Applicazioni: Inferenza Bayesiana Corso di Laurea Magistrale in Biostatistica

Prof.ssa Fulvia Pennoni fulvia.pennoni@unimib.it

A.A. 2022-2023

Contents

Queste dispense riguardano la parte delle applicazioni del corso di **Inferenza Bayesiana** ed integrano le dispense di teoria. Si prega di riportare adeguatamente la citazione del presente materiale didattico come segue:\

Pennoni, F. (2022). *Dispense dell'insegnamento di inferenza Bayesiana: Applicazioni con R.* Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

Le seguenti sono state scritte in R Markdown, utilizzando la libreria knitr.

```
citation("knitr")
```

Modello Beta-Binomiale: Esempio ipertensione

In un campione di 20 (n) pazienti 13 (k) pazienti sono ipertesi. Si ipotizza un modello Beta-Binomiale per stimare la proporzione di pazienti ipertesi nella popolazione di riferimento.

Si disegna la distribuzione a posteriori in base a diverse formulazioni della distribuzione a priori.

Si considerano i seguenti tre senari: * (1) distribuzione a priori uniforme $\alpha = 1, \beta = 1$;

- (2) distribuzione a priori di Jeffrey con $\alpha = \beta = 0.5$;
- (3) distribuzione a priori con $\alpha, \beta > 1$

Scenario 1

Suppendo che la distribuzione a priori per il parametro θ che rappresenta la probabilità di successi (ipertesi) sia una Beta (α,β) tale che $\alpha = \beta = 1$.

La verosimiglianza è una Binomiale il cui kernel è proporzionale alla Beta con parametri $(\alpha = k + 1)$ e $(\beta = n - k + 1)$ (cf. Dispense di teoria).

La distribuzione a posteriori è ancora una Beta con parametri $(\alpha + k)$ e $(\beta + n - k)$.

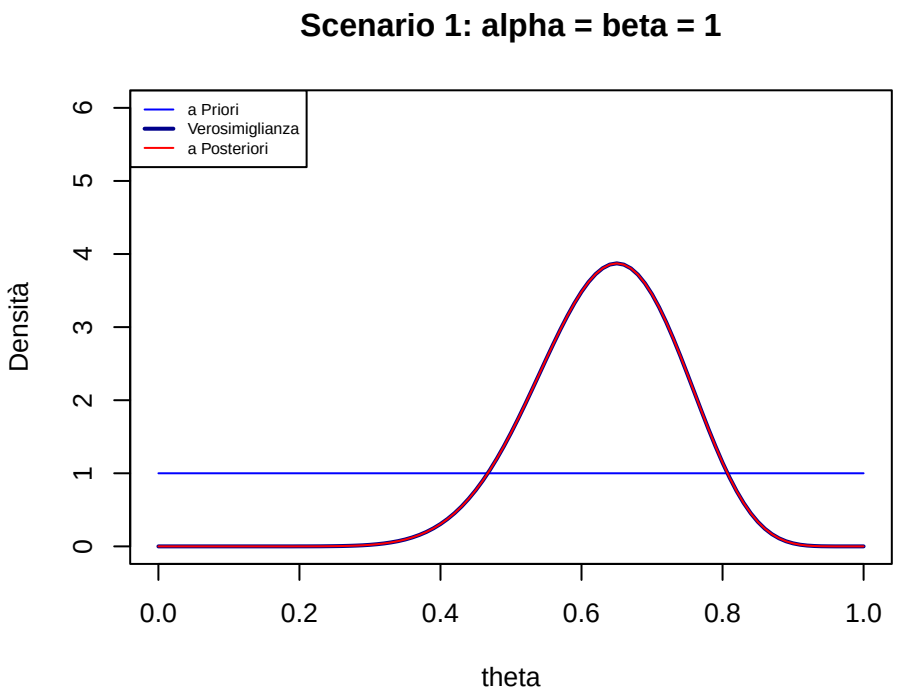
Nel seguente grafico si visualizzano le tre distribuzioni.

```
alpha <- 1
beta<- 1

pval<-seq(0,1,by=0.01)
##  distribuzione a priori
plot(pval, dbeta(pval,alpha,beta),
     type = "l", col = "blue", ylim = c(0,6),
     ylab = "Densità", xlab = "theta",
     main = "Scenario 1: alpha = beta = 1")
```

```r
## verosimiglianza
n <- 20
k <- 13
lines(pval, dbeta(pval, k+1, n-k+1),
      lwd = 2,
      col = "darkblue")

## distribuzione a posteriori
lines(pval, dbeta(pval, k+alpha, n-k+beta),
      lwd = 1,
      col = "red")
legend("topleft",
       c("a Priori","Verosimiglianza","a Posteriori"),
       lty=c(1,1,1),
       lwd=c(1,2,1),
       col=c( "blue","darkblue","red" ), cex=0.6)
```



La numerosità campionaria è esigua. Si evince che nel caso in cui la prior *non è informativa* (scenario 1) la distribuzione a posteriori coincide con la verosimiglianza.

Il valore atteso della distribuzione a posteriori costistituisce la stima puntuale Bayesiana per il parametro

```
(alpha + k)/(n+alpha + beta)
#> [1] 0.6363636
```

la proporzione di ipertesi nella popolazione è 0.64.

La moda della distribuzione a priori è un'altra stima puntuale per il parametro e coincide con la stima di massima verosimiglianza che è pari a k/n

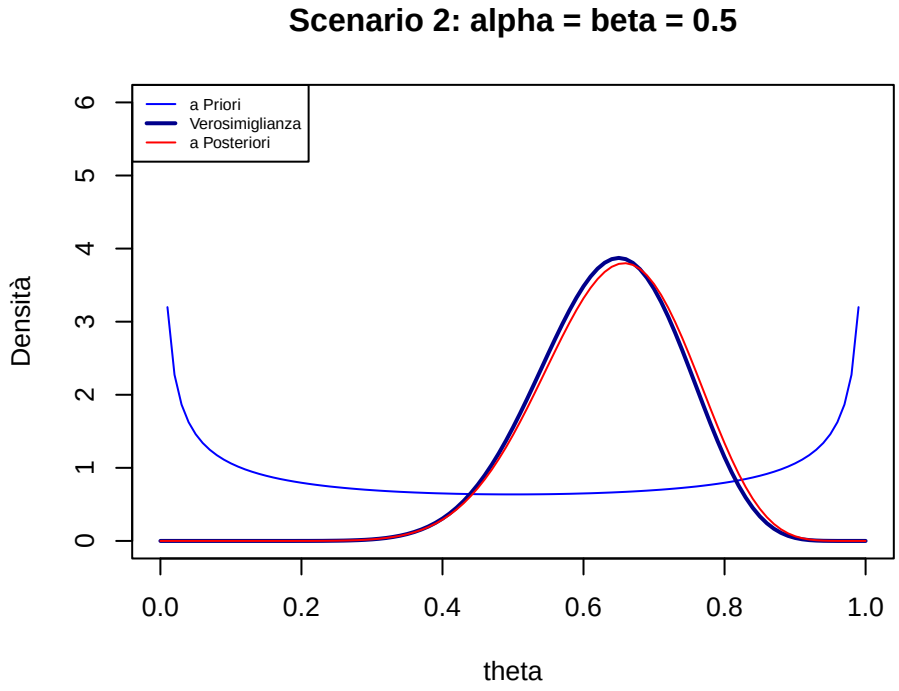
```
(alpha + k -1)/(n + alpha + beta -2)
#> [1] 0.65
```

Scenario 2

Scegliendo come *iperparametri* della distribuzione a priori quelli proposti da Jeffrey ovvero $\alpha = 0.5$ e $\beta = 0.5$ la prior non è molto informativa

```
alpha <- 0.5
beta <- 0.5
# distribuzione a priori
pval<-seq(0,1,by=0.01)
plot(pval, dbeta(pval,alpha,beta),
     type = "l", col = "blue", ylim = c(0,6),
     ylab = "Densità", xlab = "theta",
     main = "Scenario 2: alpha = beta = 0.5")
#  verosimiglianza

lines(pval, dbeta(pval, k+1, n-k+1),
      lwd = 2,
      col = "darkblue")
# distribuzione a posteriori
lines(pval, dbeta(pval, k+alpha, n-k+beta),
      lwd = 1,
      col = "red")
legend("topleft",
       c("a Priori","Verosimiglianza","a Posteriori"),
       lty=c(1,1,1),
       lwd=c(1,2,1),
       col=c( "blue","darkblue","red" ), cex=0.6)
```



In questo scenario in cui la prior è un po' più informativa si nota la diversa forma della distribuzione a priori e la traslazione della a posteriori verso destra rispetto alla verosimiglianza.

Il valore atteso adesso è

```
(alpha + k)/(n+alpha + beta)
#> [1] 0.6428571
```

La moda è leggermente superiore a quella osservata per lo scenario 1

```
(alpha + k -1)/(n + alpha + beta -2)
#> [1] 0.6578947
```

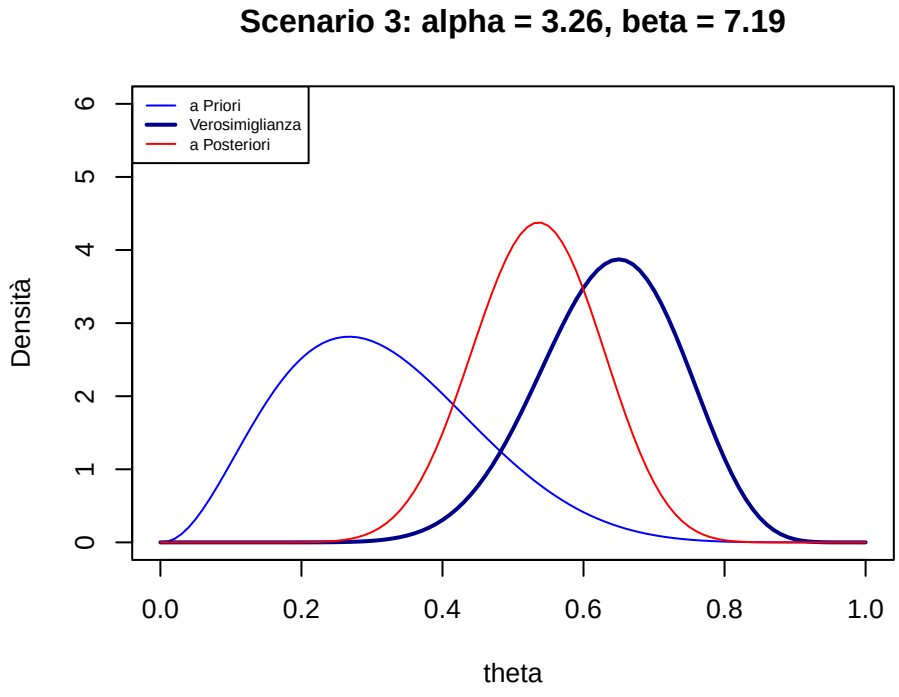
Scenario 3

Nel caso in cui si suppone una distribuzione a priori *molto informativa* ad esempio con $\alpha = 3.26$ e $\beta = 7.19$ si ha

```r
alpha <- 3.26
beta<- 7.19
# distribuzione a priori
pval<-seq(0,1,by=0.01)
plot(pval, dbeta(pval,alpha,beta),
     type = "l", col = "blue", ylim = c(0,6),
     ylab = "Densità", xlab = "theta",
     main = "Scenario 3: alpha = 3.26, beta = 7.19")

# verosimiglianza
lines(pval, dbeta(pval, k+1, n-k+1),
      lwd = 2,
      col = "darkblue")

# distribuzione a posteriori
lines(pval, dbeta(pval, k+alpha, n-k+beta),
      lwd = 1,
      col = "red")
legend("topleft",
       c("a Priori","Verosimiglianza","a Posteriori"),
       lty=c(1,1,1),
       lwd=c(1,2,1),
       col=c( "blue","darkblue","red" ), cex=0.6)
```



In questo caso la distribuzione a priori è molto infomativa e pertanto "trascina" verso la sua

moda la distribuzione a posteriori che è sempre spostata verso la verosimiglianza ma meno vicina a quest'ultima rispetto ai due scenari precenti.

La stima puntuale della probabilità di successo in base al valore atteso della distribuzione a posteriori è

```
(alpha + k)/(n+alpha + beta)
#> [1] 0.5339901
```

Anche la stima in base alla moda è inferiore rispetto ai due precedenti scenari 1 e 2

```
(alpha + k -1)/(n + alpha + beta -2)
#> [1] 0.5363796
```

Nello scenario considerato la distribuzione a priori è *piuttosto informativa* ed essendo la numerosità campionaria non troppo elevata la distribuzione a posteriori risulta influenzata maggiormente dalla distribuzione a priori, il peso della funzione di verosimiglianza è minore.

Specificazione della distribuzione a priori

Si intende stimare la proporzione di soggetti che soffrono di sintomi gravi dovuti al COVID-19 e non sono vaccinati nella popolazione italiana.

- In base alle attuali conoscenze a priori si assume 0.3 come valore plausibile per la mediana della distribuzione e 0.5 come valore plausibile per il 90-esimo percentile della distribuzione a priori.
- Si utilizza la funzione `beta.select` della libreria `LearnBayes` per determinare i parametri α_1 e β_1 della distribuzione a priori con la funzione seguente `learnBayes::beta.select`.

Le misure di posizione devono essere definite come una **lista** specificando la mediana `quantile1` ed il novantesino centile `quantile2` della distribuzione

```
quantile1 <- list(p = 0.5, x=0.3); quantile1
#> $p
#> [1] 0.5
#>
#> $x
#> [1] 0.3
quantile2 <- list(p = 0.9, x=0.5); quantile2
#> $p
#> [1] 0.9
#>
#> $x
#> [1] 0.5
```

La funzione `beta.select` restituisce i valori dei due parametri della distribuzione Beta che rispettano i vincoli imposti

```
require(LearnBayes)
beta.select(quantile1, quantile2)
#> [1] 3.26 7.19
```

da cui risulta la distribuzione a priori $p(\theta) \sim Beta(3.26, 7.19)$ e la proporzione media di ipertesi nella popolazione in base alle conoscenze a priori è

```r
M <- 3.26/(3.26+7.19); M
#> [1] 0.3119617
```

Determinazione della distribuzione a posteriori

Supponendo di osservare adesso un campione casuale di persone ricoverate per covid-19 in un'ospedale della zone e di riscontrare che su 12 pazienti complessivi (n_1) pazienti 6 non risultano vaccinati. Nel seguito si utilizza $k_1 = 6$ "successi" e $f_1 = (n_1 - k_1)$ "insucessi"

Si disegna: - la **distribuzione a priori**

- la **verosimiglianza**
- la **distribuzione a posteriori**

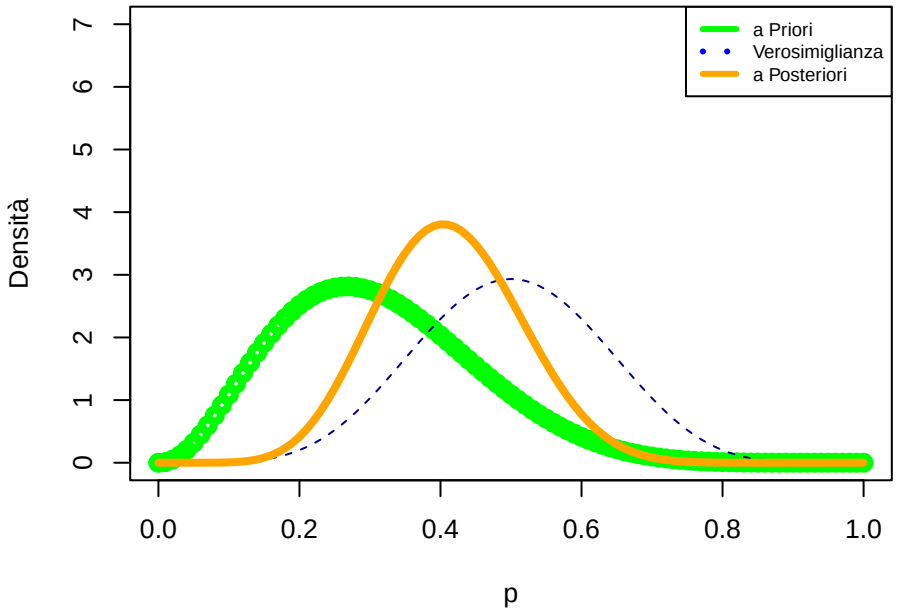
```r
alpha1 <- 3.26
beta1 <- 7.19
n1 <- 12
k1 <- 6

# a priori
pval <- seq(0,1, by = 0.01)
plot(pval, dbeta(pval,alpha1, beta1),
     xlab="p",
      ylab="Densità",
      ylim=c(0,7),
      lty=3,
      lwd=4,
      col="green")

# verosimiglianza
lines(pval,dbeta(pval, k1+1,n1-k1+1),
      lty = 2,
      lwd = 1, col = "darkblue")

# a posteriori
lines(pval, dbeta(pval,
                  k1 + alpha1,
                  n1-k1+beta1),
      lty=1,lwd=4,
      col="orange")
```

```
legend("topright",
       c("a Priori","Verosimiglianza","a Posteriori"),
       lty=c(1,3,1),
       lwd=c(3,3,3),
       col=c("green", "blue","orange" ),
       cex = 0.7)
```



- Si nota che la distribuzione di riferimento per l'inferenza ovvero la distribuzione a posteriori è un compromesso tra la distribuzione iniziale che ha portato ad ipotizzare certi valori per il parametro (e che è abbastanza informativa) e la verosimiglianza basata sulle rilevazioni effettuate con il campione che è abbastanza diversa dalla prior.
- La media della distribuzione a posteriori si trova tra la media della prior e quella della verosimiglianza.
- La distribuzione a posteriori ha una variabilità ridotta rispetto alla variabilità della distribuzione iniziale.

Aggiornamento della distribuzione a posteriori

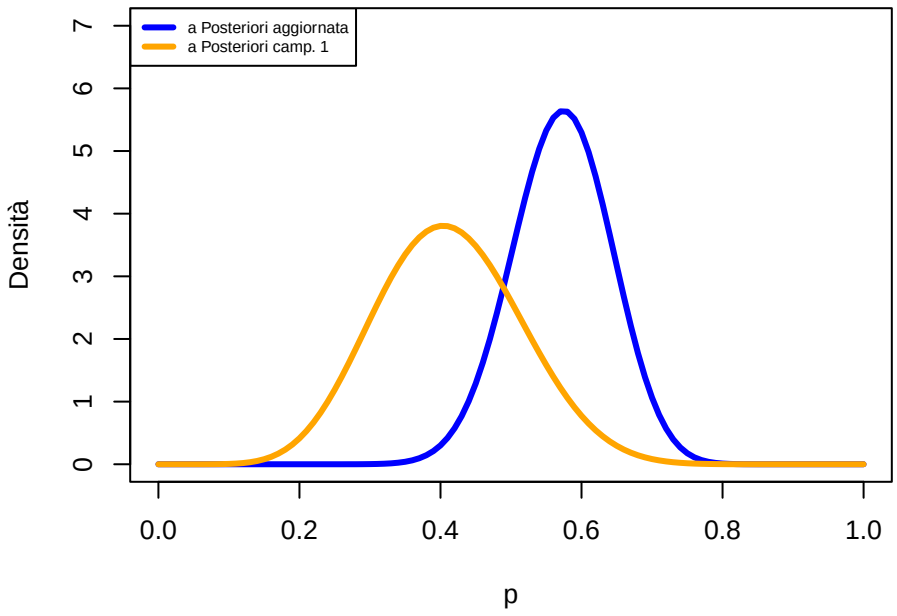
Nel seguito si applica il ragionamento sequenziale e si suppone di osservare un altro campione di pazienti ricoverati in un altro ospedale di Milano la distribuzione a posteriori **viene aggiornata** in base alle nuove evidenze empiriche. Su 27 pazienti ricoverati per covid-19, 19 risultano senza vaccino $n = 27+12 = 39$, $k = 19+6 = 25$.

Il confronto grafico tra la prima e la seconda distribuzione a posteriori ci permette di stabilire come si modifica l'inferenza sul parametro a seguito di ulteriori evidenze campionarie

```
k <- 19 +k1
n<- 39
# a posteriori aggiornata
plot(pval, dbeta(pval,
                  alpha1 + k,
                  beta1+n-k),
      xlab="p",
      type = "l",
      ylab="Densità",
      ylim=c(0,7),
      lty=1,
      lwd=3,
      col="blue")

# a posteriori campione 1
lines(pval, dbeta(pval,
                  k1 + alpha1,
                  n1-k1+beta1),
      lty=1,lwd=3,
      col="orange",
      add=TRUE)

legend("topleft",
       c("a Posteriori aggiornata", "a Posteriori camp. 1"),
       lty=c(1,1),
       lwd=c(3,3),
       col=c("blue", "orange" ),
       cex = 0.6)
```



Il valore atteso della distribuzione a posteriori aggiornata con le ulteriori evidenze campionarie è

```
(alpha1+k)/(n+alpha1 + beta1)
#> [1] 0.5714863
```

Notiamo che la distribuzione a posteriori è proporzionale al prodotto della prior e della verosimiglianza.

Esempio Bayes Billiard Balls

Si considera l'esempio proposto a pagina 20 delle dispense di teoria. Si considera la variabile casuale che identifica il numero di palline blu che si posizionano prima della pallina gialla fissando il numero di palline blu pari a 10.

Tramite la simulazione dell'esperimento al calcolatore si verifica la prima espressione dell'equazione.

Si ripete l'esperimento di lanciare le palline 100000 volte contando ogni volta il numero dei successi in base a diversi valori della probabilità di successo generati come numeri pseudo-casuali da una distribuzione $Beta(\alpha = 1, \beta = 1)$ (uniforme in [0,1]).

- Utilizzando una **distribuzione a priori non informativa** si generano dei numeri pseudo-casuali

```
nsim <- 10^5
set.seed(1234)
p <- rbeta(nsim, shape1 = 1, shape2 = 1)
length(p)
#> [1] 100000
head(p)
#> [1] 0.8862966 0.3907253 0.1390846 0.9905042 0.3339162 0.3064087
```

- Utilizzando i valori generati si simulano i valori della **distribuzione marginale** sapendo che fissando n come numero complessivo delle prove si ha

$$\int_0^1 x^k (1-x)^{n-k} dx = \frac{k!(n-k)!}{(n+1)!}$$

l'oggetto `pxk` contiene il numero di successi ottenuti in ogni prova con diversa probabilità di successo p. Se fissiamo il seme e generiamo l'esito di una sola prova abbiamo

```
n <- 10
set.seed(1234)
pxk <- rbinom(1,n,p)
head(p)
#> [1] 0.8862966 0.3907253 0.1390846 0.9905042 0.3339162 0.3064087
head(pxk)
#> [1] 10
```

dove il primo elemento pari a 10 indica che nel primo esperimento su 10 prove sono stati ottenuti 10 successi con una probabilità di successo è 0.89.

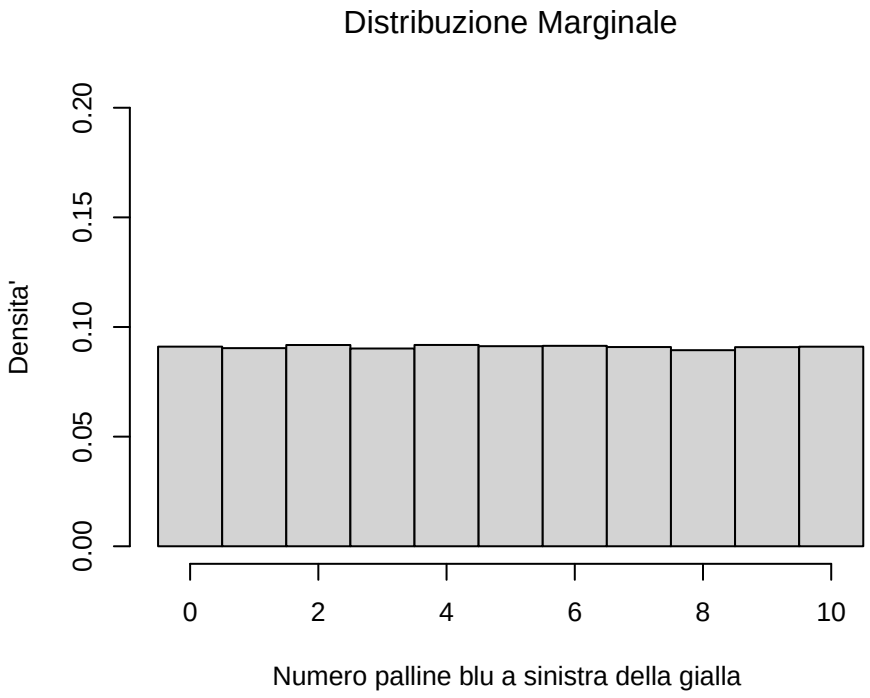
```
n <- 10
set.seed(1234)
pxk <- rbinom(nsim,n,p)
head(pxk)
#> [1] 10  4  2 10  5  4
```

Nella seconda simulazione su 10 prove sono stati ottenuti 4 successi con probabilità di successo pari a 0.39.

Nell'esempio la distribuzione marginale di X non dipende da k (numero di successi) infatti come dimostrato nelle dispense $P(X = k) = \frac{1}{n+1} = \frac{1}{11} = 0.09090909$.

Disegnando l'istogramma rispetto ai valori realizzati e salvati in `pxk` si nota che la distribuzione è assimilabile a quella di un'iniforme in (0,10) e che le densità sono proporzionali a $\frac{1}{n+1}$

```
h <- hist(pxk,
     breaks = seq(-0.5,n+0.5,1),
    freq=FALSE,
    ylim = c(0,0.20),
    xlab = "Numero palline blu a sinistra della gialla ",
     ylab= "Densita'",
     main = expression("Distribuzione Marginale")
    )
```




```
h$density
#>  [1] 0.09105 0.09036 0.09179 0.09020 0.09181 0.09125 0.09141 0.09086 0.08943
#> [10] 0.09082 0.09102
```

Modello Gaussiano: misurazione del FEV1

Occorre stimare il volume medio di aria che può essere espirato con uno sforzo massimale in un secondo (FEV1) a seguito dell'assunzione di un farmaco per l'asma, in modo da valutare la capacità dei bronchi. Si intende fare inferenza sulla media di una variabile casuale continua considerando nota la varianza della popolazione.

Specificazione dei parametri della distribuzione a priori (iper-parametri)

L'azienda che produce il farmaco dichiara che il quinto percentile è 100 e il settantesimo è 180. L'inferenza in ambito Bayesiano considera noti i parametri della distribuzione a priori. E' possibile ricavare la media e la deviazione standard della distribuzione normale che soddisfa i percentili dichiarati dall'azienda.

Utilizzando la funzione `learnBayes::normal.select` con i quantili di riferimento si ottengono i parametri della distribuzione a priori $p(\theta) \sim N(\mu, \tau^2)$

```
quantile1 <- list(p = 0.05, x=100); quantile1
#> $p
#> [1] 0.05
#>
#> $x
#> [1] 100
quantile2 <- list(p = 0.7, x=180); quantile2
#> $p
#> [1] 0.7
#>
#> $x
#> [1] 180
```

La funzione `normal.select` restituisce i valori dei due parametri (**media** e **deviazione standard**) della distribuzione di Gauss che rispettano i vincoli imposti

```
require(LearnBayes)
normal.select(quantile1, quantile2)
#> $mu
#> [1] 160.6606
#>
#> $sigma
#> [1] 36.87904
```

la distribuzione a priori per il parametro che rappresenta il volume medio espirato in un secondo è ha i seguenti iperparametri $N(161, 37^2)$.

Informazioni campionarie

Disponendo di 31 rilevazioni effettuate su un campione di pazienti a seguito dell'assunzione del farmaco si osserva un valore medio pari a 135 e si suppone che la deviazione standard nella popolazione sia pari a 12 (ovviamente anche questo parametro potrebbe essere stimato, ma per il momento lo consideriamo noto).

Pertanto $f(\bar{\mathbf{x}}; \theta) \sim N(\theta, \sigma^2/n)$.

Distribuzione a posteriori

La distribuzione a posteriori è $p(\theta|\mathbf{x}) \sim N(\mu_1, \tau_1^2)$.

La stima del valore atteso è

```
n <- 31
mu1<- (161*12^2 + (n*135*37^2))/(n*37^2 + 12^2); mu1
#> [1] 135.0879
```

dove μ_1 è una media ponderta della media a priori e della media campionaria.

Mentre la varianza della distribuzione a posteriori è

```
tau12<- ((37^2)*12^2)/(n*37^2+12^2);
tau12
#> [1] 4.629453
# oppure
1/(31/12^2 + 1/37^2)
#> [1] 4.629453
#
tau1<-sqrt(tau12)
tau1
#> [1] 2.151616
```

Pertanto la deviazione standard a priori è 36.879, quella del modello assunto per i dati è $12/\sqrt(n) = 2.155$, mentre quella della distribuzione a posteriori risulta pari a 2.15. Pertato sia la verosimiglianza che la distribuzione a posteriori hanno simile e ridotta variabilità.

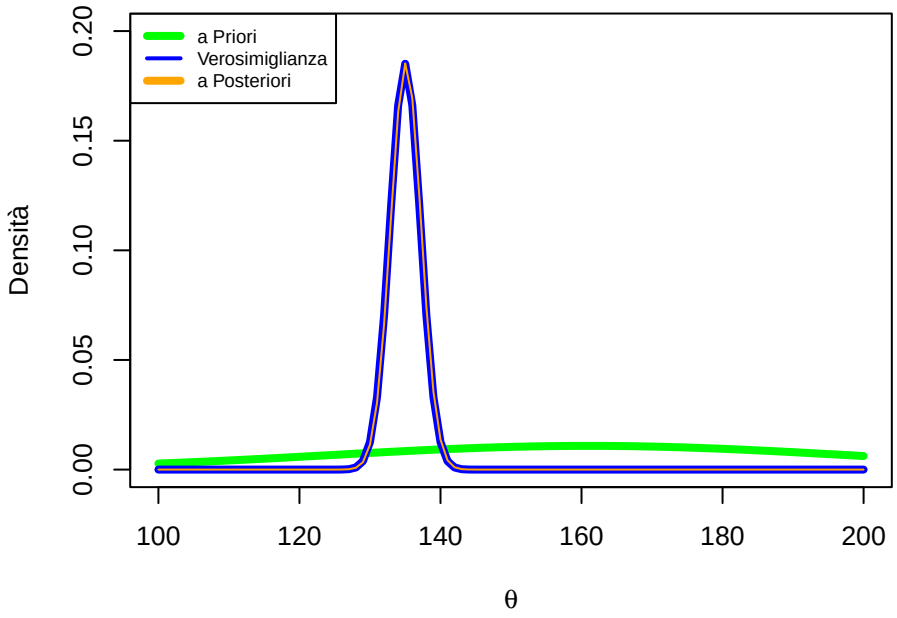
Rappresentazione grafica: scenario 1

Si disegnano le funzioni di densità per lo scenario illustrato in precedenza (si noti che la rappresentazione della verosimiglianza rigurarda il modello che ha generato i dati, per cui la varianza è da intendersi σ^2).

```
# distribuzione a priori
mu <- 161
tau <- 36.879
curve(dnorm(x,mu,tau),
      xlab = expression(theta),
      ylab="Densità",
      xlim=c(100,200),
      ylim = c(0,0.2),
      lwd=4,
      col="green")
n<-31
# verosimiglianza della media campionaria
xbar <- 135
sigma <- 12
ss<-sigma/sqrt(n)
curve(dnorm(x,xbar,ss),
      lwd=4,
      col= "blue",
      add=TRUE)

# distribuzione a  posteriori
curve(dnorm(x,mu1,tau1),
      lwd=1,
      col="orange",
      add=TRUE)

legend("topleft",
       c("a Priori","Verosimiglianza","a Posteriori"),
       lwd=c(4,2,4),
       cex = 0.7,
       col=c("green", "blue","orange" ))
```



Dato che la distribuzione iniziale non è molto informativa (la variabilità è piuttosto elevata) la distribuzione delle osservazioni è quella che determina il valore medio a posteriori.

La distribuzione a posteriori è centrata sulla media della verosimiglianza e presenta stessa variabilità.

Prior maggiormente informativa

Supponendo una distribuzione iniziale più informativa (con ridotta variabilità rispetto alla precedente, ovvero maggiormente concentrata intorno alla media) $ p( ) N(161, 1)$ si ottengono le seguenti stime per i parametri della distribuzione a posteriori

```
mu1 <- (161*12^2 + (n*135*1^2))/(n*1^2 + 12^2); mu1
#> [1] 156.3943

tau12 <- ((1^2)*12^2)/(n*1^2+12^2);
tau12
#> [1] 0.8228571
tau1 <-sqrt(tau12)
```

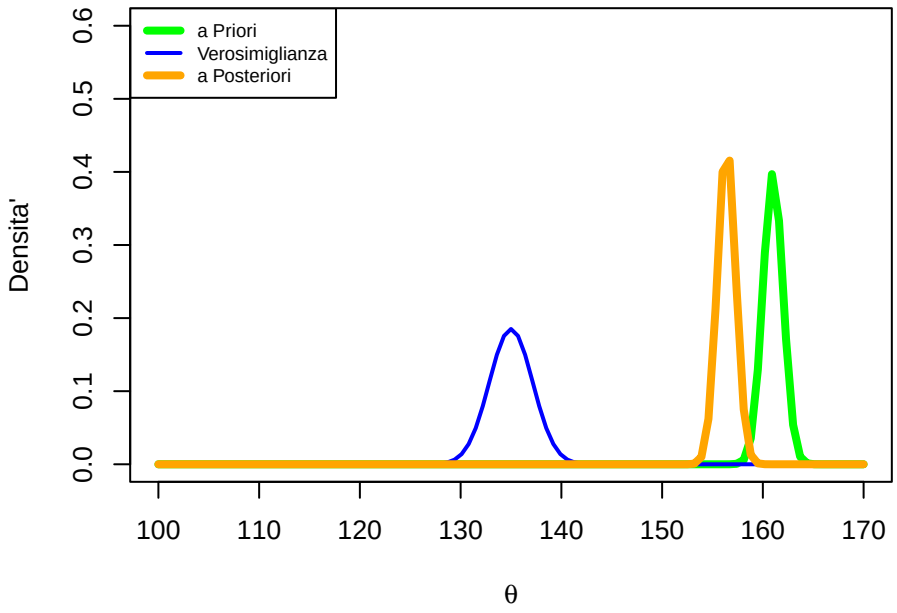
Rappresentazione grafica: scenario 2

```
# distribuzione a priori
mu <- 161
tau <- 1
curve(dnorm(x,mu,tau),
      xlab= expression(theta),
      ylab="Densita'",
      xlim=c(100,170),
      ylim = c(0,0.6),
      lwd=4,
      col="green")

# verosimiglianza
sigma <- 12
n<-31
ss<-sigma/sqrt(n)
curve(dnorm(x,xbar,ss),
      lwd=2,
      col= "blue",
      add=TRUE)

# distribuzione a posteriori
curve(dnorm(x,mu1,tau1),
      lwd=4,
      col="orange",
      add=TRUE)

 legend("topleft",
        c("a Priori","Verosimiglianza","a Posteriori"),
        lwd=c(4,2,4),
       col=c("green", "blue","orange" ), cex = 0.7)
```



Il peso delle osservazioni campionarie si è notevolmente ridotto in quanto la prior è particolarmente informativa. La distribuzione a posteriori ha una media molto più vicina a quella della distribuzioni a priori e presenta sempre minima variabilità.

Numerosità campionaria ridotta

Si considera lo scenario 1 con distribuzione a priori poco informativa e si suppone di disporre di un campione ancora più piccolo con numerosità $n = 6$.

Le stime dei parametri della distribuzione a posteriori diventano

```
n1 <- 6
mu11<- (161*12^2 + (n1*135*37^2))/(n1*37^2 + 12^2); mu11
#> [1] 135.448

tau121<- ((37^2)*12^2)/(n1*37^2+12^2);
tau121
#> [1] 23.5865
tau11<-sqrt(tau121)
```

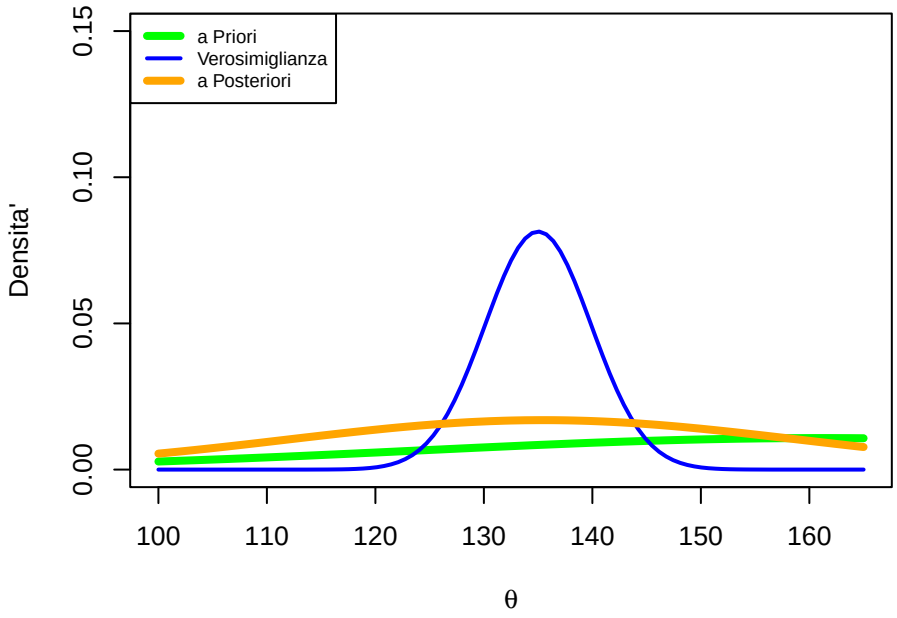
Si confrontano le figure rispetto al caso precedente di $n = 31$

```r
# distribuzione a priori
mu <- 161
tau <- 37
curve(dnorm(x,mu,tau),
      xlab = expression(theta),
      ylab="Densita'",
      xlim=c(100,165),
      ylim = c(0,0.15),
      lwd=4,
      col="green")

# verosimiglianza
sigma <- 12
ss<-sigma/sqrt(n1)
curve(dnorm(x,xbar,ss),
      lwd=2,
      col= "blue",
      add=TRUE)


# distribuzione a posteriori
curve(dnorm(x,mu11,tau121),
      lwd=4,
      col="orange",
      add=TRUE)

 legend("topleft",
        c("a Priori","Verosimiglianza","a Posteriori"),
        lwd=c(4,2,4),
       col=c("green", "blue","orange" ), cex = 0.7)
```



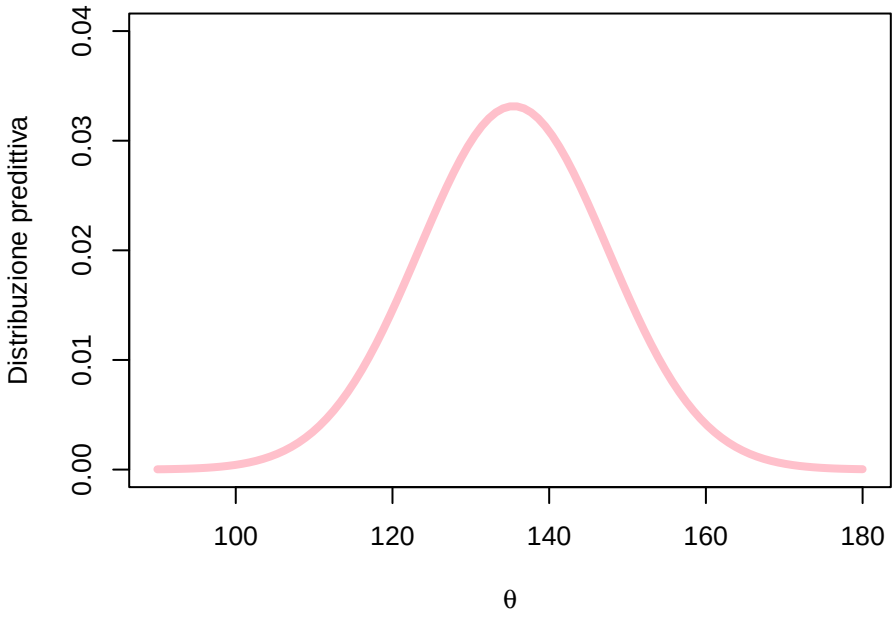
La distribuzione a posteriori come nello scenario 1 è centrata rispetto al valore medio della verosimiglianza essendo la prior non informativa ma dato che la numerosità campionaria è ridotta a parità di altre condizioni la variabilità della distribuzione a posteriori è più elevata.

In questo caso la precisione nella stima del parametro è minore.

Distribuzione predittiva

In quest'ultimo scenario considerando la *distribuzione predittiva* e sapendo che $E(\bar{x}^*|\bar{\mathbf{x}}) = \mu_1$ e $V(\bar{x}^*|\bar{\mathbf{x}}) = \sigma^2 + \tau_1^2$ la rappresentazione grafica è

```
curve(dnorm(x,mu11,sqrt((tau12 + sigma^2))),
      xlab = expression(theta),
      ylab="Distribuzione predittiva",
      xlim=c(90,180),
      ylim = c(0,0.04),
      lwd=4,
      col="pink")
```


Distribuzione predittiva
0.00
0.01
0.02
0.03
0.04
100
120
140
160
180
θ

Confronto tra distribuzioni a priori: Gauss e t di Student

Si intende ricavare la distribuzione a posteriori per parametro riferito al punteggio medio di un test utilizzato per le valutazioni di dislessia. Da rilevazioni precedenti si sa che il punteggio mediano è 100 e che un valore plausibile per il 95-esimo percentile della distribuzione a priori del parametro è 120.

Utilizzando la funzione `normal.select` si ricavano i due momenti della prior

```
require(LearnBayes)
quantile1 <- list(p=.5,x=100);
quantile2 <- list(p=.95,x=120)
ris <- normal.select(quantile1, quantile2);
mu <- ris$mu; mu
#> [1] 100
tau <- ris$sigma; tau
#> [1] 12.15914
```

Da cui $p(\theta) \sim N(100, 12.16^2)$

Supponendo di aver esaminato $n = 4$ bambini selezionati in modo casuale, ed assumendo $\sigma = 15$ (variabilità attesa nella popolazione considerata), si mostrano i momenti della distribuzione a posteriori in base al modello coniugato rispetto al modello in cui la prior è specificata come una distribuzione t di Student. In questo contesto la distribuzione t avendo code più pesanti comporta una diversa distribuzione a posteriori specialmente se i valori campionari si allontano dal valore medio della distribuzione a priori (Albert, 2009).

Supponiamo ad esempio che le rilevazioni sui 4 bambini forniscano i seguenti tre diversi scenari:

- a) $\bar{x}_1 = 110$
- b) $\bar{x}_2 = 125$
- c) $\bar{x}_3 = 140$.

Ovvero nel primo scenario la media più vicina a quella specificata a priori mentre nel terzo scenario è particolarmente lontana.

Se il modello è coniugato come illustrato in precedenza le stime dei parametri della distribuzione a posteriori per ogni scenario si ricavano come segue

```r
sigma <- 15
xnn <- c(110, 125, 140)
n<-4

# deviazione standard
tau12 <- 1/(n/sigma^2 + 1/tau^2); tau12
#> [1] 40.74708


# media
mu1 <- (mu*sigma^2 + (n*xnn*tau^2))/(n*tau^2 + sigma^2); mu1
#> [1] 107.2439 118.1098 128.9757

# sintesi

summ1 <- cbind(xnn, mu1, tau12)
summ1
#>      xnn      mu1    tau12
#> [1,] 110 107.2439 40.74708
#> [2,] 125 118.1098 40.74708
#> [3,] 140 128.9757 40.74708
```

A parità di numerosità campionaria (esigua) con una distribuzione a priori informativa la distribuzione a posteriori non è particolarmente spostata verso la verosimiglianza.

La distribuzione a priori esercita una notevole influenza sulla distribuzione a posteriori anche quando come nel caso dell'osservazione $\bar{x}_2 = 140$ il punteggio medio osservato sul campione è molto lontano dal valore medio stabilito a priori pari a 100.

Distribuzione a priori T di Student

Invece di utilizzare il modello coniugato supponiamo una distribuzione t **di Student** come prior per la media. Questa distribuzione avendo le code pesanti (ovvero c'è maggiore densità sui valori lontani dalla media rispetto alla normale) permette di tener conto di valori estremi del parametro.

Si caratterizza per i seguenti tre parametri: il valore centrale μ, la scala τ e i gradi di livertà v. Si considera una distribuzione con **2 gradi di libertà**.

La funzione di densità è la seguente dove ν sono i gradi di libertà

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}$$

Il parametro di scala τ deve essere ricavato rispetto ai valori espressi a priori ($\mu = 100$ e 95esimo percentile 120). Il parametro di scala si ottiene nel modo seguente:

$$P(\theta < 120) = 0.95$$

$$P\left(\frac{\theta - \mu}{\tau} < \frac{120 - 100}{\tau}\right) = 0.95.$$

Considerando una t di Student (si ricorda che se ν=1 la t ha la forma di una Chauchy) si ha

$$P\left(t_2 < \frac{20}{\tau}\right) = 0.95$$

si tratta di determinare il quantile di ordine p della distribuzione T_2 ($t_2(p)$)

```
qt(0.95,2)
#> [1] 2.919986
```

ed il valore del parametro di scala che si ottiene dall'equazione

$$\tau = \frac{20}{t_2(p)}$$

da cui risulta

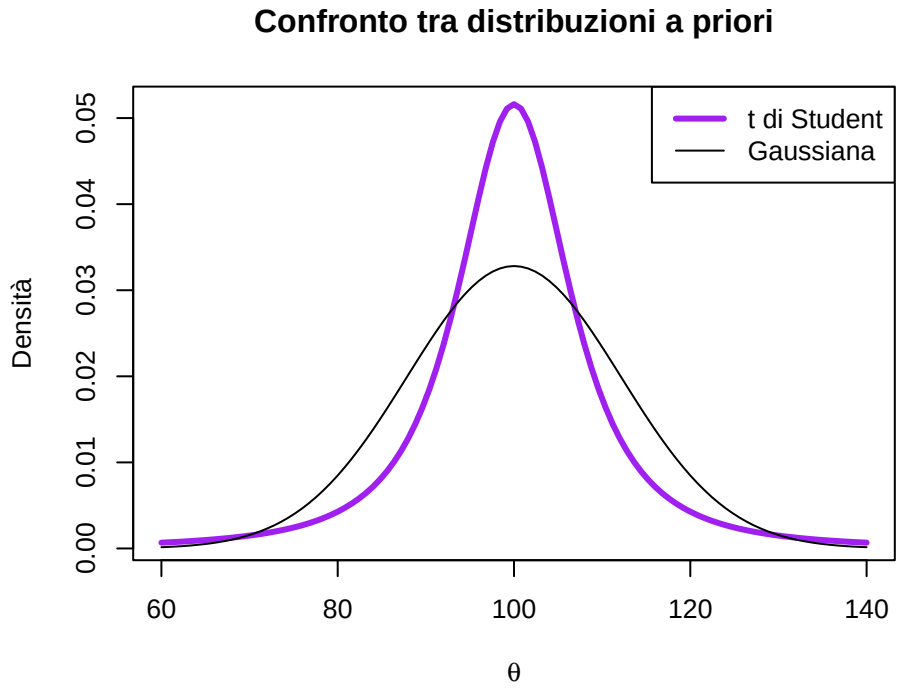
```
taut <- 20/qt(0.95,2); taut
#> [1] 6.849349
```

Confronto tra le due distribuzioni a priori

La seguente figura mette in evidenza le differenze tra le due specificazioni della prior. La distribuzione di Gauss (μ =100 e $\tau = 15$) e la distribuzine t di Student con 2 gradi di libertà

```
curve(1/taut*dt((x-mu)/taut,2),
      from=60,
      to=140,
      xlab = expression(theta),
      ylab ="Densità",
      main ="Confronto tra distribuzioni a priori",
      col ='purple',
      lwd=3)

curve(dnorm(x,mean=mu,sd=tau),
       add=TRUE,
       lwd=1)
 legend("topright",
        legend=c("t di Student","Gaussiana"),
        lwd=c(3,1),
        col=c("purple", "black"))
```



Si vede che la distribuzione ha curtosi maggiore di 0 pertanto di dice leptocurtica ovvero più appuntita di una normale. Ha inoltre code più pesanti della normale.

Determinazione della distribuzione a posteriori

La distribuzione a posterori è proporzionale al prodotto della verosimiglianza del modello e della distribuzione a priori

$$p(\theta|\mathbf{x}) \propto \phi(\bar{\mathbf{x}}|\theta, \sigma/\sqrt{n}) g_t(\theta|v, \mu, \tau).$$

La densità a posteriori si ottiene **approssimando** la distribuzione continua con una sua discretizzazione su una griglia di valori.

Si genera una sequenza di valori (griglia di valori possibili per il parametro) e si approssima la densità continua con una distribuzione discreta su questa griglia.

```
theta <-  seq(60, 180, length = 500)
summary(theta)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>      60      90     120     120     150     180
```

Si calcolano i valori della verosimiglianza riferiti al valore medio considerando $n = 4$

```
n <- 4
like <- dnorm(theta,mean=xnn,sd=sigma/sqrt(n))
summary(like)
#>      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
#> 0.000e+00 0.000e+00 1.687e-05 8.317e-03 7.223e-03 5.318e-02
```

Si determinano i valori del parametro di posizione e di scala della distribuzione a posteriori dopo averla normalizzata

```
prior <- dt((theta - mu)/taut, 2)
 post <-  prior * like
 post <- post/sum(post)
 posizione <- sum(theta * post)
 scala <- sqrt(sum(theta^2 * post) - posizione^2)
```

Il procedimento precedente può essere riassunto con la seguente funzione che in output restituisce i valori attesi (locazione e scala) della distribuzione a posteriori ed in input richiede il vettore delle medie.

```
norm.t.compute <- function(xnn){
                  theta <- seq(60, 180, length = 500)
                  like <- dnorm(theta,mean=xnn,sd=sigma/sqrt(n))
                  prior <- dt((theta - mu)/taut, 2)
                  post <- prior * like
                  post <- post/sum(post)
                  posizione <- sum(theta * post)
                  scala <- sqrt(sum(theta^2 * post) - posizione^2)
    c(xnn, posizione, scala)
}
```

Utilizzando la funzione `sapply` si applica la funzione al vettore delle medie xnn osservate e si ottengono i parametri della distribuzione a posteriori.

```
summ2<- t(sapply(c(110, 125, 140),norm.t.compute))


dimnames(summ2)[[2]] =
         c("xnn","mu1 t","tau1 t")
summ2
#>      xnn    mu1 t   tau1 t
#> [1,] 110 105.2921 5.841676
#> [2,] 125 118.0841 7.885174
#> [3,] 140 135.4134 7.973498
cbind(summ1,summ2)
#>      xnn      mu1    tau12 xnn    mu1 t   tau1 t
#> [1,] 110 107.2439 40.74708 110 105.2921 5.841676
#> [2,] 125 118.1098 40.74708 125 118.0841 7.885174
#> [3,] 140 128.9757 40.74708 140 135.4134 7.973498
```

L'inferenza sulla media (diversamente dalla mediana) è molto sensibile a cambiamenti rispetto ai valori campionari.

La scelta della distribuzione a priori Normale o t di Student comporta circa la stessa stima puntuale del parametro a posteriori per valori osservati della media campionaria non particolarmente estremi rispetto a quelli medi a priori.

Per valori campionari estremi la stima fornita dal modello coniugato è 129 mentre quella fornita dalla scelta della t di Student come prior è 135 e assicura maggiore aderenza alle osservazioni campionarie.

Modello coniugato Poisson-Gamma

Si intende stimare il tasso di occorrenza nell'unità di tempo (2 mesi) del numero di decessi avvenuti in una regione a seguito di una certa operazione chirurgica.

Sapendo che nei due mesi precedenti sono avvenuti 16 decessi in 10 ospedali che hanno operato in totale 15174 pazienti. La distribuzione a priori è

$$p(\lambda) \sim Gamma(\alpha, \beta).$$

con $\alpha = 16$ e $\beta = 15174$.

Rilevando oggi che in un altro ospedale della regione c'è stato un decesso a seguito di 66 operazioni su pazienti diversi, con il modello coniugato la distribuzione a posteriori è

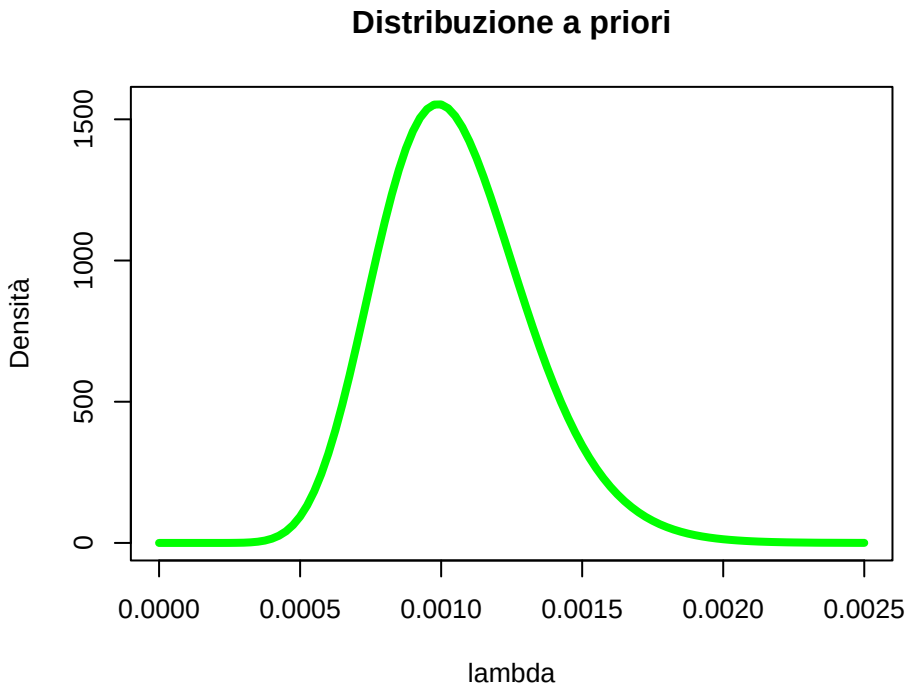
$$p(\lambda|\mathbf{y}) \sim Gamma(\alpha + \sum_i x_i, \beta + n)$$

i cui parametri $\alpha_1 = 17$ e $\beta_1 = 15174$.

Rappresentazioni grafiche

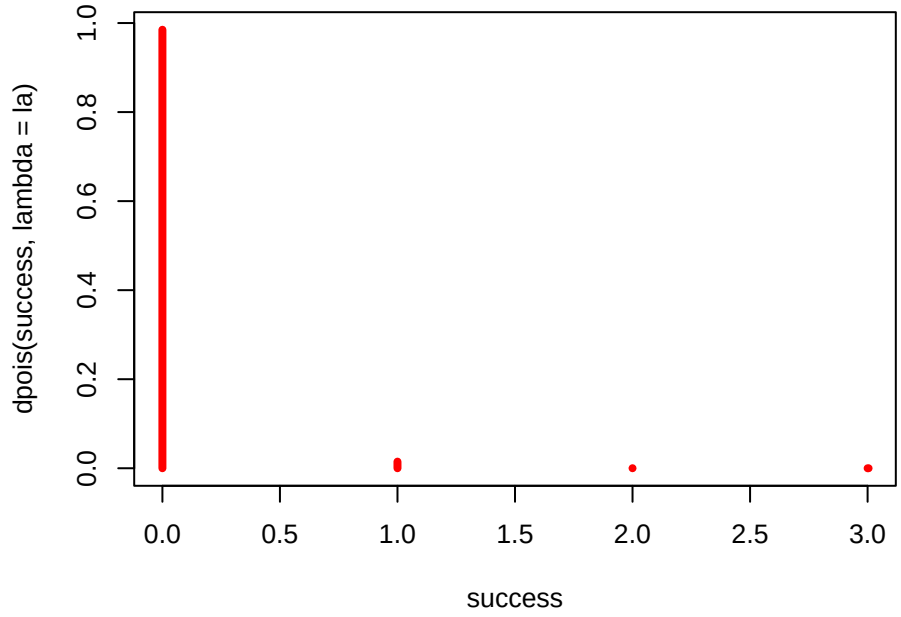
```
alpha <- 16
beta <- 15174

curve(dgamma(x,alpha,beta),
      xlab="lambda",
      ylab="Densità",
      type = "l",
      xlim=c(0,0.0025),
      lwd=4,
      col="green",
     main ="Distribuzione a priori")
```



Distribuzione di Poisson

```
success <- 0:3
la <- 1/66
plot(success, dpois(success, lambda=la), type = "h",
     lwd=4,
     col = "red")
```



Rappresentazione della distribuzione a posteriori

In base alle osservazioni campionarie la distribuzione a posteriori per il parametro è la seguente

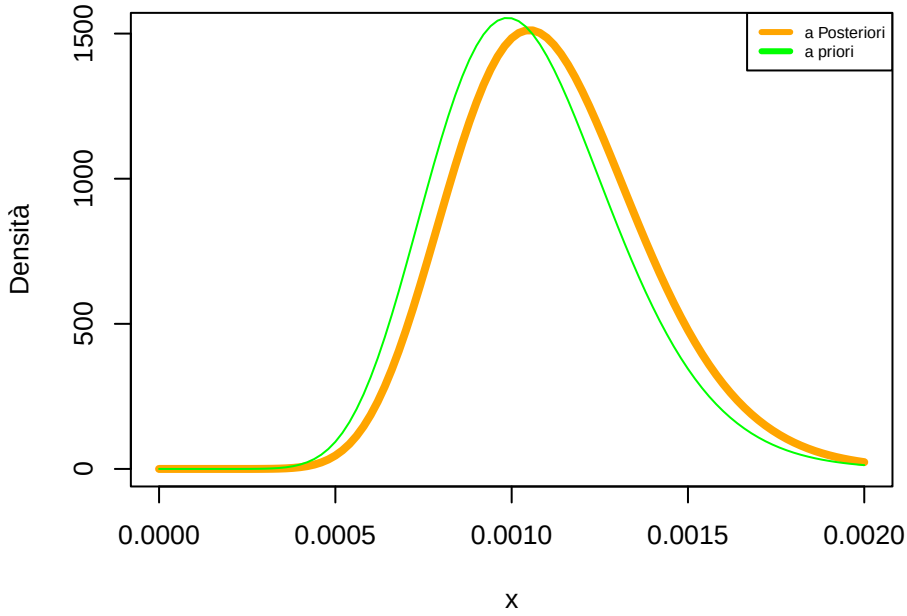
```
n<-66
alpha1<- alpha + 1; alpha1
#> [1] 17
beta1 <- beta + n; beta1
#> [1] 15240

curve(dgamma(x,alpha1,beta1),
      lty=1,
      lwd=4,
       ylab="Densità",
      xlim=c(0,0.002),
      col="orange",
      main = " ")

# prior

curve(dgamma(x, alpha, beta),
lty=1,lwd=1,
col="green",
add=TRUE)

legend("topright",
c("a Posteriori", "a priori"),
lwd=c(3,3),
cex = 0.6,
col=c("orange","green" ))
```



Il valore atteso in base alla distribuzione a posteriori per il tasso dei decessi mensili è

```
Elambda <- (alpha+1)/(beta+n); Elambda
#> [1] 0.001115486
```

con variabilità media intorno alla media pari a

```
sqrt(alpha1/((beta1)^2))
#> [1] 0.000270545
```

Distribuzione predittiva

Supponendo di voler conoscere la probabilità di 0,1,...,10 decessi prevista in base al modello Poisson-Gamma a fronte di 1767 pazienti operati si considera la distribuzione predittiva.

La **densità prevista** permette anche di validare il modello assunto controllando che i valori previsti siano coerenti con quelli possibili nel contesto di riferimento.

Dalla teoria la distribuzione predittiva è una Binomiale Negativa. Un'approssimazione è fornita nel seguente chunch in cui si calcola

$$f(x_{n+1}) = \frac{p(x_{n+1}|\lambda)p(\lambda)}{p(\lambda|\mathbf{x}, x_{n+1})}.$$

```r
ex <- 1767
ys <- 0:10

(alpha/beta)*ex
#> [1] 1.863187

#
pyn1 <- dpois(ys, (alpha/beta)*ex)*
         dgamma(alpha/beta, shape = alpha, rate = beta)/
   dgamma(alpha1/beta1, shape = alpha1 + ys,
rate = beta1 + ex)

cbind(ys, round(pyn1, 3))
#>       ys
#>  [1,]  0 0.177
#>  [2,]  1 0.296
#>  [3,]  2 0.261
#>  [4,]  3 0.163
#>  [5,]  4 0.080
#>  [6,]  5 0.033
#>  [7,]  6 0.012
#>  [8,]  7 0.004
#>  [9,]  8 0.001
#> [10,]  9 0.000
#> [11,] 10 0.000
```

Tra gli eventi previsti quello con massima probabilità è 1 decesso, seguito da 2 decessi e 0 decessi.

Essendo plausibili tutti i valori da 0 a 10 speficicati notiamo che il modello Bayesiano è stato correttamente specificato.