Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Statistická inference I

Zadání domácího úkolu – rok 2016

Stanislav Katina, Veronika Bendová

katina@math.muni.cz, 375612@math.muni.cz

12. prosince 2016

Instrukce k domácímu úkolu: Odevzdává se jeden pdf soubor nazvaný prijmeni-jmeno-text-statinf-I-2016.pdf (obsahuje řešení příkladů, obrázky, R-kód napsaný v TEXu), jeden zdrojový soubor naprogramovaných funkcí prijmeni-jmeno–source-statinf-I-2016.R a jeden soubor R-kódu konkrétních zadání z DÚ prijmeni-jmeno-priklady-statinf-I-2016.R, který používá tento zdrojový kód. Dejte si záležet na přehlednosti programovaného kódu, na doplnění komentářů a vhodného užití zavedených pravidel, které máte k dispozici v prezentaci Standards of programming in R: R style guide. Také věnujte svou pozornost a čas dostatečným popisům vašich úvah a zvolených postupů a interpretacím výsledků, ať už číselných nebo grafických. I to bude součástí celkového hodnocení úkolu. Na psaní R-kódu doporučuji TEXovský balíček listings a vytvoření prostředí v hlavičce dokumentu pomocí následujícího kódu:

```
\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
```

V textu potom kód vkládáme do prostředí \begin{lstlisting} a \end{lstlisting}.

DÚ je nutné odevzdat 7 dní před termínem zkoušky, na který se přihlásíte.

Příklad 1. Jana s Bárou a Vojtíškem dostali hořko-mléčný adventní kalendář, ve kterém je polovina čokolád hořkých a polovina čokolád mléčných, přičemž příchutě čokolád jsou v kalendáři rozmístěny náhodně. O čokolády se děti rozhodly podělit rovným dílem, ale protože je Vojtíšek nejmenší, dovolily mu sestry, aby svůj díl čokolád snědl jako první.

1. Vypočítejte, jaká je pravděpodobnost, že Vojtíšek, který vůbec nemá rád hořkou čokoládu, bude mít ve svém dílu:

   a) všechny čokolády mléčné,

   b) maximálně jednu čokoládu hořkou,

   c) více než polovinu přidělených čokolád mléčných.

   Získáné pravděpodobnosti interpretujte, vždy uveďte odpověď.

2. Nakreslete graf pravděpodobnostní a distribuční funkce rozdělení, které popisuje rozložení počtu mléčných čokolád ve Vojtíškově přídělu. Grafy řádně okomentujte.

Příklad 2. Načtěte datový soubor 01-one-sample-mean-skull-mf.txt obsahující údaje o délce a šířce lebky ze starověké egyptské populace.

1. Vytvořte tabulku základních charakteristik (průměr, rozptyl, směrodatná odchylka, medián, 1.kvartil, 3.kvartil, IQR, koeficient šikmosti a špičatosti) pro délku a šířku lebky starověké egyptské populace.

2. Nakreslete histogramy (v rel. škále) pro délku a šířku lebky starověké egyptské populace.

3. Vypočítejte Pearsonův korelační koeficient pro vztah délky a šířky lebky egyptské populace a závislost obou veličin demonstrujte pomocí tečkového grafu.

4. Vypočítejte dvourozměrný jádrový odhad hustoty délky a šířky lebky starověké egyptské populace, zakreslete jej pomocí funkce image() a superponujte jej

   a) konturovými křivkami dvourozměrného jádrového odhadu,

   b) teoretickými konturami dvourozměrného normálního rozdělení, kde střední hodnoty μ_1 a μ_2, rozptyly σ_1^2 a σ_2^2 a korelační koeficient ρ nahraďte jejich MLE odhady získanými z dat.
   Poznámka: Pohlídejte si, aby kontury přesně seděly s barevnými přechody image *grafu.*

5. Dvourozměrný jádrový odhad hustoty vykreslete také pomocí funkce persp(). Hustotu rozsekejte na 12 intervalů, kde hodnoty v těchto intervalech budou odpovídat barvám terrain.colors(12).

6. Bylo v tomto případě vhodné použít Pearsonův korelační koeficient k určení závislosti mezi délkou a šířkou lebky? Kterou informaci z datového souboru jsme při počítání korelačního koeficientu zanedbali a jak se to projevilo v grafech dvourozměrného jádrového odhadu?

7. Zohledněte informaci, kterou jsme na začátku příkladu zanedbali, a vygenerujte nový tečkový graf, oddělující barevně zanedbanou vlastnost, a superponujte jej konturami dvourozměrného jádrového odhadu. Určete nové korelační koeficienty zohledňující zanedbanou informaci z datového souboru.

Všechny výsledky a grafy okomentujte, své závěry a postupy zdůvodněte. Nezapomeňte na přesnou interpretaci Pearsonových korelačních koeficientů.

Příklad 3. 1. Nakreslete škálovaný logaritmus profilové funkce věrohodnosti normálního rozdělení pro μ. Na ose x bude μ a na ose y $\ln \mathcal{L}_P(\mu|\mathbf{x}) = l_P(\mu|\mathbf{x}) - \max(l_P(\mu|\mathbf{x}))$. Porovnejte $\ln \mathcal{L}_P(\mu|\mathbf{x})$ s kvadratickou aproximací vypočítanou pomocí Taylorova rozvoje $\ln \mathcal{L}_P(\mu|\mathbf{x}) = \ln(\frac{L_P(\mu|\mathbf{x})}{L_P(\widehat{\mu}|\mathbf{x})}) \approx -\frac{1}{2}\mathcal{I}(\widehat{\mu})(\mu - \widehat{\mu})^2$.

2. Nechť skóre funkce $S(\mu) = \frac{\partial}{\partial \mu} \ln L_P(\mu|\mathbf{x})$. Vezmeme-li derivaci kvadratické aproximace uvedené výše, dostaneme $S(\mu) \approx -\mathcal{I}(\widehat{\mu})(\mu - \widehat{\mu})$ nebo $-\mathcal{I}^{-1/2}(\widehat{\mu})S(\mu) \approx \mathcal{I}^{1/2}(\widehat{\mu})(\mu - \widehat{\mu})$. Potom zobrazením pravé strany na ose x a levé strany na ose y dostaneme asymptoticky lineární funkci s jednotkovým sklonem. Je postačující mít rozsah osy x rovný $\langle -2, 2\rangle$, protože funkce je asymptoticky (lokálně) lineární na tomto intervalu. Rozumně škálujte osu y. Zobrazte pro (a) $n = 10$, (b) $n = 100$ a (c) $n = 1000$. Použijte (1) $X \sim N(0, 1)$ a (2) $X \sim (1 - p)N(0, 1) + pN(0, 2)$, kde $p = 0.05$. Okomentujte rozdíly mezi (a), (b) a (c), stejně jako rozdíly mezi (1) a (2).

Příklad 4. maximálně věrohodný odhad μ a σ^2

Vygenerujte pseudonáhodná čísla z $X \sim N(4, 1)$, $n = 1000$.

1. Napište logaritmus odhadnuté a profilové funkce věrohodnosti pro μ a σ^2 a porovnejte maximálně věrohodné odhady parametrů μ a σ^2 získané maximalizací těchto funkcí. Nakreslete grafy $l_e(\mu|\mathbf{x})$, $l_P(\mu|\mathbf{x})$, $l_e(\sigma^2|\mathbf{x})$ a $l_P(\sigma^2|\mathbf{x})$, kde zvýrazníte polohu maxim těchto funkcí.

2. MLE odhady parametrů μ a σ^2 z bodu 1 najděte pomocí

   (a) funkce optimize(),

   (b) Newton-Raphsonovy metody (naprogramujte),

   (c) metody sečen (naprogramujte).

   Získané odhady zaokrouhlete na šest desetinných míst, uspořádejte do přehledné tabulky a vzájemně porovnejte.

	optimize()	Newton-Raphson	metoda.secen
$\widehat{\mu}_e$			
$\widehat{\mu}_P$			
$\widehat{\sigma}^2_e$			
$\widehat{\sigma}^2_P$			

   Porovnejte také numerické metody, s jejichž pomocí byly parametry odhadnuty. Která z použitých numerických metod je pro odhadování parametrů efektivnější a proč? Porovnání získaných odhadů a numerických metod slovně popište, rozeberte a své závěry zdůvodněte.

3. Napište logaritmus funkce věrohodnosti pro $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ a prověřte, zda je maximálně věrohodný odhad $\widehat{\boldsymbol{\theta}}$ dostatečně blízko k jeho skutečné hodnotě. Nakreslete graf $l(\boldsymbol{\theta}|\mathbf{x})$, kde na x-ové ose bude parametr μ a na y-ové ose parametr σ^2, použitím funkce image() a superponujte ho konturovým grafem použitím funkce contour(). Zvýrazněte polohu maxima.

Příklad 5. Simulační studie Na základě simulační studie ověřte, že pokud

a) $X \sim N(\mu, \sigma^2)$, kde $\mu = 0$, $\sigma^2 = 1$,

b) $X \sim [(1-p)N(\mu, \sigma_1^2) + pN(\mu, \sigma_2^2)]$, kde $\mu = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $p = 0.1$,

potom

1. testovací statistika
$$t_W = \frac{\bar{x} - \mu}{s}\sqrt{n}$$
má Studentovo rozdělení o $n-1$ stupních volnosti $T_W \sim t_{n-1}$,

2. testovací statistika
$$t_W^2 = \left(\frac{\bar{x} - \mu}{s}\sqrt{n}\right)^2$$
má Fisherovo rozdělení o 1 a $n-1$ stupních volnosti $T_W^2 \sim F_{1,n-1}$.

Použijte rozsahy náhodných výběrů $n = 15$ a $n = 500$. Pro každou simulaci X vypočítejte $t_{W,m}$ (resp. $t_{W,m}^2$), $m = 1, 2, \ldots, M$, kde $M = 1000$. Superponujte histogram vygenerovaných testovacích statistik v relativní škále s teoretickou křivkou hustoty Studentova (resp. Fisherova rozdělení). Vygenerované histogramy pro $n = 15$ a $n = 500$ jakožto i pro rozdělení (a) a (b) vzájemně srovnejte a srovnání slovně okomentujte.

Příklad 6. Poissonovo rozdělení: část 1 – odvození Nechť náhodná veličina X pochází z Poissonova rozdělení, $X \sim Po(\lambda)$ a realizace $X = x$.
Pro náhodou veličinu X odvoďte

1. pravděpodobnostní funkci $f(x, \lambda)$,
2. věrohodnostní funkci (+ uveďte, jak vypadá jádro) $L(\lambda|x)$,
3. logaritmus věrohodnostní funkce (+ uveďte, jak vypadá jádro) $l(\lambda|x)$,
4. skóre funkci $S(\lambda)$,
5. maximálně věrohodný odhad parametru λ, ozn. $\widehat{\lambda}$,
6. pozorovanou Fisherovu míru informace $\mathcal{I}(\widehat{\lambda})$,
7. rozptyl odhadu parametru λ, ozn. $\widehat{\mathrm{Var}[\widehat{\lambda}]}$.

Příklad 7. Poissonovo rozdělení: část 2 – maximálně věrohodné odhady Mějme početnosti úrazů mezi dělníky v továrně, kde početnosti dělníků m_n při daném počtu úrazů n jsou uvedeny v následující tabulce (Greenwood a Yule (1920)).

n	0	1	2	3	4	≥5
m_n	447	132	42	21	3	2

1. Vypočítejte maximálně věrohodný odhad parametru λ
   (a) pomocí maximalizace věrohodnostní funkce; výsledek zobrazte do grafu s křivkou věrohodnostní funkce Poissonova rozdělení,
   (b) pomocí maximalizace logaritmu věrohodnostní funkce; výsledek zobrazte do grafu s křivkou logaritmu věrohodnostní funkce Poissonova rozdělení.
2. Naprogramujte metodu sečen a s její pomocí vypočítejte maximálně věrohodný odhad parametru λ. Výsledek zobrazte do grafu s
   (a) křivkou věrohodnostní funkce Poissonova rozdělení,
   (b) křivkou logaritmu věrohodnostní funkce Poissonova rozdělení.
3. Vypočítejte odhad rozptylu odhadu parametru λ získaného pomocí
   (a) maximalizace věrohodnostní funkce,
   (b) maximalizace logaritmu věrohodnostní funkce,
   (c) metody sečen.
4. Všechny tři odhady parametru λ a k nim příslušné odhady rozptylů zaokrouhlete na šest desetinných míst a uspořádejte do přehledné tabulky

	maximalizace $L(\lambda\|\mathbf{x})$	maximalizace $l(\lambda\|\mathbf{x})$	metoda sečen
$\widehat{\lambda}$			
$\widehat{\mathrm{Var}[\widehat{\lambda}]}$			

   Získané odhady porovnejte s explicitně vyjádřenými odhady (zaokrouhlenými též na šest desetinných míst). Porovnání slovně opište, stejně jako vygenerované grafy.

Příklad 8. Poissonovo rozdělení: část 3 – kvadratická aproximace logaritmu funkce věrohodnosti

1. Pro data z příkladu 7 nakreslete škálovaný logaritmus funkce věrohodnosti Poissonova rozdělení. Na x-ové ose bude λ a na y-ové ose $\ln \mathcal{L}(\lambda|\mathbf{x}) = l(\lambda|\mathbf{x}) - \max(l(\lambda|\mathbf{x}))$. Porovnejte $\ln \mathcal{L}(\lambda|\mathbf{x})$ s kvadratickou aproximací vypočítanou pomocí Taylorova rozvoje $\ln \mathcal{L}(\lambda|\mathbf{x}) = \ln\left(\frac{L(\lambda|\mathbf{x})}{L(\widehat{\lambda}|\mathbf{x})}\right) \approx -\frac{1}{2}\mathcal{I}(\widehat{\lambda})(\lambda - \widehat{\lambda})^2$.
2. Nechť skóre funkce $S(\lambda) = \frac{\partial}{\partial\lambda} \ln L(\lambda|\mathbf{x})$. Vezmeme-li derivaci kvadratické aproximace uvedené výše, dostaneme $S(\lambda) = -\mathcal{I}(\widehat{\lambda})(\lambda - \widehat{\lambda})$ anebo
$$-\mathcal{I}^{-1/2}(\widehat{\lambda})S(\lambda) \approx \mathcal{I}^{1/2}(\widehat{\lambda})(\lambda - \widehat{\lambda}). \quad (1)$$
   Potom zobrazením pravé strany na x-ové ose a levé strany na y-ové ose dostaneme asymptoticky lineární funkci s jednotkovým sklonem. Nakreslete graf, kde na x-ové ose bude vynesena pravá strana rovnosti 1 a na y-ové ose levá strana rovnosti 1. Křivku superponujte lineární křivkou $x = y$. Rozumně škálujte x-vou a y-vou osu. Vygenerované grafy řádně okomentujte.