Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Štatistická inferencia I

Zadania domácich úloh

Stanislav Katina

katina@math.muni.cz

22. decembra 2014

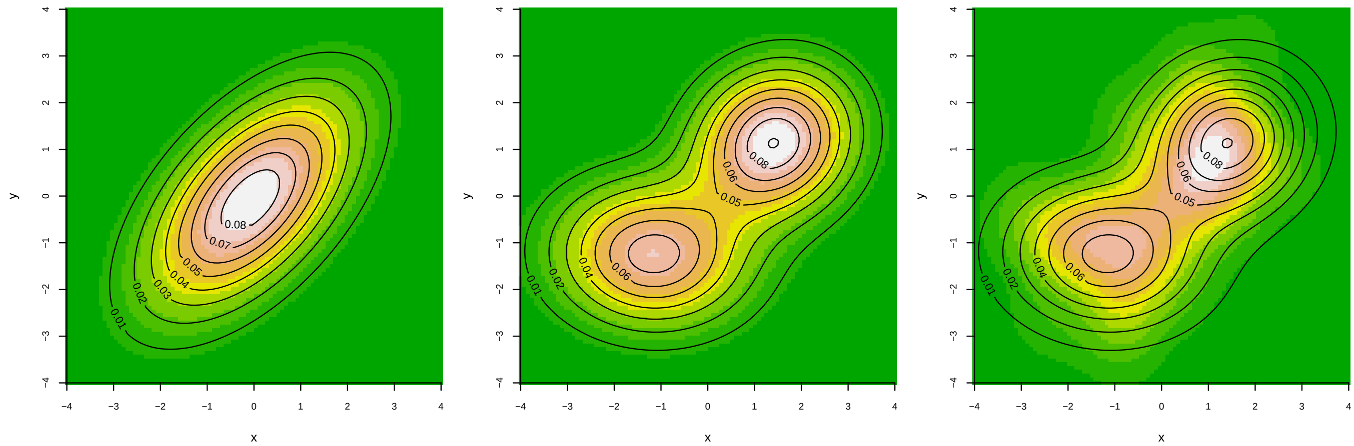
Inštrukcie k DÚ: Odovzdáva sa jeden pdf súbor nazvaný priezvisko-meno-text-statinf-I-2014.pdf (obsahuje riešenia príkladov, obrázky, R-kód napísaný v TEXu), jeden zdrojový súbor naprogramovaných funkcií priezvisko-meno–source-statinf-I-2014.r a jeden súbor R-kódu konkrétnych zadaní z DÚ priezvisko-meno-priklady-statinf-I-2014.r, ktorý používa tento zdrojový kód. Na písanie R-kódu odporúčam TEX balíček `listings` a vytvoreniu prostredia v hlavičke dokumentu ako

```
\lstset{language=R, % nastavenie jazyka R
basicstyle=\footnotesize\ttfamily, % typ pisma R-kodu
commentstyle=\ttfamily\color{farba1}, % farba  komentara k funkciam
numberstyle=\color{farba2}\footnotesize, % farba a velkost cislovania
numbers=left, % cislovanie vlavo
stepnumber=1, % cislovanie po krokoch jedna
frame=leftline, % vytvorenie lavej hranicnej ciary
breaklines=true} % zalomenie riadkov
```

a potom v texte medzi `begin` a `end`.
DÚ je potrebné odovzdať 7 dní pred termínom skúšky, na ktorý sa prihlásite.

Príklad 1 (zmes dvoch dvojrozmerných normálnych rozdelení) *Simuláciu pseudonáhodných čísel zo zmesi dvoch normálnych rozdelení $pN_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-p)N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ môžeme v R urobiť použitím jedného z alternatínych postupov z príkladu 20 (zo zadaní príkladov na cvičenia). Nasimulujte pseudonáhodné čísla X a Y (1) zo zmesi $pN_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-p)N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, kde $\boldsymbol{\theta} = (\mu_{11}, \mu_{12}, \sigma_{11}^2, \sigma_{12}^2, \rho_1, \mu_{21}, \mu_{22}, \sigma_{21}^2, \sigma_{22}^2, \rho_2)^T$ a (2) z dvojrozmerného rozdelenia $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde parametre predstavujú spoločný vektor stredných hodnôt a spoločnú kovariančnú maticu. t.j. $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$. Pre (1) vypočítajte dvojrozmerný jadrový odhad hustoty $(X, Y)^T$ pomocou funkcie **`kde2d()`**.*
*(a) Nakreslite teoretickú hustotu (2) pomocou funkcie **`image()`** a superponujte ju s kontúrovým grafom teoretickej hustoty (2) pomocou funkcie **`contour()`**.*
*(b) Nakreslite teoretickú hustotu (1) pomocou funkcie **`image()`** a superponujte ju s kontúrovým grafom teoretickej hustoty (1) pomocou funkcie **`contour()`**.*
*(c) Nakreslite dvojrozmerný jadrový odhad hustoty realizácí (1) pomocou funkcie **`image()`** a superponujte ju s kontúrovým grafom teoretickej hustoty (1) pomocou funkcie **`contour()`**.*
*Hustotu rozsekajte na 12 intervalov, kde hodnoty v týchto intervaloch budú zodpovedať farbám **`terrain.colors(12)`**. Pri simulácii použite $\boldsymbol{\theta} = (-1.2, -1.2, 1, 1, 0, 1, 1, 1, 1, 0)^T$,*
(1) $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}_{11}, \widehat{\mu}_{12}, \widehat{\sigma}_{11}^2, \widehat{\sigma}_{12}^2, \widehat{\rho}_1, \widehat{\mu}_{21}, \widehat{\mu}_{22}, \widehat{\sigma}_{21}^2, \widehat{\sigma}_{22}^2, \widehat{\rho}_2)^T$, $n_1 = n_2 = 50$ a $p = 0.5$ (odhady pochádzajú z nasimulovaných dát).
(2) $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \widehat{\rho})^T$ a $n_1 = n_2 = 50$ (odhady pochádzajú zo spoločného výberu nasimulovaných dát).
Vzorové riešenie pozri na obrázku **??**. DÚ

Príklad 2 (kvadratická aproximácia profilovej funkcie vierohodnosti) *(2.1) Nakreslite škálovaný logaritmus profilovej funkcie vierohodnosti normálneho rozdelenia pre μ. Na x-ovej osi bude μ a na y-ovej osi $\ln \mathcal{L}(\mu|\mathbf{x}) = l(\mu|\mathbf{x}) - \max(l(\mu|\mathbf{x}))$. Porovnajte $\ln \mathcal{L}(\mu|\mathbf{x})$ s kvadratickou aproximáciou vypočítanou pomocou Taylorovho rozvoja $\ln \mathcal{L}(\mu|\mathbf{x}) = \ln(\frac{L(\mu|\mathbf{x})}{L(\widehat{\mu}|\mathbf{x})}) \approx -\frac{1}{2}\mathcal{I}(\widehat{\mu})(\mu - \widehat{\mu})^2$. (2.2) Nech skóre funkcia $S(\mu) = \frac{\partial}{\partial\mu} \ln L(\mu|\mathbf{x})$. Keď zoberieme deriváciu kvadratickej aproximácie uvedenej vyššie, dostaneme $S(\mu) \approx -\mathcal{I}(\widehat{\mu})(\mu - \widehat{\mu})$ alebo $-\mathcal{I}^{-1/2}(\widehat{\mu})S(\mu) \approx \mathcal{I}^{1/2}(\widehat{\mu})(\mu - \widehat{\mu})$. Potom zobrazením pravej strany na x-ovej osi a ľavej strany na y-ovej osi dostaneme asymptoticky lineárnu funkciu s jednotkovým sklonom. Asymptoticky tiež platí $\mathcal{I}^{1/2}(\widehat{\mu})(\mu - \widehat{\mu}) \sim N(0, 1)$. Je postačujúce mať rozsah x-vej osi $\langle -2, 2 \rangle$, pretože funkcia je asymptoticky (lokálne) lineárna na tomto intervale. Rozumne škálujte y-ovú os. Zobrazte pre (a) $n = 10$, (b) $n = 100$ a (c) $n = 1000$. Použite (1)*



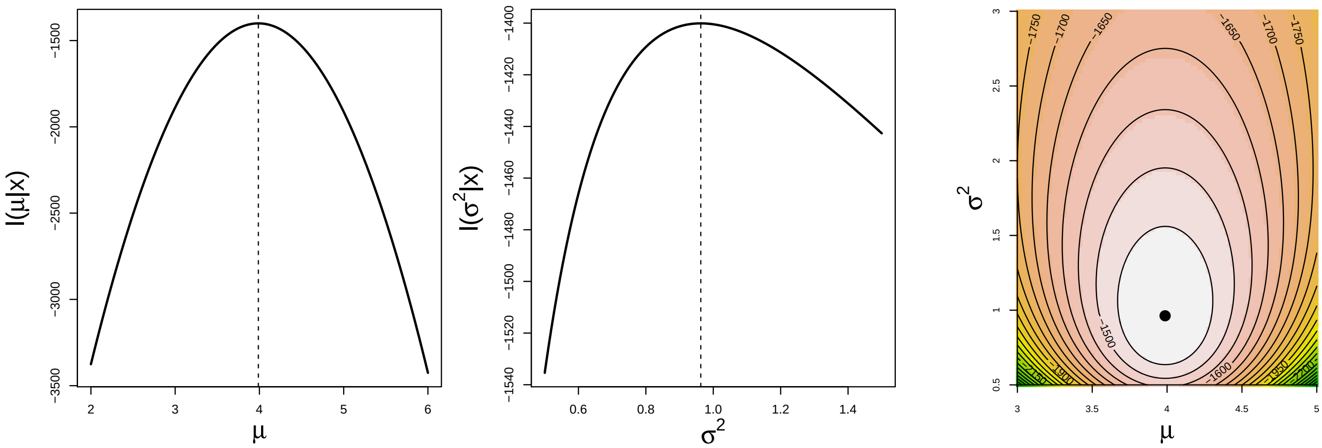
Obr. 1: Spoločná hustota dvojrozmerného normálneho rozdelenia (vľavo), hustota zmesi dvoch dvojrozmerných normálnych rozdelení (uprostred) a dvojrozmerný jadrový odhad superponovaný hustotou zmesi dvoch dvojrozmerných normálnych rozdelení (vpravo)

$X \sim N(0,1)$ a (2) $X \sim (1-p)N(0,1) + pN(0,2)$, kde $p = 0.05$. Okomentujte rozdiely medzi (a), (b) a (c), ako aj rozdiely medzi (1) a (2). (2.3) Dokážte, že ak $N(\mu, \sigma^2)$ a σ^2 je známe, potom je kvadratická aproximáca logaritmu relatívnej vierohodnosti identická s logaritmom relatívnej vierohodnosti, t.j. $\ln \mathcal{L}(\mu|\mathbf{x}) = -\frac{1}{2}\mathcal{I}(\widehat{\mu})(\mu - \widehat{\mu})^2$. (2.4) Dokážte, že ak $N(\mu, \sigma^2)$ a σ^2 je známe, potom $-\mathcal{I}^{-1/2}(\widehat{\mu})S(\mu) = \mathcal{I}^{1/2}(\widehat{\mu})(\mu - \widehat{\mu})$. DÚ

Príklad 3 (maximálne vierohodný odhad μ a σ^2) *Vygenerujte pseudonáhodné čísla X, ak platí $X \sim N(4,1)$, $n = 1000$. (a) Napíšte logaritmus profilovej funkcie vierohodnosti pre μ a σ^2 a preverte, či sú maximálne vierohodné odhady μ a σ^2 dostatočne blízko k ich skutočným hodnotám. Nakreslite grafy $l(\mu|\mathbf{x})$ a $l(\sigma^2|\mathbf{x})$, kde zvýrazníte polohu maxím týchto funkcií. (b) Napíšte logaritmus funkcie vierohodnosti pre $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ a preverte, či je maximálne vierohodný odhad $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ dostatočne blízko k jeho skutočnej hodnote. (c) Nakreslite graf $l((\mu, \sigma^2)|\mathbf{x})$ použitím funkcie `image()` a superponujte ho s kontúrovým grafom použitím funkcie `contour()`. Zvýraznite polohu maxima (pozri obrázok 14).* DÚ

Príklad 4 (maximálne vierohodné odhady; multinomické rozdelenie) *Majme dáta `more-samples-probabilities-pubis.txt`. Nakreslite logaritmus štandardizovanej funkcie vierohodnosti v parametroch p_1 a p_2 Európskej populácie ($n_1 = 30$, $n_2 = 20$ a $n_3 = 10$) pomocou funkcie `contour()`. Dokreslite do obrázku jej maximum v bode $\widehat{\boldsymbol{\theta}} = (\widehat{p}_1, \widehat{p}_2)^T$.* DÚ

Príklad 5 (argument minima) *Vygenerujte pseudonáhodné čísla $X \sim N(\mu, \sigma^2)$, $n = 1000$, $\mu = 0, \sigma^2 = 1$. Vygenerované čísla ozn. $x_i, i = 1, 2, \dots, 1000$. Nájdite numericky také c, ktoré minimalizuje (a) sumu štvorcov odchýlok $\sum_{i=1}^{1000}(x_i - c)^2$, t.j. $c_1 = \arg\min_{\forall c} \sum_{i=1}^{1000}(x_i - c)^2$ a (b) sumu absolútnych odchýlok $\sum_{i=1}^{100}|x_i - c|$, t.j. $c_2 = \arg\min_{\forall c} \sum_{i=1}^{1000}|x_i - c|$. Za c dosadzujte postupne (1) všetky $x_{(j)}$ ($x_{(j)}$ sú usporiadané x_i podľa veľkosti od najmenšieho po najväčšie) a vybrané charakteristiky polohy ako (2) aritmetický priemer, (3) nejaké kvantily $\widetilde{x}_p$, kde $p \in \langle 0, 1 \rangle$ a pod. Nakreslite obrázok závislosti (a) sumy štvorcov odchýlok na $x_{(j)}$, t.j. body $[x_j, y_j]$, kde $y_j = \sum_{i=1}^{1000}(x_i - x_{(j)})^2$ a (b) sumy absolútnych odchýlok na $x_{(j)}$, t.j. body $[x_{(j)}, y_j]$, kde $y_j = \sum_{i=1}^{1000}|x_i - x_{(j)}|$. Podobné obrázky nakreslite aj pre $\widetilde{x}_p$ namiesto $x_{(j)}$.* DÚ



Obr. 2: Profilová funkcia vierohodnosti pre μ (vľavo), σ^2 (uprostred) a funkcia vierohodnosti pre oba parametre (vpravo); $X \sim N(4,1)$); maximálne vierohodné odhady strednej hodnoty a rozptylu sú označené zvislou čiarkovanou čiarou (vľavo a uprotred) a maximálne vierohodný odhad vektora parametrov je označený $\bullet$ (vpravo)

Definícia 1 (asymptotické rozdelenie poriadkovej štatistiky) *Nech $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ sú poriadkové štatistiky náhodného výberu $X_1, X_2, \ldots, X_n$. Majme pravdepodobnosť α, kde $F(t_\alpha) = \alpha$. Asymptoticky platí, že $\sqrt{n}(\frac{j}{n} - \alpha)$ konverguje k 0. Potom je poriadková štatistika $X_{(j)}$ normálne rozdelená so strednou hodnotou $E[X_{(j)}] = t_\alpha$ a rozptylom $\sigma^2_{X_{(j)}} = \frac{\alpha(1-\alpha)}{f^2(t_\alpha)n}$. Ak $X \sim N(\mu, \sigma^2)$, potom $\sigma^2_{X_{(j)}} = \sigma^2 \frac{\pi^2}{24 \ln n}$.*

Príklad 6 (rozptyl poriadkovej štatistiky) *(a) Pomocou delta metódy odvoďte rozptyl poriadkovej štatistiky v definícii 1. (b) Pomocou definície 1 odvoďte rozptyl poriadkovej štatistiky, ak $X \sim N(\mu, \sigma^2)$.* DÚ

Definícia 2 (stredná hodnota a rozptyl mediánu) *Stredná hodnota mediánu $X_{(\frac{n+1}{2})}$ je rovná $E[X_{(\frac{n+1}{2})}] = \widetilde{\mu}$ a rozptyl mediánu $\sigma^2_{X_{(\frac{n+1}{2})}} = \frac{1}{4f^2(\widetilde{\mu})n}$, kde n je nepárne. Ak $X \sim N(\mu, \sigma^2)$, potom $\sigma^2_{X_{(\frac{n+1}{2})}} = \sigma^2 \frac{\pi}{2n}$.*

Príklad 7 (rozptyl mediánu) *(a) Pomocou delta metódy odvoďte rozptyl poriadkovej štatistiky v definícii 2. (b) Pomocou definície 2 odvoďte rozptyl poriadkovej štatistiky, ak $X \sim N(\mu, \sigma^2)$.* DÚ

Veta 1 (koeficient variácie) *Nech náhodná premenná X pochádza z normálneho rozdelenia s parametrami μ a σ^2, $X \sim N(\mu, \sigma^2)$, kde $E[X] = \mu$ je stredná hodnota a $Var[X] = \sigma^2$ je rozptyl náhodnej premennej X. Nech $g(\boldsymbol{\theta}) = \sigma/\mu$, kde $\boldsymbol{\theta} = (\mu, \sigma^2)^T$, $g(\widehat{\boldsymbol{\theta}}_n) = \frac{S_n}{\overline{X}_n}$ a $\Delta = \frac{\partial}{\partial \boldsymbol{\theta}} g(\boldsymbol{\theta}) = \left(-\frac{\sigma}{\mu^2}, \frac{1}{2\sigma\mu}\right)^T$. Potom*

$$\sqrt{n}\left(\frac{S_n}{\overline{X}_n} - \frac{\sigma}{\mu}\right) \overset{\mathcal{D}}{\sim} N_k\left(0, \Delta^T (i(\boldsymbol{\theta}))^{-1} \Delta\right),$$

kde $\Delta^T (i(\boldsymbol{\theta}))^{-1} \Delta = \left(-\frac{\sigma}{\mu^2}, \frac{1}{2\sigma\mu}\right) \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix} \left(-\frac{\sigma}{\mu^2}, \frac{1}{2\sigma\mu}\right)^T = \frac{\sigma^2}{\mu^2}\left(\frac{\sigma^2}{\mu^2} + \frac{1}{2}\right)$.

Príklad 8 (koeficient variácie) *Pomocou delta metódy odvoďte rozptyl koeficientu variácie z vety 1.* DÚ

Definícia 3 (hodnoty distribučnej funkcie v kvantiloch) *Empirická distribučná funkcia $F_n(x)$ je definovaná nasledovne*

$$F_n(x) = \begin{cases} 0, & ak\ x < X_{(1)}, \\ \frac{i}{n}, & ak\ X_{(i)} \leq x < X_{(i+1)}, \\ 1, & ak\ x \geq X_{(n)}. \end{cases}$$

Majme transformáciu $T_{(1)} = F_n(X_{(1)})$, $T_{(2)} = F_n(X_{(2)}), \ldots, T_{(n)} = F_n(X_{(n)})$. Potom $T_{(1)}$, $T_{(2)}$, $\ldots, T_{(n)}$ sú ***poriadkové štatistiky****. Potom platí*

$$\lim_{n\to\infty} \Pr(\sup_{\forall x \in \mathcal{Y}} [F_n(x) - F(x)] n^{1/2} \leq \lambda) = \Phi(\lambda),$$

kde $F(X)$ je teoretická distribučná funkcia a $\Phi(\lambda) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}$. Potom $100 \times (1-\alpha)\%$ pás spoľahlivosti pre $F(x)$ definujeme ako $F_n(x) \pm \lambda_\alpha 1/n^{1/2}$, kde $\Phi(\lambda_\alpha) = 1-\alpha$ a $Var[F_n(x)] = 1/n$. Potom môžeme tvrdiť, že $F(X)$ patrí do $100 \times (1-\alpha)\%$ pásu spoľahlivosti a zároveň je medzi nulou a jednotkou s pravdepodobnosťou $1-\alpha$.

Príklad 9 (graf distribučnej funkcie a jej IS) *Nakreslite graf distribučnej funkcie $N(\mu, \sigma^2)$, kde $\mu = 0$ a $\sigma^2 = 1$. Do grafu dokreslite 95% pás spoľahlivosti pre $F(x)$. Jeho hranice vypočítajte pomocou simulácie pseudonáhodných čísel z $N(0,1)$ pri $n = 50$, kde $F_n(x)$ je odhadnutá z dát. Teoretická distribučnú funkciu $\Phi(\lambda)$ naprogramujte v R alebo použite knižnicu `kolmim` a funkciu `pkolm()`; help k tejto funkcii je pristupný na* http://cran.r-project.org/web/packages/kolmim/index.html. DÚ