

1. sklop: Binomski model

Nina Ruzic Gorenjec

1 Primer

Izberite pravilni odgovor na spodnje vprašanje.

Vprašanje: Qskd senciljm dowdlq a?

- (a) 25
- (b) 625
- (c) 1
- (d) Nic od nastetega.

Zanima nas verjetnost, da odgovorimo pravilno.

2 Verjetnostni model za nas primer

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

- n stevilo studentov na vajah,
- X_i predstavlja pravilnost odgovora i -tega studenta, tj. $X_i = 1$, ce i -ti student odgovori pravilno, in $X_i = 0$, ce le-ta odgovori napacno.

Preucujemo $X_1 + X_2 + \dots + X_n$, tj. stevilo vseh pravilnih odgovorov, ki ga oznacimo z X (druga standardna oznaka je Y v smislu izida, anglesko *outcome*).

- $X \mid \theta \sim \text{Bin}(n, \theta)$
- $P(X = k \mid \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$
- θ (na predavanjih ϑ) je verjetnost pravilnega odgovora – **parameter, ki nas zanima**
- $E(X) = n\theta$

Nas primer:

```
n <- 26
```

Nasi podatki (oznacimo s k realizacijo X na nasem vzorcu):

```
k <- 6
```

2.1 Kako bi ocenili nas parameter s “klasicno” frekventisticno statistiko? Katere metode bi lahko uporabili?

2.2 Bayesova formula

Na ravni dogodkov:

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

Na ravni gostot z enotno oznako p (lahko bi pisali tudi f):

$$p(\theta | \text{podatki}) = \frac{p(\text{podatki} | \theta) p(\theta)}{p(\text{podatki})} \propto p(\text{podatki} | \theta) p(\theta).$$

V “standardnih” oznakah:

$$\pi(\theta | x) = \frac{f(x | \theta) \pi(\theta)}{f(x)} \propto f(x | \theta) \pi(\theta).$$

Podatki x so znani. Zanima nas θ . **Poglejmo zato na zgornje kot na funkcijo θ .**

Spomnimo se funkcije verjetja (anglesko *likelihood*) $L(\theta | x) = L(\theta; x)$ in zapisimo zgornje kot:

$$\pi(\theta | x) \propto L(\theta | x) \pi(\theta).$$

Trije gradniki Bayesove formule, ki jih bomo predstavili graficno kot funkcije θ :

- **Apriorna porazdelitev** $\pi(\theta)$ (njen integral je 1) – *nase vnaprejsnje (apriorno) vedenje/prepricanje o θ , preden zberemo podatke!*
- **Verjetje** $L(\theta | x)$ (potrebno mnoziti s konstanto, tako da bo integral enak 1) – *verjetnost podatkov pri razlicnih moznih vrednostih parametra θ .*
- **Aposteriorna porazdelitev** $\pi(\theta | x)$ – *preko Bayesove formule posodobimo nase apriorno vedenje o parametru ($\pi(\theta)$) s tem, kar nam povedo podatki ($L(\theta | x)$).*

2.3 Verjetje

Število pravih odgovorov med n študenti je enako k :

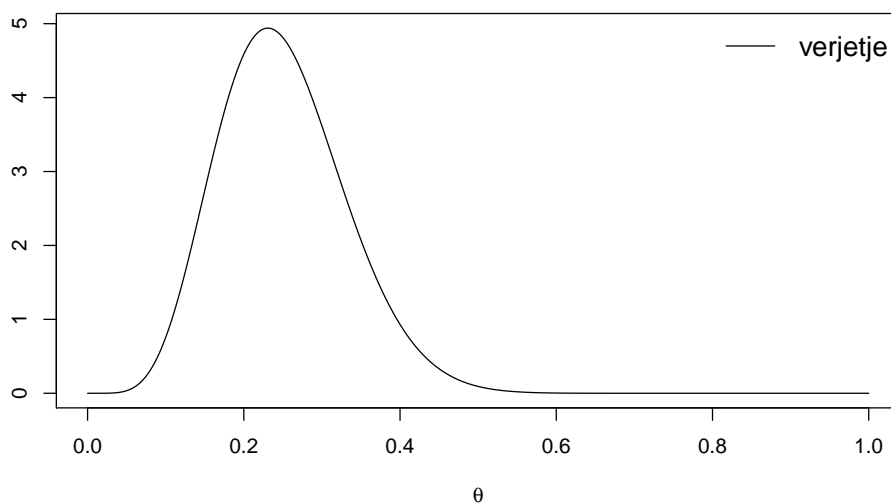
$$L(\theta \mid x) = \theta^k (1 - \theta)^{n-k}.$$

V R-u:

```
verjetje <- function(theta, k, n){  
  dbinom(k, size = n, prob = theta)  
}  
  
#Z množenjem s konst dosežemo, da je integral verjetja glede na theta enak 1.  
konst <- function(k, n){  
  theta <- seq(0.001, 1, 0.001)  
  1 / (0.001 * sum(verjetje(theta, k, n)))  
}
```

Narisemo za nas vzorec:

```
theta <- seq(0, 1, 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)
```



2.4 Apriorna porazdelitev

Za apriorno porazdelitev si izberemo beta porazdelitev, ki je v primeru binomske porazdelitve podatkov *conjugate prior* (pomeni, da apriorna in aposteriora porazdelitev pripadata enaki družini porazdelitev), zato se lahko uporablja tudi izraz **beta-binomski model**.

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

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

kjer je funkcija beta $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ in je funkcija gama $\Gamma(a) = (a-1)!$ za pozitivna cela stevila a . Spomnimo se:

- $E(\text{Beta}(\alpha, \beta)) = \frac{\alpha}{\alpha+\beta}$,
- $\text{var}(\text{Beta}(\alpha, \beta)) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Najprej poskusimo $\alpha = \beta = 1$, s čimer dobimo enakomerno zvezno porazdelitev $U[0,1]$ - **neinformativna apriorna porazdelitev**.

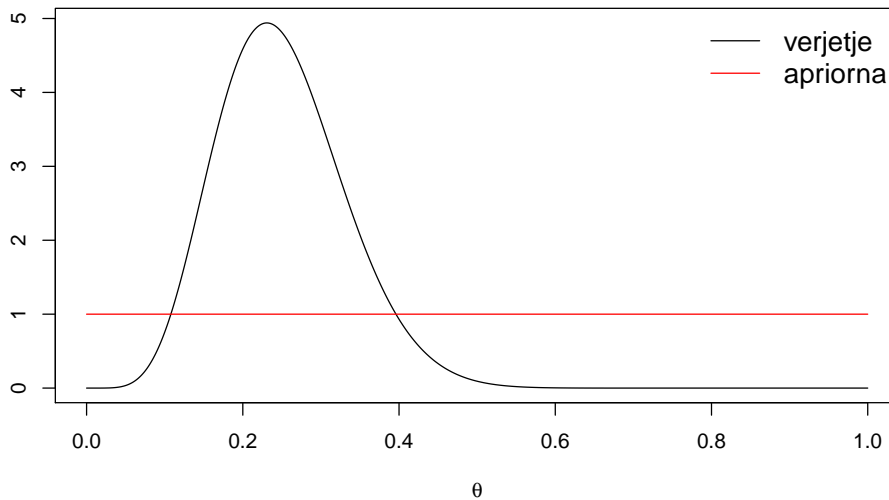
Narisemo v R-u:

```
alpha <- 1
beta <- 1

theta <- seq(0, 1, 0.001)
apriorna <- dbeta(theta, alpha, 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)
```



2.5 Aposteriorna porazdelitev

Ker smo uporabili *conjugate prior*, bo aposteriorna porazdelitev tudi iz družine beta porazdelitev, njena parametra sta enaka:

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

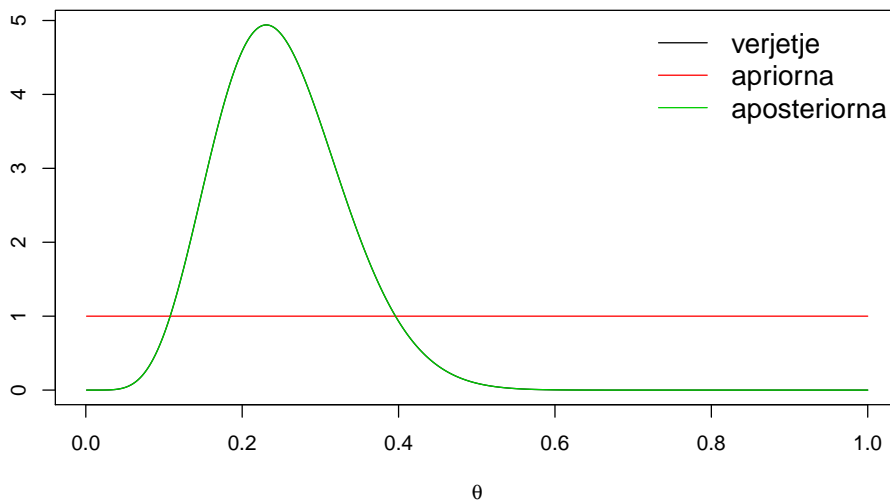
Narisemo v R-u:

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

theta <- seq(0.001, 1, 0.001)
aposteriorna <- dbeta(theta, alpha.apost, beta.apost)

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



2.6 Ocena parametra θ

Ena možnost je pričakovana vrednost aposteriorne porazdelitve:

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

```
alpha.apost / (alpha.apost + beta.apost)
```

```
## [1] 0.25
```

Ali dobimo enako kakor pri frekventističnemu pristopu?

```
k/n
```

```
## [1] 0.2307692
```

V primeru neinformativne porazdelitve $\alpha = \beta = 1$, dobimo

$$\hat{\theta} = \frac{k + 1}{n + 2}.$$

Na predavanjih ste $\hat{\theta}$, ocenjen preko pričakovane vrednosti aposteriorne porazdelitve, zapisali kot

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

kjer je $\mu = E(\text{Beta}(\alpha, \beta)) = \frac{\alpha}{\alpha + \beta}$ in $\phi = \alpha + \beta$.

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

2.7 Interval (obmocje) zaupanja v Bayesovi statistiki

Pri danih podatkih $X = x$ ima interval $[L_B(x), U_B(x)]$ 95% **Bayesovo pokritje** za θ , ce velja

$$P(L_B(x) < \theta < U_B(x) \mid X = x) = 0,95.$$

Preden zberemo podatke ima interval $[L(X), U(X)]$ 95% **frekventisticko pokritje** za θ , ce velja

$$P(L(X) < \theta < U(X) \mid \theta) = 0,95.$$

Ko poznamo podatke $X = x$, jih vstavimo v $L(X)$ in $U(X)$ ter s tem dobimo $L(x)$ in $U(x)$.

Koliko je $P(L(x) < \theta < U(x) \mid \theta)$?

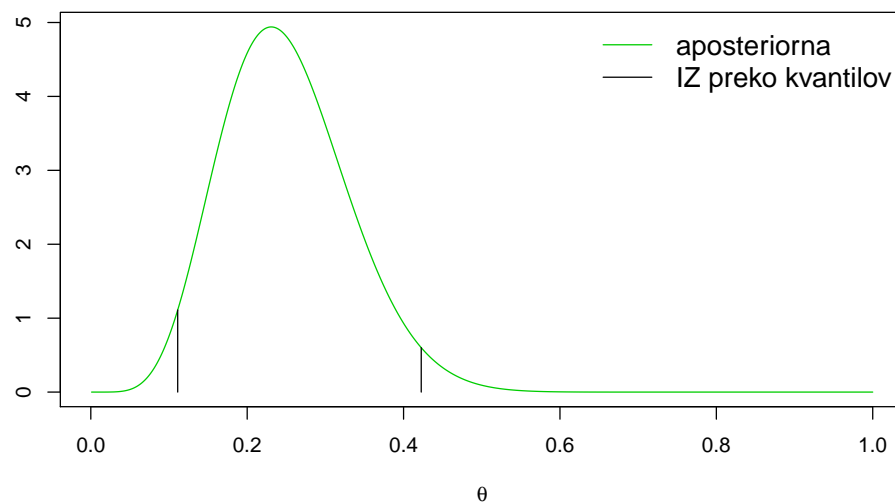
Interval zaupanja v Bayesovi statistiki (angl. pogosto *credible interval*, lahko tudi *confidence interval*) je torej katerikoli interval, v katerem “je vsebovanih 95% gostote aposteriorne porazdelitve”. Seveda pa zelimo, da je “centralen glede na porazdelitev”.

Najbolj preprosta varianta preko kvantilov porazdelitve (smiselna, ce je porazdelitev priblizno simetricna):

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

```
## [1] 0.1111446 0.4225831
```

```
plot(theta, aposteriorna, type = "l", col="green3",
      xlab = expression(theta), ylab = "")
segments(x0 = iz[1], y0 = 0,
         x1 = iz[1], y1 = dbeta(iz[1], alpha.apost, beta.apost))
segments(x0 = iz[2], y0 = 0,
         x1 = iz[2], y1 = dbeta(iz[2], alpha.apost, beta.apost))
legend("topright", legend = c("aposteriorna","IZ preko kvantilov"),
      col = c("green3","black"), lty = 1, bty = "n", cex = 1.3)
```

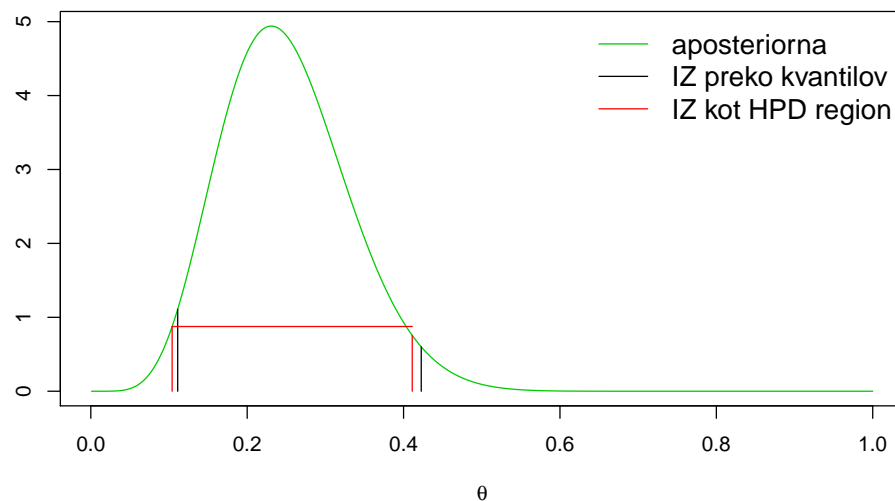


Highest posterior density (HPD) region:

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

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

```
##      lower      upper
## 0.1040288 0.4110791
## attr(,"credMass")
## [1] 0.95
```



Kaksen interval zaupanja dobimo s frekventističnim pristopom? Katere intervale zaupanja imamo na voljo?


```
prop.test(k, n, correct=F)$conf #IZ z aproksimacijo normalne porazdelitve
```

```
## [1] 0.1103385 0.4205155  
## attr("conf.level")  
## [1] 0.95
```

```
binom.test(k, n)$conf #Clopper-Pearsonov IZ
```

```
## [1] 0.08974011 0.43647510  
## attr("conf.level")  
## [1] 0.95
```

```
iz #Bayesov IZ, metoda s kvantili
```

```
## [1] 0.1111446 0.4225831
```

2.8 Testiranje hipotez

Namen starih možnih odgovorov je bil, da je verjetnost pravilnega odgovora brez kakrsnegakoli učenja dovolj majhna. Ali lahko sklepamo, da je verjetnost pravilnega odgovora manjša od 0,4?

Kako testiramo to hipotezo s Bayesovim pristopom? Kaj mora biti ničelna in kaj alternativna hipoteza?

Kako testiramo s frekventističnim pristopom?

3 Primerjava dveh neodvisnih delezev

Nase vprasanje iz zacetka navodil tega sklopa zastavimo se skupini studentov, ki se je ucila.

Od 30 studentov, jih je 21 odgovorilo pravilno.

Verjetnost pravilnega odgovora za studenta, ki se uci, naj bo θ_{uci} . Ocenimo jo z uporabo neinformativne apriorne porazdelitve (tako kakor prej, $\alpha = \beta = 1$). Dobimo aposteriorno porazdelitev

$$\theta_{\text{uci}} \sim \text{Beta}(21 + 1, 30 - 21 + 1) = \text{Beta}(22, 10).$$

Ocenimo $\hat{\theta}_{\text{uci}} = 22/(22 + 10) = 0,6875$.

Kaj smo predpostavili, da smo lahko uporabili ta model?

Prej smo ocenili verjetnost pravilnega odgovora za studenta, ki nakljucno izbere pravilni odgovor, kot

$$\theta_{\text{blef}} \sim \text{Beta}(6 + 1, 26 - 6 + 1) = \text{Beta}(7, 21).$$

Ocenimo $\hat{\theta}_{\text{blef}} = 7/(7 + 21) = 0,25$.

Zelimo primerjati ti dve verjetnosti.

Na kaksne nacine ju lahko primerjamo? Katere mere povezanosti lahko izracunamo?

- Razlika delezev (angl. *risk difference*): $\theta_{\text{uci}} - \theta_{\text{blef}}$.
- Relativno tveganje (angl. *risk ratio*): $\theta_{\text{uci}}/\theta_{\text{blef}}$.
- Razmerje obetov (angl. *odds ratio*):

$$\frac{\theta_{\text{uci}}/(1 - \theta_{\text{uci}})}{\theta_{\text{blef}}/(1 - \theta_{\text{blef}})}.$$

Kako ocenimo vsako izmed teh mer?

Ali znate s orodji frekventisticne statistike testirati, da je razlika delezev vecja od nic?

Kako bi testirali, da je relativno tveganje vecje od ena (studentje, ki se ucijo, imajo torej vecjo verjetnost pravilnega odgovora)? Kako bi izracunali interval zaupanja za relativno tveganje?

Frekventisticno?

Bayesovsko?

Moramo pri teh dveh pristopih kaj izpeljati?

4 Napovedovanje (angl. *prediction*)

Izpit je sestavljen iz desetih vprasanj (taksnih iz zacetka navodil tega sklopa).

1. Denimo, da bi pred zacetkom prvih vaj dali izpit v resevanje nekemu studentu. Kaj lahko povemo o porazdelitvi stevila njegovih pravilnih odgovorov?
2. Na prvih vajah smo pridobili vzorec, s katerim smo preizkusili, kako na vprasanje odgovarjamo, ce ne znamo cisto nic. Vzorec ste bili studentje, prisotni na prvih vajah. Izpit damo v resevanje studentu, **ki ni bil prisoten na prvih vajah** in se tudi ni ucil. Kaj lahko povemo o porazdelitvi stevila njegovih pravilnih odgovorov?

Odgovor na 1. vprasanje je **apriorna napovedna porazdelitev** (angl. *prior predictive distribution*).

Ta nas tipicno ne zanima.

Odgovor na 2. vprasanje je **aposteriorna napovedna porazdelitev** (angl. *posterior predictive distribution*).

Splosna formula za apriorno napovedno porazdelitev:

$$f(x_{\text{nov}}) = \int_{\Theta} f(x_{\text{nov}}, \theta) d\theta = \int_{\Theta} f(x_{\text{nov}} | \theta) \pi(\theta) d\theta.$$

Splosna formula za aposteriorno napovedno porazdelitev:

$$f(x_{\text{nov}} | x) = \int_{\Theta} f(x_{\text{nov}}, \theta | x) d\theta = \int_{\Theta} f(x_{\text{nov}} | \theta, x) \pi(\theta | x) d\theta = \int_{\Theta} f(x_{\text{nov}} | \theta) \pi(\theta | x) d\theta.$$

V nasem modelu (binomski model z apriorno beta porazdelitvijo) je:

- $\pi(\theta) \sim \text{Beta}(\alpha, \beta)$; izbrali smo $\alpha = 1, \beta = 1$
- $\pi(\theta | x) \sim \text{Beta}(\alpha_{\text{apost}}, \beta_{\text{apost}}) = \text{Beta}(k + \alpha, n - k + \beta)$; za nas vzorec velikosti $n = 26$ smo dobili $k = 6$
- za $x_{\text{nov}} \equiv K \in \{0, 1, \dots, N\}$ je $f(x_{\text{nov}} | \theta) = \binom{N}{K} \theta^K (1 - \theta)^{N-K}$; določili smo $N = 10$, zanimajo nas vsi možni K

Izkaze se, da je iskana apriorna ali aposteriorna napovedna porazdelitev iz družine t.i. **beta-binomske porazdelitve** (BetaBin). To je diskretna porazdelitev Y s parametri $N \in \mathbb{N}$ in $\tilde{\alpha}, \tilde{\beta} > 0$, ki lahko zavzame vrednosti $K \in \{0, 1, \dots, N\}$ in je

$$P(Y = K) = \binom{N}{K} \frac{B(K + \tilde{\alpha}, N - K + \tilde{\beta})}{B(\tilde{\alpha}, \tilde{\beta})}.$$

Apriorna napovedna porazdelitev v binomskem modelu: BetaBin(N, α, β).

Aposteriorna napovedna porazdelitev v binomskem modelu: BetaBin($N, \alpha_{\text{apost}}, \beta_{\text{apost}}$) oziroma BetaBin($N, k + \alpha, n - k + \beta$).

Beta-binomska porazdelitev v R (je vključena tudi v nekaterih paketih, ponekod drugače parametrizirana):

```
dbetabinom <- function(K, N, a, b){  
  choose(N, K) * beta(K+a, N-K+b) / beta(a, b)  
}
```

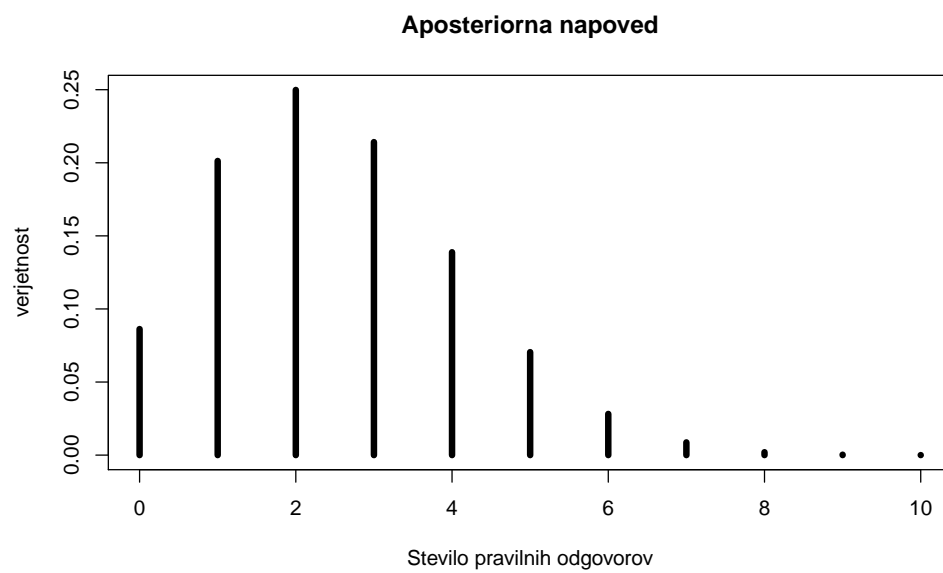
Narisemo apriorno napovedno porazdelitev.

```
plot(0:10, dbetabinom(0:10, N = 10, a = alpha, b = beta), type = "h",  
     xlab = "Število pravih odgovorov", ylab = "verjetnost",  
     main = "Apriorna napoved", ylim = c(0, 0.1), lwd = 5)
```



Narisemo aposteriorno napovedno porazdelitev.

```
plot(0:10, dbetabinom(0:10, N = 10, a = alpha.apost, b = beta.apost), type = "h",  
     xlab = "Število pravih odgovorov", ylab = "verjetnost",  
     main = "Aposteriorna napoved", lwd = 5)
```



Ali je bilo vse to racunanje res potrebno?

- Nasa ocena parametra po upostevanju podatkov nasega vzorca je $\hat{\theta} = \alpha_{\text{apost}} / (\alpha_{\text{apost}} + \beta_{\text{apost}}) = 0.25$.
- Stevilo pravih odgovorov je porazdeljeno $\text{Bin}(10, \theta)$.
- Ali je preprosto aposteriorna porazdelitev kar $\text{Bin}(10, \hat{\theta})$?

5 Jeffreyeva apriorna porazdelitev

To je **neinformativna** apriorna porazdelitev, ki je proporcionalna $\sqrt{\det \mathcal{I}(\theta)}$. Zanja je znailno, da je invariantna glede na razlicne reparametrizacije prostora parametrov.

Pri binomskem modelu je Jeffreyeva apriorna porazdelitev $\text{Beta}(\alpha = 0.5, \beta = 0.5)$.

Pri nasih podatkih dobimo:

