Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Statistická inference II

Zadání domácích úkolů – rok 2016

Stanislav Katina, Veronika Bendová

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

11. května 2016

Instrukce k DÚ: Odevzdávají se celkem tři soubory:

1. pdf soubor nazvaný prijmeni-jmeno-text-statinf-II-2016.pdf ... tento soubor obsahuje podrobně okomentovaný postup řešení příkladů doplněný stěžejními pasážemi vybranými z R-kódu, přehledně zpracované výsledky příkladů (obrázky, tabulky) a dostatečně podrobné interpretace těchto výsledků. Pdf soubou je vytvořený v TEXu.

2. R skript nazvaný prijmeni-jmeno–source-statinf-II-2016.R ... zdrojový soubor obsahující pouze vlastní naprogramované funkce, které jsou k řešení domácího úkolu využívány,

3. R skript nazvaný prijmeni-jmeno-priklady-statinf-II-2016.R... skript obsahující kompletní řešení zadaných příkladů, využívající funkce obsažené ve skriptu prijmeni-jmeno–source-statinf-II-2016.R.

   Poznámka: Snažte se ve skriptu prijmeni-jmeno-priklady-statinf-II-2016.R vyhnout zbytečnému opakování stejných částí kódu. Pokud máte v řešení zadáno provést daný postup např. pro $n = 5$, $n = 10$ a $n = 100$, naprogramujte funkci se vstupem n, tu vložte do skriptu prijmeni-jmeno–source-statinf-II-2016.R a ve skriptu prijmeni-jmeno-priklady-statinf-II-2016.R ji už pouze zavolejte.

Na psaní pasáží R-kódu doporučujeme použít TEXovský balíček listings a vytvořit 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 č.1 (metoda tečen)
Naprogramujte v R metodu tečen. Rychlost konvergence této metody demonstrujte na minimalizaci funkce $x^3 - \sin^2(x)$ a porovnejte ji s rychlostí konvergence metody bisekce, kterou máte naprogramovanou ze cvičení. K metodě tečen doprogramujte graf zobrazující konvergenci posloupnosti bodů ke kořenu funkce $x^3 - \sin^2(x)$.

Požadovaná forma výstupu příkladu:

- tabulka výsledků:

metoda	kořen	počítadlo	chyba
Newtonova metoda			
Metoda sečen			

- dva grafy zobrazující konvergenci posloupnosti bodů ke kořenu funkce: jeden pro metodu tečen, druhý pro metodu bisekce,
- komentář porovnávající obě metody na základě grafů a hodnot uvedených v tabulce.

Příklad č.2 (nezávislost μ a σ^2; pravděpodobnost pokrytí)
Nechť $X \sim [pN(\mu, \sigma_1^2) + (1-p)N(\mu, \sigma_2^2)]$, kde $p = 0.9$, $\mu = 20$, $\sigma_1^2 = 100$ a $\sigma_2^2 = 400$. Pomocí simulační studie vypočítejte Pearsonův korelační koeficient $r_{\bar{X},S}$. Nakreslete šedou barvou rozptylový graf $(\bar{x}_m, s_m)$, kde $m = 1, 2, \ldots, M$, přičemž $M = 50\,000$. Černou barvou vyznačte v grafu takové body $(\bar{x}_m, s_m)$, pro které platí $t_{W,m} = \left|\frac{\bar{x}_m - \mu}{s_m}\sqrt{n}\right| < t_{n-1}(\alpha/2)$. Dále vykreslete hranice, které jsou definovány body $(\bar{x}_m, s_m)$, jež splňují vztah $t_{W,m} = t_{n-1}(\alpha/2)$. Vypočítejte pravděpodobnost pokrytí 95 % DIS pro μ jako podíl $\sum_m I(t_{W,m} < t_{n-1}(\alpha/2))/M$. Zvolte (a) $n = 5$, (b) $n = 50$ a (c) $n = 100$.

Požadovaná forma výstupu příkladu:

- tabulka výsledků:

n	$r_{\bar{X},s}$	$\widehat{\mu}_D$	$\widehat{\mu}_H$	Pr(pokrytí)
$n = 5$				
$n = 50$				
$n = 100$				

  kde $\widehat{\mu}_D$ je dolní hranice 95 % DIS a $\widehat{\mu}_H$ je horní hranice 95 % DIS,
- tři rozptylové grafy: pro $n = 5$, $n = 50$, $n = 100$.
- komentář k výsledkům uvedeným v tabulce a k vykresleným grafům. Co pozorujeme s rostoucím n?

Příklad č.3 (konvergence ρ a ξ k normálnímu rozdělení)
Proveďte v R simulaci pseudonáhodných čísel z $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 1$, kde $n = 5$, 50 a 100, $M = 10\,000$. Použijte (a) $\rho = 0$, (b) $\rho = 0.5$ a (c) $\rho = 0.9$. Pro každé $m = 1, 2, \ldots, M$, vypočítejte Pearsonův korelační koeficient r_m a Fisherovu Z-proměnnou $z_{R,m}$. Zobrazte histogramy simulovaných r_m a $z_{R,m}$ a superponujte je teoretickými hustotami příslušných normálních rozdělení.

Požadovaná forma výstupu příkladu:

- devět histogramů pro r_m: pro $n = 5$, $n = 50$, $n = 100$; $\rho = 0$, $\rho = 0.5$ a $\rho = 0.9$,
- devět histogramů pro $z_{R,m}$: pro $n = 5$, $n = 50$, $n = 100$; $\rho = 0$, $\rho = 0.5$ a $\rho = 0.9$,
- komentář k vykresleným grafům. Co pozorujeme s rostoucím n?

Příklad č.4 (minimální rozsah souboru)
Vypočítejte v R minimální rozsah náhodného výběru pro test $H_{03} : \sigma^2 \geq \sigma_0^2$ oproti $H_{13} : \sigma^2 < \sigma_0^2$, když $\alpha = 0.05$ a $1 - \beta = 0.8$, za předpokladu, že podíl $\frac{\sigma_0^2}{\sigma^2}$ je rovný (a) 1.1, (b) 1.5 a (c) 5.

Požadovaná forma výstupu příkladu:

- tabulka výsledků:

$\frac{\sigma_0^2}{\sigma^2}$	minimální rozsah n	$1 - \beta$
1.1		
1.5		
5		

  kde $1 - \beta$ je síla testu vypočítaná pro nalezený minimální rozsah n,

- komentář k výsledkům uvedeným v tabulce.

Příklad č.5 (párový test o střední hodnotě μ_d)
Mějme data paired-means-clavicle.txt a proměnnou *vertikální průměr ve středu délky těla klíční kosti* v mm (průměr prvního a druhého měření simd.1 a simd.2) na levé (side se rovná L) a pravé (side se rovná R) straně. Tato měření vykonal první výzkumník. Druhý výzkumník změřil *vertikální průměr ve středu délky těla klíční kosti* pouze jedenkrát (simd). Zajímá nás rozdíl měření obou výzkumníků (**interindividuální chyba**) na levé straně. O tomto rozdílu X_d předpokládáme, že má normální rozdělení $N(\mu_d, \sigma_d^2)$.

1. Otestujte hypotézu o shodě střední hodnoty rozdílu vertikálního průměru ve středu délky těla klíční kosti na levé straně u prvního a druhého výzkumníka na hladině významnosti $\alpha = 0.05$.
2. Vypočítejte $100 \times (1-\alpha)$ % empirický DIS, tj. $(\widehat{\mu}_D\,;\,\widehat{\mu}_H)$ pro střední hodnotu rozdílu vertikálního průměru ve středu délky těla klíční kosti na levé a pravé straně, kde koeficient spolehlivosti $1 - \alpha = 0.95$.

Jak v části (1), tak i v části (2) použijte k otestování nulové hypotézy a ke stanovení příslušných DIS

(a) Waldovu testovací statistiku $U_W^{(d)}$,

(b) skóre testovací statistiku $U_S^{(d)}$,

(c) věrohodnostní testovací statistiku $U_{LR}^{(d)}$.

Požadovaná forma výstupu příkladu:

- tabulka výsledků:

Statistika	$\widehat{\mu}_d$	test. stat.	$\widehat{\mu}_D$	$\widehat{\mu}_H$	p-hodnota
$U_W^{(d)}$					
$U_S^{(d)}$					
$U_{LR}^{(d)}$					

- komentář k výsledkům uvedeným v tabulce + zdůvodněné rozhodnutí o tom, kterou testovací statistiku byste v praxi pro konečnou analýzu využili.

Příklad č.6 (test rozptylu σ^2)
Mějme data one-sample-variance-skull-mf.txt a proměnnou skull.B – šířka lebky (v mm) starověké egyptské mužské populace, o které předpokládáme, že má normální rozdělení $N(\mu, \sigma^2)$.

1. Otestujte hypotézu o shodě rozptylu šířky lebky této populace s rozptylem šířky lebky novověké egyptské populace 6.411^2 na hladině významnosti $\alpha = 0.05$.
2. Vypočítejte $100 \times (1-\alpha)\,\%$ empirický DIS, tj. $(\widehat{\sigma}_{\mathsf{skull.B},D}\,;\, \widehat{\sigma}_{\mathsf{skull.B},H})$ pro rozptyl šířky lebky, kde koeficient spolehlivosti $1-\alpha = 0.95$.

Jak v části (1), tak i v části (2) použijte k otestování nulové hypotézy a ke stanovení příslušných DIS

(a) exaktní testovací statistiku F_W,

(b) Waldovu testovací statistiku U_W,

(c) skóre testovací statistiku U_S,

(d) věrohodnostní testovací statistiku U_{LR}.

Požadovaná forma výstupu příkladu:

- tabulka výsledků:

Statistika	$\widehat{\sigma}_{\mathsf{skull.B}}$	test. stat.	$\widehat{\sigma}_{\mathsf{skull.B},D}$	$\widehat{\sigma}_{\mathsf{skull.B},H}$	p-hodnota
F_W					
U_W					
U_S					
U_{LR}					

- komentář k výsledkům uvedeným v tabulce + zdůvodněné rozhodnutí o tom, kterou testovací statistiku byste v praxi pro konečnou analýzu využili.

Příklad č.7 (pravděpodobnost pokrytí)
Nechť $X_i \sim Bin(N, p_i)$. Vypočítejte pravděpodobnosti pokrytí zpětně transformovaného Waldova 95 % DIS pro $g(p_i)$ s hranicemi $(d_g^{(i)}, h_g^{(i)})$ na Waldův 95 % DIS pro p_i s hranicemi $\Big((g(d_g^{(i)}))^{-1}, (g(h_g^{(i)}))^{-1}\Big)$, kde

1. $g(p_i) = \dfrac{p_i}{1-p_i}$,
2. $g(p_i) = \ln \dfrac{p_i}{1-p_i}$,
3. $g(p_i) = \arcsin(\sqrt{p_i})$.

pro každé p_i, kde p_i náležící množině $\mathcal{M}_I = \langle \frac{1}{N}, 1-\frac{1}{N} \rangle$ jsou ekvidistantně vzdálené hodnoty ležící mezi $\frac{1}{N}$ a $1-\frac{1}{N}$ a jejich počet je $M = 5\,000$. Nakreslete obrázek, kde na ose x budou pravděpodobnosti p_i a na ose y bude pravděpodobnost pokrytí $\Pr_i$(pokrytí). Zvolte (a) $N = 30$, (b) $N = 100$ a (c) $N = 1000$.

Poznámka: Pravděpodobnosti pokrytí 95 % DIS pro p_i vypočítáme následovně

$$\Pr_i(\text{pokrytí}) = \sum_j \Pr(X = Np_j : p_i \in 95\,\%\text{ DIS pro } p_j),$$

kde $p_j \in \mathcal{M}_J = \{\frac{1}{N}, \frac{2}{N}, \ldots, 1-\frac{1}{N}\}$, t.j. jde o součet takových funkčních hodnot pravděpodobnostní funkce v bodech Np_j, kde $p_i \in 95\,\%$ DIS pro p_j. Pro ty DIS, které mají pro $p = 0$ a $p = 1$ nenulovou šířku, můžeme použít $\mathcal{M}_I = \langle \frac{0}{N}, \frac{N}{N} \rangle$

Požadovaná forma výstupu příkladu:

- devět grafů pravděpodobností pokrytí: pro $n = 30$, $n = 100$, $n = 1000$; pro $g(p_i) = \dfrac{p_i}{1-p_i}$, $g(p_i) = \ln \dfrac{p_i}{1-p_i}$, $g(p_i) = \arcsin(\sqrt{p_i})$,
- devět grafů pravděpodobností pokrytí: pro $n = 30$, $n = 100$, $n = 1000$; pro Waldův DIS, skóre DIS a věrohodnostní DIS (uvedeno pouze pro srovnání; probráno a naprogramováno na cvičeních; R-kód této pasáže ve skriptech **neuvádějte**; pouze do pdf souboru vykreslete grafy),
- řádně okomentujte rozdíly v pravděpodobnostech pokrytí jednotlivých zpětně transformovaných Waldových 95 % DIS,
- srovnejte vlastnosti zpětně transformovaných Waldových 95 % DIS s vlastnostmi Waldova, skóre a věrohodnostního DIS, které jsme programovali na cvičeních (z hlediska pravděpodobnosti pokrytí),
- uveďte odpověď na otázku: Který typ ze všech studovaných intervalů byste použili na odhadování parametru p binomického rozdělení ($N \leq 30$), v případě, že byste o p předpokládali, že bude nabývat buď příliš malé (např. $p = 0.05$) nebo příliš vysoké ($p = 0.9$) hodnoty? Své rozhodnutí zdůvodněte.