UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Bayesovský odhad výskytu Downova syndromu

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **Mgr. Ondřej Vencálek, Ph.D.**
Vypracoval(a): **Dagmar Náprstková**
Studijní program: B1103 Aplikovaná statistika
Studijní obor Aplikovaná statistika
Forma studia: prezenční
Rok odevzdání: 2020

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Dagmar Náprstková

Název práce: Bayesovský odhad výskytu Downova syndromu

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: Mgr. Ondřej Vencálek, Ph.D.

Rok obhajoby práce: 2020

Abstrakt: Bakalářská práce se zabývá pravděpodobností výskytu Downova syndromu v České republice pomocí Bayesovské statistiky. Odhad je vypočten na základě znalosti zachycených případů v ČR výskytu z let 1961 až 2011.

Klíčová slova: Bayesovské metody, Bayesova věta, Downův syndrom

Počet stran: 37

Počet příloh:

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Dagmar Náprstková

Title: Bayesian estimation of the Down syndrome incidence

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Ondřej Vencálek, Ph.D.

The year of presentation: 2020

Abstract: Bechelor's work is engaged in the probability of an incidence of the Down syndrome in the Czech Republic by the help of Bayesian statistic. Estimation is calculated by the help of a dataset of intercepted causes between years 1961 and 2011.

Key words: Bayesian methods, Bayes' rule, Down syndrome

Number of pages: 37

Number of appendices:

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracovala samostatně pod vedením pana Mgr. Ondřeje Vencálka, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Na tomto místě bych chtěla poděkovat panu Mgr. Ondřeji Vencálkovi, Ph.D. za odborné vedení mé práce, za cenné poznámky a čas, který mi věnoval. Dále bych chtěla poděkovat své rodině, která mě podporovala během mého studia a bez jejíž pomoci bych tuto práci nemohla dokončit.

Úvod

Má práce se zabývá Bayesovskou statistikou, která je důležitou součástí statistické analýzy dat. Důvodem pro výběr daného tématu je osobní zájem o data s přírodním zaměřením.

První kapitola podává základní informace o Downovu syndromu. Druhá kapitola popisuje stěžejní myšlenku bayesovského přístupu ke statistické analýze dat. Třetí kapitola definuje podmíněnou pravděpodobnost a dvě další vlastnosti, které jsou nutné k zavedení Bayesovy věty. Čtvrtá kapitola vysvětluje postup Bayesovské analýzy. Seznamuje s pojmy věrohodnostní funkce, apriorní rozdělení a posteriorní rozdělení. Pátá kapitola nás seznamuje s intervaly spolehlivosti. Poslední kapitola reprezentuje uvedený postup v praxi.

Cílem mé práce je výpočet odhadu pravděpodobnosti výskytu Downova syndromu v České republice. K tomuto výpočtu budou použity Bayesovské metody a statistický software R.

Základní informace o Downovu syndromu

Jedince s Downovým syndromem můžeme označit za jedinečné osobnosti trpící jistou mírou retardace, která je pro každého postiženého svým způsobem specifická. To zahrnuje jak fyzický vzhled, tak i povahu a další vlastnosti. Veškeré informace lze nalézt v [11].

Oči mohou být zelené, hnědé modré nebo to může být nějaká kombinace zmíněných barev. Barva vlasů pokrývá barevnou škálu od blonďaté, přes rezavou až po tmavé odstíny. Pleť může být tmavá nebo světlá. Děti se podobají například matce, otci, babičkám nebo třeba strýci. Někteří jedinci jsou velmi energičtí, u druhých lze pozorovat značně méně aktivní chování. Jedni milují hudbu a u druhých se mine účinkem. Přes tyto zdánlivě běžné lidské rysy, kterými se vzájemně dva různí lidé odlišují, lze stanovit jistou množinu přesných znaků, které spojují všechny jedince postižené určitým typem syndromu.

Syndrom lze definovat jako soubor příznaků, které se společně projeví. Downův patří mezi tzv. vrozené syndromy, u nichž jsou poruchy zřejmé hned po narození dítěte.

Downův syndrom byl uznán jako samostatná diagnóza v roce 1866, kdy anglický lékař John Langdon Down (1828-1896) poprvé popsal jeho charakteristické znaky. Doktor Down však nebyl schopen ve své době jednoznačně definovat příčinu této diagnózy. Zvrat přišel roku 1932, kdy doktor Waardenburg vyjádřil svou ideu, že Downův syndrom by mohl být zapříčiněn chromozomální poruchou. A nakonec přelomovým rokem se stal rok 1959, kdy Waardenburgův francouzský kolega, doktor Lejeune, v Paříži spolu se svými spolupracovníky demonstroval

fakt, že Donwnův syndrom má přímou souvislost s konkrétním nadbytečným chromozomem.

Lidská buňka obsahuje jádro, v němž je uložena kompletní genetická informace (tzv. geny). Geny jsou nezbytnou součástí procesu zvaném buněčné dělení. Podle [6] se v jádře jedné buňky nachází zhruba 20 000 až 25 000 genů, které jsou zodpovědné za výrobu zcela specifického proteinu určujícího jednu charakteristiku lidského těla (např.: barva očí). Geny jsou v jádře uspořádány do řetězců (struktura řetězců se dá přirovnat ke korálkům na nitce) a tyto řetězce se nazývají chromozomy. Každá buňka má 46 chromozomů a ty jsou uspořádány do 23 párů. Každý pár obsahuje jednu polovinu genetické sady od otce a druhou od matky.

Downův syndrom je způsoben tehdy, když 21. pár chromozomů obsahuje jeden nadbytečný chromozom. Přítomnost třetího chromozomu způsobí produkci určitých nadbytečných bílkovin, které poruší normální vývoj plodu.

Základní myšlenka bayesovského přístupu

Následující kapitola představí principy Bayesovské statistiky, viz [7]. Zajímavé vysvětlení této metody lze ukázat na následujícím příkladě:

Ráno vyjdeme před dům a pozorujeme, že chodník je mokrý. Divíme se proč? Dále zvažujeme všechny možné příčiny mokrosti chodníku, kde jsou zahrnuty následující případy: nedávný déšť, průsak podzemní vody, nedávné zavlažování trávníku, prasklé potrubí, nápoj rozlitý kolemjdoucím atd. Po předchozí úvaze přiřadíme všem uvažovaným možnostem nějakou míru věrohodnosti na základě vlastních předchozích znalostí, např.: nedávný déšť bude mít nejspíš větší pravděpodobnost než nápoj rozlitý kolemjdoucím. Tyto znalosti označujeme za tzv. apriorní znalosti.

Z tohoto jednoduchého příkladu plyne skutečnost, že jsme schopni odhadnout pravděpodobnost sledovaného jevu (příčina mokrosti chodníku) takovým způsobem, že zohledníme své předešlé zkušenosti s tímto problémem a nové zkušenosti získané bezprostředním pozorováním.

Bayesova věta

Následující kapitola představuje pojmy přímo související s Bayesovou větou. Definice, věty a důkazy jsou čerpány z [4] a [2]. Pro přiblížení vztahů jsou uvedeny vlastní příklady.

Definice 1. *Nechť je dán pravděpodobností prostor* $(\Omega, \mathscr{A}, P)$ *a náhodný jev* $B \in \mathscr{A}, P(B) > 0$. *Potom funkce* $P(\cdot|B)$ *definovaná na* $\mathscr{A}$ *předpisem*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, A \in \mathscr{A}, \tag{1}$$

se nazývá pravděpodobnost jevu A podmíněná jevem B.

Příklad 1. *Ze 150 kuliček je 25 modrých. Náhodně vybereme 2. Jaká je pravděpodobnost jevu A, že budou obě modré?*

Řešení: Úlohu vyřešíme pomocí definice podmíněné pravděpodobnosti. Označme jev B, že v prvním tahu vybereme modrou kuličku, což odpovídá pravděpodobnosti $P(B) = \frac{25}{150}$. *Při výběru druhé modré kuličky je pravděpodobnost takového výběru rovna* $P(A|B) = \frac{24}{149}$. *Úpravou výše definovaného vztahu 1 často počítáme pravděpodobnosti průniku náhodných jevů A a B. Vyjádříme-li tedy pravděpodobnost tohoto průniku a dále předpokládáme nezávislost pravděpodobností* $P(B)$ *a* $p(A|B)$. *Potom pravděpodobnost výběru dvojice modrých kuliček je rovna* $P(A) = P(A \cap B) = P(A|B) \cdot P(B) = \frac{24}{149} \cdot \frac{25}{150} = \frac{600}{22350} \doteq 0.026846$.

Věta 1 (o násobení pravděpodobností)**.** *Nechť je dán pravděpodobnostní prostor* $(\Omega, \mathscr{A}, P)$ *a jevy* $A \in \mathscr{A}$, $B \in \mathscr{A}$. *Je-li* $P(B) > 0$, *je* $P(A \cap B) = P(B) \cdot P(A|B)$, *je-li* $P(A) > 0$, *je* $P(A \cap B) = P(A) \cdot P(B|A)$. *Nechť dále jsou dány*

jevy $A_1, A_2, \cdots, A_n$, *kde* $A_i \in \mathscr{A}$ *pro* $i = 1, 2, \cdots, n$. *Je-li* $P(\bigcap_{j=1}^{n-1} A_j) > 0$, *potom platí*

$$P(\bigcap_{1}^{n} A_j) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \ldots P(A_n|\bigcap_{1}^{n-1} A_j).$$

Příklad 2. *V továrně dělá kontrolor namátkovou kontrolu kvality tak, že vybírá pět výrobků, přičemž každý výrobek vybírá zvlášť. Po vytažení je nevrací zpět. Výrobní proces projde kontrolou, jestliže se ve výběru neobjeví ani jeden zmetek. Výrobce ví, že v právě kontrolované várce 100 výrobků je 23 zmetků. Jaká je pravděpodobnost jevu B, že kontrolor vytáhne právě jeden zmetek?*

Řešení: Úlohu vyřešíme pomocí věty o násobení pravděpodobností. Označme $A_i, i = 1, 2, 3, 4, 5$, *kde v i-tém tahu vytáhneme zmetek. Pravděpodobnosti jevů* A_i *jsou rovny* $P(A_1) = \frac{23}{100}, P(A_2) = \frac{22}{99}, P(A_3) = \frac{21}{98}, P(A_4) = \frac{20}{97}, P(A_5) = \frac{19}{96}$. *Potom pravděpodobnost jevu B rovna* $P(B) = \prod_{i=1}^{5} P(A_i) = \frac{23}{100} \cdot \frac{22}{99} \cdot \frac{21}{98} \cdot \frac{20}{97} \cdot \frac{19}{96} = \frac{4037880}{9034502400} \doteq 0,0004469$.

Věta 2 (o úplné pravděpodobnosti)**.** *Nechť* (B_n) *je konečná nebo nekonečná posloupnost neslučitelných náhodných jevů, které mají kladné pravděpodobnosti a pro které je splněn vztah* $P(\bigcup_n B_n) = 1$. *Potom pro libovolný jev* $A \in \mathscr{A}$ *platí*

$$P(A) = \sum_n P(A|B_n)P(B_n).$$

Příklad 3. *V prodejně se prodávají olivy v plechovce dodávané čtyřmi různými dodavateli, označme je* a, b, c, d. *Prodávající ví, že sklenice dodavatele dodavatele a obsahují 65 % černých oliv, ze sklenic od dodavatele b je 75 % s černými olivami, ze sklenice od dodavatele c je 45 % s černými olivami a ze sklenic dodavatele d je 35 % s černými olivami. Ve skladě jsou sklenice náhodně uloženy bez ohledu na konkrétního dodavatele. Jaká je pravděpodobnost, že zákazník koupí sklenici černých oliv, je-li na skladě 18 % sklenic dodavatele a, 35 % sklenice dodavatele b, 20 % sklenic dodavatele c a 27 % sklenic dodavatele d?*

Řešení: úlohu vyřešíme pomocí věty o úplné pravděpodobnosti. Označme A, resp. B, resp. C, resp. D, náhodné jevy, že ve skladu náhodně vybereme skle-

nici dodavatele a, resp. dodavatele b, resp. dodavatele c, resp. dodavatele d. Dále označme E náhodný jev, že vybereme sklenici černých oliv. Pomocí předchozích označení určíme pravděpodobnosti $P(A) = 0,18; P(B) = 0,35; P(C) = 0,2; P(D) = 0,27$. *Podle věty o úplné pravděpodobnosti můžeme psát* $P(E) = P(E|A)P(A) + P(E|B)P(B) + P(E|C)P(C) + P(E|D)P(D) = 0,65 \cdot 0,18 + 0,75 \cdot 0,35 + 0,45 \cdot 0,2 + 0,35 \cdot 0,27 = 0,117 + 0,2625 + 0,09 + 0,0945 = 0,564$.

Poznámka 1. *Uvedené vztahy jsou klíčové pro formulaci Bayesovy věty.*

Věta 3 (Bayesova)**.** *Nechť* (B_n) *je konečná nebo nekonečná posloupnost neslučitelných náhodných jevů, které mají kladné pravděpodobnosti a pro které platí* $P(\bigcup_n B_n) = 1$. *Nechť dále* $P(A) > 0, A \in \mathscr{A}$. *Potom* $\forall n$ *platí*

$$P(B_n|A) = \frac{P(A|B_n)P(B_n)}{\sum_j P(A|B_j)P(B_j)}. \tag{2}$$

Důkaz. Nejprve na levou stranu použijeme definici podmíněné pravděpodobnosti a pak na čitatele $P(A \cap B_n)$ použijeme větu o násobení pravděpodobností a na jmenovatele $P(B_n)$ použijeme větu o úplné pravděpodobnosti

$$P(B_n|A) = \frac{P(A \cap B_n)}{P(B_n)} = \frac{P(A|B_n)P(B_n)}{\sum_j P(A|B_j)P(B_j)}. \qquad \square$$

Příklad 4. *Mějme 3 počtáře, kteří s pravděpodobnostmi 0,45; 0,7; 0,95 vypočítají zadaný příklad správně. Jaká je pravděpodobnost, že správně vyřešený příklad počítal první student?*

Řešení: označme $B_i, i = 1, 2, 3$ *náhodný jev, že počítal i-tý student. Potom můžeme vyjádřit pravděpodobnosti* $P(B_1) = P(B_2) = P(B_3) = \frac{1}{3}$. *Označme náhodný jev A, že vyřešený příklad je správně. Dále určíme podmíněnou pravděpodobnost, že příklad správně vyřešil první student, což je rovno* $P(A|B_1) = 0,45$; *resp. že příklad správně vyřešil druhý student* $P(A|B_2) = 0,7$; *resp. že příklad vyřešil třetí student* $P(A|B_3) = 0,95$. *Nyní už můžeme přejít k Bayesovu vzorci, kde počítáme pravděpodobnost jevu* B_1 *za podmínky jevu A* $P(B_1|A) =$

$$= \frac{P(A|B_1)P(B_1)}{P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+P(A|B_3)P(B_3)} = \frac{0,45 \cdot \frac{1}{3}}{0,45 \cdot \frac{1}{3} + 0,7 \cdot \frac{1}{3} + 0,95 \cdot \frac{1}{3}} = \frac{0,15}{0,7} \doteq 0,214286.$$

Vypočtená pravděpodobnost je poměrně nízká, což je způsobeno tím, že náš první student je nejhorší počtář z vybrané trojice studentů.

Pro úplnost zde uvedeme i Bayesovu větu pro spojitý náhodný vektor $\mathbf{Y}$.

Věta 4. *Nechť* $\mathbf{Y} = (Y_1, \ldots, Y_p)', \mathbf{Z} = (Z_{p+1}, \ldots, Z_n)', 1 \leq p < n, n \in \mathbb{N}$. *Podmíněná hustota* $s(\mathbf{y}|\mathbf{z})$ *vektoru* $\mathbf{Y}$ *při daném* $\mathbf{Z}$ *je rovna*

$$s(\mathbf{y}|\mathbf{z}) = \frac{r(\mathbf{y})q(\mathbf{z}|\mathbf{y})}{\int_{R_p} r(\mathbf{y})q(\mathbf{z}|\mathbf{y})d\lambda(\mathbf{y})}, pro \int_{R_p} r(\mathbf{y})q(\mathbf{z}|\mathbf{y})d\lambda(\mathbf{y}) \neq 0. \tag{3}$$

V Bayesovské statistice je funkce r nazývá apriorní hustota. Podmíněná hustota q, která vyjadřuje závislost vektoru $\mathbf{Z}$ při daném vektoru $\mathbf{Y}$, se nazývá věrohodnostní funkce. Tyto pojmy jsou vysvětleny v následující kapitole.

Poznámka 2. *Právě použitý vzorec je pojmenován po teologu Thomasu Bayesovi. Thomas Bayes (1702-1761) byl anglický matematik a presbyteriánský kněz. Jeho slavný teorém byl zveřejněn po jeho smrti v roce 1763. Zásluhu na tomto publikování nese jeho přítel Richard Price, viz [7].*

Postup Bayesovské analýzy

Následující kapitola představí přístup Bayesovské analýzy v pěti krocích. První krok zahrnuje správnou interpretaci vybrané datové sady, která je důležitá pro další výpočty. Druhým krokem je definování vhodného modelu, který nejlépe vystihuje charakter datové sady. Třetí krok zahrnuje definování apriorního rozdělení pravděpodobností neznámého parametru. Čtvrtým krokem je určení posteriorního rozdělení pravděpodobností spojitého parametru. Pátým krokem je kontrola, zda posteriorní rozdělení pravděpodobností popisuje data s dostatečnou spolehlivostí. Veškeré informace jsou čerpány z [7], [2], [4], [5],[3] a [10]. Všechny uvedené grafy jsou vlastní, pokud není uvedeno jinak.

Binomické rozdělení

Mějme následující pokus: předpokládejme, že házíme n-krát jednou mincí a určíme, kolikrát padla hlava a orel. Pravděpodobnost, že padne hlava, je konstantní během celého pokusu. Možné počty hlav, které mohou nastat, pak jsou $0, 1, \ldots, n,$.

Nechť $X_1, X_2, \ldots, X_n$, kde $n \in \mathbb{N}$, je n-tice nezávislých náhodných veličin a parametr $\theta \in \Theta, \Theta = (0, 1)$. Dále nechť má každá náhodná veličina $X_i, \forall i \in \{1, 2, \ldots, n\}$, alternativní rozdělení. Binomické rozdělení je definováno následujícím způsobem: nechť $\theta \in (0, 1)$ a náhodná veličina $X = \sum_{j=1}^{n} X_j$ nabývá hodnot $0, 1, 2, \ldots, n$, s pravděpodobnostmi

$$P(X = k) = \binom{n}{k} \theta^k (1 - \theta)^{(n-k)}, \quad k = 0, 1, \ldots, n. \tag{4}$$

Dále budeme pracovat s podmíněnou pravděpodobnostní funkcí, která je de-

finována následujícím způsobem.

Definice 2. *Mějme diskrétní náhodné veličiny X, Y se sdruženou pravděpodobnostní funkcí $p(x_i, y_j) = P(X = x_i, Y = y_j)$, marginálními pravděpodobnostními funkcemi $p_X(x_i) = P(X = x_i)$ a $p_Y(y_j) = P(Y = y_j)$ a obory hodnot M_1, M_2. Pravděpodobnostní funkci*

$$h_x(x_i|y_j) = \frac{p(x_i, y_j)}{p_Y(y_j)}, \quad x_i \in M_1, y_j \in M_2 \tag{5}$$

pro $p_Y(y_j) \neq 0$ nazveme podmíněnou pravděpodobnostní funkcí veličiny X při daném $(Y = y_j)$.

Poznámka 3. *Alternativní (Bernoulliho) rozdělení je speciálním případem binomického rozdělení; náhodná veličina X nabývá pouze dvou hodnot 0,1 a $\theta \in (0, 1)$ je neznámý parametr. Jeho pravděpodobnostní funkce*

$$P(X = x_i|\theta) = \theta^{x_i}(1 - \theta)^{(1-x_i)}, \quad i \in 1, 2 \tag{6}$$

Vyjádřeme teď podmíněné binomické rozdělení pro náhodnou veličinu X, která nabývá hodnot $0, 1, \cdots, n$, při pevně zvoleném parametru θ

$$P(X = k|\theta) = \binom{n}{k}\theta^k(1 - \theta)^{(n-k)}, \quad k = 0, 1, \ldots, n. \tag{7}$$

Nyní předpokládejme, že jsme provedli N hodů mincí a známe počet padlých hlav. Neznámý parametr $\theta \in (0, 1)$ považujeme za náhodnou veličinu. Označme i-tý výsledek hodu mincí jako x_i a posloupnost výsledků $\{x_i\}, i = 1, 2, \ldots, N$. Předpokládáme, že výsledky libovolných dvou hodů jsou nezávislé. Dále označme počet padlých hlav $z = \sum_{i=1}^{N} x_i$ a počet rubů $N - z = \sum_{i=1}^{N}(1 - x_i)$. Potom rovnici (7) vyjádříme za předpokladu nezávislosti jednotlivých výsledků hodů

$$P(\{x_i\}|\theta) = \prod_{i=1}^{N} P(X = x_i|\theta) =$$

dále pro vyjádření $P(X = x_i|\theta)$ využijeme rovnici (6)

$$= \prod_{i=1}^{N} \theta^{x_i}(1 - \theta)^{1-x_i} =$$

využijeme pravidel pro práci s mocninami

$$= \theta^{\sum_i^N x_i}(1-\theta)^{\sum_i^N (1-x_i)}$$

a nakonec využijeme nově zavedené značení

$$= \theta^z(1-\theta)^{(N-z)} \tag{8}$$

Funkci (8) nazýváme věrohodnostní funkcí. Tato funkce obsahuje dostupné informace stanovené z dat. Uvedený tvar bude užitečný při odvození posteriorního rozdělení.

Apriorní rozdělení

Hledáme takovou hustotu rozdělení pravděpodobností f_x, která co nejlépe odráží naši apriorní informaci. Tomuto požadavku však odpovídá velké množství hustot. Pokud chceme použít Bayesovu větu, tak musíme své požadavky pro zavedení apriorní hustoty více upřesnit. Vyjdeme ze vztahu (2).

První požadovaná vlastnost se týká čitatele $P(A|B_n)P(B_n)$. Její podstatou je, že uvedené pravděpodobnosti by měly být stejnou formou jedné funkce. Toto omezení na výběr funkce je velmi důležité v případech, kdy konstruujeme posteriorní rozdělení přidáním nových informací z dat. Posteriorní rozdělení bude popsáno opět stejnou obdobou jedné funkce.

Druhá požadovaná vlastnost se týká jmenovatele $\sum_j P(A|B_j)P(B_j)$. Požadujeme, aby právě zmíněný výraz byl analyticky řešitelný. Tato vlastnost závisí na tom, v jakém vztahu jsou vůči sobě funkce $P(A|B_j)$ a $P(B_j)$ pro $j = 1, 2, \ldots, n$. S tímto vztahem souvisí pojem konjugovaný systém hustot, který je definován následujícím způsobem.

Definice 3. *Řekneme, že P je systém hustot konjugovaných s $f(x|\theta)$, pokud $\forall f(\theta) \in P$ je $f(\theta|x) \in P$ pro skoro všechna x.*

Poznámka 4. *Definovaný systém je velmi užitečný tehdy, pracujeme-li s rozumně velkými třídami hustot. Práce s konjugovanými rozděleními má velkou výhodu*

v tom, že při přechodu od apriorního k aposteriorního rozdělení pravděpodobností se jen adekvátně změní parametry. To znamená, že tzv. aktualizace rozdělení pravděpodobností parametru θ je snadná.

Beta rozdělení

Vysloveným požadavkům odpovídá beta rozdělení $beta(\theta|a,b)$, jehož hustota má pro dané parametry $a > 0, b > 0$, následující podobu

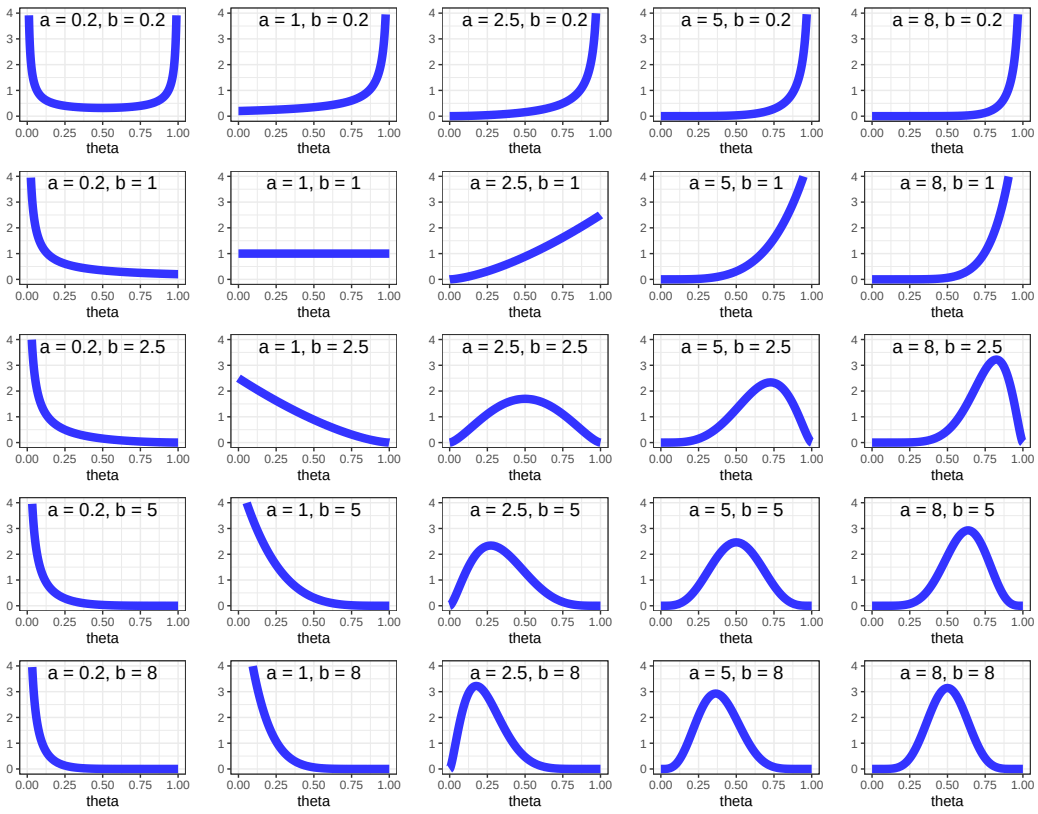
$$beta(\theta|a,b) = f(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \theta \in (0,1), \tag{9}$$

kde funkce $B(a,b)$ se nazývá beta funkce, která patří mezi tzv. speciální funkce, a je pro dané parametry $a > 0, b > 0$, definována

$$B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\mathrm{d}\theta, \tag{10}$$

a tato hodnota $B(a,b)$ ve vztahu (9) má charakter normalizační konstanty, díky níž je integrál s mezemi od 0 do 1 ze vztahu (9) roven 1.

Hustoty beta rozdělení mají odlišné průběhy na základě různých kombinací parametrů $a > 0, b > 0$, jak ukazuje následující obrázek 1.



Obrázek 1: Příklady hustot beta rozdělení pro různé hodnoty kladných parametrů a, b. Hodnota parametru a roste zleva doprava a hodnota parametru b roste shora dolů. Z grafu je patrné, že hustoty ležící na hlavní diagonále grafu jsou symetrické, což je dáno shodnou hodnotou obou parametrů a, b. Podíváme-li se na průběhy takových hustot, kde $\{a, b \in (0, \infty) : a \neq b, a \leq 1 \Rightarrow b \geq 1; b \leq 1 \Rightarrow a \geq 1\}$, tak tyto hustoty jsou prosté funkce. A nakonec hustoty, jejichž oba parametry nabývají hodnoty větší než 1, jsou různě zešikmené.

Stanovení parametrů beta rozdělení

Následující podkapitola popisuje určení parametrů a a b, které zcela přesně definují tvar hustoty beta rozdělení a zároveň nejlépe vystihují apriorní znalost o neznámém parametru θ. Učiníme předpoklad, že parametr a představuje počet padlých hlav a parametr b určuje padlých rubů a dále $n = a + b$, kde n je počet

hodů mincí.

Předpokládejme, že mince, kterou házíme, je férová, ale nejsme si tím jisti. Představme si, že jsme mincí hodili desetkrát, a výsledek byl 5 hlav a 5 rubů. Chceme tento výsledek vyjádřit pomocí hustoty beta rozdělení. Tato situace je znázorněna v předchozím obrázku 1 na pozici $(4, 4)$. Tato hustota dosahuje svého maxima v bodě $\theta = 0.5$. Nejistotu prvotního odhadu vyjadřují nižší a vyšší hodnoty kolem maxima, které jsou méně pravděpodobné. Tento příklad řeší velmi jednoduchý případ. Nyní představíme univerzální způsob výpočtu parametrů a, b, který se užívá při velkých hodnotách n.

Je známo, že střední hodnota μ a modus ω beta rozdělení pro parametry $a > 1, b > 1$ jsou definovány následujícími vztahy

$$\mu = \frac{a}{a+b} \qquad\qquad \omega = \frac{a-1}{a+b-2}. \tag{11}$$

Poznámka 5. *Z teorie pravděpodobnosti víme, že výběrový průměr je nestranným odhadem střední hodnoty a v tomto smyslu budeme tento pojem dále používat.*

Jestliže oba parametry a, b nabývají stejných hodnot, tak střední hodnota i modus se rovnají hodnotě 0.5. Jestliže platí $a < b$, tak obě číselné charakteristiky μ a ω jsou menší než 0.5. Analogicky při opačné nerovnosti $a > b$ jsou zmíněné charakteristiky větší než 0.5. Vzhledem k uvažovaným hodnotám parametrů a, b je zřejmé, že tvar hustoty bude obdobný jako například grafy na pozici (5,3) a (3,5) v obrázku 1.

Šíře takové hustoty závisí na číslu κ, které je určeno pro tyto jednodušší varianty vztahem $\kappa = a + b$ a vyjadřuje míru nejistoty pro odhad neznámého parametru $\hat{\theta}$. Podíváme-li se na obrázek 1 na pozici (5,3), tak tato hustota je užší a více koncentrovaná.

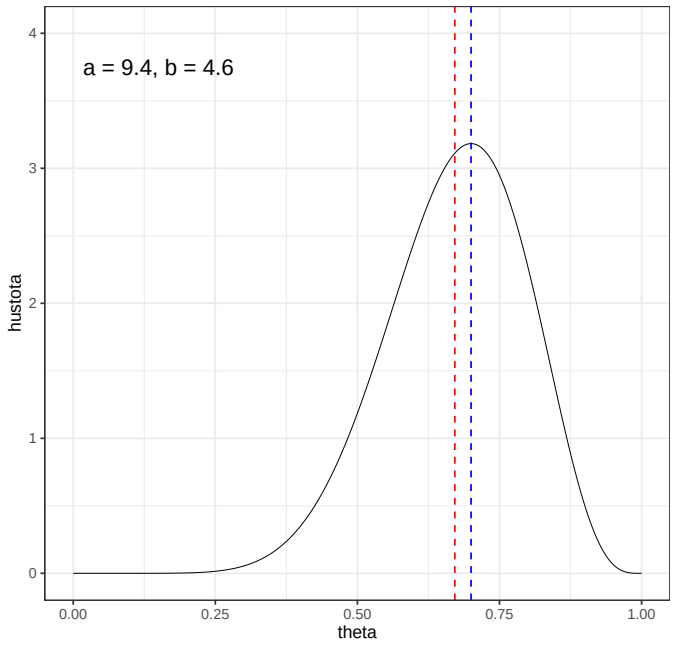
Řešení následujících rovnic pro výpočet parametrů a, b pomocí charakteristik μ, ω a vhodně zvolenou hodnotou κ:

$$a = \mu\kappa \qquad b = (1-\mu)\kappa \qquad \kappa = \{1, 2\} \tag{12}$$

$$a = \omega(\kappa - 2) + 1 \qquad b = (1-\omega)(\kappa - 2) + 1 \qquad \kappa > 2 \tag{13}$$

Jediným zbývajícím problémem zůstává, jak zvolit apriorní hodnotu κ. Hodnotu vybereme následujícím způsobem. Určíme počet nových hodů mincí, které bychom potřebovali k vyrovnání rozdílu mezi novou informací z dat a apriorní znalostí o hodnotě μ. Pokud bychom potřebovali malý počet nových hodů k přehodnocení apriorní domněnky, potom by apriorní rozdělení mělo být reprezentováno malou hodnotou κ. Jestliže bychom potřebovali naopak velký počet nových hodů k přehodnocení apriorní domněnky, potom apriorní rozdělení mělo být určeno velkou hodnotou κ.

Po vykreslení hustoty beta rozdělení s takto vypočtenými parametry a, b, se naskýtá otázka, která ze dvou charakteristik, modus a výběrový průměr, má pro nás větší vypovídající hodnotu. Odpověď můžeme vyčíst z následujícího obrázku 2. Graf je vykreslen na základě předem určených hodnot modu $\omega = 0.7$ a $\kappa = 14$, kde parametry a, b jsou vypočteny podle vzorců (13).



Obrázek 2: Hustota beta rozdělení s hodnotou modu $\omega = 0.7$; modus je v grafu vyznačen modrou vertikální čarou. Střední hodnota $\mu = 0.67$ je vyznačena červenou vertikální čarou.

Z grafu je patrné, že hodnota modu je více intuitivní než výběrový průměr,

protože právě v hodnotě modu dosahuje hustota svého maxima. Vysvětlení, proč výběrový průměr a modus nabývají různých hodnot, tkví v asymetrii vykreslené hustoty. Výběrový průměr je vzdálen od modu ve směru delšího chvostu hustoty.

Právě popsaný způsob výpočtu parametrů a, b je jeden z nejpoužívanějších metod výpočtu. Následující kapitola se věnuje posteriornímu rozdělení pravděpodobností.

Posteriorní rozdělení

Posteriorní hustota vyjadřuje „aktualizovanou“ domněnku o neznámém parametru θ, kde využíváme informaci z předchozího experimentu. Předpokládejme, že jsme provedli následující experiment: nechť N je počet hodů mincí a z je počet padlých hlav. Při následujícím odvození využijeme apriorní hustotu (9) a věrohodnostní funkci (8).

Cílem je určit podmíněnou hustotu $f(\theta|N,z)$, k čemuž nejprve využijeme Bayesův vzorec (3)

$$f(\theta|N,z) = \frac{f(N,z|\theta)f(\theta)}{\int_0^1 f(N,z|\theta)f(\theta)d(\theta)} =$$

dále využijeme zmíněné vztahy (9) a (8)

$$= \frac{\frac{\theta^z(1-\theta)^{(N-z)}\theta^{(a-1)}(1-\theta)^{(b-1)}}{B(a,b)}}{\frac{\int_0^1 \theta^z(1-\theta)^{(N-z)}\theta^{(a-1)}(1-\theta)^{(b-1)}d(\theta)}{B(a,b)}} =$$

funkční hodnoty beta funkce $B(a,b)$ jsou stejné v obou jmenovatelích složeného zlomku, takže je lze zkrátit

$$= \frac{\theta^z(1-\theta)^{(N-z)}\theta^{(a-1)}(1-\theta)^{(b-1)}}{\int_0^1 \theta^z(1-\theta)^{(N-z)}\theta^{(a-1)}(1-\theta)^{(b-1)}d(\theta)} =$$

využijeme pravidel pro práci s mocninami o stejném základu

$$= \frac{\theta^{(z+a-1)}(1-\theta)^{(N+b-z-1)}}{\int_0^1 \theta^{(z+a-1)}(1-\theta)^{(N+b-z-1)}d(\theta)} =$$

nyní zbývá upravit jmenovatele do jednoduššího tvaru, kde využijeme podobu beta funkce (10)

$$= \frac{\theta^{(z+a-1)}(1-\theta)^{(N+b-z-1)}}{B(a+z-1, N+b-z-1).} \tag{14}$$

Odvození ukazuje následující důležitou informaci: Je-li naše apriorní informace vyjádřena beta rozdělením (9) a víme-li z dat, že při N hodech padlo z hlav, tak potom posteriorní rozdělení je opět beta rozdělení tvaru (14). Vidíme, že aktualizace apriorní informace je velice jednoduchá, jak ukazuje následující příklad.

Příklad 5. *Předpokládejme, že apriornímu beta rozdělení přísluší hustota* $beta(\theta|1,1)$*, která je v obrázku 1 na pozici* $(2,2)$ *a odpovídá situaci, kdy jsme dvakrát hodili mincí a jednou padla hlava. Předpokládejme, že jsme hodili další čtyři hlavy, potom posteriorní hustota má podobu* $beta(\theta|5,1)$ *a v obrázku 1 ji najdeme na pozici* $(2,4)$*. Dále předpokládejme, že jsme sedmkrát po sobě hodili rub, potom naše nová posteriorní hustota má podobu* $beta(\theta|5,8)$ *a tuto hustotu najdeme v obrázku 1 na pozici* $(5,4)$*.*

Číselné charakteristiky posteriorního beta rozdělení

Následující podkapitola definuje posteriorní číselné charakteristiky posteriorní modus a posteriorní výběrový průměr.

Pro úplnost uvedeme vztahy (13) pro apriorní rozdělení, z nichž budeme vycházet

$$\mu = \frac{a}{a+b} \qquad\qquad \omega = \frac{a-1}{a+b-2}.$$

Dále předpokládejme, že jsme provedli N hodů mincí a označme počet hlav z; zlomek $\frac{z}{N}$ vyjadřuje poměr počtu hlav ve všech hodech. Posteriorní průměr je potom ve tvaru

$$m' = \frac{a+z}{(a+z)+(N-z+b)} = \frac{a+z}{N+b+a}. \tag{15}$$

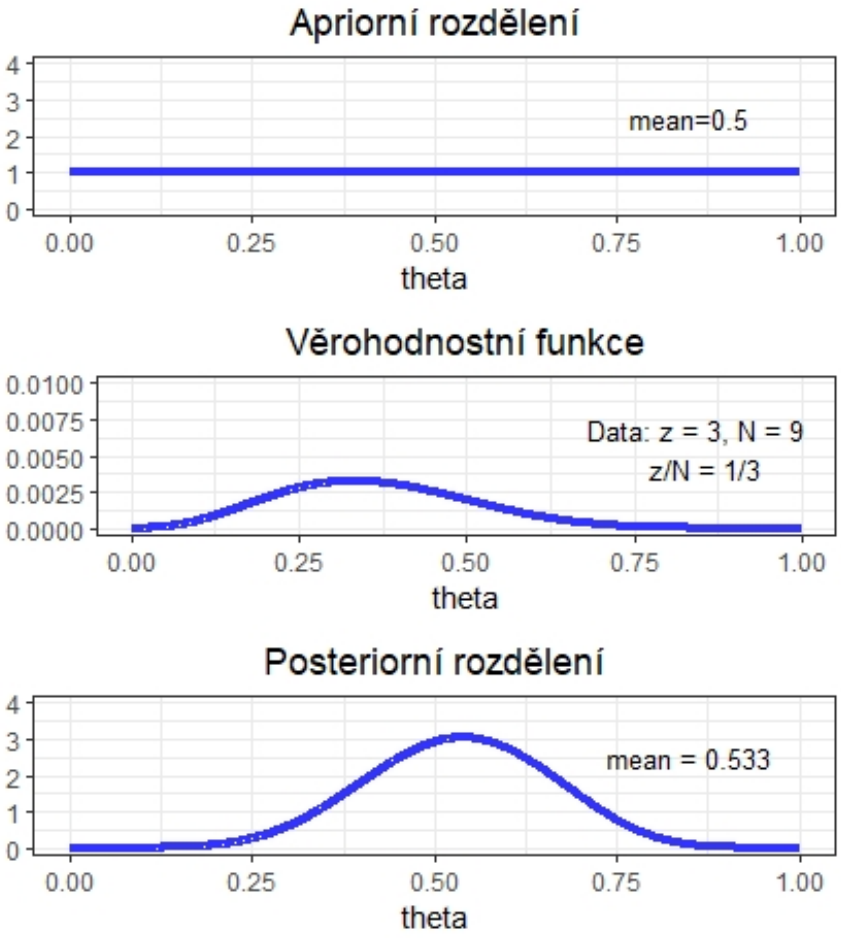
Takto vypočtenou hodnotu považujeme za Bayesovský odhad neznámého parametru θ. „Aktualizované“ parametry a', b' můžeme vyčíst ze vztahu (15)

$$a' = a + z \qquad\qquad b' = N + b - z. \tag{16}$$

Posteriorní průměr m' můžeme také vyjádřit jako vážený průměr apriorního výběrového průměru μ a poměru počtu hlav $\frac{z}{N}$, jak ukazuje následující vztah

$$\frac{a+z}{N+b+a} = \frac{z}{N}\frac{N}{N+a+b} + \frac{a}{a+b}\frac{a+b}{N+a+b} \tag{17}$$

Poměry $\frac{N}{N+a+b}, \frac{a+b}{N+a+b}$ vyjadřují příslušné váhy. Z rovnice (17) je zřejmé, že vliv apriorní informace klesá se zvyšujícím se počtem dat. Posteriorní průměr se tedy se zvyšujícím se počtem dat blíží poměru $\frac{z}{N}$. Tuto vlastnost vidíme na následující trojici grafů, kde ilustrujeme situaci z příkladu 5, kde jako apriorní rozdělení bereme $beta(\theta|1,1)$.



Obrázek 3: Grafické zobrazení rovnice (17), které zobrazuje posteriorní rozdělení jako kombinaci věrodnostní funkce (8) a příslušného apriorního rozdělení.

Intervaly spolehlivosti

V následující kapitole se budeme zabývat intervaly spolehlivosti, zejména pak kredibilními intervaly, které se používají v Bayesovské statistice. Představíme dvě metody výpočtu, jejichž pomocí kredibilní intervaly sestrojíme. Informace jsou čerpány z [7], [4],[9] a [1]. Všechny grafy jsou vlastní.

Vlastnosti intervalů spolehlivosti

Obecná definice intervalu spolehlivosti má následující podobu.

Definice 4. *Nechť* $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$. *Uspořádanou dvojici výběrových funkcí* $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ *nazýváme oboustranný intervalový odhad parametru* θ *(konfidenční interval, interval spolehlivosti), platí-li*

$$P_\theta[T_1(X_1, \ldots, X_n) \leq \theta \leq T_2(X_1, \ldots, X_n)] \geq 1 - \alpha, \forall \theta \in \Theta \tag{18}$$

Číslo 1-$\alpha, \alpha \in (0, 1)$, *se nazývá spolehlivost odhadu parametru* θ.

Důležitá je správná interpretace. Interval spolehlivosti je náhodný interval, který odhadovanou hodnotu θ obsahuje nebo ne. Ovšem pravděpodobnost, že interval odhadovanou hodnotu skutečně pokrývá, je $1 - \alpha$ (předepsaně velká).

Interval spolehlivosti je důležitým pojmem ve statistice, protože vyjadřuje nejistotu bodového odhadu θ. Šíře intervalu ovlivňují zejména tři faktory.

Prvním aspektem je rozsah náhodného výběru N. Předpokládejme, že pravděpodobnost, že padne líc, je $0.4 = \frac{40}{100} = \frac{400}{1000}$. Je velký rozdíl, jestli jsme měli 10 hodů nebo 1000. Pokud jsme měli k dispozici 1000 hodů, tak je naše nejistota mnohem menší než při 10 hodech. S rostoucím rozsahem výběru N by tedy měl

být interval kratší. Druhým aspektem je spolehlivost odhadu $1-\alpha, \alpha \in (0,1)$. Požadujeme-li spolehlivost odhadu 0.99, tak odpovídající interval bude mnohem delší, než když chceme spolehlivost menší, např.: 0.9. Třetím důležitým aspektem je rozptyl $var(X)$. Pokud je velký, budou odhady nepřesné a vypočtené intervaly široké.

Bayesovský přístup

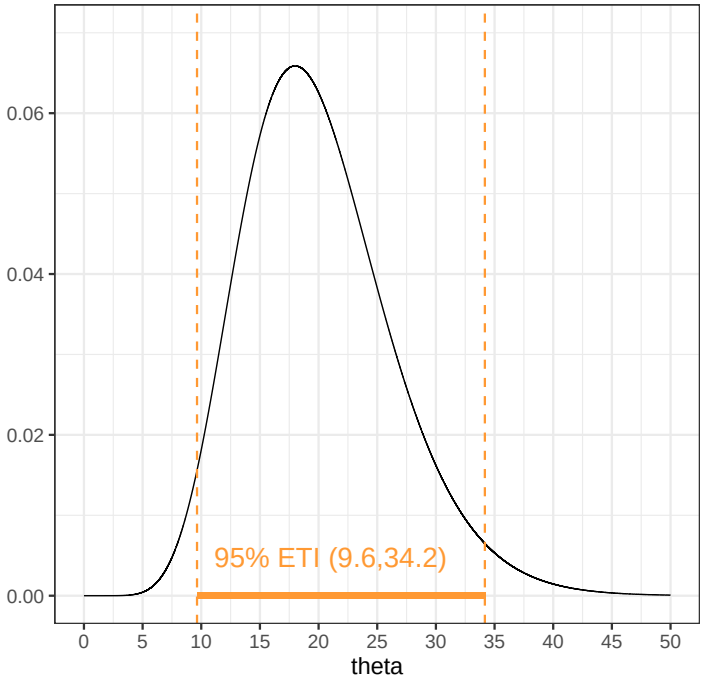
Mějme dané posteriorní rozdělení pravděpodobností neznámého parametru $\theta \in \Theta$. Rozsah hodnot, které mají vysokou posteriorní pravděpodobnost (požadujeme spolehlivost odhadu 1-$\alpha, \alpha \in (0,1)$), že obsahují neznámý parametr θ, označujeme Bayesovský kredibilní interval spolehlivosti. Existuje mnoho metod výpočtů oboustranných intervalů. Uvedeme dva nejznámější.

Interval ETI

Pro výpočet intervalu se stejně dlouhými chvosty (z angličtiny známý jako Equal-Tailed Interval, ETI), používáme kvantilovou metodu. Tato metoda spočívá ve vynechání 2.5 % z obou stran posteriorní hustoty rozdělení, jak vidíme na následující rovnici

$$P(\theta \leq \theta_d | a, b) = P(\theta \geq \theta_h | a, b) = \frac{\alpha}{2}, \tag{19}$$

kde θ_d a θ_h představují po řadě dolní intervalový odhad a horní intervalový odhad parametru θ. Tato metoda se často používá pro svou výpočetní jednoduchost. Odhady hranic tohoto intervalu byly vypočteny pomocí funkce $eti()$ ve statistickém softwaru R, která náleží do balíčku bayestestR. Na následujícím grafu je příklad tohoto intervalu pro asymetrické rozdělení.



Obrázek 4: Hustota $\chi^2(20)$. Čerchované vertikální čáry znázorňují hranice nalezeného kredibilního intervalu.

Interval HDI

Druhý často používaný interval spolehlivosti je HDI (z anglického highest density interval). Tento interval zahrnuje nejvíce věrohodné hodnoty neznámého parametru θ. Jeho rozsah pokrývá většinu hodnot rozdělení (nejčastěji 95 %), což znamená, že každá hodnota uvnitř intervalu má větší věrohodnost než jakákoli hodnota vně intervalu.

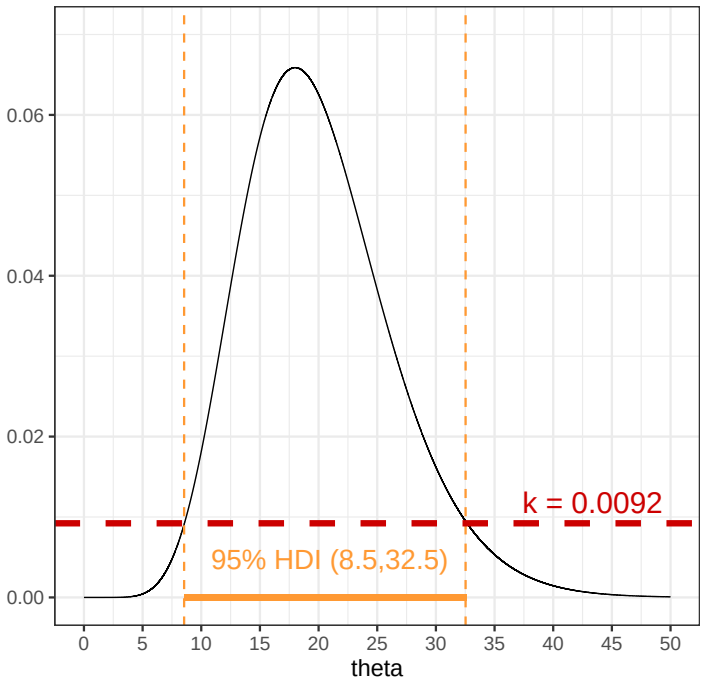
Uvažujme hustotu posteriorního rozdělení$f(\theta|N, z)$ definovanou vztahem (14). Označme C_k množinu hodnot parametru θ

$$C_k = \{\theta : f(\theta|N, z) \geq k\}, k \in \mathbf{R} \tag{20}$$

a pro konstantu k navíc předpokládáme platnost následující rovnosti

$$\int_{\theta:f(\theta|N,z)\geq k} f(\theta|N, z)d(\theta) = 1 - \alpha. \tag{21}$$

Uvedeným postupem získáme hodnotu konstanty k, která posteriorní hustotu protne ve dvou bodech. Poté výpočtem hodnot posteriorní hustoty v obou průsečících získáme x-ové souřadnice takových hodnot parametru θ, které odpovídají hledaným hranicím kredibilního intervalu. Nalezení 95% intervalu HDI můžeme vidět na následujícím obrázku 5.



Obrázek 5: Hustota $\chi^2(20)$. Čerchované vertikální čáry znázorňují vypočtené hranice hledaného intervalu. Červená čerchovaná horizontální přímka znázorňuje konstantu k, vypočtenou způsobem uvedeným výše.

Poznámka 6. *Obě uvedené metody dávají stejné hranice intervalů v případě, že je počítáme ze symetrického rozdělení. Kredibilní interval HDI je více užívaný, protože jeho rozsah je podmíněn vysokou věrohodností každé hodnoty neznámého parametru θ. Kvantilová metoda při výpočtu kredibilního intervalu ETI tuto podmínku nezohledňuje.*

Aplikace postupu na lékařská data

V této kapitole jsme aplikovali právě uvedený postup lékařské záznamy. Před samotnou analýzou jsme si museli data vhodně upravit.

Lékařská data zahrnovala dvě velké tabulky. Jedna obsahovala informace o všech narozených dětech na území ČR. Jednotlivé sloupce odpovídaly rokům 1961-2011, časovému intervalu, kdy byla data sbírána. Jednotlivé řádky odpovídaly věku matky při porodu dítěte. Věk matky se pohyboval od 14 do 50 let. Druhá tabulka vypadala obdobně, jen v ní byly zachyceny diagnostikované případy Downova syndromu.

Vybrali jsme ženy v konkrétním věku, kterým jsme se dále věnovali. Ve zdroji číslo [8] jsme se zaměřili na tabulku číslo 1, která obsahovala vybrané věky rodiček a příslušné pravděpodobnosti výskytu trizomie 21. Zvolili věk matky 25 let, takže v lékařských datech jsme se zabývali v obou tabulkách jen takovým řádkem, který obsahoval počty narozených dětí, jejichž matkám bylo právě 25 let. Ze zmíněného zdroje [8] jsme ještě využili prevalenci $\frac{1}{1350}$ (vyjadřuje podíl počtu nemocných v celkové populaci), která odpovídala našemu apriornímu bodovému odhadu neznámého parametru $\theta, \theta \in \Theta = (0, 0.5)$. Volba vhodného parametrického prostoru se vztahovala k hodnotě prevalence, která byla velmi malá. Na tomto intervalu jsme vytvořili ekvidistantní síť bodů, jejichž počet byl zvolen 10 000.

Vytvořili jsme tabulku o rozměrech 51x2, kde každý z dvaapadesáti řádků odpovídal příslušnému roku. V každém řádku se nacházely dva údaje: počet všech narozených dětí a počet diagnostikovaných případů Downova syndromu. Počtem

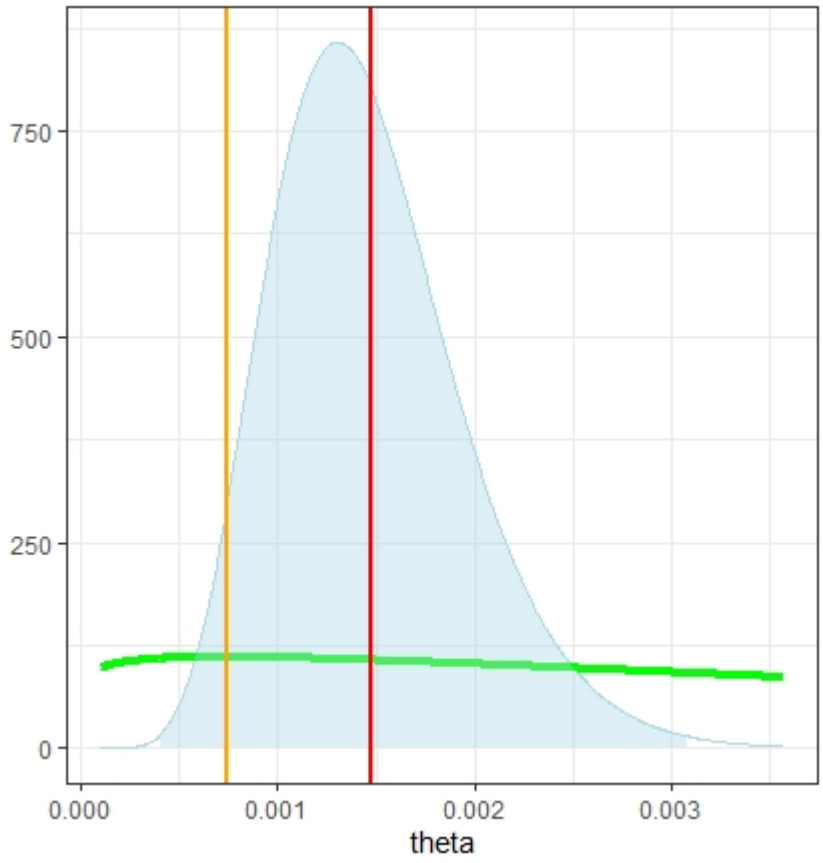
diagnostikovaných případů se rozumí součet případů prenatálně i postnatálně diagnostikovaných v příslušném roce. Prenatální diagnostika se rutinně používá od roku 1987.

Označili jsme X jako náhodnou veličinu, která nabývala hodnot

$$X = \begin{cases} 1, & \text{dítě má diagnostikován Downův syndrom} \\ 0, & \text{dítě nemá diagnostikován Downův syndrom} \end{cases}$$

Označili jsme N počet všech narozených dětí v roce 1961 a z počet dětí postižených Downovým syndromem v tomtéž roce. Dále jsme označili prevalenci $p = \frac{1}{1350}$. Hodnotu míry nejistoty κ jsme zvolili 150.

Vypočetli jsme věrohodnostní hodnoty pro každý bod z intervalu $(0, 0.5)$ pomocí věrohodnostní funkce (8). Dále jsme vypočetli první dvojici parametrů a, b pomocí vztahů (13). Takto získané parametry jsme použili pro výpočet hodnot hustoty apriorního beta rozdělení (9). Následně jsme provedli „aktualizaci“ parametru a, b podle vztahů (16) a pro tyto nové parametry a', b' jsme vypočetli hodnoty hustoty posteriorního beta rozdělení (14) a příslušný posteriorní průměr (odpovídá „aktualizované“ prevalenci) (15). Pro posteriorní rozdělení jsme na závěr vypočetli kredibilní interval spolehlivosti HDI, kterému jsme se věnovali na straně 28. První iterace byla vykreslena do následujícího obrázku 8.

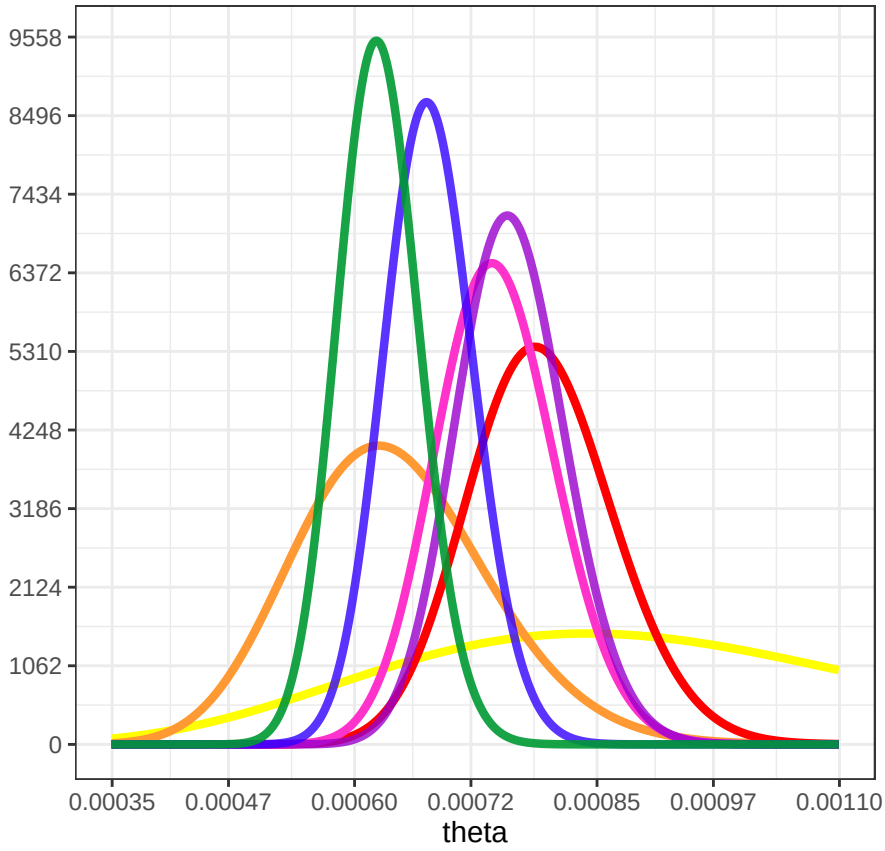


Obrázek 6: Na tomto obrázku je první krok „aktualizace“ odhadu parametru θ. Zelená křivka představuje hustotu apriorního rozdělení $beta(1.10963, 148.8904)$. Modrá křivka znázorňuje posteriorní beta rozdělení $beta(9.10963, 6171.89)$. Modrá plocha pod křivkou znázorňuje 95 % interval HDI. Červená vertikální čára znázorňuje posteriorní průměr (prevalenci) $m' = 0.00147$ a oranžová vertikální čára znázorňuje apriorní prevalenci $\mu = 0.00074$.

Z právě uvedeného grafu je patrné, že naše nejistota byla velmi vysoká, což se dalo po první iteraci očekávat.

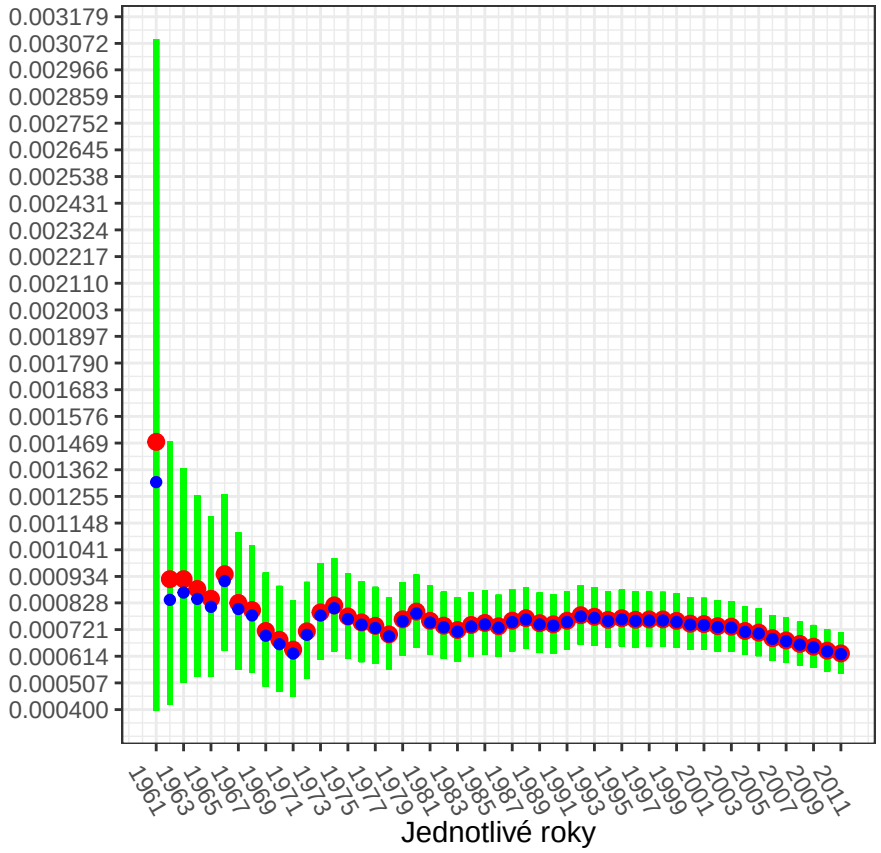
V druhé iteraci jsme přistoupili k novým parametrům a', b' jako k parametrům apriorního rozdělení a dále pro výpočet nových posteriorních parametrů a', b' vztahy (16). Tímto způsobem jsme vypočetli všech 51 posteriorních hustot.

Pro zevrubnou představu o vývoji rizika byl sestrojen graf na obrázku 7.



Obrázek 7: Na tomto obrázku jsou zobrazeny vybrané posteriorní hustoty, které znázorňují hrubý vývoj odhadu rizika výskytu Downova syndromu. Žlutá křivka zobrazuje hustotu ve druhé iteraci, oranžová znázorňuje hustotu v jedenácté iteraci, červená znázorňuje hustotu ve dvacáté iteraci, růžová znázorňuje hustotu ve devětadvacáté iteraci, fialová znázorňuje hustotu v osmatřicáté iteraci, modrá znázorňuje hustotu v sedmačtyřicáté iteraci a zelená znázorňuje poslední jednapadesátou iteraci. Z grafu je patrné, že nejistota odhadu postupně klesá se zvyšujícím se počtem iterací.

Výsledky všech výpočtů byly znázorněny v následujícím obrázku.



Obrázek 8: Postupné představy o pravděpodobnosti výskytu Downova syndromu v ČR v letech od 1961 do 2011. Červené body znázorňují posteriorní průměry a modré body znázorňují posteriorní mody ve všech iteracích. Svislé zelené čáry znázorňují délku HDI intervalu v příslušné iteraci. V grafu je patrný mírně klesající trend. Bayesovský odhad pravděpodobnosti výskytu Downova syndromu pro rok 2011 nabývá hodnoty 0.0006252817 a jeho příslušný interval spolehlivosti $(0.0005438044, 0.0007083208)$.

Závěr

V této práce jsme se zabývali odhadem pravděpodobnosti výskytu Downova syndromu v ČR na základě dat z let 1961 až 2011.

Nejdříve jsme uvedli základní informace o Downovu syndromu. Přiblížili jsme myšlenku bayesovského přístupu ke statistice na příkladu ze života. Další kapitola se věnovala Bayesově větě, která hraje v bayesovské statistice zcela zásadní roli. Dále jsme popsali postup výpočtu odhadu za předpokladu, že pracujeme s diskrétní náhodnou veličinou. Ukázali jsme důležitost beta rozdělení. Vysvětlili jsme pojmy věrohodnostní funkce, apriorní rozdělení pravděpodobností a posteriorní rozdělení pravděpodobností. V poslední kapitole jsme popsali postup aplikovaný na lékařská data.

Cílem práce bylo stanovení odhadu pravděpodobnosti výskytu Downova syndromu v ČR, k čemuž došlo v poslední kapitole. Zjistili jsme, že pravděpodobnost výskytu Downova syndromu v případě pětadvacetileté matky je nejspíš velmi malá, tj. asi $\frac{1}{1598}$.

Literatura

[1] ALEXON-FISK, M. *Statistical methods in Data Science and Al: DAT405, Module 3-4.*Chalmers University of Technology Invoice Service SE-412 96 Gothenburg Sweden - Mathematical Sciences, 2019.

[2] ANDĚL, J. *Základy matematické statistiky.* Vyd. 3. Praha: Matfyzpress, 2011. ISBN 978-80-7378-162-0.

[3] BOLSTAD, W. M. *Introduction to Bayesian statistics.* 2nd ed. New Jersey: John Wiley, 2007. ISBN 978-0-470-14115-1.

[4] HRON, J., KUNDEROVÁ P., VENCÁLEK O. *Základy počtu pravděpodobnosti a metod matematické statistiky.* 3. přepracované vydání. Olomouc: Univerzita Palackého v Olomouci, 2018. ISBN 978-80-244-5398-9.

[5] HUŠKOVÁ, M. *Bayesovské metody.* Praha: Univerzita Karlova v Praze, 1985, 93 s.

[6] INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. Finishing the euchromatic sequence of the human genome. *Nature.* 2004, **431**,931-945. DOI:`https://doi.org/10.1038/nature03001`

[7] KRUSCHKE, J. D. Doing Bayesian data analysis. *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan.* Edition 2. Boston: Academic Press, 2015, s. 30. ISBN 9780124058880.

[8] LOUCKÝ, J. Prenatální screening screening vrozených vývojových vad plodu a biochemická diagnostika. In: *Ultrazvuk v porodnictví a gynekologii: portál Sekce ultrazvukové diagnostiky ČGPS ČLS JEP* [online].[cit. 2020-07-22]. Dostupné z:`http://www.gynultrazvuk.cz/data/novinky/15/dokumenty/raabe-16el-pln-c5-5-bez.pdf`

[9] MAKOWSKI,D. BEN-SHACHAR,M.S.,LÜDECKE,D. 2019.*bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. Journal of Open Source Software,* 4(40),1541. Dostupné z: `https://doi.org/10.21105/joss.01541`

[10] PROKEŠOVÁ, M. *Markov Chain Monte Carlo (MCMC) metody.* Univerzita Karlova Matematicko-fyzikální fakulta, Ke Karlovu 3, 121 16 Praha 2, 2016. Dostupné také z: `http://www.karlin.mff.cuni.cz/~prokesov/2016/mcmc_2.pdf`

[11] SELIKOWITZ,Mark. : *Downův syndrom: definice a příčiny, vývoj dítěte, výchova a vzdělání, dospělost.* Vyd. 2. Praha: Portál, 2011, s. 37-49. Rádci pro zdraví. ISBN 978-80-7367-882-1.