

Vysoká škola ekonomická v Praze  
Fakulta informatiky a statistiky



**Odhad relativní četnosti binomického  
rozdělení pomocí klasického a  
bayesovského přístupu v jazyce R**

**BAKALÁŘSKÁ PRÁCE**

Studijní program: [Data Analytics]

Autor: [Bc. Michal Lauer]

Vedoucí práce: [Ing. Ondřej Vilikus, Ph.D.]

Praha, Prosinec 2024

## **Poděkování**

Děkuji svému vedoucímu za odborné vedení práce a průběžné konzultace a své přítelkyni za neocenitelnou podporu.

## **Abstrakt**

Abstrakt.

## **Klíčová slova**

Bayesovská statistika, odhad relativní četnosti, jazyk R

## **Abstract**

Abstract.

## **Keywords**

Bayesian statistics, relative frequency estimation, R language

# Table of contents

|   |           |
|---|-----------|
| <b>Úvod</b>   | <b>9</b>  |
| <b>1 Statistické metody</b>                                 | <b>10</b> |
| 1.1 Inference . . . . .                                     | 10        |
| 1.1.1 Problematika výběrových šetření . . . . .             | 10        |
| 1.2 Frekventistická inference . . . . .                     | 10        |
| 1.2.1 Testování hypotéz . . . . .                           | 10        |
| 1.2.2 Metriky při testování hypotéz . . . . .               | 10        |
| 1.2.3 Jednovýběrový odhad poměru s velkým vzorkem . . . . . | 10        |
| 1.2.4 Jednovýběrový odhad poměru s malým vzorkem . . . . .  | 10        |
| 1.3 Bayesovská inference . . . . .                          | 11        |
| <b>2 Monte Carlo generování</b>                             | <b>12</b> |
| 2.1 Vyhodnocení generovaného rozdělení . . . . .            | 12        |
| 2.1.1 Vyhodnocení hypotéz . . . . .                         | 12        |
| 2.1.2 Odhad poměru . . . . .                                | 12        |
| <b>3 Praktické odhady</b>                                   | <b>13</b> |
| 3.1 Balíčky pro frekventistickou inferenci . . . . .        | 13        |
| 3.1.1 Klasické test poměru . . . . .                        | 13        |
| 3.2 Software pro bayesovskou statistiku . . . . .           | 15        |
| 3.2.1 Balíček R2WinBUGS . . . . .                           | 15        |
| 3.2.2 Balíček jags . . . . .                                | 18        |
| 3.2.3 stan . . . . .  | 18        |
| 3.3 Simulace . . . . .                                      | 18        |
| 3.3.1 Malý vzorek . . . . .                                 | 19        |
| 3.3.2 Velký vzorek . . . . .                                | 19        |
| 3.3.3 Porovnání výsledků . . . . .                          | 19        |
| <b>Závěr</b>  | <b>20</b> |
| 3.4 Jak citovat v textu . . . . .                           | 20        |
| <b>Použitá literatura</b>                                   | <b>21</b> |
| <b>A Bayesovské modely</b>                                  | <b>23</b> |

# List of Figures

# List of Tables

# Seznam zdrojových kódů

|     |                        |    |
|-----|------------------------|----|
| 3.1 | Moje caption . . . . . | 13 |
| 3.2 | NENÍ CAPTION . . . . . | 13 |
| 3.3 | NENÍ CAPTION . . . . . | 14 |
| 3.4 | NENÍ CAPTION . . . . . | 15 |
| 3.5 | NENÍ CAPTION . . . . . | 16 |
| 3.6 | NENÍ CAPTION . . . . . | 16 |
| 3.7 | NENÍ CAPTION . . . . . | 17 |
| 3.8 | NENÍ CAPTION . . . . . | 17 |
| A.1 | NENÍ CAPTION . . . . . | 23 |

# Seznam použitých zkratek

**BCC** Blind Carbon Copy

**CC** Carbon Copy

**CERT** Computer Emergency Response  
Team

**CSS** Cascading Styleheets

**DOI** Digital Object Identifier

**HTML** Hypertext Markup Language

**REST** Representational State Transfer

**SOAP** Simple Object Access Protocol

**URI** Uniform Resource Identifier

**URL** Uniform Resource Locator

**XML** eXtended Markup Language



# Úvod

Tohle je **úvodní** *text*.

# 1. Statistické metody

Krátký úvod do historie, bayes, inferenční bayes (rozdělení) vs. inference (bod) citace Karla

## 1.1 Inference

proč to používáme, výběr vs. populace, reprezentativnost

### 1.1.1 Problematika výběrových šetření

reprezentativnost, definice populace, čas sběru, organizace sběru...

## 1.2 Frekventistická inference

Jak to funguje, jak to spoléhá na sampling distributions

### 1.2.1 Testování hypotéz

hladina významnosti, úroveň spolehlivosti, Testovací statistika, kritický obor, 1/2 stranný test p-hodnota, interval spolehlivosti

### 1.2.2 Metriky při testování hypotéz

Chyba I. a II. druhu, síla testu, velikost efektu

### 1.2.3 Jednovýběrový odhad poměru s velkým vzorkem

použití, předpoklady, poměrový Z test, binomický test, síla testu, velikost efektu

### 1.2.4 Jednovýběrový odhad poměru s malým vzorkem

Proč jsou důležité speciální metody, nějaké typy (wiki)

## 1.3 Bayesovská inference

Odvození bayesova vzorce, popis likelihood/aprior/data, druhy aprior/posterior

## 2. Monte Carlo generování

Halsing, Gibbs, HMC

### 2.1 Vyhodnocení generovaného rozdělení

korelace, ESS, monte carlo error...

#### 2.1.1 Vyhodnocení hypotéz

Interval kredibility, ROPE, Bayesův faktor

#### 2.1.2 Odhad poměru

# 3. Praktické odhady

## 3.1 Balíčky pro frekventistickou inferenci

### 3.1.1 Klasické test poměru

Test

test

stats::t.test()

test

Jednoduchý T-test

```
1 set.seed(78)
2 x <- rnorm(n = 100, mean = 1, sd = 1)
3 t.test(x = x, mu = 0,
4       alternative = "two.sided", conf.level = 0.95)
```

Výpis 3.1: Moje caption

One Sample t-test

data: x

t = 8.8438, df = 99, p-value = 3.621e-14

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

0.7158758 1.1300282

sample estimates:

mean of x

0.922952

3.1

Simulace alfa = chyba 1. druhu

```
1 library(dplyr)
```

Výpis 3.2: NENÍ CAPTION

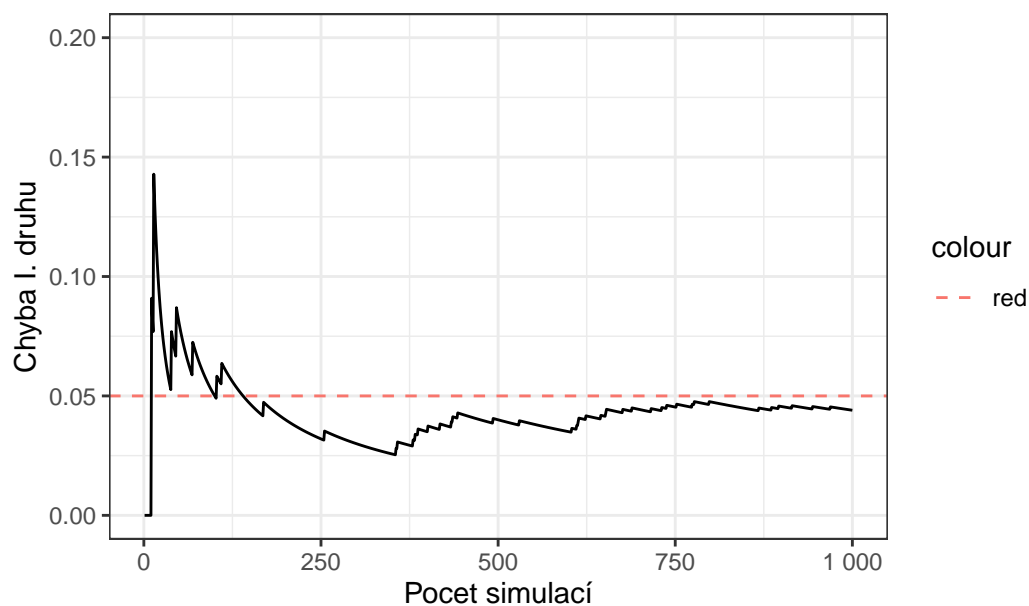
```

1 library(ggplot2)
2 # Nastavení
3 K <- 1000 # Pocet simulaci
4 n <- 200 # Velikost vzorku
5 mu0 <- 0 # Skutecny prumer
6 sd0 <- 2 # čPopulání ěsmrodatná odchylka
7 alpha <- 0.05 # Hladina významnosti
8 set.seed(639)
9 vzorky <- tibble()
10
11 # Simulace
12 for (i in seq_len(K)) {
13   x <- rnorm(n = n, mean = mu0, sd = sd0)
14   test <- t.test(x = x, mu = mu0, conf.level = 1 - alpha)
15   vzorky <-> bind_rows(vzorky, tibble(n = i, vysledek = test$p.value <= alpha))
16 }
17 vzorky$cvysledky <- cummean(vzorky$vysledek)
18 ggplot(vzorky, aes(x = n, y = cvysledky)) +
19   geom_line() +
20   geom_hline(aes(yintercept = .05, color = "red"), linetype = "dashed") +
21   scale_y_continuous(limits = c(0, .2)) +
22   scale_x_continuous(labels = scales::label_number()) +
23   theme_bw() +
24   labs(
25     title = "Procento šfalených zamítnutí h0 se žblíí ěhladin významnosti",
26     y = "Chyba I. druhu",
27     x = "čPoet simulací"
28   )

```

Výpis 3.3: NENÍ CAPTION

### Procento falešných zamítnutí $h_0$ se blíží hladině významnosti



## 3.2 Software pro bayesovskou statistiku

### 3.2.1 Balíček R2WinBUGS

podporuje WinBUGS, OpenBUGS

```
1 set.seed(123)
2 x <- rbinom(10, 1, .6)
3
4 bugs <- R2WinBUGS::bugs(
5   data = list(
6     N      = length(x), # počet pozorování
7     x      = x,         # vstupní data
8     alpha  = 0.01,      # hodnota parametru alpha
9     beta   = 0.01       # hodnota parametru beta
10  ),
11  # číselné hodnoty
12  inits = list(
13    list(p = 0.5),
14    list(p = 0.5)
15  ),
16  n.chains = 2, n.iter = 5000, n.burnin = 1000, n.thin = 1,
17  # parametry, které žuloit
18  parameters.to.save = c("p"),
19  # cesta k modelu
```

```

20 working.directory = "kapitoly/prakticka",
21 model.file = "r2winbugs.txt",
22 # Cesta k programu WinBUGS
23 bugs.directory = r"(C:\Users\Mike\Downloads\WinBUGS14\WinBUGS14)",
24 # ůOdstra pracovní soubory
25 clearWD = T,
26 # Replikovatelnost
27 bugs.seed = 123
28 )

```

Výpis 3.4: NENÍ CAPTION

Výsledek

```

1 print(bugs)

```

Výpis 3.5: NENÍ CAPTION

```

Inference for Bugs model at "r2winbugs.txt", fit using WinBUGS,
  2 chains, each with 5000 iterations (first 1000 discarded)
n.sims = 8000 iterations saved

      mean  sd 2.5%  25%  50%  75%  97.5% Rhat n.eff
p          0.6 0.1  0.3  0.5  0.6  0.7   0.9    1  4500
deviance 14.5 1.5 13.5 13.6 13.9 14.8  18.8    1  8000

```

For each parameter, n.eff is a crude measure of effective sample size, and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule,  $pD = \bar{D} - \hat{D}$ )

$pD = 1.0$  and  $DIC = 15.6$

DIC is an estimate of expected predictive error (lower deviance is better).

Odhad parametru p.

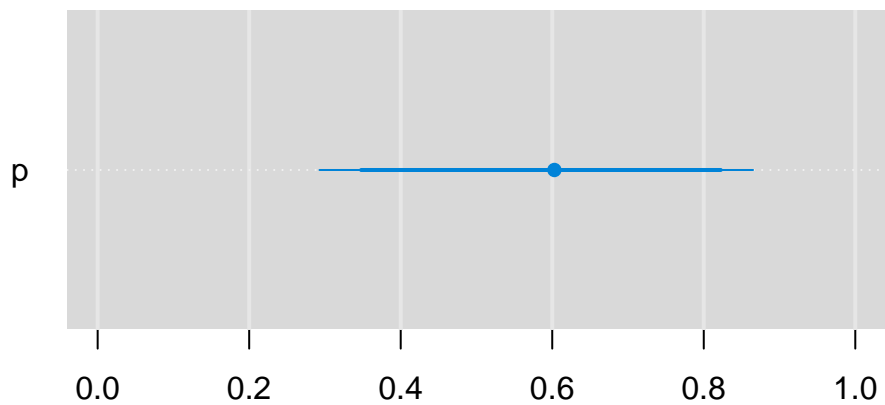
```

1 mcmcplots::caterplot(mcmcout = bugs,           # Výstup modelu
2                      parms = "p",             # Vybraný parametr
3                      val.lim = c(0, 1),        # Limity na ose X
4                      quantiles = list(
5                        outer = c(0.025, 0.975), # 95% interval kredibility
6                        inner = c(0.055, 0.945) # 89% interval kredibility
7                      )
8 )

```

Výpis 3.6: NENÍ CAPTION

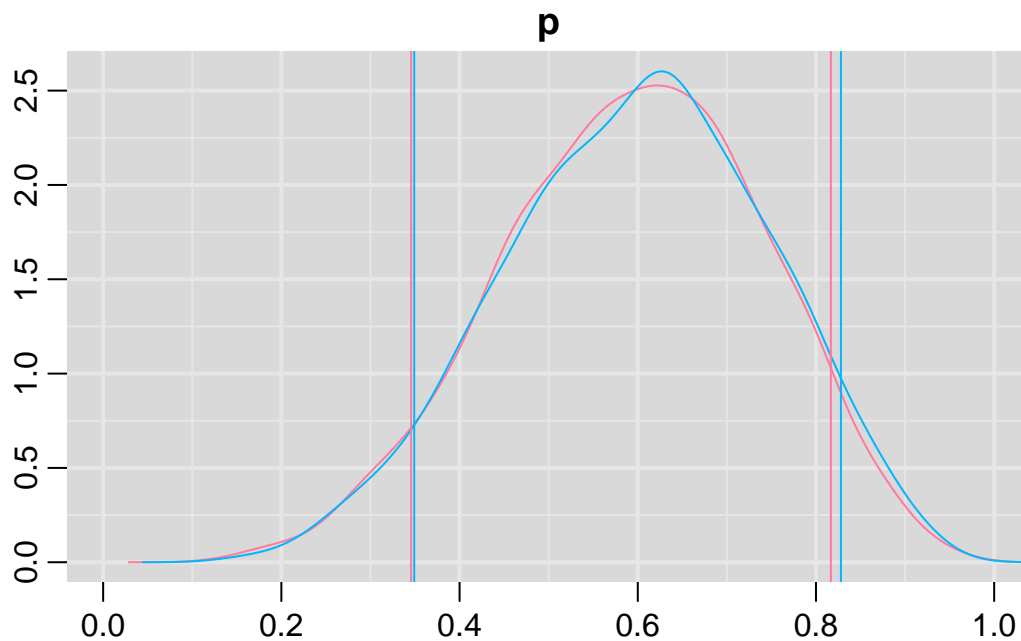




Posteriorní rozdělení jednotlivých chainů.

```
1 mcmcplots::denplot(mcmcout = bugs, # Výstup modelu
2     parms = "p", # Vybraný parametr
3     xlim = c(0, 1), # Limity na ose X
4     ci = 0.89 # 89% Interval credibility
5 )
```

Výpis 3.7: NENÍ CAPTION



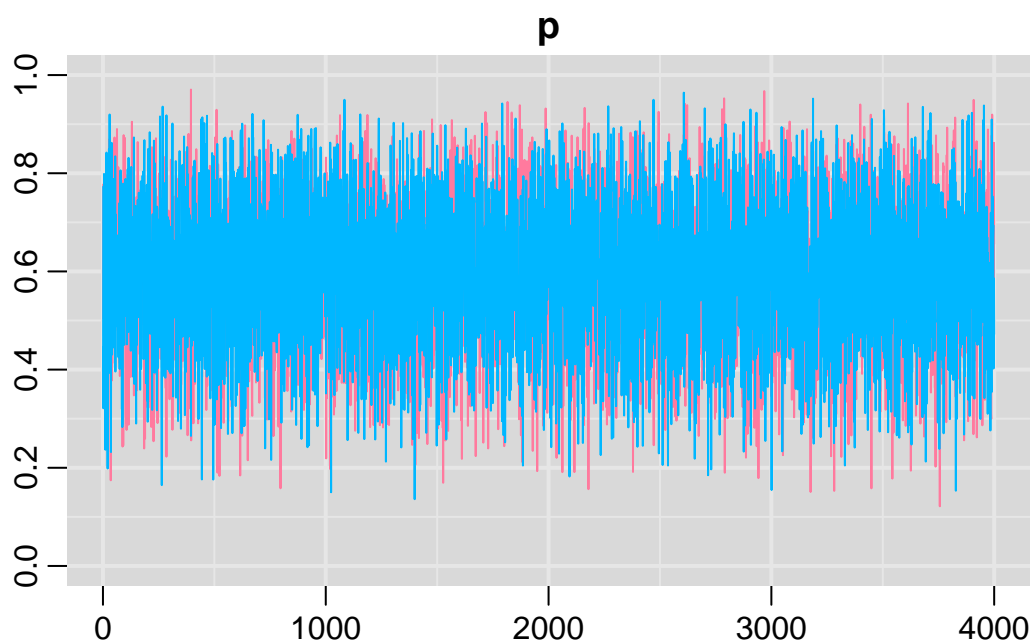
Vývoj jednotlivých chainů.

```

1 mcmcplots::traplot(mcmcout = bugs, # Výstup modelu
2                       parms = "p",   # Vybraný parametr
3                       ylim = c(0, 1) # Limity na ose Y
4 )

```

Výpis 3.8: NENÍ CAPTION



### 3.2.2 Balíček jags

aplikace, R implementace, výhody/nevýhody, používá gibse

### 3.2.3 stan

aplikace, R implementace, výhody/nevýhody, používá hmc

## 3.3 Simulace

jak budou simulace provedné, jak budou vyhodnocené, nastavení ROPE/alternativ. pro odhad chyb

### **3.3.1 Malý vzorek**

Bayes vs. vybraný vzorec vs. binomic

### **3.3.2 Velký vzorek**

Bayes vs. vybraný vzorec vs. binomic

### **3.3.3 Porovnání výsledků**

Jak testy dopadly

# Závěr

Konec práce, závěr.

## 3.4 Jak citovat v textu

`\parencite {Cermak2018}` → (Čermák & Smutný, 2018)  
`\parencite {Hladik2018,Jasek2018}` → (Hladík & Černý, 2018; Jašek et al., 2018)  
`\parencite [kap. 3]{Pecakova2018}` → (Pecáková, 2018, kap. 3)

# Použitá literatura

- Čermák, R., & Smutný, Z. (2018). A Framework for Cultural Localization of Websites and for Improving Their Commercial Utilization. In *Global Observations of the Influence of Culture on Consumer Buying Behavior* (s. 206–232). IGI Global. <https://doi.org/10.4018/978-1-5225-2727-5.ch013>
- Hladík, M., & Černý, M. (2018). The Shape of the Optimal Value of a Fuzzy Linear Programming Problem. *Fuzzy Logic in Intelligent System Design*, 281–286. [https://doi.org/10.1007/978-3-319-67137-6\\_31](https://doi.org/10.1007/978-3-319-67137-6_31)
- Jašek, P., Vraná, L., Šperková, L., Smutný, Z., & Kobulský, M. (2018). Modeling and Application of Customer Lifetime Value in Online Retail. *Informatics*, 5(1). <http://www.mdpi.com/2227-9709/5/1/2/pdf>
- Pecáková, I. (2018). *Statistika v terénních průzkumech*. Professional Publishing.

## **Přílohy**

# A. Bayesovské modely

```
1 print(1 + 1)
```

Výpis A.1: NENÍ CAPTION

```
[1] 2
```