

Informační Bulletin



České Statistické Společnosti

číslo 2., květen 1993, ročník 4.

Vlastenecké muzeum v Čechách a rozvoj naší statistické vědy

Prokop Závodský

15. dubna vzpomněla česká kulturní veřejnost 175. výročí založení *Wlasteneckého Muzeum w Čechách* (dnešního Národního muzea) r. 1818. Méně je známo, že již v prvních ročnících muzejních časopisů byly otištěny významné příspěvky k rozvoji politické aritmetiky, předchůdkyně dnešní statistické vědy.

Poměry na pražské universitě v předbreznové době nikterak nepovzbuzovaly rozvoj originálního vědeckého myšlení, vědecká práce profesorů nebyla nutná, nýbrž spíše na obtíž. Jako povinné učebnice byly stejně zpravidla předepisovány práce vídeňských autorů, nezdědka desítky let staré. Profesor statistiky G.N. Schnabel (1791 – 1857), tehdy jedna z nejvýznamnějších osobností právnické fakulty, byl autorem dlouhé řady pozoruhodných statistických publikací. Přesto byl nucen každoročně přednášet podle zastaralých oficiálních učebnic vídeňských profesorů Zizia a Bisingerů.

České muzeum, založené skupinou šlechtických nadšenců jako zemská vědecká instituce, vydávalo od r. 1827 německý a český *Muzejník* (iniciátorem a dlouholetým redaktorem byl F. Palacký) a částečně suplovalo funkci university jako organizátora a podporovatele rozvoje vědy (často bývá citováno příznivé hodnocení obou časopisů J.W.Goethem v berlínském kritickém časopise). Reprezentačním orgánem Muzejní společnosti byl zprvu zejména německý *Monatschrift der Gesellschaft des vaterländischen Museums in Böhmen*, ale svou vědeckou úroveň rychle zvyšoval i český časopis (od 5. ročníku pod názvem *Časopis Českého Museum*), zatímco německý *Měsíčník* (později jen čtvrtletník), vedený ideou zemského patriotismu (ustupujícího novodobému německému a českému nacionalismu) přestal koncem r. 1831 vůbec vycházet. V prvních ročnících *Muzejníků* nacházíme recenze a zprávy o novinkách statistické literatury (od prof. Schnabela), pravidelná hlášení o meteorologických pozorováních v Praze, zpracovaných elementárními statistickými metodami (z pera prof. Hallaschky) i příspěvky k hospodářské statistice Čech (autory jsou guberniální rada K.A.Neumann, známý přírodovědec hrabě Kašpar ze Sternberka, prof. Schnabel i další).

Typeset by \LaTeX

Dva pozoruhodné rozborů statistických údajů o obyvatelstvu Čech uveřejnil v německém *Muzejníku* i F. Palacký (*Gradation der Bevölkerung Böhmens seit den letzten 60 Jahren, Statistisch-topographische Notizen über die Bevölkerung Böhmens im J. 1830*). Leckterého čtenáře zde překvapí autorova znalost soudobé statistické literatury i jednodušších metod politické aritmetiky (Palacký píše *politische Rechenkunst*).

Teoretickými otázkami politické aritmetiky se u nás poprvé zabýval mladý absolvent pražské a vídeňské university Karl Czoernig (1804-1889). V německém *Muzejníku* uveřejnil r. 1831 komentovaný překlad stati z anglického časopisu doplněný vlastním obsáhlým úvodem.

Zajímavý je již Czoernigův pokus o vymezení politické aritmetiky jako vědní disciplíny: „Politická aritmetika se zabývá zobrazením těch jevů v životě obyvatel státu, které lze vyjádřit kvantitativními vztahy . . .“ Zařazuje pak politickou aritmetiku mezi „státní vědy“ vedle její „starší sestry“ – statistiky. Czoernig správně rozpoznal i některé limity rozvoje politické aritmetiky jako vědního oboru – v době před její integrací do moderní statistické vědy, k níž došlo v příštích desetiletích. Poukazuje zde nejen na nedostatek přesných a spolehlivých údajů (pokud byly zjištěny, často se utajují), ale především na neujasněnou koncepci této disciplíny, rozvíjející se převážně mimo universitní katedry - zásluhou amatérů a poradců pojišťovacích společností.

Czoernig, který použil pozoruhodné metody analýzy dat již ve své starší práci o Liberci (očišťování časové řady od kalendářních variací, sezónní indexy), zde dále vysvětluje princip konstrukce úmrtnostních tabulek i výpočtu některých odvozených charakteristik (pravděpodobnost dožití určitého věku, pravděpodobná délka života, střední délka života) a pojednává o metodách, které ke konstrukci úmrtnostních tabulek použili západoevropští političtí aritmetici (W. Kersseboom, A. Déparcieux, R. Price, J. Milne). V Czoernigově práci je též (patrně poprvé u nás) naznačena i možnost zkoumání závislosti mezi numerickými proměnnými (v citaci šetření francouzského statistika L.R. Villermé o závislosti úmrtnosti v jednotlivých pařížských obvodech na podílu chudiny).

Za připomenutí i dnes snad stojí skutečnost, že (na rozdíl od jiných státních úřadů) za šéfa státní statistické služby jmenoval císařský dvůr r. 1841 neurozeného, leč na rakouské poměry vysoce kvalifikovaného K. Czoerniga. Rakouskou statistiku řídil téměř 25 let. (Od r. 1852 jako svob. pán z Czernhausen – podle rodné obce Černousy u Frýdlantu. Jako devizu do erbu si vybral heslo F. Bacona *Scientia est potentia*.)

Plodným autorem v oboru politické aritmetiky byl ve 20. a 30. letech i staroměstský úřední lékař Franz Alois Stelzig (1784 – 1856), absolvent pražské lékařské fakulty. Jmenujme zde alespoň dvoudílnou publikaci *Versuch einer medizinischen Topographie von Prag* (1824), první a z hlediska statistiky nejzajímavější z řady tehdy módních medicínských topografií u nás, a velmi rozsáhlou stať v *Muzejníku*, věnovanou výhradně analýze demografických údajů (*Resultate der Geburts – und Sterbeverhältnisse – 1830*).

Výsledkem stati je sestavení prvních úmrtnostních tabulek pro Čechy (za období 1800 – 1828), které autor konfrontuje především s tabulkami slavného pruského politického aritmetika XVIII. století J. P. Süßmilcha. Obsáhle též polemizuje se Süßmilchovým pře-

svědčením o stálosti objevených demografických zákonitostí v čase i prostoru („božský řád v proměnách lidského rodu“) poukazováním na důsledky pokroku medicíny (mj. očkování, různá hygienická opatření) i na rozdílné reprodukční poměry v různých oblastech (velkoměsto v. venkov), u různých skupin obyvatelstva apod.

Stelzig vysvětluje i princip různých druhů životního pojištění, doživotních rent, tontin, vdovských, sirotčích a dalších zaopatřovacích institucí a pojednává o užití některých ukazatelů úmrtnostních tabulek i dalších statistických metod pro činnost těchto institucí.

Váženým kolegům, zajímajícím se o starší dějiny naší statistiky, mohu poskytnout ještě několik autorských výtisků své podrobné studie na toto téma. Přihlaste se, prosím, na katedře statistiky a pravděpodobnosti VŠE (tel. 21 25 131).

Korelace, která nás šatí i obouvá

Stanislav Komenda

Nejenom statistikům je známo, že vztahy a souvislosti existující mezi jevy a veličinami poskytují možnost ze známého stavu jedné veličiny (jedněch veličin) usuzovat na stav veličiny jiné. Statistikům je navíc známo, jak tuto možnost učinit tělem — tj., jak tento úsudek materializovat návodem, vzorcem, výpočetním algoritmem. Podstatnou část příslušných metod shrnují učebnice v kapitole „Regrese a regresní analýza“.

Statistické metody přitom respektují fakt, že výpověď jedněch empiricky měřených veličin o jiných veličinách nikdy není dokonalá; prakticky vždy ponechává taková výpověď určitý prostor nejednoznačnosti, nejistotě. Přenos informace má charakter kanálu se šumem.

Využívání vztahů existujících mezi veličinami se neomezuje na tok a předávání informace. Aplikací je nepřeberně — a některé z nich mohou mít výrazný ekonomický dopad. Příkladem je antropometrická standardizace v přípravě hromadné výroby produktů, jejichž rozměry musejí respektovat tělesné rozměry potencionálního uživatele. Zvlášť důležitými situacemi tohoto typu jsou oděvní konference a průmyslová výroba obuvi.

Individuální zakázková výroba oděvu krejčím a bot ševcem nepochybně také respektovala jisté vztahy a zákonitosti mezi tělesnými rozměry, jak byly tyto staletou zkušeností rozpoznány a registrovány. Podstatné ovšem je, že krejčí i švec si potřebné tělesné rozměry zákazníka mohli přeměřit a ještě v jistém stádiu rozpracovanosti výrobku ověřit jeho padnutí a provést žádoucí korekce.

Ze zcela odlišné situace musí vycházet projekt výrobku v případě masové konfekční výroby — u nás reprezentovaná už před válkou prostějovským Rolným nebo Neherou či zlínským Baťou. Konfekce totiž vyrábí pro anonymního zákazníka. Není předem známo, kdo si právě šité šaty obleče či boty obuje. Součástí přípravy výroby musí proto být racionální úvaha o tom, jaké „velikosti“ vyrábět, aby se jimi výrobce „trefil“ do prostoru velikostních typů reálně existujících postav v populaci potenciálních zákazníků. A právě v této části hry vstupuje na scénu korelace — bez jejíž existence by byla ona baťovská či neherovská úloha neřešitelná.

Pro zhotovení výrobku toho typu, o němž uvažujeme, je třeba brát v úvahu 10 – 20 (v případě šatů, sak, kalhot, sukní a kabátů), resp. 5 – 7 (v případě bot) tělesných rozměrů. Zkušenost říká, že přijatelné padnutí výrobku může být zaručeno jenom v jistém rozmezí tolerancí každého z těchto rozměrů. Tyto tolerance činí například asi 6 cm u výšky těla, 4 - 6 cm u obvodu hrudníku či pasu atp. Kdyby sestavovala příroda (genetika, anatomie, fyzická antropologie) naše postavy metodou libovolných kombinací těchto rozměrových tolerancí, dostali bychom už při pouhých pěti tělesných rozměrech řádově tisíce typových postav. Kdyby tomu tak opravdu bylo, žádná hromadná, konfekční výroba by byla nikdy nevznikla.

Za to, že vzniknout mohla, vděčí korelaci. Přesněji řečeno — nenulové korelaci. Měl jsem kdysi možnost s panem Tomášem Baťou mluvit — když dostal v roce 1990 čestnou medaili Palackého univerzity a když mu o rok později udělila VŠE v Praze čestný doktorát — nějak však tehdy na to nedošlo, abych ho přesvědčil, že by měla rodina Baťů zahrnout symbol korelačního koeficientu do svého rodového štítu.

Právě existující (a dostatečně vysoká) korelace tělesných rozměrů způsobuje, že množina typových postav, jimiž je možno pokrýt — při daných tolerancích jednotlivých rozměrů — populaci v úvahu přicházejících zákazníků, má podstatně méně prvků. Většina rozměrových kombinací se v reálných populacích nevyskytuje anebo vyskytuje s četností zanedbatelně malou. Výpočty, respektující tolerance padnutí rozměrů a jejich populační variabilitu ukazují, že hromadná výroba konfekce vystačí uspokojivě s několika desítkami velikostních typů; počet typových nohou je ještě nižší. Podobné výpočty musí ovšem respektovat rovněž pohlaví a věk zákazníka.

Antropometrické analýzy tělesných rozměrů člověka prokazují, že délkové rozměry (délka horních a dolních končetin a jejich segmentů, délka trupu atp.) silně a kladně korelují s výškou těla, zatímco šířkové, hloubkové a obvodové rozměry mají vysokou kladnou korelaci s obvodem hrudníku, případně s obvodem pasu. Regrese uvažující výšku a obvod hrudníku jako regresory (nezávislé veličiny) a další rozměry potřebné ke zhotovení oděvu jako regresandy (závislé proměnné) ukazuje, že stačí definovat typové postavy kombinací hodnot těchto dvou základních rozměrů — protože reziduální rozptyly dalších konstrukčně důležitých rozměrů (jako třeba délka horní končetiny, obvod stehna či lýtka apod.) jsou už dostatečně nízké, aby interval dvou či čtyř reziduálních směrodatných odchylek kolem hodnoty regresní funkce nevybočil podstatně z tolerance padnutí. O hromadné výrobě obuvi platí totéž — s tím, že počet rozměrů v konstrukci výrobku uvažovaných je ve srovnání se součástmi oděvu menší.

Na tomto místě se sluší přičinit poznámku. Uvažujice o reziduálním rozptylu, máme na mysli jeho hodnotu v jistém bodě (bylo-li vzato jako model vícerozměrné normální rozdělení, nezáleží na tom, ve kterém) nezávisle proměnných (regresorů). Řešení úlohy, která je zde načrtnuta, však vyžaduje, abychom se zajímali o reziduální rozptyl regresandu nikoli v bodě, ale v jistém intervalu hodnot regresorů. Tento rozptyl bude zřejmě vyšší. Dá se spočítat, nebo alespoň odhadnout, o kolik. Před několika lety jsem byl konfrontován s jedním neobvyklým systémem zhotovení oděvů: jistý varšavský krejčí, pan A. Elert (prý krejčí pro lepší lidi), publikoval v polském odborném časopise článek, v němž vysvětloval, že kterýkoli rozměr lidského těla se dá vyjádřit jako jistý násobek poloměru hypotetické koule téže hmotnosti a hustoty, jaké má lidské tělo (ta hustota činí asi $1,1 \text{ g} \times \text{cm}^{-3}$). Z tohoto faktu pak odvozoval systém konstrukce oděvu; stačilo mu znát výšku a hmotnost jedince.

Měl jsem na článek napsat recenzi — a tak se mi po jisté době, během níž mi tento projekt připadal velice bizarní, došlo, že vlastně jde o to, přezkoušet informační bohatost hmotnosti těla jako regresoru. A ejhle, ono se kupodivu ukázalo, že hmotnost dokáže hloubkové, šířkové a obvodové rozměry těla predikovat velice spolehlivě! Reziduální rozptyly obvodových konstrukčních rozměrů (obvodu krku, obvodu stehna, obvodu lýtka atp.) se při známé hmotnosti zmenšovaly na zlomek rozptylů nepodmíněných. Délkové rozměry však dokázala hmotnost predikovat jen bídňě. Jestliže se však jako regresorů použilo společně výšky a hmotnosti těla, byla jejich predikční schopnost srovnatelná s predikční schopností trojice regresorů — výška, obvod hrudníku a obvod pasu (sedu). Takže — šít konfekční oděvy na základě hmotnosti (měřené vážením) je počínání docela rozumné.

Pokud jde o ony Elertovy kánony (podle nichž je každý tělesný rozměr konstantním násobkem onoho poloměru koule odvozené z tělesné hmotnosti), jde o představu idealizovanou a naivní, protože nestatistickou. Z antropometrických dat se snadno zjistí, že v subpopulaci jedinců téže hmotnosti (daného pohlaví a stejné věkové kategorie) má každý tělesný rozměr svou variabilitu. Jak už bylo řečeno, obvodové rozměry jsou v takové subpopulaci proměnlivé relativně málo, délkové naopak relativně velmi. Představa takových kánonů je prakticky uplatnitelná jenom v omezené centrální podmnožině prostoru tělesných rozměrů - a i tak pouze s jistým přiblížením.

Jen tak mimochodem, určité populační konstanty se na lidském těle najít dají. Tak například index počítaný jako podíl obvodu hlavice femoru a jejího diametru (průměru) definuje číslo π — protože tato hlavice je s dostatečným přiblížením kulová.

Abych doložil a podepřel svoje úvahy alespoň trochu kvantitativně, uvádím pro několik oděvářsky zajímavých tělesných rozměrů (každý z nich je jako regresand označován symbolem W), jejich korelace s regresory M (hmotnost těla) a X (výška těla). Souběžně jsou uvedeny reziduální rozptyly těchto regresandů pro různé systémy regresorů: (M) , (X, M) , (X, Y, Z) , kde Y je obvod hrudníku a Z je obvod pasu (u mužů) či obvod sedu (u žen). Omezujeme se přitom na populaci dospělých mužů a žen; řešení úlohy v populacích dětí a mládeže má některé specifické rysy.

Jak je patrné, korelace konstrukčních oděvních rozměrů s výškou těla X a hmotností těla M se vzájemně komplementárně doplňují — každý rozměr je „informačně syčen“ buď pře-

vážně výškou nebo převážně hmotností — a jenom výjimečně oběma. Informační účinnost obou systémů (X, M) a (X, Y, Z) je pak pro prakticky všechny rozměry W srovnatelná.

MUŽI

| Rozměr W | r_{WM} | r_{WX} | s_{WW} | $s_{W.M}$ | $s_{W.XM}$ | $s_{W.XYZ}$ |
|--------------------------|----------|----------|----------|-----------|------------|-------------|
| Obvod krku | 0,70 | 0,20 | 4,76 | 2,41 | 2,31 | 2,09 |
| Obvod stehna | 0,79 | 0,29 | 17,65 | 6,72 | 6,59 | 7,56 |
| Obvod nadloktí | 0,78 | 0,13 | 7,91 | 3,08 | 2,56 | 2,23 |
| Výška pasu | 0,38 | 0,91 | 26,87 | 23,05 | 4,75 | 3,50 |
| Výška kolena | 0,36 | 0,76 | 7,50 | 6,54 | 3,13 | 2,87 |
| Výška 7. krčního obratle | 0,50 | 0,97 | 36,04 | 27,21 | 2,21 | 0,34 |
| Výška podhýžďové rýhy | 0,33 | 0,84 | 18,09 | 16,06 | 5,30 | 4,69 |
| Výška sedu | 0,38 | 0,80 | 21,26 | 18,26 | 7,50 | 6,76 |
| Čelní šířka ramen | 0,51 | 0,38 | 4,93 | 3,67 | 3,53 | 2,94 |
| Čelní šířka sedu | 0,68 | 0,45 | 4,64 | 2,51 | 2,39 | 1,80 |
| Profilová šířka hrudníku | 0,77 | 0,18 | 4,69 | 1,91 | 1,72 | 1,19 |

ŽENY

| Rozměr W | r_{WM} | r_{WX} | s_{WW} | $s_{W.M}$ | $s_{W.XM}$ | $s_{W.XYZ}$ |
|--------------------------|----------|----------|----------|-----------|------------|-------------|
| Obvod krku | 0,70 | 0,11 | 5,45 | 2,81 | 2,78 | 2,81 |
| Obvod stehna | 0,82 | 0,15 | 29,10 | 9,60 | 9,50 | 7,79 |
| Obvod nadloktí | 0,86 | -0,13 | 11,57 | 2,99 | 2,26 | 2,43 |
| Výška pasu | 0,34 | 0,88 | 22,52 | 19,89 | 4,72 | 4,65 |
| Výška kolena | 0,21 | 0,71 | 7,73 | 7,39 | 3,82 | 3,81 |
| Výška 7. krčního obratle | 0,31 | 0,97 | 31,98 | 28,82 | 1,99 | 1,98 |
| Výška hýžďové rýhy | 0,15 | 0,86 | 16,93 | 16,55 | 4,29 | 4,21 |
| Výška sedu | 0,25 | 0,76 | 21,55 | 20,19 | 9,03 | 8,85 |
| Čelní šířka ramen | 0,52 | 0,41 | 2,92 | 2,12 | 1,88 | 1,87 |
| Čelní šířka sedu | 0,80 | 0,22 | 7,75 | 2,75 | 2,75 | 1,71 |
| Profilová šířka hrudníku | 0,82 | 0,04 | 8,70 | 2,82 | 2,54 | 1,66 |

Tabulka

Korelace, rozptyly a reziduální rozptyly srovnatelných antropometrických systémů. X = výška těla, M = hmotnost těla, Z = obvod pasu (resp. sedu). Rozptyly jsou v cm^2 .

Jak závisí výsledek sportovního utkání na jeho počáteční fázi

Jiří Anděl

Najdete-li mezi autory nějakého článku jméno Fredericka Mostellera, můžete si být jisti, že jde o zajímavé čtení. To potvrzuje i článek Cooper H., DeNeve K. M., Mosteller F.: Predicting professional sports game outcomes from intermediate game scores, který byl publikován v časopise *Chance*, Vol. 5 (1992) v čísle 3–4 na str. 18–22.

Víme z vlastní zkušenosti, že fanouškové mnohdy opouštějí stadion ještě před ukončením utkání, pokud mají za to, že o výsledku je už rozhodnuto. (Někteří sportovní činovníci tvrdí, že v poslední době mnozí fanouškové na stadion vůbec nepřijdou.) Méně je známo, že (prý) v USA řada diváků přichází na basket až ke konci, protože se traduje, že teprve poslední čtvrt hodina hry (či dokonce poslední dvě minuty hry) přináší zvrát ve vývoji skóre a je nejdramatičtější. Co tomu říká statistika?

Bylo sledováno 200 utkání v basketu, 100 v baseballu, 100 v hokeji a 100 ve fotbalu. Tato utkání byla zvolena tak, aby je z praktického hlediska bylo možno pokládat za náhodný výběr z her odehraných v sezóně 1991/92. Pokud jde o fotbal, domnívám se, že se to týká amerického fotbalu. Definujme nyní termíny „na začátku“ a „před koncem“ takto:

| hra | na začátku | před koncem |
|----------|------------------|-------------------|
| basket | po 1/4 hry | po 3/4 hry |
| baseball | po 3 směnách | po 7 směnách |
| hokej | po první třetině | po dvou třetinách |
| fotbal | po 1/4 hry | po 3/4 hry |

U basketu, hokeje a fotbalu se ukázalo, že asi ve 20% případů dokázalo mužstvo, které před koncem prohrávalo, ještě průběh hry otočit a nakonec vyhrát. U baseballu to bylo jen v 6% případů.

Zato souvislost mezi začátkem hry a konečným výsledkem je trochu pestřejší. Mužstvo, které na začátku prohrávalo, dokázalo nakonec vyhrát ve 30% případů v basketu, v 19% případů v baseballu, v 31% případů v hokeji a v plných 45% případů ve fotbale.

Ukázalo se, že to je hlavně domácí mužstvo, které častěji dokáže zvrátit nepříznivý stav hry před koncem utkání, alespoň pokud jde o basket. Domácím se to totiž podařilo ve 33% případů, kdežto hostům jen v 10%. U ostatních sledovaných sportů už k tak výrazným rozdílům nedochází. Rozdíl je ještě trochu patrný u hokeje, zatímco u fotbalu jsou šance na zvrát stejné pro domácí jako pro hosty. Baseball nebyl z tohoto hlediska hodnocen vzhledem k malému počtu takovýchto „zvrátových“ utkání.

Který model je ten pravý aneb vyberte si koeficient determinace.

Karel Zvára

K popisu skutečnosti můžeme navrhnout dva v mnoha ohledech ekvivalentní modely (stejný reziduální součet čtverců, jednoznačný vztah mezi bodovými odhady parametrů i jejich standardními chybami), avšak s různými hodnotami koeficientu determinace.

1. Úvod

Uvažujeme-li běžný normální lineární model $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2)$ s odhadem parametru $\boldsymbol{\mu}$ metodou nejmenších čtverců tvaru $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, pak je reziduální součet čtverců RSS dán čtvercem délky vektoru $\mathbf{y} - \hat{\mathbf{y}}$, tedy výrazem $RSS = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Když popíšeme velikost variability hodnot závisle proměnné pomocí $\|\mathbf{y} - \bar{y}\mathbf{1}\|^2 = \sum_{i=1}^n (y_i - \bar{y})^2$, vyjadřuje reziduální součet čtverců tu část výchozí variability závisle proměnné, kterou se pomocí regresní závislosti nepodařilo vysvětlit. Koeficient determinace R^2 udává, jaký díl výchozí variability jsme uvažovanou závislostí naopak vysvětlili:

$$R^2 = 1 - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Na rozdíl od reziduálního součtu čtverců je koeficient determinace bezrozměrnou veličinou. V této souvislosti připomeňme, že uvedená interpretace koeficientu determinace má smysl pouze v případě, kdy uvažovaná závislost obsahuje (aspoň jako lineární kombinaci sloupců matice \mathbf{X}) absolutní člen, což budeme dále bez výjimky předpokládat.

Připomeňme si možné úpravy modelu, které právě zavedené charakteristiky neovlivní. Reziduální součet čtverců závisí pouze na vzdálenosti vektorů \mathbf{y} a $\hat{\mathbf{y}}$. Zůstane tedy beze změny, když k oběma vektorům přičteme stejný pevný vektor nebo když místo matice \mathbf{X} použijeme matici $\mathbf{X}^* = \mathbf{X}\mathbf{D}$, kde \mathbf{D} je regulární matice. Vektor $\hat{\mathbf{y}}$ je totiž průmětem vektoru \mathbf{y} do lineárního obalu sloupců matice \mathbf{X} , který je ovšem shodný s lineárním obalem matice \mathbf{X}^* . Speciálně tedy nezáleží na volbě měřítka u regresorů odpovídajících jednotlivým sloupcům matice \mathbf{X} . Změna měřítka u závisle proměnné hodnotu reziduálního součtu čtverců ovlivní, ovšem u koeficientu determinace se druhá mocnina transformační konstanty vykrátí, takže zde i tato operace zůstane bez následků.

2. Příklady

Ve svém článku se Radek a Partyková zabývají závislostí výnosu brambor na měsíčních údajích o průměrné teplotě a úhrnných vodních srážkách. I když nechceme jejich práci nějak vylepšovat, zmíníme se o pokusu vysvětlit variabilitu ve výnosech pomocí pouhých srážek. Při hledání vhodného modelu krokovou regresí z údajů za 19 sezón si ponechal Statgraphics z 12 nabídnutých regresorů pouze dva, totiž srážky v září a říjnu (tedy v předvegetačním období).

Z našeho hlediska je zajímavé, že u predikce výnosů pomocí

$$\hat{v} = 21,0102 + 0,0595s_9 - 0,0585s_{10}$$

nebo

$$\hat{v} = 21,0102 - 0,0585(s_{10} - s_9) + 0,0010s_9$$

je v obou případech $RSS = 4,3536$ a $R^2 = 57,8\%$. Snad je tento příklad dobrou ilustrací toho, že pouhá změna báze prostoru, kam promítáme \mathbf{y} , abychom dostali $\hat{\mathbf{y}}$, hodnotu dvou našich statistik nezmění. Stranou nechme zajímavou skutečnost, že regresní koeficient u s_9 ve druhém vztahu má standardní chybu 0,0358 a je tedy statisticky zcela nevýznamný.

Pokusme se nyní ze stejných dat předpovídat říjnové srážky pomocí srážek za září. Metoda nejmenších čtverců dá odhad

$$\widehat{s_{10}} = 59,0833 - 0,3915s_9$$

se standardní chybou 0,2057 u směrnice, reziduálním součtem čtverců $RSS = 529,52$ a koeficientem determinace $R^2 = 17,6\%$. Daří se tedy vysvětlit tak malý díl variability říjnových srážek, že na 5% hladině nemůžeme při oboustranné alternativě zamítnout hypotézu o nulové směrnici. Co kdybychom však chtěli pomocí zářijových srážek vysvětlit hodnotu $s_{10} - s_9$, tedy oč jsou říjnové srážky větší než srážky zářijové (když právě na tomto rozdílu tolik záleží vlastní výnos)? Místo vztahu $s_{10} = \alpha + \beta s_9$ tedy uvažujeme vztah $s_{10} - s_9 = \alpha + (\beta - 1)s_9$. Geometricky to vlastně znamená, že ve vztahu $\mathbf{y} = \alpha\mathbf{1} + \mathbf{x}\beta + \mathbf{e}$ odečítáme od obou stran rovnice pevný vektor zářijových srážek. To znamená, že se právě o tento vektor změni oba vektory \mathbf{y} i $\hat{\mathbf{y}}$ a reziduální součet čtverců zůstane nezměněn. Nezmění se ani reziduální rozptyl a protože délka vektoru u regresního koeficientu $\beta^* = \beta - 1$ se rovněž nezměnila, musí být standardní chyba odhadu tohoto koeficientu stejná jako u koeficientu β . Skutečně, citovaná data dají

$$\widehat{y_{10}} - y_9 = 59,0833 - 1,3915s_9$$

se standardní chybou 0,2057, reziduálním součtem čtverců $RSS = 529,52$, ale s koeficientem determinace $R^2 = 72,9\%$. Regresní koeficient je tentokrát významně nenulový, ať už zvolíme hladinu jakkoliv rozumně. Koeficient determinace se tedy dramaticky změnil. Je to tak pokaždé?

Zkusíme pracovat s jinými daty. V jistém lékařském výzkumu byla sledována delší dobu skupina chlapců. Pro nás podstatnou je skutečnost, že máme k dispozici jejich výšky v 10 a ve 12 letech věku. Předpověď výšky ve 12 letech na základě známé výšky v 10 letech dala

$$\hat{v}_{12} = 43,0441 + 0,7684v_{10}$$

se standardní chybou 0,2729 u směrnice, reziduální součtem čtverců $RSS = 143,2542$ a koeficientem determinace $R^2 = 53,1\%$. Když podobně jako výše hledáme předpověď pro přírůstek výšky, dostaneme

$$\widehat{v_{12}} - v_{10} = 43,0441 - 0,2316v_{10}$$

opět se standardní chybou 0,2729 u směrnice, opět s reziduálním součtem čtverců $RSS = 143,2542$, ale s koeficientem determinace $R^2 = 9,3\%$.

Pokusme se naznačený problém vysvětlit.

3. Koeficient mnohonásobné korelace

Ano, koeficient determinace je čtvercem výběrového koeficientu mnohonásobné korelace. Pro jednoduchost uvažujme pouze jednoduchou lineární regresi. Koeficient determinace je roven čtverci výběrového korelačního koeficientu r_{yx}^2 . Pomocí odhadů rozptylu a kovariance lze pro pevné c psát

$$\begin{aligned} r_{y-cx,x}^2 &= \frac{\text{cov}^2(y-cx, x)}{\text{var}(y-cx) \text{var}(x)} = \frac{[\text{cov}(x, y) - c \text{var}(x)]^2}{[\text{var}(y) - 2c \text{cov}(x, y) + c^2 \text{var}(x)] \text{var}(x)} \\ &= \frac{\text{cov}^2(x, y)}{\text{var}(x) \text{var}(y)} \frac{\text{var}(y) [\text{cov}(x, y) - c \text{var}(x)]^2}{\text{cov}^2(x, y) [\text{var}(y) - 2c \text{cov}(x, y) + c^2 \text{var}(x)]} \\ &= r_{yx}^2 \frac{\text{var}(y) [\text{cov}(x, y) - c \text{var}(x)]^2}{\text{cov}^2(x, y) [\text{var}(y) - 2c \text{cov}(x, y) + c^2 \text{var}(x)]}. \end{aligned}$$

Zřejmě tedy je v modelu pro $y - cx$ je koeficient determinace $r_{y-cx,x}^2$ větší než koeficient determinace r_{yx}^2 v modelu pro y , právě když je druhý zlomek v posledním vzorci větší než 1. Je to právě když

$$\begin{aligned} \text{var}(y) [\text{cov}(x, y) - c \text{var}(x)]^2 &> \text{cov}^2(x, y) [\text{var}(y) - 2c \text{cov}(x, y) + c^2 \text{var}(x)] \\ -2c \text{var}(x) \text{cov}(x, y) \text{var}(y) + c^2 \text{var}^2(x) \text{var}(y) &> -2c \text{cov}^3(x, y) + c^2 \text{var}(x) \text{cov}^2(x, y) \\ \text{var}(x) \text{var}(y) [-2c \text{cov}(x, y) + c^2 \text{var}(x)] &> \text{cov}^2(x, y) [-2c \text{cov}(x, y) + c^2 \text{var}(x)]. \end{aligned}$$

Závorka, která je na obou stranách poslední nerovnosti, musí být kladná, neboť jinak bychom dostali neplatnou nerovnost $\text{var}(x) \text{var}(y) < \text{cov}^2(x, y)$. Musí tedy být

$$c [2 \text{cov}(x, y) - c \text{var}(x)] < 0.$$

Odhad směrnice regresní přímky je samozřejmě roven $b_{y|x} = \text{cov}(x, y) / \text{var}(x)$, takže poslední nerovnost lze přepsat jako $c[b_{y|x} - c/2] < 0$. Speciálně pro $c = 1$ dostaneme v našich příkladech nerovnost $r_{y-x,x}^2 > r_{yx}^2$ právě když je $b_{y|x} < 0,5$.

Lapidárně lze rozdíl mezi dvěma hodnotami koeficientu determinace v našich příkladech vysvětlit také takto. Koeficient determinace je u jednoduché lineární regrese čtvercem korelačního koeficientu. Ten je významně nenulový, právě když je významně nenulový regresní koeficient (směrnice) u nezávisle proměnné. Hypotéza o nulové směrnici v modelu pro $y - cx$ je ekvivalentní hypotéze, že směrnice je rovna c v modelu pro samotné y . Protože jsou v obou případech stejné standardní chyby odhadu směrnice, záleží pochopitelně hodnota testové statistiky (korelačního koeficientu, koeficientu determinace) na vztahu mezi odhadem $b_{y|x}$ směrnice a hodnotou c . Koeficient determinace v modelu pro y bude větší než koeficient determinace v modelu pro $y - cx$, právě když bude odhad $b_{y|x}$ blíže k 0 než k c .

Literatura

- A. K. Shah (1991), *Relationship between the coefficients of determination of algebraically related models*, The American Statistician **45**, 300-301.
- J. Radek, E. Partyková (1984), *Závislost výnosu brambor na povětrnostních podmínkách*, Rostlinná výroba **30**, 729-738.

Autorova adresa: KPMS MFF UK, Sokolovská 83, 186 00 Praha 8-Karlín.

Akademická licence SPSS

Jan Řehák

Rozhodování o výběru jednoho programového balíku není prostou záležitostí. Software je drahý a navíc již dávno minuly doby, kdy téměř každý statistik usiloval o to, aby měl všechny dostupné programy na disketách, a tím zvyšoval svoji prestiž před kolegy (jeden exoticky znějící neznámý titul, který mohl být dán k lepšímu ovšem bodoval vždy). Zjistili jsme (či snad alespoň většina z nás), že je škoda disket a času i místa na disku a že nakonec stejně používáme jen jeden z nich. Byly doby, kdy rozšířenost programu byla dána v podstatě jeho nahodilým získáním někým od někoho a pak už interakcí přátel a známých. Byli i takoví, kteří takto získaný software dokázali využít tržně. Dnes už se software (většinou) kupuje. Nehodí se učit na nelegálním programu a veřejná vědecká publikace jaksi předpokládá solidnost autora (o komerčním využívání ani nemluví, tam je to prostě i dost nebezpečné). Ještě z jednoho důvodu se software kupuje: od některých firem dostáváme vysoké nebo alespoň solidní slevy. A přišli bychom o ně, kdybychom programy šířili nelegálně dál. Akademický software je drahý -v relaci k dnešním rozpočtům - a také proto, že se při jeho nákupu většinou čeká až na prostředky, které někde zbydou. Stále ještě není samozřejmé, že koupit drahý počítač bez softwaru je vlastně vyhazování peněz.

Myšlenkou akademického celonárodního pronájmu jednoho programového balíku bylo vysoké kupní náklady překonat tím, že by bylo možné je rozpočítat na mnoho uživatelů v Akademii a na vysokých školách. Tedy získat nejen přímou slevu, ale také slevu nepřímou tím, že by se spojili jinak nezávislí zákazníci v jeden zákaznický subjekt. Rozhodování o takovém statistickém softwaru se konalo u nás před dvěma lety. Uživatelem měl být vědec, výzkumník, statistik-profesionál, pedagog, student. I když nám šlo pochopitelně především o potřeby společenských věd, balík měl být k dispozici pro jiné vědy a vyhovovat jim stejně dobře jako vědám společenským. Bylo to pro nás (a pro mně speciálně) obtížné. Dlouholeté zkušenosti s SPSS vedly k rozpornému stanovisku. Po takové době, obzvláště když s ním pracujete často a na velmi rozmanitých úlohách, nejen oceňujete všechny přednosti systému, ale také důvěrně znáte (a máte protrpěné) všechny jeho nedostatky. A tak jsem se rozjel na Sofstat91 provést podrobný průzkum nabídky.

Rozhodování o softwarové volbě bylo váženo několika kritérii:

1. uživatelská vlídnost programu při jeho řízení,
2. cenové relace a možnosti,
3. zajištění správnosti postupů v matematických procedurách,
4. kompatibilita s počítačovým i výzkumným okolím (navazující programy, přechod na sálový počítač, domácí i mezinárodní partneři ve výzkumu),
5. univerzálnost pro práce různého typu, vhodnost pro práce větších objemů, opakování běžných analýz, snadná manipulace se soubory, ale i možnost používat moderní statistické metody pro kategorizovaná i číselná data,
6. dynamika a odhad budoucnosti systému.

Tato kritéria byla hlavní, pro nás znamenala konkrétně:

1. Běžným uživatelem ve výzkumu je sociolog, psycholog, ekonom ap., dokonce i středoškolsky vzdělaný pracovník. Ti neumí ani se nechtějí učit programování, či složité ovládání programu a bylo by také nesmyslné to na nich vyžadovat. Program musí být snadno ovladatelný a to tak, aby uživatel nebyl v běžné práci s daty na nikom závislý, aby všechny prostředky pro řešení standardních úloh měl k dispozici nekomplikovaným ovládaním jednoho programu. Po obvykle nákladném nákupu vyžaduje zákazník jistou míru uživatelského komfortu, nezávislost na programátorech a nechce samozřejmě dále investovat další prostředky do vlastních dodatečných úprav k zajištění uživatelské přístupnosti. Jednoduché a pohodlné ovládání ovšem ocení i statistik, i když snad pro něj ani složitější jazyk zadávání netvoří principiální problémy. (Jak rychle jsem se přizpůsobil lepicímu menu SPSS a jak nerad se vždycky znovu učím GLIM.)

2. Cenové relace jsou rozhodující nejen v uvažování okamžitých vlastních limitů, ale především vzhledem k využitelnosti (efektivní cena) a ke koncepční postupnosti softwarového i hardwarového vybavení pracoviště. Cena není jen otázkou přítomnosti, ale jde o zajištění dlouhodobě výhodné koupě (obzvláště nyní, při krátkodobém nasazování penitračních cen není možné srovnávat ceny okamžité, ale dlouhodobý výhled); na druhé straně se současná cena softwaru může snadno postupně zhodnotit nepřímo ve vyšším a rychlejším růstu znalostí a možností pracoviště. V cenových úvahách je nutno vzít i celkovou sumu nákladů s provozem spojených, nákladů na dopravu či poštovné následných vyšších verzí, za manuály ap.

3. Málokteré pracoviště má kapacitu na kontrolu správnosti algoritmů, numerické efektivnosti a přesnosti; a ne vždy míváme s programy dobré zkušenosti. U velkých systémů je tato kontrola zajištěna permanentním veřejným zájmem odborníků z univerzit; výsledky jsou publikovány.

4. Kompatibilita se týká několika aspektů: a) je výhodné (a při domácích i mezinárodních společných projektech i nutné) pracovat s jedním společným systémem: výměna souborů, jejich zpracovatelské vybavení, společné a paralelní analýzy, zjednodušení výzkumné komunikace, výměna zkušeností apod. Je proto vhodné se rozhodovat mezi systémy, které jsou majoritně používány v daném oboru; to má i tu výhodu, že jsou k dispozici zkušenosti uživatelé pro poradenství i vzájemnou výpomoc. b) SOÚ potřebuje hladký přenos dat z PC na sálový počítač a zpět, i přenos na jiné operační systémy (UNIX, Windows); c) ve výzkumné a vědecké činnosti využíváme různé speciální analytické programy a potřebujeme hladký přechod do nich i zpět (např. tabulkové kalkulátory, databázové programy, speciální statistické programy - rychlý přechod do GLIM).

5. Pro výzkumnou práci je nesmírně důležité, aby co nejvíce manipulačních i výzkumných činností bylo prováděno "pod jednou střešou", abychom nemuseli neustále přecházet z programu do programu. Statistická práce není jen výpočet statistik a třídění, ale především jsou to úpravy souboru (redukce, výběry, spojování, tvorba a úpravy proměnných), jejichž podíl čerpá velkou většinu času při zpracování dat. Program musí také obsahovat postupy i pro profesionální analýzu dat vysoké komplexnosti a složitosti a musí poskytovat nejpokročilejší statistické metody. I když není hlavním a nejfrekventovanějším uživatelem,

musí tu nalézt své odpovídající prostředky i statistik a matematik.

6. Do každého systému proniká uživatel postupně, po krocích objevuje jeho možnosti a nachází optimální postupy. Po jistém čase už možnosti programu náhle nestačí. Proto s vývojem systému a s jeho novými verzemi roste nejen uživatelská zkušenost a zručnost, ale i analytická a metodologická kvalifikace. Vzniká také určitá metodologická a pracovní kultura, ovlivňující styl práce, přístupy k analýze apod. Statistický software tedy nekupujeme na rok, ale předpokládáme dlouhodobé využívání. Z toho plyne požadavek nejen na dobrý systém, ale i na systém dynamický, perspektivní a přizpůsobivý novým trendům jak v programování tak ve statistice, operativně reagující na velmi rychle se proměňující potřeby výzkumu a statistických analýz, pružně implementující nové metodologie a metody. To úzce souvisí s úvahami o investičním charakteru vložených prostředků. K rozvoji patří také otázka módnosti a někdy i líbivých efektů, které jsou podmiňovány konkurenčním bojem. Tyto na první pohled lákavé momenty však je lépe posuzovat až po důkladném zvážení jejich užitečnosti a frekvence využití (to se týká např. grafiky, exoticky nazvaných metod, efektně popsáno programového zázemí apod.).

Přes široké spektrum nebyl (a dosud není) žádný z nabízených systémů po všech stránkách ideální, i když vývoj je tu velmi rychlý, a všechny prestižní programy se snaží v rámci svých strategií o preciznost, vhodnost, úplnost, uživatelskou vlídnost. Liší se však typem cílového uživatele, směrem vývoje a důrazem na jednotlivé aspekty, i, pochopitelně, úspěšností realizace a konečnou podobou. Velké a tradiční systémy mají velikou setrvačnost svých tradic. Geny, které jim dali jejich duchovní otcové do vínku je provázejí stále a určují neodvratně jejich vývoj, změny nejsou možné (a asi ani žádoucí). Přes veškeré vyrovnávání mají stejné přednosti a stejné nedostatky jako na začátku (ale neříkejte jim to, nemají to rádi). Je pozoruhodné, jak prezentace velkých firem jsou v postupu času stále stejné, a to bez ohledu na konkrétního mluvčího a dobu a obsah. Je to prostě v nich. Je to zákonité a je to dobře. Jen těžko nás mohou překvapit a jen těžko nás mohou zklamat (a to je asi nejdůležitější). Z některých cítíte zanícenou profesionalitu, z některých nadřazenou suverenitu, u jiných pragmatičnost. Někteří ukazují, jak jsou nejlepší, někteří prezentují nové myšlenky a jiní prostě předvedou, co je hotovo. Bylo to tak na Compstatech, na Sofstatech i při jednotlivých prezentacích. Je to styl a image značky. Všichni ale mají společné to, že ukazují to, co je u nich lepší, a nemluví o tom, co v jejich systémech chybí. Za rok, na Compstatu ve Vídni to bude stejné.

Nikdy nestačí rozhodovat podle firemní prezentace a podle předvedených (někdy velmi líbivých) efektů. V sázce je víc: dlouhé hodiny u stroje, každodenní práce, většinou nudná a nepříjemná manipulace s datovými soubory nebo do nekonečna odlaďovaný a opakovaný model a urputný boj o desetinky procenta vysvětlené variance. Pak nejde o krásný graf, ale o jednoduchý postup k (jakémukoliv) jednoduchému obrázku (je-li krásný tím lépe). Této zásadě dal za pravdu Sofstat93, kde bylo možno pozorovat kandidáty pro naši licenci s odstupem a porovnat tehdejší sliby s dnešní realitou. Jsou tu stále v rozhodování aspekty jako cena vs. možnosti, rychlost vs. objem, šíře možností vs. uživatelská jednoduchost. Některé systémy předpokládají velmi intenzivní i extenzivní speciální znalosti, které vyžadují téměř specifickou kvalifikaci (mají např. vlastní jazyk), některé vyžadují enormní paměti apod. Proto je výběr obtížný, neboť je to dnes již výběr z několika dobrých produktů.

Hlavním hlediskem se tak stává právě ona uživatelská vřídnost, vhodnost, jednoduchost, snadné ovládání, univerzálnost, rychlé zaškolení v používání a následná nezávislost uživatele. Je též důležité, aby nákupcím a rozhodovatelem nebyl zprostředkovatel, ale uživatel sám, či ten, kdo s těmito systémy opravdu pracuje a má analytické zkušenosti, tj. ten, kdo má představu o tom, co bude potřebovat výzkumník, učitel, praktický statistik.

Opětné rozhodnutí pro SPSS/PC+ (DOS) bylo v Sociologickém ústavu provedeno po podrobném zvážení všech těchto aspektů, neboť se ukazuje, že ač není ve všech bodech ještě optimální, vyhovuje těmto kritériím rovnoměrně nejlépe a že jeho cena je vzhledem k obsahu a možnostem velmi výhodná. Při jednání s firmou se navíc podařilo získat ony přímé i nepřímé výhody, o nichž byla zmínka výše, a které podstatně překračují rámec SOÚ a dokonce i ČSAV, a které, podle našeho přesvědčení, podstatně pomáhají vybavit naše pedagogická a vědecká pracoviště i jednotlivce. Tyto výhody jsou: bezplatná legalizace, padesátiprocentní slevy, dnes aplikované na verzi pro Windows, sleva pro první rok pronájmu.

Pro akademickou obec však byla poskytnuta navíc velmi podstatná cenová sleva tím, že bylo umožněno sdružit pod globální licenci SOÚ nejen ústavy ČSAV, ale i katedry a fakulty všech vysokých škol v ČR (obdobná smlouva je možná i pro SR). To nám umožnilo, po průzkumu zájmů (SOÚ rozeslalo informaci o této možnosti na všechny fakulty všech vysokých škol a všem vědeckým ústavům ČSAV) a zhodnocení všech rizik, uzavřít smlouvu, která otevírá přístup k pronájmu programu širokému okruhu uživatelů ve vědě a ve školství, a to za velmi výhodných platebních podmínek. Tak vznikla „Akademická licence SPSS“, kterou organizuje a koordinuje SOÚ (koordinátorem je autor tohoto článku, který je zároveň autorizovaným partnerem SPSS). Informace o existenci této smlouvy je zároveň nabídkou všem fakultám či katedrám i všem ústavům AVČR k připojení se k naší licenci a k jejím výhodám.

Obsah programů SPSS/PC+ aplikovaných pod systémem DOS a zahrnutých v licenční smlouvě je dostatečně bohatý pro běžnou i náročnou analytickou práci:

Základní licenční sestava modulů obsahuje:

BASE: manipulace se soubory, transformace dat, základní tabelace četností a průměrů, přehledy, EDA;

STATISTICS: regresní analýza, analýza rozptylu, neparametrické techniky, t-testy, korelace, analýza reliability, faktorová analýza, hierarchická seskupovací analýza;

ADVANCED STATISTICS: diskriminační analýza, MANOVA, log-lineární modely, logistická regrese, nelineární regrese, logitové a probitové modely, analýza tabulek délky života;

TABLES: vytváření tabulek nejrůznějších typů, vhodných pro přímou publikaci;

DATA ENTRY II: vstup, přímá i logická kontrola, přímé i automatické úpravy dat, čištění souborů, úpravy definice systémového souboru.

Volitelné moduly licence obsahují:

CATEGORIES: jednoduchá i násobná korespondenční analýza, analýza spojeného měření založeného na skórech i pořadí, příprava experimentu pro spojená měření;

TRENDS: procedury pro analýzu časových řad zahrnují vyhlazování, dekompozici, regresní postupy, ARIMA, spektrální analýzu, X11ARIMA, speciální transformace.

V případě zájmu je možné rozšířit licenční smlouvu o **GRAPHICS** (přímé napojení z SPSS do známých grafických systémů, jako Axum nebo Harvard Graphics), **MAPPING** (přímé napojení na mapové grafické systémy MapInfo nebo MapMaster), **CHAID** (Chí-kvadrátová automatická detekce interakcí pro kategorizovaná data), **Lisrel**, **DBMS/Co-py**.

Globální charakter licence umožňuje také získat dalším zájemcům se slevou pronájem verzí SPSSX pro sálové počítače, pracovní stanice, a jiné typy strojů, pro UNIX, OS/2. I tato informace je nabídkou pro případné zájemce.

Akademická licence funguje již druhý rok, letos se očekává nová, pátá, verze SPSS/PC+, která bude údajně ovladatelná myší, bude pracovat s extended memory (a požadovat 2MB RAM), bude napojitelná na statistický grafický systém Axum. Má spreadsheet v modulu Base, řadu zlepšených procedur, je podstatně rychlejší a řeší větší úlohy.

Mezitím již mnoho uživatelů získalo také zkušenosti s verzí SPSS pro Windows, která sice vyžaduje o hodně lepší PC, ale má obsah a rychlost verze velkého počítače. V licenční smlouvě obsažena není, avšak řada uživatelů si jí doplňuje své programové prostředky. I když stále ještě nejsem obdivovatelem Windows, efektivita a možnosti SPSS pod nimi mně přesvědčuje. Pro častého uživatele s velkými soubory se vyplatí.

Inzerát

SC&C, Statistické konzultace a výpočty, spol. s r. o.,

hledá statistiky

- pro spolupráci na zpracování a analýzu datových souborů (pomocí systému SPSS)
- statistické a softwarové školení
- tvorba příkazových maker v SPSS
- rozšiřování statistického softwaru.

Firma SC&C nabízí příležitostnou i stálou spolupráci. Zájemci se mohou obrátit písemně na adresu

SC&C, Statistické konzultace a výpočty, spol. s r. o.
K Milíčovu 835, 149 00 Praha 4,

faxovat na číslo (02) 795 18 61,

nebo nechat vzkaz na telefonním čísle (02) 25 88 60.

*** Co potřeboval duševní pracovník ke své práci před 150 lety**

Zápis v knize desiderat lékařské knihovny Medizinisches Lesemuseum založené v Praze v roce 1841: "Podepsaný si dovoluje otázat, zda by fond Lesemusea dostačoval k tomu, aby obstaral do první místnosti kanape nebo sofa. Duševní pracovníci nemohou dobře pracovat bez takzvaného komfortu a pro mnohé je velmi obtížné při čtení hodiny vysedávat na tvrdé židli."

7.8.1850

Dr. Hasner.

Odpověď vedení Lesemusea:

"Bylo obstaráno."

(Dr. Hasner z Arthy [1819-1892], od r. 1849 docent, od r. 1852 profesor očního lékařství na pražské lékařské fakultě.)

[Ludmila HLAVÁČKOVÁ - Dějiny a současnost 6/1991, s.60]

*** Mladý vědec si už na počátku dráhy přináší lásku k třídění a zároveň i první zkušenosti s ním.** Za dlouhých studií si rozsáhlé učivo, mnohdy podstatu stastránkových knih, shrnuje před zkouškou přehledně a stručně na drobnou papírky, jež lid nazývá obyčejně taháky. To je nejpradávnější forma zpracování literatury. Její podstata se nemění ani později, jen se s tím nemusí člověk tajit.

(T. DEVENYI)

*** Není nic hloupějšího, než předpokládat, že počet vědeckých publikací je mírou talentu a užitečnosti vědce.**

(J. BERNAL)

*** V zářijovém čísle Sofwarových novin se mj. implicitně objevila diskuse o školství.** Vyjímáme z ní:

Kdykoliv se nesměle a pokorně zmíním světlušce a hvězdám našeho školství (základního, středního nebo vysokého), že by bylo nanejvýš účelné učit a vést žáky (studenty, posluchače) k vytříbenému umění řečnickému a vyjadřovacímu, ke kvalitním projevům ústním i písemným - světla a hvězdy školství (po pravdě) opáčí, že mají plno jiných důležitějších starostí a problémů. Dříve se zmíněné dovednosti se slušným úspěchem na školách učily; ve shonu a chvatu dnešní cválající reality na takové prkotiny není čas ani nálada.

Myslíte, že přeceňuji význam umění „jak se vyjadřovat“? Už se na nás valí svět, kde platí „publish or perish“, kde se odborník musí prezentovat referáty proslovenými s perfektním přednesem a nadto v cizí řeči, kde musíme umět odpovídat bez použití „-é-é-é jak bych řekl -é-é-é“, kde bez dobrého informativního textu neprodáš sebelepší výrobek. Musíme na sobě hodně pracovat, abychom i na tomto poli obstáli.

(Richard BÉBR, SWN, září 1992, s. 90)

K poznámce pana Bébra bych připomněl, že oficiální trend vede i dnes k potlačení kultury ústního vyjadřování, stačí uvést preferenci písemných testů a zkoušek před ústními (prý to tak dělají i ve světě) maturitou počínaje a zkouškami na progresivních vysokých školách konče. Já za tím vidím jen lenost zkoušejících.

(Ivan STRAKA, SWN, září 1992, s. 91)

[vybral -jc-]

Skripta, knihy

Nová skripta na VŠE - fakulta informatiky a statistiky

Hebák P.: *Rozhodování podnikatelů při riziku a nejistotě*
(94 stran, cena 10,- Kč), 1993.

- skriptum je určeno pro kurs „Statistické rozhodování při riziku a nejistotě.“

Novák I. - Seger J. - Zychová L.: *Statistika B*
(165 stran, cena 33,- Kč), 1992.

- obsahuje tři kapitoly: Regresní analýza a analýza rozptylu, Analýza časových řad, Indexy a diference
- skriptum je určeno jako základní učební pomůcka ke kursu „Statistika B“
- autorský kolektiv řídil a celkovou redakci celé učební pomůcky provedl Prof. Ing. Václav Čermák, DrSc.

Jílek J.: *Hospodářská statistika. Aktualizující teze a materiály*
(248 stran, cena 25,- Kč), 1993

- skriptum je určeno pro kurs „Hospodářská statistika“

Řezanková H., Žvábek J.: *dBASE IV – Control Center*
(115 stran, cena 16,- Kč), 1993

- pomůcka ke kursu „Organizace statistických dat v databázovém systému“
- popisována je verze 1.5.

Česká statistická společnost zve své členy i všechny ostatní zájemce na seminář

„ SOUČASNÁ STATISTIKA “.

Seminář se koná dne 17. září 1993 v budově Vysoké školy ekonomické v Praze od 9.30.

Kontaktní adresy:

prof. RNDr. Jiří Anděl, DrSc.
katedra pravděpodobnosti
a matematické statistiky
MFF UK
Sokolovská 83
186 00 Praha 8

doc. Ing. Richard Hindls, CSc.
katedra statistiky
a pravděpodobnosti
VŠE - FIS
nám. W. Churchilla 4
130 67 Praha 3

Případní další zájemci o vystoupení s aktuálními problémy na semináři se mohou přihlásit do 6. září 1993.

Oznámení

Fakulta informatiky a statistiky VŠE, matematicko-fyzikální fakulta UK a Česká statistická společnost pořádají v pátek dne 3. 12. 1993 oslavy stého výročí narození

prof. PhDr. Jaroslava Janko, DrSc.

Prof. Janko patřil k předním odborníkům v oblasti pojistné matematiky a byl jedním ze zakladatelů moderních matematicko-statistických metod u nás.

Oslavy se konají v aule Centra manažérského a doktorandského studia VŠE, Praha 6 – Veleslavín, ul. J. Martího č. 2. Slavnostní zahájení bude v 10 hodin.

Počítá se s tím, že se na tomto shromáždění setkají bývalí studenti a spolupracovníci prof. Janko. Prosím, upozorněte na tuto akci i své známé.

Od účastníků nebude vybíráno vložné. Potřebujete-li ubytování (hradí účastník), obraťte se nejpozději do 15. listopadu 1993 na děkanát FIS VŠE, nám. W. Churchilla 4, 130 67 Praha 3.

Za přípravný výbor

prof. ing. Jiří Likeš, DrSc.

děkan FIS VŠE

Stipendium ČStS

Veden cílem zvýšit zájem o ČStS mezi mladšími kolegy, jakož i umožnit jejich odborný růst, výbor ČStS se rozhodl vypsát pravidelné stipendium usnadňující účast na odborných konferencích pořádaných v České republice. Pro tento účel výbor prozatím vyčlenil 1000,- Kč pro každý z roků 1994 a 1995. Umožní-li to v budoucnosti finanční situace Společnosti, výbor by rád v této aktivitě pokračoval.

Podmínky pro udělení stipendia jsou následující:

- žadatel(ka) musí být členem ČStS
- věk žadatele(ky) by v roce konání akce neměl překročit 35 let
- žadatel(ka) na akci aktivně vystoupí s referátem.

Žádosti o stipendium, doplněné stručným *odborným* životopisem, seznamem publikací a odborných aktivit, je třeba zaslat k rukám předsedy ČStS nejpozději 1.11.1993 pro akce konané v roce 1994, resp. do 1.11.1994 pro akce konané v roce 1995. Všechny žádosti posoudí komise jmenovaná výborem ČStS, a to nejdéle do 15.12. daného roku. Všichni žadatelé budou o výsledku informováni. Výbor ČStS si vyhrazuje právo rozdělit stipendium nejvýše na dvě části v případě doporučení komise.

Podpora bude poskytnuta bezprostředně po ukončení akce a po předložení

- potvrzení o aktivní účasti na akci
- potvrzení o zaplacení konferenčního poplatku, cestovného apod.

V případě, že skutečné náklady na účast budou vyšší než stipendium poskytnuté ČStS, rozdíl musí být uhrazen stipendistou, resp. tento si musí na jejich krytí nalézt podporu z jiných zdrojů.

V Praze, dne 11.5.1993

Prof. Ing. V. Čermák, DrSc.
předseda ČStS

Vážení kolegové,

sdělujeme Vám, že byl zrušen účet ČStS u Komerční banky, na který jste dosud posílali členské příspěvky. Nový účet byl dne 22.2.1993 založen u České spořitelny, a.s., Sladkovského nám. 4, Praha 3. Jeho číslo je 8024551-938/0800. Členský příspěvek na rok 1993 zůstává nezměněn – 60,- Kč.

Obsah

| | |
|--|----|
| <i>Prokop Závodský</i> , Vlastenecké muzeum v Čechách a rozvoj naší statistické vědy | 1 |
| <i>Stanislav Komenda</i> , Korelace, která nás šatí i obouvá | 2 |
| <i>Jiří Anděl</i> , Jak závisí výsledek sportovního utkání na jeho počáteční fázi | 7 |
| <i>Karel Zvára</i> , Který model je ten pravý | 8 |
| <i>Jan Řehák</i> , Akademická licence SPSS | 11 |
| Co píší jiní | 16 |
| Skripta, knihy | 17 |
| Ze společnosti | 19 |

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání a jednou v roce v anglické verzi. Předseda společnosti: Prof. Ing. V. Čermák, DrSc., VŠE, nám. W. Churchila 3, 130 00 Praha 3, E-mail: vaac@vse.cz.

Redakce: Dr. Gejza Dohnal, Jeronýmova 7, 130 00 Praha 3, E-mail: dohnal@fsik.cvut.cz.