



Univerza v Ljubljani
Fakulteta za računalništvo
in informatiko

Akademija FRI

Bayesova statistika s programskim jezikom Stan

Erik Štrumbelj
Jure Demšar

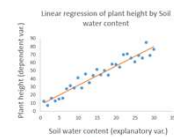
<https://github.com/bstatcomp/Stan-Intro-Workshop>

30. 5. 2019

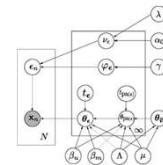
1

Zakaj naj mi bo mar za probabilistično programiranje?

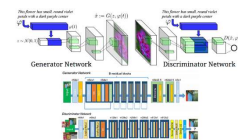
- Temelj statističnega modeliranja in probabilističnega strojnega učenja.
- Prihodnost "podatkovnega inženirstva".
- **Obvezno orodje** za vsakega, ki se želi resno ukvarjati s kvantitativno analizo podatkov!



uporabna statistika



probabilistični grafični modeli



(generativno) globoko učenje

2

Oris vsebine

- 1** Negotovost in probabilistično razmišljanje,
- 2** statistično modeliranje,
- 3** probabilistično programiranje,
- 4** programski jezik Stan,
- 5** praktični del.

Predpostavljamo znanje programiranja in osnovno razumevanje verjetnosti.



3

Interaktivni test
opreme za delavnico

4

1 del

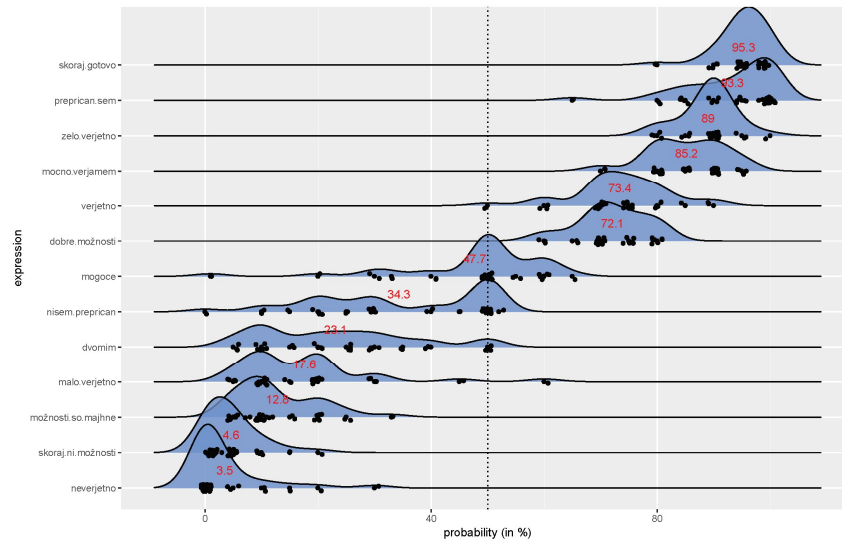
Negotovost in
probabilistično razmišljanje

5

Q: Ali bo naslednji teden
v Ljubljani deževalo?

6

Probabilistični izrazi v naravnem jeziku



7

Q: Kako toplo ($^{\circ}\text{C}$) bo jutri opoldne v Ljubljani?

8

Naravni jezik je nekonsistenten, nenatačen in premalo ekspresiven za resno kvantitativno delo!

- **Dobra novica** Primeren jezik so že razvili!
- **Slaba novica** Gre za teorijo verjetnosti – matematiki se ne moremo izogniti.
- **Dobra novica** Ni se nam potrebno naučiti niti vse dodiplomske verjetnosti¹ – potrebujemo le verjetnost kot jezik, računal pa bo računalnik.

¹ Kar pa ne pomeni, da nam ne bo koristilo! Verjetnost je osnova kvantitativne analize podatkov.

9

Gramatika verjetnosti

Verjetnost P (pogosto Pr) je funkcija, ki dogodkom prireja numerične vrednosti in zadošča tem aksiomom:

$$A1 \quad P(A) \geq 0.$$

$$A2 \quad P(\Omega) = 1.$$

$$A3 \quad P(A_1 \cup A_2 \cup A_3 \cup \dots) = \sum_{i=1}^{\infty} P(A_i),$$

za poljubno sekvenco disjunktnih dogodkov.

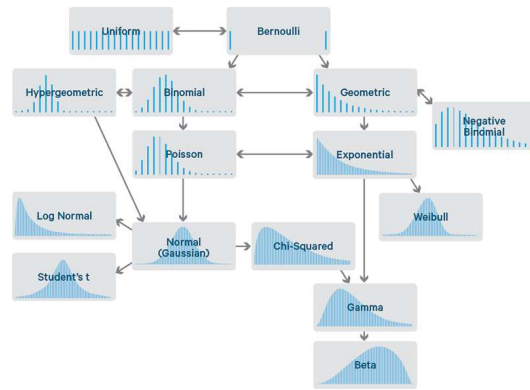
Definicija **pogojne verjetnosti**:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

10

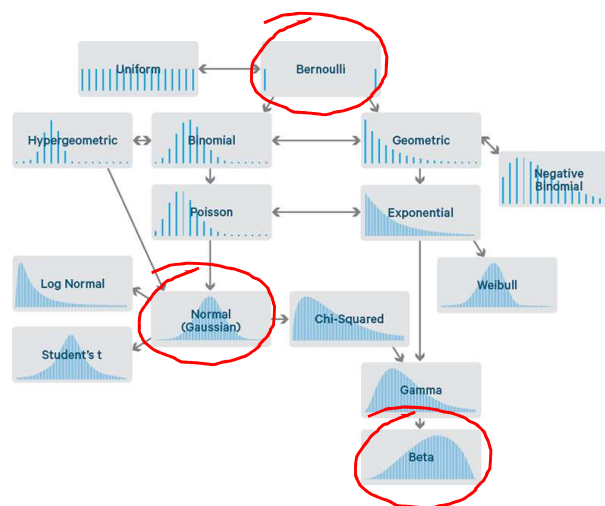
Porazdelitve

- Porazdelitve so elementarni izrazi probabilističnega razmišljanja in
- osnovni gradniki statističnih modelov.
- Porazdelitve so v skladu s pravili teorije verjetnosti, zato so **konsistentne** in **natančne** probabilistične izjave.
- Več, kot vemo o porazdelitvah, bolj bogato se lahko izražamo.



11

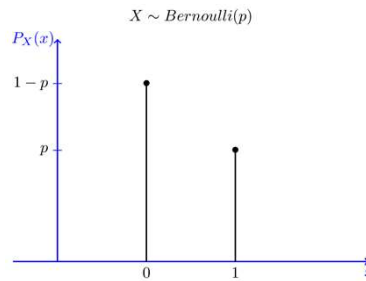
Beseda na dan...



12

Bernoullijeva porazdelitev

distribution	pmf	mean	variance
Bernoulli(p)	$p^x(1-p)^{1-x}; x = 0, 1; p \in (0, 1)$	p	$p(1-p)$

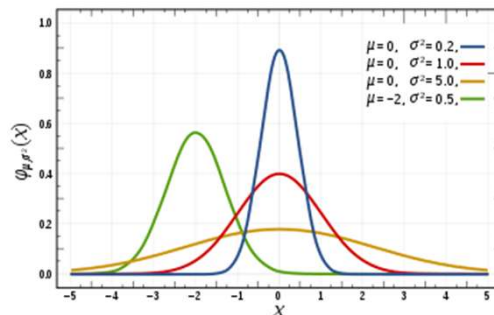


Q: Ali bo naslednji teden v Ljubljani deževalo?

13

Normalna (Gaussova) porazdelitev

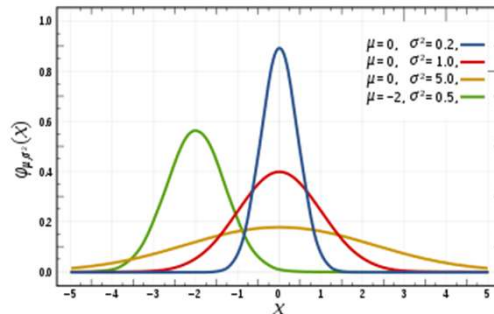
distribution	pdf	mean	variance
Normal(μ, σ^2)	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; \sigma > 0$	μ	σ^2



14

Normalna (Gaussova) porazdelitev

distribution	pdf	mean	variance
Normal(μ, σ^2)	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; \sigma > 0$	μ	σ^2

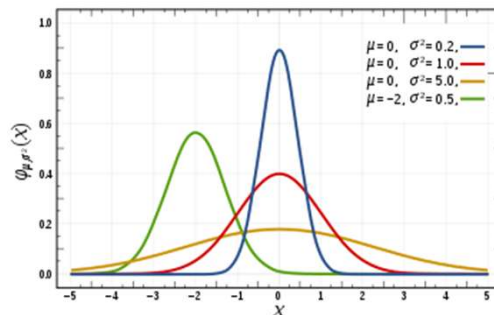


Q: Kako toplo (°C) bo jutri opoldne v Ljubljani?

15

Normalna (Gaussova) porazdelitev

distribution	pdf	mean	variance
Normal(μ, σ^2)	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; \sigma > 0$	μ	σ^2

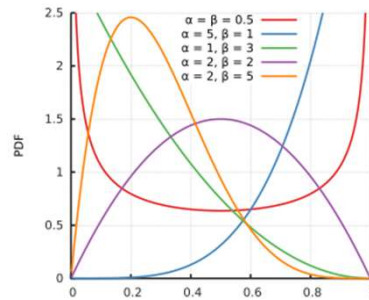


Q: Kako toplo (°C) je bilo na današnji dan pred 50 leti?

16

Porazdelitev Beta

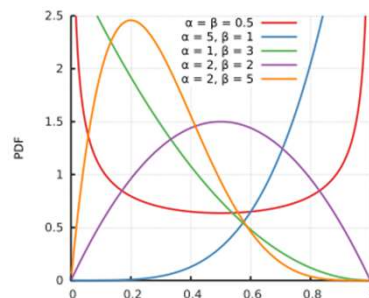
distribution	pdf	mean	variance
$\text{Beta}(\alpha, \beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}; x \in (0, 1), \alpha, \beta > 0$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$



17

Porazdelitev Beta

distribution	pdf	mean	variance
$\text{Beta}(\alpha, \beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}; x \in (0, 1), \alpha, \beta > 0$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$



Q: Kolikšna je verjetnost, da naslednji teden v LJ dežuje?

18

Preizkus probabilističnega razmišljanja

To so izidi 10 metov (morda nepoštenega) konvanca:

c c g c c g c c c g (?)

Q1: Je enajsti met **c**ifra ali **g**rb?

Q2: Kolikšna je verjetnost p , da na tem kovancu pade grb?

Q3: Je kovanec pošten? Poštenost je npr., da je p med 48% and 52%.

19

V razmislek...

Verjetnost je koherenten in natančen jezik za izražanje negotovosti:

- Če ne sledimo zakonom verjetnosti, nas nihče ne bo razumel!
- Sicer pa so probabilistične izjave lahko subjektivne ali navidez popolnoma nesmiselne.
- Precej naravno nam je, da imamo verjetnostno mnenje o stvareh, ki niso naključne. Naključje je samo eden izmed virov negotovosti (in ne preveč pogost).

Uporaba verjetnosti za izražanje negotovosti je bistvo bayesovskega pogleda na statistično sklepanje!

20

2 del

Statistično modeliranje

21

Model =
Hipoteza, kako so
nastali naši podatki.

22

**Ni modeliranja
brez modela.**

23

Ne, resno, ni.

24

Q: Zapišite 1 metodo iz statistike ali strojnega učenja, ki se uporablja za napovedovanje, razpoznavanje vzorcev, gručenje, testiranje hipotez, ipd.

25

Zaporedje enic in ničel (= podatki):

100100101011011001111111111101

Statistični model (= poskus statistične interpretacije):

Zaporedje je nastalo s 30 neodvisnimi meti kovanca z neznano verjetnostjo enice θ .

Predhodno mnenje o parametrih modela:

Nimam pojma, koliko je θ , zato ne bom izrazil preference do nobene vrednosti θ .

Statistično sklepanje (= učenje):

Pri vseh the predpostavkah in upoštevajoč zakone verjetnosti, kakšno mora biti moje mnenje o θ , ko vidim podatke?

26

Zaporedje enic in ničel (= podatki):

10010010101101100111111111101

 y_1, \dots, y_n $y_i \in \{0, 1\}$ **Statistični model (= poskus statistične interpretacije):**

Zaporedje je nastalo s 30 neodvisnimi meti kovanca z neznano verjetnostjo enice θ .

$$y_1, y_2, \dots, y_n | \theta \sim_{\text{iid}} \text{Bernoulli}(\theta)$$

Predhodno mnenje o parametrih modela:

Nimam pojma, koliko je θ , zato ne bom izrazil preference do nobene vrednosti θ .

$$\theta \sim \text{Beta}(1, 1)$$

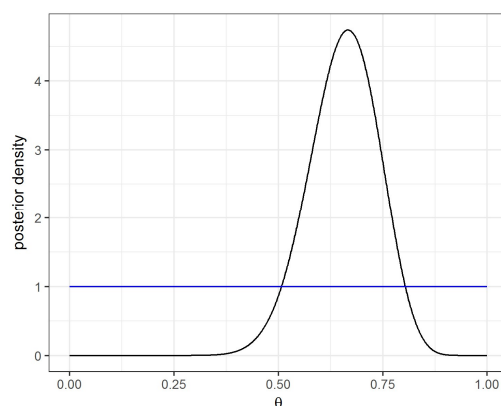
Statistično sklepanje (= učenje):

Pri vseh the predpostavkah in upoštevajoč zakone verjetnosti, kakšno mora biti moje mnenje o θ , ko vidim podatke?

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}$$

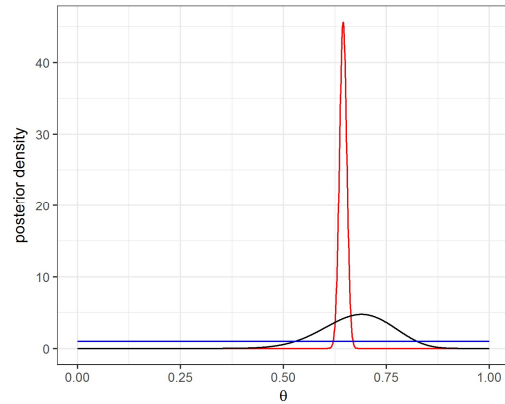
$$\theta|y_1, \dots, y_n \sim \text{Beta}(\sum y_i + 1, n - \sum y_i + 1)$$

27



Naše mnenje o θ **prej** in **potem**, ko smo videli zaporedje, ki vsebuje 20 enic in 10 ničel.

28



Naše mnenje o θ **potem**, ko smo videli zaporedje, ki vsebuje 20 enic in 10 ničel.

29

3 del

Probabilistično programiranje

30

Probabilistični programski jezik (PPL) je programski jezik, ki je zasnovan za opisovanje probabilističnih modelov in računsko sklepanje iz teh modelov.

Vir: Wikipedia

31

Probabilistični programski jezik nam omogoča, da se osredotočimo na modeliranje in preskočimo matematične in računske probleme pri sklepanju.

32

Dva primera imperativnega programiranja

```

1 # Bubble Sort
2 sort <- function(x) {
3   n <- length(x)
4
5   for (k in n:2) {
6     i <- 1
7     while (i < k) {
8       if (x[i] > x[i+1]) {
9         temp <- x[i+1]
10        x[i+1] <- x[i]
11        x[i] <- temp
12      }
13      i <- i + 1
14    }
15  }
16  x
17 }

```

```

1 # Generate 30 Bernoulli variables
2 bernoulli <- function(p) {
3   x <- c()
4
5   for (i in 1:30) {
6     if (runif(1) > p) {
7       x <- c(x, 0)
8     } else {
9       x <- c(x, 1)
10    }
11  }
12  x
13 }
14

```

33

Imperativno programiranje

in

Statistično modeliranje

```

1 # Generate 30 Bernoulli variables
2 bernoulli <- function(p) {
3   x <- c()
4
5   for (i in 1:30) {
6     if (runif(1) > p) {
7       x <- c(x, 0)
8     } else {
9       x <- c(x, 1)
10    }
11  }
12  x
13 }
14

```

#Sklepanje o relativni frekvenci tega zaporedja
10010010101101100111111111101

$$y_1, \dots, y_n \quad y_i \in \{0, 1\}$$

$$y_1, y_2, \dots, y_n | \theta \sim_{\text{iid}} \text{Bernoulli}(\theta)$$

$$\theta \sim \text{Beta}(1, 1)$$

- Podane imamo **vhodne podatke** in **parametre**,
- sprogramiramo algoritem, ki generira zahtevane izhodne podatke.
- Podane imamo **vhodne** in **izhodne podatke**,
- opišemo generator, ki naj bi generiral podatke,
- sklepamo o najbolj verjetnih vrednostih parametrov.

34

Odmor

koda za drugi del:

<https://github.com/bstatcomp/Stan-Intro-Workshop>

35

4 del

Programski jezik

Stan

36

Kaj je Stan?

- Orodje za učinkovito Bayesovo statistično modeliranje.
- Najlažje ga uporabljamo preko vmesnikov (na primer RStan, PyStan, ...).
- Stan je "compiled" jezik, to pomeni, da se statistični model preslika v c++ kodo, ki se nato pred uporabo prevede (zato je potrebno pred uporabo modela malo počakati).



37

Obvezni bloki vsakega Stan programa

- **data** – blok v katerem s pomočjo spremenljivk deklariramo vhodne podatke. Vrednosti vhodnih podatkov pripravi uporabnik/razvijalec, običajno v programskem jeziku, ki ga uporabljamo kot vmesnik.
- **parameters** – blok v katerem deklariramo parametre, ki jih želimo oceniti (kateri parametri našega statističnega modela nas zanimajo). Stan preko vmesnika (na primer RStan) vrne vrednosti parametrov nazaj v izhodiščni programski jezik.
- **model** – opis statističnega modela.

```
/*
Primer komentarja, ki obsega
več vrstic.
*/

data {
  // tukaj definiramo vhodne podatke
}

parameters {
  // parametri modela, ki jih želimo oceniti
}

model {
  // sem spada statistično modeliranje
}
```

38

Osnovni tipi spremenljivk

- **int** – celo število `int a;`
- **real** – realno število `real b;`
- **seznam** (array) – seznam celih ali realnih števil `int a[10]; real b[n];`
- **matrika** (matrix) – 2D seznam [vrstice, stolpci] `int A[10, 10];`
- **vector** – vektor realnih števil (optimiziran seznam) `vector[n] v;`
- **simplex** – vektor pozitivnih realnih števil, ki se seštevajo v 1 `simplex[n] s;`
- (skoraj) vsem spremenljivkam lahko določimo zgornjo in spodnjo mejo `real<lower=0> sigma;`
`real<lower=0, upper=1> success_rate;`

39

Porazdelitve

- **Bernoulli** `y ~ bernoulli(theta);`
 y je vektor "uspehov" (1) in "neuspehov" (0)
 θ (theta) predstavlja verjetnost uspeha
- **beta** `y ~ beta(alpha, beta);`
 y je vektor realnih števil med 0 in 1
 α, β parametra porazdelive
- **normal** `y ~ normal(mu, sigma);`
 y je vektor realnih števil
 μ, σ sta upanje oziroma varianca
- porazdelitve uporabimo tudi za vnanašnje predznaka o določenih parametrih modela `theta ~ beta(1,1);`

40

5 del

Praktični primeri

41

Izjemno kratek uvod v R

42

Problem #1

V datoteki **basketball_shots.csv** so podatki o metih na koš. Vsak košarkaš je 60 krat vrgel na koš običajne velikosti, nato pa še 60 krat na koš z manjšim obsegom.

Zanima nas:

1) Primerjava med uspešnostjo košarkaša #1 ter košarkaša #2:

- *kdo je boljši,*
- *kako prepričani smo v to ugotovitev,*
- *kolikšna je razlika v uspešnosti?*

2) Primerjava med metanjem na običajni ter manjši obroč za košarkaša #1.

Namigi:

Za modeliranje uspešnosti uporabi Bernoullijevo porazdelitev.

Predznanje v model lahko vnesemo s pomočjo Beta porazdelitve.

43

Naš prvi model – meti na koš

44

Naš prvi model – meti na koš

- **data**

n – število metov
y – rezultat meta (0 – nespešen met, 1 – uspešen)

```
data {
  int<lower=1> n;
  int y[n];
}
```

- **parameters**

theta – parameter Bernoulli distribucije, ki ocenjuje uspešnost

```
parameters {
  real<lower=0, upper=1> theta;
}
```

- **model**

predznanje (prior)
opis modela

```
model {
  // prior
  theta ~ beta(1,1);

  for (i in 1:n) {
    y[i] ~ bernoulli(theta);
  }
}
```

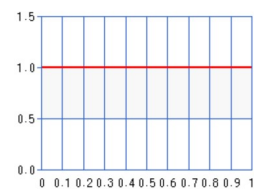
45

Prior za parameter theta

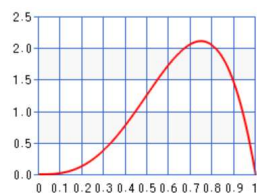
- **brez**

"flat" porazdelitev na intervalu $[-\infty, \infty]$

- **beta(1, 1)**



- **beta(4, 2)**



46

Problem #2

V datoteki **temperature.csv** so podatki o temperaturi, za Slovenijo imamo podatke o povprečni temperature za vsak mesec med 1901 in 2015.

Zanima nas ali je bila temperature julija (najbolj vroč mesec) med leti 1970-1985 nižja kot med leti 2000 in 2015?

- *Kako prepričani smo v to trditev?*
- *Za koliko je bila temperature nižja?*

Namig:

Za modeliranje uspešnosti uporabi normalno porazdelitev – $N(\mu, \sigma)$.

47

Problem #3

V datoteki **temperature.csv** so podatki o temperaturi. Za Slovenijo imamo podatke o povprečni temperaturi za vsak mesec med 1901 in 2015, za Finsko pa med 2000 in 2015.

Zanima nas:

1) Ali julijska temperatura skozi čas na Finskem narašča?

- *Kakšno je naše zaupanje v napovedi modela, zakaj?*
- *Kako lahko ta problem rešimo?*

2) Ali julijska temperatura skozi čas v Sloveniji narašča?

3) Kakšna bo temperatura leta 2019, kakšna 2070? Kakšna je verjetnost, da bo leta 2019 oziroma 2070 pričakovana temperatura višja od 25°C?

Namig:

Za modeliranje uspešnosti uporabi normalno linearno regresijo – normalni linearni model, kjer je μ porazdelitve linearno odvisno ($a + bx$) od leta (x).

48

Problem #4

Novo nastalo zagonsko podjetje nas je najelo, da jim pomagamo pri dveh pomembnih odločitvah:
(ob predpostavki, da želijo maksimizirati svoj dobiček)

- 1) Kam vlagati sredstva (razvoj, marketing ali administracija) ?
- 2) Kje naj imajo svoje prostore (na voljo imajo Florido, Kalifornijo in New York)?

S pomočjo analiziranja podatkov iz datoteke **50_startups.csv** jim pomagaj pri odločitvi. V datoteki se nahajajo podatki o 50 zagonskih podjetjih (kako so vlagali denar, kje imajo pisarno ter koliko so zaslužili).

Namig:

Nadgradi linearno normalno regresijo iz prejšnjega problema.

Odvisna spremenljivka (profit) je odvisna od več atributov (research, marketing, ...) – vhod X je torej matrika.

Za vsak atribut želimo svojo beta (b) vrednost – parameter b je torej vektor.

Iz kategorične spremenljivke (state) naredimo več binarnih spremenljivk.

49

Problem #5

Parlamentarne stranke se odločajo ali bi razpisale predčasne volitve ali ne, na podlagi zadnje javno mnenjske raziskave (**elections.csv**) jim pomagaj pri tej odločitvi. Naročnika zanima:

- 1) Ocena uspešnosti za vse stranke.
- 2) Kakšna je verjetnost, da bo stranka po volitvah imela več sedežev v državnem zboru?
- 3) Kakšna je verjetnost, da ima po volitvah trenutna koalicija (LMSŠ, SAB, SD, DeSUS in SMC) večino?
- 4) Kakšna je verjetnost, da ima po volitvah trenutna opozicija (SDS, SLS, NSi in SNS) večino?

Namigi:

Uporabite Bernoulli-beta model (problem #1) – tukaj ima vsaka stranka svoj parameter θ .

Namesto vektorja za θ ete lahko uporabite simplex – ta poskrbi, da se verjetnosti seštejejo v 1.

Večina v slovenskem državnem zboru pomeni več kot 44 sedežev (algoritem za izračun sedežev je v R datoteki).

Večina algoritmov (izračun sedežev, primerjava s trenutnim stanjem) je že pripravljenih.

50