Inferencia Bayesiana

Marco Galliani

Contents

Settings

```r
rm(list = ls())
bomb_data <- read.csv("../data/geocoded_bomb_data.csv")

bomb_data$Time <- as.POSIXct(bomb_data$Time)

bomb_data$district <- as.factor(bomb_data$district)
bomb_data$Type.of.bomb <- as.factor(bomb_data$Type.of.bomb)

knitr::kable(head(bomb_data))
```

X	Time	Location	Type.of.bomb	Damage.or.other	lat	lon	district
1	1940-09-07 00:08:00	43 Southwark Park Road, SE16, London, UK	IB	Grocers: 3x2 roof damaged	51.49225	-0.0621761	Southwark
2	1940-09-07 00:10:00	49 Southwark Park road, Bermondsey, SE16, London, UK	IB	Bakers: 3x2 roof damaged	51.49269	-0.0653908	Southwark
3	1940-09-07 00:15:00	84 Southwark Park Road, SE16, London, UK	IB	front room on 1st floor and contents slightly damaged. 3x2 rood damage	51.49225	-0.0621761	Southwark
4	1940-09-07 00:18:00	141 Braidwood Road, Catford SE6, London, UK	IB	10x6 roof damage	51.44085	-0.0053336	Lewisham
5	1940-09-07 00:20:00	129 Killearn Road, Catford SE6, London, UK	IB	Front room on 1st floor severely damaged	51.44151	-0.0054617	Lewisham
6	1940-09-07 00:20:00	27 Crutchley Road, Downham, London, UK	IB	IB on enclosed ground at rear of premises	51.43670	0.0052611	Lewisham

```r
summary(bomb_data)
```

```
##        X                Time                              Location
##  Min.   :  1.0   Min.   :1940-09-07 00:08:00.000   Length:843
##  1st Qu.:211.5   1st Qu.:1940-09-07 18:00:00.000   Class :character
##  Median :422.0   Median :1940-09-07 18:28:00.000   Mode  :character
##  Mean   :422.0   Mean   :1940-09-07 18:08:30.253
##  3rd Qu.:632.5   3rd Qu.:1940-09-07 20:56:00.000
##  Max.   :843.0   Max.   :1940-09-07 23:59:00.000
##
##                Type.of.bomb Damage.or.other         lat
##  IB                 :478   Length:843         Min.   :50.72
##  EB                 :220   Class :character   1st Qu.:51.47
##  EB.and.IB          : 59   Mode  :character   Median :51.49
##  COB                : 10                      Mean   :51.50
##  Unknow.enemy.action:  6                      3rd Qu.:51.51
##  (Other)            :  3                      Max.   :55.16
##  NA's               : 67                      NA's   :8
##       lon                     district
##  Min.   :-3.52493   Tower Hamlets :192
##  1st Qu.:-0.08988   Lewisham      :139
##  Median :-0.04334   Southwark     :102
##  Mean   :-0.06730   City of London: 86
##  3rd Qu.:-0.01223   Greenwich     : 78
##  Max.   : 0.14640   (Other)       :218
##  NA's   :8          NA's          : 28
```

Modelacion

Variable aleatoria

Muestra antes de la 6:34 a.m.

```r
sample_EB <- ifelse(bomb_data$Type.of.bomb == "EB" | bomb_data$Type.of.bomb == "EB.and.IB", 1, 0)

sunrise_hour <- as.POSIXct("1940-9-7 06:34:00")

sample_before_sunrise <-
  sample_EB[bomb_data$Time <= sunrise_hour]

sample_before_sunrise <- na.omit(sample_before_sunrise)
n_before_sunrise <- length(sample_before_sunrise)
```

Detalle: Incluimos el caso "EB.and.IB" en el caso general de bombas explosivas, esto puede cambiar mucho los resultados, sobre todo en ultimo apartado

Definimos la siguiente variable aleatoria:

$$X = \begin{cases} 1 & \text{si BOMBA EXPLOSIVA(EB)} \\ 0 & \text{en otro caso} \end{cases}$$

Calculo de la aposteriori

Funcion de verosimilitud

La variable X es distribuida como una Bernoulli, tomando como muestra $X_1, \ldots, X_n$, podemos definir la funcion de verosimilitud

$$L(p|\mathbb{X}) = \prod_{i=1}^{n} f(x_i|p) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

Calculo de la apriori

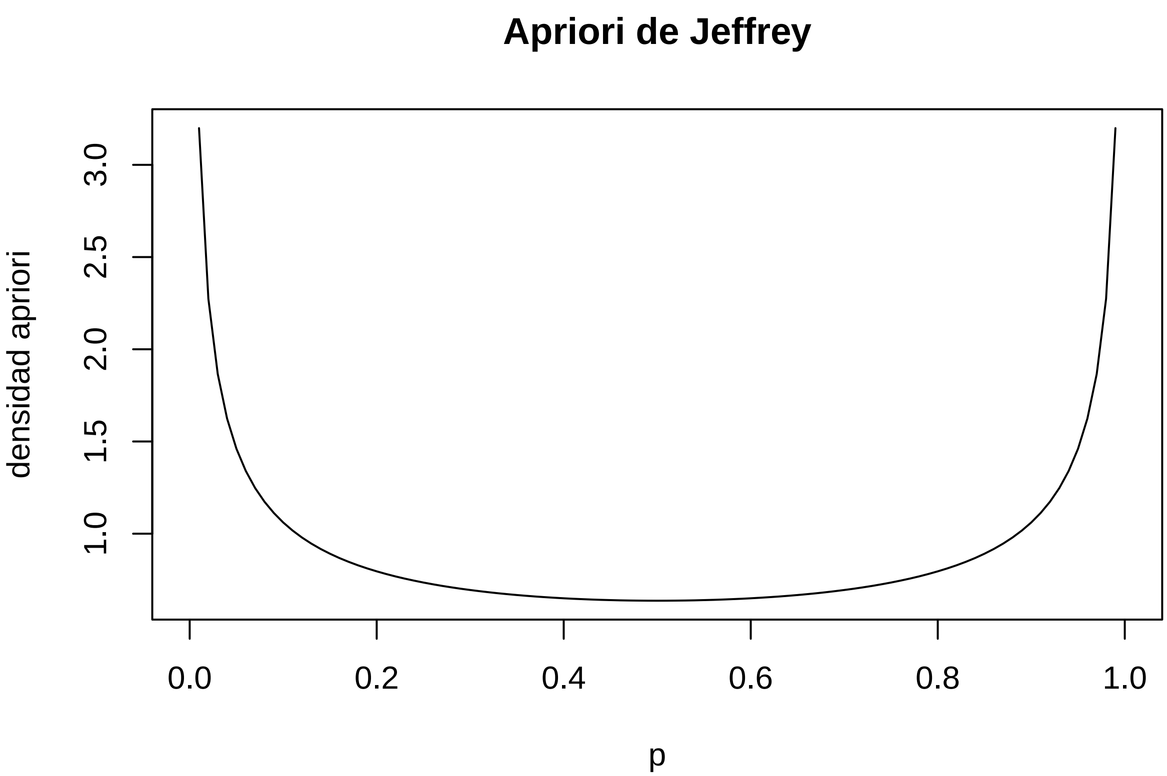
La apriori de Jeffrey se calcula como:

$$\pi(p) = \frac{1}{\sqrt{p(1-p)}}, \quad p \in (0,1)$$

Es decir que

$$p \sim Beta(\frac{1}{2}, \frac{1}{2})$$

```
curve(dbeta(x, shape1 = 1/2, shape2 = 1/2), xlab = "p", ylab = "densidad apriori")
title("Apriori de Jeffrey")
```



Calculo de la aposteriori

Usando la apriori y la verosimilitud se puede calcular la distribucion aposteriori de p

$$\begin{aligned}\pi(p|\mathbb{X}) &\propto L(p|\mathbb{X})\pi(p) \\ &\propto p^{\sum_{i=1}^{n} x_i - 1/2}(1-p)^{n-\sum_{i=1}^{n} x_i - 1/2}\end{aligned}$$

Es decir que

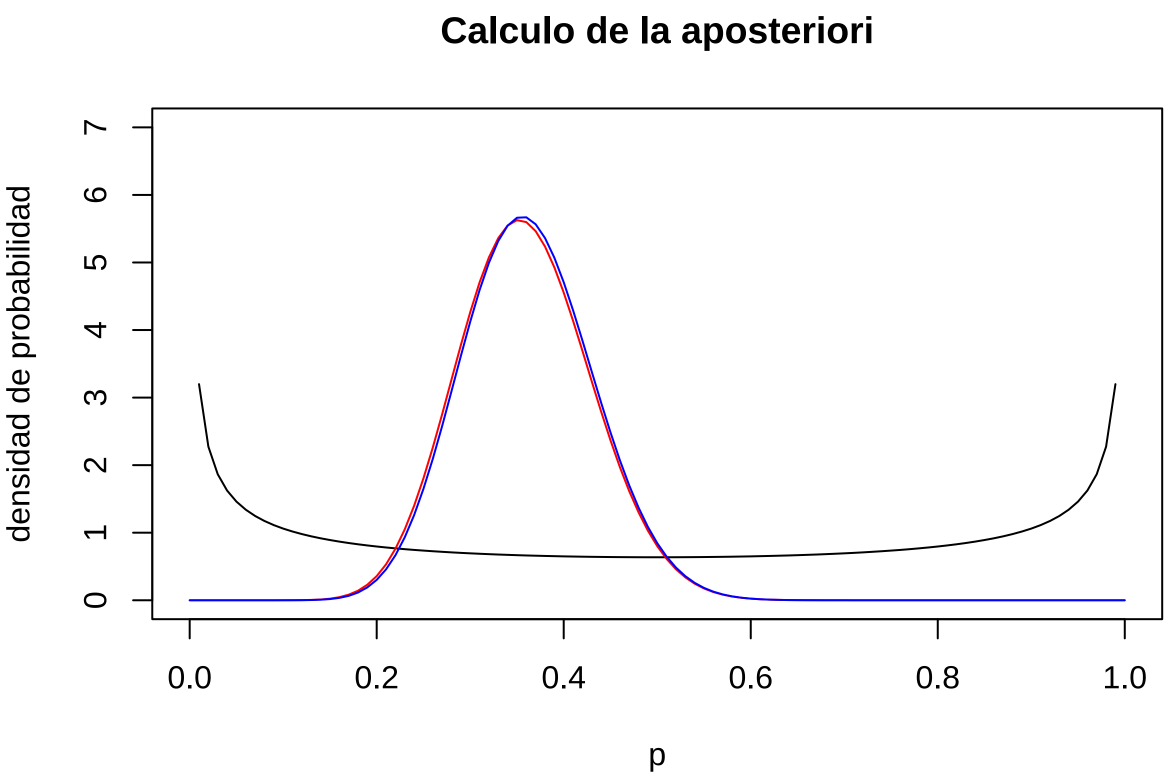
$$p|\mathbb{X} \sim Beta(\sum_{i=1}^{n} x_i + 1/2, \quad n - \sum_{i=1}^{n} x_i + 1/2)$$

```
# apriori
curve(dbeta(x, shape1 = 1/2, shape2 = 1/2),
      xlab = "p", ylab = "densidad de probabilidad",
      xlim = c(0,1), ylim = c(0,7))

# aposteriori
curve(dbeta(x,
            shape1 = sum(sample_before_sunrise) + 1/2,
            shape2 = n_before_sunrise - sum(sample_before_sunrise) + 1/2),
      add = T, col = "red")

# funcion de verosimilitud
curve(dbeta(x,
            shape1 = sum(sample_before_sunrise) + 1,
            shape2 = n_before_sunrise - sum(sample_before_sunrise) + 1),
      add = T, col = "blue")

title("Calculo de la aposteriori")
```



Predicciones en la noche siguiente

Muestra despues del anochecimiento

```
sunset_hour <- as.POSIXct("1940-9-7 19:16:00")

sample_after_sunset <-
```

```r
  sample_EB[bomb_data$Time >= sunset_hour]

sample_after_sunset <- na.omit(sample_after_sunset)
n_after_sunset <- length(sample_after_sunset)
```

Estimador de Bayes

Media de la distribucion aposteriori (considerando funcion de perdida cuadratica)

```r
estimador_bayes <-
  (sum(sample_before_sunrise) + 1/2) /
  (sum(sample_before_sunrise) + 1/2 + n_before_sunrise - sum(sample_before_sunrise) + 1/2)

estimador_bayes
```

```
## [1] 0.3586957
```

Intervalo de Credibilidad

```r
library(HDInterval)

cred_interval <- hdi(
  qbeta(c(0.05, 0.95),
  shape1 = sum(sample_before_sunrise) + 1/2,
  shape2 = n_before_sunrise - sum(sample_before_sunrise) + 1/2))

cred_interval
```

```
##     lower     upper
## 0.2470991 0.4773550
## attr(,"credMass")
## [1] 0.95
```

Valor observado despues del anochecimiento

```r
val_obs <- sum(sample_after_sunset)/n_after_sunset
val_obs
```

```
## [1] 0.3478261
```

Vemos che el valor predicido por el estamador de Bayes es muy cercano al valor observado y es entre el intervalo de confianza

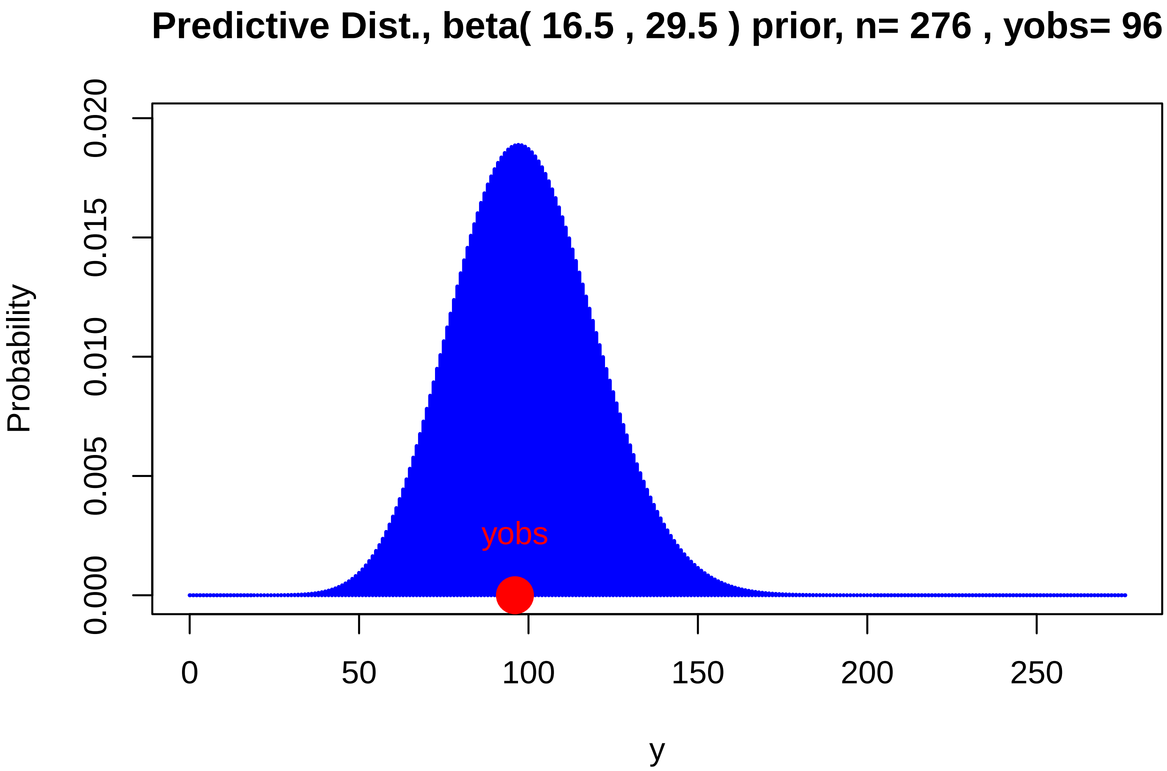
Densidad predictiva

```r
library(LearnBayes)

prior_beta_params <- c(sum(sample_before_sunrise) + 1/2,
                       n_before_sunrise - sum(sample_before_sunrise) + 1/2)

# densidad predictiva
predictions <- pbetap(prior_beta_params, n_after_sunset, 0:n_after_sunset)

predplot(prior_beta_params, n_after_sunset, sum(sample_after_sunset))
```



La distribucion confirma la bondad de la prediccion.

Resumiendo: (porcentaje de bombas EB sobre el numero total de bombas)

```
knitr::kable(
  data.frame(
    prediction = estimador_bayes,
    cred_int_left = cred_interval[1],
    cred_int_right = cred_interval[2],
    observed_value = val_obs
  )
)
```

	prediction	cred_int_left	cred_int_right	observed_value
lower	0.3586957	0.2470991	0.477355	0.3478261

Constraste de hipotesis sobre el tipo de bombas en Greenwich

Como antes, usamos una apriori de Jeffrey, calculamos la aposteriori usando todos los datos del dia (primero dia del bombardeo)

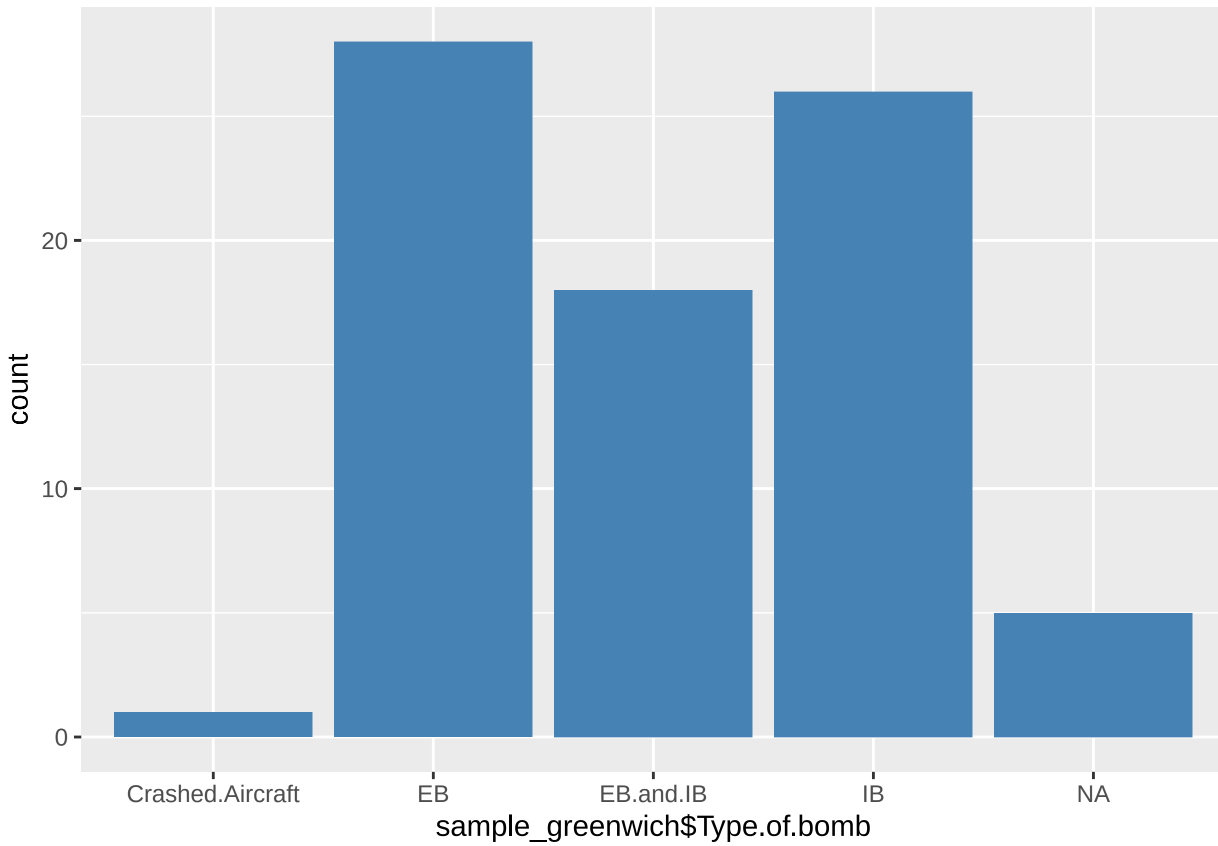
Grafico descriptivo

```
#descriptive
ind_Greenwich <- bomb_data$district == "Greenwich"
ind_Greenwich <- replace(ind_Greenwich, is.na(ind_Greenwich), FALSE)

sample_greenwich <- bomb_data[ind_Greenwich,]

library(ggplot2)
```

```r
ggplot(mapping = aes(x = sample_greenwich$Type.of.bomb)) +
  geom_bar(fill = "steelblue")
```



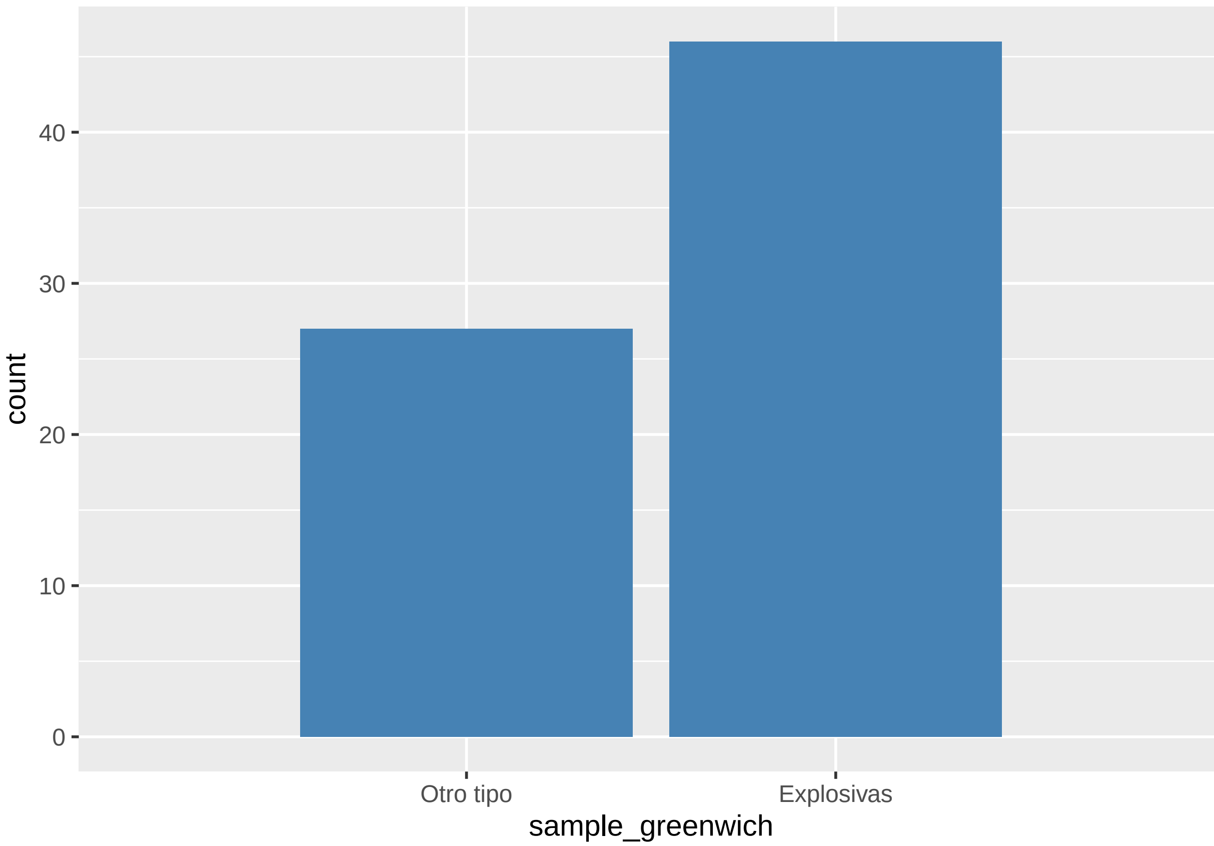
Desde el grafico podemos ver que en este caso la decision de considerar "EB.and.IB" como "EB" podria tener mucha influencia

Quitamos los NAs y consideramos como eventos "EB" tambien los eventos "EB.and.IB"

```r
sample_greenwich <- ifelse(sample_greenwich$Type.of.bomb == "EB" | sample_greenwich$Type.of.bomb == "EB
sample_greenwich <- na.omit(sample_greenwich)


library(ggplot2)
ggplot(mapping = aes(x = sample_greenwich)) +
  geom_bar(fill = "steelblue") +
  scale_x_discrete(limit = c(0,1), labels = c("Otro tipo", "Explosivas"))
```

```
## Warning: Continuous limits supplied to discrete scale.
## i Did you mean `limits = factor(...)` or `scale_*_continuous()`?
```



Calculo de la aposteriori

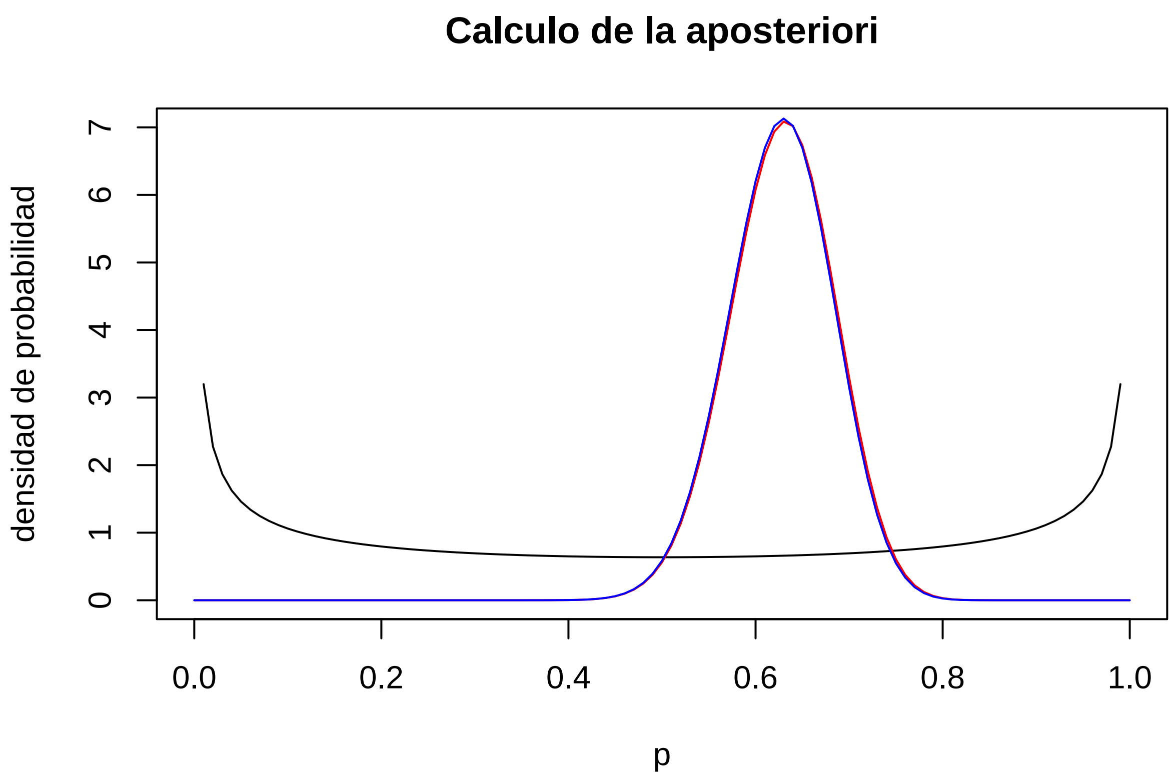
```
n_greenwich <- length(sample_greenwich)

# apriori
curve(dbeta(x, shape1 = 1/2, shape2 = 1/2),
      xlab = "p", ylab = "densidad de probabilidad",
      xlim = c(0,1), ylim = c(0,7))

# aposteriori
curve(dbeta(x,
            shape1 = sum(sample_greenwich) + 1/2,
            shape2 = n_greenwich - sum(sample_greenwich) + 1/2),
      add = T, col = "red")

# funcion de verosimilitud
curve(dbeta(x,
            shape1 = sum(sample_greenwich) + 1,
            shape2 = n_greenwich - sum(sample_greenwich) + 1),
      add = T, col = "blue")

title("Calculo de la aposteriori")
```



Contraste de hipotesis

$$H_0 : p > 0.5$$
$$H_1 : p \leq 0.5$$

Calculo del Factor de Bayes

```r
# posteriors
posterior_H1 <- pbeta(0.5, sum(sample_greenwich) + 1/2, n_greenwich - sum(sample_greenwich) + 1/2)
posterior_H0 <- 1 - pbeta(0.5, sum(sample_greenwich) + 1/2, n_greenwich - sum(sample_greenwich) + 1/2)

# priors
prior_H1 <- pbeta(0.5, 1/2, 1/2)
prior_H0 <- 1 - pbeta(0.5, 1/2, 1/2)

# Factor de Bayes
FB01 <- (posterior_H0/posterior_H1)*(prior_H1/prior_H0)
FB01
```

```
## [1] 77.48713
```

Evidencia muy fuerte para H0, es decir tenemos evidencia fuerte para decir que la mayoria de bombas que cayeron en Greenwich eran explosivas (EB).