Informační Bulletin

České Statistické Společnosti *číslo 2, ročník 16, duben 2005*

Vážené kolegyně, vážení kolegové,

jak mnozí z Vás vědí, klíčovým řečníkem konference STAKAN 2003 byl pan prof. RNDr. Jiří Anděl, DrSc. Jeho, jak jinak než velmi pěkně připravená a poutavě přednesená, přednáška byla přijata s velkým zájmem. Jak se říká, kdo přišel ten určitě nelitoval. Příspěvek byl později publikován v časopise *Statistika*. Díky laskavosti šéfredaktora pana Ing. J. Křováka, CSc., Vám jej v přepracované verzi předkládáme i na stránkách našeho Bulletinu. Dovolujeme si Vám však nabídnout mnohem více. Jedná se o „filmový přepis" přednášky. Tento „akční film" včetně titulků by nebyl možný bez pomoci mnoha kolegů. Na tomto místě bychom rádi poděkovali především pánům J. Dohnalovi, T. Magyárovi a J. Strouhalovi.

Na přiloženém CD však naleznete mnohem více. Jsou to především texty věnované výuce statistiky v České republice, na jejichž přípravě se naše společnost podílela ať již jako (spolu)pořadatel či (spolu)vydavatel. Dále jsme zařadili elektronickou verzi „sebraného vydání" sborníků konferencí ROBUST 1980 – 2004 jakož i kompletní Bulletiny naší společnosti. Vše je doplněno dvěma volně šiřitelnými programy, tj. `MuPAD` a `R`, spolu s řadou doprovodných textů ať již výukové povahy či usnadňujícími použití programů.

Pokud naleznete ve svém okolí někoho, komu by toto CD udělalo radost, napište nám. ČStS mu jej ráda zašle ve své snaze propagovat statistiku a její výuku v České republice, své hlavní cíle otevřeně deklarované v minulém čísle.

redakce

VOLBA PŘÍKLADŮ VE VÝUCE MATEMATICKÉ STATISTIKY

Jiří Anděl

1. Úvod

Tento příspěvek byl přednesen na konferenci STAKAN, která se konala ve dnech 23. – 25. května 2003 v Bystřici pod Hostýnem. Modifikovaná verze referátu pak byla publikována v časopisu Statistika (viz Anděl 2003). Autor děkuje výkonné radě časopisu Statistika za souhlas otisknout části článku Anděl (2003) v rámci tohoto textu.

Budeme se zabývat hlavně otázkou, jak volit numerické příklady na ilustraci statistických metod při výuce matematické statistiky. Při jejich výběru by se mělo přihlížet i k tomu, aby prezentované příklady byly pokud možno atraktivní. Tím se může vzbudit pozornost studentů a také probíraná látka se lépe zapamatuje. Možná i z tohoto důvodu je v moderních učebnicích často připomínán Simpsonův paradox. Dvě jeho méně známé varianty popíšeme v odst. 2.. V odst. 3. se budeme věnovat jednomu příkladu z regresní analýzy.

Je třeba zdůraznit, že příklady jsou založeny na určitých modelech. Posluchačům se ve výuce vždy zdůrazňuje, že model nemůže být ztotožňován s realitou. Často se model přirovnává k mapě a v rámci tohoto přirovnání se dá poukázat na jeho užitečnost. V literatuře se o modelech můžeme dočíst takovéto výroky (viz Faraway 2000, str. 40):

- A model can be no more than a good portrait.[1] (J. J. Faraway)
- All models are wrong but some are useful.[2] (George Box)
- So far as theories of mathematics are about reality, they are not certain; so far as they are certain, they are not about reality.[3] (A. Einstein)

[1] Žádný model nemůže být ničím víc než dobrým portrétem.
[2] Všechny modely jsou špatné, ale některé jsou užitečné.
[3] Pokud jde o matematické teorie reality, nejsou jisté; pokud jsou jisté, pak nejsou o realitě.

2. Simpsonův paradox

2.1. Měření výkonnosti

Předpokládejme, že lékař A v prvním týdnu uzdravil 3 pacienty ze 4, kteří k němu přišli. Lékař B v této době vyléčil 2 pacienty ze 3. V následujícím týdnu A vyléčil 4 z 11, kdežto B vyléčil 1 ze 3. Výsledky zaneseme do tabulky.

Lékař	První týden	Druhý týden	Celé období
A	$\frac{3}{4} = 0{,}750$	$\frac{4}{11} = 0{,}364$	$\frac{7}{15} = 0{,}467$
B	$\frac{2}{3} = 0{,}667$	$\frac{1}{3} = 0{,}333$	$\frac{3}{6} = 0{,}500$

Ačkoliv první týden byl úspěšnější lékař A, druhý týden rovněž, za celé období byl úspěšnější lékař B. Tento paradox se snaží někteří lidé vyložit tím, že lékař B vyléčil méně pacientů. Ukážeme, že tomu tak není.

Představme si dva studenty, které označíme A a B. Každý z nich patří do jiné studijní skupiny. V každé skupině se píší stejně těžké písemky. V zimním semestr A uspěl ve 2 z 8, B uspěl v 1 z 5. V druhém semestru A uspěl ve 2 ze 2, B uspěl ve 4 z 5. Následující tabulka ukazuje, že A byl úspěšnější v zimním semestru i v letním semestru, ale v celém akademickém roce byl úspěšnější B. Přitom oba psali celkem 10 písemek, tedy stejný počet.

Student	Zimní semestr	Letní semestr	Celý akademický rok
A	$\frac{2}{8} = 0{,}250$	$\frac{2}{2} = 1{,}000$	$\frac{4}{10} = 0{,}400$
B	$\frac{1}{5} = 0{,}200$	$\frac{4}{5} = 0{,}800$	$\frac{5}{10} = 0{,}500$

Mohl by vzniknout pocit, jde o umělá data (což je pravda), zatímco u reálných dat takovéto paradoxy nenastávají (což není pravda). V literatuře jsou zmínky o tom, kdy se na takovou situaci narazilo. Došlo k tomu např. při porovnávání úspěšnosti hráčů košíkové i při velkých výběrových šetřeních v USA (viz Hilton a kol. 2002). Dochází k tomu i při různých dalších, zdánlivě seriózních argumentacích. Jednu z nich nyní uvedu. Dejme tomu, že se na nějakou fakultu přihlásilo 2000 uchazečů, z toho 1000 žen a 1000 mužů. V přijímacím řízení bylo přijato 400 žen a 500 mužů. I když nejde o náhodný výběr, někdo může provést oboustranný test homogenity dvou binomických rozdělení. Ve statistických programech (jako je např. program R) se za testovou statistiku bere U_b^2, kde U_b je uvedeno ve vzorci (8.8) na str. 98 v knize

Anděl (1998). Výpočet dává, že se poměry 400/1000 a 500/1000 liší na hladině 6,968e-06, tedy 7×10^{-6}. Podrobnější rozbor celé situace však ukazuje, že se uchazeči hlásili do dvou studijních oborů, řekněme A a B. Předpokládejme pro jednoduchost, že se každý uchazeč hlásil jen do jednoho z nich. Výsledky jsou uvedeny v následujících dvou přehledech.

Ženy	Obor A	Obor B	Celkem
Počet přijatých	200	200	400
Počet nepřijatých	600	0	600
Počet uchazeček	800	200	1000

Muži	Obor A	Obor B	Celkem
Počet přijatých	100	400	500
Počet nepřijatých	400	100	500
Počet uchazečů	500	500	1000

Na obor A byl přijat signifikantně větší poměr žen než mužů (poměry 200/800 a 100/500 se liší na hladině 0,037) a na obor B byl také přijat signifikantně větší poměr žen než mužů (poměry 200/200 a 400/500 se liší na hladině 8×10^{-12}). Jak jsme však viděli, celkové údaje vykazují vztah zcela obrácený. Je jasné, že takováto data mohou být interpretována diametrálně odlišně podle toho, jak podrobná informace je poskytnuta. Reálná data tohoto typu o přijímacím řízení v U. C. Berkeley v r. 1973 se najdou v publikaci Bickel a kol. (1975).

2.2. Pohádka

Příklady uvedené v předchozím odstavci už začínají být poměrně známé. Méně známý je případ, kdy postupné kolapsování tabulek vede k různým závěrům v několika krocích. Nejprve však připomeneme některé pojmy z analýzy kontingenčních tabulek.

Dejme tomu, že nějaký muž onemocní určitou chorobou. Zjistí si, že zatím touto chorobou onemocnělo 18 mužů. Někteří z nich se léčili, jiní ne; někteří přežili, jiní zemřeli. Údaje jsou v tab. 1.

Takový člověk může uvažovat následovně. Bude-li se léčit, šance na přežití se dá odhadnout na 5:6. Nebude-li se léčit, pak šance na přežití bude zhruba 3:4. Vydělením se zjistí, co je výhodnější. Protože poměr

$$\frac{5:6}{3:4} = \frac{5 \times 4}{6 \times 3} = \frac{20}{18}$$

Tabulka 1: Údaje o nemocných mužích

	Přežili	Zemřeli	Celkem
Léčeni	5	6	11
Neléčeni	3	4	7
Celkem	8	10	18

je větší než 1, bude patrně výhodnější dát se léčit. Poměr šancí (odds ratio) v obecné čtyřpolní tabulce s četnostmi n_{ij} je

$$b = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Protože n_{ij}/n je odhadem příslušné pravděpodobnosti p_{ij}, je b vlastně odhadem teoretického podměru šancí

$$\beta = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

A nyní budu (podle knihy Hilton a kol. 2002) vyprávět jednu pohádku. Byla jednou jedna mírumilovná země. Jmenovala se Borgonsko. Všichni její obyvatelé již od mládí vášnivě rádi hrávali stolní tenis. Jednou však došlo ke sporu se sousední válečnickou zemí, která se jmenovala Macenrie[4]. Místo toho, aby se spor řešil válkou, vyzval předseda vlády Borgonska předsedu vlády Macenrie k uspořádání velkého turnaje ve stolním tenise mezi občany obou zemí. Výzva byla přijata a zjistilo se, že každá země má přesně 2 090 000 registrovaných hráčů stolního tenisu. Soupeři byli určeni losem a byl stanoven termín turnaje. V turnaji zvítězila Macenrie nad Borgonskem v poměru 1 100 000 ku 990 000. Je možné, že na tom měl zásluhu ministr zdravotnictví Macenrie MUDr. Ignorant Popleta[5]. Ten totiž vynalezl lektvar[6], který podle jeho názoru zvyšoval u hráčů jejich šanci na výhru. Proto doporučoval hráčům, aby si tento lektvar koupili a před turnajem ho vypili. K jeho nelibosti ho však poslechli jen někteří. Během turnaje si vedl přesné statistiky a doufal, že výsledky potvrdí dobré účinky lektvaru. Jako vzdělaný lékař pan ministr věděl, že je vhodné, aby sledoval výkony žen a mužů odděleně. Navíc sledoval také faktor věku, a tak ženy a muže rozdělil na mladé (do 30 let) a na

[4] Názvy zemí mohou souviset se známými sportovci, jako byli Bjorn Borg a John McEnroe.

[5] V anglicky psané literatuře se tento ministr jmenuje Dr. Ignoramus Fuddle-Thought.

[6] Lektvar se nazýval Everswear.

ty ostatní. Dostal následující výsledky, které jsou pro přehlednost uvedeny v desetitisících.

Muži nad 30

	Lektvar	Bez lektvaru
Vítězství	6	8
Prohra	6	9

$(6 \times 9 = 54) > (6 \times 8 = 48)$

Muži pod 30

	Lektvar	Bez lektvaru
Vítězství	6	22
Prohra	3	12

$(6 \times 12 = 72) > (3 \times 22 = 66)$

Ženy nad 30

	Lektvar	Bez lektvaru
Vítězství	4	17
Prohra	6	27

$(4 \times 27 = 108) > (6 \times 17 = 102)$

Ženy pod 30

	Lektvar	Bez lektvaru
Vítězství	4	43
Prohra	3	33

$(4 \times 33 = 132) > (3 \times 43 = 129)$

Jelikož v každé z těchto tabulek poměr šancí vyšel větší než jedna, pan ministr Popleta radostně sděloval všem svým přátelům, že jeho lektvar výrazně přispěl k vítězství Macenrie.

Výbor proti věkové diskriminaci (VPVD) však odmítal tyto závěry s tím, že sledování výsledků obyvatel Macenrie v závislosti na věku je nezákonné. Trval na tom, aby výsledky byly shrnuty bez ohledu na věk. To vedlo k následující statistice.

Všichni muži

	Lektvar	Bez lektvaru
Vítězství	12	30
Prohra	9	21

$(12 \times 21 = 252) < (9 \times 30 = 270)$

Všechny ženy

	Lektvar	Bez lektvaru
Vítězství	8	60
Prohra	9	60

$(8 \times 60 = 480) < (9 \times 60 = 540)$

VPVD suše konstatoval ke zděšení ministra Poplety, že ve skutečnosti lektvar uškodil výkonnosti mužů i žen. To jasně vyplývá z faktu, že poměr šancí je v obou tabulkách menší než jedna.

Ministr se ještě nestačil z této rány vzpamatovat a už ho navštívila delegace zástupců Asociace proti sexuální nadřazenosti (APSN). Tito zástupci žádali, aby údaje o mužích a ženách byly sloučeny do jediné tabulky. Proti tomu MUDr. Popleta nenašel sílu protestovat. Ostatně se mu zdálo, že už na tom nemá co ztratit. Tím se dostala tato finální tabulka.

Všichni hráči

	Lektvar	Bez lektvaru
Vítězství	20	90
Prohra	18	81

$(20 \times 81 = 1620) = (18 \times 90 = 1620)$

Jelikož v ní je poměr šancí přesně roven jedné, lektvar evidentně nemá žádný účinek. Popletený ministr si vzal velkou dávku lektvaru, rozšlapal svou oblíbenou pingpongovou pálku, zařval a za stálého klení vydechl naposled.

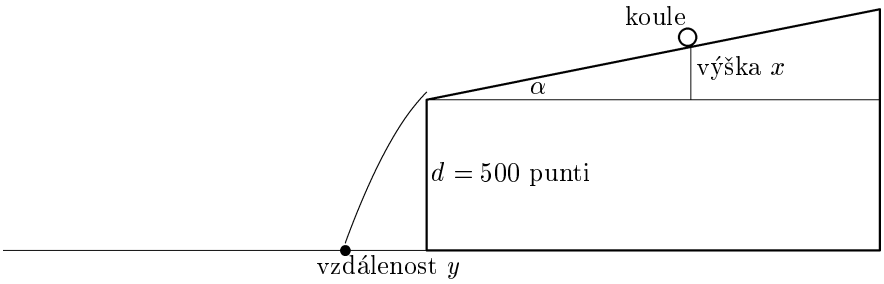
Další zajímavé informace o Simpsonově paradoxu může čtenář najít v diskusním příspěvku k článku Anděl (2003) na str. 29, který má název Kauzalita v epidemiologii a jehož autorem je RNDr. M. Malý.

3. Galileův pokus

Je mnoho názorů na to, jak se mají volit příklady ve výuce matematické statistiky. V jedné učebnici jsem četl, že zásadně má jít o reálně naměřená data, publikovaná v odborné literatuře příslušného vědního oboru. Já jsem byl vždy toho názoru, že data mají popisovat nějaký všeobecně známý jev (např. sjíždění pneumatik u automobilu) a má jich být málo, aby jejich zadávání do počítače bylo rychlé a snadné. Obzvláštní výhodou pak je, je-li všeobecně známa teoretická zákonitost, jimiž se data řídí.

Takováto vskutku ideální data jsou publikována v knize Ramsey, Schafer (1997) (autoři píší, že je převzali z článku Drake, MacLachtan 1975). V roce 1609 Galileo matematicky dokázal, že těleso mající nenulovou vodorovnou složku rychlosti padá k zemi po parabolické dráze. Galileo tento objev učinil o rok dříve při fyzikálním pokusu, který sledoval jiný cíl. Tato zákonitost se dnes probírá v základních kurzech fyziky, takže ji můžeme považovat za všeobecně známou. Při odvozování tohoto fyzikálního poznatku se činí některé zjednodušující předpoklady, např. že zemská přitažlivost má konstantní směr i velikost, že na těleso nepůsobí odpor vzduchu apod.

Galileo studoval pohyb tělesa, jehož vodorovný pohyb není podstatným způsobem ovlivněn třením. K tomu si sestrojil přístroj, jehož schéma je znázorněno na obr. 1. Na stůl umístil nakloněnou rovinu s drážkou. V drážce v určité výšce x (ve výpočtech a v obrázcích použijeme pro tuto výšku anglického označení *height*) nad stolem vypustil bronzovou kouli natřenou inkoustem. Změřil vzdálenost y (*distance*) od stolu k místu, kde na podlaze při dopadu koule udělala inkoustovou skvrnu. Pokus opakoval při různé počáteční výšce koule nad stolem. Dostal údaje zaznamenané v tab. 2. Data jsou



Obrázek 1: Galileův pokus

uvedena v jednotkách zvaných punti (body). Jedno punto je rovno 169/180 mm.

Galileo chtěl tímto pokusem zjistit, zda při zanedbatelném tření je vodorovná rychlost pohybujícího se objektu konstantní. Když si maloval trajektorie padajícího tělesa do svého zápisníku, zřejmě ho napadlo, že touto trajektorií je vždy parabola. Jakmile na tuto myšlenku přišel, nebylo už pro něj obtížné dokázat to i matematicky. Nyní Ramsey a Schafer (1997) doslova píší: *Although Galileo's experiment preceded Gauss's invention of least squares and Galton's empirical fitting of a regression line by more than 200 years, it is interesting to use regression here to see what form of trajectory the data support.*[7]

Data z tab. 2 jsou znázorněna na obr. 2. Zběžné posouzení tohoto obrázku nasvědčuje tomu, že by bylo vhodné vyrovnat data kvadratickou funkcí. Pokud tak učiníme (třeba pomocí programu R), dostaneme tyto výsledky:

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
(Intercept)	1.999e+02	1.676e+01	11.928	0.000283	***
height	7.083e-01	7.482e-02	9.467	0.000695	***
I(height)^2	−3.437e-04	6.678e-05	−5.147	0.006760	**

Residual standard error: 13.64 on 4 degrees of freedom
Multiple R-Squared: 0.9903, Adjusted R-Squared: 0.9855
F-statistic: 205 on 2 and 4 DF, p-value: 9.333e−005

[7] Ačkoliv Galileův experiment předcházel Gaussovo vynalezení metody nejmenších čtverců a Galtonovo empirické proložení o více než 200 let, je zajímavé použít zde regresi a podívat se, jaký tvar trajektorie odpovídá datům.

Tabulka 2: Výsledky Galileova pokusu

Vodorovná vzdálenost y (punti)	Počáteční výška x (punti)
253	100
337	200
395	300
451	450
495	600
534	800
573	1000

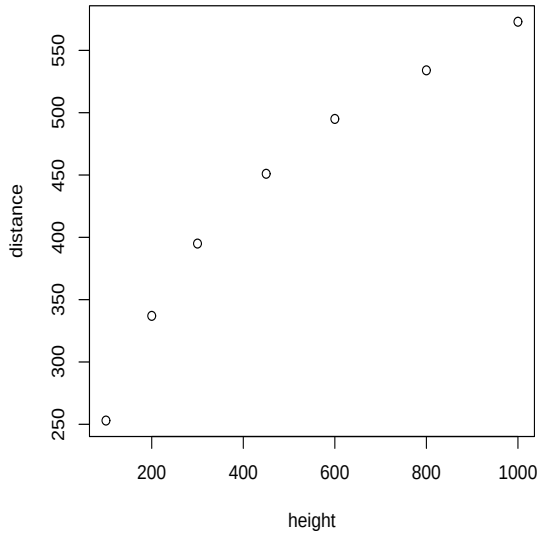
Kvadratická funkce výborně vystihuje Galileova data. Z reziduální standardní odchylky vypočteme, že reziduální součet čtverců je roven 744,1984. Na toto číslo se odvoláme později. Hodnotu R^2 interpretujeme tak, že model je schopen vysvětlit 99.03 % rozptýlenosti naměřených vodorovných vzdáleností. Přesto proložíme ještě regresní polynom třetího stupně. Tím dostaneme následující výsledky.

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	1.558e+02	8.326e+00	18.710	0.000333 ***
height	1.115e+00	6.567e-02	16.983	0.000445 ***
I(height)^2	−1.245e-03	1.384e-04	−8.994	0.002902 **
I(height)^3	5.477e-07	8.327e-08	6.577	0.007150 **

Residual standard error: 4.011 on 3 degrees of freedom
Multiple R-Squared: 0.9994, Adjusted R-Squared: 0.99987
F-statistic: 1595 on 3 and 3 DF, p-value: 2.662e−005

Z toho vyplývá, že se koeficient u kubického členu signifikantně liší od nuly (dosažená hladina testu je 0,007). Nicméně přidání kubického členu vysvětlí jen dalších 0,91 % rozptýlenosti. Nadto víme, že přítomnost kubického členu můžeme připsat na vrub odporu prostředí. Citováno doslova z knihy Ramsey, Schafer (1997): *The significance of the cubic term can be explained by the effect of resistance.*

Shrňme tedy dosavadní výsledky. Data jsou pro výuku zcela ideální, protože

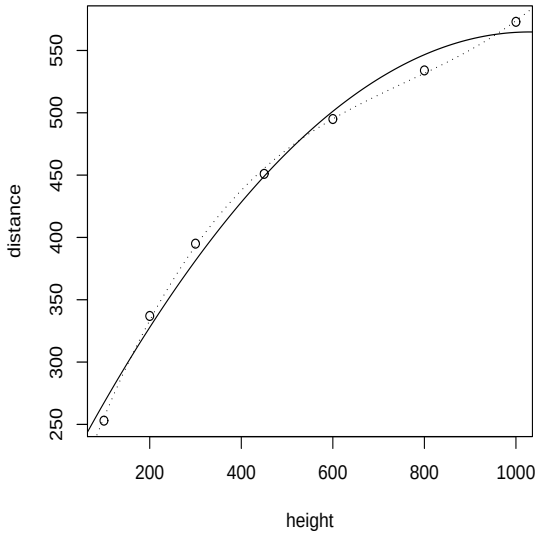


Obrázek 2: Galileova data

- jsou to reálná data,
- jsou to experimentální data (nikoliv observační),
- počet dat je malý,
- přesnost měření je velká, takže i malý počet dat stačí k vystižení fyzikálního zákona,
- data popisují jev známý každému,
- data potvrzují správnost parabolické závislosti v rámci zjednodušených předpokladů,
- statisticky je prokázáno, že parabolická závislost není úplně přesná, což lze ihned vysvětlit odporem prostředí; existence tohoto odporu je statisticky prokázána, ale jeho praktický vliv je nesmírně malý.

Přesto však s těmito výsledky nemůžeme být zcela spokojeni, protože to je celé úplně špatně. Předně může být poučné, když si prostudujeme nejen na Galileova data znázorněná na obr. 2, ale i kvadratickou regresní funkci, která je jimi proložena. To je uděláno v levé části obr. 4. Tam je kvadratická

funkce rostoucí. Jak ale všichni víme, brzy přijde její vrchol (to bude v bodě $x = 1030{,}4$) a její sestupná větev. Těžko se však smíříme s tím, že pak větší výšce x pouštěné koule bude odpovídat menší vzdálenost jejího dopadu a že při velkých výškách x bude vzdálenost y dokonce záporná.



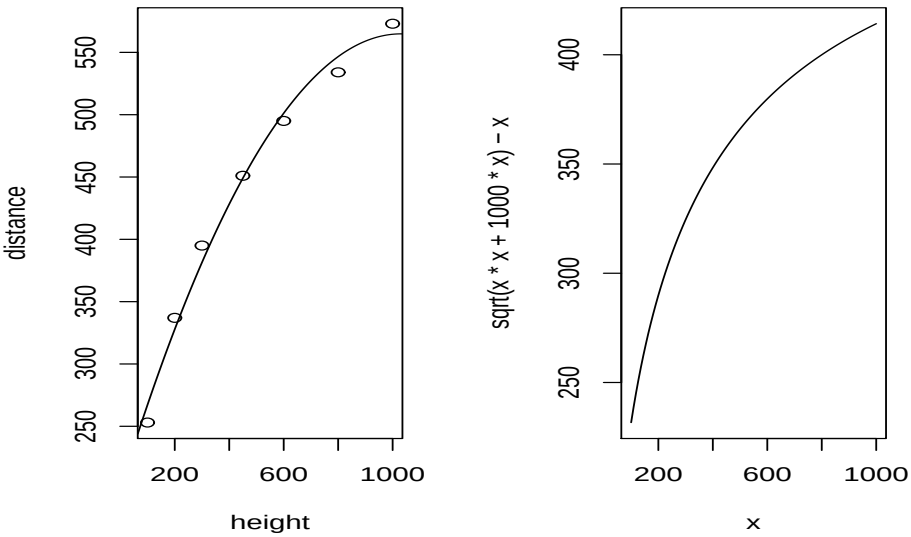
Obrázek 3: Kvadratická a kubická regrese

Situaci nezachrání ani přidání kubického členu. Na obr. 3 je plnou čarou znárorněna kvadratická regresní funkce a tečkovaně pak kubická. Poměrně náhlá změna průběhu kubické regresní funkce je nepochybně z fyzikálního hlediska zcela neopodstatněná.

Poslední zkoušku z fyziky jsem absolvoval téměř před 50 léty. Naštěstí se od té doby fyzikální zákony pohybu po nakloněné rovině a šikmého vrhu nezměnily. Proto jsem se pokusil odvodit závislost délky dopadu y na výšce x. Označil jsem si α velikost úhlu, který svírá nakloněná rovina s rovinou vodorovnou a dostal jsem výsledek

$$y = \sqrt{x^2 \sin^2 2\alpha + 4dx \cos^2 \alpha} - x \sin 2\alpha. \tag{1}$$

Není těžké ověřit, že jde o rovnici hyperboly (přesněji řečeno, její části), ale rozhodně ne o kvadratickou funkci. Abych si zkontroloval, zda někde ve výpočtu nemám hrubou chybu, dosadil jsem známou hodnotu $d = 500$. Protože



Obrázek 4: Galileova data a použité modely

obr. 1 je jen schématický, nelze určit α. Zvolil jsem $\alpha = \pi/4$ (slovy čtyřicet pět stupňů), jelikož se při této hodnotě rovnice (1) nejvíce zjednoduší. Výsledkem je

$$y = \sqrt{x^2 + 1000x} - x. \tag{2}$$

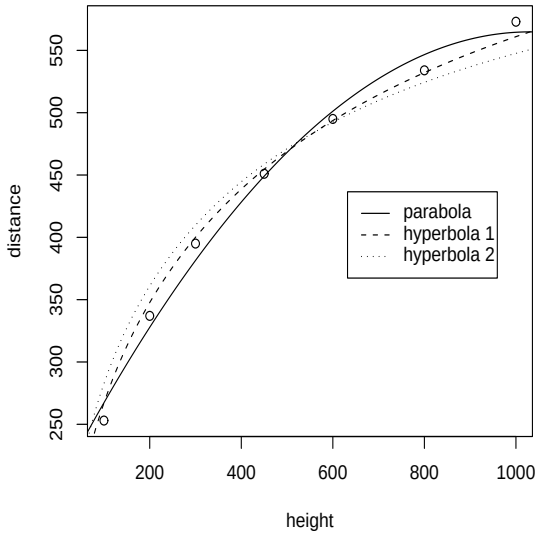
Graf této funkce je znázorněn v pravé části obr. 4. Podle mého názoru shoda s tím, co je v levé části obrázku, je lepší, než jsem doufal.

Takže všechno je jinak. Kvadratická regresní funkce je ve sledované oblasti měření jen aproximací skutečné funkce (1), signifikantní kubický člen spíše signalizuje odchylku od správného modelu než vliv odporu vzduchu na pohyb koule.

Jak by dopadlo proložení správné regresní funkce (1)? Pro zjednodušení zavedeme nový parametr $a = \sin 2\alpha$, takže jde o proložení funkce

$$y = \sqrt{a^2x^2 + 4d(1 - a^2)x} - ax, \tag{3}$$

kde $d = 500$. Použitím nelineární regrese se metodou nejmenších čtverců pro a dostane odhad $\hat{a} = 0{,}6793403$ (a reziduální součet čtverců 503,8294). Z rovnice $\sin 2\hat{\alpha} = \hat{a}$ dostaneme $\hat{\alpha} = 0.3734317$ (pochopitelně v radiánech), takže $\hat{\alpha} \doteq 21{,}4\,^\circ$. Na obr. 5 jsou znovu znázorněna Galileova data, plnou



Obrázek 5: Porovnání regresních funkcí

čárou je znázorněna dříve odvozená regresní parabola a nyní navíc čárkovaně je znázorněna regresní hyperbola. I zběžné porovnání těchto křivek ukazuje, že hyperbola vystihuje naměřená data lépe naž parabola.

Povšimněme si, že náš reziduální součet čtverců činí 503,8294, přičemž jsme odhadovali jediný parametr a. V případě kvadratické regrese bylo nutno odhadovat tři regresní parametry, a přesto byl reziduální součet čtverců větší — činil 744,1984 (jak si zajisté vzpomínáme).

Přesto však s těmito výsledky nemůžeme být zcela spokojeni, protože to je (zase) celé úplně špatně. Doc. K. Zvára si povšiml, že ve správném vzorci (1) poslední člen pod odmocninou obsahuje výraz $\cos^2 x$, zatímco vztah (3) odpovídá tomu, jako kdyby tam byl výraz $\cos^2 2x$. Prostě chyba jako vystřižená z knížky Cipra (2002).

Tak nezbývá než pokusit se o proložení regrese do třetice. Proložení křivky (1) dalo odhad $\tilde{\alpha} = 0{,}6158214 \doteq 35{,}3\,^\circ$ a reziduální součet čtverců 2 485,263. Příslušná regresní hyperbola je znázorněna na obr. 5 tečkovaně jako hyperbola 2. I když (snad) už je to (konečně) dobře, dostali jsme větší reziduální součet čtverců než v předchozích dvou (chybných) modelech.

V diskusním příspěvku nazvaném O kauzálních vztazích a analýze dat k článku Anděl (2003) na str. 18 doc. J. Tvrdík poukázal na to, že jednoduchý

model, který bere v úvahu odpor prostředí, by vedl k regresnímu vztahu

$$y = \beta \left(\sqrt{x^2 \sin^2 2\alpha + 4dx \cos^2 \alpha} - x \sin 2\alpha \right). \tag{4}$$

Odhady parametrů tohoto modelu jsou

$$\tilde{\alpha}^* = 0{,}39125, \qquad \tilde{\beta}^* = 0{,}73240.$$

Parametr $\tilde{\alpha}^*$ odpovídá přibližně 22 stupňům, reziduální součet čtverců je 33,258, koeficient determinace činí $R^2 = 0{,}99957$. Znepokojivé je, že tento model dává vyrovnání dat až příliš přesné. Nedá se očekávat, že by Galileova měření mohla takové přesnosti dosahovat.

Literatura

[1] Anděl J. (1998): Statistické metody (2. vyd). Matfyzpress, Praha.

[2] Anděl J. (2003): Statistické modely. *Statistika* **2**, 1–17, 46–47.

[3] Bickel P. J., Hammel J. W. O'Connell J. W. (1975): Sex bias in graduate admissions. Data from Berkeley. *Science* **187**, 398–403.

[4] Cipra B. (2002): Chibičky. A jak je najít dříve než učitel. Dokořán, Praha 2002. (Překlad z anglického originálu Misteaks: ... and How to Find Them Before the Teacher Does...)

[5] Drake S., MacLachtan J. (1975): Galileo's discovery of the parabolic trajectory. *Scientific American* **232**, 102–110.

[6] Faraway J. J. (2000): Practical Regression and ANOVA using R. (PDF file.)

[7] Hilton P., Holton D., Pedersen J. (2002): Mathematical Vistas From a Room With Many Windows. Springer-Verlag, New York, Berlin, Heidelberg. ISBN 0-387-95064-8, SPIN 10770592.

[8] Ramsey F. L., Schafer D. W. (1997): The Statistical Sleuth. A Course in Methods of Data Analysis. Duxbury Press, Belmont.

Poděkování: Práce na tomto článku byla podpořena výzkumným záměrem MSM 0021620839 Metody moderní matematiky a jejich aplikace.

Adresa: Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
katedra pravděpodobnosti a matematické statistiky

HLEDÁNÍ STŘEDU KULEK

Pavel Stříž

1. Úvod

Na závěr Robustu 2004 bylo možné vyslechnout přednášku pana Jana Kaliny o tom, jak automaticky nalézt v obličeji jednotlivé části. Hlavním zájmem byly oči a vyhledávání pomocí vzorů očí. Snahou bylo nalézt oči i v pootočené poloze. Po identifikaci očí se mnohem lépe hledaly další části obličeje.

Nad tímto problémem jsem se také nejednou zamýšlel z pohledu automatického zaměřování zbraní na předem daný objekt. Jednou z dalších věcí, při které jsem zkoumal toto automatické vyhledávání, bylo hledání středu kulky v terči a jejich automatické vyhodnocování v sérii ran. Je to samozřejmě velmi efektivní u nástřelu zbraní. Po jistých úpravách je možné takto automaticky vyhodnocovat i turnajovou střelbu.

2. Hledání symetrie

Nastíním řešení těchto problémů pomocí hledání symetrie objektu. Obrázek pak překlápím na tři části jako obálku a odpovídající si sloupce pixelů srovnávám pomocí škály šedé barvy. Barevný obrázek si téměř v libovolném grafickém editoru převádím na obrázek ve stupních šedi (grayscale).

Na vlastní výpočty používám `PHP` (3i), který je šířen pod licencí `GNU` (a je spouštěn z prostředí serveru; např. pod `Apachem` (2i)). Pak .php soubor spouštíme ze serveru a výsledky jsou okamžitě k dispozici, zveřejněné na Internetu na příslušné www stránce.

Zda-li máme `PHP` k použití zjistíme při spuštění .php souboru, ve kterém máme např.:

```
<br>Funguje PHP? <?php echo "Ano funguje."; ?>
```

Pokud `PHP` funguje, tak se vypíše „Ano funguje.“, jinak se nevypíše za otázkou nic (o instalaci a technických detailech více v (1i)). Pak si lze ověřit, zda-li můžeme používat grafické formáty při práci s `PHP`:

```
<?php
if (ImageTypes() & IMG_GIF) {echo "GIF lze použít.<br>";}
if (ImageTypes() & IMG_JPG) {echo "JPG lze použít.<br>";}
if (ImageTypes() & IMG_PNG) {echo "PNG lze použít.<br>";}
if (ImageTypes() & IMG_WBMP){echo "WBMP lze použít.";}
?>
```

Bohatě nám stačí jen jeden formát. Skoro v jakémkoliv grafickém editoru případně zkonvertujeme. Připravíme si .php soubor a ten v závěru spustíme. Zavedeme PHP a otevřeme si obrázek $(_1)$, zjistíme rozměry obrázku $(_2)$, počet barev $(_3)$ a nejhorší možnou hodnotu při hledání symetrie $(_4)$ (nejlepší hodnota je 0). Hodnoty si necháme následně vypsat $(_5)$.

```
1  <?php $im=ImageCreateFromPng("picture.png");
2  $iks=imagesx($im); $yps=imagesy($im); $pul=$iks/2;
3  $barev=imagecolorstotal($im);
4  $barev=$barev-1; $max=$barev*$pul*$yps; $min=$max;
5  echo "Velikost obrázku je ".$iks." x ".$yps.", barev je "
       .++$barev." s maximální ztrátou ".$max.".<br><br>";
```

Nyní již budeme počítat odchylky barev při překlopení. Na řádku $(_6)$ začíná a na řádku $(_{23})$ končí cyklus prohledávající veškerá zalomení. Levá část zalomení se zkoumá mezi řádky $(_7)$ až $(_{12})$. Pravá část se počítá v řádcích $(_{13})$ až $(_{19})$. Dále se jen testuje a hledá nejmenší součet barevných odlišností příslušných pixelů.

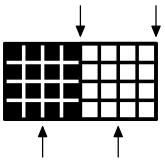
```
6  for ($delka=1;$delka<=$pul;$delka++): $citac=0;
7  for ($i=0;$i<=$delka-1;$i++){
8  for ($j=0;$j<=$yps-1;$j++){
9  $k=2*$delka-$i-1;
10 $citac=$citac+abs(imagecolorat($im,$i,$j)-
       imagecolorat($im,$k,$j));
11           }
12     }
13 for ($i=$delka;$i<=$pul-1;$i++){
14 for ($j=0;$j<=$yps-1;$j++){
15 $k=2*$delka+$i-$delka;
16 $k2=$iks+$delka-$i-1;
17 $citac=$citac+abs(imagecolorat($im,$k,$j)-
       imagecolorat($im,$k2,$j));
18           }
19     }
20 $doc=$delka+$pul;
21 echo "Zalomeno po ".$delka." a ".$doc." s ".$citac."<br>";
22 if ($citac<$min) {$min=$citac; $pole=$delka;}
23 endfor
```

Nyní již víme vše důležité a necháme si vypsat závěry v řádcích $(_{24})$ až $(_{27})$, kde také ukončíme PHP skript.

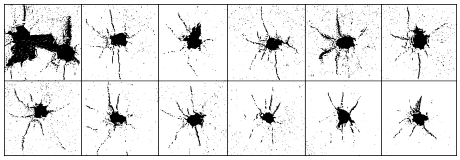
```
24 echo "<br>Zalomení po ".$pole." a ";
25 $doc=$pole+$pul; echo $doc;
26 echo " s účelovou funkcí ".$min.".<br><br>";
27 ?>
```

Takto připravený soubor vypíše výstup na obrazovku.

Pro obrázek se 4 řádky a 8 sloupci, pokud je levá polovina černou barvou (0) a pravá polovina bílou barvou (255) (obr. 1a), dostáváme zalomení po sloupcích 2 a 6 s účelovou funkcí 0. Nejhorší možnost nastává při zalomení po sloupcích 4 a 8 (nebo chceme-li sloupec 0), kdy všechny bílé se odečítají s černými (účelová funkce 4080). Velmi podobně bychom postupovali při postupném natáčení obrázku. Při otočení o 90 je symetrie v ose při všech zalomeních. Vše si lze nechat vykreslit pomocí PHP, ale to zde neřeším a neuvádím.



Obrázek 1a.
Ne/symetrie v ose.



Obrázek 1b.
Ukázky děr v terči po kulkách.

Problém však nastává při hledání středu po zásahu v terči (obr. 1b). Může se stát, že i při vrácení roztrženého papíru terče do původní polohy kousky terče odpadly. Dále je zajímavé hledat středy střel, pokud jsou blízko sebe atd. Pak jedna z možností je hledat symetrii s postupnou úpravou obrázku. To je však nad rámec tohoto článku.

3. Závěr

Příspěvek měl uvést jeden z reálných problémů a jak by se snad mohlo jít na jeho zautomatizované řešení. Problém je to netriviální a zasluhoval by si obšírnější studium teoretické literatury a praktickou verifikaci v praxi.

Zdroje Internet: Odkazy byly funkční k 1. dubnu 2005.

(1i) http://cz.php.net/tut.php
(2i) http://httpd.apache.org/
(3i) http://www.php.net/

INFORMACE O NOVÉM TEXOVÉM STYLU PRO INFORMAČNÍ BULLETIN

Jména autorů

1. Šablona – Název příspěvku a jména autorů

Příkaz `\title{Informace o novém \TeX{}ovém stylu\\ pro Informační Bulletin}` nebo
`\nazev{Informace o novém \TeX{}ovém stylu\\ pro Informační Bulletin}` byl použit k vysázení názvu příspěvku.

Jména autorů lze vysázet:
`\author{Jméno autora}, \authors{Jména autorů},`
`\autor{Jméno autora}` nebo pomocí `\autori{Jména autorů}.`

2. Šablona – ukázka \section nebo \sekce

První odstavec autor nemusí upravovat pomocí `\noindent`, protože používáme standard CSLaTeXu, který to již automaticky zajišťuje.

Zde následuje druhý odstavec, který je již odsazený.

Tento nadpis byl vysázen:
`\section{Šablona -- ukázka $\backslash$section nebo $\backslash$sekce}`

2.1. Šablona – ukázka \subsection nebo \podsekce

Opět první odstavec není zbytné odsazovat pomocí `\noindent`.

Tento nadpis byl vysázen:
`\subsection{Šablona -- ukázka $\backslash$subsection nebo $\backslash$podsekce}`

2.1.1. Šablona – ukázka \subsubsection nebo \podpodsekce

Můžeme použít až tři úrovně nadpisů. Doporučujeme zůstat jen u dvou úrovní.

Zmíníme další příkazy, které by měly sjednotit příspěvky (pokud si je autor přeje vysázet). Všechny zmíněné příkazy mají na vstupu jen jeden argument, např. `\email{honza.pesek@mujweb.cz}`:

`\abstract` nebo `\abstrakt` Používáme k vysázení abstraktu.
`\keywords` nebo `\klicovaslova` Používáme k vysázení klíčových slov.
`\thanks` nebo `\podekovani` Příkaz určený k poděkování.
`\address` nebo `\adresa` Příkaz použijeme pokud uvádíme adresu.

`\email`	Vysází emaily autorů.
`\forget` nebo `\zapomen`	Svůj argument nevysází do příspěvku.

2.1.2. Drobné příkazy Příkaz `\refname` vysází „Literatura“.
Příkaz `\registered` vysází ®.

3. Odevzdání příspěvku

Příspěvky prosím posílejte elektronicky na adresu:

`gejza.dohnal@fs.cvut.cz`

Alternativně lze poslat příspěvek na disketě nebo CD-ROM na adresu uvedenou v patičce Bulletinu.

Pokud je to možné, tak příspěvky posílejte napsané v TEXu. Z jiných typografických systémů (Writer OpenOffice.org, QuarkXPress, InDesign atp.) bude příspěvek do TEXu převeden. Bulletin je sázen pomocí formátu LATEX 2_ε.

4. Finální verze

Na Internetu se zveřejňuje `PDF` verze ve velikosti A4.
Předtisková příprava bulletinu ještě pokračuje a provádí se takto.

Bulletin jako A5 knížečka za pomocí nařezání papírů (např. IB 2-3/2004):

```
psbook.exe -s4 bull_old.ps > bull_tmp.ps
pstops 2:1L(790,400)+L(790,-25) bull_tmp.ps > bull_new.ps
```

Bulletin jako skládaná A5 brožurka (např. IB číslo 4/2004):

```
psbook.exe -s20 bull_old.ps > bull_tmp.ps
pstops 2:1L(790,400)+L(790,-25) bull_tmp.ps > bull_new.ps
```

Obecně se za parametr `-s` dává počet stránek bulletinu.

Doporučujeme před odesláním článku využít i této možnosti – ať si můžete sami, přímo a nezávisle zkontrolovat čitelnost článku po vytištění.

5. Ukázka formátování

Tento Informační Bulletin je již vysázen za použití nové šablony.

Ať se Vám články dobře píší!

Šablona Informačního Bulletinu

Šablona je připravena tak, aby autor příspěvku mohl i nadále psát běžnými příkazy TEXu nebo jejich počeštěnou variantou.

Výstup může autor vizuálně zkontrolovat stejně jako sazeč nebo tiskař Informačního Bulletinu.

Šablona je dostupná na `http://www.statspol.cz/`

Příspěvky do Informačního Bulletinu

Redakce se dle svých sil snaží i o převody článků a pozvánek z jiných typografických systémů. V takovém případě je však téměř nemožné zajistit identický převod do TEXu. Pak může být nespokojen sazeč, že se identický převod nepovedl, a autoři příspěvku mohou být znepokojeni s vysázeným příspěvkem.

Apelujeme a doporučujeme vše sázet pomocí TEXu. TEXisti, kterých není málo, jistě rádi pomohou nebo se poraďte s redakcí.

Těšíme se na Vaše příspěvky

redakce

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. ISSN 1210 – 8022.
Předseda společnosti: Prof. RNDr. Jaromír Antoch, CSc., KPMS MFF UK Praha, Sokolovská 83, 186 75 Praha 8, e-mail: `jaromir.antoch@mff.cuni.cz`
Redakce: Doc. RNDr. Gejza Dohnal, CSc. a Mgr. Pavel Stříž;
e-mail: `gejza.dohnal@fs.cvut.cz` a `striz@fame.utb.cz`