Informační Bulletin

České statistické společnosti *číslo 2, ročník 19, 1. července 2008*

MATEMATICKÉ MODELOVANIE V JAZYKOVEDE

MATHEMATICAL MODELLING IN LINGUISTICS

Gejza Wimmer

FPV UMB, Banská Bystrica; MÚ SAV, Bratislava; ÚMS PřF MU Brno[1]

Abstract: The contribution is focused

(i) on mathematical modelling in linguistic by the use of discrete probability distributions (morphological productivity of stems in languages, semantic productivity of the language, theory of word lengths) and

(ii) on an attempt toward a unified derivation of some linguistic laws.

Úvod

Príspevok je ukážkou

a) *modelovania v jazykovede pomocou diskrétnych pravdepodobnostných rozdelení* a

b) *pokusu o jednotné odvodenie (veľkej) triedy jazykovedných zákonov.*

Práca vznikla v spolupráci s mnohými jazykovedcami. Výrazná bola spolupráca s prof. G. Altmannom.

[1]Pozvaná prednáška na konferencii STAKAN 2007. Podporené grantom VEGA, 1/3016/06 a projektom MŠMT ČR č. LC06024.

1. Jazykovedné zákony a diskrétne rozdelenia pravdepodobnosti

V jazykovede rovnako ako aj v iných vedných disciplínach sa pokúšame hľadať a odhaľovať zákony (zákonitosti). Podobne ako v iných empirických vedách toto hľadanie nie je jednoduché a môže s realizovať mnohými spôsobmi. Niekedy sa hľadajú analógie s postupmi v iných vedách (fyzika, chémia), niekedy špekulatívnou, deduktívnou cestou vyjdúc z nejakej teórie prídeme k hypotéze, ktorá sa po praktickom preverení neskôr môže považovať za zákonitosť, zákon. Často sa pokúšame získať zákonitosť z nameraných údajov, ale táto cesta vo väčšine prípadov stroskotá. Z nameraných údajov obyčajne môžeme (v najlepšom prípade) dostať nejakú funkciu, ktorá dostatočne fituje (vhodne napasuje) získané údaje, dobrú predikciu alebo interpoláciu. Ale veľmi zriedkavo zákon vo všeobecnosti. Pozrime sa na tri problémy.

Predpokladajme, že všetky procesy (samozrejme aj v jazykovede) podliehajú určitým zákonitostiam.

1. Hľadajme zákon *morfologickej produktivity slovných kmeňov* v jazyku. To znamená „Ako môže byť matematicky formulovaný (modelovaný) vznik nových slov a zánik starých slov v jazyku zapríčinený morfologickými zmenami.“
2. Analogicky nás môže zaujímať sémantická produktivita jazyka (vznik nových významov slova resp. strácanie rôznych starých významov).
3. Pýtame sa, akej zákonitosti podlieha (podľa akej zákonitosti sa správa) distribúcia slov podľa ich dĺžky v jazyku (v slovníku jazyka, v textoch jednotlivého autora, atď.).

Takisto nás môže zaujímať zákonitosť výskytu resp. tvorby viet podľa ich dĺžky (v texte), atď. Dĺžkou slova, vety rozumieme počet skladajúcich jazykových jednotiek (napr. slabík, slov, atď.).

Vo všetkých vyšetrovaných analyzovaných prípadoch predpokladáme, že namerané údaje sú dané vo forme frekvenčnej tabuľky. V prvom prípade táto tabuľka vyzerá nasledovne.

kmeň, ktorý vytvára x nových slov morfologickými zmenami	f_x – relatívna frekvencia týchto kmeňov
$x = 0$	f_0
$x = 1$	f_1
$\vdots$	$\vdots$
$x \geq 50$	f_{50}

Vo všetkých skúmaných prípadoch popíšeme hľadanú zákonitosť *diskrétnym pravdepodobnostným rozdelením*. Ako sa k nemu dostaneme? Základné možnosti sú

- použijeme urnovú schému
- kombinatorické úvahy
- iné cesty.

Morfologická produktivita slovných kmeňov v jazyku

V prvom probléme (morfologická produktivita slovných kmeňov v jazyku) z teórie (predchádzajúcich úvah a usudzovaní) je známe, že slová z hľadiska morfologickej produktivity podliehajú dvom procesom:

(i) nové slovo vznikne kreativitou (invenciou, improvizáciou) tvorcu (hovoriaceho, redaktora, spisovateľa),

(ii) nové slovo zanikne

  (a) naraz (napr. v jednom roku v časopise Spiegel bolo zistených 8000 nových slov, dovtedy nevyskytujúcich sa v žiadnom nemeckom slovníku, ktoré vznikli „len tak“ aj ihneď zanikli (neujali sa))

  (b) v priebehu času (keď prestane byť aktuálne, keď sa nájde nový, priliehavejší výraz).

Zánik slova je zapríčinený

(α) príjemca (poslucháč) rozumie novému slovu, ale ho nikdy nepoužije

(β) jazyk ako taký „má nastavenú“ hornú hranicu pre tvorbu nových slov z existujúcich slovných kmeňov (keby tomu tak nebolo, bolo by možné z jedného slovného kmeňa vytvoriť morfologickými pochodmi všetky slová a všetky ostatné slovné kmene by sa mohli eliminovať).

Prípad (i) popisuje *proces vzniku (birth process)*, prípad (ii) zase *proces zániku (death process)* definované v teórii pravdepodobnosti.

(i) V triede slovných kmeňov s vlastnosťou, že z každého kmeňa (v tejto triede) je možné vytvoriť x nových slov morfologickými zmenami je intenzita procesu vzniku

$$\lambda_x = \frac{f(x)}{g(x)} = \frac{a + xc}{b + (n - x - 1)c}\,, \quad x = 1, 2, \ldots, n - 1. \tag{1.1}$$

(ii) Intenzita procesu zániku pre x-tú triedu je

$$\mu_x = \frac{x}{n-x+1}, \quad x = 1, 2, \ldots, n. \tag{1.2}$$

V jazyku existuje rovnovážny stav. To znamená, že nové slová vznikajú a tiež slová zanikajú, ale nikdy neprichádza „ku katastrofe“ v zmysle komunikačnej schopnosti jazyka. Matematicky to vyjadrujeme rovnicami pre rovnovážny stav (steady state equations)

$$\begin{aligned} \lambda_0 P_0 &= \mu_1 P_1 \\ (\lambda_i + \mu_i) P_i &= \lambda_{i-1} P_{i-1} + \mu_{i+1} P_{i+1}\,, \qquad i = 1, 2, \ldots, n-1 \\ \lambda_{n-1} P_{n-1} &= \mu_n P_n. \end{aligned} \tag{1.3}$$

(P_x je pravdepodobnosť, že slovný kmeň patrí do x-tej triedy, t.j. môže sa z neho vytvoriť morfologickými procesmi x nových slov.)

Keď vyjdeme z predpokladu, že $\{P_0, P_1, \ldots, P_n\}$ je pravdepodobnostná funkcia (určuje istú distribúciu), tak použitím niektorých kombinatorických identít dostaneme

$$P_x = \frac{\binom{\frac{a}{c}+x-1}{x}\binom{\frac{b}{c}+n-x-1}{n-x}}{\binom{\frac{a+b}{c}+n-1}{n}}\,, \qquad x = 0, 1, \ldots, n,$$

čo je dobre známa pravdepodobnostná funkcia diskrétnej náhodnej veličiny s *Pólyovym rozdelením pravdepodobnosti.*

Vskutku, keď testujeme testom dobrej zhody údaje získané z mnohých jazykov, Pólyovo rozdelenie dáva vynikajúce výsledky, t.j. test „prakticky vždy“ nezamieta, že údaje sú realizáciami náhodnej veličiny s Pólyovym rozdelením pravdepodobnosti. Toto podporuje hypotézu, že morfologická produktivita jazyka sa môže matematicky modelovať (popísať) ako proces vzniku a zániku s intenzitami (1.1) a (1.2). Zákonitosť (zákon) je vyjadrená pomocou Pólyovho rozdelenia pravdepodobnosti. Intenzity majú veľmi rozumnú lingvistickú interpretáciu. Čitateľ procesu vzniku vyjadruje vytváranie (kreáciu) nových slovných konštruktov (hovoriacim, redaktorom, spisovateľom). Menovateľ vyjadruje „brzdiacu silu“ počúvajúceho (príjemcu, poslucháča, čitateľa) ako aj samotného jazyka (možnou hornou hranicou vyjadrenou číslom n). Odhadnuté parametre a, b, c, n prinášajú informáciu o jazyku, o autorovi, o čase vzniku textu, atď. Podrobnejši výklad pozri v článku Wimmer, Altmann (1995).

Sémantická produktivita jazyka

V druhom probléme (sémantická produktivita jazyka) sú v hre (sú navrhnuté) dve cesty (dve riešenia).

Prvá cesta vedie k procesu vzniku a zániku s intenzitou vzniku

$$\lambda_x = a + x\,, \qquad x = 0, 1, 2, \ldots$$

a intenzitou zániku

$$\mu_x = a + b + x\,, \qquad x = 0, 1, 2, \ldots$$

čo vedie na *Waringovo rozdelenie pravdepodobnosti*

$$P_x = \frac{b}{a+b}\frac{a^{(x)}}{(a+b+1)^{(x)}}\,, \qquad x = 0, 1, 2, \ldots$$

pre rovnovážny stav v komunikácii.

Druhá cesta vedie k *Bissingerovmu geometrickému rozdeleniu*, ktoré patrí do triedy rozdelení čiastočných súčtov (partial-sums distributions), pričom

$$P_x = \frac{p}{1-p}\sum_{j=x}^{\infty}\frac{(1-p)^j}{j}\,, \qquad x = 0, 1, 2, \ldots, 0 < p < 1.$$

Obidva modely boli testované na údajoch z maorského jazyka. Druhá cesta sa zdá byť výhodnejšia (priaznivejšia, lepšia) lebo automaticky zahŕňa (zaraďuje) tento výsledok do širšej (všeobecnejšej) teórie, potvrdzuje (podopiera) ju a táto teória sama dostáva deduktívnu podporu pomocou nej. Spomenutá teória bude prezentovaná neskôr v mojom príspevku. Viac o sémantickej produktivite v maorskom jazyku pozri v článku Wimmer, Altmann (1999).

Teória slovných dĺžok

Pokúsme sa analyzovať tretí problém (tvorba slov v jazyku podľa ich dĺžok). Slovná dĺžka je ovplyvnená obrovským množstvom faktorov. Na základe lingvistických analýz týchto faktorov by sa mala slovná dĺžka správať „ako biely šum“. Ale opak je pravdou. Príčiny tohoto sú:

(i) dĺžka slova je ovplyvnená mnohými faktormi, ale ona sama (slovná dĺžka) ovplyvňuje mnohé ďalšie vlastnosti jazyka. V tomto zmysle vplýva na rôzne atribúty (vlastnosti, črty) jazyka, „riadi“ mnohé zákonitosti. Toto spôsobuje samoreguláciu (self-regulation) v jazyku.

(ii) každé novovytvorené alebo zaniknuté slovo vnáša poruchy (disturbancie) aspoň do jednej zákonitosti jazyka. Keď je táto porucha malá, zaúčinkuje samo-regulačný proces. Ak je porucha veľká, spôsobuje spontánnu samoreguláciu k novému rovnovážnemu stavu (k novému atraktoru, k novému pravdepodobnostnému rozdeleniu). Tieto lingvistické úvahy môžu byť matematicky vyjadrené ako

$$P_x(= \text{pravdepodobnosť, že slovo má } x \text{ slabík}) \sim P_{x-1}. \tag{1.4}$$

Nech

$$P_x = g(x)P_{x-1}. \tag{1.5}$$

Keď zvolíme $g(x) = ax^{-b}$, $a > 0$, $b > 0$ (čo je vhodná funkcia vyhovujúca tzv. Menzerathovému zákonu), dostaneme *Conwayovo-Maxwellovo-Poissonovo rozdelenie*

$$P_x = \frac{a^x}{(x!)^b} P_0\,, \qquad x = 0, 1, 2, \ldots,$$

Toto rozdelenie pravdepodobnosti sa našlo v slovenčine, kórejčine, maďarčine, poľštine. Analýza veľkého množstva textov v mnohých jazykoch ukázala, že tvorba slov podľa ich dĺžok je stále len modifikácia základného modelu (1.5), konkrétne

(i) v prípade $b = 0$

je $g(x) = a$ − dostávame z (1.5) *geometrické rozdelenie*

$$P_x = (1-a)a^x, \qquad x = 0, 1, 2, \ldots$$

Toto rozdelenie pravdepodobnosti sa doteraz neobjavilo v praxi (u slovnych dĺžok).

Variantom je $g(x) = a(R-x+1)-$ dostávame z (1.5) *Palmovo-Poissonovo r.*

$$P_x = \frac{a^x R(R-1)\ldots(r-x+1)}{\sum_{j=0}^{R} R(R-1)\ldots(r-j+1)a^j}, \qquad x = 0, 1, \ldots, R.$$

Toto rozdelenie sa našlo v taliančine.

(ii) v prípade $b = 1$

je $g(x) = \dfrac{a}{x}$ – dostávame z (1.5) *Poissonovo rozdelenie*

$$P_x = \frac{e^{-a}a^x}{x!}, \qquad x = 0, 1, \ldots$$

Toto rozdelenie sa našlo v nemčine, ruštine, poľštine, maďarčine.

Variantmi sú $g(x) = \dfrac{a + cx}{x}$ – dostávame z (1.5) *negatívne binomické rozdelenie*

$$P_x = \binom{\frac{a}{c} + x}{x}(1 - c)^{\frac{a}{c}+1}c^x, \qquad x = 0, 1, \ldots$$

Toto rozdelenie sa našlo v nemčine, dánštine, nórštine.

$g(x) = \dfrac{a - cx}{x}$ – dostávame z (1.5) *binomické rozdelenie*

$$P_x = \binom{\frac{a}{c} - 1}{x}(1 + c)^{-\frac{a}{c}+1}c^x, \qquad x = 0, 1, \ldots, \frac{a}{c} - 1$$

Toto rozdelenie sa našlo v češtine, turečtine, latinčine, poľštine, estónčine.

$g(x) = \dfrac{a}{c + x}$ – dostávame z (1.5) *hyperpoissonovské rozdelenie*

$$P_x = \frac{a^x}{c^{(x)} \sum\limits_{j=0}^{\infty} \frac{a^j}{c^{(j)}j!}}, \qquad x = 0, 1, \ldots$$

Toto rozdelenie sa našlo v slovenčine, kórejčine, nemčine, staroislandštine, gréčtine, estónčine, starohebrejčine.

$g(x) = \dfrac{a + cx}{d + ex}$ – dostávame z (1.5) *hyperpascalovské rozdelenie*

$$P_x = \frac{\binom{\frac{a}{c}+x}{x}}{\binom{\frac{d}{e}+x}{x}}\left(\frac{c}{e}\right)^x P_0, \qquad x = 0, 1, \ldots$$

Toto rozdelenie sa našlo vo fínštine.

Rozšírením vzťahov (1.4) a (1.5) na

$$P_x = g(x) \sum_{j=1}^{x} h(j) P_{x-j}, \tag{1.6}$$

pričom v (1.6) uvažujeme $g(x) = \frac{a}{x}$ a $h(j) = j\Pi_j(\{\Pi_0, \Pi_1, \ldots\})$ je nejaká pravdepodobnostná funkcia), dostávame zovšeobecnené *Poissonovo rozdelenie*. V špeciálnom prípade, keď uvažujeme $\Pi_1 = \alpha$, $\Pi_2 = 1 - \alpha$, $0 < \alpha < 1$ *(Bernoulliho rozdelenie)*, dostávame z (1.6) *Hiratovo-Poissonovo rozdelenie*

$$P_x = \sum_{i=0}^{[x/2]} \binom{x-1}{i} \frac{e^{-a} a^{x-i}}{(x-i)!} (1-\alpha)^{x-2i} \alpha^j, \qquad x = 0, 1, \ldots$$

Toto rozdelenie sa našlo vo francúzštine a nemčine.

Ak

$$\Pi_j = \frac{e^{-bj} b^j j^{j-1}}{(j-1)!}, \quad j = 1, 2, \ldots$$

(Borelovo rozdelenie), dostávame z (1.6) *Consulovo-Jainovo-Poissonovo rozdelenie*

$$P_x = \frac{a(a+bx)^{x-1} e^{-(a+bx)}}{x!}, \qquad x = 0, 1, \ldots \ .$$

Toto rozdelenie sa našlo skoro v každom jazyku.

Pri hľadaní zákonitostí sme postupovali opačne. Použili sme špeciálny softvér tzv. „Altmann Fitter“, ktorý fituje namerané údaje (vo forme frekvenčnej tabuľky) na viac ako 200 pravdepodobnostných rozdelení. Podľa vhodnosti fitu (napasovania), vždy niektoré rozdelenia boli zvolené na ďalšiu analýzu. Podľa zhody (vybraných) matematických vlastností týchto rozdelení s možnými lingvistickými interpretáciami týchto vlastností sme sa pokúšali sformulovať zákonitosť (teda vybrať vhodné rozdelenie pravdepodobnosti pre skúmanú zákonitosť).

Keď zhrnieme doterajšie úvahy, môžeme skonštatovať, že namiesto tradičnej cesty objavovania určitej zákonitosti, ktorá používa metódu

postaviť hypotézu $\Longrightarrow$ *zozbierať údaje* $\Longrightarrow$ *testovať hypotézu*

sme použili

sformulovať (počiatočný) model
(triedu rozdelení)
⇓
zozbierať údaje
⇓
vybrať model z podtriedy
(jednotlivé rozdelenie)
⇓
test dobrej zhody pre vybrané rozdelenie
⇓
sformulujeme hypotézu (zákonitosť)
(určíme stochastický mechanizmus deja)

Modelovanie pomocou diskrétnych pravdepodobnostných rozdelení vo všeobecnosti v empirických vedách môže prebiehať dvomi spôsobmi:

(i) pomocou modelovania vytvárajúcich mechanizmov, ktoré sa historicky vyvíjali a prinášajú informácie o vzniku dát,

(ii) pomocou usporiadania nameraných údajov a hľadania zákonitostí v tomto usporiadaní.

Vyššie spomenuté výskumy patrili k spôsobu (i). Spôsobom (ii) odhalíme tiež veľa lingvistických modelov – zákonitostí (spomenieme napr. *Zipfov zákon* s jeho mnohými modifikáciami). Veľmi široká trieda diskrétnych pravdepodobnostných modelov, ktorá sem tiež patrí sú vyššie spomenuté rozdelenia *čiastočných súčtov*.

Poďme teraz k druhej téme nášho príspevku.

2. Jednotné odvodenie (veľkej) triedy jazykovedných zákonov

V každej vednej oblasti začína výskum roztrúsene, teda ako membra disiecta, lebo neexistuje žiadna teória, ktorá by systematizovala vedomosti a pomocou ktorej by sa mohli formulovať hypotézy.

Sami tí, čo robia výskum, majú rozličné vedecké záujmy a najmä v začiatkoch skúmajú úzke výseky reality. Neskôr sa spájajú krok za krokom nesúrodé oblasti výskumu (napr. jednotná reprezentácia všetkých druhov pohybu v makrosvete pomocou Newtonovej teórie). Staršie teórie sa obyčajne stávajú špeciálnymi prípadmi nových teórií. Hovorí sa o epistemickej integrácii týkajúcej sa poznávania alebo vedenia (Bunge (1983): „Integrácia

prístupov, dát, hypotéz, teórií, ba dokonca celých oblastí výskumu je potrebná nielen na vysvetlenie vecí, ktoré silne interagujú so svojim okolím. Epistemická integrácia je potrebná všade, lebo neexistujú úplne izolované veci. Každá vlastnosť súvisí s inými vlastnosťami. Každá vec je systém alebo časť systému... Teda tak isto, ako rôznorodosť sveta (skutočnosti) si vyžaduje veľké množstvo disciplín, ktoré realitu skúmajú, ich integrácia je nutná z hľadiska jednotnosti sveta.“).

Kvantitatíva lingvistika stojí na začiatku takéhoto vývoja. Existujú dva „veľké“ integrujúce „cezhraničné“ prístupy – Köhlerova (1986) jazyková synergetika a Hřebíčkova (1997) teória textov. Ďalej sú tu niektoré „menšie“, ktoré spájajú rôzne jazykovedné javy. Medzi ne patria napríklad:

a) Baayen (1989), Chitashvili a Baayen (1993), Zörnig a Boroda (1992), Balasubrahmanyan a Naranan (1997) ukazujú, že distribúcie, ktoré dostaneme usporiadaním frekvencií môžu byť transformované a vyšetrované ako nerankované, čo bolo neformálne naznačené už Rapoportom (1982).
b) Altmann (1990) ukázal, že Bühlerova „teória“ je len špeciálnym prípadom Zipfovej teórie (1949), ktorý videl „princíp najmenšej námahy (sily)“ za každým ľudským javom (úkazom, činnosťou).
c) Viac integrujúci je Menzerathov zákon, ktorého účinky môžeme pozorovať nielen v rôznych oblastiach lingvistiky, ale aj v molekulárnej biológii, sociológi a psychológii (Altmann, Schwibbe (1989)). Je to paralela k alometrickému zákonu a môžeme ho nájsť aj v teórii chaosu (Schroeder (1990), Hřebíček (1997)) alebo v muzikológii (Boroda, Altmann (1991)).
d) Orlov, Boroda a Nadarejšvili (1982) hľadali spoločné črty vyskytujúce sa v jazykovede, hudbe aj vo výtvarnom umení. Objavili platnosť Zipfovho-Mandelbrotovho zákona.
e) Krylov, Naranan a Balasubrahmanyan, všetko fyzici, prišli zhodne a nezávisle k poznatku, že princípom maximalizácie entropie sa dajú vynikajúco odvodiť distribúcie niektorých lingvistických entít.

Mohli by sme dlho pokračovať vo vymenovaní príkladov „zjednocovania oblastí“. Vyššie sme uviedli len niekoľko príkladov. Na každom prípade môžeme vidieť spoločný poznatok, že v podstate všetko vedie k teórii systémov. Všetky veci sú systémy. Spájame dva oblasti ak nachádzame izomorfizmy, podobnosti, paralely medzi príslušnými systémami alebo ak zistíme, že ony sú špeciálnymi prípadmi všeobecnejšieho systému. Z času na čas treba spraviť takúto integráciu, aby sme dostali jednotnejšie teórie a lepšie utriedenie vedomostí o skúmanom objekte. V tomto príspevku chceme ukázať prístup

ktorý zjednocuje mnohé známe lingvistické hypotézy, ľahko sa dá zovšeobecniť a je veľmi jednoduchý – aj keď jednoduchosť nepatrí k nutným cnostiam vedy (pozri. Bunge (1983)). Tento prístup je logické rozšírenie „synergetického“ prístupu (pozri. Wimmer, Köhler, Grotjahn, Altmann (1994), Wimmer, Altmann (1996), Altmann, Köhler (1996)). Jednotlivé hypotézy patriace k tomuto systému boli skôr sformulované ako empirické krivky (funkcie), ktoré dobre fitujú určité úkazy (javy) alebo boli odvodené z rôznych iných prístupov.

Spojitý prístup

V jazykovede sa môžeme stretnúť so spojitými premennými predovšetkým vo fonetike, ale si musíme uvedomiť, že „premenná“ je iba konštrukt (koncept) nášho nástroja – matematiky, ktorým sa snažíme vystihnúť (zachytiť) stupne skutočných vlastností vecí. Mnohokrát ich transformujeme z „diskrétnych“ na „spojité“ (napr. priemer) alebo naopak (napr. rozdelením spojitej stupnice na intervaly), podľa toho, ako to potrebujeme. Toto nie je nič neobvyklé vo vede. V tomto zmysle neurobíme nič zlého, ak modelujeme spojité javy použitím diskrétnych modelov alebo naopak. „Spojitý“ a „diskrétny“ sú vlastnosti našich koncepcií, prvé aproximácie nášho epistemického snaženia (snaženia poznávať, vedieť).

Začneme z dvoch predpokladov, ktoré sú v jazykovede veľmi rozšírené a akceptované. Najprv spojitý prípad:

(i) Nech y je spojitá premenná. Zmena dy tejto premennej je regulovaná (ovládaná) priamo jej veľkosťou (závisí od jej veľkosti), lebo každá lingvistická premenná je konečná a je časťou samoregulujúceho sa systému. Preto pri modelovaní môžeme vždy použiť relatívny pomer zmeny dy/y.

(ii) Každá lingvistická premenná y je spojená s najmenej jednou inou premennou x ktorá ovplyvňuje jej správanie sa a ktorú budeme považovať v danom prípade za nezávislú premennú. Nezávislá premenná x ovplyvňuje závislú premennú aj cez jej zmenu dx, ktorá je spätne riadená (ovplyvnená) rôznymi mocninami hodnôt x, ktoré sú prepojené s rôznymi inými faktormi (silami, vplyvmi, atď.).

Predpokladajme, že x a y sú rôzne škálované a preto dva vyššie uvedené predpoklady môžeme formálne napísať ako

$$\frac{dy}{y-d} = \Big(a_0 + \sum_{i=1}^{k_1} \frac{a_{1i}}{(x-b_{1i})^{c_1}} + \sum_{i=1}^{k_2} \frac{a_{2i}}{(x-b_{2i})^{c_2}} + \dots\Big)dx \qquad (2.1)$$

s $c_i \neq c_j$, $i \neq j$. (Poznamenávame len, že pre $k_s = 0$ je $\sum_{i=1}^{k_s} \frac{a_{si}}{(x-b_{si})^{c_s}} = 0$.) Konštanty a_{ij} musia byť v každom jednotlivom prípade rozlične interpretované. Reprezentujú vlastnosti, „sily“, „príkazové (riadiace) parametre“, požiadavky systému, atď., ktoré sa aktívne zúčastňujú pri prepojení premenných x a y (pozri Köhler (1986, 1987, 1989, 1990)), ale zostávajú konštatné pre podmienku *ceteris paribus* (modelujeme za predpokladu, že ostatné premenné sú konštantné). V diferenciálnej rovnici (2.1) sú premenné už separované. Riešenie (2.1) je

$$y = Ce^{a_0 x} \prod_{i=1}^{k_1} (x - b_{1i})^{a_{1i}} e^{\left(\sum_{j \geq 2} \sum_{i=1}^{kj} \frac{a_{ji}}{(1-c)(x-b_{ji})^{c_j - 1}} \right)} + d \qquad (2.2)$$

Najčastejšie a najznámejšie riešenia tohto prístupu sú

(a) počet všetkých slov – počet odlišných slov v texte, tzv. type-token krivky,

(b) Menzerathov zákon, pri ktorom ide o zákonitosti vzťahu velkosti konštruktu a konštituent,

(c) Piotrowského-Bektajeva-Piotrowskej zákon o raste slovníka,

(d) Narananov-Balasubrahmanyanov model výskytu slov,

(e) Geršićovo-Altmannovo rozdelenie pre trvanie samohlások, ktore je identicke s jednym Narananovym-Subrahmanyanovym rozdelenim

(f) Jobov-Altmannov model zmeny jednotlivých foném, modelujuci pravdepodobnost zmeny foném v závislosti na ich norme a komplexite

(g) Tuldavov zákon polysémie,

(h) Uhlířovej zákon vyjadrujúci závislosť výskytu podstatných mien na danej pozícii vo vete,

(i) spojitá varianta Zipfovho-Mandelbrotovho zákona a jej špeciálne prípady (pozri napr. Zörnig-Altmann (1993)) atd. atd.

Spojitý dvojrozmerný prístup

Doterajší prístup samozrejme nie je vo všeobecnosti postačujúci. V synergetickej lingvistike existuje množstvo vzťahov, ktoré nemožno vystihnúť pomocou funkcie jednej premennej, keď ostatné premenné „zatajíme“ podmienkou „ceteris paribus“ (považujeme ich za konštantné). Niekedy ich musíme vziať do úvahy.

Najprv uvažujme jednoduchý špeciálny prípad vzorca (2.1), a síce

$$\frac{dy}{y} = \Big(a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \dots\Big)dx \tag{2.3}$$

ktorého riešenie je

$$y = Ce^{a_0 x} x^{a1} e^{-a_2/x}. \tag{2.4}$$

Vyjadruje napr. Geršićov-Altmannov (1988) model trvania samohlásky. V (2.3) predpokladáme, že všetky ostatné faktory (okrem x) sú slabšie ako x a môžu byť považované za konštanty vzhľadom k mocninám x (napr. k a_2/x^2, a_3/x^3 atď.). Ale v synergetickej lingvistike toto nie je pravidlom. V mnohých modeloch sa ukázalo, že treba uvažovať závislosť jednej premennej od mnohých iných premenných. Dostávame v prvom priblížení sústavu

$$\frac{\partial y}{\partial x} = y\Big(a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \dots\Big);\ \frac{\partial y}{\partial z} = y\Big(b_0 + \frac{b_1}{z} + \frac{b_2}{z^2} + \dots\Big), \tag{2.5}$$

ktorej riešenie je

$$y = Ce^{a_0 x + bz} x^{a_1} z^{b_1} \exp\left(-\sum_{i=1}^{\infty} \frac{a_{i+1}}{i x^i} - \sum_{i=1}^{\infty} \frac{b_{i+1}}{i z^i} \right). \tag{2.6}$$

Špeciálne prípady (2.6) sa často vyskytujú v synergetickej lingvistike, pričom sa uvažuje o viac ako dvoch premenných. Takýto zovšeobecnený systém, ktorý môže obsahovať ľubovoľné (konečné) množstvo premenných, môže obsiahnuť v podstate celú synergetickú lingvistiku a je aplikovateľný na veľmi zložité systémy. Niektoré dobre známe prípady zo synergetickej lingvistiky sú

$$y = Cx^a z^b \tag{2.7}$$

$$y = Ce^{ax+bz} \tag{2.8}$$

$$y = Ce^{ax+bz} x^a z^b \tag{2.9}$$

atď.

Diskrétny prístup

Ak X je diskrétna premenná (čo zvyčajne v lingvistike býva), potom použijeme namiesto infinitezimálneho prírastku dx diferenciu $\Delta x = x-(x-1) = 1$. Obyčajne sa jedná o celočíselné nezáporné náhodné premenné s pravdepodobnostnou funkciou $\{P_0, P_1, \ldots\}$, preto uvažujeme relatívnu zmenu

$$\frac{\Delta P_{x-1}}{P_{x-1}} = \frac{P_x - P_{x-1}}{P_{x-1}} \tag{2.10}$$

a dostávame diskrétnu analógiu vzťahu (2.1), síce

$$\frac{\Delta P_{x-1}}{P_{x-1}} = a_0 + \sum_{i=1}^{k_1} \frac{a_{1i}}{(x-b_{1i})^{c_1}} + \sum_{i=1}^{k_2} \frac{a_{2i}}{(x-b_{2i})^{c_2}} + \ldots \ . \tag{2.11}$$

Ak $k_1 = k_2 = \ldots = 1$, $d = b_{11} = b_{21} = \ldots = 0$, $c_i = i$, $a_{i1} = a_i$, $i = 1, 2, \ldots$, ekvivalentná forma (2.11) je

$$P_x = \Big(1 + a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \ldots\Big) P_{x-1}. \tag{2.12}$$

Systém, ktorý sa najviac používa v lingvistike, je

$$P_x = \Big(1 + a_0 + \frac{a_1}{x-b_1} + \frac{a_2}{x-b_2}\Big) P_{x-1}. \tag{2.13}$$

Jeho riešenie je

$$\begin{aligned} P_x = \ & \frac{(1+a_0)^x \binom{C-B+x}{x} \binom{D-B+x}{x}}{\binom{-b_1+x}{x}\binom{-b_2+x}{x}} \, {}_3F_2^{-1}(1, C-B+1, D-B+1; \\ & -b_1+1, -b_2+1; 1+a_0), \qquad x = 0, 1, 2, \ldots, \end{aligned} \tag{2.14}$$

kde ${}_3F_2$ je zovšeobecnená hypergeometrická funkcia,

$$B = \frac{b_1 + b_2}{2}\,,$$

$$C = \frac{a_1+a_2-\left(2(1+a_0)^2(b_1-b_2)^2-2(1+a_0)(a_1-a_2)(b_1-b_2)+(a_1+a_2)^2\right)^{1/2}}{2(1+a_0)},$$

$$D = \frac{a_1+a_2+\left(2(1+a_0)^2(b_1-b_2)^2-2(1+a_0)(a_1-a_2)(b_1-b_2)+(a_1+a_2)^2\right)^{1/2}}{2(1+a_0)}.$$

Z rekurentných vzťahov (2.12) a (2.13) dostávame množstvo známych rozdelení, ktoré sa používajú v lingvistike, napr. geometrické rozdelenie, Katzovu

triedu rozdelení, diverzifikačné rozdelenia, rozdelenia usporiadaných frekvencií, rozdelenia vzdialeností, Poissonovo rozdelenie, negatívne binomické rozdelenie, hyperpoisonovo rozdelenie, hyperpascalovo rozdelenie, Yuleovo rozdelenie, Simonovo rozdelenie, Waringovo rozdelenie, Johnsonovo-Kotzovo rozdelenie, negatívne hypergeometrické rozdelenie, Conwayovo-Maxwellovo-Poissonovo rozdelenie, atď., atď.

Zákony (zákonitosti), ktoré sa dajú formulovať pomocou tohto systému rozdelení sú napr. Frumkinin zákon, rôzne zákony rozdelenia dĺžok slabík, slov a viet, niektoré formy Zipfovho zákona, zákony usporiadania, rozdelenia syntaktických vlastností, Krylovov sémantický zákon, atď., atď.

Diskrétny dvojrozmerný prístup

Rovnakým spôsobom ako v spojitom prípade, môžeme zovšeobecniť aj diskrétny prípad na niekoľko premenných. Pretože doterajšia lingvistická analýza v tomto smere je málo početná (článok Uhlířovej a Wimmera (2003) a článok o slabikovej štruktúre od Zörniga a Altmanna (1993)), ukážeme len metódu.

V jednorozmernom diskrétnom prístupe sme mali rekurentné vzorce (napr. (2.12) alebo (2.13)), ktoré sa môžu písať ako

$$P_x = g(x)P_{x-1}, \tag{2.15}$$

kde $g(x)$ bola časť nekonečnej postupnosti. Pretože tentokrát máme dve premenné, môžeme sformulovať model nasledovne

$$P_{i,j} = g(i,j)P_{i,j-1},\ P_{i,j} = h(i,j)P_{i-1,j}, \tag{2.16}$$

kde $g(i,j)$ a $h(i,j)$ sú rôzne funkcie i a j. Rovnice sa musia riešiť súčasne. Výsledok samozrejme závisí od zvolených funkcií $g(i,j)$ a $h(i,j)$. Takto Uhlířová a Wimmer (2003) dostali dvojrozmerné binomické rozdelenie, kým Zörnig a Altmann dostali dvojrozmerné Conwayovo-Maxwellovo-Poissonovo rozdelenie.

Záver

Skutočnosť, že týmto spôsobom môžeme integrovať rôzne hypotézy, má niekoľko dôsledkov:

(i) Ukazuje sa, že v pozadí mnohých dejov v lingvistike existuje jednotný mechanizmus – reprezentovaný vzťahmi (2.1), (2.5), (2.11), (2.16). V rámci tohto mechanizmu môžeme kombinovať premenné a „sily“.

(ii) Vzorce (2.1), (2.5), (2.11), (2.14) predstavujú systémy ktoré môžu obsahovať aj mimosystémové faktory.

(iii) Tento prístup dovoľuje induktívne testovať nové, doteraz neznáme vzťahy a systemizovať ich do teórie s korektnou interpretáciou faktorov; toto obyčajne nie je možné ak postupujeme induktívne. Exploratívna časť práce sa preto môže urýchliť vhodným softvérom. Nedá sa predpokladať, že použitím tohto prístupu budeme môcť všetko v jazyku vysvetliť, ale môžeme uspokojivo zjednotiť a interpretovať aposteriórne mnoho rôznorodých javov. Podrobnejšie pozri Wimmer, Altmann (2005).

Literatúra

[1] Altmann, G. (1990). *Bühler or Zipf? A re-interpretation.* In: Koch, W.A. (Hrsg.), Aspekte einer Kultursemiotik: 1-6. Bochum: Brockmeyer.

[2] Altmann, G., Köhler, R. (1996). *„Language Forces" and synergetic modeling of language phenomena.* Glottometrika 15, 62-76.

[3] Altmann, G., Schwibbe, M.H. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim: Olms.

[4] Baayen, R.H. (1989). *A corpus-based approach to morphological productivity.* Amsterdam: Centrum voor Wiskunde en Informatica.

[5] Balasubrahmanyan, V.K., Naranan, S. (1997). *Quantitative linguistics and complex system studies.* Journal of Quantitative Linguistics 3, 177-228.

[6] Boroda, M.G., Altmann, G. (1991). *Menzerath's law in musical texts.* Musikometrika 3, 1-13.

[7] Bunge, M. (1983). *Understanding the world.* Dordrecht: Reidel.

[8] Chitashvili, R.J., Baayen, R.H. (1993). *Word frequency distributions of texts and corpora as large number of rare event distributions.* In: Hřebíček, L., Altmann, G. (eds.), Quantitative Text Analysis: 54-135. Trier: WVT.

[9] Geršić, S., Altmann, G. (1988). *Ein Modell für die Variabilität der Vokaldauer.* Glottometrika 9, 49-58.

[10] Hřebíček, L. (1997). *Lectures on text theory.* Prague: Oriental Institute.

[11] Köhler, R. (1986). *Zur linguistischen Synergetik.* Struktur und Dynamik der Lexik. Bochum: Brockmeyer.

[12] Köhler, R. (1987). *Systems theoretical linguistics.* Theoretical Linguistics 14, 241-257.

[13] Köhler, R. (1989). *Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation.* Glottometrika 11, 1-18.

[14] Köhler, R. (1990). *Elemente der synergetischen Linguistik.* Glottometrika 12, 179-187.

[15] Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š. (1982). *Sprache*, Text, Kunst. Quantitative Analysen. Bochum: Brockmeyer

[16] Rapoport, A. (1982). *Zif's law re-visited.* In: Guiter, H., Arapov, M.V. (eds.), Studies on Zipf's Law: 1-28. Bochum: Brockmeyer.

[17] Schroeder, M. (1990). *Fractals, chaos, power laws. Minutes from an infinite paradise.* New York: Freeman.

[18] Uhlířová, L., Wimmer, G. (2003). *A contribution to word length theory.* In: Festschrift für Werner Lehfeldt zum 60. Geburtstag (Kempgen, S., Schweier, V., Berger, T. (Eds.)). München: Verlag Otto Sagner, 524-530.

[19] Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G. (1994). *Towards a theory of word length distribution.* Journal of Quantitative Linguistics 1, 98-106.

[20] Wimmer, G., Altmann, G. (1995) *A model of morphological productivity.* Journal of Quantitative Linguistics 2, 212-216.

[21] Wimmer, G., Altmann, G. (1996). *The theory of word length: Some results and generalizations.* Glottometrika 15, 112-133.

[22] Wimmer, G., Altmann, G. (1999) *Rozdelenie polysemie v maorijčine* (Distribution of Polysemy in Maori), Pange Lingua; Genzor, J., Ondrejovič, S. (Eds.). Bratislava: Veda, 17-25.

[23] Wimmer, G., Altmann, G. (2005). *Unified derivation of some linguistic laws.* In: Quantitative Linguistics. An International Handbook. (Köhler, R., Altmann, G., Piotrowski, R., G. (Eds.)). Berlin: Walter de Gruyter, 791-807.

[24] Zipf, G.K. (1949). *Human behavior and the principle of least effort.* Reading: Addison-Wesley.

[25] Zörnig, P., Altmann, G. (1993). *A model for the distribution of syllable types.* Glottometrika 14, 190-196.

[26] Zörnig, P., Boroda, M.G. (1992) *The Zipf-Mandelbrot law and the interdependencies between frequency structure and frequency distribution in coherent texts.* Glottometrika 13, 205-218.

VÝUKA JEDNOROZMĚRNÉ A DVOUROZMĚRNÉ ANALÝZY KATEGORIÁLNÍCH DAT

Hana Řezanková

Abstract: The paper is focused on teaching one-way and two-way analysis of categorical data. It is based on experiences with teaching at the University of Economics, Prague. The preparing of teaching, the choice of the software and data files and the comparison of some software packages are dealt with. The possibilities of systems SAS Enterprise Guide, SPSS, S-PLUS. STATGRAPHICS and STATISTICA are compared in the area of categorical data analysis.

1. Příprava předmětu

Tento příspěvek se věnuje jednak obecně předmětům zaměřeným na analýzu dat, jednak konkrétně výuce základů analýzy kategoriálních dat v bakalářském, event. magisterském studiu. Předpokládá se tedy, že buď celá nebo část výuky se koná na počítačové učebně vybavené vhodným programovým vybavením. Mnohé úvahy vycházejí ze specifické organizace studia na Vysoké škole ekonomické v Praze. Avšak vzhledem k tomu, že kreditní způsob studia, včetně ECTS kreditů, má jistý obecný základ, neměly by být mezi vysokými školami zásadní odlišnosti v oblastech, o kterých bude dále pojednáno.

Připravuje-li pedagog na vysoké škole výuku předmětu, je často limitován různými faktory. Málokdy má plnou volnost, aby předmět akreditoval podle svých představ. Někdy je již předmět akreditován a pedagog má k dispozici obsah, který v rámci stanoveného počtu hodin naplňuje výkladem teorie a řešením příkladů.

Pokud je pedagog garantem a připravuje podklady pro akreditaci předmětu, pak je třeba zohlednit zejména následující:

- zda je pro předmět určeno, *kolik hodin týdně* má být vyučován, případně je-li dána dolní či horní hranice této hodinové dotace (pokud je v této oblasti určitá volnost, pak záleží na rozsahu a hloubce látky, kterou by v rámci daného tématu měli zvládnout studenti určitého oboru, stupně studia, případně semestru či ročníku),
- zda je u předmětu vymezeno *členění na přednášky a cvičení* a pokud ano, zda se bude na počítačové učebně konat také přednáška, či nikoli

(nemusí být součástí akreditace, lze měnit operativně například podle počtu přihlášených studentů v semestru, viz níže),

- jaký *software* je pro analýzu dat k dispozici (v době akreditace nemusí být žádný, může být jeden, či více specializovaných programových produktů, používaný software lze měnit operativně podle znalostí či zájmů studentů účastnících se kurzů),
- zda existuje vhodný *studijní materiál* dostupný v potřebném počtu studentům.

K prvnímu bodu zřejmě není potřeba žádný komentář. Přejděme tedy k bodu druhému. Pokud má smysl rozlišovat přednášky a cvičení, pak můžeme uvažovat dvě základní situace. Je-li studentů více, než předpokládaná kapacita učebny na cvičení, a je menší kapacita pedagogů, pak je výhodné zorganizovat přednášku pro všechny studenty daného předmětu v semestru (varianta A). Na druhou stranu, pokud kapacita počítačových učeben s vhodným programovým vybavením stačí pojmout všechny zájemce o předmět a je dostatečná kapacita pedagogů, je výhodné, aby se i přednášky konaly na počítačové učebně, případně se přednášky a cvičení nerozlišovaly, tj. forma výuky může být buď označena jako přednáška nebo jako cvičení (varianta B). Pak lze kombinovat teoretický výklad s bezprostředním praktickým využitím určité metody. Nevýhodou může být někde menší tabule na počítačových učebnách, což lze kompenzovat například promítáním vzorců z dokumentu na počítači pomocí datového projektoru.

Není-li zatím k dispozici *specializovaný software*, je vhodné podniknout kroky k jeho získání (zakoupit z prostředků školy, resp. fakulty, požádat o grant FRVŠ apod.). Otázkou je, který produkt pořídit. Problematika výběru vhodného softwaru bude probrána v dalším odstavci a následně později podrobně v souvislosti s výukou analýzy kategoriálních dat. Na většině škol[1] je hlavním hlediskem cena. Ovšem když vezmeme v úvahu, že lze o pořízení programového systému zažádat prostřednictvím grantu, nemusí být nutně cena limitujícím faktorem. Je důležité zohlednit, zda by mohl být software využit i v jiných předmětech. Pokud by se nepodařilo software získat do začátku výuky, pak lze provádět některé výpočty v systému MS Excel (zejména transformace dat, tabulky rozdělení četností, dosazování do vzorců), využívat prostředků na Internetu (viz [7]) aj.

Je-li softwarových produktů k dispozici více, je třeba zvážit, který je pro studenty daného oboru a stupně studia nejvhodnější. Buď ten, který již znají,

[1] Termínem „škola“ bude nadále označován vysoká škola či fakulta, přesněji řešeno subjekt, který spravuje počítačové učebny, zabezpečuje nákup a instalaci programových systémů aj.

nebo ten, ve kterém jsou vyučované metody zastoupeny co nejvíce, a to jak pokud jde o počet metod, tak co se týká použitých postupů, dílčích možností, grafických výstupů apod. Nejsou-li studenti se softwarem obeznámeni, dalším zvažovaným faktorem by měla být snadnost ovládání. Studenti by se měli soustředit na analýzu dat a interpretaci výsledků a neměli by se příliš „rozptylovat“ výukou programového systému (s výjimkou situace, kdyby byl software využíván v dalších navazujících předmětech a součástí stávajícího předmětu by měla být výuka samotného programového systému).

Kromě samotných metod je třeba vzít v úvahu možnosti programového systému v oblasti přípravy dat, jejich popisu, transformací a práce s chybějícími údaji. Důležitým faktorem je také to, jaké *existující datové soubory* by měly být při výuce využívány. Převést samotná data z jednoho systému do druhého problém obvykle není. Problémem je to, že se třeba nepřevedou popisy proměnných, popisy použitých kódů (číselníky), či identifikace chybějících údajů, což může znesnadnit výuku.

Se softwarem úzce souvisí dostupný studijní materiál. Pokud jsou k dispozici skripta nebo existuje cenově dostupná kniha zaměřená na vyučované téma a obsahuje příklady s využitím určitého programového systému, pak při možnosti výběru produktu by tato skutečnost měla být zohledněna. Jinak je vhodné nějaký materiál připravit, alespoň formou dílčích dokumentů poskytovaných pouze účastníkům příslušných kurzů. I když si v současné době ještě někteří studenti zapisují při výuce poznámky, stále více se dožadují skript nebo knih obsahově přesně korespondujících s vyučovaným předmětem.

Dosud nerozsáhlejším studijním materiálem k předmětům zaměřeným na analýzu kategoriálních dat, který je dostupný v češtině, je kniha [4]. Na VŠE jsou používána například skripta [5]. Částečný výklad některých pasáží je též k dispozici elektronicky na Internetu v rámci interaktivní učebnice IASTAT, viz [7]. Text však již nebyl delší dobu aktualizován a některé úpravy by bylo vhodné provést. Při přípravě studijních materiálů se lze kromě anglické a české literatury (seznam základní viz [5]) inspirovat i na Slovensku, viz např. [2], [3] a [8].

2. Dopad volné tvorby studijního plánu na obsah a způsob výuky

Vezměme konkrétnější situaci týkající se VŠE v Praze. Studijní plán studenta je individuální, jediným omezujícím faktorem je splnění studijních povinnosti ve formě počtu získaných kreditů a počtu složených zkoušek. I když s novým způsobem studia založeným na ECTS kreditech je v prvním semestru tzv. pevný rozvrh, již ve druhém a dalších semestrech se může skladba předmětů

jednotlivých studentů lišit vzhledem k tomu, že někteří neuspěli u zkoušek, jiným z důvodu nemoci byly některé předměty omluveny, další přerušili studium.

Na přednáškách a cvičeních se pak setkávají studenti různých semestrů a ročníků jednoho oboru. Pokud je předmět určen jako oborově povinný, nebo je součástí skupiny předmětů, z níž si student musí něco vybrat (oborově volitelný), jeho volná kapacita (volná místa v učebně po zapsání studentů určitého oboru) je dána k dispozici studentům ostatních oborů v podobě celoškolně volitelného předmětu. Vzhledem k malým počtům studentů v jednotlivých oborech se také často stává, že je předmět akreditován jako povinně či volně volitelný pro více studijních oborů, resp. specializací[2]. Na přednáškách a cvičeních se pak setkávají také studenti různých oborů.

Nejde samozřejmě o to, že se tam studenti setkají, jde o to, že je potřeba pro ně připravit výuku tak, aby byla pro všechny srozumitelná a aby se pokročilejší nenudili. Je potřeba vzít v úvahu, že někteří studenti již absolvovali různé předměty a jiní ne. To souvisí i se znalostí určitého softwarového produktu. V případě, že je na škole k dispozici více programových systémů, které je možno využít, každý student může znát jiný a někdo třeba žádný. Ideální řešení pro tuto situaci asi nalézt nelze a pedagog je ve velmi obtížné pozici.

Na VŠE se navíc situace zkomplikovala tím, že byl zaveden nový způsob studia s jinými pravidly a vzhledem k malým počtům studentů v jednotlivých oborech je potřeba slučovat výuku pro různé obory. Například v původním způsobu studia byl akreditován předmět se dvěma hodinami přednášek a dvěma hodinami cvičení za 14 dní, což bylo realizováno střídavou výukou přednášek a cvičení podle lichých a sudých týdnů. V novém způsobu studia se již liché a sudé týdny nerozlišují, takže vzhledem k tomu, že jsou při výuce využívány počítače, je nový předmět akreditován se dvěma hodinami cvičení týdně. Jednoho konkrétního kurzu by se tedy měli účastnit studenti různých oborů s tím, že u některého by rozdělení na přenášky a cvičení bylo již pouze formální (výuka ve variantě B). Bohužel výše uvedené nelze zobecnit a v některých případech jsou buď nové předměty akreditovány s jiným počtem hodin výuky, nebo jsou akreditovány předměty zcela odlišné.

3. Výběr softwaru a datových souborů

Problematiku výběru softwaru lze ilustrovat na předmětu, jehož cílem je seznámit studenty především bakalářského stupně studia se základy analýzy dat uspořádaných do kontingenčních tabulek a se související problematikou.

[2]Na VŠE existuje tzv. vedlejší specializace, kterou si volí studenti magisterského stupně studia. Z této vedlejší specializace skládají státní zkoušku.

Na VŠE je již léta takový předmět vyučován s využitím programového systému SPSS. Dále jsou k dispozici starší systém STATGRAPHICS a systém SAS. Před několika lety byl výpočetním centrem zaveden nový způsob evidence softwaru, který vystřídal způsob sledování počtu současně spuštěných licencí s možností používání produktů s omezeným počtem licencí pouze na určitých učebnách. To je potřeba zohlednit při sestavování rozvrhu. V minulém semestru jsem jednak zapomněla dát takový požadavek na učebnu se systémem SPSS, jednak bylo z důvodu velkého počtu zájemců o předmět přidáno cvičení. Přesunutí výuky na potřebnou učebnu bylo z různých důvodů nerealizovatelné (hlavní roli hrál fakt, že počítačové učebny mají rozdílné kapacity). Při výuce byl tedy využíván především systém SAS Enterpise Guide (v dalším textu bude zkratka SAS používána výhradně pro tento produkt) a částečně též STATGRAPHICS.

Pro analýzy jsou používány různé *datové soubory*. Především to jsou datové soubory dodávané se systémem SPSS, jejichž součástí jsou popisy proměnných, popisy použitých kódů (číselníky) a specifikace kódů pro chybějící údaje. Dále je to soubor odpovědí řešitelů chemického korespondenčního semináře na několik jednoduchých otázek. Dotazník pro anketu připravili synové, kteří mi poskytli získané údaje. Soubor ve formátu SPSS je řádně popsán, ke kategoriálním proměnným existují číselníky. Na základě této tématiky jsem pro studenty připravila vzorové příklady k probírané látce, které s upravenými texty ze skript poskytuji studentům jako PDF soubory.

Zmíněné datové soubory lze v jiných souborech používat pouze omezeně, protože se převedou pouze data, nikoli popisy. Postupně jsem v systému SAS přidala popisy alespoň k proměnným souboru týkajícího se korespondenčního semináře. Ve stejném semestru mi jeden student kurzu nabídl data, která získal pro svou diplomovou práci od jedné cestovní kanceláře. Data byla pořízena na základě dotazníku, který nebyl profesionální a volba otázek a nabízených odpovědí nebyla v některých případech příliš šťastná. Také vložení do systému MS Excel nebylo provedeno způsobem odpovídajícím datovým souborům vytvořeným na základě profesionálních sociologických šetření. Přesto jsem začala tento soubor s oblibou používat, dokonce i v následujícím semestru, abych jednak poukázala na rozdíly mezi programovými systémy, jednak odůvodnila, proč je v případě kategoriálních dat je vhodnější vložit do tabulky pouze kódy a používat číselník.

Omezená možnost používání zavedených datových souborů vedla k tomu, že jsem ve větší míře začala používat zadání ve formě kontingenční tabulky. Tím se mohou studenti seznámit s rozdíly mezi systémem STATGRAPHICS, do jehož datového editoru lze tuto tabulku zadat přímo, a ostatními systémy (SAS, příp. SPSS), kde je třeba data vložit jako kombinaci kódů a jejich

četností. Zadání v podobě kontingenční tabulky používám například u souborů, které nemohu studentům poskytnout. Je to zejména datový soubor, který jsem na základě smlouvy získala z archivu Sociologického ústavu AV ČR. K tomu později přibyl datový soubor z průzkumu o uplatnění absolventů vysokých škol. Dále lze tímto způsobem vkládat různá data publikovaná v literatuře, například známé příklady na paradoxy poměru šancí, viz [1].

Doplňkem k výše uvedeným pomůckám je již dříve zmíněná interaktivní učebnice IASTAT, pomocí níž jsou ilustrovány zejména míry variability pro nominální a ordinální proměnné, které jsou základem při zkoumání asymetrické závislosti. Možnosti jsou ovšem omezené, výpočty lze provádět pouze pro maximálně pět kategorií.

4. Porovnání programových systémů a datových souborů

V předchozí části byly zmíněny programové systémy SPSS, SAS a STATGRAPHICS (dále SG), které lze využít pro analýzu kategoriálních dat. Do dalšího porovnání bude navíc zahrnut systém STATISTICA (dále ST) a částečně také systém S-PLUS (možnosti nabídkového režimu), tj. systémy s jednoduchým ovládáním vhodným pro výuku.

Hledisek pro porovnání programových systémů lze vymyslet velké množství. V tomto příspěvku budou vybrána pouze některá. Výše byl již například zmíněn ***způsob vstupu dat***. Programové systémy obvykle umožňují vstup zdrojových dat, tj. vstup datové matice, v níž řádky odpovídají statistickým jednotkám a sloupce statistickým znakům (proměnným). Není to však pravidlem.

Pokud jde *o jednorozměrnou analýzu*, pak k provedení *binomického testu* některé systémy požadují zadat již zjištěné četnosti. Systém SG vyžaduje relativní četnost sledované kategorie (předtím tedy musí být vytvořena tabulka četností jiným způsobem), pro S-PLUS jsou vstupními parametry absolutní četnost sledované kategorie a celkový rozsah výběru. Systém ST tímto testem zřejmě nedisponuje, pro soubory většího rozsahu může být pro použit *chí-kvadrát test dobré shody*. Pokud jde o chí-kvadrát test, pak v SG existuje pouze jeho aplikace na shodu s konkrétním pravděpodobnostním rozdělením, systém S-PLUS touto nabídkou také přímo nedisponuje. V systému ST je potřeba zadat do jedné „proměnné“ (sloupce) zjištěné četnosti a do druhé četnosti očekávané. To je však z hlediska vstupu dat v systému ST výjimka.

Obecně ST, stejně jako systémy SAS a SPSS, umožňují jak vstup zdrojových dat, tak *vstup četností*. Ve druhém případě se do jedné proměnné zadají varianty hodnot a do druhé četnosti odpovídající jednotlivým varian-

tám. V SPSS se pak najednou před všemi analýzami specifikuje, že proměnná s četnostmi obsahuje *váhy*. V systému SAS se obdobná specifikace musí provést před každou analýzou tím, že se proměnná obsahující četnosti definuje jako *frequency variable*. V systému ST je tato možnost obsažena v rámci každé analýzy s tím, že lze specifikovat, zda mají být váhy použity pouze pro danou analýzu, nebo i pro všechny další (bohužel tuto možnost nelze využít pro chí-kvadrát test).

V případě *dvourozměrné analýzy* lze v systémech SAS, SPSS a ST postupovat zcela stejným způsobem. Zadáváme tedy buď zdrojová data, nebo kombinace variant hodnot do dvou proměnných a do třetí četnosti odpovídající jednotlivým kombinacím. Systém ST navíc v případě čtyřpolních tabulek umožňuje zadávat přímo sdružené četnosti, a to do čtyř speciálních políček. V systémech SG a S-PLUS může datový editor obsahovat buď zdrojová data, nebo sdružené četnosti uspořádané do kontingenční tabulky. V systému SG je třeba zvolit vhodnou proceduru, v S-PLUS lze při analýze specifikovat, že datový editor obsahuje kontingenční tabulku.

Zdánlivě podružnější se může jevit *možnost popisu jednotlivých variant hodnot*, která je nejvíce propracována v SPSS. Tento systém je také určen ke zpracování dat z dotazníků, kde převažují kategoriální proměnné. Vytvoření datového souboru pomocí číselných kódů odpovědí a číselníků těchto kódů poskytuje řadu výhod. Jedna výhoda se týká vkládání a uchování dat, kdy jsou data snadněji kontrolovatelná a menší co do objemu. Druhá výhoda je při analýzách, zejména v případě ordinálních proměnných. Tabulky a grafy četností se v případě textových hodnot vytvářejí tak, že se kategorie uspořádají podle abecedy, což nemusí odpovídat jejich pořadí na ordinální škále. Z toho vyplývá potřeba použití číselných kódů. Bez odpovídajících číselníků však jsou výsledné tabulky a grafy obtížně interpretovatelné, analytik je musí v konečné fázi jednotlivě popsat. Číselníky v systému SPSS umožňují výsledné tabulky a grafy popisovat automaticky.

Dalším specifikem datových souborů vytvořených na základě dotazníků je *velký podíl chybějících údajů*. Pro proměnné obsahující číselné kódy (a čísla) umožňují některé systémy specifikovat kódy pro chybějící údaje. Výhodou systému SPSS je, že umožňuje specifikovat až tři takové kódy, případně celý interval, jehož hodnoty lze označit jako chybějící údaje. Je-li datový soubor tvořen textovými hodnotami a některý údaj chybí, pak v políčku není vložena žádná hodnota. V *tabulce četností* je pak v SPSS tato varianta uvedena jako první a její četnost se zahrnuje do výpočtu relativních a kumulativních četností. Číselné kódy jsou pro vytvoření tabulky četností přímo nutné. Při jejich použití a definování kódů pro chybějící údaje se v jednorozměrné ta-

bulce počítají dvě varianty relativních četností, jedna pro všechny hodnoty včetně chybějících údajů a druhá pouze pro tzv. platné hodnoty.

V systému SAS se v případě chybějících údajů nevkládá nic. Pro jednorozměrnou tabulku četností se řádky s takovými políčky nezahrnují do analýzy, pouze se pod tabulkou vypíše jejich počet. Jsou-li v proměnné vyjadřující určité aktivity pouze jedničky a prázdná políčka, pak výsledná tabulka četností obsahuje pouze jeden řádek a nelze přímo vyčíst, kolik procent respondentů na danou otázku odpovědělo kladně (řešením je nahradit prázdná políčka nulovou hodnotou). Obdobný výsledek pro data obsahující prázdná políčka získáme v systému SG s tím, že se nevypisuje počet chybějících údajů. Ten si tedy musíme zjistit odečtením zjištěného počtu platných hodnot od celkového rozsahu výběru (počtu řádků v tabulce).

Systém ST chybějící údaje bere do úvahy. Tabulku četností můžeme získat dvěma různými způsoby. Při způsobu, kdy systém sám navrhuje intervaly (bez ohledu na počet variant hodnot), se zobrazují dvě varianty četností stejně jako v SPSS, tj. pouze pro platné hodnoty a včetně chybějících údajů. Při zobrazení četností pro jednotlivé kategorie se při výpočtu relativních četností vychází z celého rozsahu výběru, tj. včetně chybějících údajů.

Pokud jde o porovnání systémů z hlediska samotné analýzy kategoriálních dat, pak se v tomto příspěvku zaměříme pouze na její základy. V oblasti ***jednorozměrné analýzy*** to jsou kromě tabulek a grafů četností již výše binomický test a chí-kvadrát test dobré shody. V oblasti dvourozměrné analýzy pak testy a míry závislosti pro kontingenční tabulky, případně některé další neparametrické testy.

Binomický test se v programových systémech vyskytuje ve třech implementacích, a to jako exaktní s využitím binomického rozdělení, jako asymptotický s využitím aproximace normovaným normálním rozdělením bez korekce a jako asymptotický s korekcí (při výpočtu testové statistiky se v čitateli přičítá hodnota 0,5). Podrobnější popis této problematiky lze nalézt v [6]. SAS zahrnuje první dvě možnosti s tím, že jejich výběr zcela závisí na uživateli. SPSS dle popisu algoritmů disponuje možností první a třetí, přičemž pro rozsah výběru do 25 včetně je na výstupu uvedeno, že je použito binomické rozdělení, a pro větší výběry se uvádí, že byla použita aproximace. Výsledky pro tyto větší výběry však odpovídají hodnotám distribuční funkce binomického rozdělení. Systémy SG a S-PLUS aplikují pouze exaktní test. Tyto systémy také umožňují, aby si uživatel zvolil jednu ze tří variant alternativní hypotézy. SAS zobrazuje výsledky pro relevantní jednostrannou a pro oboustrannou hypotézu, SPSS zobrazuje výsledky pouze jedné varianty dle kontextu, viz [6].

Chí-kvadrát test dobré shody může (stejně jako binomický test) v různých systémech vyžadovat různý vstup dat, viz výše. SAS umožňuje testovat pouze shodu s diskrétním rovnoměrným rozdělením (tj. shodu relativních četností pro všechny kategorie). SPSS má tuto možnost sice prioritně nastavenu, ale umožňuje též zadat seznam očekávaných četností. Systém ST vychází z četností zadaných do datového editoru, viz výše. SAS na rozdíl od ostatních systémů nabízí navíc exaktní variantu, jejíž výsledek je v případě dvou kategorií shodný s výsledkem exaktního binomického testu.

Zajímavé je také zařazení výše uvedených testů do nabídek systému. V SPSS jsou oba testy zařazeny do skupiny neparametrických testů. V systému SAS jsou testy nabízeny v rámci možností jednorozměrné tabulky četností. Binomický test je v systému SG zařazen k testování hypotéz pro jednu proměnnou (tedy do stejné skupiny jako parametrické testy) a v S-PLUS ve stejné skupině jako kontingenční tabulky, viz níže. Chí-kvadrát test v systému ST najdeme v nabídce *Neparametrická statistika*.

Kontingenční tabulky zahrnují vždy dvě základní oblasti, a to charakteristiky políček tabulky a charakteristiky závislostí dvou sledovaných kategoriálních proměnných. V prvním případě jde kromě sdružených a marginálních zjištěných absolutních četností též o různé varianty relativních četností (řádkové, sloupcové a na základě celé tabulky), o četnosti očekávané, rezidua a dílčí výpočty pro chí-kvadrát test o nezávislosti. I když možnosti jednotlivých uvažovaných systémů se i v této oblasti liší, nebudou zde podrobně rozvedeny, protože je jde pouze o pomocné nástroje pro sledování závislostí.

Ve druhém případě jde o různé testy a o výběrové míry závislosti. *Z testů* jde především o testy nezávislosti v kontingenční tabulce, případně o McNemarův test shody četností v políčkách na (vedlejší) diagonále ve čtyřpolní tabulce. Další testy se týkají testování nulovosti některých koeficientů závislosti (resp. logaritmu odhadu míry, jako v případě poměru šancí). Pokud jsou však tyto testy implementovány, jejich výsledky jsou zobrazovány spolu s příslušnými výběrovými koeficienty.

První dva typy testů systémy obvykle nabízejí odděleně od měr závislosti (výjimkou je systém ST). Základem u *testů nezávislosti* je Pearsonova statistika chí-kvadrát. Pro čyřpolní tabulku bývá navíc součástí výstupu statistika s Yatesovou korekcí a výsledek Fisherova exaktního testu (pro relevantní jednostrannou a pro oboustrannou alternativní hypotézu). Tyto možnosti zahrnují všechny zde uvažované programové systémy. Odlišnosti existují, ale nejsou rozsáhlé. S výjimkou S-PLUS je v systémech zahrnut věrohodnostní poměr, v systémech SAS a SPSS navíc Mantelova-Haenszelova statistika chí-kvadrát. SAS zobrazuje u Fisherova exaktního testu výsledky pro obě jednostranné alternativní hypotézy.

Test McNemarův není obsažen v systému SG. Protože tento test je v podstatě speciálním případem binomického testu pro shodu četností, rozdíly odpovídají rozdílům binomického testu. Můžeme tedy rozlišit exaktní test s využitím binomického rozdělení, asymptotický s využitím chí-kvadrát rozdělení bez korekce a asymptotický s korekcí (při výpočtu testové statistiky se v čitateli před umocnění odečítá hodnota 1). Podrobnější popis této problematiky lze nalézt v [6]. SAS zahrnuje první dvě možnosti s tím, že jejich výběr zcela závisí na uživateli. SPSS disponuje možností první a třetí, přičemž pro součet četností do 25 včetně je použito binomické rozdělení, a pro větší výběry aproximace. Tak je tomu ovšem pouze v případě implementace zařazené k neparametrickým testům. V rámci kontingečních tabulek je implementován McNemarův-Bowkerův test, který umožňuje porovnávat četnosti v políčkách označených vzájemně opačným pořadím indexů. McNemarův test je tedy speciálním případem pro čtyřpolní tabulku a vždy je prováděn jako exaktní. V S-PLUS je nabízena druhá a třetí varianta (standardně nastavená je možnost s korekcí, lze vypnout). V systému ST je použita třetí varianta, tj. asymptotický chí-kvadrát s korekcí. Zvláštností je, že se kromě shody četností v políčkách na vedlejší diagonále testuje také shoda četností v políčkách na hlavní diagonále. Jak je vidět, každý systém zaujímá jiný přístup.

Větší odlišnosti se vyskytují u *měr závislosti*. Nabídkový režim systému S-PLUS nenabízí žádné. U ostatních jsou samozřejmostí *míry založené na Pearsonově chí-kvadrát statistice*, jako Pearsonův kontingenční koeficient a Cramérovo V, příp. koeficient fí. V systému SAS jsou tyto koeficienty součástí výstupu týkajícího se výsledku chí-kvadrát testu, v systému SG jsou zařazeny k symetrickým mírám. V systémech SPSS a ST je třeba vybrat je z nabídky.

Dále můžeme míry rozlišit jednak podle typu proměnných, jednak podle toho, zda jde o závislost vzájemnou (symetrické míry), nebo jednostrannou (asymetrické míry). SAS a ST neuplatňují žádnou z těchto klasifikací. Lze pouze odlišit asymetrické míry, u nichž jsou v systému SAS uváděny symboly C|R, resp. R|C (závislost sloupcové proměnné na řádkové nebo řádkové proměnné na sloupcové) a v systému ST symboly X|Y, resp. Y|X. Systém SG organizuje výstup do dvou částí, přičemž první zahrnuje asymetrické míry (včetně jejich symetrických variant – pokud existují) a druhá míry symetrické (včetně měr založených na Pearsonově chí-kvadrát statistice). Nabídka SPSS je členěna podle typů proměnných, výstup pak primárně podle symetrických a asymetrických měr a v rámci těchto skupin pak podle typů proměnných. Jsou rozlišeny míry pro dvě nominální, dvě ordinální a dvě kvantitativní (intervalové) proměnné a míra pro závislost kvantitativní (intervalové) proměnné na nominální (odmocnina z koeficientu determinace počítaného při analýze rozptylu).

Z *asymetrických měr pro nominální proměnné* obsahují všechny čtyři systémy koeficient nejistoty. V systému ST je to jediná míra pro tento typ proměnných. Ostatní tři systémy obsahují ještě koeficient lambda a SPSS navíc koeficient tau (podrobněji viz [5]). Největší zastoupení je u *měr pro ordinální proměnné*. Všechny čtyři systémy zahrnují symetrické koeficienty gama, Kendallovo tau-b a tau-c a asymetrické Somersovo d. Kromě systému SG dále obsahují Spearmanův korelační koeficient. *Míra pro kvantitativní proměnné* je zastoupena jedna, a to Pearsonův korelační koeficient (v SPSS jsou korelační koeficienty nabízeny odděleně). Tento koeficient chybí v systému ST. Dále můžeme v systémech nalézt koeficient éta určující *míru jednostranné závislosti kvantitativní vysvětlované proměnné na proměnné nominální*. Je obsažen pouze v systémech SPSS a SG.

Systém SPSS a zvláště systém SAS poskytují ještě některé další analýzy. Jsou to především charakteristiky souhlasu ve čtvercových tabulkách. Kromě již výše uvedeného McNemarova testu sem patří koeficient souhlasu kappa a analýza poměru šancí ve čtyřpolní tabulce. Ta je velmi detailně implementována právě v systému SAS.

Jak již bylo uvedeno dříve, některé koeficienty lze *testovat na nulovost* (netýká se koeficientů určených pro kvantitativní proměnné, u nichž vzhledem k diskrétnímu charakteru není splněn předpoklad normality). SPSS uvádí výsledky testů pro všechny koeficienty určené pro nominální a ordinální proměnné automaticky, v systému SAS lze tuto možnost zvolit. Výsledky jsou uvedeny pouze pro koeficienty týkající se ordinálních proměnných a koeficientu kappa. Systém SG uvádí výsledky u koeficientů korelace, tj. Pearsonova a Kendallova tau-b, systém ST u korelačního koeficientu Spearmanova. SAS na rozdíl od ostatních systémů uvádí navíc dolní a horní meze intervalového odhadu.

Závěrem této problematiky si naznačme, pod jakými nabídkami se analýza kontingenčních tabulek skrývá. V systému SAS je to jedna z hlavních analýz pod názvem *Table Analysis*. V systému SPSS jde o dílčí analýzu v rámci popisných metod, tj. v části analýz se vybírá *Descriptive Statistics* a *Crosstabs*. Obdobné je to v systému SG, kde jde o nabídky *Describe*, *Categorical Data* a *Crosstabulation* (resp. *Contingency Tables* při zadávání již zjištěných sdružených četností). V systému ST v české verzi vybereme Základní statistiky/tabulky a v rámci nich *Kontingeční tabulky* (nabídka testů a měr závislosti je k dispozici až po specifikaci proměnných a přechodu do druhé fáze analýzy). Poněkud jinak je tomu v S-PLUS, kde v nabídce statistických metod je třeba zvolit *Compare Samples*, *Counts and Proportions* a *Chí-square Test*, případně jiný test. K dispozici jsou Fisherův exaktní a McNemarův test.

Další *neparametrické testy* použitelné pro kategoriální data a neuvedené výše jsou implementovány v systémech SPSS a ST. V oblasti dvourozměrné analýzy jde o testy pro dva či více nezávislých výběrů (sleduje se závislost ordinální proměnné na nominální) a testy pro dva závislé výběry (pro dvě ordinální proměnné, v SPSS zahrnut i McNemarův test).

Při obecném porovnání programových systémů by bylo dalším hlediskem organizace a formát výstupů, možnost jejich úpravy a převodu do systémů pro přípravu textových dokumentů a prezentací. Z hlediska výuky jde však o záležitost podružnou. Hraje sice důležitou úlohu při zpracování seminárních prací, ale ve vztahu k výše uvedeným faktorům je její význam menší.

Při výuce je kladen důraz především na získání výsledků a jejich interpretaci. Z tohoto hlediska lze drobné nedostatky vytknout systému ST v oblasti kontingenčních tabulek, kde například u koeficientu nejistoty jsou 3 varianty označeny jako X, Y a X|Y, kde poslední symbol je určen pro vzájemnou závislost, kdežto u Somersova d symboly X|Y a Y|X označují závislost jednostrannou. U výsledků testů je jeden ze sloupečků nadepsán „sv“ (stupně volnosti), ale obsahy políček zůstaly nepřeloženy, takže se vyskytuje např. „df = 1“. Dále u procedury určené pouze pro čtyřpolní tabulky je výsledek McNemarova testu označen názvem „Chí-kvadrát“ (tento název se tedy ve výstupu objevuje dvakrát, v prvním případě označuje Pearsonovu statistiku). Pod názvem „McNemarův chí-kvadrát“ se skrývá neobvyklá varianta testování shody četností v políčkách na hlavní diagonále.

5. Závěr

Pokud pedagog vyučuje již akreditovaný předmět, je omezen počtem hodin, případně rozdělením výuky na přednášky a cvičení. Obvykle je omezen také softwarovým vybavením. To však nemusí být trvalý stav, škola může získat finanční prostředky a vybavit počítačovou učebnu jiným systémem, či novější verzí stávajícího. I když pro základní výuku statistiky disponují široce zaměřené statistické systémy potřebnými metodami, v určitých oblastech se mohou implementace lišit, jak bylo uvedeno výše.

Z hlediska výuky je třeba se zaměřit na to, zda testy jsou prováděny jako exaktní, či jako aproximace, event. zda je při této aproximaci použita korekce, či nikoli. To se týká například binomického a McNemarova testu. Pokud jsou přednášky organizovány odděleně od cvičení, je vhodné výklad látky zaměřit primárně na možnosti používaného softwaru (a v rámci časových možností zmínit také možnosti jiné).

Pokud jde o výuku analýzy kategoriálních dat, tak z porovnávaných systémů je co do rozsahu metod je nejvíce vybaven systém SAS, v některých

směrech je ovšem omezen. Chí-kvadrát test dobré shody umožňuje testovat pouze shodu četností, nejsou používány korekce při aproximaci binomického rozdělení normálním a chí-kvadrát. Dále SAS neobsahuje všechny koeficienty závislosti, které jsou zahrnuty v SPSS, a neuvádí výsledky testů na nulovost těchto koeficientů v případě měr pro nominální proměnné. Na druhou stranu jsou součástí výstupu intervaly spolehlivosti.

Systém SPSS je zase uživatelsky příjemnější v oblasti práce s popisy kódů a s chybějícími údaji. Každý systém má své určité přednosti, ať už co se týká možnosti vstupu dat, či práce s výstupy. K výuce lze tedy použít různé systémy. Ideální jsou alespoň dva, aby si studenti uvědomili, v čem se mohou programové systémy lišit a co je potřeba zohlednit při interpretaci výsledků.

Reference

[1] Anděl, J.: Statistické modely. *Statistika,* 2003, č. 2, s. 1-17.

[2] Luha J.: Metódy štatistickej analýzy kvalitatívnych znakov. *EKOMSTAT´93.* SŠDS, Trenčianske Teplice 1993.

[3] Luha J.: Analýza nominálnych a ordinálnych znakov. *EKOMSTAT 2000.* SŠDS, Trenčianske Teplice 2000

[4] Řehák, J., Řeháková, B.: *Analýza kategorizovaných dat v sociologii.* Academia, Praha 1986.

[5] Řezanková, H.: *Analýza kategoriálních dat.* Oeconomica, Praha 2005.

[6] Řezanková, H.: Testy pro alternativní proměnné ve statistických programových systémech. *Forum Statisticum Slovacum,* 2005, č. 2, s. 114-118.

[7] Řezanková, H., Marek, L., Vrabec, M.: *IASTAT – Interaktivní učebnice statistiky.* http://iastat.vse.cz/.

[8] Stankovičová, I.: Ako robiť štatistiku v systéme SAS. *Výpočtová štatistika 2000,* SŠDS, Bratislava 2000, s. 74-78.

Adresa: doc. Ing. Hana Řezanková, CSc.
Katedra statistiky a pravděpodobnosti, Vysoká škola ekonomická v Praze, Nám. W. Churchilla 4, 130 67 Praha 3

E-mail: `rezanka@vse.cz`

VYDALO SE...

POČÍTAČOVÉ ZPRACOVÁNÍ DAT V PROGRAMU MATLAB: 1. ČÁST

Autor knihy: Martin Kovářík

Knihu recenzovali: Pavel Stříž, Petr Klímek, Petr Nevřiva a Lubor Homolka

Vydalo nakladatelství Martin Stříž, http://striz.cz/

Tato kniha s 278 stranami podporuje v podstatě výuku tří základních kurzů matematik na ekonomických vysokých školách bez určitých partií (optimalizace, rozvoj v Taylorovu řadu).

V první části dává přehled o komerčním a open-source software. Podává základní popis jednotlivých matematických, statistických a vizualizačních nástrojů. Kniha pracuje s Matlabem od nuly. Zmiňuje jeho výhody a nevýhody, instalaci a první krůčky. Postupně se dostává k práci s konstantami, proměnnými až k práci s funkcemi. Kniha zmiňuje i základní postupy a programovací techniky. Velkou část knihy tvoří 2D a 3D grafika, včetně tvorby animovaných sekvencí. Závěr knihy tvoří řešené a neřešené příklady a užitá a doporučená literatura.

Kniha se bude výborně hodit do nově vznikajícího kurzu „Statistická výpočetní prostředí“, kde se touto knihou kurz začne. Uplatnění však najde u odborné i zaujaté laické veřejnosti.

APLIKOVANÁ STATISTIKA: PŘEDNÁŠKY

Autor skript: Petr Klímek

Skripta recenzoval: Pavel Stříž

Vydala Univerzita Tomáše Bati ve Zlíně, http://utb.cz/

Tato skripta s 202 stranami shrnují poznámky a komplexně nahlíží na aplikovanou statistiku z pohledu posluchače vysoké školy ekonomické fakulty.

Skripta najdou uplatnění v navazujících statistických kurzech (Aplikovaná statistika, někde pojmenováno jako Statistika 2 nebo Statistika B) na ekonomických fakultách vysokých škol. Skripta lze doporučit odborné i zaujaté laické veřejnosti.

OBSAH BULLETINU

Gejza Wimmer Matematické modelovanie v jazykovede 1
Hana Řezanková Výuka jednorozměrné a dvourozměrné analýzy kategoriálních dat 18
Vydalo se 31

Navštívili jste nebo plánujete konferenci a chcete o ní informovat odbornou veřejnost?

Informační Bulletin k tomu dává ideální příležitost!

Chcete zveřejnit zajímavé netisknutelné materiály či záznamy videopřednášek?

ČStS dostupný FTP server k tomu dává možnost!

Vychází vám skripta, učebnice, kniha nebo jste někomu dopsali recenzi?

I my, členové ČStS, se o tom rádi dozvíme!

ISSN 1210 – 8022. Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo.

Předseda společnosti: Doc. RNDr. Gejza Dohnal, CSc., ÚTM FS ČVUT v Praze, Karlovo náměstí 13, 121 35 Praha 2, e-mail: gejza.dohnal@fs.cvut.cz

Ediční rada: Prof. Ing. Václav Čermák, DrSc. (předseda), Prof. RNDr. Jaromír Antoch, CSc., Doc. Ing. Josef Tvrdík, CSc., RNDr. Marek Malý, CSc., Doc. RNDr. Jiří Michálek, CSc., Doc. RNDr. Zdeněk Karpíšek, CSc. a Prof. Ing. Jiří Militký, CSc.

Techničtí redaktoři: Doc. RNDr. Gejza Dohnal, CSc., gejza.dohnal@fs.cvut.cz a Ing. Pavel Stříž, Ph.D., striz@fame.utb.cz

Pokyny autorům: <http://www.statspol.cz/bulletiny/sablony.htm>

FTP: exp.uis.fame.utb.cz; uživatel: csts; heslo: csts

WEB server: <http://www.statspol.cz/>