Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Statistická inference I

Zadání domácího úkolu
podzimní semestr 2021

Stanislav Katina
Veronika Horská

katina@math.muni.cz

4. prosince 2021

Příklad 1.1. Směs dvourozměrných normálních rozdělení
Mějme k dispozici datový soubor faithful z knihovny datasets obsahující údaje o době čekání na erupci (waiting) a o době trvání erupce (eruption) gejzírů v Yellowstonském národním parku, Wyoming, USA.

1. Nakreslete histogram pro dobu čekání na erupci a superponujte jej křivkou jádrového odhadu hustoty získaného z dat. Analogický histogram vykreslete pro dobu trvání erupce.
2. Naprogramujte funkci hustoty smíšeného dvourozměrného normálního rozdělení mixdnorm2().
3. Pomocí funkce densityMclust(), jejíž syntax si samostatně nastudujte, odhadněte parametry smíšeného dvourozměrného normálního rozdělení. Získané odhady zaokrouhlete na čtyři desetinná místa a vložte do přehledné tabulky. Všechny získané hodnoty řádně interpretujte.
4. Nakreslete tečkový diagram doby čekání na erupci a doby trvání erupce a superponujte jej konturami:
   a. hustoty smíšeného dvourozměrného normálního rozdělení;
   b. jádrového odhadu hustoty získaného pomocí funkce kde2d().
5. Pomocí funkce persp() vykreslete 3D-diagram:
   (a) hustoty smíšeného dvourozměrného normálního rozdělení;
   (b) jádrového odhadu hustoty.

   Vytvořte animaci zobrazující náhled na oba 3D-diagramy.

Požadovaná forma výstupu příkladu:

- Dvojice histogramů superponovaných jádrovými odhady hustoty.
- Vlastnoručně naprogramovaná funkce mixdnorm2(). Vstupem funkce budou hodnoty x, y, p, μ_{11}, μ_{12}, σ_{11}, σ_{12}, μ_{21}, μ_{22}, σ_{21}, σ_{22} a výstupem bude hodnota hustoty smíšeného dvourozměrného normálního rozdělení v bodě (x, y).
- Tabulka odhadů parametrů smíšeného dvourozměrného normálního rozdělení zaokrouhlených na čtyři desetinná místa. Explicitní interpretace odhadu každého parametru.

$\widehat{p}$	$\widehat{\mu}_{11}$	$\widehat{\mu}_{12}$	$\widehat{\sigma}_{11}$	$\widehat{\sigma}_{12}$	$\widehat{\rho}_1$	$\widehat{\mu}_{21}$	$\widehat{\mu}_{22}$	$\widehat{\sigma}_{21}$	$\widehat{\sigma}_{22}$	$\widehat{\rho}_2$

- Dva tečkové diagramy zobrazující vztah mezi dobou čekání na erupci a dobou trvání erupce. První z diagramů bude superponován křivkou hustoty smíšeného dvourozměrného normálního rozdělení. Parametry smíšeného rozdělení odhadněte pomocí funkce densityMclust() z balíčku mclust. K získání hodnot hustoty použijte funkci mixdnorm2(). Druhý z diagramů bude superponován jádrovým odhadem hustoty. Počet bodů (x, y), v nichž budete hustotu počítat, zvolte 2500 ($n_x = n_y = 50$).
- Animace zobrazující dva 3D-diagramy otáčející se okolo své svislé osy. První z diagramů zobrazuje tvar hustoty smíšeného dvourozměrného normálního rozdělení, druhý z diagramů zobrazuje jádrový odhad hustoty. Hustotu rozsekejte na $k = 12$ intervalů, kde hodnoty v těchto intervalech budou odpovídat barvám sequential_hcl(12) z balíčku colorspace. Oba diagramy jsou animovány zároveň, otáčí se souběžně a poskytují náhled ze stejné strany. Animaci vytvoříte vhodným nastavením argumentů phi a theta v příkazu persp(). Argument phi nastavte pevně phi = 30, argument theta bude v rámci animace proměnný a bude nabývat hodnot od -180 do 180 po kroku 5.
- Komentář popisující postup řešení příkladu, popis funkce densityMclust() a argumentů, které jste do funkce zadali, detailní popis všech grafů a komentář k dění v obou animacích.

Příklad 1.2. Multinomický a součinový multinomický model
Mějme datový soubor 25-one-sample-probability-dermatoglyphs.txt obsahující údaje o dermatoglyfickém vzoru na deseti prstech pravé a levé ruky (*vír* (*whorl*), *smyčka* (*loop*), *obloučeк* (*arch*)) u 470 jedinců (235 mužů a 235 žen) bagathské populace z Araku Valley (Reddy, 1975; viz tabulka 1).

Tabulka 1: Dermatoglyfický vzor na deseti prstech levé a pravé ruky u jedinců bagathské populace

sex	vír	smyčka	obloučeк
m	1053	1246	51
f	880	1349	121

1. Předpokládejme, že náhodný vektor $\mathbf{X} = (X_1, \ldots, X_6)^T$ popisující dermatoglyfický vzor (s variantami vír (V), smyčka (S) a obloučeк (O)) u dospělých jedinců (muži (m) a ženy (f)), kde X_1 značí V-m, X_2 značí S-m, …, X_6 značí O-f, pochází z multinomického rozdělení, tj. $\mathbf{X} \sim \text{Mult}_6(N, \mathbf{p})$, kde $N = 4700$.

   (a) Stanovte odhad parametru $\mathbf{p}$ a vizualizujte jej pomocí dvourozměrného tečkového diagramu (např. pomocí funkce scatterplot3d() z knihovny scatterplot3d).

2. Zaměřte se na rozložení četností variant dermatoglyfického vzoru podmíněného pohlavím. Předpokládejme, že matice náhodných vektorů $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^T$, kde $\mathbf{X}_1 = (X_{11}, \ldots, X_{13})^T$, X_{11} značí V|m, …, X_{13} značí O|m a $\mathbf{X}_2 = (X_{21}, \ldots, X_{23})^T$, X_{21} značí V|f, …, X_{23} značí O|f, popisující typ dermatoglyfického vzoru podmíněný pohlavím, pochází ze součinového multinomického rozdělení, tj. $\mathbf{X} \sim \text{ProdMult}_3(\mathbf{N}, \mathbf{P})$, kde $\mathbf{N} = (N_1, N_2)$ a $\mathbf{P}$ je pravděpodobnostní matice.

   (a) Stanovte odhad matice $\mathbf{P}$ a vizualizujte jej pomocí dvourozměrného tečkového diagramu.

   (b) Vykreslete sloupcový diagram relativních četností jednotlivých variant dermatoglyfického vzoru podmíněných pohlavím.

   Všechny uvedené grafy a výsledky řádně okomentujte.

Požadovaná forma výstupu příkladu

- Odhad vektoru pravděpodobností $\mathbf{p}$, zapsaný jako kontingenční tabulka velikosti 2×3 a ukázka interpretací dvou libovolných odhadů.
- Dvourozměrný tečkový diagram pro odhad vektoru pravděpodobností $\mathbf{p}$. Na ose x budou vyneseny varianty darmatoglyfického vzoru, na ose y varianty pohlaví. V diagramu barevně odlišeny pravděpodobnosti pro muže a pravděpodobnosti pro ženy.
- Odhad pravděpodobnostní matice $\mathbf{P}$ a ukázka interpretací dvou libovolných odhadů.
- Dvourozměrný tečkový diagram pro odhad pravděpodobnostní matice $\mathbf{P}$. Na ose x budou vyneseny varianty dermatoglyfického vzoru, na ose y varianty pohlaví. V diagramu budou barevně odlišeny pravděpodobnosti pro muže a pravděpodobnosti pro ženy.
- Sloupcový diagram relativních četností. Diagram bude zobrazovat zastoupení jednotlivých variant dermatoglyfického vzoru. Diagram bude mít dva sloupce (levý pro muže, pravý pro ženy), přičemž každý sloupec bude mít výšku 1. Každá varianta bude mít vlastní barevný odstín a četnostní zastoupení uvedené v absolutní i procentuální škále. Součástí diagramu bude legenda variant dermatoglyfických vzorů. Diagram můžete vykreslit například pomocí funkce relBarplotTwo(), který najdete v RSkriptu SI-I-relBarplotTwo.R.
- Komentář popisující řešení příkladu, popis grafů a jejich propojení s vypočítaným odhadem parametru $\mathbf{p}$, resp. pravděpodobnostní matice $\mathbf{P}$.

Příklad 1.3. Maximálně věrohodné odhady v ZIP modelu
V rámci národní studie byly zaznamenány četnosti primárních a revizních operací kyčelního kloubu provedených na Slovensku v období od 1. ledna 2003 do 31. prosince 2012. Celkem byly získány údaje o 12 349 operacích, přičemž v 11 317 případech šlo o operace primární (operaci nepředcházela žádná operace kyčelního kloubu) a v 1 032 případech šlo o operaci revizní (operaci předcházela alespoň jedna operace kyčelního kloubu). Přesné údaje o počtu předcházejících operací u každého pacienta jsou uvedeny v tabulce 2.

Tabulka 2: Absolutní četnosti operací kyčelního kloubu provedených na Slovensku v letech 2003–2012

n	0	1	2	3	4	$\sum$
m_{obs}	11 317	907	106	18	1	12 349

Za předpokladu, že náhodná veličina X popisující početnosti výskytu primárních a revizních operací kyčelního kloubu se řídí ZIP modelem s parametry λ, p, tj. $X \sim ZIP(\lambda, p)$:

1. Odvoďte:
   a. tvar jádra věrohodnostní funkce $L((\lambda, p)^T|x)$;
   b. tvar jádra logaritmu věrohodnostní funkce $\ell((\lambda, p)^T|x)$;
   c. skóre funci pro parametr λ;
   d. skóre funkci pro parametr p;
   e. tvar Fisherovy informační matice.
2. Pomocí maximalizace logaritmu věrohodnostní funkce $\ell((\lambda, p)^T|x)$ ZIP modelu nalezněte maximálně věrohodný odhad parametrů λ a p. Maximalizaci proveďte:
   a. pomocí funkce `optim()`;
   b. pomocí vlastnoručně naprogramované dvourozměrné Newton-Raphsonovy metody;
   c. pomocí vlastnoručně naprogramované Broydenovy metody.
3. Vykreslete:
   a. křivku logaritmu věrohodnostní funkce ZIP modelu spolu s maximálně věrohodnými odhady parametrů λ a p odhadnutými pomocí (A) funkce `optim()`; (B) Newton-Raphsonovy metody; (C) Broydenovy metody;
   b. 3D-diagram logaritmu věrohodnostní funkce ZIP modelu;
   c. animaci zobrazující konvergenci Newton-Raphsonovy metody (resp. Broydenovy metody) k maximu logaritmu věrohodnostní funkce.

Požadovaná forma výstupu příkladu

- Vzorce a odvození:
  - Vzorec věrohodnostní funkce $L((\lambda, p)^T|x)$;
  - Odvození vzorce pro logaritmus věrohodnostní funkce $\ell((\lambda, p)^T|x)$;
  - Odvození vzorce skóre funkce pro parametr λ;
  - Odvození vzorce skóre funkce pro parametr p;
  - Odvození pozorované Fisherovy informační matice.
- Dvě vlastnoručně naprogramované funkce: `NRzip(x01, x02, t)` a `BMzip(x10, x11, x20, x21, t)`. Výstupem funkcí budou MLE odhady parametrů λ a p, hodnota logaritmické věrohodnostní funkce $\ell((\lambda, p)|\mathbf{x})$ v MLE odhadech $\widehat{\lambda}$ a $\widehat{p}$, hodnota počítadla k reprezentující počet kroků iterace a matice `body` obsahující hodnoty x_{01}, x_{02}, ..., resp. x_{10}, x_{11}, ... pro každý k-tý krok iterování. Tyto body použijeme níže při vykreslení animací.
- Tabulka odhadů parametrů λ a p zaokrouhlených na šest desetinných míst.

	$\widehat{\lambda}$	$\widehat{p}$
exaktni vypocet		
funkce optim()		
Newton-Raphsonova metoda		
Broydenova metoda		

- Tři grafy obsahující dvourozměrnou křivku logaritmu věrohodnostní funkce ZIP modelu spolu s maximálně věrohodnými odhady parametrů λ a p odhadnutými pomocí (A) funkce optim(); (B) Newton-Raphsonovy metody; (C) Broydenovy metody. K vykreslení křivky použijte příkaz image() v kombinaci se škálou $k = 15$ barev z palety heat_hcl() z knihovny colorspace. Do grafu dále dokreslete kontury příkazem contour(). Ohlídejte, aby vykreslené kontury ohraničovaly barevně oddělené vrstvy.
- 3D-diagram zobrazující 3D pohled na logaritmus věrohodnostní funkce normálního modelu. K vykreslení křivky použijte příkaz persp() v kombinaci se škálou $k = 15$ barev z palety heat_hcl() z knihovny colorspace.
- Animace zobrazující konvergenci Newton-Raphsonovy metody k maximu logaritmu věrohodnostní funkce. Součástí animace bude popisek umístěný pod osou x obsahující měnící se hodnoty parametrů λ a p a měnící se hodnotu k reprezentující počítadlo iteračních kroků funkce.
- Animace zobrazující konvergenci Broydenovy metody k maximu logaritmu věrohodnostní funkce. Součástí animace bude popisek umístěný pod osou x obsahující měnící se hodnoty parametrů λ a p a měnící se hodnotu k reprezentující počítadlo iteračních kroků funkce.
- Podrobné komentáře porovnávající výsledky v tabulce a popisující všechny čtyři grafy i obě animace.

Příklad 1.4. Kvadratická aproximace v Poissonově modelu
Načtěte datový soubor 2021-SI-I-DU-data-poiss.Rdata obsahující tři pseudonáhodné výběry hodnot pocházejících z rozdělení Poiss(λ), kde $\lambda = 6$. Rozsahy náhodných výběrů jsou (a) N = 10; (b) N = 50; (c) N = 100).

1. Naprogramujte funkci KvadrAproxPoiss().
2. Pomocí funkce KvadrAproxPoiss() nakreslete pro každý pseudonáhodný výběr (a)–(c) křivku škálovaného logaritmu funkce věrohodnosti Poissonova rozdělení superponovanou křivkou kvadratické aproximace.
3. Pomocí funkce KvadrAproxPoiss() nakreslete pro každý pseudonáhodný výběr (a)–(c) graf zobrazující asymptotický lokálně lineární vztah mezi funkcemi $\mathcal{I}^{1/2}(\widehat{\lambda})(\lambda - \widehat{\lambda})$ a $-\mathcal{I}^{-1/2}(\widehat{\lambda})S(\lambda)$.
4. Rozdíly v grafech pro situace (a), (b) a (c) řádně okomentujte.

Požadovaná forma výstupu příkladu

- Vlastnoručně naprogramovaná funkce KvadrAproxPoiss(). Povinnými vstupními argumenty funkce budou lambda, x, a plot. Výstupem funkce bude graf zobrazující kvadratickou aproximaci funkce $\ln \mathcal{L}(\lambda)$ pro náhodný výběr $X_1, \dots X_N$, pokud argument plot == aproximace, nebo graf zobrazující asymptotický lokálně lineární vztah mezi funkcí $-\mathcal{I}^{-1/2}(\widehat{\lambda})S(\lambda)$ a funkcí $\mathcal{I}^{1/2}(\widehat{\lambda})(\lambda - \widehat{\lambda})$ pro náhodný výběr $X_1, \dots X_N$, pokud argument plot == linearita.
- Tři grafy zobrazující škálovaný logaritmus funkce věrohodnosti Poissonova rozdělení pro pseudonáhodné výběry (a), (b) a (c). Na x-ové ose bude λ a na y-ové ose $\ln \mathcal{L}(\lambda) = \ell(\lambda|\mathbf{x}) - \ell(\widehat{\lambda}|\mathbf{x}))$. V grafu bude dále zakreslena modrá přerušovaná křivka reprezentující kvadratickou aproximace vypočítanou pomocí Taylorova rozvoje $\ln \mathcal{L}(\lambda) = \ln\left(\frac{L(\lambda|\mathbf{x})}{L(\widehat{\lambda}|\mathbf{x})}\right) \approx -\frac{1}{2}\mathcal{I}(\widehat{\lambda})(\lambda - \widehat{\lambda})^2$. Rozsah osy x bude pro všechny tři grafy 1 až 10, rozsah osy y bude -30 až 10. Součástí každého grafu bude legenda popisující obě křivky vykreslené v grafu a popisek pod osou x s údaji o rozsahu náhodného výběru N.
- Tři grafy zobrazující asymptotický lokálně lineární vztah mezi funkcí $\mathcal{I}^{1/2}(\widehat{\lambda})(\lambda - \widehat{\lambda})$ a funkcí $-\mathcal{I}^{-1/2}(\widehat{\lambda})S(\lambda)$ pro pseudonáhodné výběry (a), (b) a (c). Funkce $\mathcal{I}^{1/2}(\widehat{\lambda})(\lambda - \widehat{\lambda})$ bude zobrazená na ose x a funkce $-\mathcal{I}^{-1/2}(\widehat{\lambda})S(\lambda)$ na ose y. V každém grafu bude dále vykreslená modrá přerušovaná přímka $y = x$. Rozsah osy x i osy y bude pro všechny tři grafy v rozmezí -4 až 4. Součástí každého grafu bude legenda popisující křivku i přímku a popisek pod osou x s údaji o rozsahu náhodného výběru N.
- Podrobné komentáře porovnávající vzájemně trojici grafů zobrazujících kvadratickou aproximaci pro situace (a), (b) a (c) a dále podrobné komentáře porovnávající vzájemně trojici grafů zobrazujících asymptotický lokálně lineární vztah pro situace (a), (b) a (c).