UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

BAKALÁŘSKÁ PRÁCE

Elementy bayesovské statistiky s R

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Tereza Löfflerová**
Studijní program: B1103 Aplikovaná matematika
Studijní obor Aplikovaná statistika
Forma studia: prezenční
Rok odevzdání: 2019

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Tereza Löfflerová

Název práce: Elementy bayesovské statistiky s R

Typ práce: BAKALÁŘSKÁ PRÁCE

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2019

Abstrakt: Tato práce se zabývá základními pojmy bayesovské statistiky s využitím statistického softwaru R. Bayesovská statistika je důležitou alternativou ke standardní (tzv. frekventistické) statistice. V práci se seznámíme se základními principy tohoto přístupu. Hlavním tématem je odhadování parametrů u binomického a normálního rozdělení pomocí této metodiky. Odhady parametrů pro reálná data jsou vypočteny pomocí statistického softwaru R.

Klíčová slova: Bayesova věta, apriorní rozdělení, aposteriorní rozdělení, konjugované rozdělení

Počet stran: 58

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Tereza Löfflerová

Title: Bayesian essentials with R

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2019

Abstract: This thesis aims to describe basics of Bayesian statistics with R. Bayesian statistics is an important alternative to the standard (so called frequentist) statistics. In the thesis basic principles of this approach are reviewed first. The main goal is then to estimate parameters of the binomial and normal distributions using the Bayesian metodology. Estimates of the parameters for real data are computed by statistical software R.

Key words: Bayes theorem, prior distribution, posterior distribution, conjugate prior

Number of pages: 58

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracovala samostatně pod vedením pana doc. RNDr. Karla Hrona, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Ráda bych poděkovala panu doc. RNDr. Karlu Hronovi, Ph.D., svému vedoucímu bakalářské práce, za rady, trpělivost a čas, který mi věnoval.

Úvod

Tato práce se zabývá základními pojmy bayesovské statistiky. Bayesovská statistika je jedním z odvětví statistiky, které je v dnešní době populární. Jak už z názvu vyplývá, tento přístup využívá hlavně Bayesovu větu. Nejistota je zde vyjádřena pomocí podmíněné pravděpodobnosti a při výpočtech vždy vycházíme z tzv. apriorního rozdělení. Lze tedy spojit více zdrojů informací (apriorní informaci a data samotná) o dané události v jednom výpočtu.

V této práci se seznámíme s pojmem *bayesovská statistika* a dozvíme se její výhody či nevýhody a proč je tak oblíbená. Připomeneme si základní tvar Bayesovy věty aplikované na konkrétním příkladu, podíváme se na obecný postup bayesovské inference a seznámíme se s novými pojmy, které se v této oblasti užívají, jako je například konjugované apriorní rozdělení.

Hlavní částí této práce jsou odhady parametrů vybraných rozdělení pravděpodobností. Budeme odhadovat parametr θ u binomického rozdělení $Bi(n, \theta)$ a paramer μ u normálního rozdělení $N(\mu, \sigma^2)$.

V kapitolách, kde se budeme věnovat daným rozdělením, si vždy uvedeme apriorní rozdělení odhadovaného parametru a poté odvodíme i aposteriorní rozdělení. Nabyté znalosti z této části si ukážeme na příkladech s reálnými daty, kde inferenci provedeme pomocí statistického softwaru R. Ten je užíván pro statistické a grafické analýzy, umožňuje manipulovat s daty, provádět různé výpočty nebo grafické výstupy. Jedná se jak o prostředí, tak i o programovací jazyk. V oboru se samozřejmě užívají i jiné softwarové nástroje, jako například Python. Seznámit se s R v kontextu bayesovské statistiky ale není ztráta času, protože je ve statistické komunitě oblíbený.

Kapitola 1

Seznámení s bayesovským přístupem

V první kapitole se seznámíme s bayesovskou metodikou, dozvíme se její výhody a nevýhody, proč je v dnešní době oblíbená a v čem se liší oproti klasickému přístupu. V této kapitole jsem čerpala ze zdrojů [3], [10], [11], [14], a [21].

Ve statistice můžeme využít dvou hlavních přístupů, klasické (frekventistické) statistiky nebo bayesovské statistiky. Bayesovský přístup užívá apriorní informaci o datech. Bayesova věta nám dovoluje využít oba zdroje informací (data i apriorní informaci) a potom se dále „učit“ z předchozí zkušenosti. Například aposteriorní pravděpodobnosti, resp. aposteriorní rozdělení, můžeme v dalším kroce využít jako apriorní. Bayesovský přístup je výbornou alternativou i v situacích, když si jiný přístup neumí poradit s malým rozsahem výběru.

Protože ve statistice pracujeme vždy s nejistotou, musíme ji umět vyjádřit i pro toto odvětví. K vyjádření nejistoty bayesovský přístup využívá podmíněnou pravděpodobnost, která se přibližuje k běžnému užití slova pravděpodobnost (slovo „pravděpodobnost“ u klasického přístupu je bráno jako objektivní vlastnost). Nejistotu lze vyjádřit přímo pomocí rozdělení pravděpodobností, což je pro nás výhodné. Rozdělení pravděpodobností nám říká, jaké jsou přijatelné hodnoty pro parametr rozdělení, který nás zajímá. Tedy parametr, který chceme odhadnout.

Můžeme samozřejmě najít i nevýhody tohoto přístupu. Bayesovská statistika, jak již bylo zmíněno výše, vždy požaduje znalost apriorního rozdělení. Musíme vzít v úvahu něco, co bylo ještě před tím, než jsme získali data. Apriorní rozdělení tak vlastně vyjadřuje naši nejistotu o určitém parametru. Ta může být buď objektivní nebo subjektivní. Když je apriorní informace k dispozici, bayesovský přístup nám ji umožní efektivně využít, což u tradičního přístupu často nejde. Apriorní informaci můžeme získat například ze zkušenosti. Zřejmě se ale nevyhneme určitému vlivu subjektivity. Za další nevýhodu můžeme považovat to, že výsledek analýzy může být komplexnější, než když použijeme tradiční přístup. Také výpočty jsou složitější a těžší.

Statistická inference nám pomáhá se rozhodnout, čemu věřit. Spoléháme přitom na data z pozorování. Na základě těchto dat se ale naše přesvědčení může změnit. Představme si například minci. Mince má dvě strany, tudíž věříme, že máme pravděpodobnost rovnu 1/2, že nám padne hlava. Ale co když z 10 pokusů padnou 3 hlavy?

V této oblasti statistiky je parametr, který nás zajímá, považován za náhodný a data jsou pevně stanovena. To je však naopak, než u klasického přístupu, kde parametr je pevně dán a data jsou brána jako náhodná.

Bayesovská statistika je v dnešní době populární. Bayesovské metody kombinují mnoho zdrojů informací v jednom modelu a pro výpočet či simulaci aposteriorního rozdělení se dá často využít technika Markovových řetězců Monte Carlo, čímž se vyhneme složitému odvozování. To nám umožňuje konstruovat sofistikované statistické modely, které budou odrážet spletitost parametrů, které budeme chtít odhadnout.

1.1. Bayesova věta

Mějme jevy A a B, které prezentují nějaké události. Podmíněnou pravděpodobnost[1]

jevu A za podmínky jevu B, jestliže $P(B) > 0$, definujeme vztahem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1.1}$$

V kontextu bayesovské statistiky chápeme prvky výše uvedeného vztahu následovně:

- $P(A)$ je marginální pravděpodobnost jevu A, neboli apriorní pravděpodobnost
- $P(A|B)$ je podmíněná pravděpodobnost jevu A za podmínky jevu B , neboli aposteriorní pravděpodobnost
- $P(B|A)$ je podmíněná pravděpodobnost jevu B za podmínky jevu A, která se nazývá věrohodnost
- $P(B)$ je marginální pravděpodobnost jevu B.

Většinou je vhodné pracovat s větou o úplné pravděpodobnosti, která má v nejjednodušším případě tvar

$$P(B) = P(A)P(B|A) + P(A^c)P(B|A^c).$$

Když na jmenovatele vztahu 1.1 aplikujeme větu o úplné pravděpodobnosti, dostaneme Bayesovu větu

$$P(A|B) = \frac{P(B|A)P(A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}.$$

Podívejme se na příklad, který byl převzatý z [17], str. 112 a 114.

Příklad 1 (Test HIV/AIDS): Budeme provádět testy na HIV/AIDS. Víme, že tento test má vysokou jak specificitu (pravděpodobnost, že test vyjde negativní, když je člověk zdravý), tak senzitivitu (pravděpodobnost, že test vyjde pozitivní, když je člověk nemocný). Senzitivita i specificita jsou rovny 0,99. To znamená, že test ze 100 nemocných lidí odhalí nemoc u 99 z nich a ze 100 zdravých určí (chybně) jednoho jako nemocného. Předpokládejme, že počet nakažených lidí

[1] Více o podmíněných hustotách a podmíněných rozděleních obecně je uvedeno ve skriptu [5].

v populaci je 0,4 % (prevalence nemoci). Otázka tedy zní, jaká je pravděpodobnost, že je člověk infikovaný, když mu test vyšel pozitivní?

Už ze zadání můžeme vyčíst mnoho informací. Označme jevy:

- $\oplus$ ... pozitivní výsledek testu
- $\ominus$ ... negativní výsledek testu (doplněk k $\oplus$)
- N ... nemocný člověk
- Z ... zdravý člověk (doplněk k N)

Dále víme:

- $P(\ominus|Z) = 0{,}99$
- $P(\oplus|N) = 0{,}99$
- $P(N) = 0{,}004$

- $P(\oplus|Z) = 1 - P(\ominus|Z) = 0{,}01$
- $P(Z) = 1 - P(N) = 0{,}996$

a zajímá nás $P(N|\oplus)$. Dosaďme do Bayesovy věty:

$$P(N|\oplus) = \frac{P(\oplus|N)P(N)}{P(\oplus|N)P(N) + P(\oplus|Z)P(Z)} = \frac{0{,}99 \times 0{,}004}{0{,}99 \times 0{,}004 + 0{,}01 \times 0{,}996} =$$

$$= 0{,}2845.$$

Odpověď na otázku tedy je, že když je test pozitivní, člověk je nemocný jen s pravděpodobností 28,45%. Přestože specificita a senzitivita testu jsou vysoké a test by se neměl moc mýlit, ovlivnila ho nízká prevalence, a tím nám vzniklo mnoho falešně pozitivních výsledků testu.

Výpočet v R je ovšem mnohem jednoduší. Potřebujeme k tomu pouze knihovnu *LaplacesDemon* [19].

```
library(LaplacesDemon)
aprior = c(0.004,1-0.004)
verohodnost = c(0.99,1-0.99)
asposterior = BayesTheorem(aprior,verohodnost)

> asposterior
[1] 0.2844828 0.7155172
attr(,"class")
[1] "bayestheorem"
```

První výsledek výstupu kódu je přesně situace, kterou jsme spočítali výše. Druhý výsledek výstupu je pro případ, kdy $P(Z|\oplus)$.

1.2. Postup bayesovské inference

Mějme datovou sadu $\boldsymbol{y}$ (proměnné označeny tučně vyjadřují náhodný výběr). Chceme udělat závěr o neznámé hodnotě θ, což může být například nějaký parametr rozdělení, chybějící data, prediktivní hodnoty a tak dále. Bayesovská statistika začíná, jako klasická statistická analýza, specifikací modelu.

Z bayesovského pohledu je θ neznámá, takže bychom měli mít rozdělení pravděpodobností, které zobrazuje naši nejistotu o hodnotách tohoto parametru, než jsme dostali data. To znamená specifikovat apriorní rozdělení vyjádřené hustotou $p(\theta)$.

Věrohodnostní funkce, symbolicky vyjádřena pomocí podmíněné hustoty $p(\boldsymbol{y}|\theta)$, dohromady s apriorním rozdělením, tvoří model úplné pravděpodobnosti

$$p(\boldsymbol{y}, \theta) = p(\boldsymbol{y}|\theta)p(\theta).$$

Když užijeme Bayesovu větu, dostaneme podmíněné rozdělení pravděpodobností pro parametr θ, který chceme odhadnout, vzhledem k datům $\boldsymbol{y}$,

$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{\int_{-\infty}^{\infty} p(\theta)p(\boldsymbol{y}|\theta)d\theta} \propto p(\theta)p(\boldsymbol{y}|\theta), \tag{1.2}$$

$p(\theta|\boldsymbol{y})$ se nazývá aposteriorní rozdělení pro θ. Symbol $\propto$ značí proporcionální vztahy, tedy rovnost až na násobek kladnou konstantou. Pro náhodný výběr $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ platí $p(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$.

1.3. Konjugované apriorní rozdělení

Výběr relevantního apriorního rozdělení je v bayesovské statistice důležitý. Jestliže je apriorní infomace o datech či modelu k dispozici, měla by být použita ke stanovení apriorního rozdělení. Protože výběr apriorního rozdělení má vliv na výsledek našich výpočtů, měl by být tento krok prováděn s určitou opatrností.

Z výpočtového hlediska je nejvhodnější apriorní rozdělení to, které nějakým způsobem napodobuje strukturu věrohodnosti, apriorní a aposteriorní rozdělení zůstanou ve stejné „rodině“ rozdělení. Takové apriorní rozdělení nazýváme konjugované k věrohodnosti.

Konjugované apriorní rozdělení existuje ovšem pouze pro málo typů věrohodností. V následující tabulce můžeme vidět příklady konjugovaných apriorních rozdělení.

Rozdělení y	Parametr	Konjugované apriorní rozdělení
binomické	pravděpodobnost úspěchu	beta
poissonovo	střední hodnota	gama
exponenciální	převrácená střední hodnota	gama
normální (σ^2 známé)	střední hodnota	normální
normální (μ známé)	rozptyl	inverzní gama

Tabulka 1.1: Vybraná konjugovaná apriorní rozdělení k věrohodnosti.

1.4. Inference pro binární data

Inferenci pro binární data můžeme krásně pochopit na „školním“ příkladu s mincí.

Příklad 2 (Hod mincí): Uvažujme, že v krabici jsou 3 mince. První je falešná, tedy pravděpodobnost, že padne hlava, je 0,25. Druhá mince je spravedlivá, to

znamená, že pravděpodobnost padnutí hlavy je stejná jako pravděpodobnost padnutí orla, 0,5. Třetí mince je zase falešná, ale tentokrát je pravděpodobnost padnutí hlavy 0,75. Náhodně vybereme minci a jednou hodíme, dostaneme hlavu. Jaká je pravděpodobnost, že jsme vybrali třetí minci?

Označme náhodnou veličinu y, která bude vyjadřovat počet úspěchů, tedy počet padnutí hlavy, na minci. Padnutí hlavy v pokuse odpovídá realizaci náhodné veličiny $y = 1$. Nechť θ udává pravděpodobnost padnutí hlavy: $\theta \in (0{,}25; 0{,}5; 0{,}75)$. Apriorní pravděpodobnost se pak rovná $p(\theta = 0{,}25) = p(\theta = 0{,}5) = p(\theta = 0{,}75) = 1/3$.

Tento výběr má binomické rozdělení s parametry 1 a θ, kde 1 je počet pokusů a $\theta \in (0, 1)$ je pravděpodobnost výběru mince,

$$y|\theta \sim Bi(1, \theta).$$

Speciálně bychom mohli hovořit i o alternativním rozdělení s parametrem θ. Věrohodnost, odpovídající vlastně pravděpodobnostní funkci y při daném θ, je ve tvaru

$$p(y|\theta) = \theta^y(1 - \theta)^{1-y}.$$

Výpočet aposteriorního rozdělení parametru θ není nijak složitý, pro všechny tři případy hodnoty θ by vypadal stejně, jen by se měnila věrohodnost. Ukážeme si tedy výpočet pro třetí minci, kdy $\theta = \theta_3 = 0{,}75$. Nejprve spočítáme pravděpodobnost $p(y = 1)$ z pravděpodobnostní funkce pomocí Bernoulliho schématu,

$$p(y = 1) = \binom{1}{1}\theta^1(1 - \theta)^0 = \theta^1 = \theta.$$

Teď můžeme přejít k výpočtu toho, co nás zajímá nejvíce, a to podmíněného rozdělení $p(\theta|y = 1)$,

$$p(\theta_3|y = 1) = \frac{p(y = 1|\theta_3)p(\theta_3)}{p(y = 1|\theta_3)p(\theta_3) + p(y = 1|\theta_2)p(\theta_2) + p(y = 1|\theta_1)p(\theta_1)} =$$

$$= \frac{\frac{3}{4}\frac{1}{3}}{\frac{3}{4}\frac{1}{3} + \frac{1}{4}\frac{1}{3} + \frac{1}{2}\frac{1}{3}} = \frac{1}{2},$$

kde $p(\theta_1)$ značí pravděpodobnost padnutí hlavy na první minci, $p(\theta_2)$ na druhé a $p(\theta_3)$ na třetí minci.

Výpočet v R provedeme obdobně jako u příkladu v podkapitole 1.1

```
library(LaplacesDemon)
apriorni_pst_theta = c(rep(1/3,3))
verohodnost_theta = c(1/4,1/2,3/4)
asposteriorni_pst_theta = BayesTheorem(apriorni_pst_theta,
      verohodnost_theta)

> asposteriorni_pst_theta
[1] 0.1666667 0.3333333 0.5000000
attr(,"class")
[1] "bayestheorem"
```

V následující tabulce můžete vidět výsledky pro všechny mince.

Mince	θ	$p(\theta)$	$p(y=1\|\theta)$	$p(\theta\|y=1)$
1	0,25	0,33	0,25	0,167
2	0,50	0,33	0,50	0,333
3	0,75	0,33	0,75	0,500
$\sum$	1,50	1,00	1,50	1,00

Z tabulky můžeme vidět, že pravděpodobnost, že jsme házeli první mincí, jestliže padla hlava, je 16,7% a pravděpodobnost, že jsme házeli druhou mincí, je přibližně 33,3%.

1.4.1. Aposteriorní prediktivní rozdělení

Prediktivní aposteriorní rozdělení pro nové pozorování y^* je dáno vztahem

$$p(y^*|\boldsymbol{y}) = \int_{-\infty}^{\infty} p(y^*|\boldsymbol{y},\theta)p(\theta|\boldsymbol{y})d\theta.$$

Za předpokladu, že minulé a budoucí pozorování jsou nezávislá na θ, uvedený vztah můžeme zjednodušit,

$$p(y^*|\boldsymbol{y}) = \int_{-\infty}^{\infty} p(y^*|\theta)p(\theta|\boldsymbol{y})d\theta.$$

Pro diskrétní případy θ jsou integrály nahrazeny součtem,

$$p(y^*|\boldsymbol{y}) = \sum_{\theta_i} p(y^*|\theta_i)p(\theta_i|\boldsymbol{y}), \tag{1.3}$$

kde $p(\theta_i|\boldsymbol{y})$ můžeme uvažovat jako „aposteriorní váhy“.

Vraťme se k Příkladu 2. Uvažujme, že chceme předpovědět pravděpodobnost toho, že v dalším hodu padne zase hlava. To jednoduše obdržíme pomocí vztahu 1.3.

$$p(y^* = 1|y = 1) = \sum_{i=1}^{3} \theta_i p(\theta_i|y = 1) =$$

$$= (0{,}25 \times 0{,}167) + (0{,}50 \times 0{,}333) + (0{,}75 \times 0{,}500) = 0{,}5833.$$

V R-ku spočítáme pomocí 2 řádků:

```
apr.pst = asposteriorni_pst_theta[1:3]
sum(verohodnost_theta*apr.pst)
```

```
> sum(verohodnost_theta*apr.pst)
[1] 0.5833333
```

1.4.2. Sekvenční učení

Mějme data $\boldsymbol{y}_1$ ve formě náhodného výběru. Z aposteriorního rozdělení $p(\theta|\boldsymbol{y}_1)$ uvažujeme další data $\boldsymbol{y}_2$. Aposteriorní rozdělení založené na $\boldsymbol{y}_1$ a $\boldsymbol{y}_2$ vypadá následovně,

$$p(\theta|\boldsymbol{y}_1, \boldsymbol{y}_2) \propto p(\boldsymbol{y}_2|\theta) \times p(\theta|\boldsymbol{y}_1).$$

Výsledné aposteriorní rozdělení je stejné, i když dostaneme data $\boldsymbol{y}_1$ a $\boldsymbol{y}_2$ současně,

$$p(\theta|\boldsymbol{y}_1, \boldsymbol{y}_2) \propto p(\boldsymbol{y}_1, \boldsymbol{y}_2|\theta) \times p(\theta).$$

Je důležité si uvědomit, že nynější aposteriorní rozdělení parametru θ je budoucí apriorní rozdělení tohoto parametru.

Navažme opět na Příklad 2. Teď budeme uvažovat, že jsme po datech $y_1 = 1$ pozorovali ještě data $y_2 = 1$ (tedy výsledkem hodu náhodně zvolenou mincí je opět hlava) a budeme pozorovat změnu naší důvěry. Výpočet provedeme zase pro třetí minci a budeme vycházet z následující tabulky.

Mince	θ	$p(\theta)$	$p(y = 1\|\theta)$
1	0,25	0,167	0,25
2	0,50	0,333	0,50
3	0,75	0,500	0,75
$\sum$	1,50	1,00	1,50

$$p(\theta_3|y_2 = 1) = \frac{p(y_2 = 1|\theta_3)p(\theta_3)}{p(y_2 = 1|\theta_3)p(\theta_3) + p(y_2 = 1|\theta_2)p(\theta_2) + p(y_2 = 1|\theta_1)p(\theta_1)} =$$

$$= \frac{0{,}75 \times 0{,}5}{0{,}75 \times 0{,}5 + 0{,}50 \times 0{,}33 + 0{,}25 \times 0{,}167} = 0{,}644.$$

Obdobně bychom dopočítali výsledky pro zbylé dvě mince. Ty vypočítáme ale pomocí jednoho jediného řádku v R, protože všechny hodnoty máme uložené z dřívějších výpočtů.

```
BayesTheorem(apr.pst,verohodnost_theta)
```

```
> BayesTheorem(apr.pst,verohodnost_theta)
[1] 0.07142857 0.28571429 0.64285714
attr(,"class")
[1] "bayestheorem"
```

Rozdíl, který nám vyšel mezi ručním výpočtem a výpočtem v R pro třetí minci může být zapříčiněn zaokrouhlováním při výpočtu. Není ovšem nijak veliký.

Původní tabulku můžeme tedy rozšířit o naše výsledky.

Mince	θ	$p(\theta)$	$p(y=1\|\theta)$	$p(\theta\|y=1)$
1	0,25	0,167	0,25	0,071
2	0,50	0,333	0,50	0,286
3	0,75	0,500	0,75	0,643
$\sum$	1,50	1,00	1,50	1,00

Po hození druhé hlavy je zde 64,3% pravděpodobnost, že jsme házeli třetí mincí a pouze 7,1% pravděpodobnost, že jsme vybrali první minci.

Kapitola 2

Bayesovské metody pro binomické rozdělení

Tato kapitola se věnuje bayesovským metodám tam, kde naše data mají binomické rozdělení. To znamená, že každé pozorování může skončit pouze jedním ze dvou výsledků. Hodnoty těchto pozorování jsou nominální. To znamená, že je nelze uspořádat a nemají ani žádný číselný význam. Parametr θ pak odpovídá pravděpodobosti výskytu daného jevu při realizaci náhodného pokusu, resp. jeho proporcí v dané populaci. Příkladem této situace je proporce narozených holčiček ze všech narozených dětí, proporce zmetků v rámci výrobků na výrobní lince nebo proporce pacientů po operaci srdce, kteří přežijí více než rok po operaci.

V této kapitole jsem čerpala z [10] a [21].

2.1. Inference proporcí

2.1.1. Věrohodnost

Představme si, že máme n pokusů, například hodů mincí. Z toho dostaneme n pozorování $y_1, ..., y_n$, kde y_i může nabývat pouze dvou hodnot, které můžeme označit jako 0 a 1, kde $y = 1$ bude značit padnutí hlavy a $y = 0$ padnutí orla. Předpokládejme, že pokusy jsou vzájemně nezávislé. S neznámým parametrem θ, který značí pravděpodobnost úspěchu (padnutí hlavy), nás tyto předpoklady

vedou k binomické věrohodnosti

$$p(r|n,\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r} \propto \theta^r(1-\theta)^{n-r}, \tag{2.1}$$

kde $r = \sum_i y_i$.

Pravděpodobnost padnutí hlavy na minci dostaneme z pravděpodobnostní funkce $p(y = 1|\theta) = f(\theta)$. Budeme předpokládat, že funkce $f(\theta)$ je identita, v tom případě dostaneme $p(y = 1|\theta) = \theta$. Jev, že padne orel, je doplňkem k jevu, kdy padne hlava, tedy $p(y = 0|\theta) = 1 - \theta$. Když předchozí úvahy shrneme do jednotného vyjádření, dostaneme výraz

$$p(y|\theta) \propto \theta^y(1-\theta)^{1-y}$$

pro $y \in \{0, 1\}$ a $\theta \in (0, 1)$ (prováděli jsme pouze jeden pokus, to tedy odpovídá podmíněné pravděpodobnostní funkci alternativního rozdělení). Tento vztah přitom odpovídá vztahu (2.1), který je pouze zobecněný pro n pokusů s r úspěchy.

Věrohodnost je funkce parametru θ při pevně daných hodnotách $\boldsymbol{y}$, takže roli proměnné hraje θ. Věrohodnostní funkce je funkce spojitého parametru, zatímco binomické rozdělení je diskrétní. Navíc věrohodnostní funkce nevyjadřuje rozdělení pravděpodobností. Představme si například, že $y = 1$. Potom

$$\int_0^1 \theta^y(1-\theta)^{1-y}d\theta = \int_0^1 \theta d\theta = 0,5 \neq 1,$$

a jak víme, tak pro hustotu rozdělení pravděpodobností musí platit, že

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Je důležité si uvědomit, že stejné vyjádření dané funkce může znamenat něco jiného. Pro pevně dané θ jde o rozdělení pravděpodobností, zatímco při pevně daných hodnotách $\boldsymbol{y}$ jde o věrohodnostní funkci.

2.1.2. Apriorní rozdělení

K odhadu parametru budeme potřebovat nějakou apriorní informaci. V principu můžeme užít jakékoli rozdělení pravděpodobností, ale jestli chceme využít Bayesovu větu, měli bychom splnit určité požadavky.

Jako první by bylo dobré, kdyby nám součin $p(\boldsymbol{y}|\theta)$ a $p(\theta)$ (což je čitatel v Bayesově větě) dal funkci ve stejném tvaru (jinak řečeno, aby výsledné rozdělení bylo ze stejné rodiny rozdělení) jako $p(\theta)$. Když se tak stane, potom apriorní a aposteriorní rozdělení jsou popsána funkcemi lišícími se pouze hodnotami parametrů. To nám umožňuje přidávat další data a odvozovat další aposteriorní rozdělení, které má zase stejný tvar. Nezáleží na tom, kolik dat zahrneme, pořád budou apriorní a aposteriorní rozdělení pocházet ze stejné rodiny rozdělení.

Zadruhé chceme, aby jmenovatel Bayesovy věty šel vypočítat analyticky. To také záleží na tom, v jakém vztahu jsou rozdělení, kterým odpovídají funkce (hustoty, resp. pravděpodobnostní funkce) $p(\boldsymbol{y}|\theta)$ a $p(\theta)$.

Když tyto funkce zkombinujeme tak, že aposteriorní rozdělení bude ze stejné rodiny jako apriorní, potom to znamená, že $p(\theta)$ je konjugované apriorní rozdělení k $p(\boldsymbol{y}|\theta)$. To je výhodné, protože se nám tím výpočet velmi usnadní. Hledáme tedy takové rozdělení parametru θ, které bude konjugované k binomické věrohodnostní funkci.

Když nemáme k dispozici žádná data, věříme, že všechny hodnoty θ jsou stejně pravděpodobné (což je nerealistické). Potom by měl mít parametr θ rovnoměrné rozdělení,

$$\theta \sim Ro(0, 1).$$

Takové apriorní rozdělení se nazývá neinformativní[1]. Pro rovnoměrné apriorní rozdělení platí, že $p(\theta) = 1$ na intervalu (0,1) a $p(\theta) = 0$ jinde. Aposteriorní pravděpodobnost potom v tomto případě vypadá následovně,

$$p(\theta|r, n) \propto \theta^r(1 - \theta)^{n-r} \times 1.$$

[1]Existuje mnoho neinformativních rozdělení, která se dají využít i u dat s jiným rozdělením. Jde o plochá rozdělení. Většinou můžeme použít konjugovaná apriorní rozdělení a vhodnou volbou parametrů je učinit neinformativními. Neinformativní apriorní rozdělení použijeme tehdy, když o parametru, který chceme odhadnout, nic nevíme.

Toto vyjádření má formu beta rozdělení

$$Beta(r+1, n-r+1).^{2} \tag{2.2}$$

Abychom mohli využít toho, že máme k dispozici apriorní informaci, je dobré použít jako apriorní rozdělení beta rozdělení,

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1} \propto Beta(a, b),$$

abychom mohli určit, které hodnoty parametru θ jsou více pravděpodobné a které méně. Parametry a a b jsou ekvivalentní k pozorování $a-1$ úspěchů v $a+b-2$ pokusech.

Když zkombinujeme tuto apriorní informaci s binomickou věrohodností, dostaneme aposteriorní rozdělení,

$$\begin{aligned} p(\theta|r,n) = \frac{p(r,n|\theta)p(\theta)}{p(r,n)} = \frac{\theta^r(1-\theta)^{n-r} \times \theta^{(a-1)}(1-\theta)^{b-1}}{B(a,b)p(r,n)} = \\ = \frac{\theta^{r+a-1}(1-\theta)^{n-r+b-1}}{B(a,b)p(r,n)} \sim \\ \sim Beta(r+a, n-r+b), \end{aligned} \tag{2.3}$$

kde

$$B(a,b)p(r,n) = B(r+a, n-r+b). \tag{2.4}$$

Dostali jsme opět beta rozdělení. To znamená, že apriorní rozdělení je konjugované k věrohodnosti, protože je ze stejné rodiny rozdělení jako aposteriorní rozdělení.

$Beta(a, b)$ rozdělení má střední hodnotu

$$E(\theta) = \frac{a}{a+b} \tag{2.5}$$

a rozptyl

$$var(\theta) = \frac{ab}{(a+b)^2(a+b+1)}.$$

[2] $Beta(a, b)$ rozdělení má hustotu $p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$, kde parametry $a, b > 0$. $B(a, b)$ je normalizační konstanta, která zajišťuje, že oblast pod hustotou bude rovna 1. Jinými slovy, $B(a, b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta$.

Odtud střední hodnota aposteriorního rozdělení,

$$E(\theta|r,n) = \frac{r+a}{n+a+b}, \tag{2.6}$$

kterou budeme používat jako odhad parametru θ. Jde tzv. o bodový odhad aposteriorní střední hodnotou.

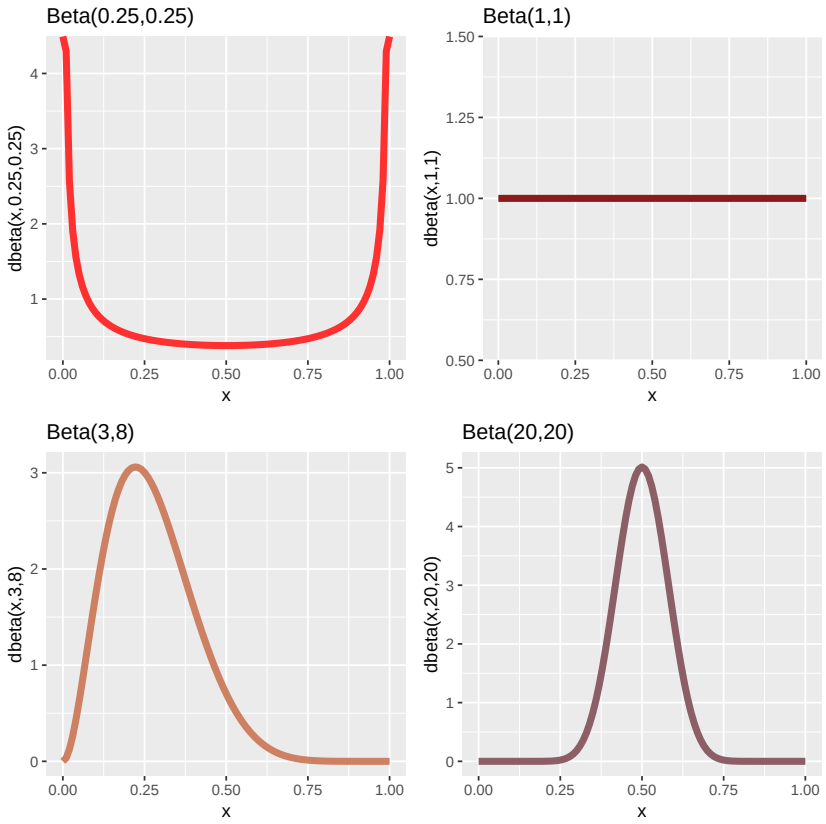
Ještě je důležité zmínit, že $Beta(1,1)$ rozdělení je ekvivalentní s rovnoměrným rozdělením $Ro(0,1)$, takže i v tomto případě jde o konjugované apriorní rozdělení k věrohodnosti.

Pomocí kódu v R se můžeme podívat, jak vypadá beta rozdělení pro různé parametry. Použijeme zde balíček *ggplot2*[23] pro vizualizaci grafů *gridExtra* [2], který nám umožní vykreslit více grafů do jednoho okna. Kód může vypadat například následovně.

```
library(ggplot2)
library(gridExtra)

x1 = seq(0,1,0.01)

h.beta1 = dbeta(x1,0.25,0.25)
h.beta2 = dbeta(x1,1,1)
h.beta3 = dbeta(x1,3,8)
h.beta4 = dbeta(x1,20,20)

y = t(matrix(c(h.beta1, h.beta2, h.beta3, h.beta4), byrow = T,
     nrow = 4))

gg.b.1= ggplot(data.frame(x1,h.beta1), aes(x1)) +
     geom_line(aes(y = h.beta1), colour = "firebrick1", lwd = 2) +
     xlab("x") + ylab("dbeta(x,0.25,0.25)") +
     ggtitle("Beta(0.25,0.25)")
gg.b.2 = ggplot(data.frame(x1,h.beta2), aes(x1)) +
```

```
      geom_line(aes(y = h.beta2),colour = "firebrick4", lwd = 2) +
      xlab("x") + ylab("dbeta(x,1,1)") + ggtitle("Beta(1,1)")
gg.b.3 = ggplot(data.frame(x1,h.beta3), aes(x1)) +
      geom_line(aes(y = h.beta3), colour = "lightsalmon3", lwd = 2) +
      xlab("x") + ylab("dbeta(x,3,8)") + ggtitle("Beta(3,8)")
gg.b.4 = ggplot(data.frame(x1,h.beta4), aes(x1)) +
      geom_line(aes(y = h.beta4), colour = "lightpink4", lwd = 2) +
      xlab("x") + ylab("dbeta(x,20,20)") + ggtitle("Beta(20,20)")

grid.arrange(gg.b.1, gg.b.2, gg.b.3, gg.b.4, ncol = 2)
```

Hustoty rozdělení s jednotlivými parametry můžeme vidět na obrázku 2.1. Z grafu vpravo nahoře vidíme, že opravdu $Beta(1,1)$ odpovídá rozdělení $Ro(0,1)$.



Obrázek 2.1: Beta rozdělení pro různé hodnoty parametrů a a b.

2.2. Příklad na reálných datech - odhad parametru θ

Ukažme si bayesovskou inferenci pro parametr θ binomického rozdělení na příkladu s reálnými daty. Data pocházejí z internetové stránky Kaggle [9]. Data s názvem *food choices* obsahují odpovědi studentů z Mercyhurst University a zahrnují informace o preferencích v jídle a pohybových aktivitách. Tento datový soubor o životním stylu obsahuje 125 pozorování a 61 proměnných. My se budeme věnovat pouze proměnné *sports*.

Tato proměnná vyjadřuje, jestli studenti sportují. Hodnoty této proměnné jsou tedy 1 (Ano) a 2 (Ne). Je zřejmé, že tato proměnná má binomické rozdělení, s parametry $n = 125$ a θ neznámým. Ten se budeme snažit odhadnout. Jako příznivé výsledky budeme chápat odpovědi, kde studenti uvedli, že sportují, tzn. hodnoty 1 u proměnné *sports*.

Protože nemáme k dispozici apriorní informaci, použijeme pro parametr θ rovnoměrné rozdělení $Ro\ (0,1) = Beta(1,1)$. To pro nás znamená, že využijeme vztah (2.2).

```
n = length(b$sports)
r = length(b$sports[b$sports==1])
a = 1
b = 1
apr.E = a/(a + b)
apost.a = r + 1
apost.b = n - r + 1
```

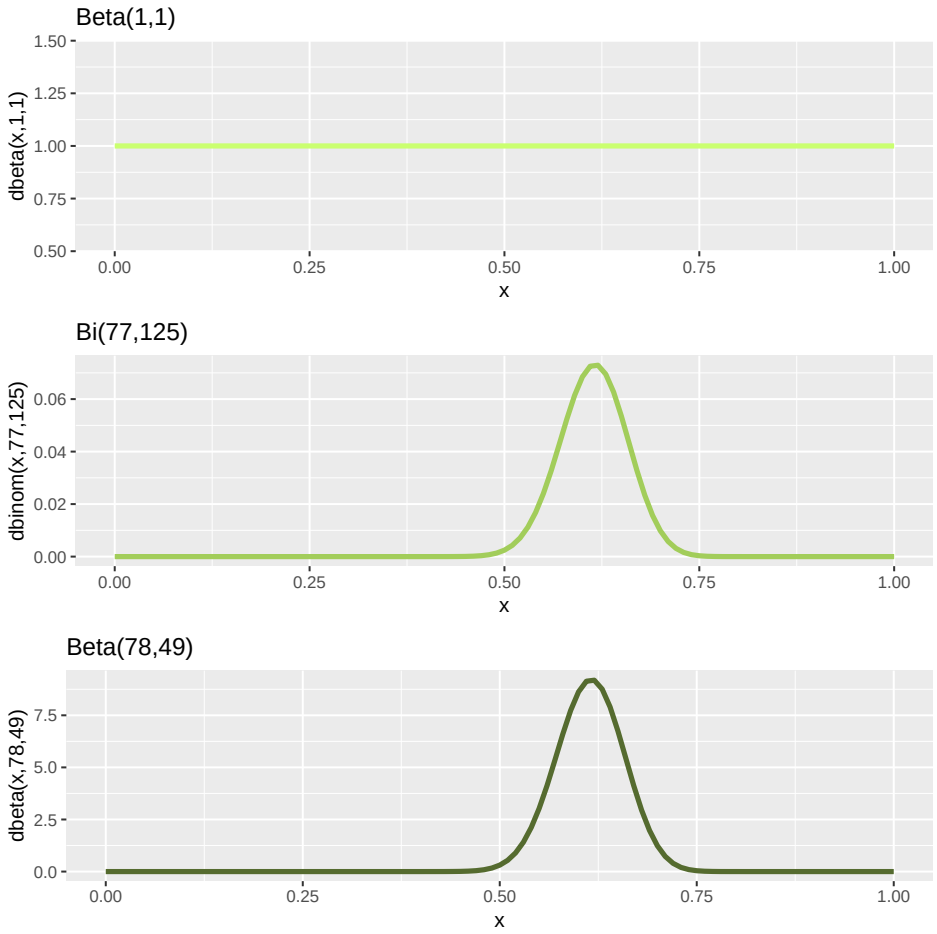
```
> n
[1] 125
> r
[1] 77
> apr.E
```

```
[1] 0.5
> apost.a
[1] 78
> apost.b
[1] 49
```

Spočítali jsme si, kolik je zde pozorování a kolik jich je pro nás příznivých (studenti sportují). Vyšli jsme z toho, že parametry $a = 1$ a $b = 1$. Pomocí vztahu (2.5) jsme si spočítali apriorní střední hodnotu, která vyšla $1/2$. Potom jsme vypočítali parametry pro aposteriorní beta rozdělení pomocí vztahu (2.2). Naše aposteriorní rozdělení parametru θ tedy odpovídá $Beta(78; 49)$. Jednotlivá rozdělení (apriorní, věrohodnost, aposteriorní) si můžeme vykreslit, viz Obrázek 2.2.

```
library(grid)

x1 = seq(0,1,0.01)

apr.beta = dbeta(x1, a, b)
ver.beta = dbinom(r, n, x1)
apost.beta = dbeta(x1, apost.a, apost.b)

y = t(matrix(c(apr.beta, ver.beta, apost.beta),
     byrow = T, nrow = 3))

apr.b = ggplot(data.frame(x1, apr.beta), aes(x1)) +
    geom_line(aes(y = apr.beta),
    colour = "darkolivegreen1", lwd=1.3) +
    xlab("x") + ylab("dbeta(x, 1, 1)") + ggtitle("Beta(1,1)")
ver.b = ggplot(data.frame(x1, ver.beta), aes(x1)) +
    geom_line(aes(y = ver.beta),
    colour = "darkolivegreen3", lwd=1.3) +
```

```
    xlab("x") + ylab("dbinom(x, 77, 125)") +
    ggtitle("Bi(77,125)")
apost.b = ggplot(data.frame(x1, apost.beta), aes(x1)) +
    geom_line(aes(y = apost.beta),
    colour="darkolivegreen", lwd=1.3) +
    xlab("x") + ylab("dbeta(x, 78, 49)") + ggtitle("Beta(78,49)")

graf = grid.arrange(apr.b, ver.b, apost.b, ncol = 1)
```



Obrázek 2.2: Jednotlivé hustoty apriorního rozdělení, věrohodnosti a aposteriorního rozdělení. Vycházíme z apriorního rozdělení $Ro(0,1) \equiv Beta(1,1)$.

Bodový odhad parametru θ získáme pomocí střední hodnoty aposteriorního beta rozdělení. Použijeme tedy vztah (2.6).

```
apost.E = (apost.a)/(apost.a + apost.b)
round(apost.E, 4)
```

```
> apost.E
[1] 0.6141732
> round(apost.E, 4)
[1] 0.6142
```

Výsledek tedy můžeme interpretovat tak, že 61,42 % studentů této univerzity sportuje.

Krok s výpočtem parametrů aposteriorního rozdělení jsme v tomto případě mohli v podstatě vynechat a mohli jsme rovnou dosadit do vztahu pro aposteriorní střední hodnotu.

Můžeme si zkusit změnit apriorní rozdělení a podíváme se, jak se změní aposteriorní rozdělení, respektive bodový odhad. Pracujme například s rozdělením $Beta(50, 20)$. Postup budeme opakovat, budeme ale vycházet ze vztahu (2.3).

```
a2 = 50
b2 = 20
apost.a.2 = r + a2
apost.b.2 = n - r + b2
```

```
> apost.a.2
[1] 127
> apost.b.2
[1] 68
```

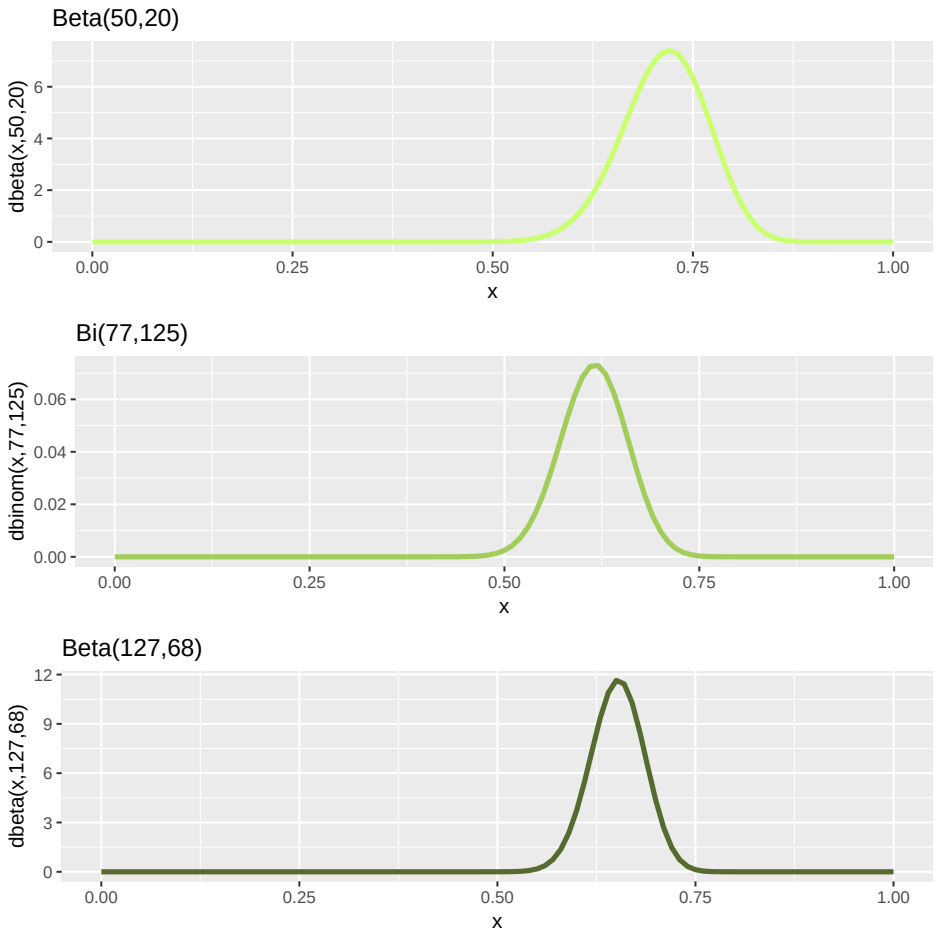
Aposteriorní rozdělení má teď tvar $Beta(127, 68)$. Opět si necháme vykreslit jednotlivá rozdělení. Podívejme se tedy na jednotlivé hustoty (Obrázek 2.3).

```
apr.beta2 = dbeta(x1, a2, b2)
ver.beta2 = dbinom(r, n, x1)
```

```
apost.beta2 = dbeta(x1, apost.a.2, apost.b.2)

y = t(matrix(c(apr.beta2, ver.beta2, apost.beta2),
     byrow = T, nrow = 3))

apr.b.2 = ggplot(data.frame(x1, apr.beta2), aes(x1)) +
     geom_line(aes(y = apr.beta2),
     colour = "darkolivegreen1", lwd=1.3) +
     xlab("x") + ylab("dbeta(x, 50, 20)") + ggtitle("Beta(50,20)")
ver.b.2 = ggplot(data.frame(x1, ver.beta2), aes(x1)) +
     geom_line(aes(y = ver.beta2),
     colour = "darkolivegreen3", lwd=1.3) +
     xlab("x") + ylab("dbinom(x, 77,125)") +
     ggtitle("Bi(77, 125)")
apost.b.2 = ggplot(data.frame(x1, apost.beta2), aes(x1)) +
     geom_line(aes(y=apost.beta2),
     colour = "darkolivegreen", lwd=1.3) +
     xlab("x") + ylab("dbeta(x,127, 68)") + ggtitle("Beta(127,68)")

grid.arrange(apr.b.2, ver.b.2, apost.b.2, ncol = 1)
```

Z obrázku 2.3 z grafu pro apriorní rozdělení vidíme, že jsme počítali s tím, že studenti nejčastěji sportují s modem rozdělení 0,72. Věrohodnost vypadá pořád stejně. Změnilo se i aposteriorní rozdělení. Spočítejme tedy aposteriorní střední hodnotu.

```
apost.E.2 = (apost.a.2)/(apost.a.2 + apost.b.2)
round(apost.E.2, 4)

> apost.E.2
[1] 0.6512821
```



Obrázek 2.3: Jednotlivé hustoty apriorního rozdělení, věrohodnosti a aposteriorního rozdělení. Vycházíme z apriorního rozdělení $Beta(50, 20)$.

```
> round(apost.E.2, 4)
[1] 0.6513
```

Odhad parametru θ nám vyšel 0,6513. Když to srovnáme s předchozím odhadem, který vyšel 0,6142, vidíme, že je vyšší. Teď se ovšem nabízí otázka, který model je tedy lepší, když věrohodnost vypadá stejně? Podíváme se na to v následující podkapitole.

2.3. Porovnávání modelů

K porovnávání modelů využijeme poměr aposteriorních rozdělení. Vyjdeme tedy z Bayesovy věty (1.2). Nejprve si ji upravíme do vhodného tvaru,

$$p(\theta|\boldsymbol{y}, m) = \frac{p(\boldsymbol{y}|\theta, m)p(\theta|m)}{p(\boldsymbol{y}|m)},$$

kde $p(\boldsymbol{y}|m) = \int_{-\infty}^{\infty} p(\boldsymbol{y}|\theta, m)p(\theta|m)d\theta$. Symbol m zde představuje daný model.

Mějme modely m_1 a m_2, pro které platí,

$$p(m_1|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|m_1)p(m_1)}{p(\boldsymbol{y})}, \qquad p(m_2|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|m_2)p(m_2)}{p(\boldsymbol{y})}, \tag{2.7}$$

kde $p(\boldsymbol{y}) = \sum_i p(\boldsymbol{y}|m_i)p(m_i)$. Můžeme zkonstruovat poměr aposteriorních rozdělení

$$\frac{p(m_1|\boldsymbol{y})}{p(m_2|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|m_1)}{p(\boldsymbol{y}|m_2)} \times \frac{p(m_1)}{p(m_2)}.$$

Poměr aposteriorních rozdělení je tedy roven součinu poměru věrohodností vztahů (2.7) a apriorních rozdělení. Většinou nedáváme přednost jednomu modelu před druhým, poměr apriorních rozdělení bude tedy nejčastěji 1. V takovémto případě nám stačí pracovat pouze s poměrem

$$B_{12} = \frac{p(\boldsymbol{y}|m_1)}{p(\boldsymbol{y}|m_2)} = \frac{\int_{-\infty}^{\infty} p(\boldsymbol{y}|\theta, m_1)p(\theta|m_1)d\theta}{\int_{-\infty}^{\infty} p(\boldsymbol{y}|\theta, m_2)p(\theta|m_2)d\theta}, \tag{2.8}$$

který se nazývá Bayesův faktor.

Data $\boldsymbol{y}$ a model m jsou vlastně vyjádřena pomocí r a n. Když využíváme binomickou věrohodnost a apriorní rozdělení beta, potom $p(\boldsymbol{y}|m)$ můžeme vyjádřit jako $p(r, n)$ a to už umíme jednoduše spočítat. Ze vztahu (2.4) plyne, že

$$B(a, b)p(r, n) = B(r + a, n - r + b).$$

Když tento výraz vyřešíme pro $p(r, n)$, dostaneme

$$p(r, n) = \frac{B(r + a, n - r + b)}{B(a, b)}. \tag{2.9}$$

Pro srovnání těchto dvou situací, kdy v prvním případě vycházíme z $Beta(1,1)$ a v druhém z $Beta(50,20)$, si tedy spočítáme hodnotu $p(r,n)$ pomocí vztahu (2.9), musíme ale přidat informaci i o apriorním rozdělení.

```
x = seq(0,1,0.001)
mb_1 = beta(apost.a, apost.b)/beta(a, b)*sum(dbeta(x, a, b))
mb_2 = beta(apost.a.2, apost.b.2)/beta(a2, b2)*sum(dbeta(x, a2, b2))
pomer_post = mb_1/mb_2
```

```
> pomer_post
[1] 0.5052346
```

Poměr aposteriorních rozdělení vyšel 0,505. Hodnota vyšla menší než 1, měli bychom tedy upřednostnit model m_2. Kdyby nám hodnota vyšla větší než 1, potom bychom preferovali model m_1.

Kapitola 3

Bayesovské metody pro normální rozdělení

V minulé kapitole jsme probrali bayesovské metody pro odhad parametru θ pro binomické rozdělení. V této kapitole se podíváme, jak to bude vypadat pro data s normálnm rozdělením $N(\mu, \sigma^2)$. Připomeňme si, že hustota normálního rozdělení $N(\mu, \sigma^2)$ je ve tvaru

$$p(y|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

Věrohodnostní funkce pro náhodný výběr $\boldsymbol{y} = (y_1, \ldots, y_n)$ bude mít tvar

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}, \tag{3.1}$$

kde $\boldsymbol{\theta} = (\mu, \sigma^2)$.

V této kapitole se tedy podíváme na případ normálního rozdělení, kdy neznámým parametrem bude nejprve jen μ a potom budeme mít neznámé oba parametry.

V této kapitole jsem pracovala se zdroji [8], [11], [12] a [21].

3.1. Bayesovská analýza při známém rozptylu a neznámé střední hodnotě

Předpokládejme, že máme složky náhodného výběru $y_i \sim N(\mu, \sigma^2)$, které se řídí normálním rozdělením, kde $i = 1, \ldots, n$, σ^2 známe, ale μ neznáme. Konjugo-

vané apriorní rozdělení μ bude opět normální

$$\mu \sim N(\mu_0, \tau_0^2). \tag{3.2}$$

Odvodíme si aposteriorní rozdělení (proporcionálně) pomocí Bayesovy věty. Už víme, že aposteriorní rozdělení je proporcionální k součinu věrohodnosti a apriorního rozdělení

$$p(\mu|\boldsymbol{y}) \propto p(\boldsymbol{y}|\mu)p(\mu).$$

Dosadíme tedy věrohodnost a apriorní rozdělení pro naši situaci, to znamená věrohodnost ze vztahu (3.1) a apriorní rozdělení (3.2).

$$p(\mu|\boldsymbol{y}) \propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2} \times e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2} = e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2+\frac{-1}{2\tau_0^2}(\mu-\mu_0)^2}.$$

Nyní pro jednoduchost budeme pracovat pouze s exponentem

$$\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2+\frac{-1}{2\tau_0^2}(\mu-\mu_0)^2 = -\frac{1}{2}\left(\frac{\sum_{i=1}^{n} y_i^2 - 2n\overline{y}\mu + n\mu^2}{\sigma^2} + \frac{\mu^2 - 2\mu\mu_0 + \mu_0^2}{\tau_0^2}\right),$$

kde $\overline{y}$ značí výběrový průměr. Každý člen, kde se nevyskytuje parametr μ, bereme jako konstantu, tudíž ho můžeme vynechat a dostaneme

$$\propto -\frac{1}{2}\left(\frac{-2\tau_0^2 n\overline{y}\mu + \tau_0^2 n\mu^2 + \sigma^2\mu^2 - 2\sigma^2\mu\mu_0}{\sigma^2\tau_0^2}\right) =$$

$$= -\frac{1}{2}\left(\frac{\mu^2(n\tau_0^2+\sigma^2) - 2\mu(\sigma^2\mu_0 + \tau_0^2 n\overline{y})}{\sigma^2\tau_0^2}\right)$$

$$= -\frac{1}{2}\left(\frac{\mu^2 - 2\mu\frac{\sigma^2\mu_0+\tau_0^2 n\overline{y}}{n\tau_0^2+\sigma^2}}{\frac{\sigma^2\tau_0^2}{n\tau_0^2+\sigma^2}}\right).$$

Teď už jen upravíme na úplný čtverec a vynecháme přebytečné konstatny,

$$-\frac{1}{2}\left(\frac{(\mu - \frac{\sigma^2\mu_0+\tau_0^2 n\overline{y}}{n\tau_0^2+\sigma^2})^2}{\frac{\sigma^2\tau_0^2}{n\tau_0^2+\sigma^2}}\right),$$

dostaneme tedy výsledný tvar

$$p(\mu|\boldsymbol{y}) \propto exp\left\{\frac{-(\mu - \frac{\sigma^2\mu_0+\tau_0^2 n\overline{y}}{n\tau_0^2+\sigma^2})^2}{2(\frac{\sigma^2\tau_0^2}{n\tau_0^2+\sigma^2})}\right\}.$$

Střední hodnota aposteriorního rozdělení parametru μ je rovna

$$\mu_n = \frac{\sigma^2\mu_0 + \tau_0^2 n\overline{y}}{n\tau_0^2 + \sigma^2} \tag{3.3}$$

a aposteriorní rozptyl

$$\tau_n^2 = \frac{\sigma^2\tau_0^2}{n\tau_0^2 + \sigma^2}. \tag{3.4}$$

Aposteriorní střední hodnotu můžeme vyjádřit více způsoby.

1. způsob

$$\mu_n = \frac{\sigma^2\mu_0 + \tau_0^2 n\overline{y}}{n\tau_0^2 + \sigma^2} = \frac{\sigma^2\mu_0 + \tau_0^2 n\overline{y} + \tau_0^2 n\mu_0 - \tau_0^2 n\mu_0}{n\tau_0^2 + \sigma^2} = \mu_0 + \frac{\tau_0^2 n\overline{y} - \tau_0^2 n\mu_0}{n\tau_0^2 + \sigma^2} =$$

$$= \mu_0 + (\overline{y} - \mu_0)\frac{n\tau_0^2}{n\tau_0^2 + \sigma^2},$$

2. způsob

$$\mu_n = \frac{\sigma^2\mu_0 + \tau_0^2 n\overline{y}}{n\tau_0^2 + \sigma^2} = \frac{\sigma^2\mu_0 + \tau_0^2 n\overline{y} + \sigma^2\overline{y} - \sigma^2\overline{y}}{n\tau_0^2 + \sigma^2} = \overline{y} - \frac{\sigma^2\overline{y} - \sigma^2\mu_0}{n\tau_0^2 + \sigma^2} =$$

$$= \overline{y} - (\overline{y} - \mu_0)\frac{\sigma^2}{n\tau_0^2 + \sigma^2}.$$

Z prvního způsobu vyjádření můžeme vidět, že došlo k posunutí aposteriorní střední hodnoty směrem k apriorní střední hodnotě.

Střední hodnota aposteriorního rozdělení je jiná, než jak ji klasicky odhadujeme pomocí aritmetického (výběrového) průměru. Důvodem je to, že naše apriorní informace tento odhad do jisté míry ovlivňuje. Kdyby apriorní informace byla, že μ je blízko deseti, potom aposteriorní rozdělení posune μ k deseti.

3.1.1. Příklad na reálných datech - odhad parametru μ

K aplikaci na reálný problém jsem použila data s názvem *Fat*. Data byla získána z knihovny *UsingR* [22] přímo z R. Datový soubor zahrnuje 19 antropologických a s nimi souvisejících údajů od 252 mužů. Budeme se zabývat proměnnou *body fat*. Tato proměnná vyjadřuje procento tělesného tuku pomocí Brozekovy rovnice [1].

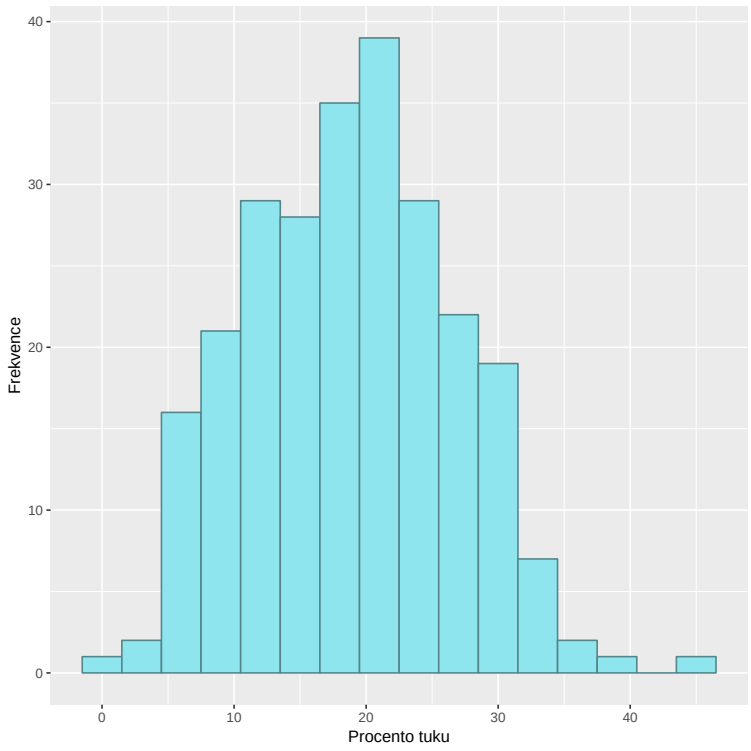
Protože předpokládáme, že data pocházejí z normálního rozdělení, měli bychom to ověřit. Nejprve se podíváme na histogram a ten potom doplníme testem normality. Ověřování normality provedeme pomocí klasického přístupu, přestože se jedná o práci na téma bayesovských metod. Testování hypotéz ale není tématem této práce, hlavním úkolem je odhadnout parametry za přepokladu normality.

```
library(UsingR)
f = fat
ggplot(f, aes(x = body.fat)) + geom_histogram(binwidth = 3,
     fill = "cadetblue2", col = "cadetblue4") + ylab("Frekvence") +
     xlab("Procento tuku")
shapiro.test(f$body.fat)
```

```
> shapiro.test(f$body.fat)

Shapiro-Wilk normality test

data:  f$body.fat
W = 0.99292, p-value = 0.2747
```

Když se podíváme na histogram (Obrázek 3.1), tak bychom zde normální chování předpokládali, protože je rozdělení symetrické. P-hodnota je větší než zvolená hladina testu $\alpha = 0{,}05$, tudíž hypotézu o tom, že je rozdělení proměnné

[1] Procento tělesného tuku pomocí Brozekovy rovnice dostaneme $(457/\rho - 414{,}2)$, kde ρ vyjadřuje hustotu v g/cm^3.



Obrázek 3.1: Histogram proměnné body fat

body fat normální, nelze zamítnout. Ještě provedeme odhad rozptylu, který budeme brát za skutečnou hodnotu, protože se nacházíme v případě, kdy σ^2 je známé.

```
var(f$body.fat)
```

```
> var(f$body.fat)
[1] 60.07576
```

Můžeme pokračovat v odhadu parametru μ. Když se podíváme na tabulku 3.1 [13], vidíme, že jsou zde čtyři kategorie. My budeme vycházet z kategorie pro průměrné muže, protože chceme odhadnout střední hodnotu procentuálního zastoupení tělesného tuku. Apriorní rozdělení tedy bude mít hustotu $\mu \sim N(21; 9)$. Paramter τ_0^2 má hodnotu 9, protože když se podíváme na rozpětí procentuálního zastoupení tělesného tuku pro průměrné muže, tak od hodnoty 21 % se hodnoty pohybují o 3 % níže i výše.

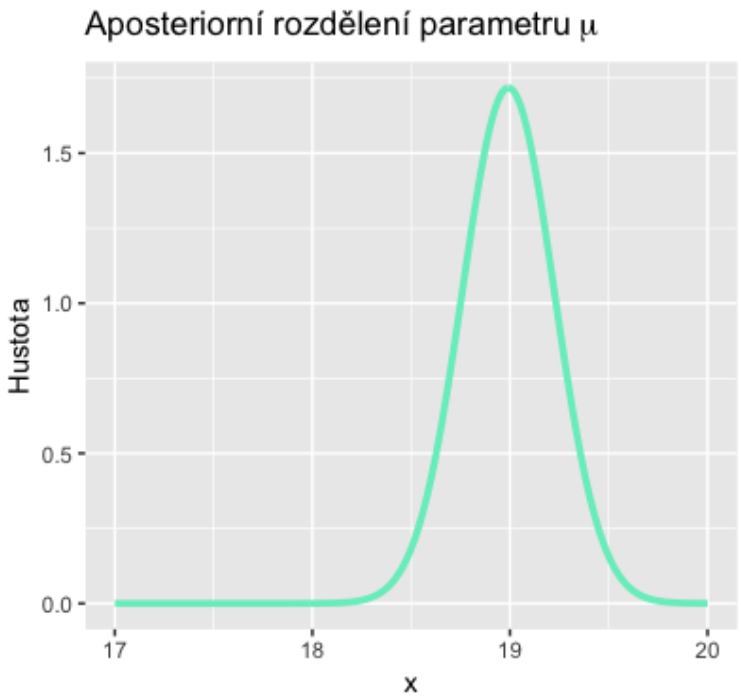
Klasifikace	Muži	Zastupující hodnota
atleti	6-13 %	9,5 %
fitness	14-17 %	15,5 %
průměrný	18-24 %	21 %
obézní	25 % +	37 %

Tabulka 3.1: Tabulka procentuálního zastoupení tělesného tuku.

K výpočtu využijeme vztahy (3.3) a (3.4).

```
mu_0 = 21
tau_0 = 9
n = length(f$body.fat)
apost_mu = (sigma_2*mu_0 + tau_0*n*mean(f$body.fat))/
     (n*tau_0 + sigma_2)
apost_tau = (sigma_2*tau_0)/(n*tau_0 + sigma_2)

x = seq(17,20,.01)
Hustota = dnorm(x, n.prum.mí, n.var.mí)
ggplot(data.frame(Hustota,x), aes(x)) + geom_line(aes(y = Hustota),
     colour = "aquamarine2", lwd = 1.3) +
     ggtitle("Aposteriorní rozdělení parametru" ~ mu)

> n
[1] 252
> apost_mu
[1] 18.99169
> apost_tau
[1] 0.2322441
```

Odhad parametru μ formou aposteriorní střední hodnoty vyšel 18,99. To nám tedy říká, že průměrný muž z tohoto souboru má zhruba 18,99 % tělesného tuku.



Obrázek 3.2: Aposteriorní rozdělení parametru μ při známém σ^2 je normální s parametry $\mu_n = 18{,}99$ a $\sigma_n^2 = 0{,}23$.

3.2. Bayesovská analýza při neznámém rozptylu a neznámé střední hodnotě

Případ normálního rozdělení se známým rozptylem je ve většině praktických situací nerealistický. Předpokládejme tedy výběr z normálního rozdělení $y_i \sim N(\mu, \sigma^2)$, kde $i = 1, \ldots, n$. Tentokrát neznáme ani σ^2 ani μ. Většinou nás zajímá hlavně hodnota parametru μ.

V tomto případě budeme pracovat s převrácenou hodnotou parametru σ^2, kterou budeme značit $\tau = \frac{1}{\sigma^2}$. Parametr τ představuje přesnost napozorovaných hodnot. Věrohodnostní funkce pro tento případ má tvar

$$p(\boldsymbol{y}|\mu, \tau) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} (\tau)^{1/2} e^{-\frac{\tau}{2}(y_i-\mu)^2} = \frac{1}{(2\pi)^{n/2}} \tau^{n/2} e^{-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2}.$$

Potřebujeme vhodné konjugované apriorní rozdělení pro tyto dva parametry

μ, τ. Pro tuto situaci je vhodné vyjít z normálního-gama rozdělení[2], které vzniká součinem normálního rozdělení a gama rozdělení[3]. Předpokládejme tedy apriorní rozdělení

$$\mu|\tau \sim N(\mu_0, (\kappa_0\tau)^{-1}), \tag{3.5}$$

$$\tau \sim Gamma(\alpha_0, \beta_0). \tag{3.6}$$

Sdružené konjugované apriorní rozdělení odvodíme pomocí hustot jednotlivých rozdělení doplněním parametrů ze vztahů (3.5) a (3.6),

$$p(\mu, \tau) = N(\mu|\mu_0, (\kappa_0\tau)^{-1}) \times Gamma(\tau|\alpha_0, \beta_0) =$$

$$= \frac{1}{\sqrt{2\pi}}(\tau\kappa_0)^{1/2} e^{-\frac{(\mu-\mu_0)^2}{2/\kappa_0\tau}} \times \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\tau^{\alpha_0-1}e^{-\beta_0\tau} =$$

$$= \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\left(\frac{\kappa_0}{2\pi}\right)^{1/2} \times \tau^{1/2} e^{-\frac{\kappa_0\tau}{2}(\mu-\mu_0)^2} \times \tau^{\alpha_0-1}e^{-\beta_0\tau} \propto$$

$$\propto \tau^{1/2} e^{-\frac{\kappa_0\tau}{2}(\mu-\mu_0)^2} \times \tau^{\alpha_0-1}e^{-\beta_0\tau} \propto$$

$$\propto \tau^{\alpha_0-1/2} e^{-\frac{\tau}{2}[\kappa_0(\mu-\mu_0)^2+2\beta_0]} \sim NG(\mu, \tau|\mu_0, \kappa_0, \alpha_0, \beta_0).$$

Marginální apriorní rozdělení parametru μ dostaneme integrací sdruženého apriorního rozdělení přes parametr τ,

$$p(\mu) = \int_0^\infty p(\mu, \tau)d\tau.$$

Po integraci této funkce bychom došli k vyjádření

$$p(\mu) \propto \left(1 + \frac{1}{2\alpha_0} \times \frac{\alpha_0\kappa_0(\mu-\mu_0)^2}{\beta_0}\right)^{-\frac{2\alpha_0+1}{2}},$$

z čehož vyplývá, že apriorní rozdělení parametru μ má zobecněné Studentovo t-rozdělení[4][7],

$$p(\mu) \sim t_{2\alpha_0}\left(\mu_0, \frac{\beta_0}{\alpha_0\kappa_0}\right). \tag{3.7}$$

[2]Hustota normálního-gama rozdělení má tvar $f(x, \tau|\mu, \lambda, \alpha, \beta) = \frac{\beta^\alpha\sqrt{\lambda}}{\Gamma(\alpha)\sqrt{2\pi}}\tau^{\alpha-1/2}e^{-\beta\tau}e^{-\frac{\lambda\tau(x-\mu)^2}{2}}$, kde parametry $\alpha, \beta, \lambda > 0$, $\mu, x \in \boldsymbol{R}$ [4].

[3]Gama rozdělení je definováno hustotou $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$, $x \in (0, \infty)$, $\alpha, \beta > 0$.

[4]Hustota zobecněného t-rozdělení má tvar $f(x|\nu, \mu, \sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\sigma^2}}(1+\frac{1}{\nu}(\frac{x-\mu}{\sigma})^2)^{-\frac{\nu+1}{2}}$, kde parametr ν určuje počet stupňů volnosti, parametr μ střední hodnotu a parametr σ^2 rozptyl rozdělení.

Aposteriorní rozdělení parametru $\boldsymbol{\theta} = (\mu, \tau)$ dostaneme jako součin apriorního rozdělení a věrohodnosti,

$$p(\mu, \tau|\boldsymbol{y}) \propto p(\mu, \tau) \times p(\boldsymbol{y}|\mu, \tau) \propto \tau^{\alpha_0 - 1 + 1/2} e^{-\frac{\tau}{2}(\kappa_0(\mu-\mu_0)^2 + \beta_0)} \times \tau^{n/2} e^{-\frac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

$$\propto \tau^{\alpha_0 - 1 + 1/2 + n/2} e^{-\frac{\tau}{2}[\kappa_0(\mu-\mu_0)^2 + 2\beta_0 + \sum_{i=1}^{n}(y_i-\mu)^2]}.$$

Podívejme se na výraz $\sum_{i=1}^{n}(y_i - \mu)^2$,

$$\sum_{i=1}^{n}(y_i-\mu)^2 = \sum_{i=1}^{n}[(y_i-\overline{y})-(\mu-\overline{y})]^2 = \sum_{i=1}^{n}(y_i-\overline{y})^2 - 2\sum_{i=1}^{n}(y_i-\overline{y})(\mu-\overline{y}) + \sum_{i=1}^{n}(\mu-\overline{y})^2 =$$

$$= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2(\mu - \overline{y})(\sum_{i=1}^{n} y_i - n\overline{y}) + \sum_{i=1}^{n}(\mu - \overline{y})^2 =$$

$$= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2(\mu - \overline{y})(n\overline{y} - n\overline{y}) + \sum_{i=1}^{n}(\mu - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 + \sum_{i=1}^{n}(\mu - \overline{y})^2.$$

Nyní se vraťme ke sdruženému aposteriornímu rozdělení. Úpravami exponentu

$$e^{-\frac{\tau}{2}[\kappa_0(\mu-\mu_0)^2 + 2\beta_0 + \sum_{i=1}^{n}(y_i-\mu)^2]},$$

kde bychom využili skutečnosti uvedené výše (rozepsání výrazu $\sum_{i=1}^{n}(y_i - \mu)^2$), bychom se dostali k vyjádření

$$\propto \tau^{1/2} e^{-\frac{\tau}{2}(\kappa_0+n)(\mu-\mu_n)^2} \times \tau^{\alpha_0 + n/2 - 1} e^{-\frac{\tau}{2}\left[2\beta_0 + \sum_{i=1}^{n}(y_i-\overline{y})^2 + \frac{\kappa_0 n(\overline{y}-\mu_0)^2}{(\kappa_0+n)}\right]}$$

$$\sim N(\mu|\mu_n, ((\kappa_0 + n)\tau)^{-1}) \times Gamma(\tau|\alpha_n, \beta_n).$$

Parametry sdruženého aposteriorního rozdělení představují hodnoty

$$\mu_n = \frac{\kappa_0\mu_0 + n\overline{y}}{\kappa_0 + n}, \tag{3.8}$$

$$\alpha_n = \alpha_0 + \frac{n}{2},$$

$$\beta_n = \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \overline{y})^2 + \frac{\kappa_0 n(\overline{y} - \mu_0)^2}{2(\kappa_0 + n)}.$$

Marginální aposteriorní rozdělení obou parametrů dostaneme aktualizací marginálních apriorních rozdělení,

$$\mu|\boldsymbol{y} \sim t_{2\alpha_n}\left(\mu_n, \frac{\beta_n}{\alpha_n(\kappa_0+n)}\right), \tag{3.9}$$

$$\tau|\boldsymbol{y} \sim Gamma(\alpha_n, \beta_n). \tag{3.10}$$

Kdybychom chtěli znát aposteriorní rozdělení parametru σ^2 místo τ, vezmeme místo gama rozdělení inverzní gama rozdělení[5].

V tomto případě zde máme parametr κ_0, se kterým jsme se ještě nesetkali. Interpretace tohoto parametru je, že střední hodnota normálního rozdělení μ je odhadnuta z κ_0 pozorování se střední hodnotou μ_0. Parametr κ_0 vlastně určuje míru naší jistoty v apriorním rozdělení parametru μ. Když zvolíme $\kappa_0 = 1$, potom si jisti nejsme, protože hodnota rozptylu ze vztahu (3.5) bude velká. Zatímco když budeme hodnotu κ_0 zvyšovat, rozptyl se bude snižovat, a tím naše jistota bude větší.

Parametr τ byl určen z $2\alpha_0$ pozorování se střední hodnotou μ_0 a rozptylem $2\beta_0$.

3.2.1. Příklad na reálných datech - odhad μ při neznámém rozptylu

Příklad provedeme opět na datech pod názvem *Fat*. Normalitu už ověřovat nemusíme, tu jsme ověřili v příkladu 3.1.1. Odhad parametru μ provedeme pomocí vztahu (3.8). Rozptyl odhadu parametru μ počítáme, abychom mohli vykreslit graf aposteriorního rozdělení. Vykreslíme hustotu aposteriorního rozdělení tohoto případu společně s aposteriorním rozdělením v situaci, kdy jsme znali σ^2.

Parametry apriorního rozdělení zvolíme $\mu_0 = 21$, $\kappa_0 = 20$. U apriorního rozdělení parametru τ použijeme gama rozdělení s parametry $\alpha_0 = 40$ a $\beta_0 = 5$.

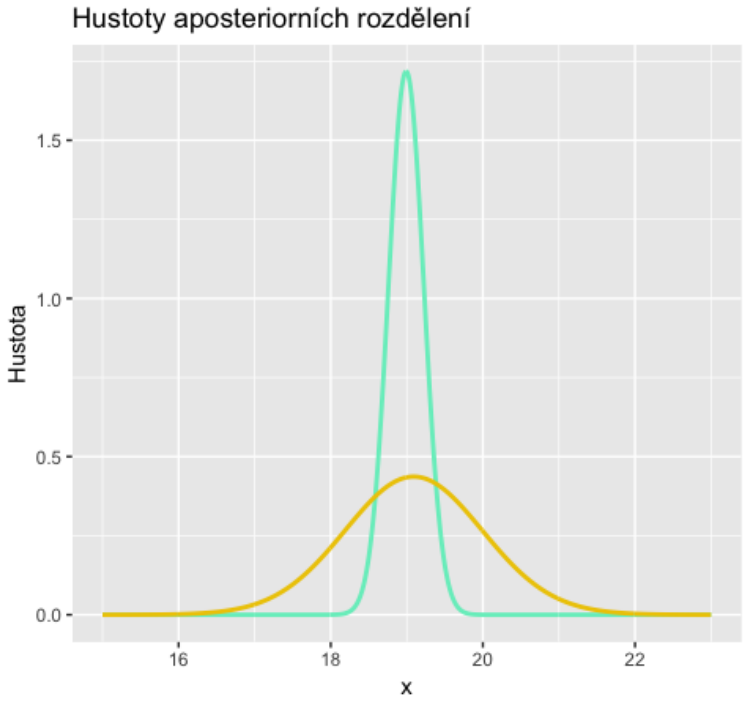
```
kappa_0 = 20
alpha_0 = 40
```

[5]Inverzní Gama rozdělení má hustotu $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}}$, $x \in (0, \infty)$, $\alpha, \beta > 0$.

```
beta_0 = 5

mu_n = (kappa_0*mu_0+n*mean(f$body.fat))/(kappa_0 + n)
beta_n = beta_0 + 1/2*sum((f$body.fat-mean(f$body.fat))^2) +
     (kappa_0*n*(mean(f$body.fat)-mu_0)^2)/(2*(kappa_0 + n))
alpha_n = alpha_0 + 1/2
rozpt_n = beta_n/(alpha_n*(kappa_0 + n))

library(mnormt)
#pro vykreslení hustoty t rozdělení s jiným polohovým a
#škálovým parametrem (funkce dmt)

x = seq(17, 21, 0.01)
Hustota = dnorm(x, apost_mu, apost_tau)
Hustota.t = dmt(x.t, mean = mu_n, S = sqrt(rozpt_n),
   df = 2*alpha_n)
df=data.frame(x,Hustota,Hustota.t)

ggplot(df, aes(x)) +
  geom_line(aes(y = Hustota), colour = "aquamarine2", lwd=1) +
  geom_line(aes(y = Hustota.t), colour = "gold2", lwd = 1) +
  ggtitle("Hustoty aposteriorních rozdělení")



> mu_n
[1] 19.34671
> rozpt_n
[1] 1.85273
```

Odhad střední hodnoty μ se liší jen málo. V tomto případě jsme dostali hodnotu 19,35. Rozptyl parametru μ je ale větší. Pro ilustraci jsme aposteri-

orní hustoty v tomto případě a v případě, kdy jsme σ^2 znali, nechali vykreslit (Obrázek 3.3).



Obrázek 3.3: Hustota aposteriorního rozdělení parametru μ při neznámém σ^2 je zobrazena žlutě. Hustotu aposteriorního rozdělení, kdy jsme parametr σ^2 znali, je zobrazena zelenou barvou.

Kapitola 4

Intervalové odhady

V předchozích kapitolách jsme ukázali, jak provést bodové odhady parametru θ. Někdy je dobré znát i odhad intervalový. V klasické statistice tento interval nazýváme konfidenčním nebo intervalem spolehlivosti. Interpretace tohoto intervalu zní, s jakou pravděpodobností pokrývá skutečnou (fixní) hodnotu daného parametru θ. V bayesovských metodách mluvíme o intervalu věrohodnostním. Protože odhadovaný paramtr zde vnímáme jako náhodnou veličinu, interpretace se liší. Zde můžeme říci, s jakou pravděpodobností leží hodnota parametru v daném intervalu. Věrohodnostní interval je tedy bayesovskou obdobou pro interval spolehlivosti.

V této kapitole jsem čerpala z [6], [11], [16] a [20].

Stejně jako vše ostatní, odvození intervalového odhadu je založeno opět na aposteriorním rozdělení $p(\theta|\boldsymbol{y})$. Mějme tedy náhodnou veličinu θ (odhadovaný parametr). $100(1-\alpha)\%$ věrohodnostní interval $C_{1-\alpha}(\boldsymbol{y})$ parametru θ je určen jako

$$p(\theta \in C_{1-\alpha}(\boldsymbol{y})|\boldsymbol{y}) = 1-\alpha, \tag{4.1}$$

kde α je námi určená mez, například 0,05 a $\boldsymbol{y}$ je náhodný výběr. Ve vícerozměrných případech mluvíme o věrohodnostní množině. Interval můžeme psát i jako

$$C_{1-\alpha}(\boldsymbol{y}) = (c,d),$$

$$P(c \leq \theta \leq d|\boldsymbol{y}) = 1-\alpha,$$

$$\int_c^d p(\theta|\boldsymbol{y})d\theta = 1-\alpha.$$

Tato definice intervalu může vést k mnoha různým volbám (c, d). Proto se často pracuje se specifičtějšími intervaly, například se symetrickým nebo s věrohodnostním intervalem o nejvyšší aposteriorní hustotě.

4.1. Symetrický věrohodnostní interval

Uvažujme, že chceme dostat symetrický interval $C^S_{1-\alpha}(\boldsymbol{y}) = (c, d)$, to znamená že

$$P(\theta < c|\boldsymbol{y}) = \frac{\alpha}{2} \qquad \wedge \qquad P(d < \theta|\boldsymbol{y}) = \frac{\alpha}{2},$$

respektive

$$\int_{-\infty}^{c} p(\theta|\boldsymbol{y})d\theta = \frac{\alpha}{2} \qquad \wedge \qquad \int_{d}^{\infty} p(\theta|\boldsymbol{y})d\theta = \frac{\alpha}{2}.$$

Potom je interval $C^S_{1-\alpha}(\boldsymbol{y}) = (c, d)$ $100(1-\alpha)\%$ symetrický věrohodnostní.

4.2. Věrohodnostní interval o nejvyšší aposteriorní hustotě

Věrohodnostní interval o nejvyšší aposteriorní hustotě (anglicky *highest posterior density (HPD)*) má tu vlastnost, že každý bod v intervalu má vyšší hustotu, než jakýkoli jiný bod mimo tento interval. Je to také nejkratší $100(1-\alpha)\%$ věrohodnostní interval. Předpokládejme, že chceme takový věrohodnostní interval. Takový interval $C^{HPD}_{1-\alpha}(\boldsymbol{y})$ dostaneme, když budou splněny následující podmínky:

- $C^{HPD}_{1-\alpha}(\boldsymbol{y})$ je $100(1-\alpha)\%$ věrohodnostní interval,
- $p(\theta_{in}|\boldsymbol{y}) \geq p(\theta_{out}|\boldsymbol{y}), \qquad \forall\theta_{in} \in C^{HPD}_{1-\alpha}(\boldsymbol{y}), \qquad \forall\theta_{out} \notin C^{HPD}_{1-\alpha}(\boldsymbol{y}).$

4.3. Příklady na reálných datech

Abychom doplnili odhadování parametru θ úplně, ke všem příkladům na reálných datech provedeme ještě oba intervalové odhady.

4.3.1. Binomické rozdělení - rovnoměrné apriorní rozdělení

Budeme vycházet z podkapitoly 2.2. Vycházeli jsme z apriorního rozdělení $Ro(0,1) \equiv Beta(1,1)$. Dostali jsme aposteriorní rozdělení $Beta(78,49)$. Naším úkolem bude sestavit 95% věrohodnostní interval. Začneme symetrickým. Potřebujeme tedy vyřešit

$$\int_{-\infty}^{c} Beta(78,49)d\theta = 0{,}025 \qquad \wedge \qquad \int_{d}^{\infty} Beta(78,49)d\theta = 0{,}025.$$

Takové integrály ale neumíme spočítat ručně. Pomůže nám k tomu následující kód v R.

```
c = qbeta(0.025, apost.a, apost.b)
d = qbeta(1-0.025, apost.a, apost.b)
```

```
> c
[1] 0.5282941
> d
[1] 0.6966455
```

Můžeme učinit závěr, že se zde vyskyuje 95% pravděpodobnost, že proporce studentů z Mercyhurst University, kteří sportují, je mezi hodnotami 0,53 a 0,70.

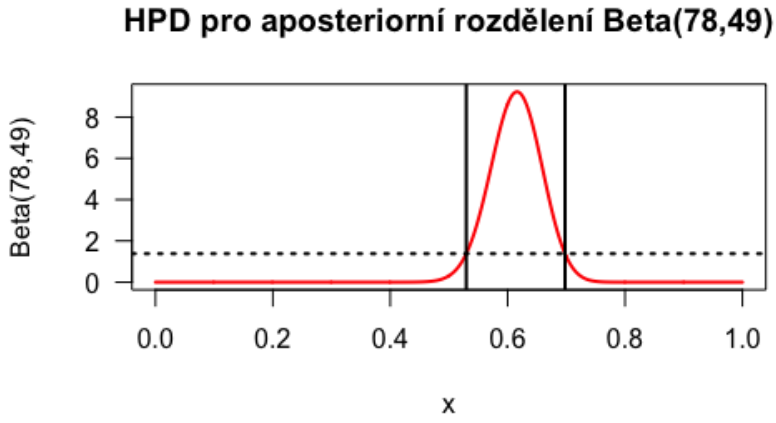
Nyní zkusme věrohodnostní interval o nejvyšší aposteriorní hustotě. V tomto případě bychom museli řešit integrál

$$\int_{c}^{d} Beta(78,49)d\theta = 0{,}95.$$

To ale opět analyticky neumíme. S využitím balíčku *Smisc* [18] a funkcí *hpd* to lze jednoduše provést.

```
library(Smisc)
apost.hustota = function(x) dbeta(x, apost.a, apost.b)
apost.distr.fce = function(x) pbeta(x, apost.a, apost.b)
hpd.int = hpd(apost.hustota, c(0,1), cdf = apost.distr.fce,
```

```
     prob = 0.95)
print(hpd.int)
plot(hpd.int, las = 1, main = "HPD pro aposteriorní rozdělení
     Beta(78,49)", lwd = 2, ylab = "Beta(78,49)")

> print(hpd.int)
$lower
[1] 0.5295197

$upper
[1] 0.6978049

$prob
[1] 0.95
```



Obrázek 4.1: Vyobrazení věrohodnostního intervalu o nejvyšší aposteriorní hustotě při aposteriorním rozdělení parametru $\theta \sim Beta(78, 49)$.

Můžeme říci, že s pravděpodobností 95 % se odhad parametru θ, tedy počet studentů z Mercyhurst University, kteří sportují, vyskytuje v intervalu (0,53;0,70).

Po zaokrouhlení na dvě desetinná místa v tomto případě ve výsledcích těchto dvou různých věrohodnostních intervalů nepoznáme rozdíl.

4.3.2. Binomické rozdělení - apriorní rozdělení beta

V případě, kdy jsme náhodně zvolili apriorní rozdělení parametru

$$\theta \sim Beta(50, 20),$$

jsme dostali aposteriorní rozdělení $\theta|\boldsymbol{y} \sim Beta(127, 68)$. Výpočet je analogický, jako v podkapitole 4.3.1. V R provedeme pouze malé úpravy kódu.

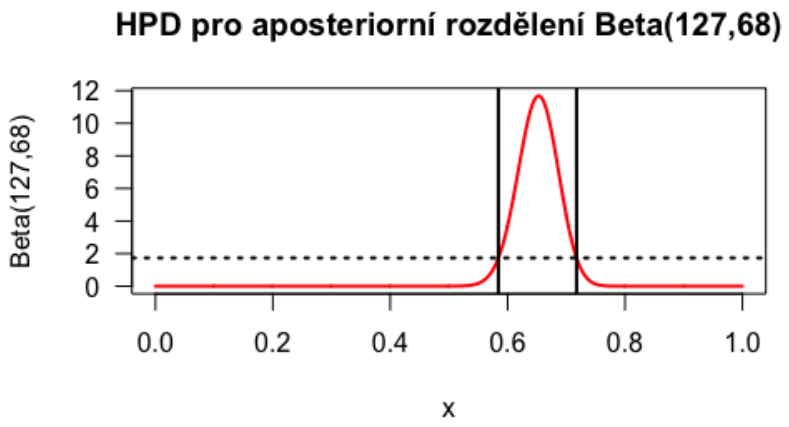
```
c2 = qbeta(0.025, apost.a.2, apost.b.2)
d2 = qbeta(1-0.025, apost.a.2, apost.b.2)


> c2
[1] 0.5831792
> d2
[1] 0.7164453
```

Symetrický interval pro aposteriorní rozdělení parametru $\theta|\boldsymbol{y} \sim Beta(127, 68)$ je (0,58;0,72).

Věrohodnostní interval o nejvyšší aposteriorní hustotě provedeme také pouze malými změnami stávajícího kódu.

```
apost.hustota.2 = function(x) dbeta(x, apost.a.2, apost.b.2)
apost.distr.fce.2 = function(x) pbeta(x, apost.a.2, apost.b.2)
hpd.int.2 = hpd(apost.hustota.2, c(0,1), cdf = apost.distr.fce.2,
     prob = 0.95)
print(hpd.int.2)
plot(hpd.int.2, las = 1, main = "HPD pro aposteriorní rozdělení
     Beta(127,68)", lwd=2,ylab="Beta(127,68)")


> print(hpd.int.2)
$lower
[1] 0.5842426
```

```
$upper
[1] 0.717446

$prob
[1] 0.949999
```



Obrázek 4.2: Vyobrazení věrohodnostního intervalu o nejvyšší aposteriorní hustotě při aposteriorním rozdělení parametru $\theta \sim Beta(127, 68)$.

Po zaokrouhlení na dvě desetinná místa budou opět oba intervaly shodny.

4.3.3. Normální rozdělení - známý rozptyl

Vrátíme se ke kapitole 3.1.1, kde jsme předpokládali apriorní rozdělení $\mu \sim N(21; 9)$. Aposteriorní rozdělení parametru μ mělo tvar $N(18{,}95; 0{,}23)$. Výpočty provedeme nejprve pro symetrický věrohodnostní interval.

```
c=qnorm(0.025,apost_mu,apost_tau)
d=qnorm(1-0.025,apost_mu,apost_tau)
```

```
> c
[1] 18.5365
> d
```

```
[1] 19.44688
```

Věrohodnostní interval o nejvyšší aposteriorní hustotě dostaneme pomocí následujícího kódu.

```
apost.hustota.norm = function(x) dnorm(x, apost_mu, apost_tau)
apost.distr.fce.norm = function(x) pnorm(x, apost_mu, apost_tau)
hpd.int.norm = hpd(apost.hustota.norm, c(16,21),cdf =
     apost.distr.fce.norm, prob = 0.95)
print(hpd.int.norm)
plot(hpd.int.norm, las = 1,main = "HPD pro aposteriorní rozdělení
     N(18,94;0,24)", lwd = 2, ylab = "N(18,94;0,24)")
```

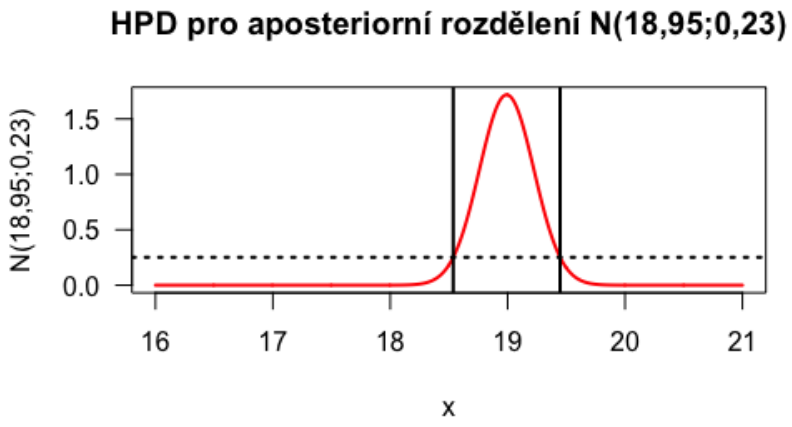
```
> print(hpd.int.norm)
$lower
[1] 18.5365

$upper
[1] 19.44688

$prob
[1] 0.9499999
```

Symetrický věrohodnostní interval a věrohodnostní interval o nejvyšší aposteriorní hustotě jsou oba rovny $(18{,}54; 19{,}45)$. Je zde 95% pravděpodobnost, že průměrné procento tělesného tuku u mužů, je mezi hodnotami 18,48 a 19,41.

4.3.4. Normální rozdělení - neznámý rozptyl

Zbývá poslední příklad, kdy jsme neznali jak střední hodnotu, tak rozptyl našich dat. Zde jsme dostali aposteriorní rozdělení $T_{254}(19{,}02; 2{,}81)$. Intervalové odhady provedeme následovně.

```
c = qt(0.025, df=2*alpha_n)*sqrt(rozpt_n) + mu_n
```



Obrázek 4.3: Vyobrazení věrohodnostního intervalu o nejvyšší aposteriorní hustotě při aposteriorním rozdělení parametru $\mu \sim N(18{,}95; 0{,}23)$.

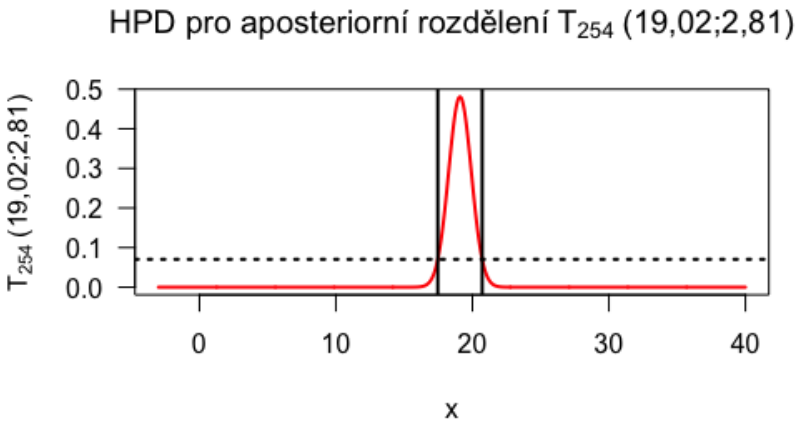
```
d = qt(1-0.025, df=2*alpha_n)*sqrt(rozpt_n) + mu_n
```

```
> c
[1] 17.43918
> d
[1] 20.74096
```

V tomto případě je symetrický věrohodnostní interval širší, než v případě, kdy jsme parametr σ^2 znali, což bychom očekávali. Věrohodnostní interval o nejvyšší aposteriorní hustotě, jak ukážou výpočty, bude o trochu užší.

```
apost.hustota.t = function(x) dmt(x, mu_n, rozpt_n)
apost.distr.fce.t = function(x) pmt(x, mu_n, rozpt_n)
hpd.int.t = hpd(apost.hustota.t, c(-3,40), cdf = apost.distr.fce.t,
     prob = 0.95)
print(hpd.int.t)
plot(hpd.int.t, las = 1, main = bquote("HPD pro aposteriorní
     rozdělení"~T[254]~"(18,95;39,74)"),
     lwd = 2, ylab = bquote(T[254]~"(18,95;39,74)"))
```

```
> print(hpd.int.t)
$lower
[1] 17.46374

$upper
[1] 20.7164

$prob
[1] 0.9500143
```



Obrázek 4.4: Vyobrazení věrohodnostního intervalu o nejvyšší aposteriorní hustotě při aposteriorním rozdělení parametru $\mu \sim T_{254}(19{,}02; 2{,}81)$.

Symetrický věrohodnostní interval je (17,44; 20,74). Věrohodnostní interval o nejvyšší aposteriorní hustotě je (17,46; 20,72).

Závěr

Cílem této práce bylo seznámení se se základními principy bayesovských metod a jejich aplikací v R. Začali jsme od jednoduchých příkladů na základní podobu Bayesovy věty a došli jsme k odhadům parametrů u binomického a normálního rozdělení. V práci jsme popsali, jak bayesovské metody fungují, co je to konjugované apriorní rozdělení a jak ho vybrat, dále jak spočítat aposteriorní rozdělení a odhadnout samotný parametr daného rozdělení.

Bayesovská statistika nám umožňuje využít apriorní informaci o datech, například zkušenost z minulosti. Když žádnou takovou informaci nemáme, můžeme užít neinformativní apriorní rozdělení, což jsou plochá rozdělení. Například u binomického rozdělení lze použít rovnoměrné apriorní rozdělení pro parametr θ. K vyjádření nejistoty bayesovské metody užívají podmíněnou pravděpodobnost. Ta je mnohem blíže k našemu běžnému smyslu užití slova pravděpodobnost. Parametr, který chceme odhadovat, považujeme za náhodný, data jsou pro nás pevná. U klasického přístupu je tato situace opačná.

Tato metodika má ovšem i svá úskalí. Vždy po nás požaduje apriorní rozdělení, odvozování aposteriorních rozdělení je dost složité. Výsledky samotné analýzy mohou být komplexnější, než kdybychom užili frekventistický přístup.

Nejtěžší na tomto tématu bylo porozumění principu bayesovských metod. Samotná práce není tak složitá, jak se na první pohled mohlo zdát, důležité je však tématu dobře porozumět. Hodně času mi také zabralo pochopit různé úpravy v odvozování, které nejsou jednoduché.

S bayesovskou statistikou jsem se dříve nesetkala, téma mi přijde velice zajímavé a myslím si, že tato práce může být vhodnou inspirací pro ty, kteří by se chtěli

seznámit s jejími základy a umět je aplikovat ve statistickém softwaru R. Práci doplnily i četné příklady tohoto softwaru, kde bylo jeho použití demonstrováno.

Literatura

[1] Azzalini, A. and Genz, A. (2016). *The R package 'mnormt': The multivariate normal and 't' distributions (version 1.5-5).* URL http://azzalini.stat.unipd.it/SW/Pkg-mnormt.

[2] Baptiste Auguie (2017). *gridExtra: Miscellaneous Functions for "Grid"Graphics.* R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

[3] Bernardo, J. M.: *Bayesian Statistics.* [online]. [cit. 2018-07-20]. Dostupné z: https://www.uv.es/bernardo/BayesStat.pdf.

[4] Bernardo, J. M.; Smith, A.F.M. *Bayesian Theory*, Wiley, 1993.

[5] Hron, K., Kunderová, P., Vencálek, O.: *Základy počtu pravděpodobnosti a metod matematické statistiky (3. vydání).* Univerzita Palackého, Olomouc, 2018.

[6] Chuchel, K.: *Základy bayesovské statistiky.* Univerzita Karlova v Praze Matematicko-fyzikální fakulta, Praha, 2012.

[7] Jackman, S.: *Bayesian Analysis for the Social Sciences.* Wiley, Stanford, 2009.

[8] Junker, B.: *Basics of Bayesian Statistics.* [online]. [cit. 2018-07-20]. Dostupné z: www.stat.cmu.edu/ brian/463-663/week09/Chapter%2003.pdf

[9] Kaggle - Datasets [online]. [cit. 2017-04-23]. Dostupné z: https://www.kaggle.com/borapajo/food-choices.

[10] Kruschke, J. K.: *Doing Bayesian Data Analysis: A Tutorial with R and BUGS.* Academic Press, Orlando, 2010.

[11] Marin, J. M., Robert, Ch.: *Bayesian Essentials with R (2. vydání).* Springer, New York, 2014.

[12] Murphy, K. P., *Conjugate Bayesian analysis of the Gaussian distribution.* [online]. [cit. 2019-03-27]. Dostupné z: https://www.cs.ubc.ca/ murphyk/Papers/bayesGauss.pdf.

[13] Muth, N. D.: *What are the guidelines for percentage of body fat loss?* [online]. [cit. 2019-01-25]. Dostupné z: https://www.acefitness.org/education-and-resources/lifestyle/blog/112/what-are-the-guidelines-for-percentage-of-body-fat-loss.

[14] Nazir, N., Athar Ali Kahn A. H. Mir, Maqbool, S.: *Applications of R Software in Bayesian Data Analysis.* [online]. [cit. 2018-07-20]. Dostupné z: https://www.acefitness.org/education-and-resources/lifestyle/blog/112/what-are-the-guidelines-for-percentage-of-body-fat-loss.

[15] R Core Team (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[16] Pham-Gia, T., Turkkan, N.: *Computation of the highest posterior density interval in bayesian analysis.* [online]. [cit. 2019-02-04]. Dostupné z: https://doi.org/10.1080/00949659308811461.

[17] Revelle, W.: *An introduction to psychometric theory with applications in R.* [online]. [cit. 2018-11-20]. Dostupné z: http://www.personality-project.org/r/book/chapter4.pdf.

[18] Sego, L. H. (2016). *Smisc: Sego Miscellaneous.* A collection of functions for statistical computing and data manipulation in R. Pacific Northwest National Laboratory. https://pnnl.github.io/Smisc.

[19] Statisticat, LLC. (2018). *LaplacesDemon: Complete Environment for Bayesian Inference.* Bayesian-Inference.com. R package version 16.1.1. https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software.

[20] Steorts, R. C.: *More on Bayesian Methods: Part II.* [online]. [cit. 2019-02-04]. Dostupné z: https://www.coursehero.com/file/17506487/lecture5-more-bayes/.

[21] Verde, V. E.: *Intrudoction to Bayesian Data Analysis using R and WinBUGS.* [online]. [cit. 2018-10-09]. Dostupné z: https://is.muni.cz/el/1431/jaro2013/M8BDA/um/Lecture 1 - 9.

[22] Verzani, J. (2018). *UsingR: Data Sets, Etc. for the Text "Using R for Introductory Statistics, Second Edition.* R package version 2.0-6. https://CRAN.R-project.org/package=UsingR.

[23] Wickham, H.: (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag, New York. http://ggplot2.org.