1. sklop: Normalni model z znano varianco

Nina Ruzic Gorenjec

1 Primer

Podan imamo naslednji vzorec visin (metri) studentov moskega spola:

```
x <- c(1.91, 1.94, 1.68, 1.75, 1.81, 1.83, 1.91, 1.95, 1.77, 1.98,
       1.81, 1.75, 1.89, 1.89, 1.83, 1.89, 1.99, 1.65, 1.82, 1.65,
       1.73, 1.73, 1.88, 1.81, 1.84, 1.83, 1.84, 1.72, 1.91, 1.63)
```

Zanima nas povprecna visina studentov, kjer privzamemo, da je standardni odklon $\sigma = 0.1$.

```
sigma <- 0.1
```

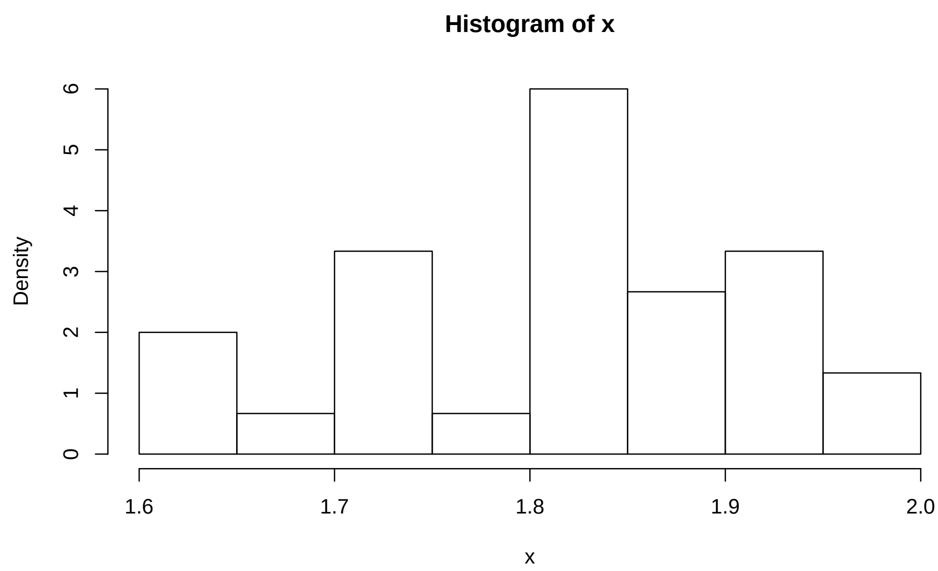
2 Verjetnostni model za nas primer

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

- $n = 30$ stevilo studentov,
- X_i predstavlja visino i-tega studenta,
- $X_i \mid \theta \sim N(\theta, \sigma^2 = 0.1^2)$,
- $f(x \mid \theta) = \frac{1}{\sqrt{2\pi}\,0.1} e^{-\frac{(x-\theta)^2}{2\cdot(0.1)^2}}$.

Ali je zgornji model smiseln za nase podatke?

```
hist(x, prob = TRUE)
```




```r
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.96666, p-value = 0.4523
```

```r
sd(x)
```

```
## [1] 0.09857374
```

3 Ocenjevanje v frekventisticni statistiki

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

```r
mean(x)
```

```
## [1] 1.820667
```

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.

$$L(\theta \mid x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,0.1} e^{-\frac{(x_i-\theta)^2}{2\cdot(0.1)^2}}$$

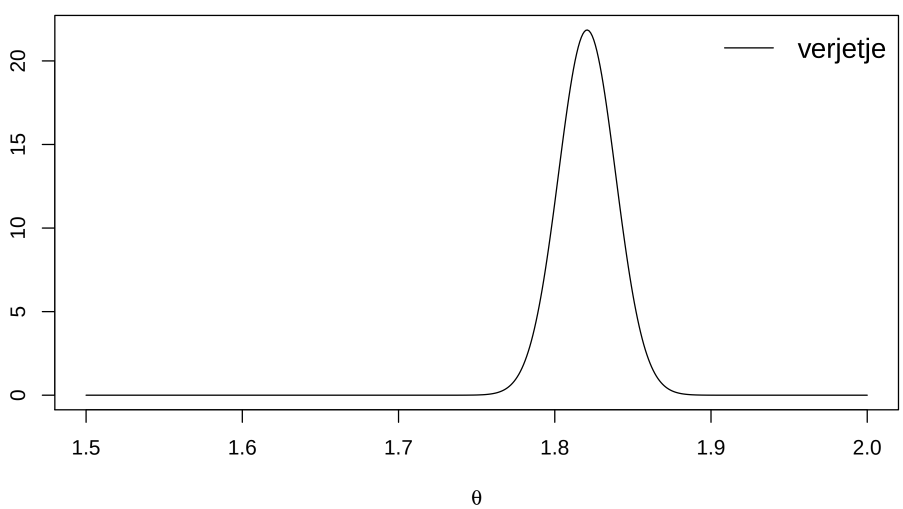
V R-u:

```
verjetje <- function(theta, x, sigma = 0.1){
  prod(dnorm(x, mean = theta, sd = sigma))
}

#Z mnozenjem s konst dosezemo, da je integral verjetja glede na theta enak 1.
konst <- function(x, from = 1.5, to = 2, by = 0.001, sigma = 0.1){
  theta <- seq(from = from, to = to, by = by)
  1 / (by * sum(sapply(theta, FUN = verjetje, x = x, sigma = sigma)))
}
```

Narisemo za nas vzorec:

```
theta <- seq(1.5, 2, 0.001)
konst.verjetje <- konst(x) * sapply(theta, FUN = verjetje, x = x, sigma = sigma)
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

V tem modelu je konjugirana porazdelitev normalna porazdelitev, njena parametra bomo oznacili z μ_0 in σ_0^2.

Jeffrejeva apriorna porazdelitev v tem modelu je $\pi(\theta) \propto \sqrt{1/\sigma^2} \propto 1$, kar si lahko interpretiramo kakor gostoto $N(\mu_0 = 0, \sigma_0^2 =$ "zelo velik"). Ker je $\int_{-\infty}^{\infty} 1\, d\theta = \infty$, je to *improper prior*.

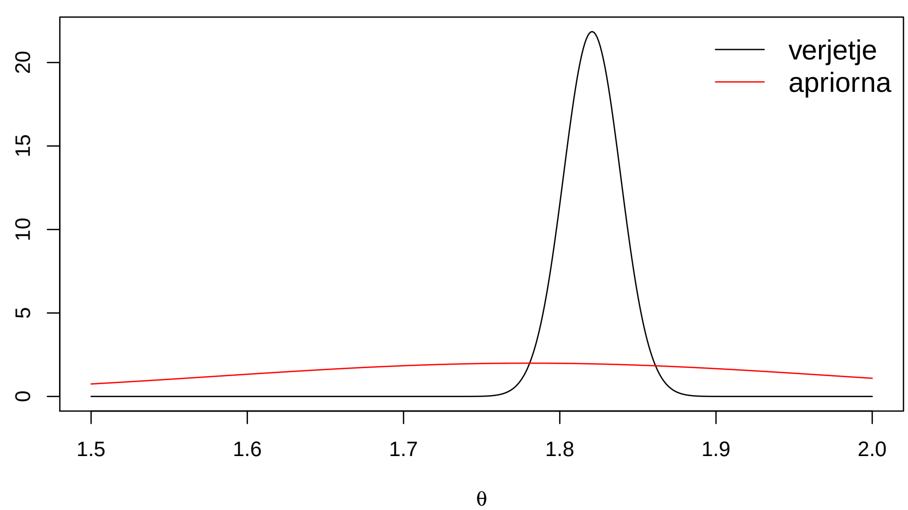
Na spletnih straneh SURS-a (Statisticni urad republike Slovenije) lahko najdemo podatek, da je povprecna visina moskih 178 cm (leto 2015), zaradi cesar se odlocimo za $\mu_0 = 1.78$. Odlocimo se za $\sigma_0^2 = 0.2^2$, tj. 95% referencni interval apriorne porazdelitve bo priblizno 178 cm $\pm$ 40 cm oz. [138 cm, 218 cm] (sibko informativna porazdelitev).

Narisemo v R-u:

```r
mu0 <- 1.78
sigma0 <- 0.2

theta <- seq(1.5, 2, 0.001)
konst.verjetje <- konst(x) * sapply(theta, FUN = verjetje, x = x, sigma = sigma)
apriorna <- dnorm(theta, mean = mu0, sd = sigma0)

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 konjugirano porazdelitev, bo tudi aposteriorna porazdelitev normalna.

Njena parametra, ki ju oznacimo z μ_n in σ_n^2, sta enaka:

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2},$$

$$\mu_n = \frac{1/\sigma_0^2}{1/\sigma_0^2 + n/\sigma^2}\,\mu_0 + \frac{n/\sigma^2}{1/\sigma_0^2 + n/\sigma^2}\,\bar{x},$$

kjer je $\sigma = 0.01$.

Aposteriorna pricakovana vrednost μ_n je torej utezeno povprecje apriorne pricakovane vrednosti μ_0 in vzorcnega povprecja $\bar{x}$, kjer preko *precision* apriorne porazdelitve $1/\sigma_0^2$ kontroliramo, kako mocno verjamemo apriorni pricakovani vrednosti.

V primeru Jeffrejeve apriorne porazdelitve dobimo $\mu_n = \bar{x}$ in $\sigma_n^2 = \sigma^2/n$. Ali je to skladno s frekventisticno statistiko? Zakaj?

Narisemo v R-u:

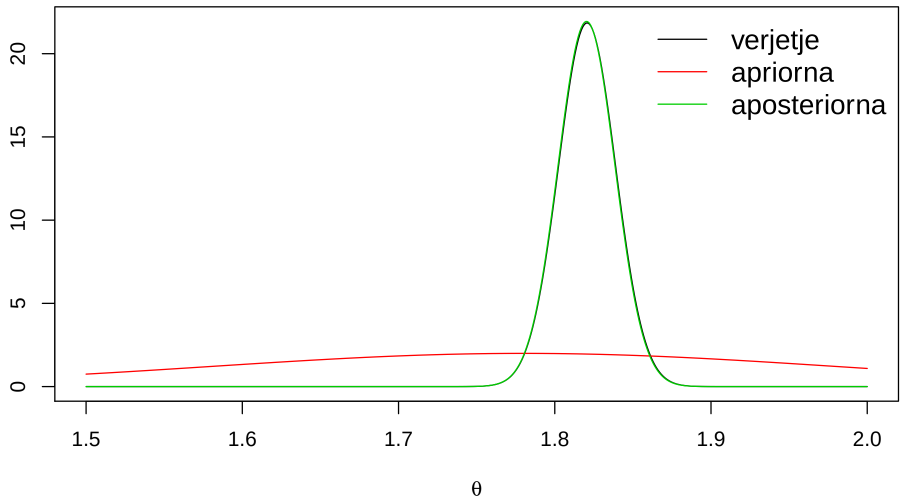
```
n <- length(x)
prec <- 1/sigma^2
prec0 <- 1/sigma0^2

prec.n <- prec0 + n*prec
sigma.n <- sqrt(1/prec.n)

mu.n <- prec0/prec.n * mu0 + n*prec/prec.n * mean(x)

theta <- seq(1.5, 2, 0.001)
konst.verjetje <- konst(x) * sapply(theta, FUN = verjetje, x = x, sigma = sigma)
apriorna <- dnorm(theta, mean = mu0, sd = sigma0)
aposteriorna <- dnorm(theta, mean = mu.n, sd = sigma.n)

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

Ocenimo parameter θ s pricakovano vrednostjo aposteriorne porazdelitve:

$$\hat{\theta} = \mu_n.$$

```
mu.n
```

```
## [1] 1.820331
```

4.5 Interval zaupanja

Izracunamo 95% interval zaupanja za θ.

Preko kvantilov porazdelitve:

```
(iz <- qnorm(c(0.025, 0.975), mean = mu.n, sd = sigma.n))
```

```
## [1] 1.784695 1.855966
```

Highest posterior density (HPD) region:

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

aposteriorna.sample <- rnorm(1000000, mean = mu.n, sd = sigma.n)
(iz.hdi <- hdi(aposteriorna.sample, credMass = 0.95))
```

```
##    lower    upper
## 1.784735 1.855984
## attr(,"credMass")
## [1] 0.95
```

Katera metoda se vam zdi pri tem modelu boljsa? Zakaj?

4.6 Napovedovanje

Zanima nas, kaj lahko povemo o visini novega studenta ob upostevanju podatkov 30 studentov, tj. zanima nas **aposteriorna napovedna porazdelitev**.

(Ce bi nas zanimala visina studenta brez upostevanja podatkov 30 studentov, potem bi nas zanimala **apriorna napovedna porazdelitev**.)

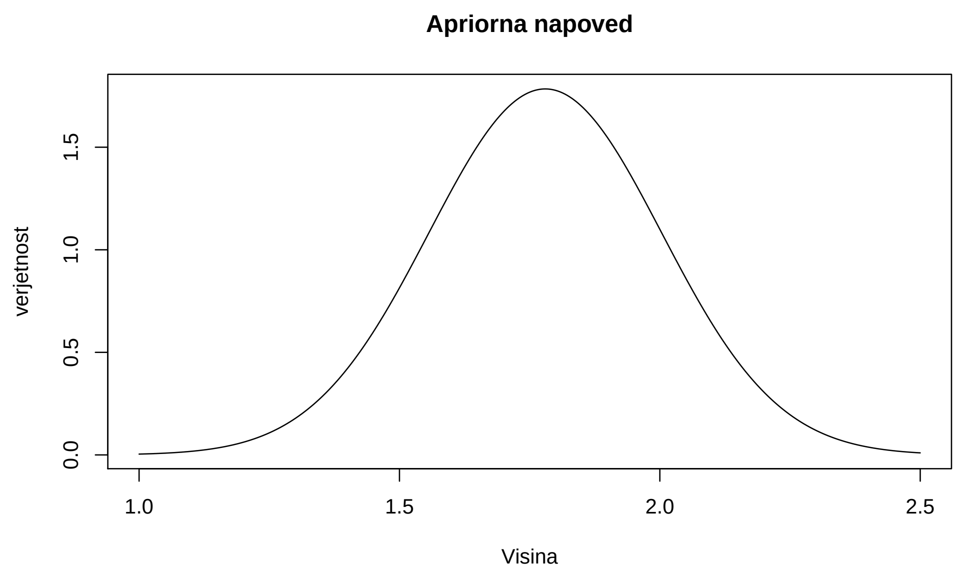
V tem modelu je apriorna/aposteriorna napovedna porazdelitev normalna z naslednjimi parametri:

- apriorna napovedna porazdelitev: povprecje μ_0, varianca $\sigma_0^2 + \sigma^2$,
- aposteriorna napovedna porazdelitev: povprecje μ_n, varianca $\sigma_n^2 + \sigma^2$.

Ne glede na to, kako velik vzorec imamo oz. kako natancna je nasa aposterorna porazdelitev (majhen σ_n^2), bo varianca aposteriorne napovedne porazdelitve vsaj σ^2.

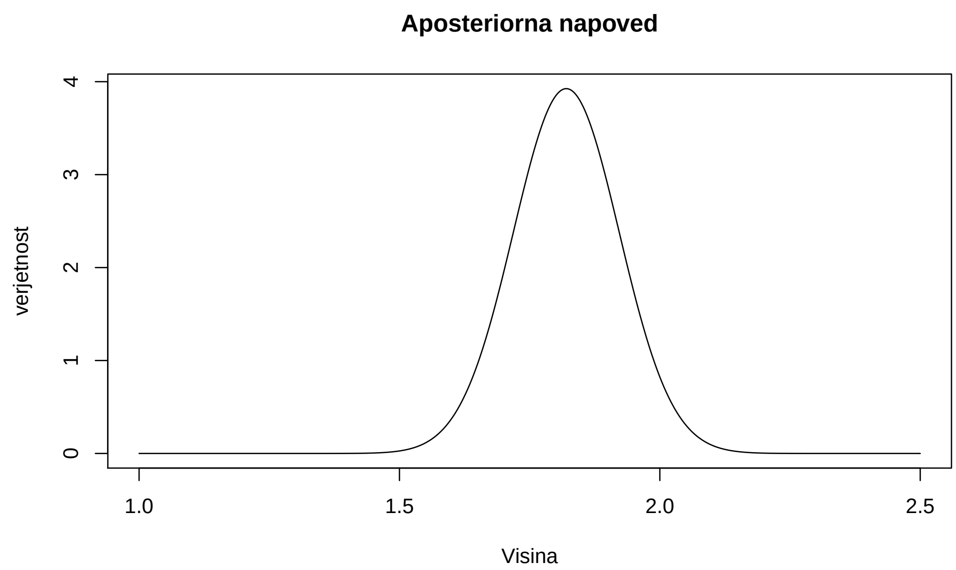
Narisemo apriorno napovedno porazdelitev.

```
theta <- seq(1, 2.5, 0.001)
plot(theta, dnorm(theta, mean = mu0, sd = sqrt(sigma0^2 + sigma^2)), type = "l",
     xlab = "Visina", ylab = "verjetnost",
     main = "Apriorna napoved")
```



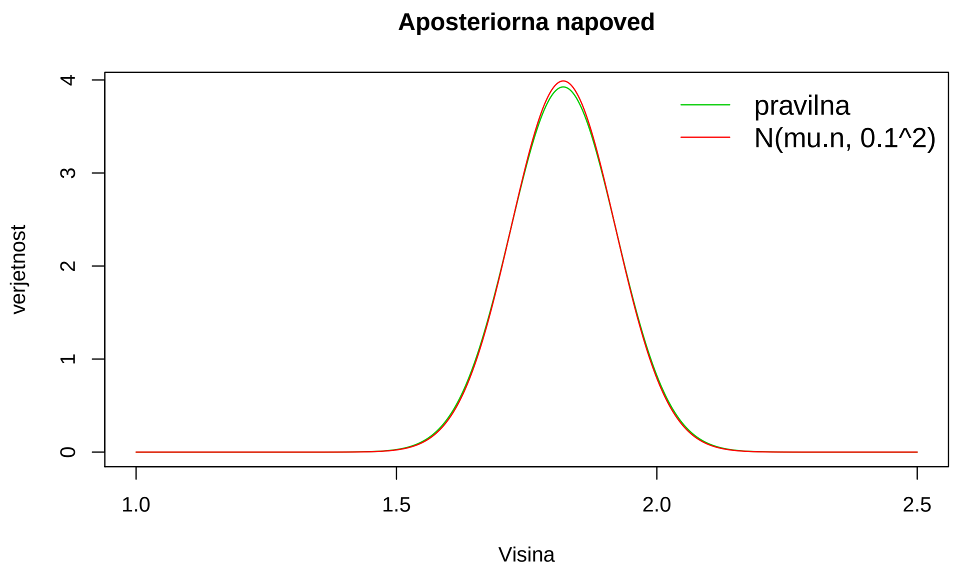
Narisemo aposteriorno napovedno porazdelitev.

```
theta <- seq(1, 2.5, 0.001)
plot(theta, dnorm(theta, mean = mu.n, sd = sqrt(sigma.n^2 + sigma^2)), type = "l",
     xlab = "Visina", ylab = "verjetnost",
     main = "Aposteriorna napoved")
```



Poglejmo si se, kaksna je razlika med pravilno izracunano aposteriorno napovedno porazdelitvijo in tisto, ki jo dobimo, ce v normalno porazdelitev z znano varianco $\sigma^2 = 0.1^2$ vstavimo naso oceno parametra $\hat{\theta} = \mu_n$, torej primerjamo s porazdelitvijo $N(\mu_n, \sigma^2)$.

```
theta <- seq(1, 2.5, 0.001)
plot(theta, dnorm(theta, mean = mu.n, sd = sqrt(sigma.n^2 + sigma^2)), type = "l",
     xlab = "Visina", ylab = "verjetnost",
     main = "Aposteriorna napoved", col="green3")
lines(theta, dnorm(theta, mean = mu.n, sd = sigma), col = "red")
legend("topright",  lty = 1,
       c("pravilna", "N(mu.n, 0.1^2)"), col = c("green3","red"), bty = "n", cex = 1.3)
```



Poudarimo bistveno razliko med aposteriorno porazdelitvijo povprecne visine in aposteriorno napovedno porazdelitvijo za visino novega studenta:

```
theta <- seq(1, 2.5, 0.001)
plot(theta, dnorm(theta, mean = mu.n, sd = sigma.n), type = "l",
     xlab = "", ylab = "verjetnost", col="purple")
lines(theta, dnorm(theta, mean = mu.n, sd = sqrt(sigma.n^2 + sigma^2)), col="green3")
legend("topleft",  lty = 1,
       c("aposteriorna napovedna", "aposteriorna"), col = c("green3","purple"),
       bty = "n", cex = 1.3)
```