

2. sklop: Poissonov model

Nina Ruzic Gorenjec

1 Primer

Zgodovinski podatki o številu rojstev četverčkov 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

Zanima nas povprečno število rojstev četverčkov na leto.

2 Verjetnostni model za nas primer

Vzorec X_1, X_2, \dots, X_n , kjer je:

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

Kaj so v našem 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  
## [39] 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:

- članek: [http://ims.mf.uni-lj.si/archive/17\(2\)/31.pdf](http://ims.mf.uni-lj.si/archive/17(2)/31.pdf)
- pripadajoče izračune lahko najdete na: [http://ims.mf.uni-lj.si/archive/17\(2\)](http://ims.mf.uni-lj.si/archive/17(2))

3 Ocenjevanje v frekventistični statistiki

Kako bi ocenili naš parameter s frekventistično statistiko? Katere metode bi lahko uporabili?

Cenilka po metodi največjega verjetja in po metodi momentov je povprečje 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 plosčina pod narisano krivuljo enaka ena. Predrugajte lahko spodnjo kodo, ki smo jo uporabili za binomski model.

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

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

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

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

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

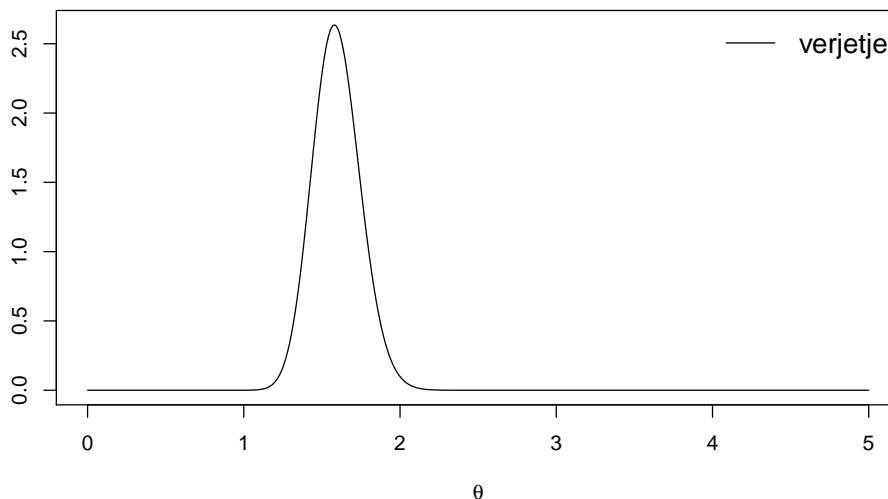
```
## [1] 109
```

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 bile vrednosti verjetja tako majhne, da bi bila  
# pri teh podatkih ploscina pod krivuljo v R enaka 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.  
  
#Z mnozenjem s konst dosežemo, 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)))  
}
```

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)
```



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 družini 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:

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

Denimo, da se po nasih izkusnjah rodi eni do dvoji četvorci letno, zato se odločimo, da bo povprečje apriorne porazdelitve enako 1.5.

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

Izračunajte α in β nase apriorne porazdelitve.

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

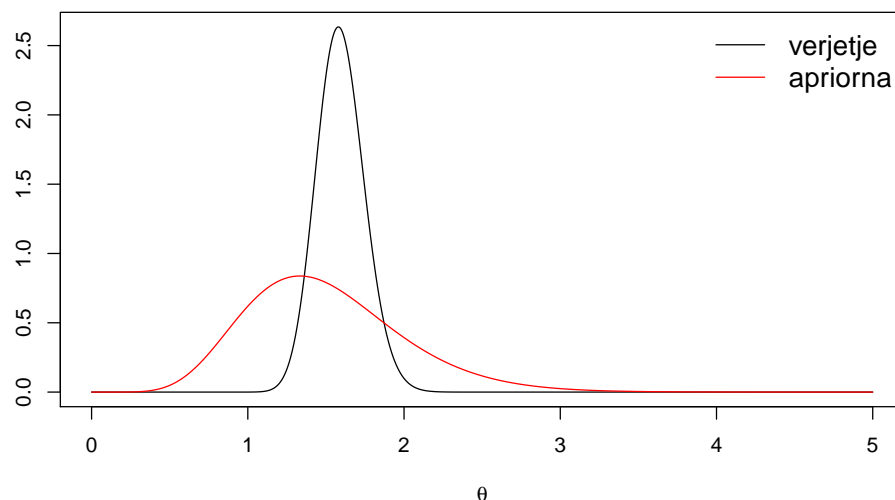
Narisemo v R-u:

```
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)
```



4.3 Aposteriorna porazdelitev

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

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

Njena parametra sta enaka:

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

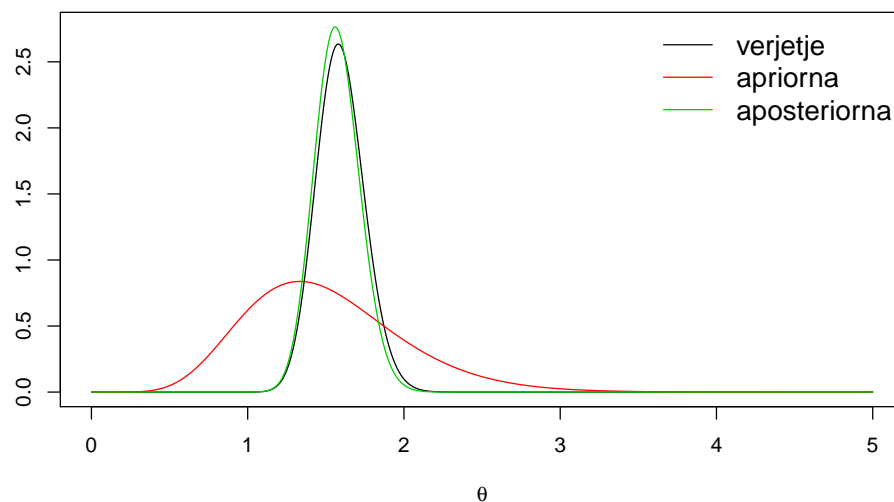
Narisemo v R-u:

```
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)
```



4.4 Ocena parametra θ

Ocenite parameter θ .

Ena možnost je pričakovana 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 $\mu = E(\text{Gama}(\alpha, \beta)) = \frac{\alpha}{\beta}$.

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

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

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

4.5 Interval zaupanja

Izračunajte 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  
## 1.296258 1.860958  
## 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:

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

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

4.7 Napovedovanje

Zanima nas, kaj lahko povemo o številu cetvorckov v prihajajočem letu ob upoštevanju podatkov zadnjih 69 let, tj. zanima nas **aposteriorna napovedna porazdelitev**.

(Ce bi nas zanimalo število cetvocekov v prihajajočem letu brez upoštevanja 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:

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

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**.

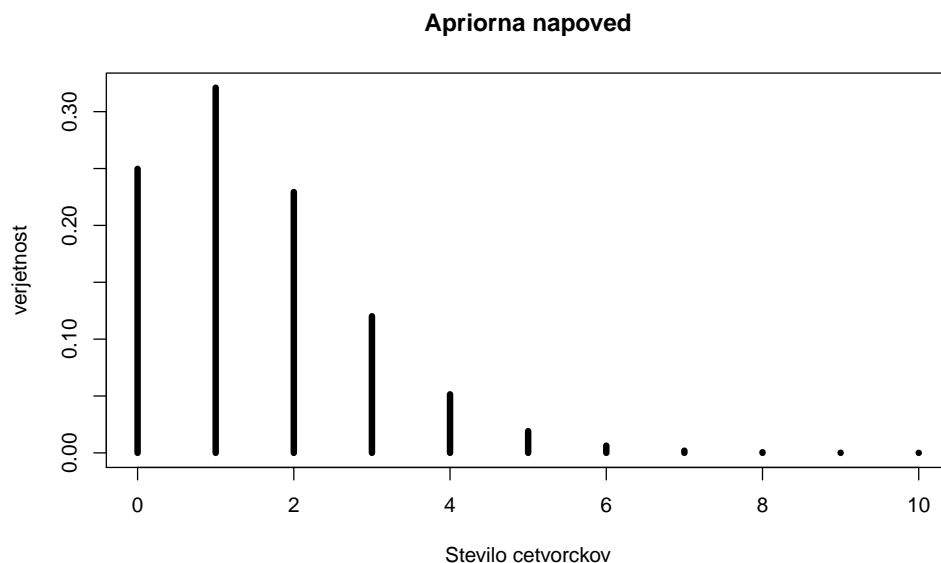
Za $\tilde{\alpha}, \tilde{\beta}$ vstavimo primerna parametra gama apriorne oz. aposteriorne porazdelitve.

Gama-Poissonova porazdelitev:

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

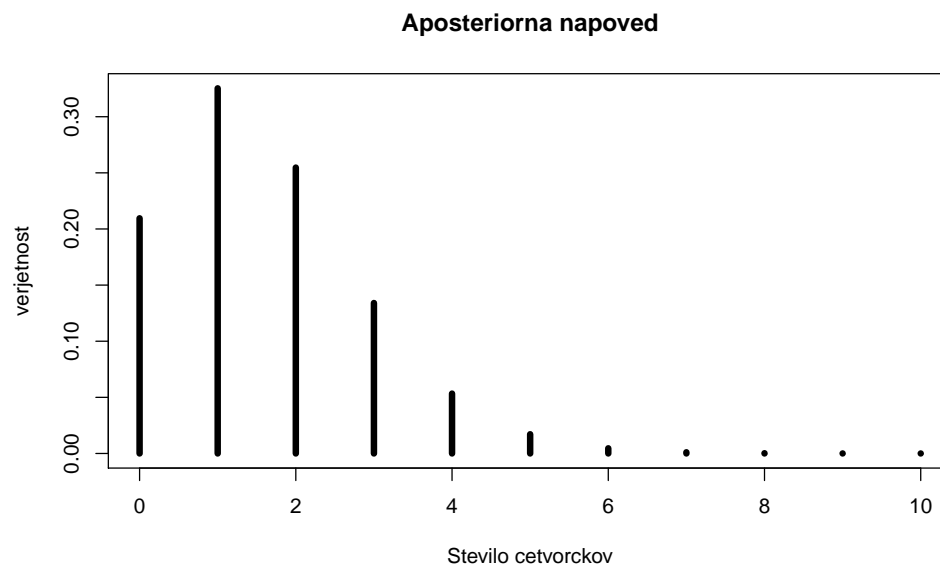
Narisemo apriorno napovedno porazdelitev.

```
plot(0:10, dgammapoiss(0:10, a = alpha, b = beta), type = "h",  
     xlab = "Število 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)
```



5 Jeffrejeva apriorna porazdelitev

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

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

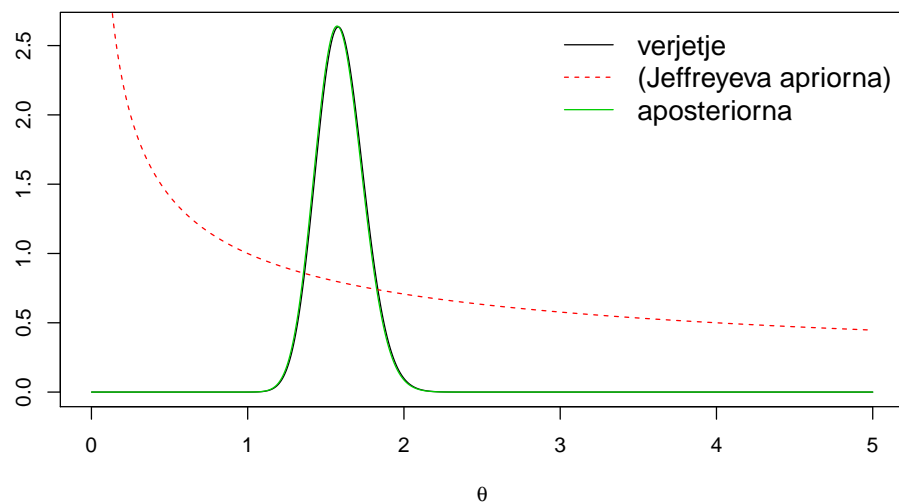
Pri naših podatkih dobimo naslednje, kjer Jeffrejeve apriorne porazdelitve dejansko ne bi smeli narisati, saj njen integral ni enak 1.

```
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", "(Jeffrejeva apriorna)", "aposteriorna"),
      col = c("black", "red", "green3"), lty = c(1, 2, 1), bty = "n", cex = 1.3)
```



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

