Informační Bulletin

České Statistické Společnosti *číslo 1, ročník 19, únor 2008*

Zpráva o činnosti České statistické společnosti v roce 2007, která byla přednesená a projednaná na výroční schůzi společnosti dne 31. 1. 2008.

1. **Základní údaje o společnosti.** Uplynulý rok byl prvním rokem dvouletého funkčního období výboru České statistické společnosti (ČStS), který byl zvolen na valné hromadě dne 8. 2. 2007. Předsedou byl Doc. RNDr. Gejza Dohnal, CSc. (FS ČVUT v Praze), místopředsedou Ing. Jan Fischer, CSc. (ČSÚ) a hospodářkou doc. Ing. Dagmar Blatná, CSc. (VŠE Praha). K dnešnímu dni má ČStS 234 členů, z toho 17 vstoupilo do společnosti v roce 2007 a 3 v roce 2008. V roce 2007 ukončili 2 členové členství na vlastní žádost, 1 zemřel. U dalších 2 bylo členství ukončeno pro neplacení členských příspěvků. Na vyřazení kvůli neplacení je nyní 10 kandidátů (kteří nezaplatili za 2005, 2006 a 2007).
2. **Činnost výboru společnosti.** V průběhu roku se konala tři zasedání výboru České statistické společnosti. O každém z nich byl pořízen zápis, který je všem zájemcům k dispozici. V mezidobí byli členové výboru v kontaktu prostřednictvím e-mailu a diskutovali všechny důležité záležitosti, zejména přípravu akcí a bulletinů. Kromě toho proběhla řada neformálních setkání a porad při jednotlivých akcích. Při příležitosti společné konference STAKAN se Slovenskou štatistickou a demografickou spoločnosťou proběhlo společné jednání členů výborů obou společností. 22. – 29. 8. 2007 se v Lisabonu konal 56. kongres ISI, kterého se zúčastnilo několik členů výboru (Antoch, Bartošová, Blatná, Fischer, Löster, Picek, Řezanková). Jednu se sekcí, kde jsme se účastnili, organizovala Viszegradská skupina národních statistických společností (Maďarsko, Rakousko, Česko, Slovensko, Slovinsko a Rumunsko). Předseda společnosti se zúčastnil 3. setkání předsedů národních statistických společností této skupiny ve Slovinské Ljubljani.

3. **Odborná aktivita společnosti.** Valná hromada v roce 2007 se konala v Praze dne 8. února 2007 v zasedací síni ČSÚ. Na valné hromadě přednesl odbornou přednášku předseda ČSÚ Ing. Jan Fischer, CSc. na téma Problémy statistické služby. Zabýval se v ní problematikou práce na ČSÚ a aspekty, které přináší současná doba a technika nejen v ČR, ale i v mezinárodním kontextu. Společnost se podílela na organizaci konference Centra pro jakost a spolehlivost výroby REQUEST v Praze ve dnech 30. 1. – 1. 2. 2007 Česká statistická společnost a Slovenská statistická a demografická společnost uspořádaly společně v květnu (25. – 27. 5.) v Rusavě v Hostýnských vrších odborný seminář o výuce a aplikacích statistiky STAKAN 2007. Sborník z této konference vyšel jako zvláštní číslo Forum Statisticum Slovacum na podzim spolu s DVD. ČStS převzala záštitu nad konferencí TIES'2007, jež se konala 16. – 20. 8. 2007 v Mikulově. 6. 12. se v Balbínově poetické hospůdce v Praze konal Mikulášský statistický den, kde zaznělo celkem osm příspěvků. Vedle konferencí a seminářů je třeba zmínit tyto další odborné aktivity: Česká statistická společnost se stala signatářem deklarace ke vzniku oborového seskupení Jakost a spolehlivost v rámci připravované České technologické platformy Strojírenství. V roce 2007 byla vydána čtyři čísla Informačního bulletinu a dvě DVD (STAKAN a GISAK) Internetové stránky společnosti byly pravidelně udržovány a aktualizovány. ČStS spolupracovala na vydávání časopisu Statistika.
4. **Plán aktivit pro rok 2008.** V dubnu se v Liberci uskuteční další, tentokrát dvoudenní statistické dny V červnu 2008 proběhne v Praze mezinárodní symposium ISBIS 2008 věnované ekonomické a průmyslové statistice, na jehož organizaci se naše společnost podílí (členové ČStS mají slevu na vložném) V létě se bude ČStS podílet na organizaci konference o jakosti a spolehlivosti výroby v Brně, jejímž hlavním organizátorem bude CQR 5. – 7. 9. 2008 bude naše společnost organizovat v Praze mezinárodní studentskou statistickou konferenci, spojenou se 4. setkáním předsedů národních statistických společností. 8. – 12. 9. 2008 se bude konat další ROBUST, tentokrát ve spolupráci se Slovenskou statistickou a demografickou společností.

BLAHOPŘÁNÍ

V těchto dnech se dožívá významného životního jubilea náš člen a kolega, doc. RNDr. Karel Zvára, CSc., významný odborník v oblasti regrese a aplikované statistiky. Kolega Zvára věnoval převážnou část svého života výuce statistiky, především pro nestatistiky, jakož i aplikacím statistiky v přírodovědě a medicíně. Výbor ČStS, johož byl kolega Zvára po řadu let členem, mu přeje mnoho zdraví a spokojenosti v dalším životě.

NĚKOLIK SLOV O RELIABILITĚ SLOŽENÝCH DICHOTOMNÍCH MĚŘENÍ

ON RELIABILITY OF COMPOSED DICHOTOMOUS MEASUREMENTS

aneb doktorandkou pana docenta Zváry

Patrícia Martinková

Adresa: EuroMISE centrum UK a AV ČR, ÚI AV ČR, Praha

E-mail: martinkova@euromise.cz

Abstract This remark concentrates on generalization of popular Cronbach alpha for the case when the measurements are dichotomous. Main result is a new definition of reliability for this type of measurements based on conditional expectation and conditional variance.

V jednom z předchozích čísel Informačního Bulletinu (viz [1]) pojednal pan docent Zvára o reliabilitě měření a o Cronbachově alfa, které se k jejímu odhadu často používá. V závěru článku nastínil otázku, zda máme právo použít postup založený na představě o spojitých veličinách i v případě, kdy položky složeného měření jsou výhradně nulajedničkové. V takovém případě autor navrhl nahradit Cronbachovo alfa, jehož odhad lze ve smíšeném modelu analýzy rozptylu vyjádřit pomocí testové statistiky F, jeho obdobou z logistické regrese, využívající testovou statistiku jinak sloužící k testování analogické hypotézy.

Za dobu posledních čtyř let jsem měla tu čest pod vedením pana docenta Zváry bádat právě nad definováním a odhadováním reliability v případě složených dichotomních měření.

Dovolte mi tu zmínit některé výsledky tohoto bádání. Za hlavní výsledek práce považuji navržení obecnější definice reliability pomocí podmíněné střední hodnoty a podmíněného rozptylu

$$\mathrm{rel}(Y) = \frac{\mathrm{var}\left[\mathrm{E}(Y|A)\right]}{\mathrm{var}\left[\mathrm{E}(Y|A)\right] + \mathrm{E}\left[\mathrm{var}(Y|A)\right]} = \frac{\mathrm{var}\left[\mathrm{E}(Y|A)\right]}{\mathrm{var}(Y)}. \qquad (1)$$

Nová definice, stejně jako ta klasická, vyjadřuje relativní díl celkové variability měření Y způsobený variabilitou měřené vlastnosti A. V případě smíšeného modelu analýzy rozptylu obě definice splývají. Navíc však novou definici využijeme u modelů, v nichž nevystupuje chyba měření. Takovým

modelem je i Raschův model, běžně používaný pro popis vlastností didaktických testů s nulajedničkovými položkami. Díky tomu, že se nám podařilo vyjádřit reliabilitu v Raschově modelu a dalších modelech vhodných pro popis složených dichotomních měření, bylo pak možné, zatím alespoň pomocí simulací, posoudit použitelnost odhadu navrženého v [1]. Zdá se, že v některých případech nově navržené logistické alfa odhaduje reliabilitu lépe než alfa Cronbachovo. Výsledky byly publikovány v článku [2].

Výrazem (1) navazujeme na práci [3], jejíž tvrzení o ekvivalentní definici pro modely se společným koeficientem vnitrotřídní korelace se nám podařilo uvést na pravou míru – najít protipříklady a dokázat tvrzení správné. Podařilo se díky tomu také nahradit požadavky klasické τ-ekvivalence tak, že Spearmanova-Brownova formule pro reliabilitu měření složeného z m položek zůstává i nadále v platnosti.

Postgraduální studium pod vedením pana docenta pro mne bylo velice přínosné. Školitel se mi stal velkým vzorem nejen jako vědec, ale také jako pedagog s výtečně propracovanou přípravou (jak pro studenty tak pro své cvičící), jako praktický statistik s mnoha zkušenostmi a v neposlední řadě jako nesmírně schopný, ochotný a férový člověk. Cením si všech těch mnoha hodin konzultací o to víc, že mi byly věnovány nesmírně vytíženým člověkem. Bylo mi až s podivem, kolik různých činností pan docent zvládá. Jednou přicházel s rolí papírů pod paží se slovy „Projekt rekonstrukce v Karlíně, Ferda Mravenec, práce všeho druhu!“ Jindy zase překládal na stole své pracovny štosy s různými úkoly „Tak kde Vás mám!“ Snad pro to velké pracovní vytížení, snad pro pocit, že to bádání je až příliš aplikované, jsem občas mohla slyšet „To víte, já mnoho doktorandů nevedl.“ Myslím si, že to je velká škoda. A přála bych ještě alespoň jednomu doktorandovi tohoto výtečného školitele.

Nezbývá mi než závěrem tohoto příspěvku poděkovat panu docentovi za všechen čas, který mi věnoval, i za trpělivost, kterou se mnou měl během celého mého studia, a popřát oslavenci mnoho zdraví, štěstí, a spokojenosti do dalších let.

Reference

[1] Zvára K. (2003) Reliabilita měření aneb bacha na Cronbacha. *Informační bulletin České Statistické Společnosti* **13(2)**, 13–20.

[2] Martinková P., Zvára K. (2007) Reliability in the Rasch model. *Kybernetika* **43(3)**, 315–326.

[3] Commenges D., Jacqmin H. (1994): The intraclass correlation coefficient distribution-free definition and test. *Biometrics* **50(2)**, 517–526.

VOLBA REGRESNÍHO MODELU
HOW TO CHOSE REGRESSION MODEL

Jiří Anděl
Adresa: MFF UK, KPMS, Praha
E-mail: jiri.andel@mff.cuni.cz

Abstract This contribution concentrates on typical errors connected with the choice of the regression model. Most frapant erros are illustrated using two examples. The first one shows influence of the graphical representation of the data. The second one shows how important is not to neglect the additional information about the data and their genesis. All calculations were done using the program R.

1. Úvod

V tomto příspěvku je pojednáno o chybách, které se dělají při volbě regresního modelu. Tyto chyby jsou ilustrovány na dvou numerických příkladech. V prvním z nich se posuzuje vliv grafického znázornění dat na konstrukci modelu. Ve druhém příkladě je poukázáno na důležitost využití dodatečné informace o datech. Výpočty jsou prováděny pomocí programu `R`, který lze získat na adrese `http://www.R-project.org/`.

2. Volba modelu založená na grafickém znázornění dat

Grafy ve statistice hrají velmi důležitou úlohu. Everitt (2005) na str. 16 cituje výrok převzatý z publikace Chambers a kol. (1983): „. . . there is no statistical tool that is as powerful as a well-chosen graph“[1]. Zdůrazněme, že mezi autory posledně citované knihy jsou tak slavní statistici jako je Cleveland či Tukey. Odhaduje se, že se ročně tiskne asi 10^{12} statistických grafů. Jedním z důvodů grafického znázornění dat je to, že je člověk schopen vyčíst z nich zákonitosti. Platí však varování Carla Sagana: „Humans are good at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent.“[2]

V tabulce 1 jsou uvedena data, která budeme analyzovat. Původ a skutečný mechanismus vzniku těchto dat je znám a bude uveden později pro porovnání s dosaženými výsledky. Ostatně i kdyby například výzkumník sdělil, že

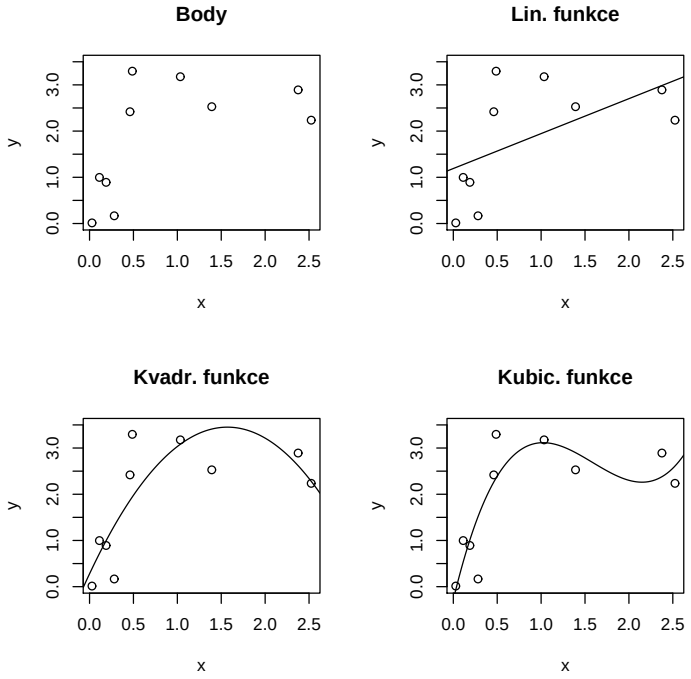
[1] Žádný jiný statistický nástroj není tak mocný jako správně zvolený graf.
[2] Lidé dobře dokáží rozeznávat subtilní zákonitosti, které tam opravdu jsou, ale zrovna tak dobře si je dokáží představit, i když tam vůbec nejsou.

třeba nezávisle proměnná udává koncentraci hexametyléntetramínu a závisle proměnná koncentraci pentaerytritolu, asi by to většině z nás nepřineslo víc informace než to, že x_i jsou hodnoty nezávisle proměnné a y_i jsou hodnoty závisle proměnné.

Poznamenejme, že v tabulce 1 jsou uvedeny zaokrouhlené hodnoty. Další výpočty byly provedeny s původními daty, která byla prezentována na víc desetinných míst.

i	1	2	3	4	5	6	7	8	9	10
x_i	2.38	1.03	0.19	0.49	2.52	0.11	0.46	0.28	1.39	0.03
y_i	2.89	3.18	0.89	3.30	2.24	1.00	2.42	0.17	2.53	0.01

Tab. 1. Data, která je třeba statisticky analyzovat.

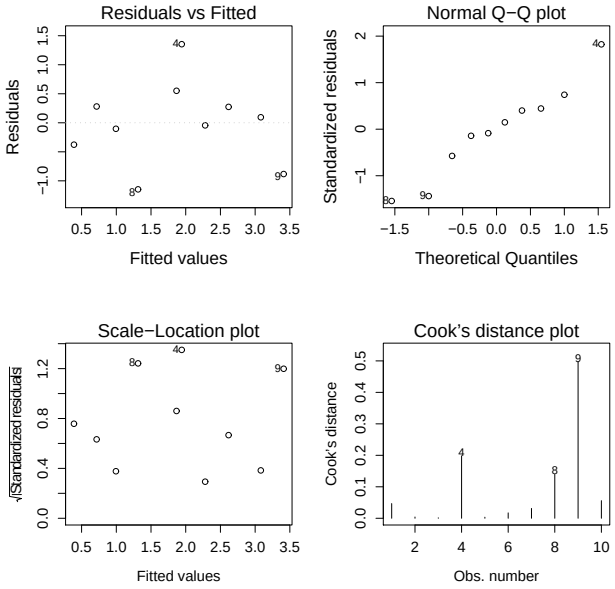


Obr. 1. Data a regresní funkce.

Tato data jsou znázorněna na obrázku 1, kde jsou také prezentovány grafy některých regresních funkcí. Výsledky, které se týkají výpočtu regresní přímky, jsou:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1900     0.4855   2.451   0.0399 *
x             0.7558     0.3886   1.945   0.0877 .

Residual standard error: 1.079 on 8 degrees of freedom
Multiple R-Squared: 0.321,      Adjusted R-squared: 0.2361
F-statistic: 3.782 on 1 and 8 DF,  p-value: 0.08772
```



Obr. 2. Diagnostické grafy ke kvadratické regresi.

Regresní koeficient sice není statisticky signifikantní na běžné hladině 5 %, protože jeho p-hodnota je 0.088, ale data spíš odpovídají kvadratické regresi. Výsledky kvadratické regrese jsou

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2786     0.4948   0.563   0.5910
x             4.0411     1.2475   3.239   0.0143 *
I(x^2)       -1.2865     0.4751  -2.708   0.0303 *

Residual standard error: 0.8062 on 7 degrees of freedom
Multiple R-Squared: 0.6684,     Adjusted R-squared: 0.5736
F-statistic: 7.054 on 2 and 7 DF,  p-value: 0.02100
```

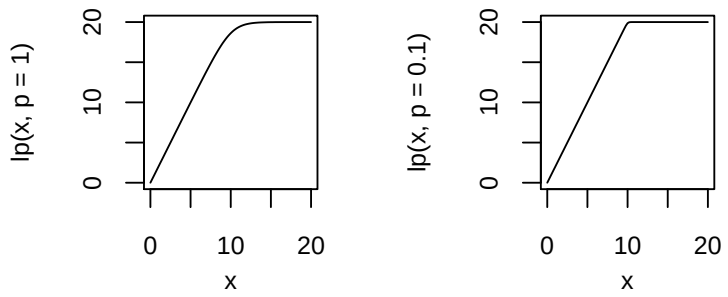
Zde je na hladině 5 % významný lineární i kvadratický člen, takže bychom se mohli přiklonit k tomu, že datům odpovídá kvadratická regrese. Poznamenejme, že v případě kubické regrese bude signifikantní jen lineární člen, kdežto ani kvadratický ani kubický člen signifikantní nebudou. To nadále svědčí ve prospěch kvadratické regrese. Přidáme ještě diagnostické grafy (viz obrázek 2), jež nasvědčují tomu, že regresní model odpovídá datům.

Na druhé straně však kvadratická funkce uvedená na obrázku 1 není monotónní. Pokud bychom věděli, že má jít např. o růstovou křivku, monotónie by se měla nutně vyžadovat. Pro ilustraci zde uvedeme jednu málo známou *růstovou křivku*

$$lp(x, a, b, p, c) = a - bp \ln\left[1 + \exp\left\{\frac{c - x}{p}\right\}\right],$$

která se nazývá *linear-plateau regression function* (česky by se snad mohlo říci *lineární regresní funkce se stabilní hladinou*). Graf této funkce připomíná dvě navazující přímky. Jedna z nich je rostoucí a druhá konstantní. Parametry této regresní funkce mají následující interpretaci

- a ... hodnota závisle proměnné v bodě změny,
- b ... směrnice rostoucí přímky,
- c ... hodnota nezávisle proměnné v bodě změny,
- p ... hladkost přechodu mezi oběma přímkami.

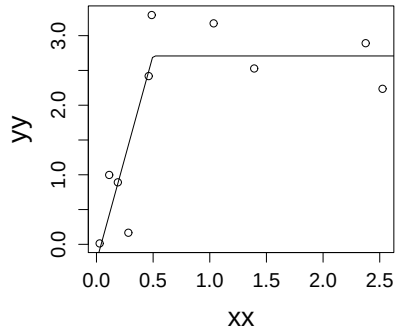


Obr. 3. Průběh $lp(x)$ s parametry a=20, b=2, p=1, c=10 (vlevo) a s parametry a=20, b=2, p=0.1, c=10 (vpravo).

Proložení této funkce našimi daty vedlo k výsledku

```
     a       b        p     c
 2.708  5.9282 0.000694 0.498
```

Body a proložená funkce jsou zobrazeny na obrázku 4.



Obr. 4. Funkce $lp(x)$ proložená metodou nejmenších čtverců.

Je však na čase uvést, jak byla výchozí data získána. Byla generována na počítači jako nezávislé náhodné veličiny. Přitom $x_i \sim \mathsf{N}(1,1)$, $y_i \sim \mathsf{N}(2,1)$. Nastavení generátoru náhodných čísel pomocí příkazu `set.seed(1203)` bylo provedeno z toho důvodu, že tuto konstantu používá ve svých ilustracích Everitt (2005). Proč tedy došlo k tak signifikantnímu prokázání nesprávného modelu? Důvodem může být některá z následujících příčin.

- Generátor není dostatečně kvalitní.
- Při statistickém hodnocení se pracuje s určitou hodnotou pravděpodobnosti chyby prvního druhu, nejčastěji to je 0.05. Počítá se tedy s tím, že zhruba jednou ve dvaceti případech vyjde signifikantně výsledek, který by ve skutečnosti signifikantní být neměl.

Generátory náhodných čísel bývají podrobně testovány. Použitá část generátoru byla v literatuře, jak již bylo výše zmíněno, mnohokrát použita. Pokud jde o druhý argument, je dobré připomenout, že se dosažené p-hodnoty blíží hladině 0.01.

Získaný signifikantní výsledek je nejspíš výsledkem toho, že jsme si hypotézu vytvořili teprve na základě získaných dat. To je principem činnosti nazývané „data mining“. Ta vede k vytváření hypotéz o modelu, kterým se data řídí. Statistické ověření modelu se však musí provádět na zcela nových datech. Pořizování dat bývá v experimentálních vědách nákladné a časově náročné. Z tohoto důvodu se někdy obě fáze rozboru, tedy jak „data mining“, tak i statistické ověřování, provádějí na témž souboru dat. Tím se snadno mohou prokázat zákonitosti, které vůbec neexistují. To jsme na výše uvedeném umělém příkladě předvedli.

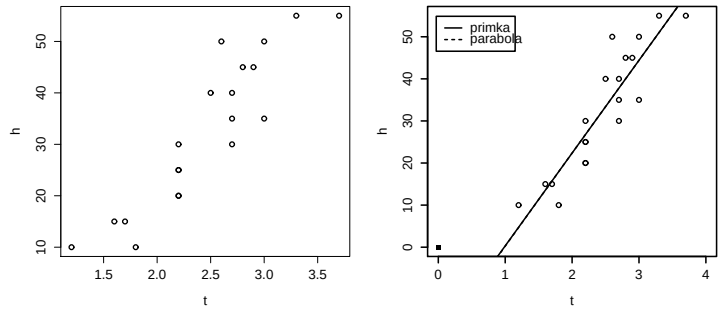
3. Upřesnění modelu pomocí dodatečné informace

Van Belle (2002) uvádí následující statistický příklad. Představme si, jak Galileo zkoumá vztah času a délky volného pádu. Z určité výšky h na věži v Pise pouští těžkou dělovou kouli a zjišťuje dobu t, po kterou koule padá k zemi. Výsledky jsou uvedeny v tabulce 2.

h	t	h	t	h	t	h	t
10	1.8	20	2.2	35	2.7	45	2.8
10	1.2	25	2.2	35	3.0	50	2.6
15	1.6	25	2.2	40	2.7	50	3.0
15	1.7	30	2.7	40	2.5	55	3.3
20	2.2	30	2.2	45	2.9	55	3.7

Tab. 2. Výška h v metrech a doba pádu t v sekundách.

Analyzujme nejprve závislost h na t běžnými regresními metodami, aniž bychom brali v úvahu znalost vzorců pro volný pád nebo některé další informace.



Obr. 5. Galileova data (vlevo) a lineární a kvadratická regrese (vpravo).

Pokud proložíme regresní přímku, dostaneme

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -21.734      6.040  -3.598  0.00206 **
          t   22.046      2.387   9.235 2.99e-08 ***
Residual standard error: 6.32 on 18 degrees of freedom
Multiple R-Squared: 0.8257, Adjusted R-squared: 0.8161
F-statistic: 85.29 on 1 and 18 DF,  p-value: 2.995e-08
```

Výsledkem je rovnice regresní přímky $s = -21.734 + 22.046t$, která je znázorněna na obrázku 5 vpravo. Oba parametry regresní přímky jsou signifikantní.

Provedeme Durbinův-Watsonův test a dostaneme

```
 lag Autocorrelation D-W Statistic p-value
   1          0.304         1.271   0.046
Alternative hypothesis: rho != 0
```

Výsledek je signifikantní, což signalizuje porušení předpokladů regresní analýzy. Proložíme kvadratickou regresní funkci a dostaneme

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -21.967     18.596  -1.181    0.254
t             22.251     15.558   1.430    0.171
I(t^2)        -0.042      3.165  -0.013    0.990

Residual standard error: 6.503 on 17 degrees of freedom
Multiple R-Squared: 0.8257, Adjusted R-squared: 0.8052
F-statistic: 40.28 on 2 and 17 DF,  p-value: 3.551e-07
```

Získali jsme kvadratickou regresi $s = -21.9668 + 22.2505t - 0.0421t^2$. Její graf je znázorněn na obrázku 5 vpravo a prakticky se překrývá s grafem regresní přímky. To je zřejmé i z porovnání parametrů. Je překvapující, že žádný z parametrů kvadratické regresní funkce není signifikantní. Pro kontrolu provedeme opět Durbinův-Watsonův test s výsledkem

```
lag Autocorrelation D-W Statistic p-value
  1          0.3043         1.272   0.026
Alternative hypothesis: rho != 0
```

Test vyšel signifikantně, což se rovněž dalo čekat vzhledem k tomu, že díky shodě regresní přímky a regresní kvadratické funkce jsou rezidua v obou případech prakticky stejná.

Nyní vezmeme v úvahu, že za nulový čas musí být dráha volného pádu také rovna nule. Proto znázorníme Galileova data i se zdůrazněným bodem (0,0), kterým musí každá regresní funkce procházet (viz obrázek 5 vpravo a obrázek 6 vlevo). Nejdřív zase proložíme přímku a máme

```
  Estimate Std. Error t value Pr(>|t|)
t   13.695      0.713   19.21  6.6e-14 ***

Residual standard error: 8.065 on 19 degrees of freedom
Multiple R-Squared: 0.9511, Adjusted R-squared: 0.9485
F-statistic: 369.2 on 1 and 19 DF, p-value: 6.594e-14
```

Rovnice této regresní přímky je $s = 13.6949t$. Pak proložíme regresní kvadratickou funkci bez absolutního členu

```
        Estimate Std. Error t value Pr(>|t|)
t          4.204      2.972   1.414    0.174
I(t^2)     3.482      1.069   3.256    0.004 **
```

Ta má tedy rovnici $s = 4.204t + 3.482t^2$. Koeficient u lineárního členu není signifikantní. Kromě toho dnes již víme, že platí

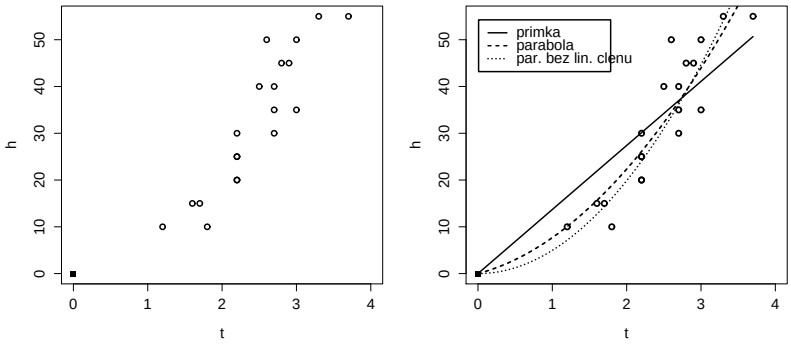
$$s = \frac{g}{2}\, t^2,$$

kde $g = 9.81$ m/sec^2 je zemské zrychlení. Oba důvody vedou k proložení kvadratické regrese bez absolutního i bez lineárního členu. Tím získáme

```
        Estimate Std. Error t value  Pr(>|t|)
I(t^2)     4.966      0.215   23.15  2.19e-15 ***

Residual standard error: 6.745 on 19 degrees of freedom
Multiple R-Squared: 0.9658, Adjusted R-squared: 0.964
F-statistic: 536.1 on 1 and 19 DF, p-value: 2.191e-15
```

Tím jsme dostali rovnici $s = 4.9656t^2$. Tu můžeme porovnat s teoretickou závislostí, která zní $s = 4.905t^2$. Všechny tři poslední regresní funkce jsou uvedeny na obrázku 6 vpravo. Interval spolehlivosti s koeficientem spolehlivosti 0.95 pro koeficient u kvadratického členu je [4.516692; 5.414459]. To znamená, že interval spolehlivosti pro g je [9.033; 10.829].



Obr. 6. Galileova data (vlevo) a regrese procházející počátkem (vpravo).

Ve skutečnosti však lze očekávat, že dráha s byla stanovena přesně, zatímco čas t byl stanoven s chybou. Proto by byla na místě závislost

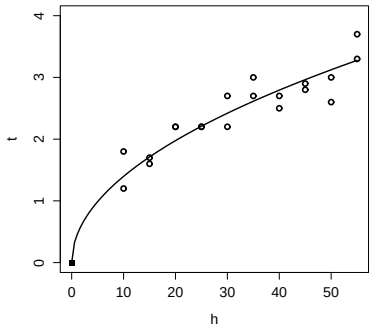
$$t = \sqrt{\tfrac{2}{g}}\sqrt{s}.$$

Pokud proložíme tuto regresní funkci, dostaneme

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
sq 0.441755   0.009832   44.93   <2e-16 ***

Residual standard error: 0.2507 on 19 degrees of freedom
Multiple R-Squared: 0.9907, Adjusted R-squared: 0.9902
F-statistic:  2019 on 1 and 19 DF,  p-value: < 2.2e-16
```

Regresní funkce je znázorněna na obrázku 7. Interval spolehlivosti pro koeficient při $\sqrt{s}$ je $[0.4211758, 0.4623348]$. Skutečná hodnota tohoto koeficientu je $\sqrt{2/g} = 0.4515236$. Z toho, že známe interval spolehlivosti pro $\sqrt{2/g}$, dostaneme, že interval spolehlivosti pro g je $[9.357; 11.275]$.



Obr. 7. Regrese času na vzdálenosti.

Poděkování: Příspěvek vznikl za pomoci grantu MSM 0021620839.

Reference

[1] Belle van G. (2002) *Statistical Rules of Thumb*. Wiley, New York.

[2] Everitt B. (2005) *An R and S-PLUS Companion to Multivariate Analysis*. Springer-Verlag, London.

[3] Chambers J. M., Cleveland W. S., Kleiner B. and Tukey P. A. (1983) *Graphical Methods for Data Analysing*. Belmont, CA, Wadsworth.

PŘIJÍMACÍ ZKOUŠKY NA MFF UK Z MATEMATIKY V ROCE 2007

ENTRY EXAMS FROM MATHEMATICS AT MFF UK IN 2007

Jiří Anděl, Jaromír Antoch

Adresa: MFF UK, KPMS, Praha

E-mail: {jiri.andel,jaromir.antoch}@mff.cuni.cz

Abstract This contribution analyze the results of entry exams from mathematics at the faculty of mathematics and physics of the Charles University of Prague.

1. Zadání

Při přijímacích zkouškách z matematiky na MFF dne 11. června 2007 byla uchazečům zadána písemná práce s následujícími úlohami. Jejich řešení uvádíme v odstavci 3.

1. Určete všechny hodnoty reálného parametru p, pro který má soustava rovnic

$$7x + 3y = p^2 \quad \text{a} \quad 5x + 2y = 20$$

   řešení $x > 0$, $y > 0$. (10 bodů)

2. V oboru reálných čísel $\mathbb{R}$ řešte rovnici

$$2 \cdot \frac{\sin x + \sin 2x}{\cos x + \cos 2x} = 3 \cdot \frac{\sin x}{\cos x}.$$

   (10 bodů)

3. Určete první člen a kvocient geometrické posloupnosti, je-li součet prvních tří členů roven 62 a součet dekadických logaritmů těchto tří členů je roven 3. (15 bodů)

4. Napište rovnice tečen paraboly

$$(y-3)^2 = 16(x+3),$$

   které procházejí bodem $[-3, -1]$. (15 bodů)

Upozornění: U každé úlohy je nutno uvést celý postup řešení, nestačí napsat pouze výsledky[1]. Uvedený počet bodů je maximální počet, který můžete za danou úlohu získat.

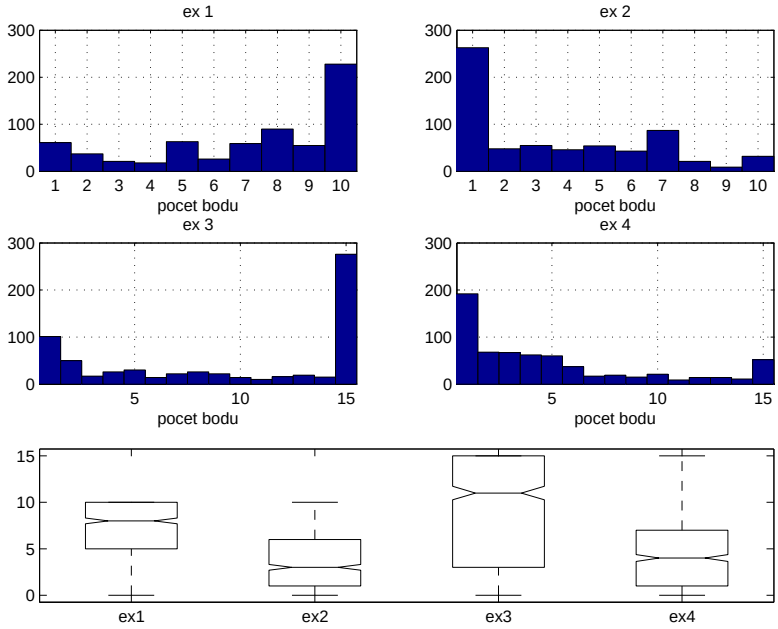
[1] Poznamenejme, že tento požadavek silně omezuje možnost použití programu Mathematica, které je diskutováno v odstavci 4.

2. Hodnocení výsledků

Celkem bylo odevzdáno 658 písemek. Pro stručnost budeme první úlohu označovat jako `ex1`, druhou `ex2` atd.

2.1. Hodnocení jednotlivých úloh

Četnostní histogramy a krabicové grafy výsledků jednotlivých úloh jsou na obrázku 1. Výsledek může čtenáři připadat poněkud zvláštní, zvláště pak ve srovnání s obrázkem 2, letitá zkušenost prvního z autorů však říká, že s podobnými výsledky se setkal prakticky každoročně.



Obr. 1. Četnostní histogramy a krabicové grafy úloh `ex1` – `ex4`.

```
      ex1              ex2              ex3              ex4
 Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
 1st Qu.: 5.00   1st Qu.: 1.00   1st Qu.: 3.00   1st Qu.: 1.00
 Median : 8.00   Median : 3.00   Median :11.00   Median : 4.00
 Mean   : 6.99   Mean   : 3.52   Mean   : 9.32   Mean   : 4.81
 3rd Qu.:10.00   3rd Qu.: 6.00   3rd Qu.:15.00   3rd Qu.: 7.00
 Max.   :10.00   Max.   :10.00   Max.   :15.00   Max.   :15.00
```

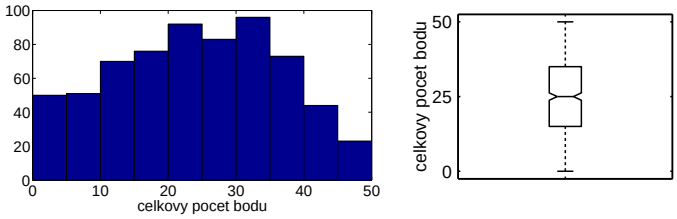
úloha	ex1	ex2	ex3	ex4
průměr	6.99	3.52	9.32	4.81
směr. odchylka	3.25	2.99	5.87	4.65

Tab. 1. Popisné statistické charakteristiky jednotlivých úloh.

2.2. Hodnocení celkového výsledku

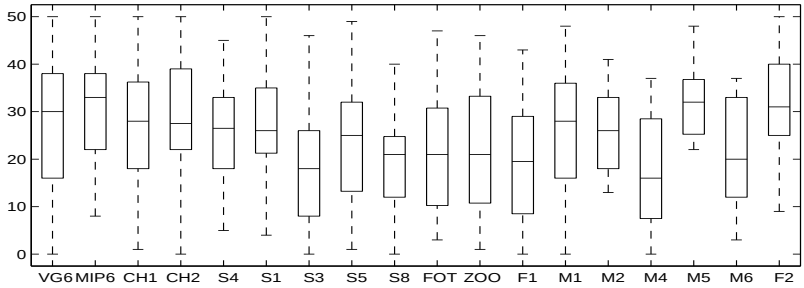
Celkový výsledek přijímací zkoušky z matematiky je dán součtem bodů za všechny čtyři úlohy. Popisné statistické charakteristiky tohoto součtu jsou

```
Min. 1st Qu.  Median    Mean  St. dev.  3rd Qu.    Max.
0.00   15.00   25.00   24.64    12.34     34.75   50.00
```



Obr. 2. Četnostní histogram a krabicový graf celkového výsledku z matematiky.

Průměry a směrodatné odchylky výsledků uváděné s přesností na jedno desetinné místo v jednotlivých posluchárnách jsou uvedeny v tabulce 2. Krabicový graf celkového výsledku v závislosti na umístění uchazečů v posluchárnách je na obrázku 3. Posluchárny jsou seřazeny podle zájmu studentů o určitý typ studia, jejich pořadí je stejné jak na obrázku 3 tak v tabulce 2. Zkratky oborů uvedené v tabulce 2 jsou vysvětleny v tabulce 3.



Obr. 3. Krabicový graf celkových výsledků v závislosti na posluchárnách.

posluchárna	pozváni	přišli	průměr	sd	obor
VG6	84	61	26.7	12.71	BM - P
MIP6	33	26	30.0	11.71	BM - P
Ch1	97	69	27.9	11.04	BM - P
Ch2	27	16	28.2	13.14	BM - P
S4	44	22	25.1	10.84	BM - K
S1	25	19	27.9	10.81	BI - P
S3	70	50	19.8	12.40	BI - P
S5	72	55	23.6	12.59	BI - P
S8	31	19	19.1	9.73	BI - P
FOT	83	63	20.9	11.45	BI - P
ZOO	72	49	21.2	13.10	BI - P
F1	90	36	19.1	12.52	BI - K
M1	110	83	26.8	12.37	BF - P
M2	25	10	25.6	9.97	BF - K
M4	27	17	18.6	12.34	MIU, FMU2
M5	13	7	32.3	8.96	MDU
M6	17	14	21.2	12.24	FMU
F2	61	42	31.8	8.94	duplicity
Celkem	981	658	24.6	12.34	

Tab. 2. Přehled poslucháren.

Zkratka	Obor
BF	bakaláři fyziky
BI	bakaláři informatiky
BM	bakaláři matematiky
FMU	učitelství fyzika — matematika
FMU2	učitelství fyzika — matematika pro 2. stupeň ZŠ
MDU	učitelství matematika — deskriptivní geometrie
MIU	učitelství matematika — informatika
duplicity	uchazeči, kteří podali přihlášku na víc programů či oborů
P	jen prezenční studium
K	jen kombinované studium

Tab. 3. Zkratky studijních programů a studijních oborů.

Případnou závislost výsledku na zařazení do poslucháren posoudíme pomocí analýzy rozptylu. Výsledkem je tabulka

```
             Df Sum Sq Mean Sq F value    Pr(>F)
room         17  10242     602  4.2937 2.223e-08 ***
Residuals   640  89799     140
```

z níž vyplývá, že rozdíly mezi posluchárnami jsou vysoce signifikantní. Přitom Leveneův test na shodnost rozptylů dává p-hodnotu 0.31, takže shodnost rozptylů nezamítáme. Pomocí Tukeyovy metody se zjistí, že na obvyklé pětiprocentní hladině se signifikantně liší F2 od M4, S8, F1, S3, FOT, ZOO, a že se každá z poslucháren MIP6 a Ch1 signifikatně liší od F1 i od S3.

Korelační matice mezi jednotlivými úlohami je

```
       ex1    ex2    ex3    ex4
ex1  1.000  0.388  0.433  0.328
ex2  0.388  1.000  0.431  0.310
ex3  0.433  0.431  1.000  0.349
ex4  0.328  0.310  0.349  1.000
```

Z ní vyplývá, že výsledky jednotlivých úloh jsou kladně korelovány, ale tato korelace není příliš velká. Analýza hlavních komponent založená na korelační matici dává tyto výsledky

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4
Standard deviation         1.457     0.845     0.782     0.743
Proportion of Variance     0.531     0.178     0.153     0.138
Cumulative Proportion      0.531     0.709     0.862     1.000

  Loadings:
      Comp.1   Comp.2   Comp.3   Comp.4
  ex1  0.510   -0.217    0.732   -0.397
  ex2  0.504   -0.368   -0.674   -0.395
  ex3  0.530   -0.195             0.825
  ex4  0.452    0.883
```

První hlavní komponenta odpovídá součtu bodů za jednotlivé úlohy. Druhá pak odpovídá rozdílu bodů za čtvrtou úlohu a prvních tří úloh. Jasnou interpretaci má i třetí hlavní komponenta, která porovnává pomocí rozdílu bodů první a druhou úlohu.

Analýza hlavních komponent aplikovaná na kovarianční matici (tedy na nestandardizovaná data) dává

```
Importance of components:
                        Comp.1  Comp.2  Comp.3  Comp.4
Standard deviation       6.693   4.070   2.869   2.417
Proportion of Variance   0.594   0.220   0.109   0.077
Cumulative Proportion    0.594   0.813   0.923   1.000

Loadings:
      Comp.1   Comp.2   Comp.3   Comp.4
ex1   -0.287            0.785   -0.547
ex2   -0.254            0.488    0.835
ex3   -0.812   -0.496   -0.304
ex4   -0.441    0.867   -0.231
```

Zde můžeme první hlavní komponentu vynásobit faktorem -1. Vidíme, že největší váhu má třetí úloha, další největší váhu má čtvrtá úloha.

3. Řešení

1. Jelikož $x = 60 - 2p^2, y = 5p^2 - 140$ mají být kladná, proto

$$28 < p^2 < 30.$$

Vyhovují právě všechny hodnoty p, pro něž platí

$$2\sqrt{7} < |p| < \sqrt{30}, \quad \text{tj.} \quad p \in (-\sqrt{30}, -\sqrt{28}) \cup (\sqrt{28}, \sqrt{30}).$$

2. Rovnici upravíme na tvar

$$2\sin x \cos x(1 + 2\cos x) = 3\sin x(\cos x + 2\cos^2 x - 1),$$

odkud $\sin x = 0$ nebo $2\cos^2 x + \cos x - 3 = 0$.

$x \in l\pi,\ l \in \mathbb{Z}$ $\qquad$ $\cos x = \begin{cases} \nearrow 1 \\ \searrow -\frac{3}{2} \end{cases}$ nevyhovuje

$$x = 2k\pi, k \in \mathbb{Z},$$

avšak pro liché l je $\cos l\pi = -1$, $\cos 2l\pi = 1$, jmenovatel prvního zlomku by se rovnal nule. Rovnici vyhovují právě jen hodnoty $x = 2k\pi$, k celé číslo.

3. Má platit

$$a_1(1+q+q^2)=62 \quad \text{a} \quad 3\log a_1+3\log q=3, \quad \text{odkud} \quad a_1\cdot q=10.$$

Vyloučením a_1 dostaneme pro q rovnici $5q^2-26q+5=0$ s kořeny $5,\ \frac{1}{5}$, k nim dostaneme $a_1=2,\ a_1=50$. Úloha má právě dvě řešení: $\{a_1=2,\ q=5\}$ a $\{a_1=50,\ q=\frac{1}{5}\}$.

4. Jednou tečnou je přímka $x=-3$, rovnici druhé tečny hledáme v směrnicovém tvaru $y=k(x+3)-1$. Dosazením dostaneme pro x kvadratickou rovnici ($k=0$ nevyhovuje)

$$k^2x^2+2(3k^2-4k-8)x+9k^2-24k-32=0.$$

Její diskriminant se rovná nule pouze pro $k=-1$, takže druhá tečna má rovnici

$$x+y+4=0.$$

4. Může studentům pomoci „Mathematica“?

Studenti si ke zkoušce mohli přinést jakékoliv pomůcky, včetně přenosného počítače a libovolného programového vybavení. Předpokládejme, že měli nainstalován program `Mathematica` a že s ním umí alespoň trochu zacházet; nepředpokládáme nicméně žádnou „přehnanou znalost“ tohoto programu. Podívejme se, zda a jak nám takovýto program může pomoci přenést se přes úskalí přijímacího písemky.

Příklad 1. Zde se zdá být přirozené použít příkaz `Solve` určený pro řešení systémů rovnic. Napíšeme-li

```
Solve[{7 x + 3 y == p^2, 5 x + 2 y == 20}, {x, y}]
```

dostaneme jako výsledek

$$\left\{\left\{x\to -2\left(p^2-30\right), y\to 5\left(p^2-28\right)\right\}\right\}$$

To sice příklad 1 plně neřeší, může nám ale usnadnit hledání definitivního řešení.

Pokud si student uvědomí, že místo `Solve` může použít příkaz `Reduce`, tj. napsat například

```
Reduce[{7x+3y == p^2 && 5x+2y == 20 && x>0 && y>0}, {p,x,y}]
```

jako výsledek dostane

$$\left(\left(-\sqrt{30} < \Re(p) < -2\sqrt{7} \wedge \Im(p) = 0 \wedge x = 60 - 2\Re(p)^2\right) \vee \right.$$
$$\left.\left(2\sqrt{7} < \Re(p) < \sqrt{30} \wedge \Im(p) = 0 \wedge x = 60 - 2\Re(p)^2\right)\right) \wedge y = \frac{1}{2}(20 - 5x)$$

odkud již hledané řešení „vyčíst“ lze.

Příklad 2. Použijeme-li opět „přirozený“ příkaz pro řešení rovnic

```
Solve[2 (Sin[x] + Sin[2 x])/(Cos[x] + Cos[2 x]) ==
      3 Sin[x]/Cos[x], x]
```

dostaneme obratem řešení

$$\left\{\{x \to 0\}, \left\{x \to -\cos^{-1}\left(-\frac{3}{2}\right)\right\}, \left\{x \to \cos^{-1}\left(-\frac{3}{2}\right)\right\}\right\}$$

a hlášku

```
Solve:Inverse functions are being used by Solve, so some
solutions may not be found; use Reduce for complete solution
information. More ...
```

Zvědavý student jistě nápovědu zkusí. Napíše-li

```
Reduce[2 (Sin[x] + Sin[2 x])/(Cos[x] + Cos[2 x]) ==
       3 Sin[x]/Cos[x], x]
```

dostane

$$c_1 \in \mathbb{Z} \wedge \left(x = 2\pi c_1 \vee x = 2\pi c_1 - 2i \tanh^{-1}\left(\sqrt{5}\right)\right.$$
$$\left. \vee\, x = 2\pi c_1 + 2i \tanh^{-1}\left(\sqrt{5}\right)\right)$$

odkud již hledané řešení jistě „vyčíst“ lze.

Není nám nicméně jasné, kolik studentů si uvědomí, že `ArcCos[-3/2]` sice vrátí `ArcCos[-3/2]`, ale že na druhé straně `N[ArcCos[-3/2]]` vrátí `3.14159-0.962424i`.

Příklad 3. Použijeme-li opět přirozený příkaz pro řešení rovnic

```
Solve[{a + a q + a q^2 == 62,
Log[10, a] + Log[10, a q] + Log[10, a q^2] == 3}, {a, q}]]
```

dostaneme jako výsledek

$$\left\{\left\{a\to 2, q\to 5\right\}, \left\{a\to 50, q\to \frac{1}{5}\right\},\right.$$
$$\left\{a\to \frac{1}{31}\left(\frac{2077}{2}+\frac{155i\sqrt{3}}{2}-\frac{181}{8}\sqrt{-\frac{801}{50}-\frac{1271i\sqrt{3}}{50}}-\frac{25}{4}\left(-\frac{801}{50}-\frac{1271i\sqrt{3}}{50}\right)^{3/2}\right.\right.$$
$$\left.\left.-\frac{651}{8}i\sqrt{3\left(-\frac{801}{50}-\frac{1271i\sqrt{3}}{50}\right)}\right), q\to -\frac{41}{20}+\frac{31i\sqrt{3}}{20}+\frac{1}{2}\sqrt{-\frac{801}{50}-\frac{1271i\sqrt{3}}{50}}\right\},$$
$$\left\{a\to \frac{1}{31}\left(\frac{2077}{2}+\frac{155i\sqrt{3}}{2}+\frac{181}{8}\sqrt{-\frac{801}{50}-\frac{1271i\sqrt{3}}{50}}+\frac{25}{4}\left(-\frac{801}{50}-\frac{1271i\sqrt{3}}{50}\right)^{3/2}\right.\right.$$
$$\left.\left.+\frac{651}{8}i\sqrt{3\left(-\frac{801}{50}-\frac{1271i\sqrt{3}}{50}\right)}\right), q\to -\frac{41}{20}+\frac{31i\sqrt{3}}{20}-\frac{1}{2}\sqrt{-\frac{801}{50}-\frac{1271i\sqrt{3}}{50}}\right\},$$
$$\left\{a\to \frac{1}{31}\left(\frac{2077}{2}-\frac{155i\sqrt{3}}{2}+\frac{181}{8}\sqrt{-\frac{801}{50}+\frac{1271i\sqrt{3}}{50}}+\frac{25}{4}\left(-\frac{801}{50}+\frac{1271i\sqrt{3}}{50}\right)^{3/2}\right.\right.$$
$$\left.\left.-\frac{651}{8}i\sqrt{3\left(-\frac{801}{50}+\frac{1271i\sqrt{3}}{50}\right)}\right), q\to -\frac{41}{20}-\frac{31i\sqrt{3}}{20}-\frac{1}{2}\sqrt{-\frac{801}{50}+\frac{1271i\sqrt{3}}{50}}\right\},$$
$$\left\{a\to \frac{1}{31}\left(\frac{2077}{2}-\frac{155i\sqrt{3}}{2}-\frac{181}{8}\sqrt{-\frac{801}{50}+\frac{1271i\sqrt{3}}{50}}-\frac{25}{4}\left(-\frac{801}{50}+\frac{1271i\sqrt{3}}{50}\right)^{3/2}\right.\right.$$
$$\left.\left.\left.+\frac{651}{8}i\sqrt{3\left(-\frac{801}{50}+\frac{1271i\sqrt{3}}{50}\right)}\right), q\to -\frac{41}{20}-\frac{31i\sqrt{3}}{20}+\frac{1}{2}\sqrt{-\frac{801}{50}+\frac{1271i\sqrt{3}}{50}}\right\}\right\}$$

Nalezení správného řešení necháme na čtenáři.

Příklad 4. Zde nám Mathematica asi řešení jen tak sama nenabídne. V každém případě nám však může pomoci alespoň zkontrolovat naše „ruční“ výpočty a malovat za nás grafy. Dobře nám již známý příkaz `Solve` zkontroluje, zda umíme vyřešit rovnici paraboly. Skutečně, napíšeme-li

```
Solve[(y - 3)^2 == 16(x + 3), y]
```

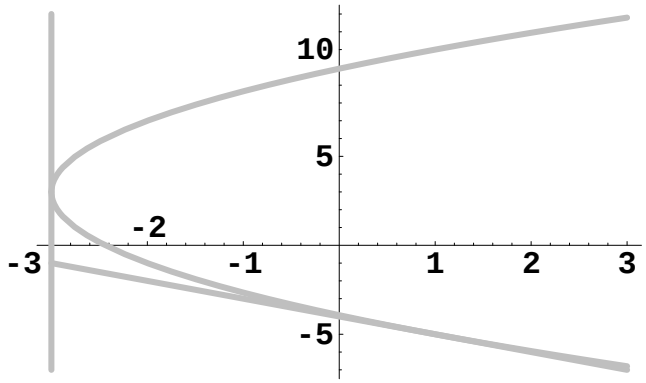
dostaneme jako výsledek

$$\left\{\left\{y\to 3-4\sqrt{x+3}\right\}, \left\{y\to 4\sqrt{x+3}+3\right\}\right\}$$

Nyní již nezbývá nic jiného, než si předchozí řešení namalovat a vzít rozum do hrsti. Zatímco tečna $x = -3$ nás asi napadne ihned, k nalezení druhé tečny už přeci jenom potřebujeme více. Nakonec si řešení namalujeme, například pomocí

```
res = Solve[(y - 3)^2 == 16(x + 3), y];
res = {res, {y -> -x - 4}};
o1  = Plot[Evaluate[y /. res], {x, -3, 3},
       PlotStyle -> {{Thickness[0.01], GrayLevel[0.75]}}];
o2  = Graphics[{Thickness[0.01], GrayLevel[0.75],
       Line[{{-3, -7}, {-3, 13}}]}];
Show[o1, o2];
```

Dostaneme tak to, co očekáváme, totiž



5. Post scriptum

Pro jistou dobu se jedná o poslední přijímací zkoušky na MFF UK, neboť Akademický senát MFF UK na návrh vedení fakulty schválil, že v roce 2008 se odborné přijímací zkoušky na bakalářské studium konat nebudou. Blíže viz

`http://www.mff.cuni.cz/studium/uchazec/prijriz.htm`

POSUZOVÁNÍ BIMODALITY NA ZÁKLADĚ HISTOGRAMU

JUDGEMENT ON BIMODALITY BASED ON HISTOGRAM

Šárka Došlá

Adresa: MFF UK, KPMS, Praha

E-mail: dosla@karlin.mff.cuni.cz

Abstract In this paper we try to respond the following question, i.e., *Does two-modal histogram really indicate two modal distribution?* The response is, according to the expectation, negative. We will show the reasons for and several alternative approaches enabling to decide more reliably on the number of modes.

Úvod

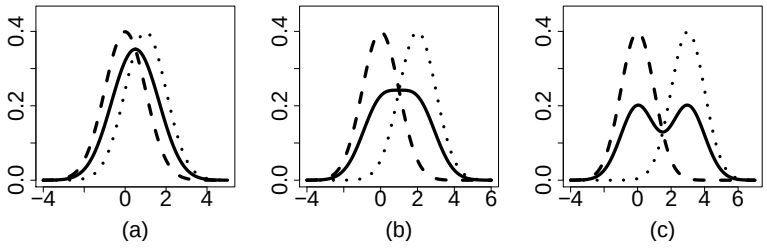
Normovaný histogram je zřejmě nejznámějším odhadem hustoty náhodného výběru. Jelikož je jeho konstrukce velmi jednoduchá a intuitivní, patří mezi oblíbené nástroje statistické analýzy dat. Jeho grafické znázornění nám pomáhá získat lepší představu o chování zkoumaného rozdělení. Avšak vždy bychom měli mít na paměti, že vlastnosti a tvar histogramu mohou být interpretovány a přeneseny na jeho „teoretický protějšek“ pouze přiměřeně, s přihlédnutím k možným náhodným odchylkám.

Bimodální rozdělení ve většině případů vzniká jako směs dvou jednovrcholových rozdělení. V situaci, kdy pracujeme s daty pocházejícími ze směsi dvou rozdělení, můžeme mít tendenci bimodalitu jistým způsobem očekávat. Pokud navíc histogram vykazuje dvě maxima, zdá se být naše podezření potvrzeno.

V následujícím textu se podíváme na to, jak je to s posuzováním bimodality rozdělení na základě histogramu. Indikuje-li histogram dva vrcholy, může být pro nás tento jev dostatečným „důkazem“, že je odpovídající hustota bimodální? Zřejmě nikoliv. Ukážeme, proč může být takový postup velmi zavádějící. Nakonec popíšeme alternativní možnost, kterou lze využít, chceme-li rozhodnout o počtu vrcholů zkoumaného rozdělení.

1. Směsi dvou rozdělení a jejich bimodalita

Již v úvodu jsme použili výraz „směs“. I když je tento pojem zřejmě všeobecně znám, připomeňme pro přesnost, že *směsí dvou rozdělení* s hustotami f_1 a f_2 rozumíme rozdělení s hustotou f, pro kterou platí $f = pf_1 + qf_2$, kde $p, q \in [0, 1]$, $p + q = 1$. V takovém případě je f směsí složek (komponent) f_1 a f_2 a parametry p, q jsou váhy těchto složek.



Obr. 1. Směsi dvou hustot f_1 a f_2 normálních rozdělení $\mathsf{N}(0, 1)$ a $\mathsf{N}(\mu, 1)$ s váhami $p = q = \frac{1}{2}$ pro (a) $\mu = 1$, (b) $\mu = 2$ a (c) $\mu = 3$.

Příklady směsí dvou normálních rozdělení jsou graficky znázorněny na obrázku 1. Pro tento jednoduchý případ směsí $\mathsf{N}(0, 1)$ a $\mathsf{N}(\mu, 1)$ s váhami $p = q = \frac{1}{2}$ můžeme vidět, že tvar výsledné hustoty evidentně závisí na volbě parametru μ, tj. na vzdálenosti vrcholů složek. Leží-li tyto vrcholy velmi blízko sebe, je hustota směsi unimodální. Postupným vzdalováním komponent, tj. zvětšováním μ, se hustota f pomalu „zplošťuje“, až po překročení určité meze vznikne rozdělení bimodální. Toto naše pozorování lze zobecnit a jednoduše shrnout, že směs dvou unimodálních hustot je bimodální pouze v případě, že jsou vrcholy jejích složek „dostatečně“ vzdáleny. Přesně zformulované podmínky pro unimodalitu je možné nalézt např. v [3] či [4].

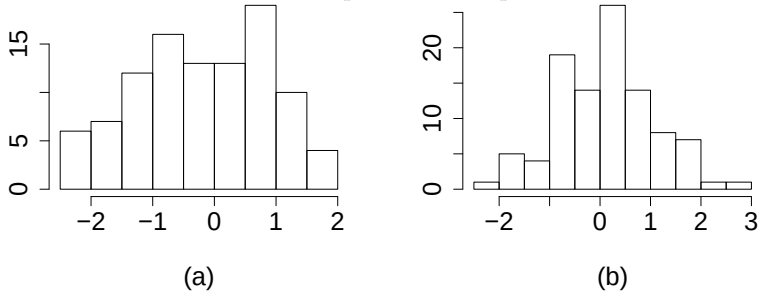
Jak tedy v praxi rozhodnout o počtu vrcholů rozdělení daného náhodného výběru? Předpokládejme, že víme, že naše data pocházejí z nějaké směsi dvou rozdělení. V situaci, že známe váhy a parametry jejích složek nebo jejich odhady, můžeme o unimodalitě, resp. bimodalitě, rozhodnout na základě citovaných teoretických kritérií. Bohužel, většinou však máme k dispozici pouze data a parametry komponent nejsme schopni odhadnout. Zmíněná kritéria pak nelze aplikovat, a tak přichází na řadu histogram...

2. Histogramy a posuzování jejich bimodality

Je všeobecně známo, že tvar histogramu závisí na parametrech, z nichž některé sami, často spíše subjektivně, volíme. Počet tříd či jejich šířka ovlivňují

hladkost a výskyt případných vrcholů. Kromě toho, čím více pozorování máme k dispozici, tím je histogram „hladší a přesnější“.

Vědomi si všech těchto skutečností, vykreslíme histogram našeho výběru. Odhlédněme nyní od možnosti měnit počet jeho tříd a předpokládejme, že jsme použili optimální volbu dle některého ze známých kritérií (např. Sturgesova). Na základě vytvořeného histogramu se snažíme získat představu o tvaru skutečného rozdělení našich dat: Mohlo by se jednat o normální či jiné unimodální rozdělení? Nebo bude naopak hustota spíše dvouvrcholová?



Obr. 2. Histogramy náhodných výběrů simulovaných z rozdělení $\mathsf{N}(0,1)$ o rozsahu 100 pozorování s nastavením (a) `set.seed(89)` a (b) `set.seed(59)`.

Na chvilku ještě počkejme se svým rozhodnutím a podívejme se na následující možnou situaci. Na obrázku 2(a) je znázorněn histogram výběru simulovaného z normálního rozdělení $\mathsf{N}(0,1)$ o rozsahu 100 pozorování[1]. Tento histogram má dvě lokální maxima, tj. dva vrcholy. Vyvodili bychom z tohoto jevu, že zkoumaný výběr pochází z bimodálního rozdělení? Či bychom jej spíše připsali jakési „nepřesnosti“ odhadu? Nebo bychom se soudit neodvážili?

Odpověď asi není jednoznačná. V tomto konkrétním příkladě jsme věděli, že jde o výběr generovaný z normálního rozdělení, a proto bychom se zřejmě zdrželi unáhlených soudů. Co ale v případě našich reálných dat? Problémem je, že v situaci, kdy víme, že data pocházejí ze směsi dvou jednovrcholových rozdělení, bimodalitu jaksi očekáváme. A tak se na základě dvouvrcholového histogramu můžeme snadno nechat přesvědčit o tom, že je zkoumané rozdělení bimodální, a učinit tak možná chybný závěr.

Posuzování tvaru histogramu je evidentně záležitost subjektivního rázu. Navíc, ne každé jeho lokální maximum vnímáme jako „potenciální vrchol“

[1] Simulace provedena v programu `R` s nastavením `set.seed(89)`.

hustoty. Velmi často jsou četnosti několika prostředních tříd histogramu výrazně vyšší než četnosti zbývajících tříd. Při zkoumání modality si pak všímáme pouze vrcholů indikovaných v těchto prostředních třídách a případná další lokální maxima pomineme. V případě histogramu na obrázku 2(b) budeme zřejmě brát v úvahu pouze vrcholy, které indikuje na intervalech $(-1, -0.5]$ a $(0, 0.5]$ a lokální maximum ve třídě $(-2, -1.5]$ budeme chápat spíše jako „náhodnou odchylku“.

Proto se nadále omezíme pouze na studování několika prostředních tříd histogramů a budeme sledovat maxima indikovaná pouze zde. Ostatní třídy nebudeme brát při posuzování bimodality v úvahu.

3. Případ rozdělení s „tupým“ vrcholem

U některých směsí nejsou vrcholy jejich složek vzdáleny natolik, aby byla výsledná hustota dvouvrcholová. Může tak nastat případ, kdy je sice rozdělení unimodální, ale tento jeho jediný vrchol je velmi „neostrý“. Tak je tomu například u směsi (b) na obrázku 1, jejíž hustota je na jakémsi okolí svého vrcholu téměř konstantní. V následujícím textu se zaměříme na taková unimodální rozdělení s „tupým“ vrcholem a podíváme se na odhad pravděpodobnosti, s jakou se histogram výběru z takového rozdělení jeví jako bimodální.

Pro ilustraci vezměme nejprve konkrétní směs dvou normálních rozdělení $\mathsf{N}(0, 1)$ a $\mathsf{N}(2, 1)$ s váhami $p = q = \frac{1}{2}$ (viz obrázek 1(b)) a uvažujme náhodnou veličinu X s tímto rozdělením. Zaměřme se pouze na interval $[0, 2]$. Rozdělíme-li jej na šest stejně velkých podintervalů $I_1, \ldots, I_6$, je pravděpodobnost, že X padne do intervalu I_i, přibližně stejná pro všechna $i = 1, \ldots, 6$. Podmíněné pravděpodobnosti $\mathsf{P}(X \in I_i | X \in [0, 2])$, $i = 1, \ldots, 6$, jsou postupně 0.1630, 0.1680, 0.1690, 0.1690, 0.1680 a 0.1630. V případě, že se zaměříme na veličinu X pouze na intervalu $[0, 2]$, tj. podmíníme-li její rozdělení jevem $\big[X \in [0, 2]\big]$, dostaneme tak přibližně rovnoměrné rozdělení na $[0, 2]$.

Podobnou úvahu můžeme snadno aplikovat na rozdělení s „tupým“ vrcholem obecně. Docházíme k následujícímu závěru: Jelikož jsme se při posuzování histogramu omezili pouze na zkoumání několika jeho prostředních tříd, stačí nám dívat se na danou hustotu jen na nějakém okolí jejího vrcholu. Rozdělení, jehož vrchol je dostatečně „tupý“, můžeme na tomto intervalu dostatečně dobře aproximovat rovnoměrným rozdělením. Okamžitě se tudíž nabízí následující zjednodušení celého problému: Najdeme-li odhad pravděpodobnosti, s jakou se histogram výběru z rovnoměrného rozdělení jeví jako bimodální, budeme jej pak moci použít i pro jakékoliv unimodální rozdělení s „tupým“ vrcholem.

4. Histogramy výběrů z rovnoměrného rozdělení

Kdy tedy chápeme histogram jako bimodální? Zcela intuitivně to bude v případě, že má „právě dva vrcholy“. Připomeňme, že v tomto momentě se již díváme pouze na několik, řekněme N, prostředních tříd histogramu a četnosti ostatních necháváme stranou. Bimodální tak bude takový histogram, který má mezi těmito N třídami právě dvě „maxima“, tj. splňuje podmínku:

Označme zvolených N prostředních tříd histogramu jako $1, 2, \ldots, N$ a jejich odpovídající četnosti $n_1, n_2, \ldots, n_N$, kde $n_i \geq 0$ pro všechna $i = 1, \ldots, N$. Dodefinujme $n_0 = n_{N+1} = 0$. Řekneme, že daný *histogram je bimodální*, jestliže existují přirozená čísla M_1, M_2, M_3 taková, že platí $0 < M_1 < M_2 < M_3 < N + 1$ a

$$\begin{aligned}
&n_{i-1} \leq n_i \quad \text{pro } i = 1, \ldots, M_1, && n_{M_1} > n_{M_1+1},\\
&n_{i-1} \geq n_i \quad \text{pro } i = M_1 + 2, \ldots, M_2, && n_{M_2} < n_{M_2+1},\\
&n_{i-1} \leq n_i \quad \text{pro } i = M_2 + 2, \ldots, M_3, && n_{M_3} > n_{M_3+1},\\
&n_{i-1} \geq n_i \quad \text{pro } i = M_3 + 2, \ldots, N + 1.
\end{aligned}$$

V takovém případě budeme i příslušnou posloupnost čísel $\{n_i\}_{i=1}^{N}$ nazývat bimodální. Permutaci čísel $1, \ldots, N$ nazveme *bimodální permutací*, jestliže je tato posloupnost čísel bimodální.

Pro histogramy výběrů z rovnoměrného rozdělení můžeme dokázat následující tvrzení popisující jejich chování[2]: *Je-li $X_1, \ldots, X_M$ náhodný výběr z rovnoměrného rozdělení na intervalu $[a, b]$, $a, b \in \mathbb{R}$, a $N \in \mathbb{N}$, potom pro $M \to \infty$ se pravděpodobnost, s jakou je histogram tohoto náhodného výběru s N třídami bimodální, blíží k pravděpodobnosti, že je náhodná permutace čísel $1, \ldots, N$ bimodální.*

V tabulce 1 jsou uvedeny četnosti bimodálních permutací čísel $1, \ldots, N$ pro $N = 4, \ldots, 8$. Vyčeteme z ní například, že mezi permutacemi čísel $1, \ldots, 6$ je přibližně 57.8 % bimodálních. Podle výše uvedeného tvrzení můžeme hodnotu 0.578 brát jako odhad pravděpodobnosti, s jakou histogram náhodného výběru pocházejícího z rovnoměrného rozdělení $\mathsf{R}[0, 1]$ s šesti třídami vykazuje dva vrcholy. Jestliže tedy obecně bereme při posuzování modality v úvahu jen prostředních šest tříd histogramu, lze hodnotu 0.578 brát i jako odhad pravděpodobnosti, s jakou se nám histogram výběru z rozdělení s „tupým“ vrcholem jeví jako bimodální.

[2] Důkaz uvedeného tvrzení viz [1].

Četnosti bimodálních permutací					
N	4	5	6	7	8
počet všech permutací	24	120	720	5040	40320
počet bimodálních permutací	16	88	416	1824	7680
podíl bimodálních permutací	$0.\overline{6}$	$0.7\overline{3}$	$0.5\overline{7}$	0.362	0.191

Tab. 1. Počty bimodálních permutací čísel $1, \ldots, N,\ N = 4, \ldots, 8$.

Jak tedy můžeme vidět, tato pravděpodobnost rozhodně není zanedbatelná. Proto posuzování bimodality rozdělení na základě histogramu není ani v nejmenším vhodné a mohlo by velmi často vést k nesprávným a zavádějícím závěrům.

5. Když ne histogram, tak co tedy?

Co tedy použít v situaci, kdy potřebujeme zjistit, zda naše data pocházejí z rozdělení s jedním či více vrcholy? Histogram zjevně není dobrý nástroj. Naštěstí existují jiné možné postupy.

V programu `R` je implementován dip test (viz [2]), pomocí kterého můžeme testovat, zda daný náhodný výběr pochází z unimodálního rozdělení. Testovou statistikou je tzv. dip, který je jakousi mírou vzdálenosti empirické distribuční funkce daného výběru a třídy všech unimodálních distribučních funkcí. Funkce `dip` z knihovny `diptest` spočítá pro naše data dip statistiku a porovnáním její hodnoty s příslušným empirickým kvantilem (tabulka `qDiptab` z téže knihovny) pak můžeme učinit závěr, zda na zvolené testovací hladině zamítáme nulovou hypotézu unimodality či nikoliv.

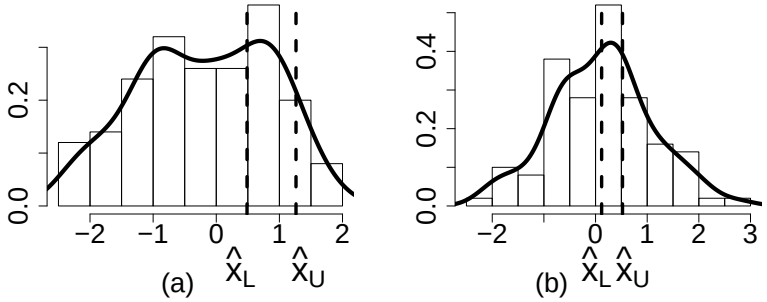
Při konstrukci testu je nutné zvolit konkrétní unimodální rozdělení za nulové hypotézy. Zřejmě však neexistuje takové, pro něž by byla dip statistika stochasticky větší než pro všechna ostatní unimodální rozdělení. Proto se volí za nulové hypotézy rovnoměrné rozdělení. Tato volba je velmi jednoduchá, ale vede k testu, který je asymptoticky konzervativní (viz [2]). Pro ilustraci jsou v tabulce 2 uvedeny relativní četnosti výběrů generovaných z rovnoměrného a normálního rozdělení, pro něž byla hypotéza unimodality dip testem na hladině 0.05 zamítnuta. Pro výběry z normálního rozdělení se zdá být chyba prvního druhu znatelně menší než 0.05 a pro rostoucí rozsah se dokonce blíží k 0. Tato skutečnost potvrzuje asymptotické vlastnosti ukázané v [2] a zmíněnou konzervativnost testu.

rozdělení	rozsah výběru 50	100	1000	5000
rovnoměrné $\mathsf{R}[0,1]$	0.04995	0.04867	0.04834	0.04946
normální $\mathsf{N}(0,1)$	0.00292	0.00109	0.00004	0

Tab. 2. Relativní četnost výběrů, pro něž byla hypotéza unimodality dip testem na hladině 0.05 zamítnuta: V prvním řádku jsou výsledky dip testu pro 100 000 náhodných výběrů simulovaných z rovnoměrného rozdělení, druhý řádek odpovídá výběrům generovaným z normálního rozdělení $\mathsf{N}(0,1)$. Počáteční nastavení `set.seed(1023)`.

Mohlo by nás zajímat, jak dip test posoudí rozdělení výběrů, jejichž histogramy z obrázku 2 jsme diskutovali v předchozích odstavcích. Připomeňme, že jde o data simulovaná z normálního rozdělení $\mathsf{N}(0,1)$ o rozsahu 100 pozorování a jejich histogramy vykazovaly více než jeden vrchol.

V prvním případě jsme simulace provedli s nastavením `set.seed(89)` a histogram indikoval dvě maxima. Dip statistika spočtená pro tento výběr vychází 0.0408. Jelikož kritická hodnota na hladině významnosti 0.05 pro rozsah výběru 100 je 0.0511, dip test hypotézu unimodality nezamítá. Na obrázku 3(a) je vykreslen histogram a neparametrický odhad hustoty obdržený funkcí `density`. Dále je znázorněn odhad $(\hat{x}_L, \hat{x}_U)$ intervalu, ve kterém by se měl nacházet vrchol rozdělení. Pro druhý výběr, generovaný z $\mathsf{N}(0,1)$ s nastavením `set.seed(59)`, vychází dip roven 0.0256, takže stejně jako v předchozím případě hypotézu unimodality na hladině 0.05 nezamítáme. Grafické znázornění viz obrázek 3(b). V obou případech nám tedy dip test dává na naši otázku o unimodalitě rozdělení „správnou odpověď“.



Obr. 3. Histogram, odhad hustoty (funkce `density`) a modálního intervalu rozdělení náhodného výběru o rozsahu 100 pozorování simulovaného z $\mathsf{N}(0,1)$ v programu `R` s nastavením (a) `set.seed(89)` a (b) `set.seed(59)`.

Při zkoumání histogramů jsme se zabývali především směsmi dvou unimodálních rozdělení. Podívejme se proto nyní na to, jak dip test funguje v takových případech.

K tomuto účelu jsme v programu `R` simulovali náhodné výběry ze směsi dvou normálních rozdělení $\mathsf{N}(0,1)$ a $\mathsf{N}(\mu,1)$ s váhami $p = q = \frac{1}{2}$ s různými rozsahy a volbami parametru μ a sledovali jsme, jaké výsledky dává dip test. Není obtížné ukázat (viz [4]), že směs dvou normálních rozdělení $\mathsf{N}(0,1)$ a $\mathsf{N}(\mu,1)$ s váhami $p = q = \frac{1}{2}$ je unimodální pro $|\mu| \leq 2$ a bimodální pro $|\mu| > 2$. Tudíž bychom zřejmě pro $\mu > 2$ očekávali zamítnutí nulové hypotézy unimodality. V tabulce 3 jsou uvedeny výsledky dip testu pro 100 000 generovaných výběrů s rozsahy $M = 100, 1000$ a 5000 pro volby $\mu = 2, 2.5, 2.8, 3, 3.5$ a iniciální nastavení `set.seed(1023)` v programu `R`. Vidíme, že při rostoucím rozsahu výběru roste i síla testu. Ale například pro $\mu = 2.5$ a pro rozsah 5000 pozorování jsme stále u 70 % výběrů hypotézu unimodality nezamítli, přestože se jednalo o data z bimodálního rozdělení.

Při použití dip testu se tak dostáváme do opačného problému než tomu bylo u histogramů. Na základě nich jsme mohli s nezanedbatelnou pravděpodobností považovat unimodální rozdělení za bimodální. Naopak, pomocí dip testu bychom mohli bimodální rozdělení mylně označit jako unimodální. Rozhodně je však vhodnější při posuzování bimodality použít formální dip test než dělat nepodložené závěry na základě histogramu indikujícího dva možné vrcholy.

	rozsah výběru M		
μ	100	1000	5000
2.0	0.00458	0.00061	0.00008
2.5	0.02092	0.04888	0.30210
2.8	0.05634	0.42790	0.99584
3.0	0.06856	0.82634	1
3.5	0.38187	0.99998	1

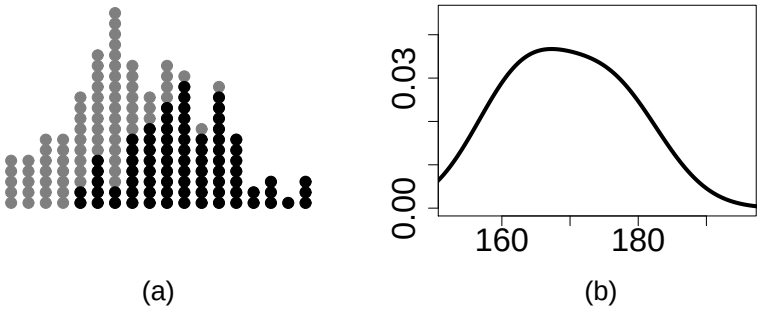
Tab. 3. Výsledky dip testu pro 100 000 náhodných výběrů simulovaných ze směsi dvou normálních rozdělení $\mathsf{N}(0,1)$ a $\mathsf{N}(\mu,1)$ s váhami $p = q = \frac{1}{2}$ pro různé hodnoty μ a různé rozsahy výběrů M. V tabulce jsou uvedeny relativní četnosti výběrů, pro něž byla hypotéza unimodality zamítnuta. Pro $\mu = 2$ je daná směs unimodální a pro $\mu > 2$ je směs bimodální. Vždy iniciální nastavení `set.seed(1023)` v programu `R`.

6. Reálný příklad — „živý“ histogram

Na začátku našeho textu, v části 2., jsme diskutovali o subjektivním postoji při posuzování histogramů. Ukázali jsme, že daný dvouvrcholový histogram na nás ve dvou různých situacích může působit zcela jiným dojmem. V prvním případě jsme větší počet vrcholů automaticky připsali nepřesnosti odhadu, jelikož jsme věděli, že data pocházejí z normálního rozdělení. Naopak ve druhém případě jsme měli data pocházející ze směsi dvou rozdělení, a tak jsme dva vrcholy možná i trochu očekávali a nechali se proto přesvědčit o bimodalitě odpovídající hustoty. Příkladem takového jednání, kdy byl tvar histogramu shledán jako dostatečný důkaz bimodality, je následující situace pocházející ze článku [5].

Během jedné přednášky ze statistiky seřadil vyučující své studenty na školním hřišti do skupin dle jejich výšky a zkonstruoval tak jakýsi „živý“ histogram. Jeho tvar působil „bimodálně“ (viz obrázek 4(a)), a tak bylo zábavnou formou studentům ilustrováno, že rozdělení lidské výšky, jakožto směs dvou unimodálních rozdělení, má dva vrcholy. Bezpochyby se jednalo o velmi zdatný didaktický počin. Avšak problém je v tom, že takové tvrzení není pravdivé.

Autoři článku [5] se podívali na rozdělení výšky studentů více teoreticky. Na základě dat pocházejících z šetření státního zdravotního centra USA odhadli parametry rozdělení výšky mužů a výšky žen v odpovídajícím věku. Aplikací teoretických kritérií potom zjistili, že výsledné společné rozdělení výšky by mělo být unimodální, viz obrázek 4(b), a nikoliv bimodální!



Obr. 4. (a) Struktura „živého“ histogramu studentů: Znázorněné tečky odpovídají jednotlivým studentům, dívky a chlapci jsou barevně odlišeni. (b) Hustota rozdělení výšky studentů spočtená na základě odhadnutých parametrů.

Závěr z celého experimentu je tedy spíše rozpačitý. Místo toho, aby vyučující studentům ukázal příklad bimodálního rozdělení, dopustil se chyby a sdělil jim nepravdivou informaci. Navíc svým žákům (nechtěně) přímo demonstroval nekorektní postup, který ho dovedl k nesprávným závěrům. A tak můžeme jen doufat, že žádný ze zmíněných studentů nepoužije podobnou nepodloženou úvahu při nějaké skutečně důležité analýze dat.

7. Závěr

Závěrem lze shrnout, že posuzování bimodality či unimodality dané hustoty pouze na základě tvaru histogramu může často vést k nesprávným závěrům. V situaci, kdy nás skutečně zajímá počet vrcholů zkoumaného rozdělení, je vhodnější použít jiné postupy. Rozhodně bychom se neměli nechat ovlivnit našimi očekáváními a dát se strhnout k unáhleným a nepodloženým soudům, tak jako tomu bylo v uvedeném příkladě vyučujícího a výšky jeho studentů.

Poděkování: Příspěvek vznikl za pomoci grantu MSM 0021620839.

Reference

[1] Došlá Š. (2006) Bimodální rozdělení. *Diplomová práce*, Univerzita Karlova, Praha.

[2] Hartigan J.A., Hartigan P.M. (1985) The dip test of unimodality. *Ann. Statist.* **13**, 70–84.

[3] Kemperman J.H.B. (1991) Mixture with a limited number of modal intervals. *Ann. Statist.* **19**, 2120–2144.

[4] Robertson C.A., Fryer J.G. (1969) Some descriptive properties of normal mixtures. *Skand. Aktuarietidskr.* **52**, 137–146.

[5] Schilling M.F., Watkins A.E., Watkins W. (2002) Is human height bimodal? *Amer. Statist.* **56**, 223–229.

MIKULÁŠSKÝ STATISTICKÝ DEN 2007

Marek Malý

Adresa: SZÚ, Praha

E-mail: maly@szu.cz

Rok 2007 zakončila Česká statistická společnost 6. prosince přednáškovým seminářem v příjemném prostředí Balbínovy poetické hospůdky na Vinohradech v Praze. Asi 25 posluchačů vyslechlo v průběhu pětihodinového programu Mikulášského statistického dne osm přednášek, které se dotkly různých aspektů statistické teorie i praxe. Mezi přednášející se zamíchal i hodný čert, který podělil malými dárky všechny posluchače, Mikuláš osobně k nám třeba zavítá příště.

P. Praks a P. Zajac připravili přednášku o posuzování spolehlivosti softwaru (*PageRank ve statistice*). D. Hlubinka se zabýval dotazníkovými nástroji, které jsou nedílnou součástí práce každého statistika pohybujícího se v aplikacích (*O kvalitě vyplňování dotazníků v rovníkové Africe*), P. Popela se zabýval důležitými otázkami posuzování naší práce a hodnocením činnosti vysokých škol (*Jak vážíme vědu*). Po polední přestávce ukázal J. Běláček konkrétní aplikace statistiky v prostředí lékařské fakulty (*Jak jsem doloval v datech aneb O úplně normálních regresních přímkách*), J. Anděl ve velmi zajímavé přednášce ilustroval na dvou příkladech konstrukci regresních modelů jednak z pohledu možného vlivu grafického znázornění, jednak z pohledu využití dodatečné informace (*Volba regresního modelu a o chybách, které se přitom dělají*), G. Dohnal nám vysvětlil, proč se v životě tolik načekáme (*Frontové paradoxy*), Z. Fabián ve vesele laděném příspěvku pohovořil o vážném tématu (*Inferenční funkce a parametrické odhady*) a závěrem J. Klaschka na pozadí praktické aplikace poukázal na úskalí v přístupu lékařů ke statistice (*Co je statisticky nejvýznamnější?*).

Diskuse, která se rozhodně netýkala jen semináře, nýbrž i mnoha dalších zajímavých témat statistické komunity, se po semináři přesunula do přilehlé kavárny. V průběhu statistického dne měli účastníci výjimečnou možnost setkat se všemi pěti dosavadními předsedy České statistické společnosti v její sedmnáctileté historii, tedy prof. Andělem, prof. Čermákem, ing. Rothem, prof. Antochem a doc. Dohnalem. Některé z přednášek autoři připravili pro publikaci v Informačním bulletinu, takže i ti, jimž předvánoční shon neumožnil chvilku zastavení se statistikou, budou mít možnost se s probíranými tématy seznámit a třeba je to podnítí k účasti na některé z dalších akcí.

STUDENTSKÁ KONFERENCE A ČTVRTÉ SETKÁNÍ NÁRODNÍCH STATISTICKÝCH SPOLEČNOSTÍ V PRAZE

Gejza Dohnal

E-mail: gejza.dohnal@fs.cvut.cz

Počátkem září (4. - 6.9. 2008) proběhne v Praze další, v pořadí již čtvrté setkání zástupců národních statistických společností. V posledním čísle IB minulého roku jsme Vás informovali o 3. setkání, které se uskutečnilo na podzim 2007 ve Slovinské Ljubljani. Letošní stekání bude spojeno s mezinároidní studentskou konferencí o matematické statistice a pravděpodobnosti, na níž předpokládáme účast studentů ze všech zúčastněných zemí, tj. z Česka, Maďarska, Slovenska, Slovinska, Rakouska a Rumunska (skupina V6). Konference bude mít dvě sekce, jednu pro studenty magisterského studia a druhou pro doktorandy. Účast studentů na této konferenci bude finančně podpořena jejich národními statistickými společnostmi. Pro řadu studentů by to mohla být jejich první příležitost vystoupit před mezinárodním fórem. Studenti obou typů studia (magisterského i postgraduálního) mohou již teď posílat své přihlášky na adresu tajemníka České statistické společnosti. Přihláška by měla obsahovat kromě jména studenta a kontaktu i název příspěvku, krátkou anotaci, název školy, obor, ročník a případně doporučení vedoucího diplomové práce či školitele. Přijaté příspěvky budou publikovány v některém z periodik, vydávaných statistickými společnostmi skupiny V6.

KONFERENCE ISBIS 2008

Mezinárodní společnost pro obchodní a průmyslovou statistiku (ISBIS) pořádá každé dva roky mezinárodní symposium, na němž vystupují přední světoví experti v uvedených oblastech. Po Severním Qeenslandu, Limě a Azorech se bude toto setkání konat letos v červenci v Praze. Symposium proběhne ve dnech 1. – 4. 7. 2008 v hotelu Anděl na Smíchově v Praze 5. Hlavními pořadateli jsou American Statistical Association, Section on Physical and Engineering Sciences a American Society for Quality, spolupořádajícími organizacemi jsou International Statistical Institute, European Network of Business and Industry Statistics a v neposlední řadě i Česká statistická společnost a Centrum pro jakost a spolehlivost výroby CQR. Členové všech zúčastněných organizací, tedy i naší společnosti, mají slevu na vložném.

Hlavní sekce budou věnovány kvantitativní analýze v bankovnictví, finančnictví a pojišťovnictví. Připravují se však i sekce týkající se statistických metod v řízení jakosti, spolehlivosti a analýzy rizik. Jejich seznam, spolu s dalšími informacemi a registračním formulářem viz http://www.action-m.com/isbis2008/index.php

Vážené kolegyně, vážení kolegové,
redakce, výbor společnosti a organizátoři si Vás dovolují pozvat na Liberecké statistické dny *Průmyslová statistika a chemometrie*, které se uskuteční ve dnech 10. – 11. dubna v Liberci. Zájemci o podrobné informace nechť se obrátí na doc. RNDr. Aleše Linku, CSc. (ales.linka@tul.cz).

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

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

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

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

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

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

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