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

Obsah

Seznam obrázků

Seznam tabulek

Seznam zdrojových kódů

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

V úvodu závěrečné práce autor vysvětlí, proč si vybral zvolené téma, tedy **motivaci** celé závěrečné práce. V úvodu nesmí chybět přesně formulovaný **hlavní cíl** závěrečné práce (popř. dílčí cíle), měla by zde být nastíněna **metodika** celé závěrečné práce (popř. výzkumné otázky či hypotézy). Zvykem bývá rovněž nastínit **hlavní výsledky/výstupy** závěrečné práce.

Po úvodu následují jednotlivé **číslované kapitoly** členěné do podkapitol.

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, Gibs, HMC

2.1 Vyhodnocení generovaného rozdělení

korelace, ESS, monte carlo error...

2.1.1 Vyhodocení hypotéz

Interval kredibility, ROPE, Bayesův faktor

2.1.2 Odhad poměru

3. Praktické odhady

Test

test

`stats::t.test()`

test

Jednoduchý T-test

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

```
    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
```

Simulace alfa = chyba 1. druhu

```
library(dplyr)
```

```
library(ggplot2)
```

```
# Nastavení
K   <- 1000   # Počet simulací
n   <- 200    # Velikost vzorku
mu0 <- 0      # Skutečný průměr
sd0 <- 2      # Populační směrodatná odchylka
alpha <- 0.05 # Hladina významnosti
set.seed(639)
```

```r
vzorky <- tibble()

# Simulace
for (i in seq_len(K)) {
    x <- rnorm(n = n, mean = mu0, sd = sd0)
    test <- t.test(x = x, mu = mu0, conf.level = 1 - alpha)
    vzorky <<- bind_rows(vzorky, tibble(n = i, vysledek = test$p.value <= alpha))
}
vzorky$cvysledky <- cummean(vzorky$vysledek)
ggplot(vzorky, aes(x = n, y = cvysledky)) +
    geom_line() +
    geom_hline(aes(yintercept = .05, color = "red"), linetype = "dashed") +
    scale_y_continuous(limits = c(0, .2)) +
    scale_x_continuous(labels = scales::label_number()) +
    theme_bw() +
    labs(
        title = "Procento falešných zamítnutí h0 se blíží hladině významnosti",
        y = "Chyba I. druhu",
        x = "Počet simulací"
    )
```

3.1 Balíčky pro frekventistickou inferenci

3.1.1 base r

Test

test

`stats::t.test()`

test

Jednoduchý T-test

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

```
    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
```

Simulace alfa = chyba 1. druhu

```
library(dplyr)
library(ggplot2)
# Nastavení
K   <- 1000   # Počet simulací
n   <- 200    # Velikost vzorku
mu0 <- 0      # Skutečný průměr
sd0 <- 2      # Populační směrodatná odchylka
alpha <- 0.05 # Hladina významnosti
set.seed(639)
vzorky <- tibble()
```

```r
# Simulace
for (i in seq_len(K)) {
    x <- rnorm(n = n, mean = mu0, sd = sd0)
    test <- t.test(x = x, mu = mu0, conf.level = 1 - alpha)
    vzorky <<- bind_rows(vzorky, tibble(n = i, vysledek = test$p.value <= alpha))
}
vzorky$cvysledky <- cummean(vzorky$vysledek)
ggplot(vzorky, aes(x = n, y = cvysledky)) +
    geom_line() +
    geom_hline(aes(yintercept = .05, color = "red"), linetype = "dashed") +
    scale_y_continuous(limits = c(0, .2)) +
    scale_x_continuous(labels = scales::label_number()) +
    theme_bw() +
    labs(
        title = "Procento falešných zamítnutí h0 se blíží hladině významnosti",
        y = "Chyba I. druhu",
        x = "Počet simulací"
    )
```

3.1.2 easystats

použití, výhody/nevýhody

3.1.3 inferencer

použití, výhody/nevýhody

3.2 Software pro bayesovskou statistiku

3.2.1 Balíček R2WinBUGS

podporuje WinBUGS, OpenBUGS

```
set.seed(123)
x <- rbinom(10, 1, .6)

bugs <- R2WinBUGS::bugs(
    data = list(
        N     = length(x), # Počet pozorování
        x     = x,         # Vstupní data
        alpha = 0.01,      # Hodnota parametru alpha
        beta  = 0.01       # Hodnota parametru beta
    ),
    # Počáteční hodnoty
    inits = list(
        list(p = 0.5),
        list(p = 0.5)
    ),
    n.chains = 2, n.iter = 5000, n.burnin = 1000, n.thin = 1,
    # Parametry, které uložit
    parameters.to.save = c("p"),
    # Cesta k modelu
    working.directory = "prakticka",
    model.file = "r2winbugs.txt",
    # Cesta k programu WinBUGS
    bugs.directory = r"(C:\Users\Mike\Downloads\WinBUGS14\WinBUGS14)",
    # Odstraň pracovní soubory
    clearWD = T,
    # Replikovatelnost
    bugs.seed = 123
)
```

Výsledek

```
print(bugs)
```

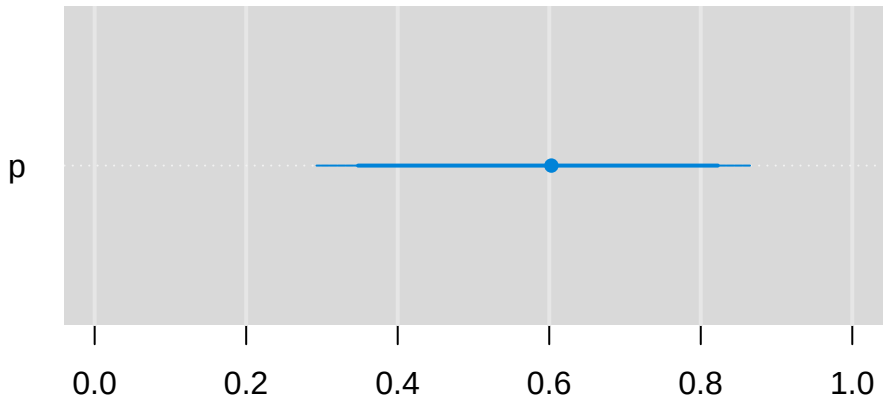
```
Inference for Bugs model at "r2winbugs.txt", fit using WinBUGS,
 2 chains, each with 5000 iterations (first 1000 discarded)
 n.sims = 8000 iterations saved
          mean  sd 2.5%  25%  50%  75% 97.5% Rhat n.eff
p         0.6 0.1  0.3  0.5  0.6  0.7   0.9    1  4500
deviance 14.5 1.5 13.5 13.6 13.9 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 = Dbar-Dhat)
pD = 1.0 and DIC = 15.6
DIC is an estimate of expected predictive error (lower deviance is better).
```

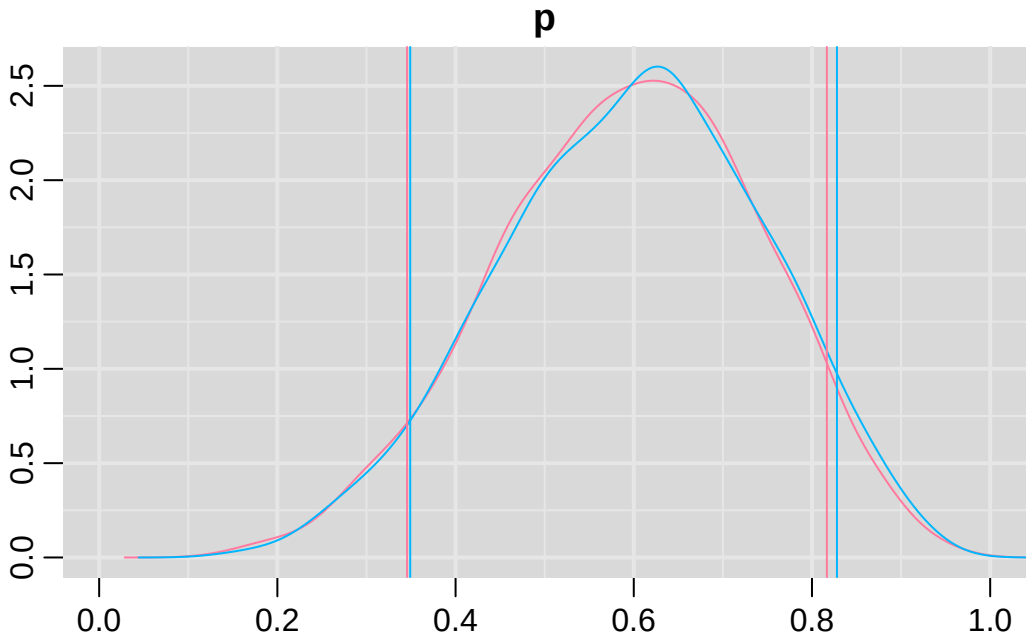
Odhad parametru p.

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



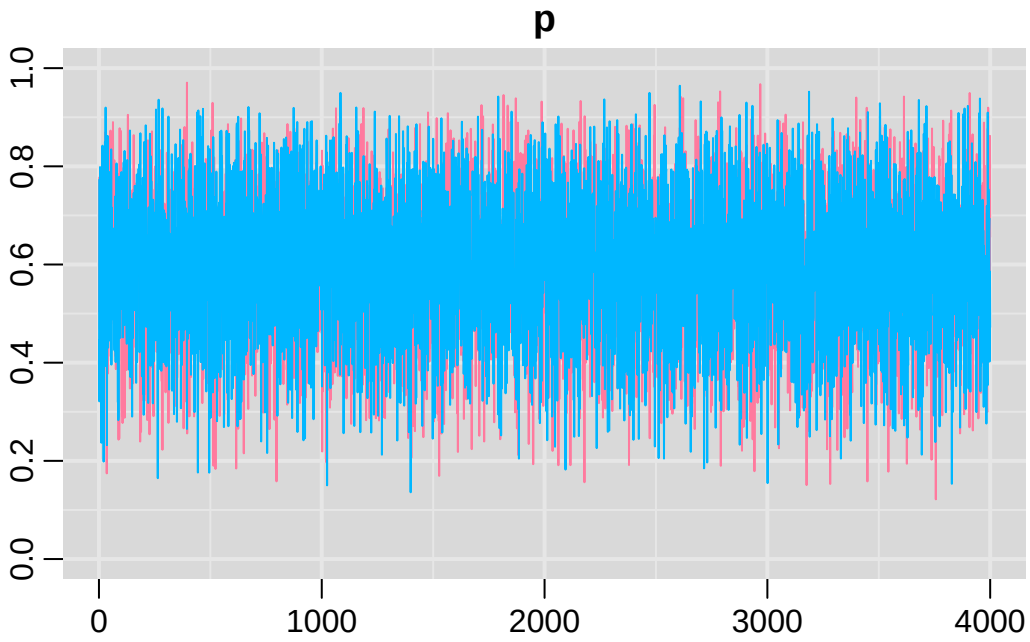
Posteriorní rozdělení jednotlivých chainů.

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



Vývoj jednotlivých chainů.

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



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