Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Štatistická inferencia II

Zadania domácich úloh

Stanislav Katina

katina@math.muni.cz

6. mája 2015

Inštrukcie k DÚ: Odovzdáva sa jeden pdf súbor nazvaný priezvisko-meno-text-statinf-II-2015.pdf (obsahuje riešenia príkladov, obrázky, R-kód napísaný v TEXu), jeden zdrojový súbor naprogramovaných funkcií priezvisko-meno–source-statinf-II-2015.r a jeden súbor R-kódu konkrétnych zadaní z DÚ priezvisko-meno-priklady-statinf-II-2015.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 (odvodenie testovacích štatistík) *Nech $X \sim N(\mu, \sigma^2)$, kde σ^2je naznáma. Majme $H_{01}: \mu = \mu_0$ vs. $H_{11}: \mu \neq \mu_0$. Odvoďte testovacie štatistiky*
(1) U_W použitím $U_W = (\widehat{\boldsymbol\theta} - \boldsymbol\theta_0)^T \mathcal{I}(\widehat{\boldsymbol\theta})(\widehat{\boldsymbol\theta} - \boldsymbol\theta_0)$ a
(2) U_S použitím $U_S = (S(\boldsymbol\theta_0))^T(\mathcal{I}(\boldsymbol\theta_0))^{-1}S(\boldsymbol\theta_0)$.

Za platnosti H_0 je pravdepodobnosť teoretickej CHPD rovná

$$\alpha_W = \Pr\left(U_W \geq \chi^2_{n-1}(\alpha)\right) = \Pr\left(F_{1,n-1} \geq \frac{n-1}{n}\chi^2_{n-1}(\alpha)\right),$$

$$\alpha_{\mathrm{LR}} = \Pr\left(U_{\mathrm{LR}} \geq \chi^2_{n-1}(\alpha)\right) = \Pr\left(F_{1,n-1} \geq (n-1)\left[\exp\left(\frac{\chi^2_{n-1}(\alpha)}{n}\right) - 1\right]\right)$$

a

$$\alpha_S = \Pr\left(U_S \geq \chi^2_{n-1}(\alpha)\right) = \Pr\left(F_{1,n-1} \geq \frac{n-1}{n - \chi^2_{n-1}(\alpha)}\chi^2_{n-1}(\alpha)\right),$$

Príklad 2 (pravdepodobnosť teoretickej CHPD pri danom n) *Odvoďte pravdepodobnosti (1) α_W, (2) α_{LR} a (3) α_S.*

Príklad 3 (Fisherova miera informácie) *Vypočítajte Fisherovu mieru informácie $\mathcal{I}(g(\widehat{p}))$, kde (a) $g(p) = \ln\frac{p}{1-p}$ a (b) $g(p) = \arcsin(\sqrt{p})$.*

Príklad 4 (Fisherova miera informácie) *Porovnajte DIS pre p použitím (a) Waldovho DIS, (b) vierohodnostného DIS, (c) skóre DIS, (d) transformovaného Waldovho DIS pre $g(p) = \frac{p}{1-p}$, (f) transformovaného Waldovho DIS pre $g(p) = \ln\frac{p}{1-p}$, (g) transformovaného Waldovho DIS pre $g(p) = \arcsin(\sqrt{p})$, kde (1) $N = 1400$ a $p = 0.05$, (2) $N = 40$ a $p = 0.05$, (3) $N = 1400$ a $p = 0.25$, (4) $N = 40$ a $p = 0.25$, (5) $N = 1400$ a $p = 0.5$, (6) $N = 40$ a $p = 0.5$.*

Príklad 5 (Fisherova miera informácie) *Majme dvojrozmerné normálne rozdelenie. Odvoďte pozorovanú Fisherovu informačnú maticu* $\mathcal{I}(\widehat{\boldsymbol{\theta}})$*, kde* $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \widehat{\rho})^T$*. Pomocou nej vypočítajte rozptyly jednotlivých elementov* $\widehat{\boldsymbol{\theta}}$*.*

Príklad 6 (pravdepodobnost pokrytia) *Nech* $X_i \sim Bin(N, p_i)$*. Vypočítajte pravdepodobnosti pokrytia:*
(a) vierohodnostného 95% DIS a (b) spätne tranformovaného Waldovho 95% DIS pre $g(p_i)$ *s hranicami* $(d_g^{(i)}, h_g^{(i)})$ *na Waldov 95% DIS pre* p_i *s hranicami* $\Big((g(d_g^{(i)}))^{-1}, (g(h_g^{(i)}))^{-1}\Big)$*, kde (1)* $g(p_i) = \frac{p_i}{1-p_i}$*, (2)* $g(p_i) = \ln \frac{p_i}{1-p_i}$ *a (3)* $g(p_i) = \arcsin(\sqrt{p_i})$ *pre každé* p_i*, kde* p_i *patria množine* $\mathcal{M}_I = \left\langle \frac{1}{N}, 1 - \frac{1}{N} \right\rangle$*, sú ekvidištantne vzdialené medzi* $\frac{1}{N}$ *a* $1 - \frac{1}{N}$ *a ich počet* $M = 5000$*. Nakreslite obrázok, kde na x-ovej osi budú* p_i *a na y-ovej osi pravdepodobnosť pokrytia* $\Pr_i$*(pokrytie). Zvoľte (a)* $N = 30$*, (b)* $N = 100$ *a (c)* $N = 1000$*. Pozn.: pravdepodobnosti pokrytia 95% DIS pre* p_i *vypočítame nasledovne*

$$\Pr_i(\textit{pokrytie}) = \textstyle\sum_j \Pr(X = Np_j : p_i \in 95\%\ \textit{DIS pre } p_j),$$

kde $p_j \in \mathcal{M}_J = \left\{\frac{1}{N}, \frac{2}{N}, \dots, 1 - \frac{1}{N}\right\}$*, t.j. ide o súčet takých funkčných hodnôt pravdepodobnostnej funkcie v bodoch* Np_j*, kde* $p_i \in 95\%$ *DIS pre* p_j*. Pre tie DIS, ktoré majú pre* $p = 0$ *a* $p = 1$ *nenulovú šírku, môžeme použiť* $\mathcal{M}_I = \left\langle \frac{0}{N}, \frac{N}{N} \right\rangle$*.*

Príklad 7 (test o pomere šancí) *Majme dáta* two-samples-probabilities-sexratio.txt*, premennú počet starších súrodencov* o.sib.N *a pohlavie* sex*. Predpokladáme, že početnosť chlapcov, ak nemajú staršieho súrodenca,* $X_{m,0} \sim Bin(N_{m,0}, p_{m,0})$*; početnosť chlapcov, ak ho majú,* $X_{m,1} \sim Bin(N_{m,1}, p_{m,1})$*; početnosť dievčat, ak nemajú staršieho súrodenca,* $X_{f,0} \sim Bin(N_{f,0}, p_{f,0})$*; početnosť dievčat, ak ho majú,* $X_{f,1} \sim Bin(N_{f,1}, p_{f,1})$*. (a) Otestujte hypotézu o pomere šancí narodenia chlapca (ak nemá staršieho súrodenca voči situácii, že ho má) použitím testu pomerom vierohodnosti na hladine významnosti* $\alpha = 0.05$*. (b) Vypočítajte* $100 \times (1 - \alpha)\%$ *empirický DIS pre tento pomer šancí pomocou metodiky 15% cut-off štandardizovanej (profilovej) funkcie vierohodnosti, kde koeficient spoľahlivosti* $1 - \alpha = 0.95$*. (c) Nakreslite funkciu vierohodnosti ako aj profilovú funkciu vierohodnosti a DIS.*

Aproximácia binomického rozdelenia Poissonovým rozdelením. Nech $X_j \sim Bin(N_j, p_j), j = 1, 2$, kde p_j sú veľmi malé čísla a N_j sú veľké čísla. Parametrom záujmu je **relatívne riziko** $\theta = p_1/p_2$. Preto môžeme predpokladať, že $X_1 \sim Poiss(\lambda_1), \lambda_1 = N_1 p_1$ a $X_2 \sim Poiss(\lambda_2), \lambda_2 = N_2 p_2$. Tiež môžeme situáciu zjednodušiť použitím $N_1 \approx N_2$. Potom funkcia vierohodnosti bude mať tvar

$$L(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2) = e^{-(N_1 p_1 + N_2 p_2)} (N_1 p_1)^{n_1} (N_2 p_2)^{n_2} = c e^{-p_2(N_1\theta + N_2)} \theta^{n_1} p_2^{n_1 + n_2}$$

kde $\boldsymbol{\theta} = (\theta, p_2)^T$, $c = N_1^{n_1} N_2^{n_2}$ (maximalizácia $L(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2)$ nezávisí na c) a profilova funkcia vierohodnosti

$$L(\theta|\mathbf{x}_1, \mathbf{x}_2) = \left(\frac{N_1\theta}{N_1\theta + N_2}\right)^{n_1} \left(1 - \frac{N_1\theta}{N_1\theta + N_2}\right)^{n_2}.$$

Príklad 8 (maximálne vierohodné odhady; nové liek vs. placebo) *Predpokladajme, že početnosť subjektov s infarktom myokardu (IM)* X_1 *v skupine A (nový liek) má binomické rozdelenie s parametrami* N_1 *a* p_1*, t.j.* $X_1 \sim Bin(N_1, p_1)$*, a počet subjektov s IM v skupine B (placebo) má tiež binomické rozdelenie s parametrami* N_2 *a* p_2*, t.j.* $X_2 \sim Bin(N_2, p_2)$*. Pozorovali sme* $n_1 = 139$ *z celkového* *počtu* $N_1 = 11037$ *subjektov a* $n_2 = 239$ *z celkového počtu* $N_2 = 11034$ *subjektov. Kedže* N_1 *a* N_2 *sú vysoké čísla a pravdepodobnosti IM malé čísla, môžeme predpokladať, že* $X_1 \sim Poiss(\lambda_1), \lambda_1 = N_1 p_1$ *a* $X_2 \sim Poiss(\lambda_2), \lambda_2 = N_2 p_2$*. Zjednodušme situáciu použitím* $N_1 \approx N_2$*. (a) Aplikujte funkciu vierohodnosti* $L(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2)$*, kde* $\boldsymbol{\theta} = (\theta, p_2)^T$ *a relatívne riziko* $\theta = p_1/p_2$ *na dáta v tabuľke – pre (1) IM ako aj (2) pre mozgovú mŕtvicu (MM). (b) Nakreslite profilové funkcie vierohodnosti pre každú skupinu. (c) Vypočítajte vierohodnostný* 95% *DIS pre (1)* θ_{IM} *a (2)* θ_{MM} *pomocou metodiky* 15% *cut-off štandardizovanej funkcie vierohodnosti.*

Tabuľka 1: Početnosti subjektov s infarktom myokardu a mozgovou mŕtvicou v skupine A a B

skupina	IM	MM	spolu
skupina A	139	119	11037
skupina B	239	98	11034