Informační Bulletin

České Statistické Společnosti číslo 3., srpen 1993, ročník 4.

Získejme ženy pro matematiku!

Jitka Dupačová

Nedávno jsem byla požádána o stručnou recenzi sborníku „Winning Women Into Mathematics“, který vydal výbor pro účast žen Americké matematické asociace. Tento výbor pracuje od roku 1987 a cílem sborníku je informovat o práci výboru, vykreslit pravdivě současnou situaci amerických matematiček a přispět ke společenskému ocenění jejich výsledků, vyvolat zájem žen o studium matematiky, o její aplikace i o profesionální karieru v této oblasti. Některé části se zabývají problémy typickými pro USA, jiné mají obecnou platnost. Sborník se dobře čte (můžete si ho půjčit v knihovně MFF UK) a na mnoha místech vás překvapí skutečnost, že autoři tak dobře vystihli a zobecnili i Vaše názory a zkušenosti.

Podobně jako členky výboru a autoři této publikace jsem přesvědčená, že matematika je oblastí bez předsudků a diskriminace (pokud taková oblast lidské činnosti vůbec existuje), přátelskou vůči ženám a poskytující jim více možností dobrého uplatnění na různých úrovních než mnohé jiné profese. Autoři hledají odpověď na otázku proč jsou přesto v americké populaci matematiků ženy relativně málo zastoupeny a proč jejich poměrné zastoupení s rostoucí kvalifikací klesá. Nacházejí 34 příčin společenských, daných výchovou v rodině i ve škole, 15 příčin souvisejících s matematickými zvyklostmi, a ty všechny mají vliv na chování a zvyky jednotlivců. Některé myšlenky mne zaujaly:

- Tradiční hry a hračky pro chlapce rozvíjejí jejich kvantitativní a geometrickou představivost. O typických hračkách pro děvčátka to říci nelze.
- Jedna ze známých tézí tvrdí, že úspěšný výzkum v matematice je doménou mládí. Tím jsou v povědomí matematické veřejnosti apriori vyloučeny úspěchy žen, které se v tomto plodném věku věnují spíše svým mateřským povinnostem.
- Muži jsou přijímáni, resp. navrhováni na další postup, pro svou slibnou perspektivu, zatímco ženy pro již vydobyté postavení a uznávané výsledky.
- Ženy jsou kritičtější ke svým výsledkům, méně publikují, procento přijatých článků mají stejné jako muži.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

Sborník obsahuje i zajímavá (většinou převzatá) data. Podíl žen mezi státními příslušníky USA, kteří získali doktorát v matematice, vzrostl z 9% v roce 1973–1974 na 24% v roce 1988–1989. Důvodem tohoto příznivého vývoje je ale i absolutní pokles počtu mužů absolvujících doktorandské studium. Přitom nejvíce doktorátů (36% všech doktorátů udělených matematičkám) získávají ženy v PRAVDĚPODOBNOSTI a STATISTICE. Z nových doktorů v roce 1989, kteří nastoupili pedagogickou karieru, se 3/5 mužů ale jen 2/5 žen umístily na institucích garantujících doktorandské studium a naopak 1/5 mužů a 2/5 žen na školách s bakalářským studiem. Stejné bylo poměrné zastoupení na školách s magisterským studiem. Přibližně stejná poměrná část obou pohlaví (13% resp. 12% ze všech absolventů) našla uplatnění v průmyslu, výzkumu a ve státních orgánech, také procenta nezaměstnaných byla srovnatelná (6% resp. 8%). Mediány nástupních platů mužů a žen se od sebe téměř nelišily, o polovinu byly vyšší pro obě skupiny v průmyslu oproti školství. Další nárůst platů a profesionálních příležitostí (habilitace, profesury) je však výrazně vychýlen ve prospěch mužů.

Vedle rozborů, zpráv a statistických údajů sborník obsahuje také osobní vzpomínky významných amerických matematiček i vtipy a úsměvné scénáře popisující situace, které mohou ženu potkat. Autoři také navrhují různé možnosti jak získávat větší počet dívek ke studiu matematiky a podporovat je v další karieře, např. organizace speciálních letních seminářů pro dívky nebo stipendia pro doktorandské studium žen.

Na závěr uvedu obsah části příspěvku „Rady oblíbené studentce, která začíná studovat matematiku“ od Patricie Clark Kenschaftové:

— Věnuj alespoň 40 hodin týdně studiu matematiky, tj. návštěvě zapsaných přednášek, získání přehledu o dalších přednáškách, náhodné četbě matematických pramenů, řešení různých matematických problémů, práci ve skupině, účasti na neformálních matematických diskusích se svými kolegy i na čajích a jiných společenských akcích, kde můžeš hovořit s profesory matematiky.
— V zájmu svého duševního zdraví se matematikou nezabývej víc než 50 hodin týdně. Tak ti zbyde dost času pro přátele, jiné zájmy a na péči o tělo a ducha vůbec.

Testování výsledků testů

Jiří Anděl

Znalosti studentů se asi už od pradávna ověřují zkoušením. V poslední době je čím dál tím větší tlak na to, aby zkoušky byly maximálně objektivní. V některých západních zemích, zejména v USA, se mimo jiné i z tohoto důvodu prosazují zkoušky ve formě písemných testů. Tento způsob se přinejmenším drápkem uchytil i u nás: Jistě víte, co to je TOEFL (TEST OF ENGLISH AS A FOREIGN LANGUAGE). Pro ty, kteří to snad ještě neslyšeli: Chcete-li jít studovat do USA, pak zpravidla nejdřív musíte dokázat, že

dostatečně ovládáte americkou angličtinu. To dokážete tím, že zaplatíte příslušný poplatek (samozřejmě v dolarech[1]) a že získáte dostatečně velký počet bodů v testu, který se jmenuje TOEFL. Každá americká univerzita má svou vlastní představu o tom, co to je dostatečný počet bodů.

Na začátku TOEFLu jsou úlohy následujícího typu. Je Vám jednou jedinkrát přečtena anglická věta. Třeba: *Please turn in your room key before you leave.* Tuto větu pouze slyšíte, nemáte možnost si ji přečíst. Je třeba rozhodnout, která z následujících vět je svým významem nejbližší tomu, co bylo přečteno:

(A) *Please lock your room when you leave.*
(B) *Turn the key to the left to enter your room.*
(C) *Please return the key to your room before leaving.*
(D) *You must leave your room by four o'clock.*

Tyto čtyři odpovědi vidíte napsané v brožuře, kterou Vám dají. A na zvláštním papíru, kterému se příléhavě říká *answer sheet*, vyznačíte jednu z možností (A) až (D). Pokud začerníte písmeno (C), znamená to, že jste se rozhodli v tomto případě odpovědět správně. Postupně se na Vás chtějí stále komplikovanější věci. Vyslechnete anglický rozhovor a teprve potom dostanete sérii otázek. U každé otázky si zas vybíráte jednu ze čtyř předdefinovaných odpovědí.

Traduje se, že k úspěšnému zvádnutí TOEFLu je třeba nejdřív několik let strávit v USA. Tam Vás však nepustí dřív, dokud TOEFL nesložíte. Řešení tohoto problému (zda dřív TOEFL nebo USA) nás však nyní nezajímá. Nás zajímá struktura testů tohoto typu.

Pro jednoduchost si mysleme, že jde o test, který má sto otázek. Dále předpokládejme, že ke každé otázce je uvedeno pět různých odpovědí, z nichž jen jedna je správná. A teď přijde statistický problém, jemuž je věnován tento příspěvek. Jestliže se výsledky dvou blízko sebe sedících studentů shodují dejme tomu v 90 odpovědích (ať už jsou správné nebo chybné, jen když jsou stejné) a jen v 10 liší, jaká je pravděpodobnost, že aspoň jeden z nich opisoval od druhého? Triviální (leč zcela špatná) odpověď se získá následující (správnou) úvahou. Pokud by volba kterékoli odpovědi na danou otázku byla stejně pravděpodobná, pak by pravděpodobnost shodné odpovědi byla 1/5 a neshodné odpovědi 4/5. Jsou-li jednotlivé odpovědi pokládány za nezávislé, pak pravděpodobnost, že se ti dva studenti shodnou nejméně v 90 odpovědích, je rovna

$$\sum_{i=90}^{100} \binom{100}{i} \left(\frac{1}{5}\right)^{i} \left(\frac{4}{5}\right)^{100-i} = 2 \times 10^{-51}.$$

Tato pravděpodobnost je zanedbatelně malá, leží pod všemi prahy, které si statistici vymysleli. A přesto tím (zatím) ani jeden z těch dvou studentů nemůže být usvědčen z podvodu. Klíčový předpoklad celé úvahy, že totiž volba kterékoli odpovědi na danou otázku je stejně

[1]Pokud se chcete nechat vyzkoušet v pátek, musíte napřed zaslat $ 45; když myslíte, že Vaším šťastným dnem pro psaní anglických testů je sobota, stačí $ 37. Aspoň takhle je to uvedeno v materiálu nazvaném „1991–92 BULLETIN OF INFORMATION for TOEFL and TSE“. Ta poslední zkratka znamená Test of Spoken English.

pravděpodobná, není splněn ani přibližně. (Navíc není splněn ani předpoklad nezávislosti.) Nemůže nás zajisté překvapit, když většina studentů označí většinou správné odpovědi. Vždyť se přece připravovali, a tak by bylo smutné, kdyby správnou odpověď zaškrtli jen náhodou. Zdálo by se, že předchozí úvahu po nějaké malé modifikaci bude možno aplikovat aspoň na ty otázky, které nebyly zodpovězeny správně. Ale ani to nejde, protože některé chybné odpovědi mohou být svůdnější než jiné (mnohdy dokonce svůdnější než ta správná). Jsou ostatně i další důvody, o nichž si něco povíme později.

Jak jsem již uvedl, studenti jsou asi zkoušeni od pradávna. A zrovna tak dlouho se někteří z nich snaží v případě vlastní neznalosti využít vnější paměť, ať už je to tahák nebo skutečná paměť blízkého spolutrpitele.

Jedním z nejjednodušších triků je záměna osob. Místo studenta X přijde student Y, který examinátorovi tvrdí, že je X. Když mu to examinátor věří a zkouška dopadne dobře, tak se na to nejspíš nepřijde. Může se však na to přijít, když zkouška dopadne až moc dobře. Nedávno měli takový skandál právě v USA. Nějaký pán se chtěl stát v Kalifornii právním obhájcem, ale nezdolal písemný test. Příště místo něj přišla ke zkoušce jeho manželka. Ačkoli každý musel prokázat svou totožnost průkazem s podobenkou a ačkoli byla momentálně v devátém měsíci těhotenství, dohlížející personál zřejmě nepojal žádné podezření. Případ se začal vyšetřovat teprve na základě toho, že test zvládla jako jedna z nejlepších v celém státě.

Proti záměně studentů bojují examinátoři, jak jen umějí. Dovolte mi na tomto místě jednu malou osobní poznámku. Když jsem byl ve druhém ročníku MFF UK, musel jsem v zimním semestru složit také zkoušku z teorie pravděpodobnosti u doc. Seitze. Je třeba konstatovat, že doc. Seitz měl velice dobrou paměť na studenty. Většinu z nich poznal jménem jen u jediné zkoušky — a v každém ročníku jich zkoušel asi 100. Někteří z nich mi pak říkali, že se s ním pak setkali až třeba po dvaceti letech. Oslovil je jménem a hned přidal, ve kterém roce u něj dělali zkoušku. Ale k věci. Chtěl jsem brzy odjet domů na vánoce, a tak jsem ke zkoušce přišel o den dřív, než jsem byl původně zapsán. Přesto mě doc. Seitz vyzkoušel, takže jsem klidně odjel. Po vánocích mi sdělil kolega Jan Ámos Kadlec, že měl u zkoušky z teorie pravděpodobnosti nepříjemnost. Doc. Seitz ho totiž nechtěl vyzkoušet, protože se domníval, že u něj už jednou byl pod jménem Anděl. Trvalo prý dost dlouho, než přesvědčil doc. Seitze, že u něj byl Anděl jako Anděl a ne Kadlec jako Anděl. Mimochodem, chudáka Kadlece brzy po studiích přejel vlak v Římě, ale to už by byla jiná historie.

Další problémy prý mají v USA při hlídání testů s miniaturními vysílačkami. Testy jsou totiž v celém státě stejné (to kvůli objektivnosti), začíná se také ve stejnou hodinu, ale vzhledem k různým časovým pásmům stejná hodina vlastně není tak úplně stejná. Pomocí faxu se pak otázky dají sdělit na jiné místo atd., podrobný návod snad dávat nemusím (a konečně v naší nedávno opět zmenšené vlasti by asi ani k ničemu nebyl).

Vraťme se ke statistice, na kterou tak netrpělivě čekáte a od které stále odbíhám (a to honorář není podle délky článku). V článku Klein S. P. (1992): Statistical evidence of cheating on multiple–choice tests, Chance, Vol. 5, No. 3–4, str.23–27, se popisuje případ dvou chirurgů, kteří v jedné místnosti psali test. Jednomu z nich budeme říkat dr. Koukal a

druhému dr. Samostatný. Tři dozírající osoby po odevzdání testu obvinily dr. Koukala, že během zkoušky nahlížel do formuláře dr. Samostatného. A co říkala dodatečná statistická evidence? Test měl celkem 277 otázek. Oba chirurgové měli 246 stejných odpovědí, z toho 193 správných a 53 chybných. Přitom dr. Koukal měl celkem 210 správných odpovědí a dr. Samostatný jen 200. Statistici zjistili, že bylo celkem 120 zkoušených (včetně dr. Koukala a dr. Samostatného), kteří měli počet správných odpovědí v rozmezí od 200 do 210. Ze 120 lidí se dá vytvořit 7 140 dvojic. Průměrný počet shodných správných odpovědí v rámci všech takto sestavených dvojic činil 160 (při směrodatné odchylce 4,6) a průměrný počet shodných chybných odpovědí činil 13 (při směrodatné odchylce 6,3). To znamená, že 53 shodných chybných odpovědí leží nad průměrem ve vzdálenosti zhruba 13 směrodatných odchylek. Pravděpodobnost takového jevu je menší než 10^{-9}. V soudním sporu, který následoval, se ještě bralo v úvahu to, zda se společné chybné odpovědi vyskytují u otázek příbuzného zaměření atd. Tyto detaily zde nebudu popisovat, případný zájemce si může přečíst výše citovaný článek. Po zvážení všech okolností bylo nakonec ve sporu rozhodnuto, že test vypracovaný dr. Koukalem je neplatný a že k dalšímu opakování může dr. Koukal přistoupit nejdřív za dva roky. K výkonu některých povolání se totiž požaduje i bezúhonnost. Pokud je někdo přistižen při opisování, může to mít dokonce za následek, že už vůbec nikdy mu nebude povoleno opravnou zkoušku skládat. Mimochodem, v uváděném příkladě dr. Koukal připustil, že se občas díval směrem k dr. Samostatnému, ale jen proto, aby se ujistil, že postupuje přiměřeným tempem. Je tedy třeba připustit, že proti dr. Koukalovi svědčilo více indicií. Jednak už samotné svědectví dozírajících, protože ti jen velmi zřídka někoho obviňují z podvodu. Pak vlastní přiznání dr. Koukala a nakonec i výše zmíněné statistické výsledky.

Uveďme ještě několik pozoruhodných případů. Stalo se také, že dva lékaři, manžel a manželka, měli mimořádně mnoho shodných chybných odpovědí. Testy trvaly tři dny a bylo zjištěno, že o jedné přestávce si vyměnili několik slov. Při vyšetřování celého případu bylo však akceptováno vysvětlení těchto manželů, že se ten krátký rozhovor týkal jen jejich nemocného dítěte. Neobyčejná shoda chybných odpovědí byla způsobena tím, že oba pocházeli z Indie, kde také studovali. Ale správné odpovědi na příčiny epidemií a já nevím, čeho ještě, jsou poněkud jiné v Indii než v USA. Tento případ měl tedy happy end.

Hůř to dopadlo se dvěma sestrami, které psaly tentýž lékařský test. Třebaže vůbec od sebe neopisovaly, měly mnoho stejných správných odpovědí. Když se na to přišlo, začalo se zjišťovat, jak je to možné. Ukázalo se, že jejich otec sám pro tuto zkoušku některé otázky připravoval a o některých dalších byl informován. Nejřív se hledalo smírné řešení. Pokud tatínek sám dobrovolně a okamžitě rezignuje z členství ve zkušební komisi, pak se budou moci dcerušky podrobit novému testu v nejbližším možném termínu. Rodina na to odmítla přistoupit, a tak o věci rozhodoval soud. A jeho verdikt byl skutečně přísný. Dívenky mohou psát nový test nejdřív za pět let (dovedete si jistě představit, že do té doby nejspíš zapomenou i to, co neuměly) a jejich otci byl odebrán lékařský diplom.

Podíl piva na rozvoji biometrie

Stanislav Komenda

"Natural selection is a mechanism for geneting an exceedingly high degree of improbability" — Ronald Aylmer Fisher

Místo úvodu

Okolnosti, v tomto případě TEMPUS Project 2194, mě přivedly na čtvrt roku do Institute of Technology v Dublinu. Slovo „přivedly" není výstižné: spíše než přivedly mě okolnosti „převezly" na ferry přes Kanál a Irské moře. Zážitek z plavby sice nevyvrátil mé přesvědčení, že Země je kulatá — upravil je však v tom smyslu, že kulatost zemského povrchu je vlastnost globální, rozhodně však nikoli lokální; příslušný důkaz provedly spojenými silami mé ústrojí rovnováhy a žaludek.

Hlavní důvod mého zdejšího pobytu není pro Bulletin ČSS statisticky významný. Protože však Dublin je pro mě spojen s pivovarem Guinness a ten zase se jménem William Sealy Gosset, jehož „Studentův" $t-$test patří k základním pilířům mého biometrického živobytí, připsal jsem do svého programu také heslo „Student". Zdá se to být případné i z důvodů časových: letos (1993) je to 85 let, co byl $t-$test publikován (1908), a loni to bylo 55 let, co W. S. Gosset zemřel (1937).

Pivovar jsem spíše než pomocí mapy našel díky nosu — slad voněl přes dvě městské čtvrti a vítr vál příznivý — a našel jsem i Svatojakubskou bránu (St. James's Gate), což je adresa pivovaru, kterou uváděl Gosset ve své korespondenci s Karlem Pearsonem a R. A. Fisherem, ale hlavně jsem našel dvousvazkový sborník prací z dějin statistiky a pravděpodobnosti, v němž jsou přetištěny také články z Biometriky, ze kterých v dalším čerpám.

Jen tak mimochodem — hlavním produktem Guinnessů je, samozřejmě, pivo. Černé chutná dost jinak než plzeňské, a málo pění. Hořké je možná díky žateckému chmelu, protože tady se asi chmel nepěstuje. Statistiky zaměstnává pivovar od časů páně Gossetových dodnes.

„Student" jako člověk

William Sealy Gosset se narodil jako nejstarší syn plukovníka Frederica Gosseta, R. E., v Canterburry roku 1876. V roce 1906 se oženil s Marjorií Surtees Phillpotsovou, s níž měl syna a dvě dcery. Zemřel 16. 10. 1937.

Základního vzdělání nabyl ve Winchesteru; na New College v Oxfordu studoval chemii a matematiku.

V roce 1899 vstoupil do služeb u Guinnessů jako pivovarský (brewer); (omlouvám se, mně tenhle termín nezní, ale ve svém pivovarnickém slovníku jsem žádný vhodný ekvivalent anglického „brewer" nevypátral).

Není přesně známo, jak a kdy se zrodil „Studentův“ zájem o statistiku; v oné době se však už začínaly vědecké metody a laboratorní zkoušky v pivovarnictví používat docela vážně, a zřejmě bylo třeba něco vědět o funkci chyb. U Guinnessů pracovala řada lidí s univerzitním vzděláním, takže je docela dobře možné, že na „Studenta“, který z nich věděl o mnatematice nejvíc, se ostatní obraceli a on se proto začal těmito otázkami zabývat. Je známo, že v roce 1903 uměl vypočítat pravděpodobnou chybu. Okolnosti, za nichž se vařilo pivo, kdy materiál je variabilní a citlivý vůči změnám teploty, a kdy jsou pokusné soubory nutně malé, omezovaly aplikaci teorie velkých výběrů a podtrhovaly potřebu metody přiměřené malým výběrům. Nebyla to tedy náhoda, ale okolnosti jeho práce, co přivedlo „Studentův“ zájem k tomuto problému a k objevu distribuce výběrové směrodatné odchylky, v její moderní podobě známé jako $t-$test. Ještě dlouho po objevu a jeho zveřejnění se použití testu omezovalo na okruh pracovníků Guinnessova pivovaru. Biometrická škola na University College se totiž věnovala téměř výhradně velkým výběrům, pro něž neměla „studentizace“, jak se někdy říkalo, zvláštní význam. Přesto byly jak statistické, tak osobně přátelské styky „Studenta“ s Karlem Pearsonem trvalé až do smrti.

Ačkoli byl „Student“ uznáván především jako statistik, zabýval se tolika jinými věcmi, že je s podivem, jak tohle všechno vtěsnal do denního rozvrhu.

V pivovaru prováděl spoustu rutinních statistických výpočtů. Dalo by se proto čekat, že počítal dobře a s přehledem. Nebylo tomu tak — v jeho výpočtech se často najdou drobnější chyby. Míval ve zvyku psát si na starých obálkách a útržcích papíru, třeba ve vlaku. Nerad si výsledky tabeloval, raději dopředu odhadoval. Tím někdy ztrácel spoustu času — dával však přednost možnosti pružně a rychle předcházet od jednoho problému k druhému.

Tato metoda asi není vhodná pro lidi, kteří nejsou tak všestranní jako byl on. Jeho mnohostrannost se podobala spíše lidské ruce než přesnému nástroji. Po mnoho let si stále dokola ručně opisoval záhlaví téhož formuláře pro nová a nová data — a na otázku, proč si nenechá formuláře natisknout, odpovídal — Na to jsem moc líný.

Pro statistiky byl „Student“ statistickým poradcem Guinnessova pivovaru, pro ostatní byl pivovarníkem, v ušetřeném čase se zabývajícím statistikou. Pravda je však především v tom, že dokázal organicky spojovat statistické zkoumání s praktickými problémy, jimiž se zabýval. Jistě si mnozí mysleli, že člověka s jeho nepochybným géniem bylo pro průmysl škoda. Byla to však právě ona návaznost na praktické problémy, co učinilo „Studentovo“ dílo jedinečným, a přes jeho nevelký rozsah tak důležitým. Nejméně jednou mu bylo nabídnuto akademické místo — dá se však předpokládat, že by býval nebyl úspěšným učitelem; na to byl příliš individualistický. Ani se nedá čekat, že by byl úspěšným univerzitním badatelem — jeho mysl fungovala jiným způsobem.

Práce související s křížením ječmene v Department of Agriculture in Ireland, na nichž se Guinnessové významně podíleli, umožnily „Studentovi“ získat bezprostřední zkušenost s pokusy měřícími výnos a se zemědělskými pokusy vůbec. Nebyl zvyklý vysedávat v kanceláři a počítat — raději vedl diskuse a obcházel pole před sklizní.

V posledních 10 letech svého irského pobytu byl vedoucí osobností celého procesu od setí odrůd ječmene, přes růst a sklizeň po sladování a vaření — a zároveň počítání nebo

dohlížení na příslušné matematické práce. Jistou dobu také křížil ječmen ve vlastní zahradě — a urychlil generační obměnu tím, že nechal jednu generaci vyrůst během evropské zimy na Novém Zélandě. Tato křížení byla známa jako Student I a II. Později byla zhodnocena jako chybná — a je typické, že on sám na tuto chybu jako první upozornil.

Přestože toho tolik dokázal udělat, „Student“ nikdy nespěchal. Dokázal přecházet bez prodlení od jedné záležitosti ke druhé, v telefonu začínal bez úvodu hovory o tématech diskutovaných řadu dní předtím. Pomalejší posluchče tím vyváděl z míry. Nikdy neřekl — I am busy.

„Student“ vedl rozsáhlou korespondenci, většinou zemědělskou a pokusnickou, do různých míst světa. Některé z jeho dopisů jsou tím nejlepším, co napsal; své myšlenky v nich vyložil jasněji než v publikovaných článcích.

Protože zvládal tolik rutiny a statistiky v pivovaře navíc ke své statistické činnosti poradenské a přípravě článků, které uveřejňoval, dalo by se čekat, že doma jenom jedl a spal. Tak však tomu ani zdaleka nebylo — měl řadu zájmů domácích i sportovních. Vášnivě pěstoval ovoce — jeho specialitou byly hrušky. Byl obratným tesařem — postavil několik člunů. Jeho tesařina se podobala jeho matematice: neměl rád složité nástroje, skoro všechno dělal kapesním nožem.

Byl vytrvalý chodec a před válkou často jezdil na kole. Úspěšně rybařil. Měl teorii, že pro úlovek je důležitá jenom velikost mouchy a její světlost či tmavost. Ostatní detaily považoval za důležité spíše pro ulovení rybáře obratným obchodníkem; oprávněnost jeho názoru novější zkušenost plně potvrdila. Dobře střílel a slušně bruslil. Až do nehody v roce 1934 téměř pravidelně hrál golf.

O dění ve světě a kolem sebe věděl tolik, co většina lidí. Je s podivem, že mu na tohle všechno stačilo 24 hodin denně.

V osobních vztazích byl velice přátelský a snášenlivý, intriky naznal. Málokdy mluvíval o osobních věcech — když však už k tomu došlo, stálo za to jeho názor vyslechnout, protože nikdy nebyl povrchní.

V létě 1934 si při autonehodě zlomil krček femuru. Tři měsíce ležel (zabývaje se statistikou) a rok potom špatně chodil. Samotná nehoda byla jen těžko vysvětlitelná, protože najel do stojanu lampy na rovné ulici.

Koncem roku 1935 opustil Irsko, aby převzal řízení nového Guinnessova pivovaru v Londýně. Do své nečekané předčasné smrti (zemřel v 61 letech) stačil ještě napsat několik článků.

Pro statistiku představovala „Studentova“ smrt ztrátu jedinečné, svérázné osobnosti.

„Student“ jako statistik

Mnoho let po uveřejnění prvního článku v Biometrice v roce 1907 obklopovalo jméno „Student“ ve statistických kruzích ovzduší romantiky. Ti, kdo ho znali jenom z písemných příspěvků, se jistě divili, kdo je ten zvláštní člověk spokojující se s anonymitou, který píše tak jasně a prostě o tak širokém okruhu základních problémů. Pro ty, kdo se s ním stýkali osobně, znalost faktu, že „Student“ je W. S. Gosset, onen romantický dojem nijak

nesmazávala.

Zvláštnost Gossetova postavení spočívala také v tom, že experimentátorům pomáhal jako matematik, zatímco matematikům zdůrazňoval význam zdravého selského rozumu. Překvapující charakteristikou jeho prací byla jednoduchost statistických nástrojů, kterých používal; stačily mu průměr, směrodatná odchylka a korelační koeficient — a přitom jimi dokázal vyřešit specializované problémy.

Lze to ukázat na jednom případu:

Uvažujme veličiny X a Y se směrodatnými odchylkami σ_X a σ_Y (X a Y mohou představovat výnosy dvou odrůd ječmene). Pro rozdíl výnosů platí vztah

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2\rho\sigma_X\sigma_Y .$$

Gosset jako první razil zásadu, že umění naplánovat dobrý pokus, spíše nežli v redukci rozptylů σ^2_X a σ^2_Y, spočívá v tom, uspořádat ho tak, aby korelace ρ veličin X a Y byla co nejvyšší.

Je třeba si uvědomit, že Gosset se považoval v první řadě za pivovarníka a teprve pak za statistika. Publikoval spíše zdráhavě — častěji na naléhání jiných než z vlastní potřeby. Službu firmě považoval za samozřejmost, takže podstatná část jeho životního díla je především přínosem průmyslovému výzkumu než teoretické statistice na stránkách časopisu Biometrika. Gosset měl mimořádnou schopnost uvidět problémy dříve než jiní.

V roce 1899 se Gosset stal jedním z pivovarníků u pánů Arthur Guinness Son and Co. Ltd — když krátce předtím firma začala do svých služeb přijímat vědce vzdělané v Oxfordu nebo Cambridgi a před těmi se náhle otevřelo nové pole k bádání. Byla tu spousta dat, která měla objektivně vysvětlit vztahy mezi kvalitou suroviny, např. ječmene a chmele, technologií výroby a kvalitou výsledného produktu — piva.

Do popředí se zcela přirozeně dostávaly otázky teorie chyb měření — také pro Gosseta. Záznamy zpráv, které psal pro vedení pivovaru od roku 1904, by mohly být jako připomínky o užitečnosti statistického přístupu v průmyslu použity ještě dnes. Gosset se seznamuje s distribuční křivkou chyb měření a všímá si, jak jsou měření ovlivněna korelací (způsobenou časovou následností nebo prostorovou blízkostí měření).

Kořeny $t-$testu lze najít v dopisu, který zasílá Gosset v roce 1905 Karlu Pearsonovi, u něhož strávil v Biometric Laboratory z pověření pivovaru studijní rok 1906 – 1907: “I found out that *P.E.* (probability error) of a certain laboratory analysis from n analyses of the same sample. This gives me a value of the *P.E.* which itself has a *P.E.* of $P.E./\sqrt{2n}$. I now have another sample analyzed and wish to assign limits within which it is a given probability that the truth must lie. E.g. if n were infinite, I could say “it is 10:1 that the truth lies within 2.6 of the result of analysis”. As however n is finite and in some cases not very large, it is clear that I must enlarge my limits, but I do not know by how much.”

Řešení tohoto problému podal Gosset o dva a půl roku později v článku „The probable error of a mean“, Biometrika **6**, 1–25 (1908).

Podstata Gossetovy úvahy je následující:

Pro výběr n pozorování $\boldsymbol{x}n$ z normální populace se směrodatnou odchylkou σ a nulovou

střední hodnotou veličiny X odvodil Gosset výběrové momenty veličiny $s^2 = \sum(x_i - \bar{x})^2/n$, kde $\bar{x}$ je výběrový průměr. Ukázal, že tyto momenty přesně souhlasí s momenty Pearsonovy křivky typu III a usoudil proto, že křivka výběrové distribuce veličiny s^2 musí mít skoro jistě tvar

$$y = \text{const.}\sigma^{-n+1}(s^2)^{(n-3)/2} \exp\left\{-\frac{ns^2}{2\sigma^2}\right\}.$$

Pak ukázal, že korelační koeficient mezi $\bar{x}$ a s^2 je nula a za předpokladu (který nutně platit nemusí, v tomto případě však platí), že z toho vyplývá absolutní nezávislost $\bar{x}$ a s^2, odvodil pravděpodobnostní rozdělení veličiny $z = \bar{x}/s$ jako

$$p(z) = \text{const.}(1 + z^2)^{-n/2}.$$

Vyšetřil vlastnosti této křivky (v článku jsou menší chyby), tabeloval hodnoty distribuční funkce (pravděpodobnostního integrálu) pro $n = 4, \ldots, 10$ a prověřil přiblížení k normální křivce se směrodatnou odchylkou $1/\sqrt{n-3}$. Pak porovnal distribuce y a $p(z)$ s výsledky výběrového pokusu pro případ $n = 4$ a konečně ilustroval použití svých výsledků na 4 příkladech.

Gossetovu práci z roku 1908 by měli — podle E. S. Pearsona — studovat všichni studenti statistiky ještě předtím než se pokusí o svou vlastní publikaci. Odvození distribucí veličin s^2 a z, nebo $t = \sqrt{n-1}.z$ v dnešním značení, se už dávno dělá jednoduššeji a přesněji; analytický postup není třeba pečlivě zkoumat — v uspořádání a provedení práce je však cosi, co si navždy zaslouží pozornost.

V první řadě, v Úvodu a Závěrech najdeme skvělou ilustraci Gossetovy moudré rady začátečníkovi o tom, jak tvořit: „Nejdříve pověz, co bys chtěl říci, pak to řekni, a nakonec řekni, žes to řekl.“ Hlavní část článku, ono „řekni to“, je rozčleněna do přehledných statí. Adekvátnost předpokladů, o něž se matematická teorie opírá, se testuje experimentálním výběrem; protože test probíhá uspokojivě, uvádí autor tabulky potřebné pro aplikaci a konečně vhodně zvolené příklady osvětlují účel zkoumání.

Je důležité vidět, co bylo hlavním záměrem autora. Jako obvykle to bylo jednoduché a praktické. Měl-li n pozorování, chtěl vědět, v jakém rozmezí pravděpodobně leží průměr populace, z níž byl výběr vzat — jak o tom psal ve zprávě pro vedení pivovaru už v roce 1904. Jeho řešení mlčky zavádí metodu inverzní pravděpodobnosti — přestože Gosset zřejmě neměl v úmyslu tohle precizovat. Poslední věta první stránky článku proto zní:

“The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample is to assume a normal distribution about the mean of the sample with standard deviation equal to $s/\sqrt{n}$, where s is the standard deviation of the sample, and to use the tables of the probability integral.”

Pro Gosseta to znamenalo, že může u malých výběrů předpokládat $z-$rozdělení pro populační průměr kolem výběrového průměru, přičemž stupnici udává výběrová směrodatná odchylka s. Ve svých příkladech používá tabulky rozdělení z ne proto, aby testoval hypotézu, že je populační střední hodnota rovna nule nebo nějaké dané hodnotě, ale aby našel pravděpodobnost, že tato střední hodnota leží uvnitř daných mezí, např. mezi 0 a

∞, tj., že je kladná. V příkladu uvádí pro 10 pacientů průměrné prodloužení spánku po aplikaci D.hyoscyamine hydrobromide $\bar{x} = 0{,}75$ hodin při směrodatné odchylce $s = 1{,}70$. Kdyby měla populační střední hodnota, řekněme ξ, $z-$rozdělení kolem výběrového průměru 0,75 hodin se směrodatnou odchylkou s, pak zřejmě pravděpodobnost, že $\xi > 0$, je úměrná ploše pod křivkou z mezi ordinátou $z = \dfrac{0 - 0{,}75}{1{,}70} = -0{,}44$ a ∞. To je totéž jako pravděpodobnost, že $z < +0{,}44$, k čemuž dává interpolace z jeho tabulek ve sloupci $n = 10$ hodnotu 0,887. Gosset proto usuzuje, že jsou šance (odds) 0,887 ku 0,113, že populační střední hodnota je kladná, tj. že podaná látka prodlužuje spánek. I když je úvaha podpírána vágně definovanou koncepcí inverzní pravděpodobnosti, pokud jde o praktické důsledky, dospěl Gosset k závěru, k němuž se dá sotva něco navíc dodat. Je pravda, že naše dnešní formulace intervalu spolehlivosti se vyhýbá odkazu na inverzní pravděpodobnost, jako praktičtí statistikové však musíme připustit, že naše závěry se shodují.

Jsou tu ještě jiné historicky zajímavé okolnosti. Gosset poznamenává, že než se mu podařilo vyřešit problém analyticky, chtěl tak učinit empiricky. Výběrový pokus, který za tím účelem provedl, spočíval ve vzetí 750 výběrů o 4 prvcích z pečlivě zamíchaných lístků sestavených podle W. R. Macdonellovy korelační tabulky (z roku 1901) obsahující distribuci výšky a délky prostředníku 3 000 zločinců. Bylo to zřejmě ve statistickém výzkumu poprvé, kdy se použilo experimentu založeného na náhodném výběru, což je dnes běžná metoda tam, kde analytický přístup selhává.

Podstatné je také to, že teprve pozornost upřená na malé výběry vedla k rozlišení v označování populačních a výběrových charakteristik.

S rozvojem statistické teorie byl význam „Studentova“ testu hodnocen z mnoha stran, o nichž se Gossetovi ani nesnilo. To je obecným znakem vědeckého pokroku; nicméně pořád platí, co řekl Neyman:

“The role of a rigorous scientific theory is frequently very modest and is reduced to explaining to the practical man — and this sometimes with a certain difficulty — how good is what he himself knew to be good long ago.”

Literatura

Pearson E. S., Kendall M. (Eds): Studies in the History of Statistics and Probability. Vol. I, Charles Griffin and Co. Ltd, London 1970, 1978. Jde speciálně o práce:

Pearson E. S.: Some reflexions on continuity in the development of mathematical statistics, 1885–1920, str. 339–354.

Mc Mullen Launce, Pearson E. S.: William Sealy Gosset, 1876-1937, (1) “Student” as a man, (2) “Student” as a statistician, str. 359–404.

Pearson E. S.: Some early correrspondence between W. S. Gosset, R. A. Fisher and Karl Pearson, with notes and comments, str. 405–418.

Znovu ke koeficientu determinace

Josef Kozák

1. Úvod

V posledních dvou číslech Informačního bulletinu České statistické společnosti (čís. 1 z února a čís. 2 z května letošního roku) RNDR. K. Zvára, CSc., ukazuje, že i tak známá charakteristika, kterou je koeficient determinace, představuje východisko zajímavých metodických úvah. Cílem těchto poznámek je v nich ještě chvíli pokračovat.

2. Základní poznatky

Uvažujme lineární model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad E(\boldsymbol{\epsilon}) = \boldsymbol{0}_n, \quad E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2 \boldsymbol{I}_n, \tag{1}$$

kde $\boldsymbol{Y}$ je známý n-členný vektor obsahující hodnoty vysvětlované proměnné, $\boldsymbol{X}$ známá nestochastická matice vysvětlujících proměnných rozměru $n \times K$, $1 \leq K < n$, s plnou hodností, $\boldsymbol{\beta}$ je neznámý K-členný vektor parametrů a $\boldsymbol{\epsilon}$ je n-rozměrný normálně rozdělený náhodný vektor s uvedenými vlastnostmi, kde $\sigma^2 > 0$ značí neznámý skalár. Předpokládejme přitom, že jak prvky vektoru $\boldsymbol{Y}$, tak sloupců matice $\boldsymbol{X}$ mají nulové průměry, takže při označení $\boldsymbol{1}_n$ pro n-členný vektor jedniček platí

$$\boldsymbol{Y}'\boldsymbol{1}_n = 0 \qquad \text{a} \qquad \boldsymbol{X}'\boldsymbol{1}_n = \boldsymbol{0}_K\,. \tag{2}$$

Pokud jde o vektor $\boldsymbol{\beta}$ a skalár σ^2, budou uvažovány jejich obvyklé odhady pořízené metodou nejmenších čtverců

$$\boldsymbol{b}(\boldsymbol{Y}, \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\,, \tag{3}$$

$$s^2 = (\boldsymbol{Y} - \boldsymbol{Y}(\boldsymbol{X}))'(\boldsymbol{Y} - \boldsymbol{Y}(\boldsymbol{X}))/(n-K)\,, \tag{4}$$

kde

$$\boldsymbol{Y}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{b}(\boldsymbol{Y}, \boldsymbol{X}) \tag{5}$$

je odhad vektoru deterministické složky $\boldsymbol{X}\boldsymbol{\beta}$. Uplatníme-li označení

$$\boldsymbol{U}(\boldsymbol{X}) = \boldsymbol{Y} - \boldsymbol{Y}(\boldsymbol{X}) \tag{6}$$

pro odhad vektoru náhodných poruch $\boldsymbol{\epsilon}$ (vektor residuí), není obtížné dokázat platnost identit

$$\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{X}) + \boldsymbol{U}(\boldsymbol{X})\,, \tag{7}$$

$$\boldsymbol{Y}'\boldsymbol{Y} = (\boldsymbol{Y}(\boldsymbol{X}))'\boldsymbol{Y}(\boldsymbol{X}) + (\boldsymbol{U}(\boldsymbol{X}))'\boldsymbol{U}\ (\boldsymbol{X})\,. \tag{8}$$

Poslední vztah je východiskem pro konstrukci koeficientu determinace

$$D(\boldsymbol{X}) = \frac{(\boldsymbol{Y}(\boldsymbol{X}))'\boldsymbol{Y}(\boldsymbol{X})}{\boldsymbol{Y}'\boldsymbol{Y}} = 1 - \frac{(\boldsymbol{U}(\boldsymbol{X}))'\boldsymbol{U}(\boldsymbol{X})}{\boldsymbol{Y}'\boldsymbol{Y}}\,; \tag{9}$$

protože $n^{-1}\boldsymbol{Y}'\boldsymbol{Y}$, resp. $n^{-1}(\boldsymbol{Y}(\boldsymbol{X}))'\boldsymbol{Y}(\boldsymbol{X})$ lze interpretovat jako rozptyl empirických hodnot, resp. rozptyl odhadů teoretických hodnot, uvedená míra říká, jaký podíl z rozptylu empirických hodnot lze vysvětlit rozptylem odhadů teoretických hodnot.

3. Změna prostoru

Jak uvedeno např. v [2], str.386, výchozí identitu (7) lze při zadaném nenulovém K-členném vektoru $\boldsymbol{c}$ zaměnit identitou

$$\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{c} = (\boldsymbol{Y}(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{c}) + \boldsymbol{U}(\boldsymbol{X})\,, \tag{10}$$

tj. na obou stranách (7) lze odečíst pevný vektor $\boldsymbol{X}\boldsymbol{c}$, což je motivováno tím, že namísto analýzy vektoru $\boldsymbol{Y}$ s nulovým průměrem je z věcných hledisek zdůvodnitelných v konkretní aplikaci považováno za racionálnější analyzovat posunutý vektor $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{c})$ mající rovněž nulový průměr. Jednoduché příklady takové transformace jsou pro $K = 1$ uvedeny v [2] i [1] ; obecnější vysvětlení však proti očekávání není snadné.

Vzhledem k (7) a (3) lze (10) využít k odvození identity

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{c})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{c}) = (\boldsymbol{Y}(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{c})'(\boldsymbol{Y}(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{c}) + (\boldsymbol{U}(\boldsymbol{X}))'\boldsymbol{U}(\boldsymbol{X}) \tag{11}$$

„konkurující“ identitě (8) a nabízející jinou variantu koeficientu determinace

$$D(\boldsymbol{X};\boldsymbol{c}) = \frac{(\boldsymbol{Y}(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{c})'(\boldsymbol{Y}(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{c})}{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{c})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{c})} = 1 - \frac{(\boldsymbol{U}(\boldsymbol{X}))'\boldsymbol{U}(\boldsymbol{X})}{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{c})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{c})}\,, \tag{12}$$

kterému lze pro odlišení s (9) s jistými rozpaky říkat „posunutý koeficient determinace“. Udává podíl, který vysvětluje rozptyl odhadů posunutých teoretických hodnot $n^{-1}(\boldsymbol{Y}(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{c})'(\boldsymbol{Y}(\boldsymbol{X}) - \boldsymbol{X}\boldsymbol{c})$ z rozptylu posunutých empirických hodnot $n^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{c}))'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{c}))$.

Jak již bylo naznačeno, neexistuje z věcného hlediska důvod vyžadující porovnávání obou zavedených koeficientů. Pokud bychom se přesto o takové porovnání zajímali, pak s ohledem na (8) s využitím (3) není obtížné dospět k výsledku

$$D(\boldsymbol{X};\boldsymbol{c}) - D(\boldsymbol{X}) = \frac{(\boldsymbol{U}(\boldsymbol{X}))'\boldsymbol{U}(\boldsymbol{X})}{(\boldsymbol{Y}'\boldsymbol{Y})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{c}))'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{c}))}\boldsymbol{c}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{c} - 2\boldsymbol{b}(\boldsymbol{Y},\boldsymbol{X}))\,, \tag{13}$$

který lze komentovat takto: vzájemný vztah obou koeficientů závisí na vztahu vektoru $\boldsymbol{c}$ a vektoru regresních koeficientů $\boldsymbol{b}(\boldsymbol{Y},\boldsymbol{X})$. Vztah (13) představuje zobecnění úvahy provedené v [1] pro $K = 1$.

4. Regresní model s $K > 1$

V praxi často přichází v úvahu regresní model se dvěma typy vysvětlujících proměnných, kdy v (1) uvažujeme

$$\boldsymbol{X} = [\boldsymbol{T}|\boldsymbol{F}]\,, \tag{14}$$

kde při $J + H = K,\ 2 \leq K\ < n,\ \boldsymbol{T},\ \boldsymbol{F}$ jsou matice rozměru $n \times J,\ n \times H,\ 1 \leq J < n,\ 1 \leq H < n$, s plnou hodností. Tomuto modelu věnujme hlubší pozornost.

(a) V prvé řadě je vhodné uvést výsledky z [3], str.3–12: zaveďme čtvercovou matici řádu n

$$\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{T}(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\,, \tag{15}$$

s vlastnostmi

$$\boldsymbol{M} = \boldsymbol{M}' = \boldsymbol{M}^2\,, \qquad \boldsymbol{M}\boldsymbol{T} = \boldsymbol{0}_{n\times J}\,, \tag{16}$$

dále definujme matici rozměru $n \times H$

$$\boldsymbol{F}_T = \boldsymbol{M}\boldsymbol{F} = \boldsymbol{F} - \boldsymbol{T}\boldsymbol{b}(\boldsymbol{F},\boldsymbol{T})\,, \qquad \boldsymbol{b}(\boldsymbol{F},\boldsymbol{T}) = (\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{F} \tag{17}$$

obsahující odchylky hodnot proměnných druhé skupiny a jejich odhadů pořízených s využitím proměnných prvé skupiny, a nakonec zaveďme dva vektory

$$\boldsymbol{b}(\boldsymbol{Y},\boldsymbol{T}) = (\boldsymbol{T}'\boldsymbol{T}) - \boldsymbol{1}_T'\boldsymbol{Y}\,, \qquad \boldsymbol{b}(\boldsymbol{Y},\boldsymbol{F}_T) = (\boldsymbol{F}_T'\boldsymbol{F}_T)^{-1}\boldsymbol{F}_T'\boldsymbol{Y}\,; \tag{18}$$

pak není obtížné dokázat, že vektor odhadů teoretických hodnot $\boldsymbol{Y}(\boldsymbol{X})$ zavedený v (5) má při označení

$$\boldsymbol{Y}(\boldsymbol{T}) = \boldsymbol{T}\boldsymbol{b}(\boldsymbol{Y},\boldsymbol{T})\,, \qquad \boldsymbol{Y}(\boldsymbol{F}_T) = \boldsymbol{F}_T\boldsymbol{b}(\boldsymbol{Y},\boldsymbol{F}_T)\,. \tag{19}$$

v posuzované situaci strukturu

$$\boldsymbol{Y}(\boldsymbol{X}) = \boldsymbol{Y}(\boldsymbol{T}) + \boldsymbol{Y}(\boldsymbol{F}_T)\,, \tag{20}$$

kterou lze komentovat takto: pracujeme-li s regresním modelem se dvěma typy vysvětlujících proměnných, má vektor odhadů teoretických hodnot $\boldsymbol{Y}(\boldsymbol{X})$ charakter součtu dvou vzájemně nezávislých vektorů, a to vektoru $\boldsymbol{Y}(\boldsymbol{T})$ obsahujícího odhady teoretických hodnot odvozené pouze od prvého typu vysvětlujících proměnných a vektoru $\boldsymbol{Y}(\boldsymbol{F}_T)$ obsahujícího odhady teoretických hodnot odvozených pouze od druhého typu proměnných s vyloučením vlivu prvého typu proměnných.

(b) S ohledem na (20) lze identitu (7) přepsat jako $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{T}) + \boldsymbol{Y}(\boldsymbol{F}_T) + \boldsymbol{U}(\boldsymbol{X})$; v analogii k (6) však $\boldsymbol{U}(\boldsymbol{T}) = \boldsymbol{Y} - \boldsymbol{Y}(\boldsymbol{T})$ značí vektor residuí odpovídající regresnímu modelu založenému na prvém typu proměnných a v důsledku toho dospíváme k identitě

$$\boldsymbol{U}(\boldsymbol{T}) = \boldsymbol{Y}(\boldsymbol{F}_T) + \boldsymbol{U}(\boldsymbol{X})\,, \tag{21}$$

která – protože $(\boldsymbol{Y}(\boldsymbol{F}_T))'\boldsymbol{U}(\boldsymbol{X}) = 0$ – vede dále k identitě

$$(\boldsymbol{U}(\boldsymbol{T}))'\boldsymbol{U}(\boldsymbol{T}) = (\boldsymbol{Y}(\boldsymbol{F}_T))'\boldsymbol{Y}(\boldsymbol{F}_T) + (\boldsymbol{U}(\boldsymbol{X}))'\boldsymbol{U}(\boldsymbol{X}) \tag{22}$$

představující dalšího „konkurenta“ výchozí identity (8). Ta nabízí sestrojit další koeficient determinace

$$D(\boldsymbol{F}_T) = \frac{(\boldsymbol{Y}(\boldsymbol{F}_T))'\boldsymbol{Y}(\boldsymbol{F}_T)}{(\boldsymbol{U}(\boldsymbol{T}))'\boldsymbol{U}(\boldsymbol{T})} = 1 - \frac{(\boldsymbol{U}(\boldsymbol{X}))'\boldsymbol{U}(\boldsymbol{X})}{(\boldsymbol{U}(\boldsymbol{T}))'\boldsymbol{U}(\boldsymbol{T})}\,, \tag{28}$$

který udává, kolik z rozptylu residuálních hodnot vylučujících vliv proměnných prvé skupiny je vysvětleno odhady teoretických hodnot pořízených na základě proměnných druhé skupiny po vyloučení vlivu proměnných prvé skupiny; s jistými rozpaky jej lze nazývat „parciálním“ koeficientem determinace.

V této souvislosti vzniká přirozená otázka týkající se vztahu mezi souhrnným koeficientem determinace $D(\boldsymbol{X})$ a „dílčími“ koeficienty $D(\boldsymbol{T})$ a $D(\boldsymbol{F}_T)$, kde v analogii k (9)

$$D(\boldsymbol{T}) = \frac{(\boldsymbol{Y}(\boldsymbol{T}))'\boldsymbol{Y}(\boldsymbol{T})}{\boldsymbol{Y}'\boldsymbol{Y}} = 1 - \frac{(\boldsymbol{U}(\boldsymbol{T}))'\boldsymbol{U}(\boldsymbol{T})}{\boldsymbol{Y}'\boldsymbol{Y}} \tag{29}$$

značí koeficient determinace odpovídající modelu sestrojenému pouze s využitím prvého typu proměnných. Protože za uvedených okolností platí identita

$$\boldsymbol{Y}'\boldsymbol{Y} = (\boldsymbol{Y}(\boldsymbol{T}))'\boldsymbol{Y}(\boldsymbol{T}) + (\boldsymbol{Y}(\boldsymbol{F}_T))'\boldsymbol{Y}(\boldsymbol{F}_T) + (\boldsymbol{U}(\boldsymbol{X}))'\boldsymbol{U}(\boldsymbol{X})\,,$$

není s ohledem na definice (9), (28) a (29) obtížné dokázat identitu

$$D(\boldsymbol{X}) = 1 - (1 - D(\boldsymbol{T}))(1 - D(\boldsymbol{F}_T)) \tag{30}$$

představující zobecnění známého vztahu mezi dílčími korelačními koeficienty a koeficientem mnohonásobné korelace pro lineární model se dvěma typy vysvětlujících proměnných. Po elementární úpravě (30) lze potom dospět k výsledku

$$1 - D(\boldsymbol{X}) = (1 - D(\boldsymbol{T}))(1 - D(\boldsymbol{F}_T))\,, \tag{31}$$

který charakterizuje míru, ve které se zmenšuje podíl variability nevysvětlené navrhovaným modelem.

Situace (14) přichází v úvahu např. při analýze časových řad ekonomických ukazatelů, kdy $\boldsymbol{T}$ je matice funkcí časové proměnné, o jejíchž zařazení do modelu není pochyb, kdežto $\boldsymbol{F}$ je matice faktorových proměnných (proměnných symptomatických či proměnných hrajících ve vztahu k vysvětlované proměnné roli příčiny), o jejíž výsledné podobě na začátku budování modelu zpravidla existují pochybnosti nejrůznějšího druhu. Za těchto okolností dosud stále není zcela jasné, podle jaké „filosofie“ o definitivním zařazení těch kterých proměnných do modelu rozhodovat. Jistou inspiraci v této souvislosti nabízí právě vztahy (30) a (31). Jde–li o prognostické využití analýzy časové řady, měli bychom se snažit o nalezení

„dobrého“ modelu vedoucího k pokud možno značně vysoké hodnotě souhrnného koeficientu determinace $D(\boldsymbol{X})$; protože v souladu s výše řečeným matice $\boldsymbol{T}$ je obvykle poměrně dobře známa a koeficient $D(\boldsymbol{T})$ se zpravidla zanedbatelně liší od jedničky, není obvykle třeba při tomto typu využití analýzy časové řady „trápit se“ s volbou matice faktorových proměnných $\boldsymbol{F}$. Jde–li naopak o ekonometrické využití časové řady, tj. snažíme–li se o nalezení modelu „chování“ vysvětlované proměnné v závislosti na faktorových proměnných, je třeba se vlivu proměnných z $\boldsymbol{T}$ jako nepotřebných a zavádějících zbavit a orientovat se na matici $\boldsymbol{F}$ maximalizující parciální koeficient determinace $D(\boldsymbol{F}_T)$, a to bez ohledu na výslednou hodnoty koeficientů $D(\boldsymbol{T})$ a $D(\boldsymbol{X})$.

Citovaná literatura

1. Zvára K., *Který model je ten pravý aneb vyberte si koeficient determinace*, Informační bulletin České statistické společnosti **4** (1993), čís.2 8–10.
2. Spanos A., *Statistical Foundations of Econometric Modelling*, Cambridge University Press, Cambridge, 1986.
3. Kozák J., *Shrnutí novějších poznatků o prognostickém využití lineárního modelu časové řady*, interní materiál katedry statistiky VŠE v Praze (1989).

Katedra statistiky a pravděpodobnosti VŠE v Praze, 30.června 1993

Software

Nová verze MS-DOS 6.0

Martin Kořínek

Už je to tak. Ani jsme se pořádně nepokochali operačním systémem MS–DOS verze 5.0 (je mezi námi teprve něco málo přes rok) a už jsme nuceni se prokousat novou verzí tohoto hojně používaného (na poli IBM PC) systému. Pojďme se proto v krátkosti podívat, co „šestka“ nabízí nového.

Poznámka: Tyto řádky píši v době, kdy se prodej ostré verze u nás teprve připravuje. Informace a první letmé seznámení jsme získal pomocí beta verze, která mi byla poskytnuta českou pobočkou společnosti Microsoft.

Instalace probíhá obvyklým způsobem. Máme možnost starou verzi operačního systému MS–DOS uložit pro případ, že bychom se k ní chtěli vrátit. Při instalaci lze nastavit rovnou automaticky české (Czechoslovak) prostředí (pětka toto automaticky neuměla). Při instalaci se dovíme, jaké nové komponenty jsou na disk kopírovány.

Manuál samozřejmě k beta verzi není, ale proslýchá se, že má obsahovat cca 260 stránek, které jsou určeny spíše pro uživatele, než pro programátory.

Při spuštění počítače se na monitoru objeví zpráva „*Starting MS-DOS...*“ a poté se počítač chová dle souborů CONFIG.SYS a AUTOEXEC.BAT (jak jinak, že).

Novinky:

DBLSPACE – Double Space nám dokáže zdvojnásobit kapacitu disků (obdoba SSTOR v DR.DOSu či programu STACKER). Prozatímní nevýhoda je, že obsah disku musíme před kompresí zálohovat, neboť kompresí přijdeme o všechna data. Rovněž není vhodné používat na zkompresované disky program COMPRESS z balíku PC Tools, neboť je s DBLSPACE nekompatibilní a může poškodit data. S DBLSPACE souvisí i nový parametr **/S** u příkazu **DIR**, který způsobí, že vedle standardních informací o souboru se vypíše i kompresní poměr. Program DBLSPACE je ovládán pomocí menu.

DEFRAG – program, který zanalyzuje a případně doporučí kompresi (setřesení) disku. DEFRAG je licence programu SPEEDDISK firmy Symantec (Peter Norton).

DELOLDOS – příkaz smaže soubory (včetně adresáře) starého operačního systému (tedy soubory v adresářích OLD_DOS.1, OLD_DOS.2 atd.).

DELTREE – příkaz maže adresář včetně všech souborů či podadresářů. (Osobně jsem ho raději z disku ihned smazal – jistota je jistota.)

FASTHELP – starý dobře známý HELP z pětky. Nový help (spouští se příkazem HELP) je hypertextový. Objeví se seznam hesel, po kterých se můžeme pohybovat a popřípadě v nich vyhledávat. Po aktivaci některého hesla se objeví podrobnější informace.

INTERLNK a **INTERSVR** jsou příkazy, které pomohou při přenášení souborů mezi počítači po kabelu.

MEMMAKER – program (podobný například QMM), který automaticky analyzuje paměť počítače a případně opraví (nebo doporučí změny) soubor CONFIG.SYS.

MOVE – příkaz přesouvající soubory či adresáře.

MSAV – program pro virovou ochranu našeho počítače. Jedná se o jednodušší verzi programu Central Point Anti–Virus. Program umí detekovat (případně se pokusí odstranit) cca 1000 virů.

MSBACKUP – konečně uživatelsky přijatelná forma zálohovacího programu. Jedná se (opět) o licenci na program Norton Backup firmy Symantec.

MSD – program, který nám předloží všechna technická data našeho počítače (obdoba SI z PC Tools či Norton Utilities). Dokáže rovněž vypsat důležité soubory, obsah paměti včetně rezidentních programů.

Volitelná konfigurace – MS–DOS disponuje nyní řadou příkazů, které můžeme zařadit do souboru CONFIG.SYS a s jejich pomocí vytvořit několik možných konfigurací systému. Čili jeden soubor CONFIG.SYS vlastně může v sobě zahrnovat několik souborů, přičemž nastavení, které zrovna potřebujeme, jednoduše vybereme při zavedení systému z menu.

WINDOWS aplikace – při instalaci DOSu si můžete nainstalovat i tři aplikace, které jsou určeny i pro prostředí WINDOWS. Jedná se o MSAV (virová ochrana), MSBACKUP (zálohovací program) a UNDELETE (program pro restauraci souborů). Po instalaci se samozřejmě ve WINDOWS objeví ikona DOS UTILITIES.

Závěr:

MS-DOS 6.0 přináší řadu důležitých novinek, které běžný uživatel jistě ocení. Pokud sháníte (či prozatím nemáte zakoupeny legální verze) programy pro virovou ochranu, pro zálohování apod., postačí koupit MS-DOS 6.0 a vše je vyřešeno. Jestliže ovšem vlastníte MS-DOS 5.0 a navíc programy jako PC Tools, Norton Utilities, Norton Backup, Central Point Anti-Vitrus (nebo Scan, Tři psy atd.) a QMM, není koupě MS-DOSu 6.0 pro vás nezbytnou.

P.S. Ještě mám dvě poznámky:

1. Microsoft hodlá oznámit, že bude svůj základní software obměňovat nejdříve po dvou letech. Čili, MS–DOS 6.0 by mezi námi měl pobývat nejméně dva roky (a to už stojí za uváženou, jestli si ho raději nekoupit). Totéž samozřejmě má platit i pro WINDOWS.
2. A co říkají hoši z Microsoftu na MS–DOS 7.0? Odpověď nepřekvapí – WINDOS. Chtějí, aby WINDOWS byly plně operačním systémem, čili aby okna obsahovala jádro DOSu.

Martin Kořínek, VŠCHT Pardubice, E-mail: KORINEK@nw1.upce.cs

Teaching Mathematics for Industry

Ve dnech 18.–20. září 1994 se bude konat na ČVUT v Praze konference „Teaching Mathematics for Industry“, kterou pořádá ČVUT, Mathematics Working Group SEFI a JČMF. Konference se bude zabývat otázkami výuky matematických základů umělé inteligence, robotiky, pravděpodobnosti a matematické statistiky.

Bližší informace o připravované konferenci získáte na adrese:
J. Černý, katedra matematiky Fsv ČVUT, Thákurova 7, 166 29 Praha 6, 332 3866, e-mail: k101cer@vm.fsv.cvut.cz

Seminář „Stochastické programování a stochastická aproximace“

se bude na pravidelných čtvrtečních schůzkách zabývat ve školním roce 1993/94 vícestupňovými úlohami, zvláště pak otázkami tvorby modelu, odhadů chyb a perspektivnami ekonomických aplikací. Program bude upřesněn na schůzce dne 14.10.93 v 9.00 hod. v posluchárně KPMS, Sokolovská 83.

Účastníkům semináře i dalším zájemcům doporučujeme přednášku **prof. Andrzeje Ruszczynského** (t. č. IIASA, Laxenburg) „*Parallel decomposition methods for large optimization problems*“, která sekoná v rámci Doktorandského týdne v budově MFF UK v Tróji v pátek 1.10.93 od 9.00 hod. Prof. Ruszczynski se mj. bude věnovat aplikacím paralelní dekompozice při řešení rozsáhlých úloh stochastického programování.

Seminář KPMS MFF UK

zahájil svoji činnost ve středu 13. 9. 1993 ve 14.00 hod. v posluchárně KPMS, Sokolovská 83 přednáškou doc. RNDr. Marie Huškové, CSc. *Nové směry v detekci změn statistických modelů.*

Seminář JČMF „Aplikovaná statistika“

se bude opět konat každý poslední čtvrtek v měsíci od 14.00 hod. v posluchárně KPMS, Sokolovská 83. Zatím je znám následující program:

28.10. Dr. J. Antoch, CSc. (MFF UK Praha) promluví o ISI sympoziu ve Florencii
25.11. Dr. J. Malý, CSc. (SZÚ, Praha) : *Míry asociace v kontingenčních tabulkách*
16.12. téma zatí není stanoveno (snad přednáška J. Militkého o ADSTATu 2.0 ???)

Seminář „Asymptotické problémy matematické statistiky“

se bude konat pravidelně každou středu v 9.00 hod. v posluchárně KPMS, Sokolovská 83. (Seminář zahájil 6. 9. 1993 přednáškou doc. RNDr. Marie Huškové, CSc.)

Seminář z aktuárských věd

pořádaný katedrou pravděpodobnosti a matematické statistiky MFF UK spolu s Českou společností aktuárů bude ve studijním roce 1993/94 zahájen 8.října 1993 přednáškou RNDr. Dany Vorlíčkové, CSc.:

Robustní metody v pojistné matematice.

Pravidelné schůzky semináře se konají v pátek od 8.10 hod. v posluchárně K7, Sokolovská 83, 1. patro.

Pro nejbližší období jsou připraveny přednášky doc. P. Mandla (Komplexní komutační čísla) a doc. T. Cipry (Zdroj zisku životních pojišťoven).

Ze společnosti

Anketa

Vážení kolegové,

Současně s tímto číslem IB připravujeme i letošní mimořádné vydání IB (v loňském roce na něj nedošlo, zato nyní bude o něco tlustší). Podařilo se nám do něho získat svolení autorů a vydavatele k otištění „Statistical Software Guide“ (tedy jakéhosi „Průvodce statistickým softwarem“), který vyšel v letošním druhém čísle Statistical Software Newsletter.

Tento materiál otiskujeme „v původním znění“ (bez titulků), neboť se nám jej nepodařilo (především z časových důvodů) přeložit do češtiny. Domníváme se však, že to není zase až takový nedostatek, neboť kdo by chtěl některé z uvedených programů používat, stejně se bez určité znalosti angličtiny neobejde.

Do vánočního čísla IB máme připraven přehled aktuálních cen některých vybraných softwarových produktů. Tedy informací o statistickém software je poměrně dost. Má však někdo přehled o tom, co z toho takový český statistik vlastně používá ? A používá vůbec počítač ?

Kolega Žváček přišel s nápadem uspořádat mezi členy naší společnosti anketu, jejímž cílem bude „zmapování“ této oblasti. Obracíme se na Vás s prosbou o vyplnění anketního lístku (který najdete v příloze tohoto vydání IB). Cílem je získat základní představu o používaném software u nás. Vzhledem k tomu, že mapujeme v podstatě neznámou oblast, volíme cestu spíše volného textu, který bude podle výsledků normalizován. Prosíme Vás o co nejsvědomitější zodpovězení všech otázek. Vaše odpovědi nebudou použity k jiným účelům, než k výše uvedeným. Nemusíte se proto obávat jakéhokoli zneužití informací v nich obsažených (nebudeme je prodávat!) a klidně můžete odpovídat podle pravdy. Navíc, anketa je anonymní, nicméně jsou tam otázky, které slouží k získání základních údajů o reprezentativnosti výběru (věk, zaměstnání, místo, obor zájmu). Výsledky ankety uveřejníme v některém z dalších vydání IB . Vyplněné anketní lístky pošlete do konce října na adresu: Ing. Hanka Řezanková, CSc., VŠE KSTP, nám. W. Churchilla 4, 130 67 Praha 3.

Obsah

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.