Bayesovsko hijerarhijsko modeliranje

Drobac, Nina

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Nina Drobac

BAYESOVSKO HIJERARHIJSKO MODELIRANJE

Diplomski rad

Voditelj rada:
doc. dr. sc. Azra Tafro

Zagreb, veljača 2023.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Baki Mariji i djedu Jozi,
koji su prevarili starost da bi mene ispratili u samostalan život

Sadržaj

Uvod

Mogli bismo reći da je još u drevnim civilizacijama postojala preteča statistike, u obliku popisivanja poljoprivrednih uroda, stanovništva i materijalnih posjeda. U svojim prvim upotrebama, riječ statistika podrazumijevala je znanost o stanju i političkom uređenju države. Danas, statistika (njem. Statistik, prema lat. status: stanje) označava granu primijenjene matematike koja se bavi prikupljanjem, analizom i tumačenjem podataka te izvođenjem zaključaka o pojavama i procesima koje ti podatci predstavljaju.

Kod donošenja zaključaka na temelju statističke analize, razlikujemo dva glavna pristupa: frekvencionistički i bayesovski. U fokusu ovoga rada nalazi se bayesovski pristup, nazvan po Thomasu Bayesu (1702.-1761.), engleskom matematičaru i teologu koji je prvi iznio jedan od najvažnijih teorema uvjetne vjerojatnosti, danas poznatog kao Bayesov teorem.

Za razliku od frekvencionističkog pristupa, gdje parametar populacijske distribucije statističkog obilježja smatramo nepoznatom, ali fiksnom vrijednošću, u bayesovskom pristupu parametar modeliramo kao slučajnu varijablu i opisujemo vjerojatnosnom distribucijom. Na početku, nepoznatom parametru pridružujemo apriornu distribuciju koja odražava našu predodžbu prije provođenja istraživanja. Nakon što smo prikupili podatke, revidiramo apriornu distribuciju slijedeći Bayesov teorem te dolazimo do aposteriorne distribucije parametra, koja reflektira nova saznanja na temelju opažanja.

Bayesovski je pristup posebno zanimljiv u okviru hijerarhijskog modeliranja. Naime, u statistici često koristimo pretpostavku da su podaci nezavisni i shvaćamo ih kao realizacije slučajnog uzorka - niza nezavisnih i jednako distribuiranih slučajnih varijabli. Takva pretpostavka nije sasvim opravdana kada radimo s podacima koji su, zbog načina na koji su prikupljeni ili prirode problema kojeg opisuju, strukturirani u nekoliko grupa. Primjer takve situacije su podaci koji dolaze iz više različitih eksperimenata ili su zabilježeni na različitim lokacijama. Premda možemo zanemariti grupiranje i na temelju svih opažanja donositi zaključke ili možemo svaku od grupa promatrati posebno i neovisno od drugih, takvim modeliranjem ne uzimamo u obzir razlike, odnosno sličnosti promatranih grupa. Kako bismo matematičkim jezikom opisali odnos parametara koji određuju populacijske distribucije obilježja grupa, koristimo hijerarhijske modele. Kod hijerarhijskog modeliranja, uvodimo

zajedničku vjerojatnosnu distribuciju nad parametrima $\theta_1, \theta_2, \ldots, \theta_J$ koji opisuju J grupa podataka. U punom bayesovskom hijerarhijskom modelu, ta je zajednička vjerojatnosna distribucija parametrizirana hiperparametrima ϕ, za koje također promatramo apriornu i aposeriornu distribuciju i za koje donosimo statističke zaključke slijedeći bayesovske principe. To nam omogućuje da pri donošenju zaključaka za jednu grupu iskoristimo i podatke o preostalim grupama.

Za kraj uvodnog dijela, navodimo kako je rad strukturiran. U prvom poglavlju izneseni su osnovni pojmovi teorije vjerojatnosti i statistike koji se koriste u ostatku rada. U drugom je poglavlju dan kratki pregled frekvencionističkog pristupa i razrađen je bayesovski pristup s glavnim teorijskim rezultatima koji su ilustrirani na primjerima. Motivacija i glavne ideje iza hijerarhijskog modeliranja iznesene su u trećem poglavlju te su dodatno objašnjene na primjeru. Zadnje, četvrto poglavlje obrađuje bayesovsko hijerarhijsko modeliranje, s posebnim naglaskom na provedbi takve analize, koja uključuje i analitički dio i računalne simulacije. Završavamo primjerima bayesovskih hijerarhijskih modela, uz grafičke prikaze i rezultate dobivene pomoću programskog jezika R (sav je kod priložen u dodatku).

Poglavlje 1

Teorijske osnove

U ovom je poglavlju cilj navesti osnovne pojmove i rezultate teorije vjerojatnosti i statistike, koji će poslužiti kao gradivni blokovi u daljnjem radu. Pri tome podrazumijevamo poznavanje teorije mjere i integrala.
Definicije i rezultati ovoga odjeljka preuzeti su iz literature: [5], [6], [8], [9], te se navedene izvore može pogledati za više detalja.

1.1 Osnove teorije vjerojatnosti

Polazni nam je objekt neprazni skup Ω kojeg nazivamo **prostor elementarnih događaja**, a njegove elemente ω nazivamo **elementarnim događajima**. **Događaje** definiramo kao elemente σ-algebre $\mathcal{F}$ na prostoru elementarnih događaja. Ovim pojmovima modeliramo idealizirani slučajni pokus, gdje svakom ishodu izvođenja pokusa odgovara točno jedan elementarni događaj.
Slijedi nekoliko formalnih definicija i rezultata.

Definicija 1.1.1. *Neka je Ω neprazan skup i $\mathcal{F}$ σ-algebra.* ***Vjerojatnost*** *na izmjerivom prostoru $(\Omega, \mathcal{F})$ je funkcija $\mathbb{P} : \mathcal{F} \to [0,1]$ koja zadovoljava sljedeća svojstva:*

(A1) (nenegativnost) Za sve $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$;

(A2) (normiranost) $\mathbb{P}(\Omega) = 1$;

(A3) (σ-aditivnost) Za svaki niz $(A_j)_{(j \in N)}$ po parovima disjunktnih događaja $A_j \in \mathcal{F}$ ($A_i \cap A_j = \emptyset$; za $i \neq j$) vrijedi $\mathbb{P}\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathbb{P}\left(A_j\right)$.

Uređenu trojku $(\Omega, \mathcal{F}, \mathbb{P})$ nazivamo ***vjerojatnosni prostor****.*

Definicija 1.1.2. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor te $B \in \mathcal{F}$ događaj takav da je $\mathbb{P}(B) > 0$.* ***Uvjetna vjerojatnost*** *događaja A uz dano B definira se formulom*

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1.1}$$

Iz ove definicije odmah slijedi:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B). \tag{1.2}$$

Definicija 1.1.3. *Neka je $(H_i)_{i\in I}$ konačna ili prebrojiva familija događaja iz $\mathcal{F}$ takva da je $\mathbb{P}(H_i)>0$ za sve $i \in I$ i neka vrijedi $H_i \cap H_j = \emptyset$ za $i \neq j$ te $\bigcup_{i\in I} H_i = \Omega$. Takvu familiju $(H_i)_{i\in I}$ zovemo* ***potpun sustav događaja****.*

Propozicija 1.1.4. *(Formula potpune vjerojatnosti) Neka je $(H_i)_{i\in I}$ potpun sustav događaja. Tada za svaki $A \in \mathcal{F}$ vrijedi*

$$\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(H_i)\mathbb{P}(A \mid H_i).$$

Dokaz.

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}\left(A \cap \left(\bigcup_{i\in I} H_i\right)\right) = \mathbb{P}\left(\bigcup_{i\in I} (A \cap H_i)\right) \\ &= \sum_{i\in I} \mathbb{P}\left(A \cap H_i\right) = \sum_{i\in I} \mathbb{P}\left(H_i\right)\mathbb{P}\left(A \mid H_i\right).\end{aligned}$$

Prva jednakost slijedi iz činjenice da je $\bigcup_{i\in I} H_i = \Omega$, druga jednakosti slijedi iz skupovnih operacija, treća zbog σ-aditivnost vjerojatnosti, a četvrta iz relacije (1.2). □

Teorem 1.1.5. ***(Bayesova formula)***
Neka je $(H_i)_{i\in I}$ potpun sustav događaja na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$. Tada za svaki $A \in \mathcal{F}$ takav da je $\mathbb{P}(A)>0$ vrijedi

$$\mathbb{P}\left(H_j \mid A\right) = \frac{\mathbb{P}\left(H_j\right)\mathbb{P}\left(A \mid H_j\right)}{\sum_{i\in I} \mathbb{P}\left(H_i\right)\mathbb{P}\left(A \mid H_i\right)}.$$

Dokaz.

$$\begin{aligned}\mathbb{P}\left(H_j \mid A\right) &= \frac{\mathbb{P}\left(H_j \cap A\right)}{\mathbb{P}(A)} = \frac{\mathbb{P}\left(H_j\right)\mathbb{P}\left(A \mid H_j\right)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}\left(H_j\right)\mathbb{P}\left(A \mid H_j\right)}{\sum_{i\in I} \mathbb{P}\left(H_i\right)\mathbb{P}\left(A \mid H_i\right)}.\end{aligned}$$

Pri tome, prva jednakost slijedi iz definicije uvjetne vjerojatnosti, druga iz relacije (1.2), a treća iz formule potpune vjerojatnosti. □

Često se pod nazivom "Bayesova formula" nalazi i sljedeći izraz:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{1.3}$$

On se lako izvodi preko definicije uvjetne vjerojatnosti i relacije (1.2).

Definicija 1.1.6. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ koja je $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$ - izmjeriva (odnosno $X^{-1}(B) \in \mathcal{F}$, $\forall B \in \mathcal{B}(\mathbb{R})$) naziva se* ***slučajna varijabla****.*

Za slučajnu varijablu X možemo definirati funkciju $\mathbb{P}_X : \mathcal{B}(\mathbb{R}) \to [0, 1]$ izrazom $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$. Provjerom svojstava iz definicije, slijedi da je $\mathbb{P}_X$ vjerojatnosna mjera na $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ te ju nazivamo **vjerojatnosna mjera inducirana s X** ili **zakon razdiobe od X**.

Definicija 1.1.7. *Neka je X slučajna varijabla na Ω.* ***Funkcija distribucije*** *od X je funkcija $F_X : \mathbb{R} \to [0, 1]$ definirana s $F_X(x) = \mathbb{P}(\{\omega \in \Omega;\ \ X(\omega) \leq x\}) = \mathbb{P}(X \leq x)$.*

Vjerojatnost je normirana mjera, slučajne varijable izmjerive su funkcije, no upravo je funkcija distribucije jedan od osnovnih pojmova teorije vjerojatnosti i čini glavni iskorak u odnosu na teoriju mjere. Može se pokazati da je funkcija distribucije F slučajne varijable rastuća, neprekidna zdesna i zadovoljava

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0,$$
$$F(+\infty) = \lim_{x \to +\infty} F(x) = 1.$$

Štoviše, vrijedi i da je svakom funkcijom F s navedenim svojstvima jednoznačno određena vjerojatnosna mjera koju možemo interpretirati kao zakon razdiobe neke slučajne varijable. Time je opravdano što se uglavnom, umjesto o slučajnim varijablama, govori o njihovim funkcijama distribucije.

Za sada se ponovo vraćamo na pojam slučajne varijable i dva glavna tipa koja razlikujemo.

Definicija 1.1.8. *Slučajna varijabla X je* ***diskretna*** *ako postoji konačan ili prebrojiv skup $D \subset \mathbb{R}$ takav da je $\mathbb{P}(X \in D) = 1$.*

Definicija 1.1.9. *Neka je X slučajna varijabla na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ i neka je F_X njena funkcija distribucije. Kažemo da je X* ***neprekidna slučajna varijabla*** *ako postoji nenegativna realna Borelova funkcija f na $\mathbb{R}$ takva da je*

$$F_X(x) = \int_{-\infty}^{x} f(t) d\lambda(t). \tag{1.4}$$

Za funkciju distribucije F_X neprekidne slučajne varijable X kažemo da je ***neprekidna funkcija distribucije****, a funkciju f iz relacije* (1.4) *nazivamo* ***funkcija gustoće*** *od X.*

Pojam slučajne varijable možemo poopćiti za više dimenzija na pojam slučajnog vektora.

Definicija 1.1.10. *Kažemo da je $X : \Omega \to \mathbb{R}$* ***k-dimenzionalna slučajna veličina*** *ako je X izmjerivo preslikavanje u paru σ-algebri $(\mathcal{F}, \mathcal{B}(\mathbb{R}^k))$ tj. ako vrijedi da je $(\forall B \in \mathcal{B}(\mathbb{R}^k))$ $\{X \in B\} \in \mathcal{F}$. Ako je k = 1, X zovemo slučajna varijabla, a ako je $k \geq 2$, X zovemo* ***slučajni vektor****.*

S obzirom na to da smo jednodimenzionalni slučaj već obradili, nastavljamo s definicijama za slučajni vektor (iako svaka od njih vrijedi i za k=1).

Definicija 1.1.11. *Neka je $X : \Omega \to \mathbb{R}$ k-dimenzionalni slučajni vektor. Induciranu vjerojatnost $\mathbb{P}_X$, definiranu na izmjerivom prostoru $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ s $\mathbb{P}_X(B) := \mathbb{P}(X \in B)$, $B \in \mathcal{B}(\mathbb{R}^k)$ zovemo* ***zakon razdiobe od X****.*

Definicija 1.1.12. *Neka je X k-dimenzionalni slučajni vektor sa zakonom razdiobe $\mathbb{P}_X$. Funkciju $F_X : \mathbb{R}^k \to \mathbb{R}$ definiranu sa $F_X(x) := \mathbb{P}_X(\langle -\infty, x])$, $x \in \mathbb{R}^k$, zovemo* ***funkcija razdiobe*** *(ili distribucije) od X.*

Svojstva funkcije distribucije slučajne varijable mogu se poopćiti na svojstva funkcije distribucije slučajnog vektora, te se ponovo može pokazati da svaka funkcija koja zadovoljava ta svojstva na jedinstveni način zadaje jedan slučajni vektor.

Definicija 1.1.13. *Kažemo da je k-dimenzionalni slučajni vektor X* ***neprekidan*** *ako postoji nenegativna Borelova funkcija f_X definirana na $\mathbb{R}^k$ takva da se funkcija razdiobe F_X može prikazati na sljedeći način:*

$$F_X(x) = \int_{\langle -\infty, x]} f_X(y) d\lambda(y), \quad x \in \mathbb{R}^k.$$

Funkciju f_X zovemo funkcija gustoće razdiobe od X, ili kraće, ***gustoća*** *od X.*

Dodatno, reći ćemo da je k-dimenzionalni slučajni vektor X **diskretan** ako postoji neki konačni ili prebrojivi skup $D \subset \mathbb{R}^k$ takav da je $\mathbb{P}\{X \in D\} = 1$.
Treba još naglasiti da, kod diskretnih slučajnih veličina, koristimo oznaku
$f_X(x) = \mathbb{P}(X = x)$ za $x \in \mathbb{R}$, tj. $x \in \mathbb{R}^k$.

Na početku smo definirali pojam uvjetne vjerojatnosti za događaje, a sada ga želimo uklopiti u kontekst slučajnih veličina, odnosno definirati i pojmove uvjetne distribucije i uvjetne gustoće.

Definicija 1.1.14. *Za diskretni slučajni vektor (X, Y) s gustoćom $f_{X,Y}$,* ***marginalna gustoća od X*** *je funkcija jedne varijable dana izrazom*

$$f_X(x) = \sum_{y \in ImY} f_{X,Y}(x, y), \quad x \in \mathbb{R}.$$

To je ujedno gustoća diskretne slučajne varijable X. Sasvim se analogno definira i marginalna gustoća od Y.

Definicija 1.1.15. *Neka je (X, Y) diskretni slučajni vektor sa zajedničkom gustoćom $f_{X,Y}$ i marginalnim funkcijama vjerojatnosti f_X, f_Y . Ako je $f_Y(y) = \mathbb{P}(Y = y)>0$ za $y \in ImY$, tada je **uvjetna funkcija vjerojatnosti od X za dano Y = y** funkcija $x \to f_{X|Y}(x \mid y)$ dana sa*

$$f_{X|Y}(x \mid y) := \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)},$$

i to je gustoća diskretne vjerojatnosne razdiobe koja odgovara uvjetnoj distribuciji od X za dano Y = y. Analogno se definira uvjetna funkcija vjerojatnosti (ili uvjetna gustoća) od Y uz dano X = x.

Definicija 1.1.16. *Za neprekidni slučajni vektor (X, Y) s gustoćom $f_{X,Y}$, **marginalna gustoća od X** je funkcija jedne varijable dana izrazom*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy, \quad x \in \mathbb{R}.$$

To je ujedno gustoća vjerojatnosne razdiobe neprekidne slučajne varijable X. Sasvim analogno definira se i marginalna gustoća od Y.

Definicija 1.1.17. *Neka je (X, Y) neprekidni slučajni vektor sa zajedničkom gustoćom $f_{X,Y}$ i marginalnim gustoćama f_X, f_Y , te neka je $f_Y(y)>0$ za $y \in ImY$. Tada je **uvjetna gustoća od X za dano Y = y** funkcija $x \to f_{X|Y}(x \mid y)$ dana s*

$$f_{X|Y}(x \mid y) := \frac{f_{X,Y}(x, y)}{f_Y(y)}, \quad x \in \mathbb{R}.$$

To je ujedno funkcija gustoće neprekidne vjerojatnosne razdiobe koja odgovara uvjetnoj distribuciji od X uz dano Y = y. Analogno se definira uvjetna funkcija vjerojatnosti (ili uvjetna gustoća) od Y uz dano X = x.

Gornje su definicije izrečene u slučaju dvodimenzionalnog vektora (X,Y) zbog jednostavnosti indeksacije. Mogu se poopćiti na slučaj kada imamo slučajni vektor $X = (X_1, X_2, \ldots, X_k)$ te promatramo marginalne i uvjetne gustoće ℓ-dimenzionalne slučajne veličine $X = (X_{j_1}, \ldots, X_{j_l})$. Sljedeći teorem služi samo kao primjer takvog poopćenja.

Teorem 1.1.18. *Neka je $X = (X_1, X_2, \ldots, X_k)$ neprekidni slučajni vektor s gustoćom f_X. Tada je svaka komponenta $X = (X_{j_1}, ..., X_{j_l})$ od X dimenzije ℓ $(1 \leq \ell < k, 1 \leq j_1 < j_2 < \ldots < j_\ell \leq k)$, neprekidna ℓ-dimenzionalna slučajna veličina i za gustoću $f_{X'}$ od X' vrijedi:*

$$f_{X'}(x_1, \ldots, x_\ell) =$$
$$= \int_{\mathbb{R}^{k-\ell}} f\left(y_1, .., y_{j_1-1}, x_1, y_{j_1+1}, .., y_{j_\ell-1}, x_\ell, y_{j_\ell+1}, .., y_k\right) d\left(y_1, .., y_{j_1-1}, y_{j_1+1}, .., y_{j_\ell-1}, y_{j_\ell+1}, .., y_k\right),$$

za λ^ℓ-g.s. $(x_1, ..., x_\ell) \in \mathbb{R}^\ell$.

Sljedeći će nam teorem također biti od operativne važnosti.

Teorem 1.1.19. *Neka je $X = (X_1, X_2, \ldots, X_n)$ neprekidni slučajni vektor s gustoćom f_X, neka je $L = \{x \in \mathbb{R}^n : f_X(x) > 0\}$ i neka je $g : \mathbb{R}^n \to \mathbb{R}^n$ Borelova funkcija takva da vrijedi:*

1. *$g : L \to T = g(L)$ je bijekcija*
2. *g^{-1} je glatka na T i $Dg^{-1} \neq 0$ $\forall y \in T$.*

Tada je $Y = g(X)$ neprekidni slučajni vektor s funkcijom gustoće

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|Dg^{-1}(y)\right|\chi_T(y) \tag{1.5}$$

Završavamo ovaj odjeljak s još jednom definicijom koja će nam biti od važnosti u modeliranju slučajnih pokusa, a to je nezavisnost slučajnih veličina. Operativno su nam bitne i njene karakterizacije preko funkcija distribucije i gustoće.

Definicija 1.1.20. *Niz $(X_n; n \in I)$ je niz **nezavisnih** slučajnih veličina (varijabli ili vektora) ako za svaki konačni podskup $I = \{n_1, n_2, \ldots, n_\ell\}$ skupa indeksa I (pri čemu $\ell = |\ I\ |$) i sve $B_1 \in \mathcal{B}(\mathbb{R}^{k_{n_1}}), B_2 \in \mathcal{B}(\mathbb{R}^{k_{n_2}}), ..., B_\ell \in \mathcal{B}(\mathbb{R}^{k_{n_\ell}})$ vrijedi*

$$\mathbb{P}(X_{n_1} \in B_1, X_{n_2} \in B_2, \ldots, X_{n_\ell} \in B_\ell) = \mathbb{P}(X_{n_1} \in B_1)\mathbb{P}(X_{n_2} \in B_2) \ldots \mathbb{P}(X_{n_\ell} \in B_\ell).$$

Teorem 1.1.21. *Neka je $X = (X_1, X_2, \ldots, X_k)$ k-dimenzionalni slučajni vektor. Komponente $X_1, X_2, ..., X_k$ vektora X su nezavisne ako i samo ako za funkciju razdiobe F_X od X vrijedi*

$$(\forall (x_1, x_2, ..., x_k) \in \mathbb{R}^k) \quad F_X(x_1, x_2, \ldots, x_k) = F_{X_1}(x_1)F_{X_2}(x_2) \ldots F_{X_k}(x_k),$$

gdje su $F_{X_1}, F_{X_2}, \ldots, F_{X_k}$ funkcije distribucije komponenti od X.

Teorem 1.1.22. *Komponente $X_1, X_2, \ldots, X_k$ neprekidnog slučajnog vektora $X = (X_1, X_2, \ldots, X_k)$ s gustoćom f_X su nezavisne ako i samo ako vrijedi*

$$f_X(x_1, x_2, \ldots, x_k) = f_{X_1}(x_1)f_{X_2}(x_2) \ldots f_{X_k}(x_k),$$

za sve $(x_1, x_2, ..., x_k) \in \mathbb{R}^k$ osim možda za točke iz skupa Lebesgueove mjere nula.

1.2 Osnove statistike

U većini znanstvenih istraživanja, proučavamo jedno ili više (statističkih) obilježja i želimo na temelju prikupljenog **uzorka** izvesti zaključke o tom obilježju u **populaciji**. Oslanjajući se na teoriju vjerojatnosti, statističko obilježje modeliramo slučajnom varijablom te nas zanima njena (populacijska) distribucija.
Sljedećim pojmovima formalizira se takav pristup.

Definicija 1.2.1. *Neka je* $(\Omega, \mathcal{F})$ *izmjeriv prostor i* $\mathcal{P}$ *množina vjerojatnosnih mjera definiranih na* $(\Omega, \mathcal{F})$*. Tada uredenu trojku* $(\Omega, \mathcal{F}, \mathcal{P})$ *zovemo* ***statistička struktura****.*

Statističku strukturu $(\Omega, \mathcal{F}, \mathcal{P})$ gdje je $\mathcal{P} = \{\mathbb{P}\}$ možemo poistovjetiti s vjerojatnosnim prostorom $(\Omega, \mathcal{F}, \mathbb{P})$. Množina $\mathcal{P}$ uglavnom je parametrizirana parametrom θ, u zapisu $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. Pri tome, Θ nazivamo parametarskim prostorom.

Definicija 1.2.2. ***Slučajan uzorak*** $X_1, X_2, \ldots, X_n$ *duljine n na statističkoj strukturi* $(\Omega, \mathcal{F}, \mathcal{P})$ *niz je slučajnih veličina na* $(\Omega, \mathcal{F})$ *koje su nezavisne i jednako distribuirane u odnosu na svaku vjerojatnost* $\mathbb{P} \in \mathcal{P}$*.*

Slučajni uzorak predstavlja niz opažanja ili mjerenja vrijednosti promatranog obilježja na jedinkama koje čine uzorak i koje su slučajno odabrane iz populacije. Pri odabiru uzorka, pretpostavlja se da sve jedinke iz populacije imaju jednaku vjerojatnost biti izabrane u uzorak te su odabrane neovisno jedna od druge. Zbog toga uzorak modeliramo slučajnim veličinama koje su nezavisne i imaju distribuciju jednaku populacijskoj distribuciji veličine X.

Definicija 1.2.3. *Uređena n-torka brojeva* $x = (x_1, x_2, \ldots, x_n)$ *koja predstavlja realizaciju slučajnog uzorka X naziva se* ***opaženi uzorak****.*

Definicija 1.2.4. ***Statistika*** *na statističkoj strukturi* $(\Omega, \mathcal{F}, \mathcal{P})$ *svaka je slučajna veličina koja je izmjeriva funkcija nekog slučajnog uzorka na toj strukturi.*

Dakle, statistika je funkcija slučajnih veličina, pa je i sama slučajna veličina te ima razdiobu koju nazivamo **uzoračkom razdiobom**.

Primjer 1.2.5. *Pretpostavimo da promatramo statističko obilježje modelirano varijablom X te neka je* $X_1, X_2, \ldots, X_n$ *slučajni uzorak. Često su nam od interesa statistike:*

- ***uzoračka sredina*** $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- ***uzoračka varijanca*** $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

Poglavlje 2

Bayesovska statistika

U ovom je poglavlju cilj objasniti dva pristupa u statističkom zaključivanju - frekvencionistički i bayesovski, te se zatim fokusirati na polazišne točke i temeljne rezultate bayesovske statistike.

2.1 Statističko zaključivanje

Jedan od glavnih doprinosa statistike kao znanstvene discipline jest **statističko zaključivanje**, koje obuhvaća teoriju, metode i postupke izvlačenja informacija, izvođenja zaključaka i ocjene novog znanja na temelju prikupljenih podataka.
S obzirom da su podaci kojima raspolažemo pri statističkom zaključivanju nepotpuni (oni čine uzorak - samo mali dio promatrane populacije) i sadržavaju slučajnu komponentu, zaključci koje donosimo **nesigurni** su. Matematički, tu nesigurnost modeliramo pomoću teorije vjerojatnosti.
Povijesni je razvoj statistike zamršen i brojni su mu matematičari doprinijeli, no promatrajući teorijske rezultate možemo primjetiti dva osnovna pristupa statističkom zaključivanju - **frekvencionistički** i **bayesovski pristup**.

2.2 Frekvencionistički pristup

Kako nije u fokusu ovog diplomskog rada, ističemo samo nekoliko značajki i primjera frekvencionističkog zaključivanja koji služe za usporedbu s bayesovskim pristupom. Više detalja moguće je pronaći u [4].

Pretpostavimo da promatramo neko statističko obilježje. Podaci kojima raspolažemo predstavljaju uzorak uzet iz populacije i nad tim se uzorkom promatra realizacija statistike

(koja ovdje predstavlja slučajnu varijablu - izmjerivu funkciju slučajnog uzorka). Najčešće želimo izvesti zaključak o distribuciji statističkog obilježja u populaciji, odnosno o njezinim parametrima.

Polazišna je ideja frekvencionističkog pristupa da je parametar vjerojatnosne distribucije koja nam je od interesa neka fiksna, ali nama nepoznata, vrijednost. Ključan je korak utvrđivanje vjerojatnosne distribucije statistike s obzirom na sve moguće realizacije slučajnog uzorka i ta se distribucija naziva distribucija uzorkovanja (eng. *sampling distribution*) statistike. Parametar populacijske distribucije obilježja uglavnom se pojavljuje u izrazu za statistiku i/ili njenu distribuciju. Zaključci o vjerojatnosti donose se za statistička obilježja, ali ne i za parametre. Parametar se procjenjuje točkovno ili se na temelju distribucije uzorkovanja statistike izvodi interval pouzdanosti za parametar.

O kvaliteti statističke procedure govori to, kako se ona ponašaja dugoročno, u beskonačno mnogo hipotetskih ponavljanja pokusa.

Primjer 2.2.1. *Zanima nas sistolički krvni tlak studenata Sveučilišta u Zagrebu. U svrhu istraživanja, na slučajan način odabrali smo n studenata i izmjerili im tlak.*

Možemo pretpostaviti da je sistolički krvni tlak u populaciji normalno distribuiran sa standardnom devijacijom $\sigma = 20$, a očekivanje μ nepoznati je parametar. Želimo na temelju uzorka izvesti zaključke o očekivanoj vrijednosti tlaka studenata Sveučilišta, odnosno o nepoznatom parametru μ.

Dakle, statističko obilježje koje promatramo sistolički je krvni tlak te ga modeliramo neprekidnom slučajnom varijablom X, $X \sim N(\mu, 20^2)$. Populacija su studenti Sveučilišta u Zagrebu, a podaci kojima raspolažemo predstavljaju realizaciju slučajnog uzorka duljine n.

Statistika koju u ovom slučaju promatramo je uzoračka sredina $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Kao što je već naglašeno, $\bar{X}$ je izmjeriva funkcija slučajnog uzorka, pa je slučajna varijabla i može se pokazati da je uzoračka distribucija ove statistike $N(\mu, \frac{\sigma^2}{n})$, tj. $N(\mu, \frac{20^2}{n})$. Upravo se $\bar{X}$ uzima kao točkovni procjenitelj za parametar μ.

Ovdje možemo ilustrirati i raniju primjedbu da za ocjenu kvalitete statističke procedure promatramo kako se ona ponaša u beskonačno mnogo ponavljanja pokusa. Zato uvodimo pojam konzistentnosti procjenitelja. Za procjenitelj T_n parametra τ za koji je $(\mathbb{P})\lim_{n\to\infty} T_n = \tau$ kažemo da je (slabo) konzistentan. T_n je jako konzistentan za τ ako je $\mathbb{P}(\lim_{n\to\infty} T_n = \tau) = 1$. Odnosno, T_n slabo je konzistentan procjenitelj za τ, ako niz slučajnih varijabli $(T_n)_{n\in\mathbb{N}}$ konvergira po vjerojatnosti prema τ, a jako je konzistentan procjenitelj ako $(T_n)_{n\in\mathbb{N}}$ konvergira g.s. prema τ. Obje tvrdnje govore o ponašanju statistike T_n u slučaju kada veličina uzorka (ponavljanje pokusa) n ide u beskonačnost. Može se pokazati da je $\bar{X}$ konzistentan procjenitelj za parametar μ.

Na temelju uzoračke distribucije $\bar{X}$, možemo odrediti i intervalnu procjenu za μ, odnosno interval sa svojstvom da sadrži pravu vrijednost od μ s pouzdanosti od 95%.

$$\bar{X} \sim N(\mu, \frac{20^2}{n}) \;\Rightarrow\; \frac{\bar{X}-\mu}{20}\sqrt{n} \sim N(0,1)$$
$$0.95 = \mathbb{P}\left(\left|\frac{\bar{X}_n-\mu}{20}\sqrt{n}\right| \leq 1.96\right) =$$
$$= \mathbb{P}\left(\bar{X}_n - 1.96\frac{20}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 1.96\frac{20}{\sqrt{n}}\right)$$

Stoga je traženi interval upravo:

$$\left[\bar{X}_n - 1.96\frac{20}{\sqrt{n}}, \bar{X}_n + 1.96\frac{20}{\sqrt{n}}\right]$$

Ovaj interval interpretiramo na sljedeći način: kako broj ponavljanja pokusa teži u beskonačnost, tako udio ovako izračunatih intervala koji sadrži pravu vrijednost parametra teži u 0.95. Da ilustriramo, kada bismo čitavi postupak prikupljanja uzoraka duljine n i gornjih izračuna ponovili još 100 puta, oko 95 intervala pouzdanosti sadržavalo bi pravu vrijednost od μ.

2.3 Bayesovski pristup

Ovaj je pristup dobio ime po (već spomenutom) Thomasu Bayesu, engleskom matematičaru iz 18. stoljeća. Za života je objavio samo jednu teološku raspravu i matematički spis, a najpoznatije mu je djelo *Esej o rješavanju jednog problema u nauku o vjerojatnosti (Essay Towards Solving a Problem in the Doctrine of Chances*, 1763) objavljeno posthumno, dvije godine nakon njegove smrti. U eseju promatra niz nezavisnih pokusa, čiji je ishod uspjeh s vjerojatnosti p ili neuspjeh s vjerojatnosti 1-p te nastoji riješiti pitanje uvjetne distribucije parametra p s obzirom na opaženi broj uspjeha i neuspjeha. Preciznije, za binomnu slučajnu varijablu $X \sim B(n,p)$, traži $\mathbb{P}(a < p < b \mid X = k)$. Pri tome, postavlja i dokazuje formulu uvjetne vjerojatnosti događaja:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

iz koje se izvodi izraz 1.3:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)},$$

koji se još naziva i Bayesov teorem za uvjetne vjerojatnosti. [10]

Iz definicije uvjetne funkcije vjerojatnosti, izvodi se i Bayesova formula za funkcije gustoće:

$$f(\theta \mid x) = \frac{f(\theta, x)}{f(x)} = \frac{f(\theta)f(x \mid \theta)}{f(x)}. \tag{2.1}$$

U ovoj notaciji, promatra se uvjetna gustoća slučajne varijable koja predstavlja nepoznati parametar θ, s obzirom na opaženu vrijednost $\mathbf{x}$ slučajnog uzorka $\mathbf{X} = (X_1, X_2, \ldots, X_n)$.

Izraz 2.1 predstavlja glavnu ideju bayesovske statistike. Kao i kod frekvencionističkog pristupa, promatra se neko statističko obilježje X i parametri njegove populacijske distribucije. U bayesovskom se pristupu o tim nepoznatim parametrima ne razmišlja kao o fiksnim vrijednostima, već su oni sami slučajne varijable i govori se o njihovim vjerojatnosnim distribucijama.

Prije opažanja, odnosno prikupljanja podataka, za parametar od interesa θ odabire se **apriorna gustoća**, u 2.1 označena s $f(\theta)$. Nakon što su za slučajni uzorak opažene vrijednosti $\mathbf{x}$, one služe kao novi izvor znanja koje se koristi za ažuriranje apriorne gustoće izvođenjem **aposteriorne gustoće**, u 2.1 označene s $f(\theta \mid \mathbf{x})$. Pri tome, $f(\mathbf{x} \mid \theta)$ označava uvjetnu gustoću slučajnog uzorka $\mathbf{X}$ za dani parametar θ i naziva se **vjerodostojnost** (eng. *likelihood*). Ovaj je pojam bitan i u frekvencionističkom pristupu, gdje se nešto drukčije označava i shvaća kao funkcija parametra za fiksnu realizaciju uzorka $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$. Oznaka $f(\mathbf{x})$ predstavlja marginalnu gustoću slučajnog uzorka u točki $\mathbf{X} = \mathbf{x}$ i naziva se **marginalna vjerodostojnost** (eng. *marginal likelihood*). Računa se integriranjem vjerodostojnosti $f(\mathbf{x} \mid \theta)$ (kao funkcije parametra) s obzirom na mjeru zadanu distribucijom parametra:

$$f(\mathbf{x}) = \sum_{\theta} f(\mathbf{x} \mid \theta) f(\theta), \tag{2.2}$$

$$f(\mathbf{x}) = \int f(\mathbf{x} \mid \theta) f(\theta) d\theta. \tag{2.3}$$

Izraz 2.2 koristi se u slučaju kada parametar ima diskretnu distribuciju, a 2.3 kada ima neprekidnu distribuciju.

S obzirom da $f(\mathbf{x})$, za opažanje $\mathbf{x}$ koje je fiksno, ne ovisi o vrijednosti parametra, to je konstanta koja služi za normalizaciju vjerojatnosne funkcije gustoće $f(\theta \mid \mathbf{x})$ pa se često piše i samo:

$$f(\theta \mid \mathbf{x}) \propto f(\theta) f(\mathbf{x} \mid \theta). \tag{2.4}$$

Ovaj nam izraz govori da su navedene funkcije proporcionalne, odnosno da postoji konstanta c (zapravo jednaka $1/f(\mathbf{x})$) takva da $f(\theta \mid \mathbf{x}) = c \cdot f(\theta) f(\mathbf{x} \mid \theta)$. Pri tome, $f(\theta) f(\mathbf{x} \mid \theta)$ još nazivamo i **nenormalizirana aposteriorna funkcija gustoće** (eng. *unnormalized posterior density function*). U nenormaliziranoj gustoći sadržana je sva potrebna informacija o aposteriornoj distribuciji, jer je konstanta c jednoznačno određena činjenicom da za $f(\theta \mid \mathbf{x})$, kao vjerojatnosnu funkciju gustoće, vrijedi:

$$1 = \int_{-\infty}^{+\infty} f(\theta \mid \mathbf{x}) d\theta = c \int_{-\infty}^{+\infty} f(\theta) f(\mathbf{x} \mid \theta) d\theta$$

U kontekstu bayesovske statistike, marginalna distribucija uzorka (čija je gustoća marginalna vjerodostojnost $f(\mathbf{x})$) naziva se još i **apriorna prediktivna distribucija** (eng. *prior predictive distribution*), jer se odnosi na distribuciju uzorka prije opažanja njegovih vrijednosti.

Ako nas nakon realizacije uzorka $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ zanima funkcija gustoće statističkog obilježja u točki $\tilde{x}$ (u nešto širem smislu, često nas zanima vjerojatnost da, nakon n opaženih vrijednosti, iduće opažanje bude određena vrijednost ili u određenom intervalu), tada to računamo na sljedeći način:

$$\begin{aligned} f(\tilde{x} \mid \mathbf{x}) &= \int f(\tilde{x}, \theta \mid \mathbf{x}) d\theta \\ &= \int f(\tilde{x} \mid \theta, \mathbf{x}) f(\theta \mid \mathbf{x}) d\theta \\ &= \int f(\tilde{x} \mid \theta) f(\theta \mid \mathbf{x}) d\theta. \end{aligned} \tag{2.5}$$

Treća jednakost slijedi iz druge po definiciji slučajnog uzorka, jer su $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i X_{n+1} nezavisni. Za uvjetnu distribuciju od X_{n+1} uz dano $\mathbf{X} = \mathbf{x}$ (čija je gustoća $f(\tilde{x} \mid \mathbf{x})$) koristi se još i naziv **aposteriorna prediktivna distribucija** (eng. *posterior predictive distribution*), jer se radi o distribuciji buduće realizacije varijable na temelju dosadašnjih opažanja. Također, iz 2.5 možemo iščitati da je funkcija gustoće od X_{n+1} uvjetno na realizaciju uzorka $\mathbf{x}$, u novom opažanju $\tilde{x}$ zapravo očekivanje slučajne varijable $f(\tilde{x} \mid \theta)$ (zapravo funkcije $f(\tilde{x} \mid \cdot)$ slučajne varijable θ), uz aposteriornu gustoću za θ, $f(\theta \mid \mathbf{x})$.

Sažeto, gore uvedeni pojovi predstavljaju korake bayesovskog zaključivanja: na početku se odabire vjerojatnosni model za sve statističke veličine (distribucija slučajne veličine X, apriorna gustoća/distribucija za nepoznati parametar), zatim se prikupljaju podaci, tj. opaža realizacija slučajnog uzorka te se početno znanje o parametru (sadržano u apriornoj gustoći) mijenja po izrazu 2.1, a novo je znanje sadržano u aposteriornoj gustoći/distribuciji parametra. Aposteriornu gustoću možemo interpretirati kao relativne težine koje pridajemo mogućim vrijednostima parametra nakon što u obzir uzmemo prikupljene podatke. Ukoliko nas zanimaju buduća opažanja, promatramo aposteriornu prediktivnu distribuciju.

Kako bismo dali motivaciju za razmišljanje o nepoznatom parametru kao o slučajnoj varijabli i za diskusiju o distribucijama parametra, navodimo sljedeću definiciju i teorem.

Definicija 2.3.1. *Kažemo da su slučajne varijable $X_1, X_2, \ldots, X_n$ **izmjenjive** ako svaka permutacija od $(X_1, X_2, \ldots, X_n)$ ima istu zajedničku distribuciju kao i bilo koja druga permutacija, tj. $(X_1, X_2, \ldots, X_n) \overset{D}{=} (X_{\pi_{(1)}}, X_{\pi_{(2)}}, \ldots, X_{\pi_{(n)}})$ za svaku permutaciju π skupa $\{1, 2, \ldots, n\}$. Niz $(X_n)_{n \in \mathbb{N}}$ slučajnih varijabli je izmjenjiv ako mu je svaki konačan podskup izmjenjiv.*

U mnogim se statističkim analizama koristi pretpostavka da su slučajne varijable nezavisne i jednako distribuirane (n.j.d.). Izmjenjivost je općenitiji pojam; svaki je niz $(X_n)_{n\in\mathbb{N}}$ nezavisnih i jednako distribuiranih slučajnih varijabli izmjenjiv, ali obrat ne vrijedi. Na primjer, ako je niz slučajnih varijabli $(X_n)_{n\in\mathbb{N}}$ n.j.d., a X_0 netrivijalna slučajna varijabla nezavisna od toga niza, tada je $(X_n + X_0)_{n\in\mathbb{N}}$ izmjenjiv, ali ne i n.j.d. (jer varijable više nisu nezavisne).

Teorem 2.3.2. *(De Finetti) Niz slučajnih varijabli $(X_n)_{n\in\mathbb{N}}$ je izmjenjiv ako i samo ako postoji mjera P na Θ takva da za svaki n vrijedi:*

$$f(x_1, x_2, \ldots, x_n) = \int \prod_{i=1}^{n} f(x_i \mid \theta)\, P(d\theta). \tag{2.6}$$

Ako pretpostavimo da vrijedi 2.6, jer je produkt $\prod_{i=1}^{n} f(x_i \mid \theta)$ invarijantan na permutacije indeksa, odmah slijedi da je niz slučajnih varijabli $(X_n)_{n\in\mathbb{N}}$ izmjenjiv. Drugi smjer teorema kaže da za svaki izmjenjivi niz slučajnih varijabli $(X_n)_{n\in\mathbb{N}}$, postoje skup (mogućih) parametara Θ, vjerodostojnost $f(x \mid \theta)$ i vjerojatnosna mjera P na Θ tako zadani da su $(X_1, X_2, \ldots, X_n)$ uvjetno na njih nezavisni, za svaki $n \in \mathbb{N}$.

S obzirom da je, po definiciji, beskonačni slučajni uzorak izmjenjivi niz slučajnih varijabli, vidimo da postoji veza između funkcije gustoće (beskonačnog) slučajnog uzorka i vjerojatnosne mjere definirane nad prostorom dozvoljenih parametara θ. Time se na neki način može odgovoriti na pitanje, zašto promatrati parametre populacijske distribucije kao slučajne varijable s vlastitim distribucijama.

Pojam apriorne gustoće jedan je od centralnih u bayesovskom pristupu i izbor odgovarajuće zna biti izazov. Stoga uvodimo nekoliko definicija koje se odnosne upravo na apriornu gustoću/ distribuciju.

Definicija 2.3.3. *Neka je $\mathcal{F}$ klasa funkcija vjerodostojnosti $f(\cdot \mid \theta)$, a $\mathcal{P}$ klasa apriornih gustoća za θ. Kažemo da je $\mathcal{P}$ **konjugirana klasa gustoća** za $\mathcal{F}$ ako*

$$f(\theta \mid x) \in \mathcal{P} \;\; za\ sve \;\; f(\cdot \mid \theta) \in \mathcal{F} \;\; i \;\; f(\cdot) \in \mathcal{P},$$

tj. ako je aposteriorna gustoća od θ u $\mathcal{P}$, za svaku vjerodostojnost iz $\mathcal{F}$ i za svaku apriornu gustoću iz $\mathcal{P}$.

Ova definicija nije sasvim formalna, jer ako za $\mathcal{P}$ uzmemo klasu svih funkcija gustoće, ona je konjugirana za svaki izbor klase vjerodostojnosti $\mathcal{F}$. Ipak, u primjeru na kraju poglavlja ilustrirano je kako se konjugirane klase gustoća često prirodno izvode te pojednostavljuju izračun aposterirorne gustoće i omogućavaju njenu lakšu interpretaciju.

Uvodimo još nekoliko pojmova koji se u kontekstu apriorne distribucije pojavljuju u literaturi.

U situacijama kada se ne raspolaže s dodatnim znanjima o populaciji i parametru, nastoje se koristiti one apriorne distribucije koje što manje utječu na izračun aposteriorne distribucije. Takve se onda nazivaju **neinformativne** (eng. *noninformative prior distributions*). Primjer neinformativne distribucije za parametar p binomne distribucije je uniformna distribucija na segmentu $[0, 1]$. **Slabo informativne** distribucije (eng. *weakly informative prior distributions*) sadržavaju neke informacije, često s ciljem da aposteriorna distribucija bude definirana na razumnom ili ograničenom skupu vrijednosti, ali se njima ne pokušava sažeti svo dosadašnje znanje o promatranom parametru. **Informativne** apriorne distribucije (eng. *informative prior distributions*) bile bi one koje su izabrane s ciljem da sadržavaju što više dosadašnjeg znanja o parametru. Ovi su pojmovi detaljnije razrađeni i potkrijepljeni primjerima u [3].

Često je prvi izazov u bayesovskom pristupu izabrati odgovarajuću apriornu ditribuciju. Osim što se često koriste nenormalizirane apriorne funkcije gustoće f (za koje vrijedi $\int f(\theta)d\theta = c$, pa je s $\frac{1}{c}f(x)$ dana vjerojatnosna funkcija gustoće), nekad se zadaju i **neprave** apriorne gustoće (eng. *improper prior*), za koje vrijedi $\int f(\theta)d\theta = +\infty$ (dakle, uopće ih se ne može smatrati vjerojatnosnim funkcijama gustoće i ne možemo reći da zadaju vjerojatnosnu distribuciju). Ipak, u nekim slučajevima, upotrebom tih apriornih nepravih gustoća možemo doći do **prave** (eng. *proper*) aposteriorne funkcije gustoće, za koju vrijedi $\int f(x) = 1$. Takva je situacija ilustrirana u Primjeru 4.3.1. .

Primjer 2.3.4. *Primjer je preuzet iz [3].*

Hemofilija je bolest koja se nasljeđuje preko X kromosoma. Kako muškarci imaju jedan X i jedan Y kromosom, a žene dva X kromosoma, oni muškarci koji naslijede X kromosom s genom koji uzrokuje hemofiliju od nje i obolijevaju, a one žene kod kojih samo jedan X kromosom sadrži gen odgovoran za bolest ne obolijevaju.

Promatramo slučaj žene čiji brat ima hemofiliju,a otac ne. Dakle, njezina majka nositeljica je gena za hemofiliju, te je gen prisutan na jednom majčinom X kromosomu, a na drugom ne. Možemo stoga zaključiti da je žena koju promatramo nositeljica gena s vjerojatnosti 0.5 i nije nositeljica s vjerojatnosti 0.5. Ako s θ označimo stanje žene, te $\theta = 0$ označava da nije nositeljica i $\theta = 1$ označava da jest, slijedi da je apriorna distribucija od θ Bernoullijeva s parametrom 0.5, tj. vrijedi $\mathbb{P}(\theta = 0) = 0.5$ i $\mathbb{P}(\theta = 1) = 0.5$.

Pretpostavimo da žena ima dva sina i neka su podaci x_1, x_2 koje smo opazili zdravstvena stanja sinova, gdje ponovo 0 označava da bolest nije prisutna, a 1 da jest. Stanja sinova izmjenjive su i, uvjetno na nepoznato majčino stanje θ, nezavisne slučajne varijable. Pretpostavimo da niti jedan od sinova nema hemofiliju. Ako je žena nositeljica, vjerojatnost da njezin sin nema hemofiliju je 0.5, a ako nije nositeljica, ta je vjerojatnost 1. Stoga je vjerojatnosti opaženog ishoda (da niti jedan od sinova nema bolest), uvjetno na θ:

$$\mathbb{P}(x_1 = 0, x_2 = 0 \mid \theta = 1) = 0.5 \times 0.5 = 0.25,$$

$$\mathbb{P}(x_1 = 0, x_2 = 0 \mid \theta = 0) = 1 \times 1 = 1.$$

Sada, koristeći 1.3, možemo izračunati i aposteriornu vjerojatnost da je žena nositeljica gena za hemofiliju, uz oznaku $\mathbf{x} = (x_1, x_2) = (0, 0)$*:*

$$\begin{aligned}\mathbb{P}(\theta = 1 \mid \mathbf{x} = (0,0)) &= \frac{\mathbb{P}(\mathbf{x} = (0,0) \mid \theta = 1)\mathbb{P}(\theta = 1)}{\mathbb{P}(\mathbf{x} = (0,0) \mid \theta = 1)\mathbb{P}(\theta = 1) + \mathbb{P}(\mathbf{x} = (0,0) \mid \theta = 0)\mathbb{P}(\theta = 0)} \\ &= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = \frac{0.125}{0.625} = 0.20.\end{aligned}$$

Intuitivno, ako su sinovi zdravi, manje je vjerojatno da im je majka nositeljica gena odgovornog za bolest. Gornji račun upravo nam pokazuje kako se nove informacije (zdravstvena stanja sinova) koriste u bayesovskom pristupu, kako bi se izmijenila apriorna distribucija obilježja (stanje majke) i kako bi se došlo do aposteriorne distribucije koja bolje odražava sveukupno znanje o obilježju. Dakle, nakon što smo uočili da niti jedan od sinova nema hemofiliju, vjerojatnost da je majka nositeljica smanjila se s 0.5 (apriorna vjerojatnost) na 0.2 (aposteriorna vjerojatnost).

Još je jedna prednost bayesovske statistike da se novi podaci mogu jednostavno uključiti u analizu, bez potrebe za ponavljanjem čitavog postupka. Ključno je jedino da se aposteriorna distribucija iz prethodnog koraka, u novom koraku koristi kao apriorna. Da bismo to ilustrirali, pretpostavimo da je žena dobila i trećeg sina, koji ne boluje od hemofilije, tj. opazili smo i $x_3 = 0$*. Sada je aposteriorna vjerojatnost da je žena nositelj gena:*

$$\mathbb{P}(\theta = 1 \mid (x_1, x_2, x_3) = (0,0,0)) = \frac{(0.125)(0.2)}{(0.125)(0.2) + (1.0)(0.8)} = 0.111.$$

Da je treći sin bolovao od hemofilije, račun bi izgledao:

$$\mathbb{P}(\theta = 1 \mid (x_1, x_2, x_3) = (0,0,0)) = \frac{(0.5 \times 0.5 \times 0.5)(0.2)}{(0.125)(0.2) + (0)(0.8)} = 1,$$

jer je tada $\mathbb{P}(x_1 = 0, x_2 = 0, x_3 = 1 \mid \theta = 0) = 0$ *(nije moguće da treći sin oboli ako majka nije nositeljica gena).*

Poglavlje završavamo primjerom u kojem koristimo većinu do sada uvedenih pojmova, te sljedeću vjerojatnosnu distribuciju.

Definicija 2.3.5. *Kažemo da neprekidna slučajna varijabla* X *ima* ***beta razdiobu*** *s parametrima* α *i* β *i pišemo* $X \sim Beta(\alpha, \beta)$*, ako joj je funkcija gustoće oblika*

$$f_X(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}\chi_{[0,1]}(y). \tag{2.7}$$

Pri tome je normalizirajuća konstanta $B(\alpha,\beta)$ vrijednost koju beta funkcija poprima u (α,β), gdje je beta funkcija definirana s:

$$\mathrm{B}(z_1,z_2) = \int_0^1 t^{z_1-1}(1-t)^{z_2-1}dt,$$

za sve kompleksne brojeve z_1, z_2 čiji su realni dijelovi strogo pozitivni.

Jedno je od glavnih svojstava beta funkcije da je usko vezana s gama funkcijom preko relacije:

$$B(z_1,z_2) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}. \tag{2.8}$$

Gama funkcija definirana je za sve imaginarne brojeve z čiji je realni dio strogo pozitivan, izrazom:

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt.$$

Parcijalnom integracijom slijedi:

$$\Gamma(z+1) = -t^z e^{-t}\Big|_0^\infty + z\int_0^\infty t^{z-1}e^{-t}dt = z\Gamma(z). \tag{2.9}$$

Sada možemo izračunati i očekivanje slučajne varijable $X \sim Beta(\alpha,\beta)$:

$$\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)dx \\
&= \int_0^1 x\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx \\
&= \frac{1}{B(\alpha,\beta)}\int_0^1 x^{(\alpha+1)-1}(1-x)^{\beta-1}dx \\
&= \frac{1}{B(\alpha,\beta)}B(\alpha+1,\beta) \quad \text{(definicija beta funkcije)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \quad \text{(relacija 2.8)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)}\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta)\cdot(\alpha+\beta)}\frac{\Gamma(\alpha)\cdot\alpha}{\Gamma(\alpha)} \quad \text{(svojstvo 2.9)} \\
&= \frac{\alpha}{\alpha+\beta}.
\end{aligned} \tag{2.10}$$

Primjer 2.3.6. *U ovom primjeru želimo procijeniti koliki udio u populaciji rođene djece čine ona ženskog spola. Taj ćemo udio označavati s θ.*

Sličan je problem rješavao i francuski matematičar Pierre-Simon Laplace, koji je na temelju podataka o djeci rođenoj između 1745 i 1770 u Parizu, od kojih su 241,945 bile djevočice i 251,527 dječaci, za θ pokazao da:

$$\Pr(\theta \geq 0.5 \mid y = 241,945, n = 251,527 + 241,945) \approx 1.15 \times 10^{-42},$$

te je ustvrdio da je "moralno siguran" da je udio $\theta < 0.5$.

Pokušajmo i mi detaljno razraditi ovaj problem.

Neka je u n rođenja x djece bilo ženskog spola. Možemo pretpostaviti da su opaženi podaci, tj. spolovi rođene djece, uvjetno nezavisni uz dano θ. Stoga je opravdano modelirati broj rođene ženske djece kao binomnu varijablu. Podsjetimo se, binomna slučajna varijabla s parametrima n i θ interpretira se kao model za broj uspjeha u n ponavljanja nezavisnih slučajnih pokusa, gdje je s θ označena vjerojatnost uspjeha jednoga pokusa (ovdje uspjehom možemo smatrati rođenje ženskog djeteta, a slučajnim pokusom jedan porod). Njena je funkcija gustoće dana s:

$$f(x \mid \theta) = Bin(x \mid n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \chi_{\mathbb{N}_0}(x). \tag{2.11}$$

Za početak analize, trebamo odabrati aprirornu distribuciju za θ. Zbog jednostavnosti, neka to za sada bude uniformna distribucija, čija je gustoća dana s $f(\theta) = \chi_{[0,1]}(\theta)$.

Izračunajmo zatim marginalnu gustoću uzorka:

$$\begin{aligned} f(x) &= \int_{-\infty}^{\infty} f(x \mid \theta) \cdot f(\theta) d\theta \\ &= \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{(n-x)} \cdot 1 d\theta \\ &= \binom{n}{x} \int_0^1 \theta^x (1-\theta)^{(n-x)} d\theta \\ &= \binom{n}{x} B(x+1, n-x+1). \end{aligned} \tag{2.12}$$

Primjenom Bayesove formule 2.1 slijedi:

$$\begin{aligned} f(\theta \mid x) &= \frac{f(\theta) f(x \mid \theta)}{f(x)} \\ &= \frac{1 \cdot \binom{n}{x} \theta^x (1-\theta)^{n-x}}{\binom{n}{x} B(x+1, n-x+1)} \\ &= \frac{\theta^x (1-\theta)^{n-x}}{B(x+1, n-x+1)}, \end{aligned}$$

što je upravo funkcija gustoće beta razdiobe s parametrima $x+1$ i $n-x+1$. Dakle, pronašli smo aposteriornu distribuciju za θ.

Treba napomenuti kako je u gornjim izračunima, u izrazu za $f(x \mid \theta)$ danu s 2.11, zbog jednostavnosti izostavljen član $\chi_{\mathbb{N}_0}(x)$. Zbog prirode problema, jasno je da su opaženi n i x nenegativni (u ovom primjeru, možemo pretpostaviti i pozitivni) cijeli brojevi, čime se pojednostavljuje i izraz za beta funkciju.

Ako iskoristimo relaciju 2.8 i činjenicu da je gama funkcija za pozitivni cijeli broj k definirana kao $\Gamma(k) = (k-1)!$, možemo doći do zatvorene forme za marginalnu vjerodostojnost:

$$\begin{aligned} f(x) &= \binom{n}{x} B(x+1, n-x+1) \\ &= \binom{n}{x} \frac{x!(n-x)!}{(x+1+n-x+1-1)!} \\ &= \binom{n}{x} \frac{x!(n-x)!}{(n+1)!} \\ &= \frac{n!}{x!(n-x)!} \cdot \frac{x!(n-x)!}{(n+1)!} \\ &= \frac{1}{n+1}, \end{aligned}$$

i za samu aposteriornu funkciju gustoće:

$$\begin{aligned} f(\theta \mid x) &= \frac{\theta^x (1-\theta)^{n-x}}{B(x+1, n-x+1)} \\ &= \theta^x (1-\theta)^{n-x} \frac{(n+1)!}{x!(n-x)!} \\ &= \theta^x (1-\theta)^{n-x} (n+1) \frac{n!}{x!(n-x)!} \\ &= \theta^x (1-\theta)^{n-x} (n+1) \binom{n}{x}. \end{aligned}$$

Rezultat $f(x) = \frac{1}{n+1}$ govori nam: ako smo uzeli uniformnu distribuciju kao apriornu za θ, svaka od mogućih realizacija $x \in \{0, 1, 2, \ldots, n\}$ binomne varijable $Bin(n, \theta)$ jednako je vjerojatna.

Sumirajmo što smo do sada napravili. Zanimala nas aposteriorna distribucija parametra θ binomne slučajne varijable $Bin(n, \theta)$ koja na realizaciji uzorka poprima vrijednost x. Za apriornu distribuciju od θ, izabrali smo uniformu distribuciju na segmentu [0,1] te smo primjenom Bayesove formule zaključili da je aposteriorna distribucija Beta(x+1, n-x+1).

Dakle, u ovom smo primjeru uspjeli dovesti do zatvorene forme sve faktore u Bayesovoj formuli za izračun aposteriorne distribucije parametra. Često, zbog složenosti računa, posebno kod računanja sume/ integrala kojim je zadana marginalna vjerodostojnost, nismo u mogućnosti to napraviti. Tada se pozivamo na 2.4 i, izostavljajući izračun f(x) te ostale konstante, pronalazimo nenormaliziranu aposteriornu gustoću.

Na primjer, za postavljeni problem možemo uočiti: $f(\theta \mid x) \propto \theta^x(1-\theta)^{n-x}$, te ponovo prepoznati da je riječ o beta distribuciji.

Vratimo se na problem izbora apriorne distribucije. S obzirom da za vjerodostojnost vrijedi: $f(x \mid \theta) \propto \theta^x(1-\theta)^{n-x}$, ako za apriornu distribuciju izaberemo distribuciju $Beta(\alpha, \beta)$, vrijedi:

$$\begin{aligned} f(\theta \mid x) &\propto \theta^x(1-\theta)^{n-x} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}, \end{aligned}$$

tj. aposteriorna distribucija je ponovo Beta($x + \alpha$,$n - x + \beta$). Dakle, pokazali smo da je klasa beta distribucija konjugirana za klasu binomnih distribucija.

Naravno, pri odabiru beta distribucije treba fiksirati i njene hiperparametre, α i β. Ovdje, pod pojmom hiperparametri podrazumijevamo sve parametre apriorne distribucije. Njih postavljaju istraživači i o njima se ne donose vjerojatnosni zaključci. Na samom kraju primjera ilustrirano je kako njihov izbor utječe na rezultate statističke analize.

Korisno se još prisjetiti kako je i na početku izabrana uniformna razdioba zapravo beta razdioba s parametrima $\alpha = 1, \beta = 1$.

Dotaknimo se još i vjerojatnosnih tvrdnji o budućim opažanjima. Pretpostavimo da, nakon opaženih n rođenja od kojih je rođeno x ženske djece i uz apriornu $Beta(\alpha, \beta)$ razdiobu, želimo izračunati aposteriornu vjerojatnost da je sljedeće rođeno dijete djevojčica:

$$\begin{aligned} \mathbb{P}(X_{n+1} = 1 \mid x) &= \int_0^1 \mathbb{P}(X_{n+1} = 1 \mid x, \theta) f(\theta \mid x) d\theta \\ &= \int_0^1 \theta f(\theta \mid x) d\theta \\ &= \mathbb{E}(\theta \mid x) = \frac{x + \alpha}{n + \alpha + \beta}, \end{aligned}$$

pri čemu smo u zadnjem koraku koristili izraz za očekivanje beta distribucije 2.10 i činjenicu da je aposteriorna distribucija parametra θ upravo Beta($x + \alpha$,$n - x + \beta$).

Za kraj, ostaje interpretirati dobivene rezultate.

Ako smo za parametar θ, koji modelira udio ženske djece u populaciji novorođenčadi, izabrali Beta(α, β) kao apriornu distribuciju, pokazali smo da je apriorno očekivanje $\frac{\alpha}{\alpha+\beta}$. Ako raspolažemo podacima iz nekog prijašnjeg istraživanja sličnog našem (na primjer, podacima koje je prikupio Laplace), te znamo da je od n_p rođene djece njih x_p bilo ženskog spola, za hiperparametar α možemo izabrati upravo x_p, a za hiperparametar β možemo izabrati $n_p - x_p$ (što predstavlja broj rođene djece muškog spola). Tada bi apriorno očekivanje za parametar θ iznosilo "prijašnji" udio $\frac{x_p}{n_p}$, a distribucija Beta$(x_p, n_p - x_p)$ bila bi informativna. U drugu ruku, ako želimo što manje apriornom utjecati na aposteriornu distribuciju, izabrat ćemo uniformnu, odnosno Beta$(1, 1)$ distribuciju, čije je apriorno očekivanje $\frac{1}{2}$. Uniformnu distribuciju onda možemo smatrati neinformativnom.

Promatrajući aposteriorno očekivanje za θ (koje je jednako vjerojatnosti da se u idućem rođenju radi o djevojčici), vidimo da iznosi $\frac{x+\alpha}{n+\alpha+\beta}$ i predstavlja kompromis između apriornog očekivanja $\frac{\alpha}{\alpha+\beta}$ i opaženog udjela u promatranoj populaciji $\frac{x}{n}$. Što je veći uzorak, opaženi će podaci više utjecati na aposteriorno očekivanje. Također, što veće hiperparametre izaberemo, apriorna informacija imat će veći utjecaj.

Ovdje smo ilustrirali kako konjugirane distribucije olakšavaju interpretaciju i na koje načine (i s kakvim ciljem) istraživači mogu birati hiperparametre.

Za dodatnu razradu bayesovskog pristupa te primjere, preporučuje se pogledati ovdje korištenu literaturu: [3], [7], [1].

Poglavlje 3

Hijerarhijsko modeliranje

U ovom se poglavlju uvodi i obrađuje pojam hijerarhijskog modeliranja u statističkoj analizi podataka. Kroz primjere, pojam se približava i razrađuje, te se postupno pojavljuju pitanja i problemi koji vode prema punom bayesovskom hijerarhijskom modelu.

3.1 Motivacija

Do sada smo se vodili pretpostavkom da su opaženi podaci, koje statističkim modelima matematički nastojimo opisati, nestrukturirani. Ipak, sa stvarnim podacima čest je slučaj da postoji grupiranje podataka i da svaka od tih grupa ima vlastita svojstva i posebnosti.

Primjerice, zanima nas vjerojatnost preživljavanja srčanog udara p i podatke smo prikupljali u N različitih bolnica. Možemo očekivati određene razlike u stopi preživljavanja srčanog udara između bolnica (primjerice, utjecaj bi mogla imati veličina ili opremljenost bolnice), što bi nas moglo navesti da za svaku od bolnica posebno promatramo vjerojatnost preživljavanja p_j, $j \in \{1, 2, \ldots, N\}$. Istovremeno, prirodno je pretpostaviti da se uočeni podaci na različitim lokacijama neće drastično razlikovati, jer ipak se radi o istoj pojavi, pa bismo mogli zanemariti činjenicu da pacijenti pripadaju različitim bolnicama i na temelju prikupljenih podataka izvesti statističke zaključke o općenitoj vjerojatnosti preživljavanja srčanog udara p. Ipak, najbolje bi bilo izabrati pristup koji uzima u obzir i sličnost grupa (bolnica) i različitosti među njima.

Posebno je zanimljiva situacija kada je dostupan vrlo mali uzorak za jednu ili više promatranih grupa. Ako smatramo grupe samostalnima i nepovezanima, tada je teško donijeti relevantne statističke zaključke za obilježje grupe za koju je prikupljeno svega nekoliko opažanja. Ako nam priroda problema sugerira da postoji veza između grupa, htjeli bismo na neki način iskoristiti i prikupljene podatke za ostale grupe, da izvedemo zaključak o onoj za koju nam nedostaje više podataka.

Dakle, kada se nalazimo u situaciji da su nam podaci, zbog prirode problema, strukturirani u više grupa, imamo tri pristupa za njihovo modeliranje. Prvi pristup je pristup **bez udruživanja** podataka (eng. *no-pooling model*), koji podrazumijeva da svaku grupu podataka promatramo i analiziramo posebno. Kao što smo već naglasili, takvo modeliranje često nije poželjno, jer ne koristi informacije o drugim grupama podataka da korigira ili poboljša zaključke za promatranu grupu. Drugi je pristup korištenje modela s **potpunim udruživanjem** (eng. *complete-pooling model*), gdje sve podatke koristimo za donošenje općenitih zaključaka o statističkom obilježju. Nedostatak tog pristupa jest što se u obzir ne uzimaju razlike između grupa. Treći je pristup **hijerarhijsko modeliranje**, kojim istovremeno uzimamo u obzir i razlike i sličnosti između grupa podataka. Hijerarhijski modeli također imaju nedostataka, ali oni nisu konceptualne prirode, već se uglavnom radi o složenosti računa i radu s većim brojem parametara.

3.2 Pristup

Promatrano s matematičke strane, statistički modeli za stvarne probleme i procese često podrazumijevaju upotrebu više parametara koji su, radi teorijske pozadine problema, u nekom odnosu. Stoga se javlja potreba za uvođenjem statističkog modela za parametre koji bi taj odnos odražavao.

Ako se vratimo na primjer vjerojatnosti preživljavanja srčanih udara, mogli bismo smatrati da smo izabrali N jedinki iz populacije bolnica i da su stope preživljavanja u tim bolnicama p_j (koje ne opažamo izravno) zapravo uzorak iz zajedničke distribucije stopa preživljavanja po bolnicama.

Ovakav nam je problem stoga prirodno modelirati hijerarhijski, gdje na jednoj razini imamo model za opažanja uvjetno na parametre $\theta_1, \theta_2, \ldots, \theta_n$, $f(\mathbf{x} \mid \theta_1, \theta_2, \ldots, \theta_n)$, a na drugoj razini i sami parametri $\theta_1, \theta_2, \ldots, \theta_n$ vjerojatnosno su opisani s $f(\theta_1, \theta_2, \ldots, \theta_n \mid \phi)$ preko drugih parametara $\phi = (\phi_1, \phi_2, \ldots, \phi_k)$, koje ovdje nazivamo hiperparametrima.

U sljedećem poglavlju pokazat ćemo kako se odnos promatranih parametara $\theta_1, \ldots, \theta_n$, modeliran uvođenjem zajedničke distribucije nad njima, može dodatno matematički opisati primjenom bayesovskog pristupa za parametre ϕ (za koje ćemo onda promatrati apriornu i aposteriornu distribuciju). Prednost je takvog pristupa što se opaženi podaci $x_{i,j}$, indeksirani s i unutar grupa koje su indeksirane po j, mogu koristiti za donošenje vjerojatnosnih zaključaka o populacijskoj distribuciji parametara $(\theta_1, \theta_2, \ldots, \theta_n)$, iako te parametre izravno ne opažamo (što je od posebne važnosti za situaciju s malim uzorcima za neke od grupa).

Kako bismo približili ideju hijerarhijskog modeliranja, navodimo sljedeći primjer, koji je preuzet iz [3]. U ovom je primjeru cilj procijeniti parametar, na temelju upravo provedenog eksperimenta i na temelju apriorne distribucije konstruirane uz pomoć podataka iz

prethodnih eksperimenata. Pri tome je ključna ideja da eksperimenti (odnosno, parametri koji opisuju svojstvo grupa obuhvaćenih eksperimentima) dolaze iz zajedničke populacije.

Primjer 3.2.1. *U laboratorijskim istraživanjima i studijama za testiranje potencijalnih lijekova prije njihovog puštanja u upotrebu, najčešće se koriste glodavci. U ovom primjeru zanima nas vjerojatnost pojave tumora u populaciji ženskih laboratorijskih štakora tipa F344 koji nisu primili lijek, odnosno čine kontrolnu skupinu u istraživanju.*

Pretpostavimo da smo u upravo provedenom eksperimentu, koji je bio nešto manjih razmjera nego što je uobičajeno, uočili da su 4 od 14 promatranih ženskih štakora razvili tumorske formacije na unutrašnjem sloju maternice. Osim ovih, na raspolaganju imamo i podatke iz 70 prijašnjih izvođenja istog eksperimenta, gdje su zabilježeni udjeli ženskih štakora kod kojih je uočena ta vrsta tumora. Svi su podaci navedeni ispod, u obliku:
x_i/n_i *= (broj štakora s tumorom) / (broj štakora koji je sudjelovao u eksperimentu).*

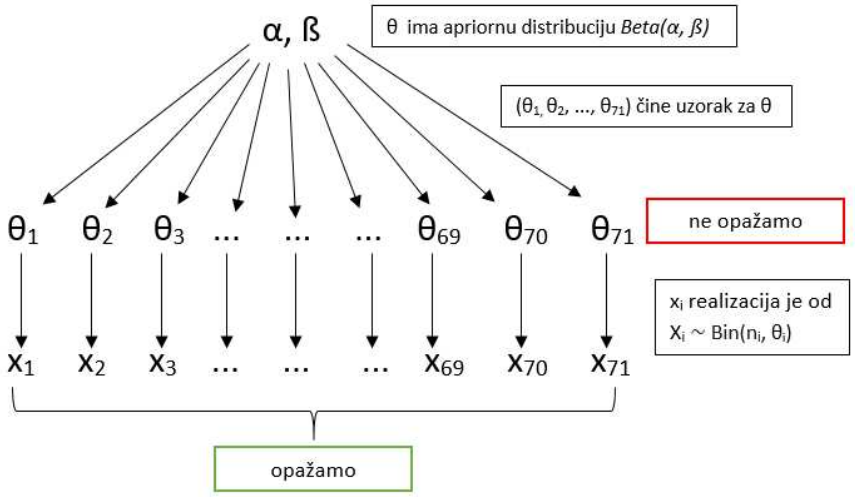
Prethodni eksperimenti:

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Upravo provedeni eksperiment: 4/14

Broj oboljelih štakora u i-tom eksperimentu modeliramo binomnom slučajnom varijablom $X_i \sim Bin(n_i, \theta_i)$. *Primarni nam je cilj donijeti statističke zaključke o vjerojatnosti razvoja tumora u populaciji ženskih štakora tipa F344 na temelju upravo provedenog pokusa, ovdje označenu s* θ_{71}, *ali želimo iskoristiti i znanje sadržano u rezultatima prethodnih pokusa.*

Prvi je korak odabrati populacijsku distribuciju parametra θ, *za kojeg parametri eksperimenata* $\theta_1, \theta_2, \ldots, \theta_{71}$ *predstavljaju slučajni uzorak. Radi jednostavnosti računanja s konjugiranim distribucijama i jer smo taj slučaj već detaljno razradili, biramo distribuciju* $Beta(\alpha, \beta)$. *Time smo uveli dva nova parametra,* α *i* β, *koje ovdje procjenjujemo točkovno, tj. za sada im ne dodjeljujemo vlastite distribucije.*



Slika 3.1: Struktura hijerarhijskog modela za primjer s laboratorijskim štakorima

Promotrimo dva načina na koja možemo doći do točkovne procjene za parametre beta distribucije.

Prvi slučaj: fiksirana apriorna distribucija za parametar θ

Pretpostavimo da na temelju do sada razvijene znanstvene teorije znamo da vjerojatnosti pojave tumora kod grupa ženskih laboratorijskih štakora tipa F344 (aproksimativno) slijede beta distribuciju s nekim poznatim očekivanjem $\tilde{\mu}$ *i standardnom devijacijom* $\tilde{\sigma}$*. S obzirom na to da za očekivanje i varijancu beta distribucije općenito vrijedi:*

$$\mu = \frac{\alpha}{\alpha + \beta},$$
$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

rješavanjem za α *i* β *dobivamo:*

$$\alpha_f = \left(\frac{1-\tilde{\mu}}{\tilde{\sigma}^2} - \frac{1}{\tilde{\mu}}\right)\tilde{\mu}^2$$
$$\beta_f = \alpha_f\left(\frac{1}{\tilde{\mu}} - 1\right)$$

U primjeru 2.3.6. pokazali smo da je klasa beta distribucija konjugirana za klasu binomnih distribucija te je uz apriornu $Beta(\alpha,\beta)$ *distribuciju za* θ *i opaženu realizaciju* x *slučajne varijable* $X \sim Bin(n,\theta)$*, aposteriorna distribucija od* θ *zapravo* $Beta(x+\alpha, n-x+\beta)$*.*

Primjenjujući to na ovaj slučaj, zaključujemo da θ_{71} ima aposteriornu Beta$(4+\alpha_f, 10+\beta_f)$ distribuciju.

Drugi slučaj: procjena parametara α i β na temelju prethodnih eksperimenata

Situacija obrađena u prvom slučaju, gdje već znamo očekivanje i standardnu devijaciju za distribuciju statističkog obilježja, vrlo je rijetka. Puno češće na raspolaganju imamo podatke iz prijašnjih eksperimenata na sličnim grupama subjekata.

U ovom primjeru s tumorima kod laboratorijskih štakora, na raspolaganju imamo rezultat upravo provedenog eksperimenta i ishode 70 prethodnih pokusa. Zapravo, raspolažemo s opažanjima $X_i = x_i$ gdje je $X_i \sim Bin(n_i, \theta_i)$ za $i = 1, 2, \ldots, 71$.

S obzirom na to da izvodimo statistički zaključak o θ_{71}, promotrimo što o parametrima α i β apriorne distribucije od θ možemo reći na temelju prošlih pokusa. Za 70 vrijednosti udjela x_i/n_i štakora koji su razvili tumor, uzoračka sredina iznosi 0.136, a uzoračka devijacija 0.103. Sada možemo ove vrijednosti iskoristiti kao $\tilde{\mu}$ i $\tilde{\sigma}$ da, na način opisan u prvom slučaju, dođemo do $\alpha_f = 1.4$ i $\beta_f = 8.6$. Aposteriorna distribucija od θ_{71} stoga je Beta$(5.4, 18.6)$, pa je aposteriorno očekivanje 0.223, a standardna devijacija 0.083.

Kada bismo, kao u frekvencionističkom pristupu, išli na temelju upravo provedenog pokusa točkovno procijeniti θ_{71}, to bismo učinili opaženim udjelom $4/14 = 0.286$.

Odmah vidimo da je aposteriorno očekivanje manje od opaženog udjela, jer nam informacije iz prethodnih pokusa (zakodirane u parametre beta distribucije) ukazuju na to da je uočeni udio neuobičajeno velik.

Ovdje je bitno naglasiti kako smo podatke iz prošlih istraživanja iskoristili samo za dobivanje početne ideje o parametrima populacijske distribucije od θ.

Sada bismo možda htjeli iskoristiti sličan pristup da dođemo do statističkih zaključaka i za ostale parametre $\theta_1, \ldots, \theta_{70}$.

Kada bismo to išli napraviti uz apriornu distribuciju Beta$(1.4, 8.6)$, koju smo koristili za θ_{71}, informacije iz prethodnih 70 pokusa upotrijebili bismo dva puta: prvo bismo iskoristili sve te podatke za procjenu parametara apriorne distribucije, a zatim bismo iskoristili podatke iz i-tog eksperimenta za izvođenje zaključaka o θ_i. Već naslućujemo da bi to dovelo u pitanje kvalitetu naših zaključaka.

Također, točkovnom procjenom parametara α i β matematički ne možemo modelirati nesigurnost u njihovom izboru, kao što bismo to mogli kada bismo uveli distribucije i za te parametre.

U ovom se primjeru hijerarhijsko modeliranje zapravo sastojalo u tome da za parametre $\theta_1, \theta_2, \ldots, \theta_{71}$ postavimo zajedničku distribuciju Beta(α, β). U sljedećem poglavlju, ovaj ćemo primjer dodatno razraditi uvođenjem apriorne i računanjem aposteriorne distribucije za α i β. To će nam omogućiti da provedemo željene analize i za ostale parametre $\theta_1, \theta_2, \ldots, \theta_{70}$, otklanjajući probleme ovdje iznesenog pristupa.

Poglavlje završavamo kratkim osvrtom na već uvedeni pojam izmjenjivosti (definicija 2.3.1.), koji u okviru hijerarhijskog modeliranja posebno dobiva na važnosti.

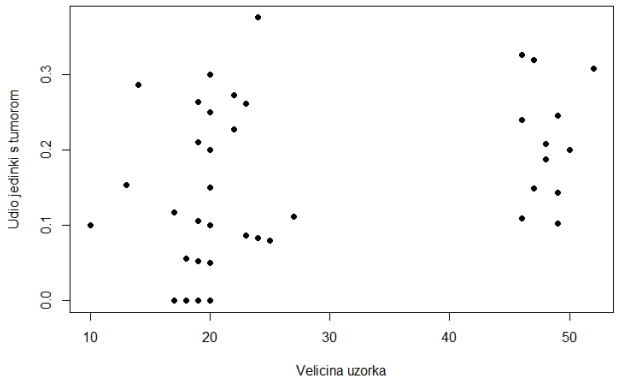
Promotrimo ponovo niz od J eksperimenata (pojam eksperimenta može se generalizirati na grupe/klastere opažanja, primjerice bolnice, škole, itd.). Neka je distribucija statističkog obilježja za eksperiment j parametrizirana s θ_j i neka je uočen vektor podataka $\mathbf{x}_j = (x_{1,j}, x_{2,j}, \ldots, x_{n_j,j})$. Ako osim podataka $\mathbf{x}_j$, nemamo dodatnih informacija zbog kojih bismo neke θ_j razlikovali od drugih ili po kojima bismo mogli grupirati i same parametre, možemo pretpostaviti "simetriju" među parametrima θ_j. Ta se "simetrija" u matematičkom smislu definira kao izmjenjivost $\theta_1, \theta_2, \ldots, \theta_J$, odnosno tada vrijedi:

$$f(\theta_1, \theta_2, \ldots, \theta_J) = f(\theta_{\pi(1)}, \theta_{\pi(2)}, \ldots, \theta_{\pi(J)}),$$

za svaku permutaciju π skupa $\{1, 2, \ldots, J\}$.

U primjeru s vjerojatnosti preživljavanja srčanog udara, ako smo podatke prikupljali u bolnicama za koje nemamo kriterije po kojima bismo za jednu rekli da je značajno bolja od druge, prirodno je pretpostaviti izmjenjivost parametara p_j. Ako pak znamo da je jedna bolnica centar za liječenje kardiovaskularnih bolesti, mogli bismo pretpostaviti da će ta bolnica imati veću stopu preživljavanja, pa ne bismo pretpostavili izmjenjivost p_j.

Za primjer s tumorima kod laboratorijskih štakora, jedina informacija na temelju koje možemo donekle razlikovati eksperimente jest veličina uzorka n_j. Na prvu nam se vjerojatno čini da n_j ne utječe na θ_j. Ako nismo sigurni postoji li poveznica i želimo to provjeriti, mogli bismo na primjer promotriti graf vrijednosti $(n_j, x_j/n_j)$ (Slika 3.2.). Kako ni iz grafa ne uočavamo vezu, nemamo informacija na temelju kojih bismo pretpostavili da će za neki eksperiment stopa jedinki s tumorom biti veća ili manja od drugih. Stoga je tu prirodno pretpostaviti izmjenjivost promatranih parametara.



Slika 3.2: Graf vrijednosti $(n_j, x_j/n_j)$

Najjednostavniji slučaj gdje su $\theta_1, \theta_2, \ldots, \theta_J$ izmjenjivi jest slučaj gdje ih promatramo kao slučajan uzorak iz neke distribucije opisane vektorom parametara $\phi = (\phi_1, \ldots, \phi_k)$. Tada vrijedi:

$$f(\theta_1, \theta_2, \ldots, \theta_J \mid \phi) = \prod_{j=1}^{J} f\left(\theta_j \mid \phi\right). \tag{3.1}$$

Upravo smo ovaj pristup izabrali u Primjeru 3.2.1. , gdje smo vjerojatnosti razvoja tumora u grupama laboratorijskih štakora odabranima za eksperimente $j = 1, 2, \ldots, 71$ interpretirali kao slučajni uzorak za općenitu vjerojatnost razvoja tumora kod te vrste štakora.

Ipak, ne možemo uvijek izabrati taj najjednostavniji pristup. Uzmimo za primjer bacanje kocke i promatrajmo vjerojatnosti da kocka padne na svaki od šest brojeva, p_i, $i = 1, 2, \ldots, 6$. Parametri p_i su izmjenjivi, ali s obzirom na to da vrijedi $\sum_{i=1}^{6} p_i = 1$, ne možemo ih modelirati kao slučajni uzorak iz neke populacije (jer nisu nezavisni).

Ovim napomenama završavamo pregled hijerarhijskog modeliranja.

Poglavlje 4

Bayesovsko hijerarhijsko modeliranje

Nastavljajući na prethodno poglavlje, kao rješenje za nedostatke koji proizlaze iz točkovne procjene hiperparametara zajedničke distribucije za parametre θ, uvodimo pojam bayesovskog hijerarhijskog modeliranja. Zadavanje apriorne i računanje aposteriorne distribucije hiperparametara modela omogućit će nam da kod donošenja zaključaka za jednu grupu iskoristimo i podatke za preostale grupe. Ovaj će nam pristup također omogućiti da matematički opišemo nesigurnost u izboru hiperparametara. U ovom su dijelu rada iznesene glavne teorijske postavke bayesovskih hijerarhijskih modela, opisan je postupak kojim se oni zadaju i način na koji se na temelju njih donose zaključci. Primjerima kojima završava poglavlje ilustriran je sam postupak i prednosti ovog pristupa. Poglavlje prati literaturu [3].

4.1 Teorijska pozadina

Za početak, ponovimo ukratko glavne ideje iza hijerarhijskog modeliranja. Promatramo neko statističko obilježje X za koje smo prikupili podatke koji su strukturirani u J grupa. Označimo ponovo s $\mathbf{x}_j = (x_{1,j}, x_{2,j}, \ldots, x_{n_j,j})$ vektor opažanja za j-tu grupu. Za svaku od grupa, definiramo prvo funkciju gustoće distribucije za koju $\mathbf{x}_j$ predstavlja slučajni uzorak, $f(x \mid \theta_j)$, pri čemu je ta gustoća parametrizirana s θ_j (koji može biti i višedimenzionalan). Ako pretpostavljamo da postoje sličnosti između grupa koje želimo iskoristiti, i parametrima $\theta_1, \theta_2, \ldots, \theta_J$ pridružujemo vjerojatnosnu distribuciju danu s funkcijom gustoće $f(\theta_1, \theta_2, \ldots, \theta_J \mid \phi)$, koja je parametrizirana s $\phi = (\phi_1, \phi_2, \ldots, \phi_k)$. Na temelju prirode problema, odlučujemo možemo li za $\theta_1, \theta_2, \ldots, \theta_J$ pretpostaviti da su izmjenjivi, tj. da je $f(\theta_1, \theta_2, \ldots, \theta_J \mid \phi)$ invarijantna na permutacije indeksa $(1, 2, \ldots, J)$. Kada bismo znali za dodatna obilježja grupa koja utječu na statističko obilježje, ne bismo mogli pretpostaviti izmjenjivost i tada je uobičajeno ta obilježja modelirati kovarijatama $\mathbf{z}_J$ i promatrati

$$f(\theta_1, \ldots, \theta_J \mid \mathbf{z}_1, \ldots, \mathbf{z}_J) = \int f(\theta_1, \ldots, \theta_J \mid \phi, \mathbf{z}_1, \ldots, \mathbf{z}_J)] f(\phi \mid \mathbf{z}_1, \ldots, \mathbf{z}_J) d\phi.$$

U ostatku ovog rada, radi jednostavnosti ćemo uzimati da je pretpostavka izmjenjivosti zadovoljena, odnosno da ne raspolažemo dodatnim informacijama na temelju kojih bismo grupe mogli razlikovati. Štoviše, koristimo i strožu pretpostavku od izmjenjivosti: da $\theta_1, \theta_2, \ldots, \theta_J$ čine slučajan uzorak iz neke zajedničke distribucije $f(\theta \mid \phi)$.

Sada u fokus stavljamo vektor parametara $\phi = (\phi_1, \phi_2, \ldots, \phi_k)$, koji nam je gotovo uvijek nepoznat. U prošlom smo poglavlju, u primjeru 3.2.1., tome doskočili točkovnom procjenom na temelju prethodno sakupljenih podataka, ali vrlo smo brzo uočili nedostatke takvog pristupa.

Puni bayesovski hijerarhijski model podrazumijeva da i za nepoznate parametre $\phi = (\phi_1, \ldots, \phi_k)$ zadamo apriornu funkciju gustoće $f(\phi)$. Sada, umjesto apriorne funkcije gustoće za $\theta = (\theta_1, \theta_2, \ldots, \theta_J)$, zadajemo uvjetnu apriornu funkciju gustoće $f(\theta \mid \phi)$. Za izvod aposteriornih gustoća, promatramo zajedničku apriornu funkciju gustoće za ϕ i θ:

$$f(\phi, \theta) = f(\phi) f(\theta \mid \phi). \tag{4.1}$$

Slijedeći Bayesovu formulu za funkcije gustoća, dolazimo do zajedničke aposteriorne gustoće:

$$\begin{aligned} f(\phi, \theta \mid \mathbf{x}) &= \frac{f(\phi, \theta, \mathbf{x})}{f(\mathbf{x})} \\ &= \frac{f(\mathbf{x} \mid \phi, \theta) f(\phi, \theta)}{f(\mathbf{x})} \\ &\propto f(\mathbf{x} \mid \phi, \theta) f(\phi, \theta) \quad = f(\mathbf{x} \mid \theta) f(\phi, \theta). \end{aligned} \tag{4.2}$$

Pri tome, zadnja jednakost vrijedi jer podaci $\mathbf{x}$ ovise o parametrima ϕ samo preko θ (preciznije, $\mathbf{X}$ i ϕ uvjetno su nezavisni uz dano θ). Dakle, vidimo da je, kao i kod nehijerarhijskih modela, dovoljno još zadati i vjerodostojnost $f(\mathbf{x} \mid \theta)$.

Kod bayesovskog hijerarhijskog modeliranja (ali i općenito bayesovskog pristupa), često izrazi za aposteriorne funkcije gustoće ne zadaju neku poznatu vjerojatnosnu distribuciju (kao što su normalna, binomna ili beta). Ako želimo predočiti funkcije gustoće tih distribucija ili iz njih želimo uzorkovati, potrebno je posegnuti za alatima numeričke matematike i računarske statistike. Ovdje ćemo primijeniti neke od tih alata, a više informacija i primjera dostupno je u [3].

Dakle, kako bismo došli do željenih zaključaka i prikaza, naš će se postupak sastojati i od analitičkih i od numeričkih i računarskih metoda.

4.2 Postupak

Za početak, moramo matematički opisati sve veličine (podatke $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$, parametre θ i parametre ϕ), što zapravo znači dodijeliti svakoj veličini apriornu vjerojatnosnu distribuciju.

Prvo, određujemo apriornu distribuciju za ϕ (koju još nazivamo i hiperapriorna distribucija). Ovisno o pozadini i dosadašnjem razumijevanju problema, možemo izabrati informativnu, slabo informativnu ili neinformativnu hiperapriornu distribuciju, ali bismo u većini slučajeva trebali raspolagati s dovoljno znanja da možemo ograničiti skup mogućih vrijednosti od ϕ i/ili precizno zadati vjerojatnosnu funkciju gustoće. Kao što smo već napomenuli, u bayesovskom se pristupu funkcije gustoće često zadaju kao nenormalizirane ili neprave, no hijerarhijsko modeliranje zahtijeva poseban oprez kod izbora apriorne hiperdistribucije, jer je bitno dobiti prave (vjerojatnosne) aposteriornu funkcije gustoće.

Zatim, zadajemo apriornu distribuciju od θ uvjetno na ϕ s funkcijom gustoće $f(\theta \mid \phi)$. Zajednička apriorna funkcija gustoće, $f(\phi, \theta)$ dana je s 4.1. Posljednje što moramo zadati jest vjerodostojnost $f(\mathbf{x} \mid \theta)$.

Kako smo pretpostavili da su parametri $(\theta_1, \theta_2, \ldots, \theta_J)$ izmjenjivi, i štoviše, da predstavljaju slučajan uzorak iz distribucije zadane funkcijom gustoće $f(\cdot \mid \phi)$, uz oznaku $\theta = (\theta_1, \theta_2, \ldots, \theta_J)$, možemo koristiti $f(\theta \mid \phi) = \prod_{j=1}^{J} f\left(\theta_j \mid \phi\right)$.

Sljedeće korake provodimo analitički:

1. Prema raspisu 4.2, dolazimo do nenormalizirane aposteriorne funkcije gustoće za ϕ i θ, koja je dana je s $f(\phi, \theta \mid \mathbf{x}) \propto f(\mathbf{x} \mid \theta) f(\phi, \theta) = f(\mathbf{x} \mid \theta) f(\theta \mid \phi) f(\phi)$.

2. Kako bismo odredili $f(\theta \mid \phi, \mathbf{x})$, aposteriornu funkciju gustoće od θ uz dano ϕ i opažanja $\mathbf{x}$, trebamo se prisjetiti pristupa obrađenog u poglavlju Bayesovska statistika. U ovom koraku, fiksiramo ϕ i slijedimo uobičajeni postupak bayesovskog zaključivanja kako bismo došli do aposteriorne gustoće $f(\theta \mid \phi, \mathbf{x}) \propto f(\theta \mid \phi) f(\mathbf{x} \mid \theta)$. Jedina je razlika što kod hijerarhijskog modeliranja, umjesto točno određenog broja imamo nepoznati parametar ϕ koji zadaje distribuciju od θ. Ovdje je ponovo korisno, zbog jednostavnosti računa, izabrati konjugirane funkcije gustoće $f(\mathbf{x} \mid \theta)$ i $f(\theta \mid \phi)$. Treba još naglasiti kako svaka od mogućih vrijednosti ϕ zadaje jednu gustoću $f(\theta \mid \phi, \mathbf{x})$ (dok je vektor opažanja $\mathbf{x}$ fiksan). Time smo izveli zaključke za θ.

3. Kako bismo bayesovskim pristupom došli do zaključaka o ϕ, potrebno je doći do aposteriorne marginalne funkcije gustoće $f(\phi \mid \mathbf{x})$. To je moguće napraviti na dva načina.

   Prvi način podrazumijeva uobičajeni izračun marginalne funkcije gustoće jedne varijable, integriranjem zajedičke funkcije gustoće po drugoj varijabli:

   $$f(\phi \mid \mathbf{x}) = \int f(\phi, \theta \mid \mathbf{x}) d\theta. \tag{4.3}$$

   Za neke od modela koji su često u upotrebi (npr. za modele obrađene u primjerima na kraju poglavlja), može se izbjeći računanje integrala i algebarski doći do željenog

izraza, upotrebom formule:

$$f(\phi \mid \mathbf{x}) = \frac{f(\phi, \theta \mid \mathbf{x})}{f(\theta \mid \phi, \mathbf{x})}, \tag{4.4}$$

koja slijedi iz raspisa:

$$f(\phi, \theta \mid \mathbf{x}) = \frac{f(\phi, \theta, \mathbf{x})}{f(\mathbf{x})} = \frac{f(\theta \mid \phi, \mathbf{x}) f(\phi, \mathbf{x})}{f(\mathbf{x})} = f(\theta \mid \phi, \mathbf{x}) f(\phi \mid \mathbf{x}).$$

Ovaj je drugi način koristan, jer smo u prethodna dva koraka već došli do nenormaliziranih izraza za $f(\phi, \theta \mid \mathbf{x})$ i $f(\theta \mid \phi, \mathbf{x})$. Ipak, treba biti oprezan, jer je nazivnik $f(\theta \mid \phi, \mathbf{x})$ funkcija u ovisnosti i o θ i o ϕ (za fiksni vektor opažanja $\mathbf{x}$) i ima normalizirajući faktor koji ovisi i o ϕ i o $\mathbf{x}$ (pa ga ne možemo jednostavno smatrati konstantom, izostaviti i govoriti o proporcionalnoj funkciji gustoće). Napomenimo još da, ako integral 4.3 nema zatvorenu formu, nije moguće iskoristiti ni formulu 4.4.

Sljedeći koraci služe za (računalnu) simulaciju slučajnog uzorka za ϕ, θ, x iz dobivenih aposteriornih distribucija, kod jednostavnijih hijerarhijskih modela obrađenih u ovom poglavlju.

1. Za početak, uzorkuje se vektor hiperparametara $\tilde{\phi}$ iz aposteriorne distribucije s funkcijom gustoće $f(\phi \mid \mathbf{x})$.

2. Zatim se simulira vektor parametara $\tilde{\theta} = (\tilde{\theta_1}, \tilde{\theta_1}, \ldots, \tilde{\theta_J})$ iz uvjetne aposteriorne distribucije uz dano $\phi = \tilde{\phi}$, s funkcijom gustoće $f(\theta \mid \tilde{\phi}, \mathbf{x})$. S obzirom da smo pretpostavili da $(\theta_1, \theta_2, \ldots, \theta_J)$ čine slučajni uzorak, komponente vektora θ_j možemo simulirati nezavisno, jednu po jednu.

3. Ako nam je od interesa, možemo simulirati i novo opažanje $\tilde{x}$. Može se raditi o novom podatku za već promatranu j-tu grupu, kojeg simuliramo iz $f(x \mid \tilde{\theta}_j)$ ili možemo simulirati opažanje koje pripada novoj grupi. U tom slučaju, prvo moramo simulirati novi podatak $\tilde{\theta}_{J+1}$ iz aposteriorne distribucije zadane s $f(\theta \mid \tilde{\phi}, \mathbf{x})$, a zatim možemo simulirati opažanje $\tilde{x}$ iz distribucije s funkcijom gustoće $f(x \mid \tilde{\theta}_{J+1})$.

Ponavljanjem gornjih koraka N puta, dobivamo slučajne uzorke duljine N za ϕ, θ i x. Već na temelju slučajnog uzorka za ϕ, možemo aproksimirati prediktivnu funkciju gustoće za θ, $f(\tilde{\theta} \mid \mathbf{x})$, i aposteriornu funkciju gustoće za θ_j, $f(\theta_j \mid \mathbf{x})$, kao i aproksimirati intervale pouzdanosti za parametre θ_j. Detaljnije objašnjenje tog postupka izneseno je u primjeru koji slijedi.

4.3 Primjena

Primjer 4.3.1. *Vratimo se na primjer s laboratorijskim štakorima. Prisjetimo se, imamo podatke o 71 eksperimentu, te za svaki znamo broj laboratorijskih štakora koji su sudjelovali (n_j) i koliko je od njih razvilo tumorske formacije (x_j). Izabrali smo sljedeći vjerojatnosni model: broj štakora koji su u j-tom eksperimentu razvili tumor modeliramo s $X_j \sim Bin(n_j, \theta_j)$, $j = 1, 2, \ldots 71$, gdje vjerojatnost razvoja tumora u j-tom eksperimentu označavamo s θ_j te $(\theta_1, \theta_2, \ldots, \theta_{71})$ prestavlja slučajni uzorak iz $Beta(\alpha, \beta)$ distribucije.*

Potpuni bayesovski hijerarhijski model podrazumijeva da i parametre α i β opišemo vjerojatnosnim distribucijama. Kako nam je bitno da izraz za aposteriornu gustoću od (α, β) zaista zadaje pravu funkciju gustoće ($\int f(\alpha, \beta \mid \mathbf{x}) d(\alpha, \beta) < \infty$) te ne bismo htjeli izborom apriorne funkcije gustoće narušiti taj zahtjev, odgađamo izbor hiperapriorne gustoće $f(\alpha, \beta)$, dok ne dođemo do izraza za aposteriornu funkciju gustoće. Navedimo preostale izabrane funkcije gustoće za model (uz $J = 71$):

- *uvjetna apriorna gustoća od θ uz dane (α, β):*

$$f(\theta \mid \alpha, \beta) = \prod_{j=1}^{J} f(\theta_j \mid \alpha, \beta) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1} \left(1-\theta_j\right)^{\beta-1}, \tag{4.5}$$

- *vjerodostojnost:*

$$f(\mathbf{x} \mid \theta) = \prod_{j=1}^{J} f(x_j \mid \theta_j) = \prod_{j=1}^{J} \theta_j^{x_j} \left(1-\theta_j\right)^{n_j - x_j}. \tag{4.6}$$

Provedimo sada postupak koji je iznesen u prethodnom potpoglavlju. Krećemo s koracima koje provodimo analitički.

1. *Nenormalizirana aposteriorna zajednička funkcija gustoće dana je s:*

$$\begin{aligned} f(\theta, \alpha, \beta \mid \mathbf{x}) &\propto f(\alpha, \beta) f(\theta \mid \alpha, \beta) f(\mathbf{x} \mid \theta, \alpha, \beta) \\ &\propto f(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1} \left(1-\theta_j\right)^{\beta-1} \prod_{j=1}^{J} \theta_j^{x_j} \left(1-\theta_j\right)^{n_j - x_j} \end{aligned} \tag{4.7}$$

2. *Za fiksirane (α, β), pokazali smo da θ_j ima $Beta(\alpha + x_j, \beta + n_j - x_j)$ distribuciju, odnosno uvjetna aposteriorna funkcija gustoće za θ dana je s:*

$$f(\theta \mid \alpha, \beta, \mathbf{x}) = \prod_{j=1}^{J} \frac{\Gamma\left(\alpha+\beta+n_j\right)}{\Gamma\left(\alpha+x_j\right)\Gamma\left(\beta+n_j-x_j\right)} \theta_j^{\alpha+x_j-1} \left(1-\theta_j\right)^{\beta+n_j-x_j-1} \tag{4.8}$$

3. *Kako bismo došli do aposteriorne marginalne funkcije gustoće* $f(\alpha,\beta \mid \mathbf{x})$*, ovdje možemo iskoristiti formulu 4.4 i u nju uvrstiti upravo izvedene izraze 4.7 i 4.8:*

$$f(\alpha,\beta \mid \mathbf{x}) \propto f(\alpha,\beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma\left(\alpha+x_j\right)\Gamma\left(\beta+n_j-x_j\right)}{\Gamma\left(\alpha+\beta+n_j\right)}. \tag{4.9}$$

Vratimo se sada na problem izbora hiperapriorne distribucije. S obzirom na to da ne raspolažemo dodatnim znanjem o parametrima (α,β)*, želimo upotrijebiti neinformativnu hiperapriornu distribuciju.*

Za razliku od parametara normalne distribucije, gdje μ *predstavlja očekivanje i* σ^2 *varijancu normalno distribuirane slučajne varijable, parametri beta distribucije nemaju tako jednostavnu i intuitivnu interpretaciju.*

Stoga uvodimo reparametrizaciju:

$$u = logit\left(\frac{\alpha}{\alpha+\beta}\right) = ln\left(\frac{\alpha}{\beta}\right),$$
$$v = \ln(\alpha+\beta).$$

Ako se prisjetimo da je očekivanje $Beta(\alpha,\beta)$ *razdiobe dano s* $\alpha/(\alpha+\beta)$*, dok se* $(\alpha+\beta)$ *može interpretirati kao veličina uzorka, vidimo da nam je nešto intuitivnije razmišljati o novouvedenim parametrima. Na očekivanje i "veličinu uzorka" primjenjujemo logit, odnosno ln funkciju, kako bismo ih sveli na* $\langle -\infty, \infty\rangle$ *skalu.*

Odaberimo za početak neinformativnu nepravu gustoću $f(u,v) \propto 1$ *i pokažimo da bismo s tim izborom dobili i nepravu aposteriornu funkciju gustoće* $f(\alpha,\beta \mid \mathbf{x})$*.*

Dovoljno je pokazati da za fiksiran $\frac{\alpha}{\beta} = c$*,* $c \neq 0$ *i* $(\alpha+\beta) \to \infty$*, imamo* $f(\alpha,\beta \mid \mathbf{x}) \to \infty$*.*

Promotrimo prvo u izrazu 4.9 dio koji se ne odnosi na $f(\alpha,\beta)$ *i označimo ga s*

$$f(\mathbf{x} \mid \alpha,\beta) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma\left(\alpha+x_j\right)\Gamma\left(\beta+n_j-x_j\right)}{\Gamma\left(\alpha+\beta+n_j\right)}.$$

Sada imamo:

$$\begin{aligned} f(\mathbf{x} \mid \alpha,\beta) &\propto \prod_{j=1}^{J} \frac{\left[\alpha \cdots \left(\alpha+y_j-1\right)\right]\left[\beta \cdots \left(\beta+n_j-y_j-1\right)\right]}{(\alpha+\beta)\cdots\left(\alpha+\beta+n_j-1\right)} \\ &\approx \prod_{j=1}^{J} \frac{\alpha^{y_j}\beta^{n_j-y_j}}{(\alpha+\beta)^{n_j}} \\ &= \prod_{j=1}^{J} \left(\frac{\alpha}{\alpha+\beta}\right)^{y_j} \left(\frac{\beta}{\alpha+\beta}\right)^{n_j-y_j}, \end{aligned}$$

što je konstanta, ako smatramo $\mathbf{x}$*,* n *i* $\frac{\alpha}{\beta}$ *fiksnima. Dakle, o apriornoj gustoći* $f(\alpha,\beta)$ *ovisi hoće li* $f(\alpha,\beta \mid \mathbf{x})$ *imati konačan integral. Kako* $f(\alpha,\beta) \propto 1$ *nema konačan integral, ni aposteriorna gustoća neće biti prava funkcija gustoće pa moramo probati s nekim drugim pristupom.*

Uzimimo zato novu reparametrizaciju:

$$\begin{aligned} t &= \frac{\alpha}{\alpha+\beta}, \\ h &= (\alpha+\beta)^{-\frac{1}{2}}, \end{aligned} \tag{4.10}$$

te zadajmo nepravu apriornu funkciju gustoće

$$f(t,h) = f\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{\frac{-1}{2}}\right) \propto 1.$$

Može se pokazati kako na ovaj način dobivamo vjerojatnosnu aposteriornu funkciju gustoće (za dokaz vidjeti [2]).

Na temelju Teorema 1.1.19., izračunom Jacobijana preslikavanja

$$g_1^{-1}(\alpha,\beta) = \left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{\frac{-1}{2}}\right) = (t,h),$$

možemo dobiti odgovarajuću apirornu funkciju gustoće izraženu u (α,β)*.*

Iz raspisa

$$Dg_1^{-1}(\alpha,\beta) = \begin{vmatrix} \frac{\beta}{(\alpha+\beta)^2} & \frac{-\alpha}{(\alpha+\beta)^2} \\ \frac{-1}{2}(\alpha+\beta)^{\frac{-3}{2}} & \frac{-1}{2}(\alpha+\beta)^{\frac{-3}{2}} \end{vmatrix} = -(\alpha+\beta)^{\frac{-5}{2}},$$

slijedi

$$f(\alpha,\beta) \propto (\alpha+\beta)^{\frac{-5}{2}}.$$

Dolazimo do potpunog izraza za aposteriornu funkciju gustoće:

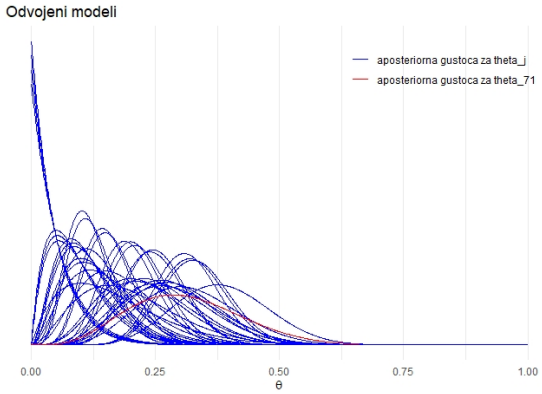
$$f(\alpha,\beta \mid \mathbf{x}) \propto (\alpha+\beta)^{\frac{-5}{2}} \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma\left(\alpha+x_j\right)\Gamma\left(\beta+n_j-x_j\right)}{\Gamma\left(\alpha+\beta+n_j\right)}. \tag{4.11}$$

Izraz 4.11 ne možemo više analitički pojednostavniti, ali se uz izračun gama funkcije za dane vrijednosti od α *i* β *on može izvrijedniti.*

Posebno, 4.11 ne predstavlja funkciju gustoće niti jedne poznate razdiobe. Zato koristimo alate računarske statistike i u programskom jeziku R nastavljamo analizu.

Promotrimo prvo dva pristupa koja smo razmatrali prije uvođenja hijerarhijskog modela i koji će nam služiti za usporedbu. Prvi je pristup bez udruživanja i podrazumijeva da za svaki za parametar θ_j posebno promatramo aposteriornu distribuciju $Beta(\alpha + x_j, \beta + n_j - x_j)$. U drugom pristupu zanemarujemo činjenicu da podaci dolaze iz različitih eksperimenata i na temelju svih podataka dolazimo do aposteriorne distribucije $Beta(\alpha + \sum_{j=1}^{71} x_j, \beta + \sum_{j=1}^{71} n_j - \sum_{j=1}^{71} x_j)$ za θ (koji ovdje predstavlja vjerojatnost razvoja tumora u cijeloj populaciji laboratorijskih štakora). Kako se ne radi o hijerarhijskom modeliranju, za parametre θ_j, odnosno za parametar θ dovoljno je izabrati apriornu distribuciju, što ovdje znači zadati parametre (α, β). Uzimimo stoga neinformativnu apriornu distribuciju $Beta(1,1) = U(0,1)$.

Na Slici 4.1 i Slici 4.2 prikazane su dobivene aposteriorne funkcije gustoće za oba pristupa.



Slika 4.1: 71 aposteriorna funkcija gustoće za svaki θ_j

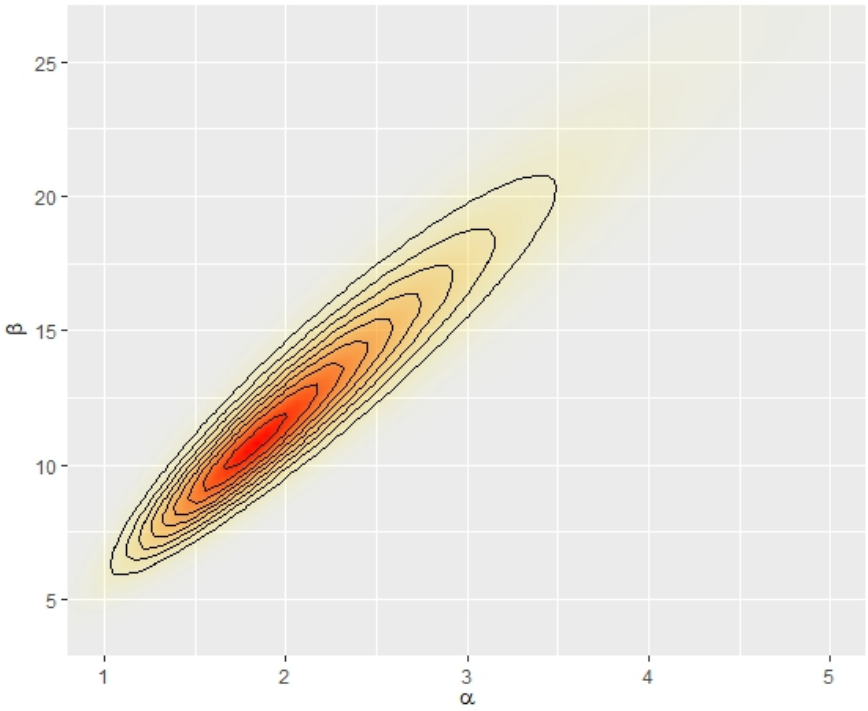
Slika 4.2: Aposteriorna funkcija gustoće za θ, na temelju svih podataka

Izložimo sada postupak kojim smo došli do grafa nenormalizirane aposteriorne funkcije gustoće za parametre (α, β)*, prikazanog na Slici 4.3. Kako se radi o dvodimenzionalnoj funkciji gustoće, računamo funkcijske vrijednosti u točkama rešetke (eng. grid) za* α *i* β *i prikazujemo funkciju preko izohipsa, koje povezuju točke s istim funkcijskim vrijednostima.*

Preciznije, kada podijelimo segment $[0.5, 6]$ *na x osi i segment* $[3, 33]$ *na y osi sa 100 ekvidistantnih točaka, Kartezijev produkt tih skupova točaka predstavlja rešetku za* (α, β) *u kojoj računamo funkciju gustoće.*

Samo izvrijednjavanje aposteriorne funkcije gustoće također zahtijeva nešto opreza: kako bismo osigurali veću numeričku preciznost, prvo logaritmiramo izraz 4.11 te njega evaluiramo u točkama rešetke, zatim od tako dobivenih vrijednosti oduzimamo najveću od njih (kako bismo spriječili "computational overflow") te na kraju primjenjujemo eksponencijalnu funkciju da bismo dobili konačne funkcijske vrijednosti. Zbog ovakvoga postupka, sve se dobivene vrijednosti nalaze između 0 i 1 (iako samu funkciju gustoće još nismo normalizirali).

Naglasimo još kako vanjska, najšira izohipsa povezuje točke u kojima gustoća ima vrijednost jednaku ($0.15\times$ *vrijedost u modu), a na unutarnjoj se izohipsi nalaze točke čija je funkcijska vrijednost (*$0.95\times$ *vrijedost u modu).*



Slika 4.3: Graf marginalne aposteriorne funkcije gustoće $f(\alpha, \beta \mid \mathbf{x})$

Možemo primijeniti isti postupak za prikaz aposteriorne funkcije gustoće za reparametrizaciju s kojom nam je bilo intuitivnije raditi, $(u,v) = \left(ln\left(\frac{\alpha}{\beta}\right), ln(\alpha+\beta)\right)$. *Pri tome, moramo samo uzeti u obzir Teorem 1.1.19. te izraz 4.11 pomnožiti s odgovarajućim Jacobijanom. Račun kojim dolazimo do Jacobijana je sljedeći:*

$$g_2(\alpha,\beta) = \left(ln\left(\frac{\alpha}{\beta}\right), ln(\alpha+\beta)\right),$$

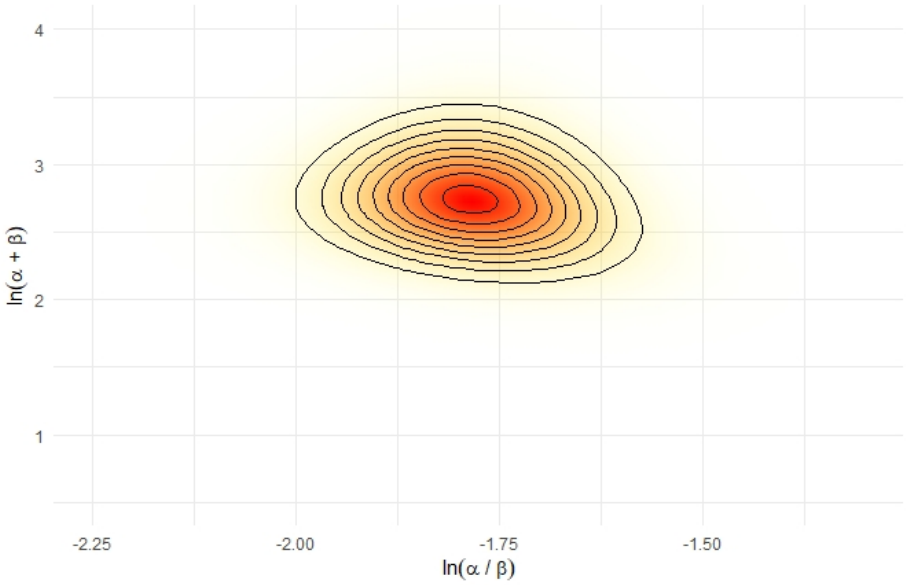
$$Dg_2(\alpha,\beta) = \begin{vmatrix} \frac{1}{\alpha} & -\frac{1}{\beta} \\ \frac{1}{\alpha+\beta} & \frac{1}{\alpha+\beta} \end{vmatrix} = \frac{1}{\alpha\beta},$$

$$Dg_2^{-1}\left(ln\left(\frac{\alpha}{\beta}\right), ln(\alpha+\beta)\right) = \alpha\beta.$$

Nakon što smo na $[-2.3,-1.3]\times[1,5]$ *uzeli rešetku točaka za* (u,v)*, u točkama* $(\alpha,\beta) = g_2^{-1}(u,v)$ *računamo vrijednost nenormalizirane aposteriorne funkcije gustoće:*

$$f\left(g_2^{-1}\left(ln\left(\frac{\alpha}{\beta}\right), ln(\alpha+\beta)\right)\right) \propto \alpha\beta(\alpha+\beta)^{\frac{-5}{2}} \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma\left(\alpha+x_j\right)\Gamma\left(\beta+n_j-x_j\right)}{\Gamma\left(\alpha+\beta+n_j\right)},$$

pa, ponavljajući iste korake kao prije, dolazimo do grafa te funkcije, prikazanog na Slici 4.4.



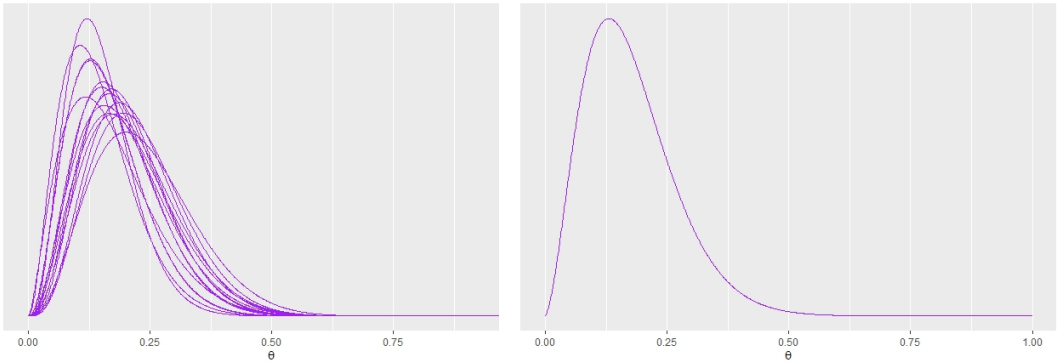
Slika 4.4: Graf aposteriorne funkcije gustoće za $\left(ln\left(\frac{\alpha}{\beta}\right), ln(\alpha+\beta)\right)$

Prisjetimo se interpretacije parametara (u, v) i parametara beta distribucije: u je logit od očekivanja Beta(α, β) razdiobe (koje bismo procijenjivali udjelom jedinki s nekim svojstvom u uzorku), a v je logaritam od "veličine uzorka" za eksperiment modeliran beta razdiobom. Već smo u poglavlju Hijerarhijsko modeliranje komentirali kako bi te dvije veličine $(x_j/n_j, n_j)$ trebale biti nezavisne. U razlici grafova aposteriornih funkcija gustoće ističe se prednost rada s transformiranim parametrima: iako o odnosu α i β nemamo neku predodžbu (a na temelju grafa vidimo da postoji neka pozitivna korelacija), graf za parametre (u, v) dobro odgovara našoj interpretaciji i pretpostavci o njihovom odnosu.

Također, ako se prisjetimo da smo (prije uvođenja bayesovskog hijerarhijskog modela) parametre na temelju podataka procijenili s $\alpha_f = 1.4$ i $\beta_f = 8.6$, vidimo da se mod dobivene aposteriorne funkcije gustoće za (α, β) ponešto razlikuje od te procjene.

Vratimo se sada na parametre (α, β). Nakon normiranja aposteriorne funkcije gustoće $f(\alpha, \beta \mid \mathbf{x})$ (koje provodimo tako da sumiramo sve funkcijske vrijednosti i zatim svaku podijelimo s dobivenom sumom), možemo simulirati slučajni uzorak duljine 100 za (α, β) iz aposteriorne distribucije tako da uzorkujemo točke iz rešetke, gdje je vjerojatnost izbora svake točke jednaka vrijednosti normirane aposteriorne funkcije gustoće u toj točki.

Na temelju simuliranog uzorka za (α, β) možemo doći i do aproksimacije za prediktivnu funkciju gustoće $f(\tilde{\theta} \mid \mathbf{x})$, na način da u svakoj od točaka kojima smo podijelili interval $\langle 0, 1\rangle$ izračunamo aritmetičku sredinu funkcijskih vrijednosti gustoća Beta(α_k, β_k) distribucije ($k = 1, 2, \ldots, 100$) te dobivene sredine shvatimo kao vrijednosti od $f(\tilde{\theta} \mid \mathbf{x})$ u tim točkama. Na Slici 4.5 nalazi se usporedba grafova funkcije gustoće Beta(α_k, β_k) razdiobe za nekoliko simuliranih vrijednosti od (α, β) te grafa prediktivne funkcije gustoće za novo opažanje od θ.



Beta(α, β) za prvih 15 elemenata uzorka iz aposteriorne distribucije za (α, β)

Prediktivna funkcija gustoće $f(\tilde{\theta} \mid \mathbf{x})$

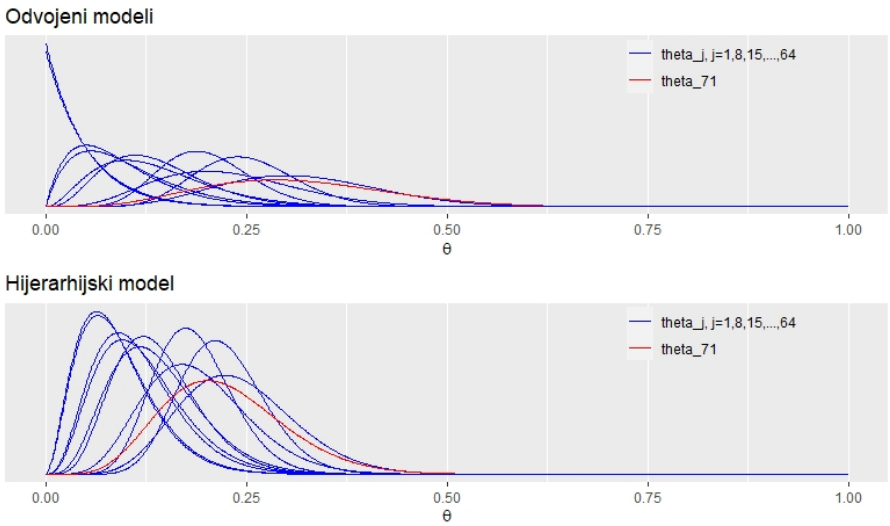
Slika 4.5

Za j-ti eksperiment, svaki od parova (α_k, β_k), $k = 1, \ldots, 100$ *simuliranog uzorka zadaje jednu aposteriornu funkciju gustoće za* θ_j, $f(\theta_j \mid \alpha_k, \beta_k, \mathbf{x})$, *koja je u našem slučaju funkcija gustoće* $Beta(\alpha_k + n_j, \beta_k + n_j - x_j)$ *razdiobe. Ako želimo sažeti te informacije, odnosno ako na temelju simuliranog uzorka za* (α, β) *želimo prikazati aproksimaciju za*

$$f(\theta_j \mid \mathbf{x}) = \int f(\theta_j \mid \alpha, \beta, \mathbf{x}) d(\alpha, \beta),$$

to ponovo radimo na način da u točkama iz intervala $\langle 0, 1 \rangle$ *izračunamo aritmetičku sredinu funkcijskih vrijednosti gustoća* $Beta(\alpha_k + n_j, \beta_k + n_j - x_j)$ *distribucije te dobivene sredine shvatimo kao vrijednost od* $f(\theta_j \mid \mathbf{x})$ *u tim točkama.*

Za usporedbu, na Slici 4.6 prikazujemo za nekoliko θ_j *funkcije* $f(\theta_j \mid \mathbf{x})$ *dobivene bez hijerarhijskog modeliranja, gdje smo svaki* θ_j *promatrali zasebno (pristup bez udruživanja), te aproksimacije za* $f(\theta_j \mid \mathbf{x})$ *dobivene gore opisanim postupkom, preko hijerarhijskog modeliranja. Pri tome smo zbog preglednosti izabrali samo one parametre* θ_j, *gdje j pri dijeljenju sa 7 daje ostatak 1 (mogli smo i na neki drugi način izabrati podskup parametara).*

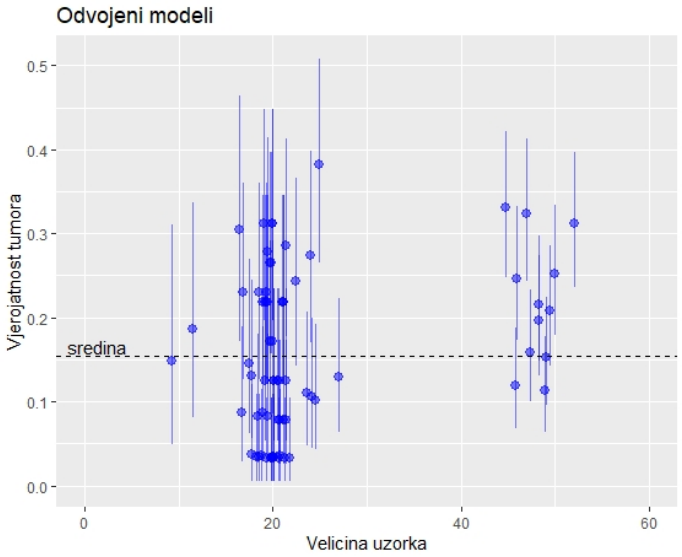


Slika 4.6: Usporedba aposteriornih funkcija gustoća za neke od parametara θ_j

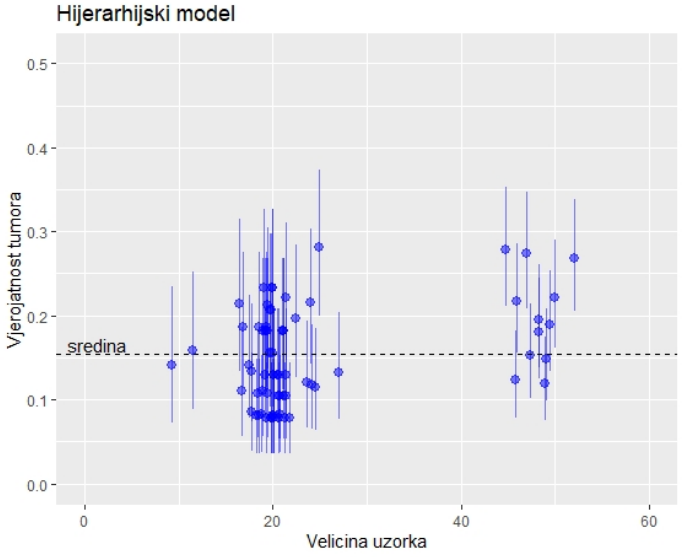
Za kraj, promotrimo 90%-pouzdane intervale za parametre θ_j. *Kod pristupa bez udruživanja podataka, rubovi intervala su odgovarajući kvantili* $Beta(1 + n_j, 1 + n_j - x_j)$ *razdiobe. Kod hijerarhijskog modeliranja, na sličan način kao i prije, za procjenu intervala rubove određujemo kao aritmetičku sredinu kvantila* $Beta(\alpha_k + n_j, \beta_k + n_j - x_j)$ *razdioba.*

Intervali pouzdanosti za odvojene modele prikazani su na Slici 4.7, a za hijerarhijski model na Slici 4.8. Kako ima više pokusa s istim brojem laboratorijskih štakora koji su

sudjelovali, zbog preglednosti smo veličinu uzorka koja je na x-osi perturbirali, dodajući na nju simuliranu realizaciju slučajne varijable $X \sim N(0,1)$. Točkom su označeni medijani aposteriornih distribucija $f(\theta_j \mid \mathbf{x})$, a horizontalna linija prikazuje aritmetičku sredinu svih podataka, $(\sum x_j)/(\sum n_j)$.



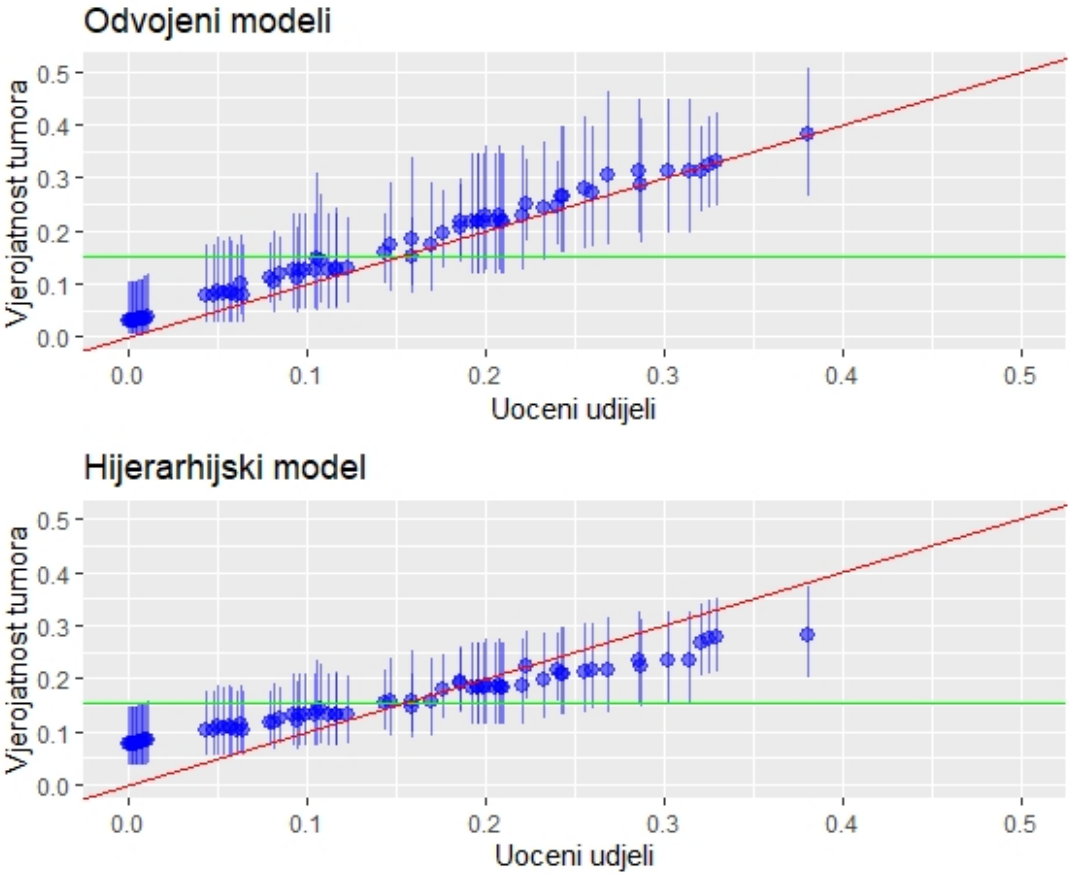
Slika 4.7: 90%-pouzdani intervali za parametre θ_j kod pristupa bez udruživanja



Slika 4.8: 90%-pouzdani intervali za θ_j kod bayesovskog hijerarhijskog modeliranja

Bitno je uočiti kako su kod odvojenih modela (bez udruživanja), intervali pouzdanosti znatno širi i kako su medijani aposteriornih distribucija kod bayesovskog hijerarhijskog modela bliži zajedničkoj aritmetičkoj sredini, nego što je to slučaj kod odvojenih modela.

Promotrimo još jedan prikaz procijenjenih intervala pouzdanosti dan Slikom 4.9, gdje su na x osi uočeni udjeli x_j/n_j. Ponovo smo zbog preglednosti perturbirali vrijednosti na x-osi. Ovdje možemo uočiti kako se medijani aposteriornih distribucija za bayesovski hijerarhijski model udaljavaju od uočenih udjela x_j/n_j (crvena dijagonalna linija), te su bliži aritmetičkoj sredini svih podataka (zelena horizontalna linija). Time je lijepo ilustrirano kako u bayesovskom hijerarhijskom modeliranju aposteriorni rezultati predstavljaju kompromis između pristupa bez udruživanja i pristupa s potpunim udruživanjem podataka.



Slika 4.9: Usporedba 90%-pouzdanih intervala za parametre θ_j

Ovim prikazom završavamo primjer.

U sljedećem primjeru uvodimo normalni bayesovski hijerarhijski model, u kojemu su opaženi podaci normalno distribuirani, pri čemu svaka od grupa ima različito očekivanje, ali sve grupe imaju zajedničku i poznatu varijancu te nepoznati parametri (očekivanja grupa podataka) dolaze iz zajedničke normalne distribucije.

Primjer 4.3.2. *Pretpostavimo da je provedeno J eksperimenata te za svaki želimo procijeniti nepoznati parametar θ_j na temelju n_j podataka $x_{\cdot j}$ koji su normalno distribuirani:*

$$X_{i,j} \mid \theta_j \sim N(\theta_j, \sigma^2), \quad i = 1, \ldots, n_j; \quad j = 1, \ldots, J.$$

Promotrimo prvo kako bismo pristupili obradi podataka $x_{i,j}$, bez primjene bayesovskog zaključivanja i hijerarhijskih modela. Vjerojatno bi nam prvo palo na pamet procijeniti parametre θ_j uzoračkom sredinom podataka $(x_{1,j}, x_{2,j}, \ldots, x_{n_j,j})$: $\hat{\theta}_j = \bar{x}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{i,j}$. Osim što su uzoračke sredine konzistentni procjenitelji za θ_j, uz pretpostavke na distribuciju podataka, znamo da vrijedi $\bar{X}_{\cdot j} \mid \theta_j \sim N(\theta_j, \sigma_j^2)$, pri čemu $\sigma_j^2 = \sigma^2 / n_j$.

Ipak, ako raspolažemo sa svega nekoliko podataka za svaku grupu i broj je grupa velik, tako dobiveni procjenitelji ne bi nam bili relevantni. U tom bismo slučaju vjerojatno udružili podatke i izračunali ukupnu aritmetičku sredinu: $\hat{\theta} = \bar{x} = \left(\sum_{j=1}^{J} n_j\right)^{-1} \sum_{i=1}^{n_j} x_{i,j}$, gdje onda podrazumijevamo $X_{i,j} \mid \theta \sim N(\theta, \sigma^2)$, $\quad i = 1, \ldots, n_j; \quad j = 1, \ldots, J$.

U frekvencionističkom pristupu, kada nismo sigurni razlikuju li se grupe "dovoljno" da ih promatramo odvojeno ili su pak "dovoljno" slične da izaberemo pristup s udruživanjem podataka, možemo provesti ANOVA test, za koji je nulta hipoteza da ne postoji statistički značajna razlika između parametara očekivanja θ_j, tj. $\theta_1 = \theta_2 = \ldots \theta_J$. Ako na temelju podataka tu hipotezu (na odabranoj razini značajnosti) možemo odbaciti, izabrat ćemo pristup bez udruživanja (gdje svaki θ_j procjenjujemo zasebno uzoračkom sredinom), a ako hipotezu ne možemo odbaciti, izabrat ćemo pristup s udruživanjem podataka (gdje onda uzoračkom sredinom podataka iz svih grupa procjenjujemo θ).

Vratimo se sada na bayesovsko hijerarhijsko modeliranje: ovdje smo za cilj postavili pronaći procjenitelje za parametre očekivanja θ_j, što kod bayesovskog zaključivanja podrazumijeva pronaći aposteriorne procjenitelje za θ_j, definirane kao očekivanje aposteriorne distribucije za θ_j, $\mathbb{E}(\theta_j \mid \mathbf{x})$. Kao što smo već istaknuli, prednost je bayesovskog hijerarhijskog modeliranja što ne moramo izabrati između promatranja grupa kao potpuno zasebnih i potpunog udruživanja podataka. Štoviše, vidjet ćemo kako se primjenom hijerarhijskog modeliranja dobiva aposteriorni procjenitelj za θ_j koji je zapravo konveksna kombinacija uzoračke sredine j-te grupe i sredine svih podataka.

Kako je svrha ovog primjera da na još jedan način ilustrira korisnost bayesovskog pristupa i hijerarhijskog modeliranja, izostavljen je raspis kojim se dolazi do aposteriornih funkcija gustoća te su samo navedeni najbitniji rezultati. Svi se detalji potrebni za taj raspis nalaze u [3], u potpoglavlju "5.4 Normal model with exchangeable parameters".

Bayesovski hijerarhijski model postavljamo na sljedeći način:

$$\begin{aligned} X_{ij} \mid \theta_j, \sigma^2 \ \ (n.j.d) \ &\sim N\left(\theta_j, \sigma^2\right) \ za\ 1 \leq i \leq n_j, \\ \theta_j \mid \mu, \tau \ \ (n.j.d) \ &\sim N\left(\mu, \tau^2\right) \ za\ 1 \leq i \leq J, \\ f(\mu, \tau) &\propto f(\mu \mid \tau) f(\tau), \\ f(\mu \mid \tau) &\propto 1, \\ f(\tau) &\propto 1. \end{aligned}$$

Dakle, i za apriornu funkciju gustoće parametra μ uvjetno na τ i za apriornu funkciju gustoće parametra τ uzimamo neinformativnu nepravu gustoću proporcionalnu 1 na cijelom $\mathbb{R}$. Može se pokazati da takvim izborom dobivamo prave aposteriorne funkcije gustoće.

Aposteriorna distribucija za parametar θ_j, uvjetno na parametre μ, τ i opažene podatke **x***, zapravo je normalna:*

$$\theta_j \mid \mu, \tau, \mathbf{x} \sim N(\hat{\theta}_j, V_j),$$

pri čemu su

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{x}_{.j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \quad i \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}.$$

Možemo iščitati da je aposteriorni procjenitelj za θ_j, definiran kao očekivanje ove normalne distribucije, upravo $\hat{\theta}_j$ koji je dan kao konveksna kombinacija uzoračke sredine $j-te$ grupe i očekivanja μ distribucije parametara $(\theta_1, \ldots, \theta_j)$. Pri tome, utjecaj uzoračke sredine i očekivanja na konačni procjenitelj ovisi o preciznosti (inverznoj vrijednosti varijance) odgovarajućih distribucija (od $\bar{X}_{.j}$ i $(\theta_1, \ldots, \theta_j)$).

Zanimljiva je još i aposteriona funkcija gustoće parametra μ, dana s:

$$\mu \mid \tau, \mathbf{x} \sim N(\hat{\mu}, V_\mu),$$

pri čemu su

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}\bar{x}_{.j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}} \quad i \quad V_\mu^{-1} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}.$$

Ponovo uočavamo da je aposteriorni procjenitelj za μ (kojeg možemo usporediti s ukupnom uzoračkom sredinom $\bar{x}$ iz frekvencionističkog pristupa), dan kao konveksna kombinacija uzoračkih sredina grupa, $\bar{x}_{.j}$, gdje je doprinos svake grupe veći, što je σ_j^2 manji (odnosno, što je n_j veći).

U potpoglavlju 5.5 u [3], može se pronaći primjena normalnog bayesovskog hijerarhijskog modela na stvarne podatke. Taj se primjer ujedno može shvatiti i kao ilustracija jednostavnog linearnog hijerarhijskog modela.

U ovom su se radu pod pojmom modela zapravo podrazumijevale vjerojatnosne distribucije (odnosno, funkcije gustoće) kojima smo matematički opisivali promatrane podatke i parametre. No, većinu pojam modela asocira na (generalizirane) linearne modele.

Bayesovski hijerarhijski linearni modeli koriste se u sličnim situacijama kao i ostali hijerarhijski modeli: kada radimo s podacima koji su strukturirani u više grupa te je radi toga narušena pretpostavka njihove nezavisnosti. Za više detalja, preporuča se pogledati 14. i 15. poglavlje u [3].

Ovom napomenom završavamo rad.

Dodatak

Na ovom se mjestu nalazi kod u programskom jeziku R kojim su dobiveni rezultati i grafovi iz prethodnog poglavlja. Kod je napisan uz pomoć [11].

```
library(ggplot2)
library(tidyr)
library(latex2exp)
library(gridExtra)


x_j <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,
       2,1,5,2,5,3,2,7,7,3,3,2,9,10,4,4,4,4,4,4,4,10,4,4,4,5,11,12,
       5,5,6,5,6,6,6,6,16,15,15,9,4)
n_j <- c(20,20,20,20,20,20,20,19,19,19,19,18,18,17,20,20,20,20,19,19,
         18,18,25,24,23,20,20,20,20,20,20,10,49,19,46,27,17,49,47,20,
         20,13,48,50,20,20,20,20,20,20,20,48,19,19,19,22,46,49,20,20,
         23,19,22,20,20,20,52,46,47,24,14)
seg <- seq(0.0001, 0.9999, length.out = 1000)


# No-pooling pristup (odvojeni modeli, za svaki theta_j zasebno)

beta_posterior <- function(n_j, x_j, seg){
  dbeta(seg, x_j+1, n_j-x_j+1)
}

df_nopool <- mapply(beta_posterior, n_j, x_j, MoreArgs = list(seg = seg)) %>%
  as.data.frame() %>% cbind(seg) %>%
  pivot_longer(cols = !seg, names_to = "indeks", values_to = "p")

oznake1  <- paste('aposteriorna gustoca za', c('theta_j', 'theta_71'))
plot_nopool <- ggplot(data = df_nopool) +
```

```
  geom_line(aes(x = seg, y = p, color = (indeks=='V71'), group = indeks)) +
  labs(x = expression(theta), y = '', title = '', color = '') +
  scale_y_continuous(breaks = NULL) +
  scale_color_manual(values = c('blue','red'), labels = oznake1) +
  theme(legend.background = element_blank(), legend.position = c(0.8,0.9))
plot_nopool


# Complete-pooling pristup (svi podaci zajedno, jedan model za sve
# eksperimente)

df_pool <- data.frame(x = seg, p = dbeta(seg,sum(x_j)+1,sum(n_j)-sum(x_j)+1))
plot_pool <- ggplot(data = df_pool) +
  geom_line(aes(x = seg, y = p, color = '1')) +
  labs(x = expression(theta), y = '', color = '') +
  scale_y_continuous(breaks = NULL) +
  scale_color_manual(values = 'red', labels = 'aposteriorna gustoca
                     za zajednicki theta') +
  theme(legend.background = element_blank(), legend.position = c(0.7,0.9))
plot_pool


# Graf marginalne aposteriorne gustoce za \alpha, \beta

gs = 100
segA <- seq(0.5, 6, length.out = gs)
segB <- seq(3, 33, length.out = gs)
gridA <- rep(segA, each = gs)
gridB <- rep(segB, gs)

log_post <- function(a, b, x, n) {
  log(a+b)*(-5/2) +
    sum(lgamma(a+b) - lgamma(a) - lgamma(b) + lgamma(a+x) + lgamma(b+n-x) -
          lgamma(a+b+n))
}

lp <- mapply(log_post, gridA, gridB, MoreArgs = list(x_j, n_j))
df_ab <- data.frame(x = gridA, y = gridB, p = exp(lp - max(lp)))


ggplot(data = df_ab, aes(x = x, y = y)) +
  geom_raster(aes(fill = p, alpha = p), interpolate = T) +
```

```
  geom_contour(aes(z = p), colour = 'black', size = 0.2,
               breaks = seq(from=0.15, to=0.95, by=0.1)) +
  coord_cartesian(xlim = c(1,5), ylim = c(4, 26)) +
  labs(x = TeX('$\\alpha$'), y = TeX('$\\beta$')) +
  scale_fill_gradient(low = 'yellow', high = 'red', guide = F) +
  scale_alpha(range = c(0, 1), guide = F)


# Graf marginalne aposteriorne funkcije gustoće za reparametrizaciju (u,v)

transf1 <- function(u,v) {
  exp(u+v)/(1+exp(u))
}

transf2 <- function(u,v) {
  exp(v)/(1+exp(u))
}

tsegA <- seq(-2.3,-1.3, length.out = gs)
tsegB <- seq(1, 5, length.out = gs)
tAB <- rep(tsegA, each = gs)
tBA <- rep(tsegB, gs)
tgridA <- transf1(tAB, tBA)
tgridB <- transf2(tAB, tBA)


t_log_post <- function(a, b, x, n) {
  log(a) +log(b) + log(a+b)*(-5/2)+
    sum(lgamma(a+b) - lgamma(a) - lgamma(b) + lgamma(a+x) + lgamma(b+n-x) -
          lgamma(a+b+n))
}

tlp <- mapply(t_log_post, tgridA, tgridB, MoreArgs = list(x_j, n_j))
df_tab <- data.frame(x = tAB, y = tBA, p = exp(tlp - max(tlp)))

ggplot(data = df_tab, aes(x = x, y = y)) +
  geom_raster(aes(fill = p, alpha = p), interpolate = T) +
  geom_contour(aes(z = p), colour = 'black', size = 0.2,
               breaks = seq(from=0.15, to=0.95, by=0.1)) +
  coord_cartesian(xlim = c(-2.25,-1.3), ylim = c(0.5,4)) +
  labs(x=TeX('$ln(\\alpha$/$\\beta)$'), y=TeX('$ln(\\alpha$ + $\\beta)$')) +
  scale_fill_gradient(low = 'yellow', high = 'red', guide = F) +
```

```
  scale_alpha(range = c(0, 1), guide = F)


# Simuliranje slučajnog uzorka iz aposteriorne distribucije za alpha, beta

n_sample <- 100
sample_i <- sample(length(df_tab$p), size = n_sample,
                   replace = T, prob = df_tab$p/sum(df_tab$p))
sample_A <- gridA[sample_i[1:n_sample]]
sample_B <- gridB[sample_i[1:n_sample]]

df_absample <- mapply(function(a, b, x) dbeta(x, a, b),
                      sample_A, sample_B, MoreArgs = list(x = seg)) %>%
  as.data.frame() %>% cbind(seg) %>%
  pivot_longer(cols = !seg, names_to = "indeks", values_to = "p")

itonum <- function(x) {
  strtoi(substring(x,2))
}

plot_absample <- ggplot(data = subset(df_absample, itonum(indeks) <= 15)) +
  geom_line(aes(x = seg, y = p, group = indeks), color='purple') +
  labs(x = expression(theta), y = '', title = '') +
  scale_y_continuous(breaks = NULL)
plot_absample

df_theta_pred <- spread(df_absample, indeks, p) %>% subset(select = -seg) %>%
  rowMeans() %>% data.frame(x = seg, p = .)

plot_theta_pred <- ggplot(data = df_theta_pred) +
  geom_line(aes(x = x, y = p), color='purple') +
  labs(x = expression(theta), y = '') +
  scale_y_continuous(breaks = NULL)

plot_theta_pred


# Usporedba no-pooling pristupa i hijerarhijskog modeliranja

oznake2 <- c('theta_j, j=1,8,15,...,64', 'theta_71')

plot_sep <- ggplot(data = subset(df_nopool, itonum(indeks)%%7==1)) +
```

```
  geom_line(aes(x = seg, y = p, color = (indeks=='V71'), group = indeks)) +
  labs(x = expression(theta), y = '', title = 'Odvojeni modeli',color = '') +
  scale_y_continuous(breaks = NULL) +
  scale_color_manual(values = c('blue', 'red'), labels = oznake2) +
  theme(legend.background = element_blank(), legend.position = c(0.8,0.9))
plot_sep

bdensity2 <- function(n, u, a, b, seg){
  rowMeans(mapply(dbeta, a + u, n - u + b, MoreArgs = list(x = seg)))
}

df_hier <- mapply(bdensity2, n_j, x_j, MoreArgs=list(sample_A,
                                                     sample_B, seg)) %>%
  as.data.frame() %>% cbind(seg) %>%
  pivot_longer(cols = !seg, names_to = "indeks", values_to = "p")

plot_hier <- ggplot(data = subset(df_hier, itonum(indeks)%%7==1)) +
  geom_line(aes(x = seg, y = p, color = (indeks=='V71'), group = indeks)) +
  labs(x = expression(theta), y = '', title = 'Hijerarhijski model',
       color = '') +
  scale_color_manual(values = c('blue', 'red'), labels = oznake2) +
  scale_y_continuous(breaks = NULL) +
  theme(legend.background = element_blank(), legend.position = c(0.8,0.9))
plot_hier

grid.arrange(plot_sep, plot_hier)


# Usporedba 90%-pouzdanih intervala za theta_j za odvojene modele i
# hijerarhijski model

mean_all <- (sum(x_j)+1)/(sum(n_j)+2)

n_j_perm <- n_j + rnorm(length(n_j),0,1)

qq_nopool <- data.frame(id=1:length(n_j), n_j=n_j, x_j=x_j,
                        q10=qbeta(0.1, x_j+1, n_j-x_j+1),
                        q50=qbeta(0.5, x_j+1, n_j-x_j+1),
                        q90=qbeta(0.9, x_j+1, n_j-x_j+1))

qhier <- function(q, n, x_j) {
  colMeans(mapply(function(q, n, x_j, a, b)
```

```
    mapply(qbeta, q, a+x_j, n-x_j+b), q, n_j, x_j,
    MoreArgs = list(sample_A, sample_B)))
}

qq_hier <- data.frame(id=1:length(n_j), n_j=n_j, x_j=x_j,
                      q10 = qhier(0.1, n, x_j), q50 = qhier(0.5, n, x_j),
                      q90 = qhier(0.9, n, x_j))

plot_sep_interval <- qq_nopool %>%
  ggplot(aes(x = n_j_perm, y = q50, ymin = q10, ymax = q90)) +
  geom_pointrange(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = mean_all, linetype = "dashed")+
  lims(x = c(0,60), y = c(0,0.51))+
  labs(x = "Veličina uzorka", y = "Vjerojatnost tumora",
       title = "Odvojeni modeli") +
  annotate("text", x = 0, y = mean_all, label = "sredina",
           vjust = -0.2, hjust = 0.3)

plot_hier_interval <- qq_hier %>%
  ggplot( aes(x = n_j_perm, y = q50, ymin = q10, ymax = q90)) +
  geom_pointrange(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = mean_all, linetype = "dashed") +
  lims(x = c(0,60), y = c(0,0.51))+
  labs(x = "Veličina uzorka", y = "Vjerojatnost tumora",
       title = "Hijerarhijski model") +
  annotate("text", x=0, y=mean_all, label = "sredina",
           vjust = -0.2, hjust = 0.3)

plot_sep_interval
plot_hier_interval

udio <- (x_j/n_j)
duljina <- length(udio[udio == 0])
udio[1:duljina] <- seq(0,by=0.0008,length.out=duljina)
udio[(duljina+1):71] <- udio[(duljina+1):71] + rnorm(71 - duljina, 0, 0.01)

plot_sep_interval2 <- qq_nopool %>%
  ggplot(aes(x = udio, y = q50, ymin = q10, ymax = q90)) +
  geom_pointrange(color = "blue", alpha = 0.5) +
  geom_hline(color="green",yintercept = mean_all) +
  geom_abline(slope=1, intercept=0, colour="red") +
  lims(x = c(0,0.5), y = c(0,0.51))+
```

```
  labs(x = "Uoceni udijeli", y = "Vjerojatnost tumora",
       title = "Odvojeni modeli")

plot_hier_interval2 <- qq_hier %>%
  ggplot( aes(x = udio, y = q50, ymin = q10, ymax = q90)) +
  geom_pointrange(color = "blue", alpha = 0.5) +
  geom_hline(color="green",yintercept = mean_all) +
  geom_abline(slope=1, intercept=0, colour="red") +
  lims(x = c(0,0.5), y = c(0,0.51))+
  labs(x = "Uoceni udjeli", y = "Vjerojatnost tumora",
       title = "Hijerarhijski model")


grid.arrange(plot_sep_interval2, plot_hier_interval2)
```

Bibliografija

[1] Bolstad, William M.: *Introduction to Bayesian Statistics*. John Wiley Sons, 2007.

[2] Gelman, A., B. Carlin, H. Stern i R. Charnigo: *Solutions to some exercises from Bayesian Data Analysis, third edition*. http://www.stat.columbia.edu/~gelman/book/solutions3.pdf.

[3] Gelman, A., B. Carlin, H. Stern, D.B. Dunson, A. Vehtari i D.B. Rubin: *Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2014.

[4] Huzak, Miljenko: *Materijali s kolegija Statistika*. https://web.math.pmf.unizg.hr/nastava/stat/index.php?sadrzaj=predavanja.php.

[5] Huzak, Miljenko: *Vjerojatnost i matematička statistika - skripta Matematičkog odsjeka Prirodoslovno-matematičkog fakulteta u Zagrebu*. 2006. http://aktuari.math.pmf.unizg.hr/docs/vms.pdf.

[6] Huzak, Miljenko: *Matematička statistika - skripta Matematičkog odsjeka Prirodoslovno-matematičkog fakulteta u Zagrebu*. 2020. https://web.math.pmf.unizg.hr/nastava/ms/index.php?sadrzaj=predavanja.php.

[7] Jordan, Michael: *Materijali s kolegija Bayesian Modeling and Inference na Berkeleyu*. http://people.eecs.berkeley.edu/~jordan/courses/260-spring10/lectures/index.html.

[8] Sandrić, Nikola i Zoran Vodraček: *Vjerojatnost - skripta Matematičkog odsjeka Prirodoslovno-matematičkog fakulteta u Zagrebu*. 2019. https://www.pmf.unizg.hr/images/50023697/vjer_predavanja.pdf.

[9] Sarapa, Nikola: *Teorija vjerojatnosti*. Školska Knjiga, 1986.

[10] Stigler, Stephen M.: *Thomas Bayes's Bayesian Inference*. Journal of the Royal Statistical Society. Series A (General), 145(2):250–258, 1982, ISSN 00359238. http://www.jstor.org/stable/2981538, posjećena 2022-12-07.

[11] Vehtari, Aki i Markus Paasiniemi: *Bayesian data analysis demo 5.1*. https://avehtari.github.io/BDA_R_demos/demos_ch5/demo5_1.html.

Sažetak

U statistici često koristimo pretpostavku da su podaci međusobno nezavisni, no ona nije sasvim opravdana kada radimo s opažanjima koja su, zbog načina na koji su prikupljeni ili prirode problema kojeg opisuju, strukturirani u nekoliko grupa. Zbog toga uvodimo hijerarhijske modele, kod kojih je distribucija podataka za svaku grupu parametrizirana s θ_j, $j = 1, 2, \ldots, J$, a odnos među parametrima modeliran je zajedničkom vjerojatnosnom distribucijom nad $\theta_1, \ldots, \theta_J$. U okviru hijerarhijskih modela, posebno se ističe prednost bayesovskog pristupa, koji omogućava da pri donošenju zaključaka za jednu grupu iskoristimo i podatke o preostalim grupama. Bayesovsko hijerarhijsko modeliranje podrazumijeva da je i zajednička distribucija od $\theta_1, \ldots, \theta_J$ parametrizirana preko hiperparametara ϕ, za koje također promatramo apriornu i aposeriornu distribuciju i za koje donosimo statističke zaključke slijedeći bayesovske principe. U ovom smo radu iznijeli osnove bayesovske statistike, motivirali uvođenje hijerarhijskih modela te obradili bayesovsko hijerarhijsko modeliranje, s naglaskom na postavljanju modela i analizi dobivenih rezultata. Na primjerima su ilustrirane prednosti i glavne značajke bayesovskog hijerarhijskog pristupa modeliranju klasteriranih podataka, kao i metode računarske statistike korisne za prikaz rezultata.

Summary

When doing statistical analysis, we often assume that data points are mutually independent. However, due to the way it was collected or the nature of the problem it came from, data is often structured into groups or clusters. In these situations, the assumption of independence comes into question. To tackle that, we introduce hierarchical models, where the probability distribution of data for each of the J groups is defined by the parameter θ_j and the relationship between these parameters is modeled by the joint distribution of $\theta_1, \ldots, \theta_J$. Bayesian inference further broadens the use of hierarchical models because it enables us to make statistical conclusions about a certain group while incorporating information contained in the data belonging to other groups. In Bayesian hierarchical models, the joint distribution of $\theta_1, \ldots, \theta_J$ is defined by the hyperparameters ϕ, for which we also set a prior and calculate the posterior distribution using Bayesian inference. In this master's thesis, we present the main theoretical results of Bayesian statistics, we discuss the motivation behind hierarchical modeling and give a detailed description of Bayesian hierarchical models, focusing on the procedure of setting up and analysing these models. We highlight the key advantages and principles of Bayesian hierarchical modeling using examples in which we also showcase computational methods for displaying results.

Zahvale

Na ovom se mjestu želim od srca zahvaliti svima koji su mi bili podrška i koji su mi, izravno i neizravno, pomogli da ovim radom privedem svoj diplomski studij kraju.

Za početak, velike zahvale i zasluge idu mojoj mentorici, doc. dr. sc. Azri Tafro, kako zbog pomoći i savjeta pri pisanju i matematičkim nedoumicama, tako i zbog ugodne i opuštene komunikacije. Veliko hvala mojim roditeljima, koji su mi pružili bezuvjetnu ljubav i uvijek vjerovali da ću ostvariti svoje ciljeve i želje. Zahvaljujem i baki i djedu, koji su sve ove godine ulagali mnogo ljubavi, vremena i truda u mene i moje obrazovanje. Veliko hvala i Zvoni, na beskonačnom razumijevanju, ljubavi i podršci. Za kraj, zahvaljujem se svim prijateljima, koji su mi uljepšali i obogatili studentske dane.

Životopis

Rođena sam 1.6.1999. u Splitu, gdje sam pohađala osnovnu školu i III. (prirodoslovno-matematičku) gimnaziju. Državnu maturu položila sam 2017. godine i nastavila sam obrazovanje na preddiplomskom studiju Matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. Na istom sam fakultetu 2020. godine upisala diplomski studij Matematičke statistike.

Nagrađena sam od strane Ministarstva znanosti i obrazovanja za uspjeh na državnoj maturi (2017.) i od strane Matematičkog odsjeka PMF-a za uspjeh na diplomskom studiju (2021.).

Tri sam godine (2019.-2022.) bila studentska predstavnica za Matematički odsjek. Od 2021.godine obavljam studentsku praksu u istraživačkom laboratoriju za analizu teksta i inženjerstvo znanja (*TakeLab*) na Fakultetu elektrotehnike i računarstva u Zagrebu.