Gremo čez vse korake 1. sklopa na Poissonovem modelu, večino kode lahko uporabimo, poudarimo le pomembnejše razlike

2. sklop: Poissonov model

Nina Ruzic Gorenjec

1 Primer

Zgodovinski podatki o stevilu rojstev cetverckov na leto v Prusiji za obdobje 69 let (Ladislaus von Bortkiewicz), za katere je znano, da se dobro prilegajo Poissonovi porazdelitvi.

```
(podatki <- data.frame(stevilo.cetverckov = 0:6,
                       stevilo.let = c(14,24,17,9,2,2,1)))
```

```
##   stevilo.cetverckov stevilo.let
## 1                  0          14
## 2                  1          24
## 3                  2          17
## 4                  3           9
## 5                  4           2
## 6                  5           2
## 7                  6           1
```

frekvenčna tabela

Zanima nas povprecno stevilo rojstev cetverckov na leto.

2 Verjetnostni model za nas primer

Vzorec $X_1, X_2, \ldots, X_n$, kjer je:

- $n = 69$ stevilo let,
- X_i predstavlja stevilo rojstev cetverckov v i-tem letu,
- $X_i \mid \theta \sim \text{Poiss}(\theta)$,
- $P(X_i = k \mid \theta) = \frac{1}{k!}\theta^k e^{-\theta}$ za $k \in \{0, 1, 2, \ldots\}$,
- $\text{E}(X_i) = \theta$ – **parameter, ki nas zanima**,
- $\text{var}(X_i) = \theta$.

originalni podatki

Kaj so v nasem primeru X_i oz. njihova realizacija?

```
(x <- rep(podatki$stevilo.cetverckov, podatki$stevilo.let))
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [39] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 4 4 5 5 6
```

Vse in se vec o Poissonovi porazdelitvi je napisal doc. dr. Gaj Vidmar:

- clanek: http://ims.mf.uni-lj.si/archive/17(2)/31.pdf
- pripadajoce izracune lahko najdete na: http://ims.mf.uni-lj.si/archive/17(2)

3 Ocenjevanje v frekventisticni statistiki

Kako bi ocenili nas parameter s frekventisticno statistiko? Katere metode bi lahko uporabili?

Cenilka po metodi najvecjega verjetja in po metodi momentov je povprecje vzorca:

```
mean(x)
```

```
## [1] 1.57971
```

4 Ocenjevanje v Bayesovi statistiki

Bayesova formula:

$$\pi(\theta \mid x) \propto L(\theta \mid x)\, \pi(\theta).$$

4.1 Verjetje

Narisite verjetje tako, da bo ploscina pod narisano krivuljo enaka ena. Predrugacite lahko spodnjo kodo, ki smo jo uporabili za binomski model.

$$L(\theta \mid x) = \prod_{i=1}^{n} \frac{1}{x_i!} \theta^{x_i} e^{-\theta} \propto \theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}$$

majhno $\Longrightarrow$ levi izraz bistveno manjši od desnega $\downarrow$ glejte (*) na naslednji strani

Odvisnost od vzorca le preko $\sum_{i=1}^{n} x_i$ (in n), ki sta na nasem vzorcu (ohranimo oznake kakor pri binomskem modelu):

```
(n <- length(x))
```

```
## [1] 69
```

```
(k <- sum(x))
```

```
## [1] 109
```

V R-u:

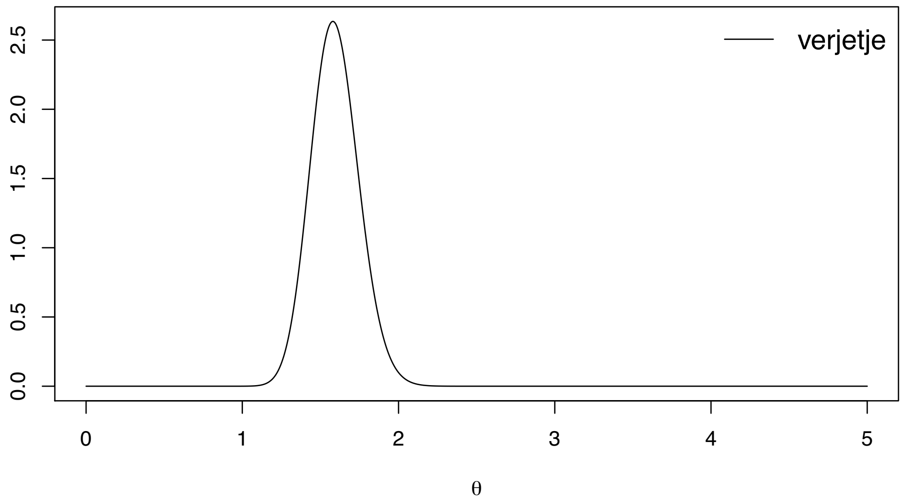
```
verjetje <- function(theta, k, n){
  theta^k * exp(-n*theta)
}
# POZOR: Verjetje bi lahko izracunali tudi kot prod(dpois(x, theta)).
# V tem primeru bi lahko imeli tezave z zelo majhnimi vrednostmi verjetja.
# Lahko bi dobili ploscino pod krivuljo v R enako 0 in s tem bi bila
# normalizacijska konstanta enaka Inf, tj. ustrezne slike ne bi mogli narisati.
# Resitev je zgornja funkcija ali pa mnozenje verjetja
# s katerokoli dovolj veliko konstanto.
#
# OB IMPLEMENTACIJI FORMUL/ALGORITMOV MORAMO BITI TOREJ PAZLJIVI:
# Vcasih je potrebno algoritem preoblikovati v ekvivalentno razlicico, da
# implementacija sploh deluje ali pa jo s tem pohitrimo.

#Z mnozenjem s konst dosezemo, da je integral verjetja glede na theta enak 1.
konst <- function(k, n){
  theta <- seq(0.001, 5, 0.001)
  1 / (0.001 * sum(verjetje(theta, k, n)))
}
```

(*)

Ob 2. DN se spomnite na to !

Narisemo za nas vzorec:

```
theta <- seq(0, 5, 0.001)
konst.verjetje <- konst(k, n) * verjetje(theta, k, n)
plot(theta, konst.verjetje, type = "l",
     xlab = expression(theta), ylab ="")
legend("topright", legend = c("verjetje"), col = c("black"),
       lty = 1, bty = "n", cex = 1.3)
```


verjetje
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3
4
5
θ

4.2 Apriorna porazdelitev

Za apriorno porazdelitev si izberemo gama porazdelitev, ki je v primeru poissonove porazdelitve podatkov *conjugate prior* (pomeni, da apriorna in aposteriorna porazdelitev pripadata enaki druzini porazdelitev), zato se lahko uporablja tudi izraz **gama-poissonov model**.

Za apriorno porazdelitev imamo torej gostoto gama porazdelitve pri parametrih $\alpha, \beta > 0$:

$$\pi(\theta) = \pi(\theta \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta},$$

kjer je funkcija gama $\Gamma(a) = (a-1)!$ za pozitivna cela stevila a. Spomnimo se:

- $\mathrm{E}(\mathrm{Gama}(\alpha, \beta)) = \frac{\alpha}{\beta}$,
- $\mathrm{var}(\mathrm{Gama}(\alpha, \beta)) = \frac{\alpha}{\beta^2}$.

Denimo, da se po nasih izkusnjah rodijo eni do dvoji cetvorcki letno, zato se odlocimo, da bo povprecje apriorne porazdelitve enako 1.5.

Tako bi si npr. lahko osmislili določitev neke apriorne porazdelitve

Pri tem mislimo, da se lahko zmotimo za priblizno 1. V primeru normalne porazdelitve je 95% vrednosti oddaljenih od povprecja za priblizno 2 standardna odklona. Standardni odklon nase porazdelitve zato nastavimo na 0.5.

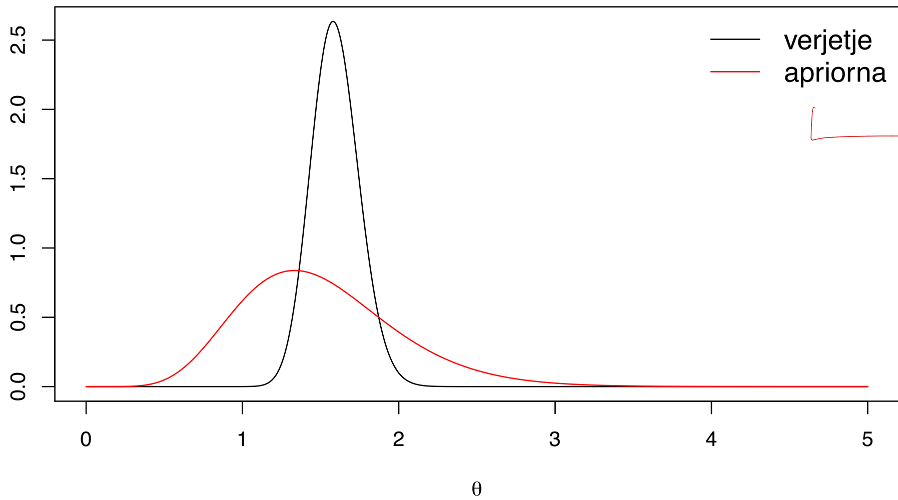
Izracunajte α in β nase apriorne porazdelitve.

→ več o dilemi „informativna VS neinformativna apriorna" pri zadnji vaji

Na graf z verjetjem dodajte gostoto apriorne porazdelitve. Predrugacite lahko spodnjo kodo, ki smo jo uporabili za binomski model.

Narisemo v R-u:

```
alpha <- 1.5*6
beta <- 6

theta <- seq(0, 5, 0.001)
apriorna <- dgamma(theta, shape = alpha, rate = beta)

konst.verjetje <- konst(k, n) * verjetje(theta, k, n)

y.max <- max(c(konst.verjetje, apriorna))
plot(theta, konst.verjetje, ylim = c(0, y.max), type = "l",
     xlab = expression(theta), ylab = "")
lines(theta, apriorna, col = "red")
legend("topright", legend = c("verjetje", "apriorna"), col = c("black", "red"),
       lty = 1, bty = "n", cex = 1.3)
```



v tem primeru je to informativna porazdelitev

4.3 Aposteriorna porazdelitev

Ker smo uporabili *conjugate prior*, bo aposteriorna porazdelitev tudi iz druzine gama porazdelitev. Koliksna sta njena parametra?

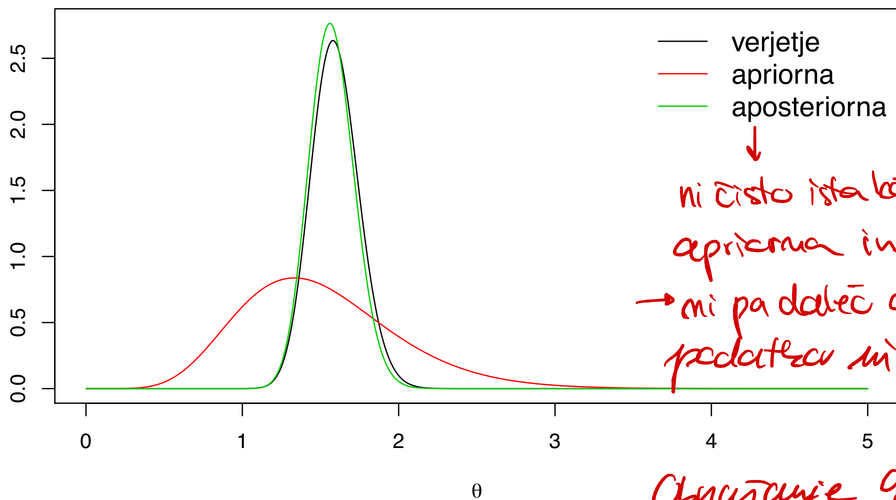
Na graf z verjetjem in gostoto apriorne porazdelitve dodajte se gostoto aposterirone porazdeitve. Predrugacite lahko spodnjo kodo, ki smo jo uporabili za binomski model.

Njena parametra sta enaka:

- $\alpha_{\text{apost}} = k + \alpha$,
- $\beta_{\text{apost}} = n + \beta$.

Narisemo v R-u:

```
alpha.apost <- k + alpha
beta.apost <- n + beta

theta <- seq(0.001, 5, 0.001)
aposteriorna <- dgamma(theta, alpha.apost, beta.apost)

konst.verjetje <- konst(k, n) * verjetje(theta, k, n)
apriorna <- dgamma(theta, alpha, beta)

y.max <- max(c(konst.verjetje, apriorna, aposteriorna))
plot(theta, konst.verjetje, ylim=c(0, y.max), type = "l",
     xlab = expression(theta), ylab = "")
lines(theta, apriorna, col = "red")
lines(theta, aposteriorna, col = "green3")
legend("topright", legend = c("verjetje", "apriorna","aposteriorna"),
       col = c("black","red","green3"), lty = 1, bty = "n", cex = 1.3)
```



↓ ni čisto ista kot verjetje (tj. podatki), ker je apriorna informativna

→ ni pa daleč od verjetja, ker se sredina podatkov in apriorne dobro ujemata

Obnašanje aposteriorne porazdelitve glede na različne izbore apriorne si boste ogledali v 1. DN.

4.4 Ocena parametra θ

Ocenite parameter θ.

Ena moznost je pricakovana vrednost aposteriorne porazdelitve:

$$\hat{\theta} = \frac{\alpha_{\text{apost}}}{\beta_{\text{apost}}} = \frac{k + \alpha}{n + \beta}.$$

Podobno kot pri binomskem modelu lahko zapisemo

$$\hat{\theta} = \frac{\beta}{\beta + n} \cdot \mu + \frac{n}{\beta + n} \cdot \frac{k}{n},$$

kjer je $\mu = \mathrm{E}(\mathrm{Gama}(\alpha, \beta)) = \frac{\alpha}{\beta}$.

→ podobno kot pri binomskem modelu

Ideja: $\hat{\theta}$ je utezeno povprecje med $\mathbf{E}$(apriorna) in $\mathbf{E}(X)$, kjer preko β kontroliramo, kako mocno verjamemo apriorni pricakovani vrednosti.

```
alpha.apost / beta.apost
```

```
## [1] 1.573333
```

4.5 Interval zaupanja

Izracunajte 95% interval zaupanja za θ. Preizkusite obe metodi, preko kvantilov in *highest posterior density (HPD) region*. Predrugacite lahko spodnjo kodo, ki smo jo uporabili za binomski model.

Preko kvantilov porazdelitve:

```
(iz <- qgamma(c(0.025, 0.975), alpha.apost, beta.apost))
```

```
## [1] 1.302290 1.869619
```

Highest posterior density (HPD) region:

```
#install.packages("HDInterval")
library(HDInterval)

aposteriorna.sample <- rgamma(100000, alpha.apost, beta.apost)
(iz.hdi <- hdi(aposteriorna.sample, credMass = 0.95))
```

```
##    lower    upper
## 1.289976 1.854661
## attr(,"credMass")
## [1] 0.95
```

4.6 Testiranje hipotez

Kako verjetna je domneva, da se v povprecju na leto rodijo eni do dvoji cetvorcki?

Verjetnost te domneve je: $= P(\theta \in (1,2))$

```
1 - pgamma(1, alpha.apost, beta.apost) -
    pgamma(2, alpha.apost, beta.apost, lower.tail = FALSE)
```

```
## [1] 0.9969729
```

4.7 Napovedovanje

1 nova meritev
→ pri binomskem smo imeli 10 novih meritev
(10 vprašanj na izpitu)

Zanima nas, kaj lahko povemo o stevilu cetvorckov v prihajajocem letu ob upostevanju podatkov zadnjih 69 let, tj. zanima nas **aposteriorna napovedna porazdelitev**.

(Ce bi nas zanimalo stevilo cetvockov v prihajajocem letu brez upostevanja podatkov 69 let, potem bi nas zanimala **apriorna napovedna porazdelitev**.)

V Poissonovem modelu z apriorno gama porazdelitvijo lahko hitro izpeljemo (predavanja) apriorno/aposteriorno napovedno porazdelitev:

upoštevamo $\int f(\theta)\, d\theta = 1$
θ gostota

$$P(Y = K) = \frac{\Gamma(K + \tilde{\alpha})}{\Gamma(\tilde{\alpha})\, K!} \tilde{\beta}^{\tilde{\alpha}} / (\tilde{\beta} + 1)^{K+\tilde{\alpha}} \quad \text{za } K \in \{0, 1, 2, \ldots\}.$$

To je ravno negativna binomska porazdelitev s parametroma $r = \tilde{\alpha}$ in $p = 1/(1+\tilde{\beta})$, zasledimo pa lahko tudi poimenovanje **Gama-Poissonova porazdelitev**.

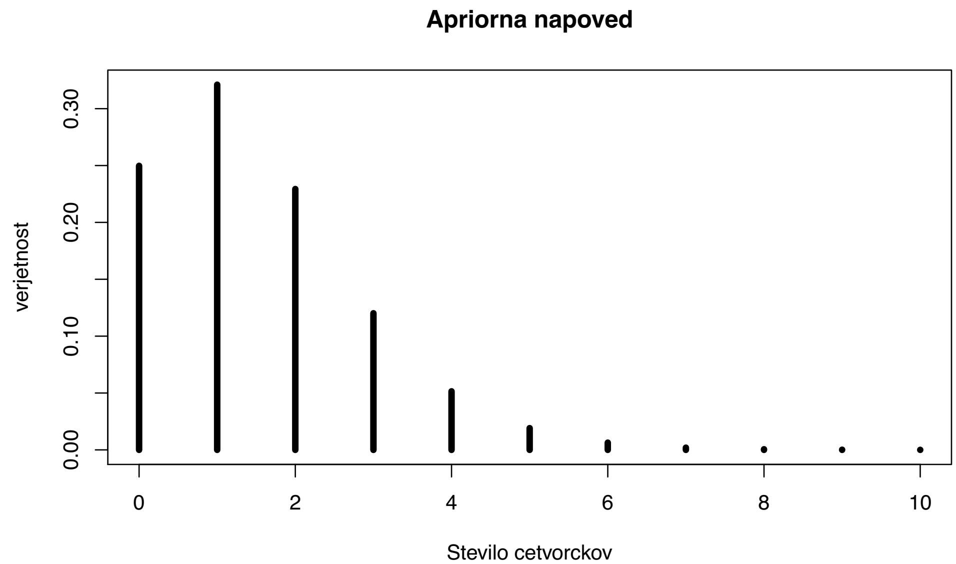
Za $\tilde{\alpha}, \tilde{\beta}$ vstavimo primerna parametra gama apriorne oz. aposteriorne porazdelitve.

Gama-Poissonova porazdelitev:

```r
dgammapoiss <- function(K, a, b){
  gamma(K+a)/(gamma(a)*factorial(K)) * b^a / (b+1)^(K+a)
}
```

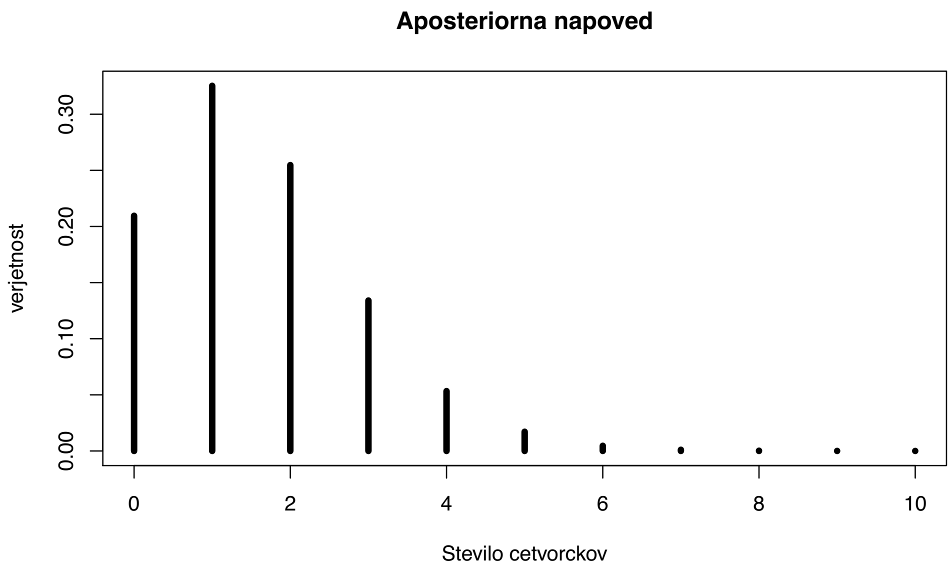
Narisemo apriorno napovedno porazdelitev.

```r
plot(0:10, dgammapoiss(0:10, a = alpha, b = beta), type = "h",
     xlab = "Stevilo cetvorckov", ylab = "verjetnost",
     main = "Apriorna napoved", lwd = 5)
```



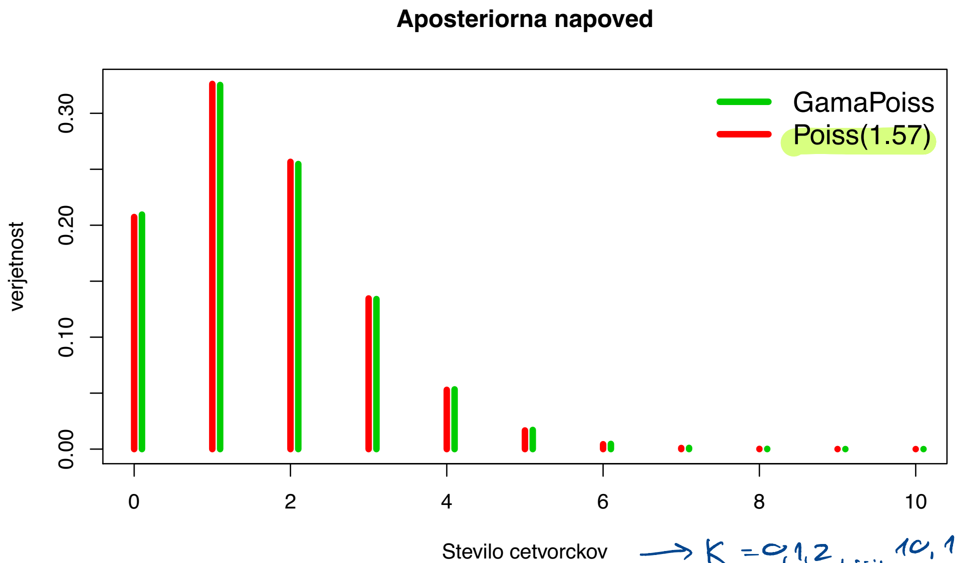
Narisemo aposteriorno napovedno porazdelitev.

```
plot(0:10, dgammapoiss(0:10, a = alpha.apost, b = beta.apost), type = "h",
     xlab = "Stevilo cetvorckov", ylab = "verjetnost",
     main = "Aposteriorna napoved", lwd = 5)
```



Poglejmo si se, kaksna je razlika med pravilno izracunano aposteriorno napovedno porazdelitvijo in tisto, ki jo dobimo, ce v Poissonovo porazdelitev vstavimo naso oceno parametra $\hat{\theta} = \alpha_{\text{apost}}/\beta_{\text{apost}} = 1.57$.

```
plot(0:10, dpois(0:10, alpha.apost / beta.apost), type = "h",
     xlab = "Stevilo cetvorckov", ylab = "verjetnost",
     main = "Aposteriorna napoved", col = "red", lwd = 5)
segments(x0 = seq(0.1,10.1,1), y0 = rep(0,11),
         x1 = seq(0.1,10.1,1), y1 = dgammapoiss(0:10, a = alpha.apost, b = beta.apost),
         lwd = 5, col = "green3")
legend("topright",  lty = 1, lwd = 5,
       c("GamaPoiss", paste("Poiss(",round(alpha.apost / beta.apost,2),")",sep="")),
       col = c("green3","red"), bty = "n", cex = 1.3)
```



$\rightarrow K = 0, 1, 2, \dots, \underbrace{10, 11, \dots \infty}$

verjetnosti za zavzemanje teh vrednosti so zelo majhne

Bin: $\theta \in (0,1)$
↳ omejeno območje

5 Jeffreyeva apriorna porazdelitev

$\theta > 0$, $\theta \in (0, \infty)$
neomejeno območje

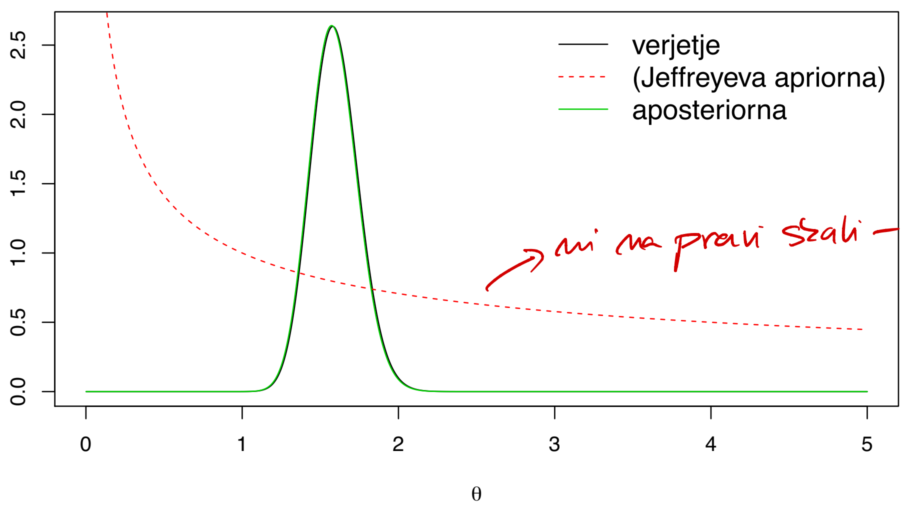
Pri Poissonovem modelu je Jeffreyeva apriorna porazdelitev $\pi(\theta) \propto \sqrt{1/\theta}$, kar si lahko interpretiramo kakor gostoto Gama($\alpha = 0.5$, $\beta = 0$). Ker je $\int_0^\infty \sqrt{1/\theta}\, d\theta = \infty$, je to **improper prior**.

$\pi(\theta)$ NI gostota, vendar mora biti takšna, da je $\pi(\theta|x)$ gostota

Pri taksni apriorni porazdelitvi bo aposteriorna porazdelitev Gama($\alpha_{\text{apost}} = k{+}0.5$, $\beta_{\text{apost}} = n$). **Nujno je, da je aposteriorna porazdelitev prava porazdelitev** (tj. integral gostote je enak 1), medtem ko to ne velja nujno za apriorno porazdelitev.

Pri nasih podatkih dobimo naslednje, kjer Jeffreyeve apriorne porazdelitve dejansko ne bi smeli narisati, saj njen integral ni enak 1.

```
alpha <- 0.5
beta <- 0

alpha.apost <- k + alpha
beta.apost <- n + beta

theta <- seq(0.001, 5, 0.001)
konst.verjetje <- konst(k, n) * verjetje(theta, k, n)
apriorna <- sqrt(1/theta)
aposteriorna <- dgamma(theta, alpha.apost, beta.apost)

plot(theta, konst.verjetje, type = "l",
     xlab = expression(theta), ylab = "")
lines(theta, apriorna, col = "red", lty = 2)
lines(theta, aposteriorna, col = "green3")
legend("topright", legend = c("verjetje", "(Jeffreyeva apriorna)","aposteriorna"),
       col = c("black","red","green3"), lty = c(1,2,1), bty = "n", cex = 1.3)
```



mi na pravi skali – nemogoče narisati, ker $\int = \infty$

na pravi skali

Spodnje dobimo, ce za parametra apriorne gama porazdelitve vzamemo $\alpha = 0.5$ in zelo majhen β, s cimer dobimo podobno kot Jeffrejeva apriorna.

približek za Jeffreyevo apriorno, zato da dobimo primernejšo sliko