá tabulka je program, který bude analogicky počítat s různými daty. Pokud si tedy „uděláme" pěkný výpočet, je možno výslednou tabulku uložit a používat jako kostru pro výpočty lišící se pouze daty (a navíc máme počáteční ilustraci).

7.2 Makra. Do políček tabulky lze také ukládat posloupnosti kroků, které s tabulkou provádíme. Je možné buď nastavit automatické zapisování příkazů ekvivalentních stisknutým klávesám, nebo tyto příkazy přímo zapsat do tabulky.

7.3 Programy. Pokud chceme připravit aplikaci s různými variantami, použijeme zřejmě navíc příkazy pro větvení programu. Pro některé složitější činnosti existují předem připravená makra, která lze využít v programech.

7.4 Uživatelské rozhraní. Většina TP obsahuje velmi silnou podporu pro vytváření aplikací. Základem jsou nejrůznější formy editorů dialogů, pomocí kterých lze vytvářet profesionálně vyhlížející rozhraní. QPW má např. User Interface Builder, ale podobné prostředky mají i Excel a 1234.

U QPW lze zejména vytvářet

- menu,
- dialogové rámečky (včetně kontroly vstupu).

Je přirozené, že tyto objekty mají všechny vlastnosti Windows.

7.5 Příklad ilustrace centrální limitní věty (CLV). Podle CLV se rozdělení průměrů a součtů hodnot nezávislých náhodných veličin rychle blíží k normálnímu rozdělení. Mnoho učitelů cítí potřebu to studentům ukázat a TP se na to velmi hodí.

Dobře je to vidět na průměrech z R[0,1] o délce 1, 2, 4 a 8. Ilustraci si mohou udělat studenti mechanicky vytvořením osmi sloupců R[0,1]:

	A	B	C	D	E	F	G	H
1	@RAND	@RAND	@RAND	@RAND	@RAND	@RAND	@RAND	@RAND

a tří sloupců průměrů:

I	J	K
@AVG(A1..B1)	@AVG(A1..D1)	@AVG(A1..H1)

které zkopírujeme do příslušného počtu řádků (např. 100). Dále vytvoříme tabulku četností pro intervaly o délce 0.1 ze sloupců H až K a zakreslíme do jednoho obrázku (pěkný je stužkový graf).

Posloupnost operací (výběry z nabídek a vyplňování dialogových rámečků) je možné uchovat si jako makro, které lze jednoduše spouštět, a tím velice rychle předvádět tuto ilustraci.

8. Závěr

Jak si jistě čtenář povšiml, zbývá ještě několik okrajových a méně používaných statistických metod, které lze (zatím) realizovat pouze v některých statistických paketech. Většinu z nich příliš nepostrádáme, ale přeci jenom jsou takové, po kterých může praxe zatoužit.

Kromě speciálních statistických grafů to jsou některé vícerozměrné a nelineární metody. Snad nejvíce chybí

- shluková analýza (je žádána, protože výsledkem je hezký graf),
- nelineární regrese (protože každý cítí povinnost vymyslet svůj originální model) a
- X11ARIMA (protože se to v ekonomii stalo zvykem).

Vesměs se jedná spíše o metody z oblasti popisné statistiky (X11ARIMA dokonce znalci překládají „pejsek a kočička vařili dort v mikrovlnné troubě").

Vzhledem k tomu, že technicky není žádný problém implementovat do tabulkového procesoru cokoliv, nemusí být pozice statistických paketů neotřesitelná.

Literatura

1. Brož, M., Brožová, P.: *Raz, dva, tři a čtyři.* CHIP, 11/1993, str. 140 – 146.
2. Gasteiger, D.: *Souboj velké trojky.* Bajt, 12/1993, str. 109 – 115 (převzato z BYTE 12/1993).
3. Luhen, J.: *Škatulky na vzorečky.* ComputerWorld, 3/1994.
4. Perratore, E.: *Tabulkový procesor nebo databáze?* Bajt, 9/1993, str. 84 – 86 (převzato z BYTE 9/1993).
5. Řezanková, H.: *Statistické výpočty v tabulkových procesorech.* Statistika, 8 – 9/1993, str. 352 – 362.
6. Stinson, C.: *Alternative Views: New Dimension in Spreadsheets.* PC Magazine, 16/1993 (September 28), str. 183 – 208.
7. Stinson, C.: *Less Is More.* PC Magazine, 1/1994 (January 11), str. 189 – 243.