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

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, 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

3.1 Balíčky pro frekventistickou inferenci

3.1.1 Klasické test poměru

Test

test

```
stats::t.test()
```

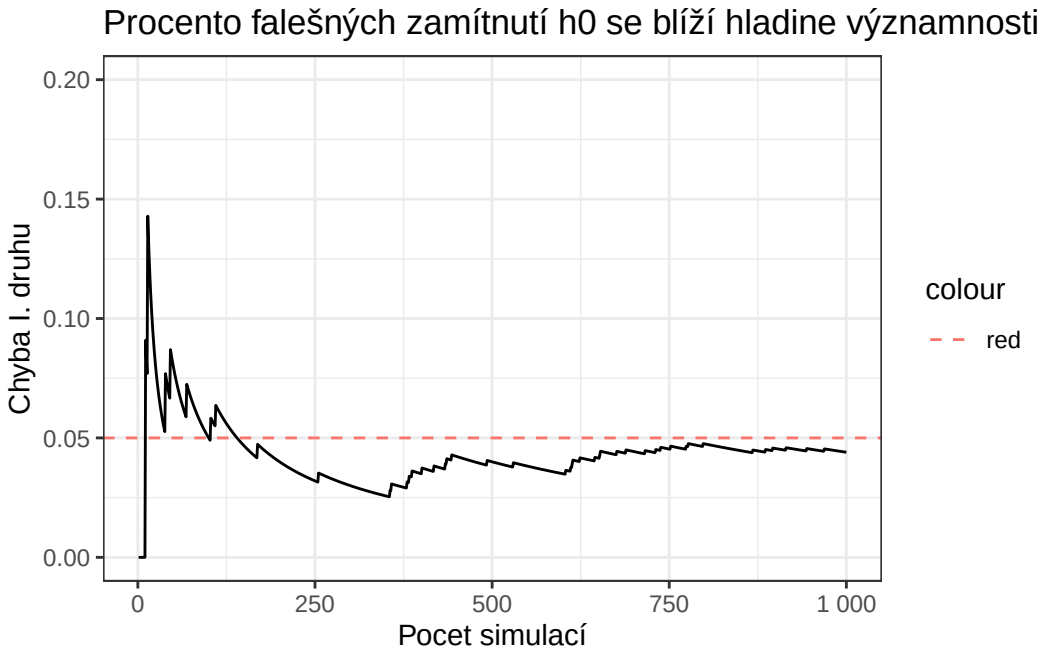
test

Jednoduchý T-test

```
    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



3.2 Software pro bayesovskou statistiku

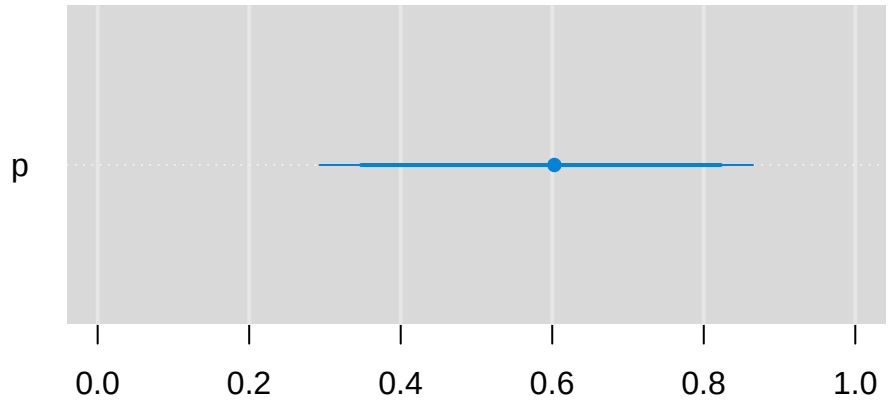
3.2.1 Balíček R2WinBUGS

podporuje WinBUGS, OpenBUGS. Při renderování se otevírá program winbugs. Nefunguje správně initializace chainů. Pro grafy pomocný balíček {mcmcplots}.

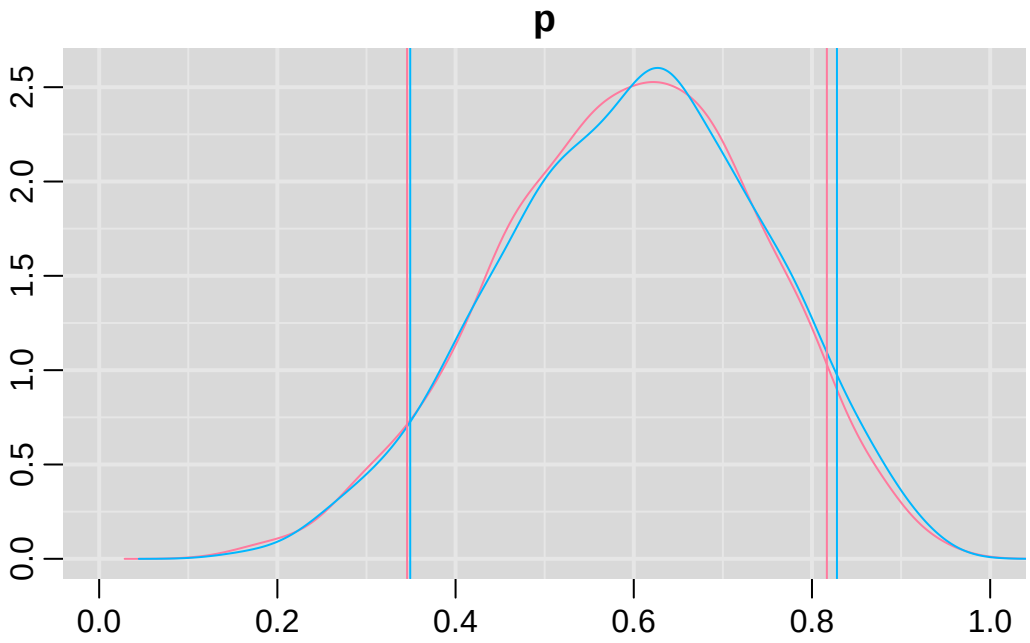
Výsledek

```
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).
```

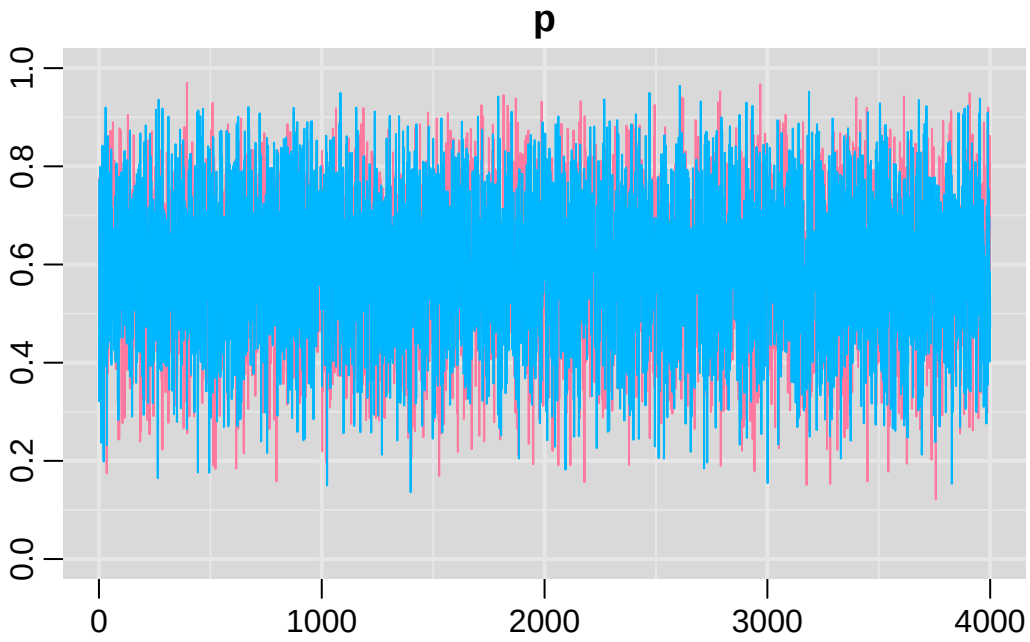
Odhad parametru p.



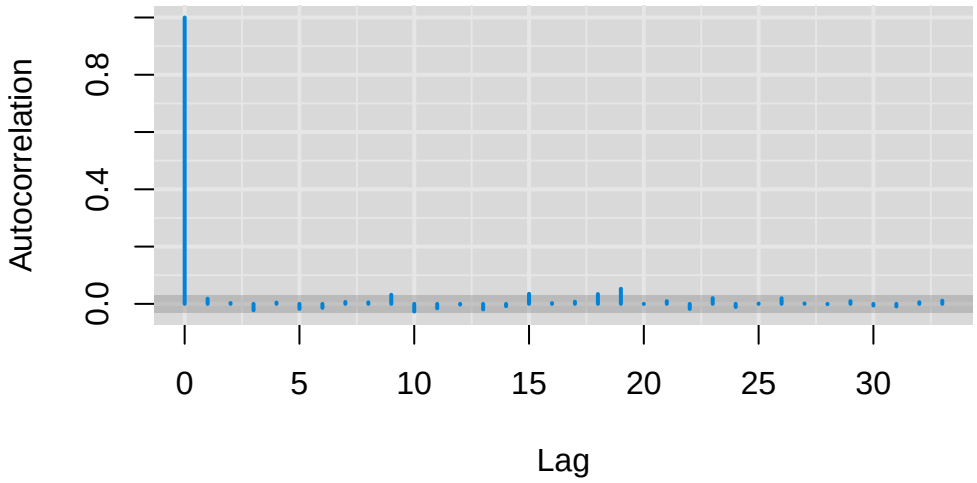
Posteriorní rozdělení jednotlivých chainů.

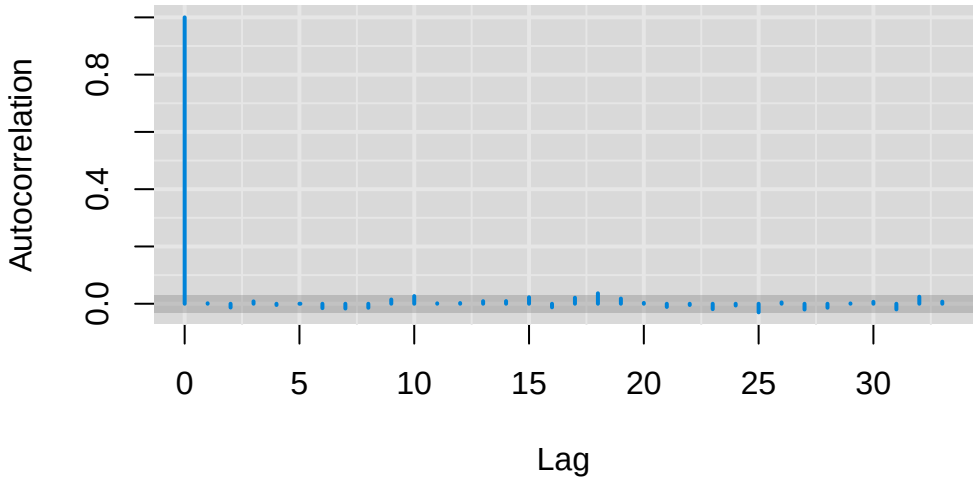


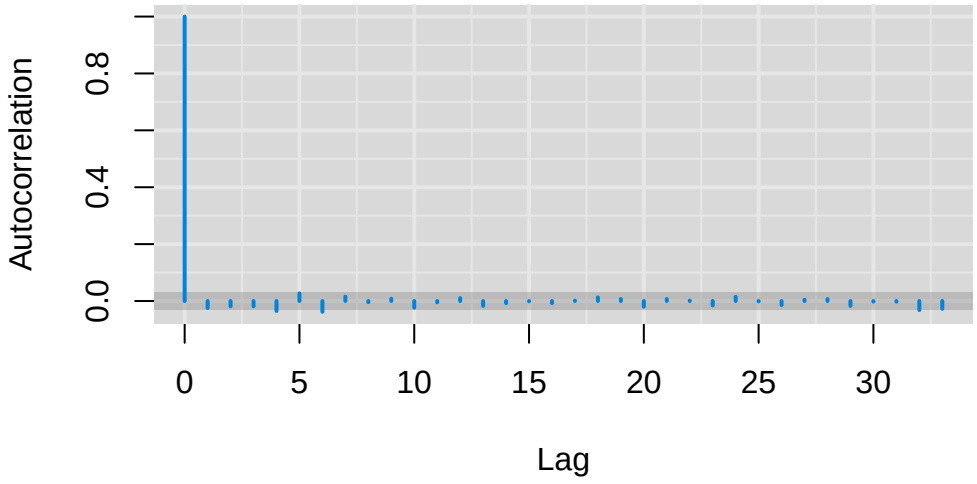
Vývoj jednotlivých chainů.

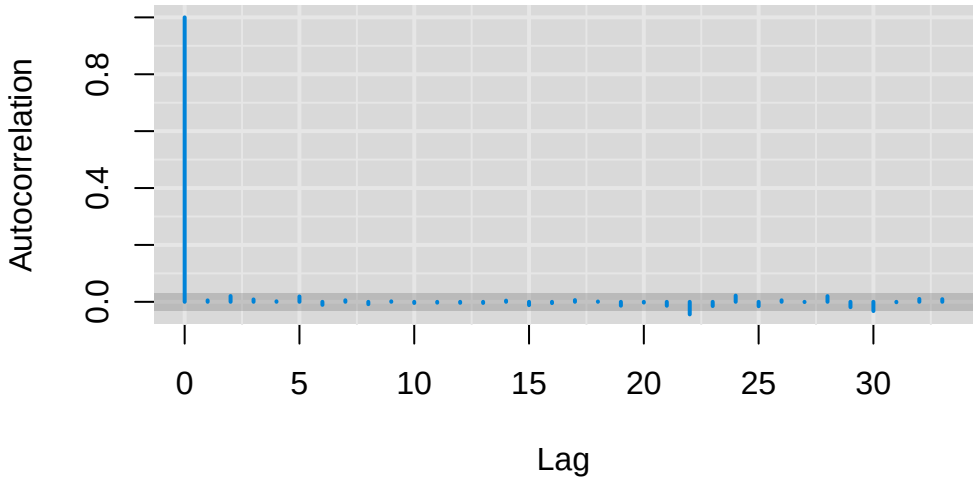


Autokorelace.




Autocorrelation
0.0
0.4
0.8
0
5
10
15
20
25
30
Lag


Autocorrelation
0.0
0.4
0.8
0
5
10
15
20
25
30
Lag



3.2.2 Balíček jags

rjags

Tvorba modelu

Adaptační doba, která se volá automaticky.

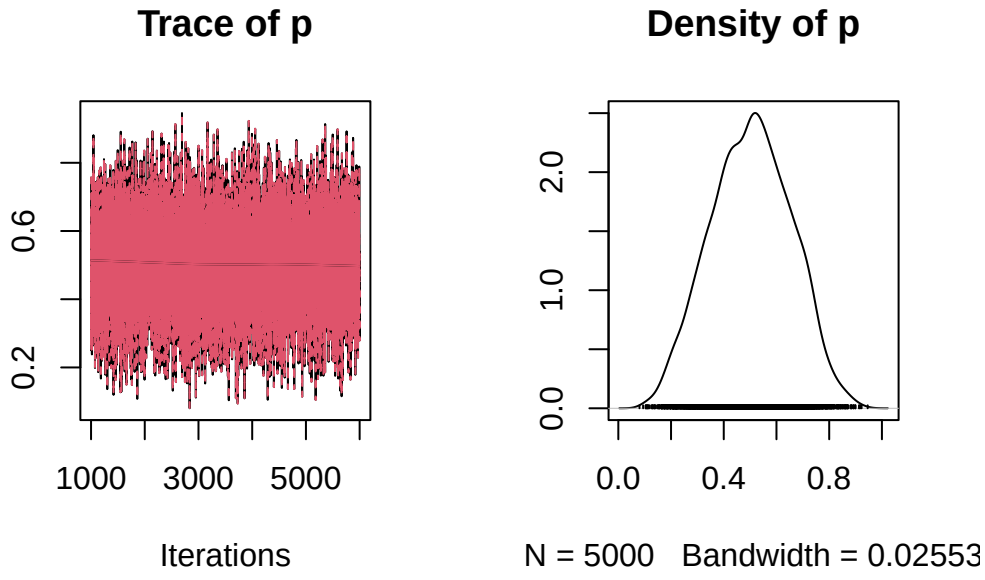
Burn-in generování, je to pro každý chainu.

Generování vzorků z každého chainu.

```
Iterations = 1001:6000
Thinning interval = 1
Number of chains = 2
Sample size per chain = 5000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean             SD       Naive SE Time-series SE
       0.50357        0.15197         0.00152        0.00152

2. Quantiles for each variable:
```

```
  2.5%    25%    50%    75%  97.5%
0.2104 0.3954 0.5061 0.6127 0.7879
```

Základní plot



R2jags

Divně spojený bugs and jags.

- Lze komplikovaně nastavit stejný seed

for (i in 1:n.chains) { init.values[[i]] <- inits[[i]] init.values[[i]].$RNG.name < -RNGnameinit.values$[[$i$]].RNG <- runif(1, 0, 2^31) }

(asi by to šlo nastavit seed a pak to generovat setjně pomocí runif i nahoře)

- adapt = burnin nebo adapt = 100

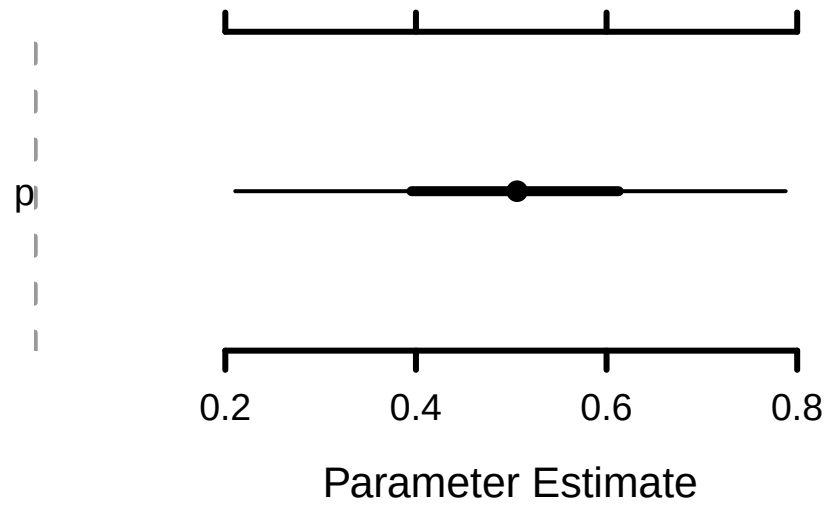
if (n.burnin > 0) { n.adapt <- n.burnin } else { n.adapt <- 100 }

- Lze paralelizovat pomocí jags.parallel

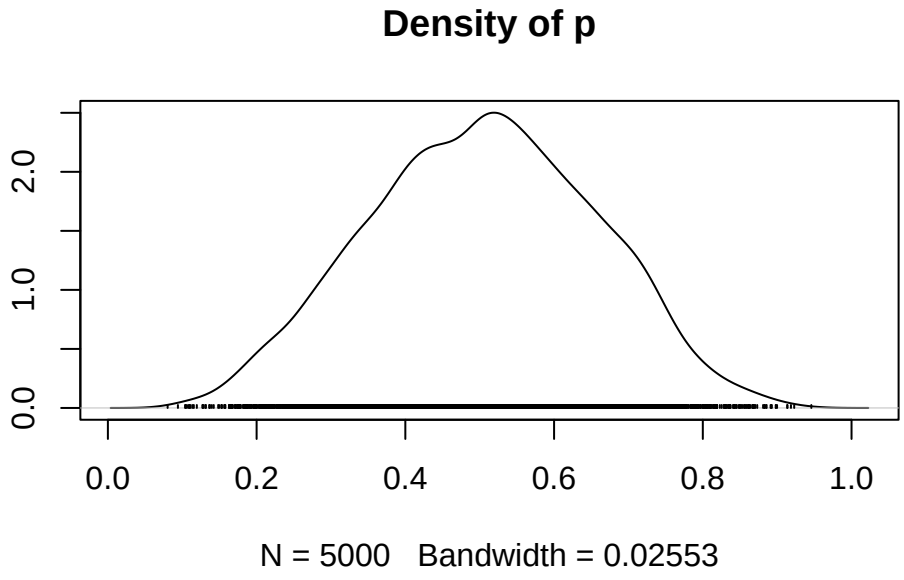
Výsledky jsou pořád ze stejného posteriorního rozdělení a jsou validní, akorát se charakteristiky nerovnají.

Visualizace

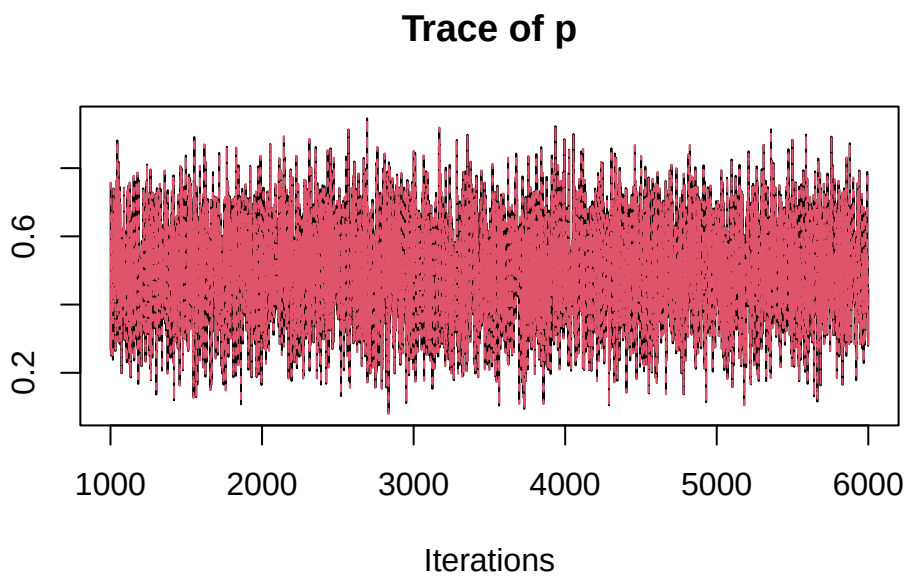
Catterplot.



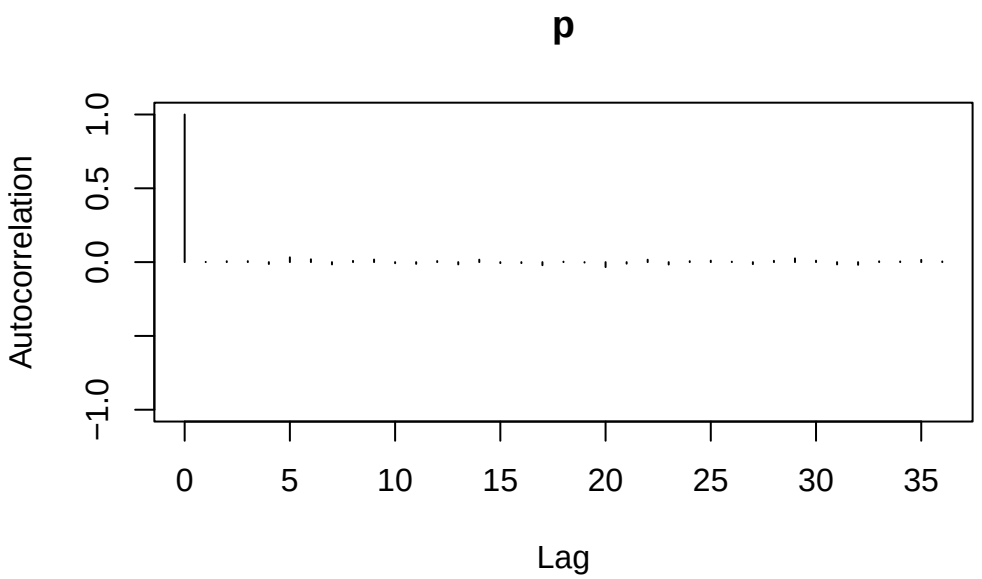
Posteriorní rozdělení.

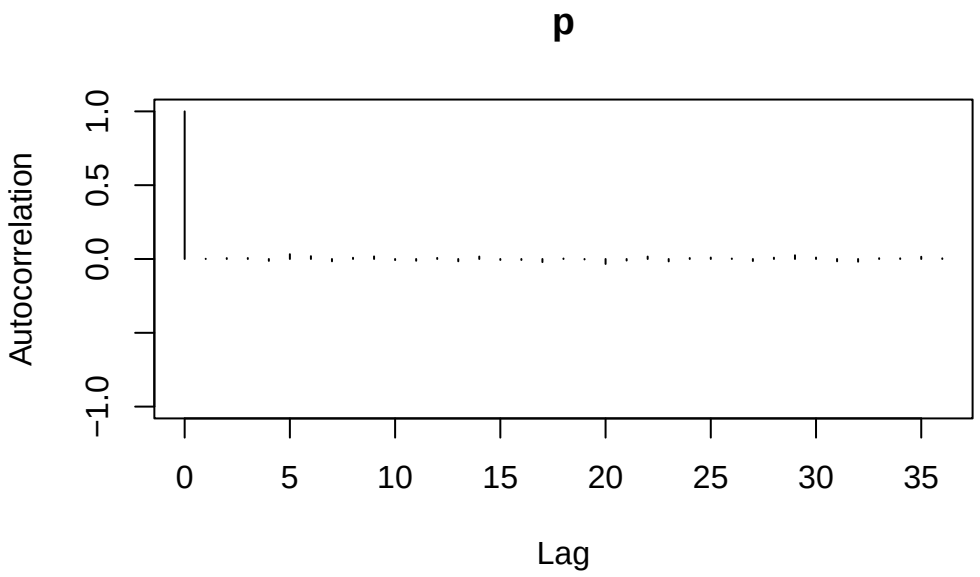


Trace plot.



Autokorelace.





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

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

```
model {
    for (i in 1:N) {
        x[i] ~ dbern(p)
    }

    p ~ dbeta(alpha, beta)
}
```

Výpis A.1: Winbugs