Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Statistická inference II

Zadání domácího úkolu – rok 2017

1. část

Stanislav Katina, Veronika Bendová

katina@math.muni.cz, xbendovav@math.muni.cz

25. května 2017

Instrukce k domácímu úkolu: Odevzdává se jeden pdf soubor nazvaný prijmeni-jmeno-text-statinf-II-2017.pdf (obsahuje řešení příkladů, obrázky, R-kód napsaný v TEXu), jeden zdrojový soubor naprogramovaných funkcí prijmeni-jmeno–source-statinf-II-2017.R a jeden soubor R-kódu konkrétních zadání z DÚ prijmeni-jmeno-priklady-statinf-II-2017.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 slovních 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}.

Kompletní řešení domácího úkolu je nutné nahrát do odevzdávárny v IS nejpozději 7 dní před termínem zkoušky, na který se přihlásíte.

Příklad 1. Vylepšená věrohodnost pomocí $g(\theta)$:

1. Nakreslete logaritmus relativní funkce věrohodnosti parametru p binomického rozdělení $\text{Bin}(N, p)$, kde $N = 10$ a $n = 8$, superponovaný jeho kvadratickou aproximací.
2. Nakreslete logaritmus relativní funkce věrohodnosti $g(p) = logit(p) = \ln \frac{p}{1-p}$ (při stejném zadání N a n jako v (1)), superponovaný jeho kvadratickou aproximací.
3. Nakreslete graf porovnávající vzájemně logaritmus relativní funkce věrohodnosti s její kvadratickou aproximací získanou na základě parametru p (ad 1) a aproximací získanou za základě parametrické funkce $g(p)$ (ad 2).
4. Vypočítejte Waldův a věrohodnostní $100 \times (1 - \alpha)\%$ empirický DIS pro p.
5. Vypočítejte Waldův a věrohodnostní $100 \times (1 - \alpha)\%$ empirický DIS pro $g(p)$ z bodu (2) a transformujte jej zpět do originální škály.
6. Vzájemně porovnejte Waldovy empirické DIS pro p a pro $g(p)$ po zpětné transaformaci do originální škály a věrohodnostní empirické DIS pro p a pro $g(p)$ po zpětné transaformaci do originální škály. Který z intervalů vykazuje lepší vlastnosti a proč?
7. Naprogramujte dvě numerické metody: metodu bisekce (funkce `bisekce()`) a metodu sečen (funkce `metoda.secen()`) ke zpřesnění hranic věrohodnostních intervalů spolehlivosti.

Požadovaná forma výstupu příkladu:

- dvě samostatně použitelné funkce `bisekce()` a `metoda.secen()` s naimplementovanými iteračními metodami;
- trojice grafů:
  - (i) graf s parametrem p na ose x a log. rel. věroh. funkcí + její kvadratickou aproximací na ose y;
  - (ii) graf s param. funkcí $g(p)$ na ose x a příslušnou log. rel. věroh. funkcí + její kvadr. aproximací na ose y;
  - (iii) graf s parametrem p na ose x a log. rel. věrohodnostní funkcí + její kvadr. aproximací pomocí parametru p a pomocí param. funkce $g(p)$ na ose y (na základě tohoto grafu porovnejte kvalitu obou kvadr. aproximací);
- tabulka hranic intervalů spolehlivosti:

parametr/p.funkce	Waldův IS – dh	Waldův IS – hh	Věroh. IS – dh	Věroh. IS – hh
p				
$g(p)$				
$g(p)$ zpětně transf. do škály p				

- tabulka přesnějších hranic věrohodnostních intervalů spolehlivosti získaných pomocí vlastnoručně naprogramovaných funkcí `bisekce()` a `metoda.secen()`:

parametr/p.funkce	pův. – dh	pův. – hh	m.bisekce – dh	m.bisekce – hh	m.sečen – dh	m.sečen – hh
p						
$g(p)$						
$g(p)$ z.t. do šk.p						

Poznámka: Hodnoty ve výsledných tabulkách i textu zaokrouhlete na šest desetinných míst.

Příklad 2. Test o směrodatné odchylce σ: Z archivních materiálů máme k dispozici původní kraniometrické údaje o délce lebky mužů a žen ze starověké egyptské populace (soubor one-sample-mean-skull-mf.csv). Současně máme k dispozici průměrné hodnoty délky lebky a hodnoty směrodatných odchylek pro muže a ženy novověké egyptské populace (délka lebky mužů $x_m = 177.568$ mm se směrodatnou odchylkou $s_m = 7.526$ mm; délka lebky žen $x_f = 171.962$ mm se směrodatnou odchylkou $s_f = 7.052$ mm; rozsah datového souboru $n_m = 88$, $n_f = 52$).

Načtěte datový soubor one-sample-mean-skull-mf.csv, kde proměnná skull.L označuje délku lebky (v mm) starověké egyptské populace a proměnná sex označuje pohlaví měřeného jedince. Zaměřte se na délku lebky žen, o které předpokládáme že má normální rozdělení $N(\mu, \sigma^2)$.

1. Otestujte nulovou hypotézu, že směrodatná odchylka délky lebky žen u starověké egyptské populace je rovna směrodatné odchylce délky lebky žen u novověké egyptské populace.

2. Vypočítejte $100 \times (1-\alpha)\%$ empirický DIS, tj. $(\widehat{\sigma}_D\,;\,\widehat{\sigma}_H)$ pro směrodatnou odchylku délky lebky žen starověké egyptské populace, 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;

(b) skóre testovací statistiku U_S;

(c) věrohodnostní testovací statistiku U_{LR}.

Požadovaná forma výstupu příkladu:

- H_0;
- H_1;
- tabulka výsledků:

Statistika	$\widehat{\sigma}$	test. stat.	$\widehat{\sigma}_D$	$\widehat{\sigma}_H$	p-hodnota
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.

Poznámka: Hodnoty ve výsledné tabulce i textu zaokrouhlete na čtyři desetinná místa.

Příklad 3. Pokračování příkladu 2: Vraťme se nyní k datovému souboru one-sample-mean-skull-mf.csv, kde proměnná skull.L označuje délku lebky (v mm) starověké egyptské populace a proměnná sex označuje pohlaví měřeného jedince.

1. Na hladině významnosti $\alpha = 0.05$ otestujte nulovou hypotézu, že směrodatná odchylka délky lebky žen u starověké egyptské populace je větší než směrodatná odchylka délky lebky žen u novověké egyptské populace. Testování proveďte pomocí:
   (a) kritického oboru;
   (b) intervalu spolehlivosti;
   (c) p-hodnoty.

Požadovaná forma výstupu příkladu:

- H_0;
- H_1;
- tabulka výsledků:

Název statistiky	$\widehat{\sigma}$	statistika	krit.obor	$\widehat{\sigma}_D$	$\widehat{\sigma}_H$	p-hodnota

- komentář k výsledkům uvedeným v tabulce + zdůvodněné rozhodnutí o nulové hypotéze.

Poznámka: Hodnoty ve výsledné tabulce i textu zaokrouhlete na čtyři desetinná místa.

Příklad 4. Simultánní oblasti spolehlivosti + elipsa spolehlivosti pro střední hodnotu a směrodatnou odchylku (dokončení ze cvičení):

1. Nakreslete simultánní množinu spolehlivosti pro $\theta = (\mu, \sigma)^T$ použitím asymptotického intervalu spolehlivosti pro μ a exaktního intervalu spolehlivosti pro σ.

2. Nakreslete simultánní množinu spolehlivosti pro $\theta = (\mu, \sigma)^T$ použitím asymptotických intervalů spolehlivosti pro μ a pro σ.

3. Do obrázku dokreslete $100(1-\alpha)\,\%$ elipsu spolehlivosti pro $\theta = (\mu, \sigma)^T$ použitím asymptotických intervalů spolehlivosti pro μ a pro σ.

Použijte (1) $n = 10$, (2) $n = 100$, (3) $n = 10000$. V (1), (2) a (3) zvolte $\mu = 0$ a resp. $\sigma^2 = 4$. Koeficient spolehlivosti simultánní množiny zvolte zvolte $1 - \alpha = 0.95$.

Požadovaná forma výstupu příkladu:

- trojice grafů obsahující vždy dvě simultánní oblasti spolehlivosti pro parametry μ a σ a jednu elipsu spolehlivosti;

- odpovědi na následující tři otázky:

  1. Proč je šířka obou oblastí spolehlivosti vzhledem k proměnné na ose x v každém grafu stejná?
  2. Oblasti spolehlivosti se vzhledem k ose y s rostoucím n čím dál více překrývají. Čím je to způsobeno? Která oblast spolehlivosti se přibližuje ke které? Svou odpověď zdůvodněte.
  3. Elipsa spolehlivosti se s rostoucím n přibližuje k simultánní množině spolehlivosti popsané v bodu (2) a to jak z vertikální, tak z horizontální strany. Jak si tento jev vysvětlujete?

Příklad 5. Minimální rozsah náhodného výběru: Předpokládejme, že $X \sim N(\mu, \sigma^2)$, kde μ neznáme. Nechť $\theta = \sigma^2$. Testujeme všechny tři typy hypotéz:

a) $H_{01} : \sigma^2 = \sigma_0^2$ oproti $H_{11} : \sigma^2 \neq \sigma_0^2$ (oboustranná);

b) $H_{02} : \sigma^2 \leq \sigma_0^2$ oproti $H_{12:}$ $\sigma^2 > \sigma_0^2$ (pravostranná);

c) $H_{03} : \sigma^2 \geq \sigma_0^2$ oproti $H_{13} : \sigma^2 < \sigma_0^2$ (levostranná);

kde $\sigma_0^2 = 2$.

Vypočítejte minimální rozsah náhodného výběru pro testy hypotéz (a)–(c) při $\alpha = 0.05$ a $1 - \beta = 0.8$, pro $\sigma^2 \in \{1.01, 1.03, 1.05, \ldots 2.99, 3.01\}$ (ad (a)), $\sigma^2 \in \{2.01, 2.03, 2.05, \ldots 2.99, 3.01\}$ (ad (b)), $\sigma^2 \in \{1.01, 1.03, 1.05, \ldots 1.97, 1.99\}$ (ad (c)). Závislost minimálního rozsahu náhodného výběru na hodnotě σ^2 zakreslete do grafu pomocí křivky (na osu x vyneste parametr σ^2, na osu y minimální rozsah náhodného výběru.). V grafech barevně odlište minimální rozsah náhodného výběru pro:

(i) $\sigma^2 = 1.4$;

(ii) $\sigma^2 = 1.85$;

(iii) $\sigma^2 = 2.2$;

(iv) $\sigma^2 = 2.8$;

je-li to možné.

Poznámka: Čím více se budeme s hodnotou σ^2 blížit k hodnotě σ_0^2, tím větší minimální rozsah souboru budeme potřebovat. V souladu s touto informací a s vhodným vykreslením grafu rozumně stanovte maximální rozsah souboru (např. $N = 3000$, příp. $N = 5000$). K našim potřebám nám stačí vědět, že pro σ^2 blízká hodnotě σ_0^2 potřebujeme více než 3000 pozorování.

Požadovaná forma výstupu příkladu:

- funkce `min.rozsah()`, která pro stanovenou spolehlivost $1 - \alpha$ a sílu β^* vypočítá pro libovolnou alternativu minimální rozsah náhodného výběru;
- tři grafy závislosti N na σ^2;
- tabulka výsledků:

Alt. hypotéza	$\sigma^2 = 1.4$	$\sigma^2 = 1.85$	$\sigma^2 = 2.2$	$\sigma^2 = 2.8$
oboustranná				
pravostranná				
levostranná				

- komentář ke grafům a k výsledkům uvedeným v tabulce.

Příklad 6. Rozdělení testovací statistiky, síla a silofunkce pro test o rozdílu středních hodnot $\mu_1 - \mu_2$

1. Nechť náhodný výběr X pochází z normálního rozdělení, $X \sim N(\mu_1, \sigma_1^2)$, kde $\mu_1 = 4$ a $\sigma_1^2 = 2.5^2$, a nechť náhodný výběr Y pochází z normálního rozdělení, $Y \sim N(\mu_2, \sigma_2^2)$, kde $\mu_2 = 2$ a $\sigma_2^2 = 2.5^2$. Rozsahy náhodných výběrů $n_1 = n_2 = 20$. Pomocí simulační studie v porovnejte rozdělení testovací statistiky pro test nulové hypotézy H_0: $\mu_1 - \mu_2 = \mu_0$ oproti alternativní hypotéza $H_{11} : \mu_1 - \mu_2 \neq \mu_0$, kde $\mu_0 = 0$.
   a. Nasimulujte $M = 10\,000$ pseudonáhodných výběrů X a Y takových, že $X \sim N(\mu_1, 2.5^2)$ a $Y \sim N(\mu_2, 2.5^2)$. Pro každé $m = 1, \ldots, 10\,000$ vypočítejte realizaci testovací statistiky $t_{W,\lambda}^{(m)}$ pro nulovou hypotézu H_0: $\mu_1 - \mu_2 = 0$. Vykreslete histogram testovacích statistik $t_{W,\lambda}^{(m)}$ a superponujte jej jednak křivkou hustoty necentrálního t-rozdělení a jednak křivkou hustoty centrálního t-rozdělení.
   b. Vypočítejte empirickou sílu testu za platnosti alternativní hypotézy $H_{11} : \mu_1 - \mu_2 = 2$.
   c. Vytvořte animaci zobrazující odchýlení rozdělení testovací statistiky od centrálního t-rozdělení. Zvolte $\mu_1 = 4$ a $\mu_2 \in \{1, 1.5, \ldots, 6, 6.5, 7\}$.
      Použijte
      i. klasický dvouvýběrový t-test;
      ii. Welchův dvouvýběrový t-test.
2. Nechť nyní X pochází ze směsi dvou normálních rozdělení, t.j. $X \sim [pN(4, 2.5^2) + (1-p)N(4, 4.5^2)]$, kde $p = 0.9$ a nechť $Y \sim N(2, 2.5^2)$. Proveďte simulační studii popsanou v bodě (1.) pro tento náhodný výběr.

Požadovaná forma výstupu příkladu:

- funkce cn.rozdeleni(), která pro zadané hodnoty μ_1, μ_2, σ_{11}^2, σ_{12}^2, σ_{21}^2, σ_{22}^2, n_1, n_2, M, α a p vykreslí dvojici grafů histogramů (jeden pro klasický t-test a druhý pro Welchův t-test);
- animace (ad 1.) zachycující posun necentrálního rozdělní k centrálnímu rozdělení $\mu_1 - \mu_2$ pro pevně zvolené $\mu_1 = 4$ a měnící se $\mu_2 \in \{1, 1.5, \ldots, 6.5, 7\}$. V animaci budou vedle sebe dva grafy histogramů (jeden pro klasický t-test a druhý pro Welchův t-test) superponované křivkou centrálního a necentrálního t-rozdělení; pod popiskem osy x bude uvedena hodnota empirické síly příslušného testu (t-test resp. Welchův test);
- animace (ad 2.) zachycující posun necentrálního rozdělní k centrálnímu rozdělení $\mu_1 - \mu_2$ pro pevně zvolené $\mu_1 = 4$ a měnící se $\mu_2 \in \{1, 1.5, \ldots, 6.5, 7\}$. V animaci budou vedle sebe dva grafy histogramů (jeden pro klasický t-test a druhý pro Welchův t-test) superponované křivkou centrálního a necentrálního t-rozdělení; pod popiskem osy x bude uvedena hodnota empirické síly příslušného testu (t-test resp. Welchův test);
- komentář ke grafům + popis okometrického srovnání dvojic histogramů;
- odpovědi na následující otázky
  - Co mají společného tvary histogramů klasického a Welchova t-testu a proč?
  - Jak se liší síla klasického a Welchova t-testu a proč?
  - Jak se mění síla testů se snižující se vzdáleností $\mu_1 - \mu_2$? Jak si tento jev vysvětlujete?

Příklad 7. Pravděpodobnost pokrytí Waldova 95% empirického DIS pro rozdíl středních hodnot $\mu_1-\mu_2$:

1. Nechť

   A. $X_j \sim N(\mu_j, \sigma^2)$, kde $j = 1, 2$, $\mu_1 = 20$, $\mu_2 = 25$ a $\sigma^2 = 100$;
   B. $X_j \sim N(\mu_j, \sigma_j^2)$, kde $j = 1, 2$, $\mu_1 = 20$, $\mu_2 = 25$, $\sigma_1^2 = 100$ a $\sigma_2^2 = 150$;
   C. $X_j \sim pN(\mu_j, \sigma^2) + (1-p)N(\mu_j, \sigma_a^2)$, kde $j = 1, 2$, $\mu_1 = 20$, $\mu_2 = 25$, $\sigma^2 = 100$ a $\sigma_a^2 = 400$;
   D. $X_j \sim pN(\mu_j, \sigma_j^2) + (1-p)N(\mu_j, \sigma_{ja}^2)$, kde $j = 1, 2$, $\mu_1 = 20$, $\mu_2 = 25$, $\sigma_1^2 = 100$, $\sigma_2^2 = 150$, $\sigma_{1a}^2 = 400$ a $\sigma_{2a}^2 = 450$.

   Pomocí simulační studie ($M = 5\,000$) vypočítejte pravděpodobnost pokrytí 95% oboustranného Waldova empirického intervalu spolehlivosti pro $\mu_1-\mu_2$ jako podíl $\sum_{i=1}^{M} I(|t_{W,m}| < t_{df}(1-\alpha/2))/M$, kde $t_{W,m}$, $m = 1, \ldots, 5\,000$ jsou

   a. testovací statistiky klasického dvouvýběrového t-testu;
   b. testovací statistiky Welchova dvouvýběrového t-testu.

   Rozsahy náhodných výběrů zvolte

   i. $n_1 = n_2 = 5$;
   ii. $n_1 = n_2 = 50$;
   iii. $n_1 = n_2 = 100$.

2. Celý postup uvedený v bodě 1. zopakujte také pro $\mu_1 = \mu_2 = 20$. Ostatní hodnoty ponechte stejné jako hodnoty zadané v bodě 1.

Požadovaná forma výstupu příkladu:

- funkce `pokryti()` která pro zadané hodnoty μ_1, μ_2, σ_1, σ_2, σ_{1a}, σ_{2a}, M, n_1, n_2, α a p vypočítá empirickou pravděpodobnost pokrytí DIS pro rozdíl středních hodnot $\mu_1 - \mu_2$;
- dvě tabulky:
  - tabulka pravděpodobností pokrytí pro náhodné výběry A.–D., testovací statistiky a.–b. a rozsahy náhodných výběrů i.–iii. pro $\mu_1 = 20$ a $\mu_2 = 25$ (ad 1.);
  - tabulka pravděpodobností pokrytí pro náhodné výběry A.–D., testovací statistiky a.–b. a rozsahy náhodných výběrů i.–iii. pro $\mu_1 = \mu_2 = 20$ (ad 2.);

	n=5	n=50	n=100
A. klasicky t-test			
A. Welch t-test			
B. klasicky t-test			
B. Welch t-test			
C. klasicky t-test			
C. Welch t-test			
D. klasicky t-test			
D. Welch t-test			

- komentář k výsledkům uvedeným v tabulkách + porovnání výsledků získaných z obou tabulek;
- odpověď na otázky:
  - S jakou pravděpodobností pokrývají Waldovy empirické DIS hodnotu 0?
  - Jak a proč se hodnota pravděpodobnosti pokrytí mění (ad 1.)?

Poznámka: Hodnoty ve výsledné tabulce i textu zaokrouhlete na čtyři desetinná místa.

Příklad 8. Nechť početnosti úmrtí X jako následek kopnutí koněm v Pruských armádních jednotkách (Bortkiewicz, 1898) mají Poissonovo rozdělení s parametrem λ, tj. $X \sim Poiss(\lambda)$. Pravděpodobnost, že někdo bude smrtelně zraněný v daném dni, je extrémně malá. Mějme 10 vojenských jednotek za 20-letou periodu (rozsah $M = 10 \times 20 = 200$), kde, při početnostech úmrtí $n = 1, 2, 3, 4, 5+$ v dané jednotce a v daném roce, zaznamenáváme také početnosti vojenských jednotek m_n při daném n, kde $M = \sum m_n = 200$ (viz tabulka).

n	0	1	2	3	4	5+
m_n	109	65	22	3	1	0

1. Vypočítejte
   a. Waldův 95% empirický DIS pro λ;
   b. skóre 95% empirický DIS pro λ;
   c. věrohodnostní 95% empirický DIS pro λ.
2. Na hladině významnosti $\alpha = 0.05$ otestujte nulovou hypotézu $H_0 : \lambda = 0.6$ oproti $H_1 : \lambda \neq 0.6$. Testování proveďte Waldovým, skóre i věrohodnostním přístupem, a to pomocí
   i. kritického oboru;
   ii. intervalu spolehlivosti;
   iii. p-hodnoty.

Požadovaná forma výstupu příkladu:

- $H_0 : \lambda = \lambda_0$, kde $\lambda_0 = 0.61$;
- $H_1 : \lambda \neq \lambda_0$, kde $\lambda_0 = 0.61$;
- tabulka výsledků:

	$\widehat{\lambda}$,	test.stat.	W	$\widehat{\lambda}_D$	$\widehat{\lambda}_H$	p-hodnota
Z_W						
U_S						
U_{LR}						

- komentář k výsledkům uvedeným v tabulce + zdůvodněné rozhodnutí o nulové hypotéze;
- Zamyslete se, jakým způsobem by se dalo zjistit, který typ testu (Waldův skóre, věrohodnostní) je nejvhodnější k otestování hypotézy o parametru λ Poissonova rozdělení. Z jakých hledisek můžeme testy vzájemně porovnávat? Vyjmenujte alespoň dvě hlediska, která vás napadají.

Poznámka: Hodnoty ve výsledné tabulce i textu zaokrouhlete na čtyři desetinná místa.