Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Statistická inference I

Zadání domácího úkolu
podzimní semestr 2022

Stanislav Katina
Veronika Horská

katina@math.muni.cz

3. prosince 2022

Příklad 1.1. Modelování dat pomocí vhodného rozdělení

Dataset: Proměnné ovlivňující stabilitu politických režimů v Sub-saharské Africe
V rámci studie (Bratton, De Walle, 1997) byly zkoumány proměnné ovlivňující stabilitu politických režimů ve státech Sub-saharské Afriky. Jednou ze zkoumaných proměnných byl počet úspěšných vojenských převratů v těchto zemích provedených v období od vyhlášení nezávislosti dané země do roku 1989. Údaje o počtu úspěšně provedených vojenských převratů ve vybraných 47 zemích jsou vloženy v datovém souboru africa, který je implementován v knihovně faraway.

Načtěte datový soubor africa obsahující údaje o počtu úspěšných vojenských převratů provedených ve vybraných 47 zemích Sub-saharské Afriky v období od vyhlášení nezávislosti dané země do roku 1989 (miltcoup).

1. Z množiny rozdělení, která jsme probírali na cvičeních Statistické inference I, nalezněte takové, které je nejvíce vhodné k modelování náhodné veličiny X popisující počet úspěšných vojenských převratů v zemích Sub-saharské Afriky. Pro toto rozdělení odhadněte jeho parametr(y).
2. Vypočítejte očekávané četnosti úspěšných vojenských převratů za předpokladu zvoleného rozdělení a porovnejte je s pozorovanými četnostmi. Na základě pozorovaných a očekávaných četností vypočítejte odhad střední hodnoty a rozptylu náhodné veličiny X. Zhodnoťte kvalitu modelu obecně i z hlediska přeceněného a podceněného rozptylu (overdisperze a underdisperze).

Požadovaná forma výstupu příkladu:

- Tvar rozdělení i s odhady parametru(ů) tohoto rozdělení.
- Tabulka pozorovaných četností a očekávaných četností za předpokladu vybraného rozdělení.

Tabulka 1: Pozorované a očekávané četnosti úspěšných vojenských převratů

	0	1	2	3	4	5	6
m_{obs}							
m_{exp}							

- Graf porovnávající pozorované a očekávané absolutní četnosti.
- Tabulka porovnávající odhad střední hodnoty a rozptylu počtu úspěšných vojenských převratů vypočítaný (a) na základě pozorovaný dat, (b) na základě očekávaných dat za předpokladu vybraného rozdělení + zhodnocení kvality modelu obecně i stran přeceněného a podceněného rozptylu.

Tabulka 2: Odhady střední hodnoty a rozptylu náhodné veličiny X

	$\widehat{E[X]}$	$\widehat{\mathrm{Var}[X]}$
pozorovaná data		
očekávaná data		

Příklad 1.2. Multinomický a součinový multinomický model
Mějme datový soubor 24-multinom-blood-groups.txt obsahující údaje o frekvenci krevních skupin AB0 systému (skupina 0, A, B, AB) u 400 obyvatelů z města Košice a 500 obyvatelů z města Praha (Vondrušková, 1983; viz tabulka 3).

Tabulka 3: Frekvence výskytu krevních skupin mezi obyvateli města Košice a města Praha

	0	A	B	AB
Košice	138	147	84	31
Praha	209	184	81	26

1. Předpokládejme, že náhodný vektor $\mathbf{X} = (X_1, \ldots, X_8)^T$ popisující krevní skupinu (s variantami 0, A, B, AB) u dvou populací (Košice (K), Praha (P)), kde X_1 značí K–0, X_2 značí K–A, …, X_8 značí P–AB, pochází z multinomického rozdělení, tj. $\mathbf{X} \sim \text{Mult}_8(N, \mathbf{p})$, kde $N = 900$.

   (a) Odhadněte vektor parametrů $\mathbf{p}$ a vizualizujte jej pomocí dvourozměrného tečkového diagramu.

2. Zaměřte se na rozložení četností variant krevní skupiny podmíněného populací. Předpokládejme, že matice náhodných vektorů $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^T$, kde $\mathbf{X}_1 = (X_{11}, \ldots, X_{14})^T$, X_{11} značí 0|K, …, X_{14} značí AB|K a $\mathbf{X}_2 = (X_{21}, \ldots, X_{24})^T$, X_{21} značí 0|P, …, X_{24} značí AB|P, popisujících krevní skupinu podmíněnou populací, pochází ze součinového multinomického rozdělení, tj. $\mathbf{X} \sim \text{ProdMult}_4(\mathbf{N}, \mathbf{P})$, kde $\mathbf{N} = (N_1, N_2)$ a $\mathbf{P}$ je matice parametrů (pravděpodobností).

   (a) Odhadněte matici parametrů $\mathbf{P}$ a vizualizujte ji pomocí dvourozměrného tečkového diagramu.

   (b) Vykreslete sloupcový diagram absolutních a relativních četností jednotlivých variant typu krevní skupiny podmíněných populací.

Všechny uvedené grafy a výsledky řádně okomentujte.

Požadovaná forma výstupu příkladu

- Odhad vektoru parametrů $\mathbf{p}$, zapsaný jako kontingenční tabulka velikosti 2×4 a ukázka interpretací dvou libovolných odhadů.
- Dvourozměrný tečkový diagram pro odhad vektoru parametrů $\mathbf{p}$. Na ose x budou vyneseny varianty krevní skupiny, na ose y varianty populace. V diagramu budou barevně odlišeny pravděpodobnosti pro obyvatele Košic a pravděpodobnosti pro obyvatele Prahy.
- Odhad matice parametrů $\mathbf{P}$ a ukázka interpretací dvou libovolných odhadů.
- Dvourozměrný tečkový diagram pro odhad matice parametrů $\mathbf{P}$. Na ose x budou vyneseny varianty krevní skupiny, na ose y varianty populace. V diagramu budou barevně odlišeny pravděpodobnosti pro obyvatele Košic a pravděpodobnosti pro obyvatele Prahy.
- Sloupcový diagram relativních četností. Diagram bude zobrazovat zastoupení jednotlivých variant krevní skupiny podmíněných populací. Diagram bude mít dva sloupce (levý pro obyvatele Košic, pravý pro obyvatele Prahy), 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 krevní skupiny. Diagram můžete vykreslit například pomocí funkce relBarplotTwo() implementovaný v RSkriptu SI-I-relBarplotTwo.R.
- Komentář popisující řešení příkladu, popis grafů a jejich propojení s vypočítaným odhadem vektoru parametrů $\mathbf{p}$, resp. s odhadem matice parametrů $\mathbf{P}$.

Příklad 1.3. Odhadnutá věrohodnostní funkce, profilová věrohodnostní funkce a věrohodnostní funkce pro parametry μ a σ^2

Vygenerujte pseudonáhodná čísla $X \sim N(4, 3)$, $n = 800$.

1. Nakreslete grafy logaritmu odhadnuté funkce věrohodnosti a logaritmu profilové funkce věrohodnosti pro μ a σ^2, tj. (A) $\ell_e(\mu|\mathbf{x})$, (B) $\ell_e(\sigma^2|\mathbf{x})$, (C) $\ell_P(\mu|\mathbf{x})$ a (D) $\ell_P(\sigma^2|\mathbf{x})$. Vypočítejte a následně porovnejte maximálně věrohodné odhady parametrů μ a σ^2 získané maximalizací funkcí (A)–(D). Ve vykreslených grafech (A)–(D) zvýrazněte polohu maxim.

2. MLE odhady parametrů μ a σ^2 z bodu (1) najděte (a) pomocí funkce optimize(); (b) pomocí vlastnoručně naprogramované metody sečen. Získané odhady zaokrouhlete na šest desetinných míst, uspořádejte do přehledné tabulky a vzájemně porovnejte.

3. Zaměřte se na 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 $\ell(\boldsymbol{\theta}|\mathbf{x})$ použitím funkce image() a superponujte ho konturovým grafem. V grafu zvýrazněte polohu maxima $(\widehat{\mu}, \widehat{\sigma}^2)^T$ dopočítaného pomocí funkce optim(). Dále nakreslete 3D graf pomocí funcke persp(). Počet bodů $(x,\ y)$, v nichž budete počítat věrohodnostní funkci, zvolte 2500 ($n_x = n_y = 50$).

Požadovaná forma výstupu příkladu

- Čtyři grafy obsahující křivku odhadnuté funkce věrohodnosti (A) pro μ, (B) pro σ^2; křivku profilové funkce věrohodnosti (C) pro μ, (D) pro σ^2. Rozsah osy x zvolte $\langle 0; 8\rangle$ v grafech (A) a (C), resp. $\langle 2; 4\rangle$ v grafech (C) a (D). V každém grafu bude vykreslena poloha maxima parametru μ, resp. σ^2 spočítaného pomocí metody sečen. Hodnota příslušného maxima (zaokrouhlená na šest desetinných míst) bude také uvedena v popisku grafu pod popiskem osy x.

- Tabulka porovnávající odhady parametrů $\widehat{\mu}_e$, $\widehat{\mu}_P$, $\widehat{\sigma}^2_e$, $\widehat{\sigma}^2_P$ získaných (a) pomocí funkce optimize(); (b) pomocí metody sečen. Hodnoty v tabulce zaokrouhlete na šest desetinných míst.

Tabulka 4: Odhady parametrů μ a σ^2 odhadnuté a profilové věrohodnosti normálního rozdělení

	$\widehat{\mu}_e$	$\widehat{\mu}_P$	$\widehat{\sigma}^2_e$	$\widehat{\sigma}^2_P$
optimize()				
Metoda sečen				

- Dva grafy: V prvním grafu budou společně vykresleny křivky odhadnuté a profilové funkce věrohodnosti (A) a (C) pro parametr μ. V druhém grafu budou společně vykresleny křivky odhadnuté a profilové funkce věrohodnosti (B) a (D) pro parametr σ^2. Rozsah osy x zvolte $\langle 0; 8\rangle$ v grafech (A) a (C), resp. $\langle 2; 4\rangle$ v grafech (C) a (D).

- Graf $\ell(\boldsymbol{\theta}|\mathbf{x})$ vykreslený pomocí funkce image() superponovaný konturovým grafem. Na osu x vyneste hodnotu parametru μ, na osu y hodnotu parametru σ^2. Barevnou paletu pro funkci image() zvolte heat.colors(12). Součástí grafu bude popisek (umístěný od popiskem osy x) s odhady obou parametrů zaokrouhlených na šest desetinných míst.

- Graf $\ell(\boldsymbol{\theta}|\mathbf{x})$ vykreslený pomocí funkce persp(). Zobrazenou plochu rozsekejte na $k = 12$ intervalů, přičemž hodnoty v těchto intervalech budou odpovídat barvám heat.colors(12). Na osu x vyneste hodnotu parametru μ, na osu y hodnotu parametru σ^2.

- Komentáře popisující všech osm vykreslených grafů a porovnávající odhady parametrů.

Příklad 1.4. Maximálně věrohodné odhady; Negativně binomický model
V rámci studie počtu úrazů v továrnách byl zaznamenán počet úrazů u každého dělníka v jedné vybrané továrně během roku 1920. Celkový počet dělníků zahrnutých do studie $M = 647$. Údaje ze studie jsou uvedeny v následující tabulce.

n	0	1	2	3	4	≥ 5	$\sum$
$m_{observed}$	447	132	42	21	3	2	647

Za předpokladu, že náhodná veličina X popisující početnosti úrazů u každého dělníka, se řídí negativně binomickým modelem s parametry k, p, tj. $X \sim \text{NegBin}(k, p)$:

1. Uveďte
   a. tvar jádra věrohodnostní funkce $L((k,p)^T|x)$;
   b. tvar jádra logaritmu věrohodnostní funkce $\ell((k,p)^T|x)$;
   c. tvar skóre funce $S_1((k,p)^T)$ pro parametr k;
   d. tvar skóre funkce $S_2((k,p)^T)$ pro parametr p;
   e. tvar Fisherovy informační matice $\mathcal{I}((\widehat{k},\widehat{p})^T)$.
2. Pomocí maximalizace logaritmu věrohodnostní funkce $\ell((k,p)^T|x)$ negativně binomického modelu nalezněte maximálně věrohodný odhad parametrů k a p. Maximalizaci proveďte
   a. pomocí funkce optim();
   b. pomocí vlastnoručně naprogramované dvourozměrné Newton-Raphsonovy metody;
   Tip: Aby dvourozměrná Newton-Raphsonova metoda zkonvergovala, je třeba vhodně zvolit startovací body, např. $x_{01} = 0.9$, $x_{02} = 0.3$.
3. Vykreslete
   a. vrstevnicový diagram logaritmu věrohodnostní funkce negativně binomického modelu spolu s maximálně věrohodnými odhady parametrů k a p odhadnutými pomocí (A) vzorců; (B) funkce optim(); (C) Newton-Raphsonovy metody;
   b. 3D-diagram logaritmu věrohodnostní funkce negativně binomického modelu;
   c. animaci zobrazující konvergenci Newton-Raphsonovy metody k maximu logaritmu věrohodnostní funkce.

Požadovaná forma výstupu příkladu

- Vzorce a odvození:
  - Vzorec věrohodnostní funkce $L((k,p)^T|x)$;
  - Vzorec pro logaritmus věrohodnostní funkce $\ell((k,p)^T|x)$;
  - Vzorec skóre funkce $S_1((k,p)^T)$ pro parametr k;
  - Vzorec skóre funkce $S_2((k,p)^T)$ pro parametr p;
  - Pozorovaná Fisherova informační matice $\mathcal{I}((\widehat{k},\widehat{p})^T)$.
- Vlastnoručně naprogramovaná funkce: NRnegbin(): Newton-Raphsonova metoda.
- Tabulka odhadů parametrů k a p zaokrouhlených na šest desetinných míst.

Tabulka 5: Odhady parametrů k a p negativně binomického modelu

	k	p
exaktní výpočet		
funkce optim()		
Newton-Raphsonova metoda		

- Vrstevnicový diagram logaritmu věrohodnostní funkce negativně binomického modelu spolu s maximálně věrohodným odhadem parametrů k a p stanovenými pomocí (A) vzorců; (B) funkce optim(); (C) Newton-Raphsonovy metody. Jednotlivé odhady od sebe barevně odlište a popište v legendě. K vykreslení křivky použijte příkaz image() v kombinaci se škálou $l = 15$ barev z palety terrain.colors(). Do grafu dokreslete kontury a ohlídejte, aby kontury ohraničovaly barevně oddělené vrstvy.
- 3D-diagram zobrazující 3D pohled na logaritmus věrohodnostní funkce negativně binomického modelu. K vykreslení křivky použijte příkaz persp() v kombinaci se škálou $l = 15$ barev z palety terrain.colors().
- Animace zobrazující konvergenci Newton-Raphsonovy metody k maximu logaritmu věrohodnostní funkce. Součástí animace bude popisek umístěný pod osou x obsahující měnící se hodnoty parametrů k a p (zaokrouhlené na šest desetinných míst) a parametr i reprezentujícího počitadlo iteračních kroků.
- Podrobné komentáře porovnávající výsledky v obou tabulkách a popisující všechny tři grafy i obě animace.

  Poznámka: Funkce, které by se mohly hodit: gamma(), digamma().