Ampliación de Inferencia Estadística

Tercero Grado de Estadística

Universidad de Sevilla

Tema 5: Inferencia Bayesiana en el modelo Normal

Práctica-R-5. Primera parte.

Problema 1.

En la práctica se tiene una información incompleta acerca de la distribución a priori de unos parámetros en el sentido de que nuestras creencias no determinan de manera única la distribución a priori. Es decir, puede que existan más de una distribución a priori que cumplan nuestras restricciones. Por ejemplo, si se cree que la a priori tiene una mediana de 30 y el percentil 80 vale 50, tendremos muchas distribuciones que cumplan con ambos requisitos. En este caso, donde diferentes a priori son posibles, es deseable que las inferencias realizadas a partir de la a posteriori no dependan de la forma funcional de la a priori. Un análisis Bayesiano se dice que es robusto si la selección de la a prior no influye de manera notoria en las conclusiones finales.

Para ilustrar esta idea vamos a suponer que estamos interesados en estimar el coeficiente intelectual IQ, denotado por θ, de una persona. Se piensa que tiene una media
que coincide con la mediana de valor 100. También se sabe que una región de credibilidad al 90% es (80,120). Esta información nos permite construir una distribución normal de media $\mu = 100$ y $\sigma = 12.16$.

Ahora sometemos a esa persona a cuatro test de niveles de IQ dando como marcadores los valores y_1, y_2, y_3, y_4. Supongamos también que un marcador y se distribuye según una normal de media $\mu = \theta$ y desviación típica $sd = 15$. Por tanto, la puntuación media de los cuatro test seguirá una distribución normal de media $\mu = \theta$, y $sd = 15/2$.

Con esta información, y por lo visto en teoría, sabemos que la distribución a posteriori de θ sigue una normal cuyos parámetros son

$$\mu_1 = \frac{(4\bar{y}/\sigma^2 + \mu/\tau^2)}{(4/\sigma^2 + 1/\tau^2)}, \text{ y } \tau_1^2 = \frac{1}{(4/\sigma^2 + 1/\tau^2)},$$

donde los parámetros a priori son μ y σ^2, y la varianza conocida es τ^2.

Vamos a ilustrar los cálculos a posteriori para tres resultados hipotéticos, por ejemplo $\bar{y}_1 = 110$, $\bar{y}_2 = 125$, $\bar{y}_3 = 140$. En cada caso calcularemos la media a posteriori y la desviación típica a posteriori de theta.

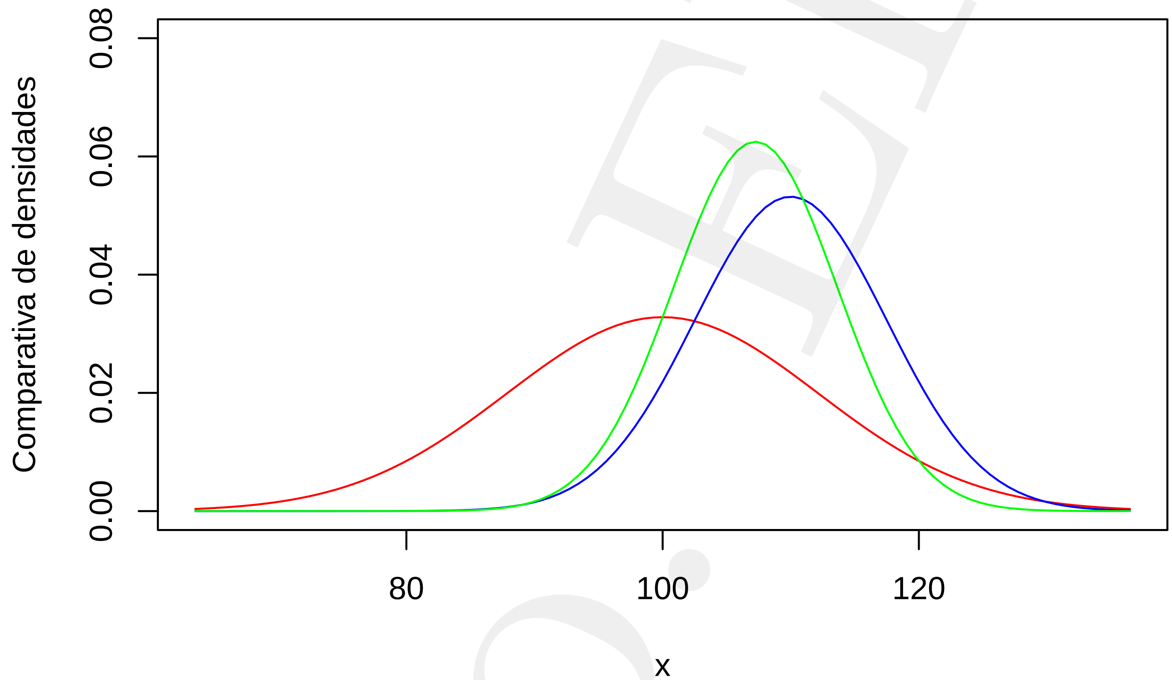
```r
mu=100
tau=12.16
sigma=15
n=4
se=sigma/sqrt(4)
ybar=c(110,125,140)
tau1=1/sqrt(1/se^2+1/tau^2)
mu1=(ybar/se^2+mu/tau^2)*tau1^2
summ1=cbind(ybar,mu1,tau1)
summ1
```

```
##      ybar      mu1     tau1
## [1,]  110 107.2442 6.383469
## [2,]  125 118.1105 6.383469
## [3,]  140 128.9768 6.383469
```

Veamos el efecto en gráficas

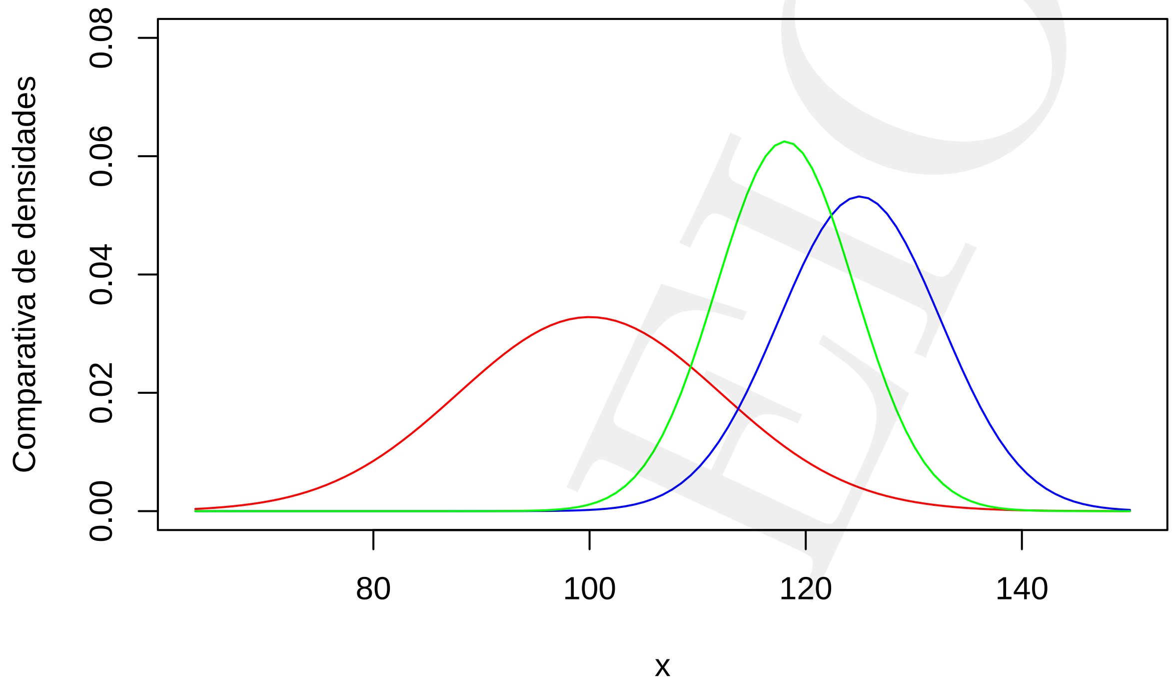
Caso 1

```
curve(dnorm(x,mean=mu,sd=tau),xlim=c(mu-3*tau,mu+3*tau),ylim=c(0,0.08),col="red",
      ylab="Comparativa de densidades")
curve(dnorm(x,mean=ybar[1],sd=se),xlim=c(mu-3*tau,mu+3*tau),ylim=c(0,0.05),col="blue",add=T)
curve(dnorm(x,mean=mu1[1],sd=tau1),xlim=c(mu-3*tau,mu+3*tau),ylim=c(0,0.05),col="green",add=T)
```



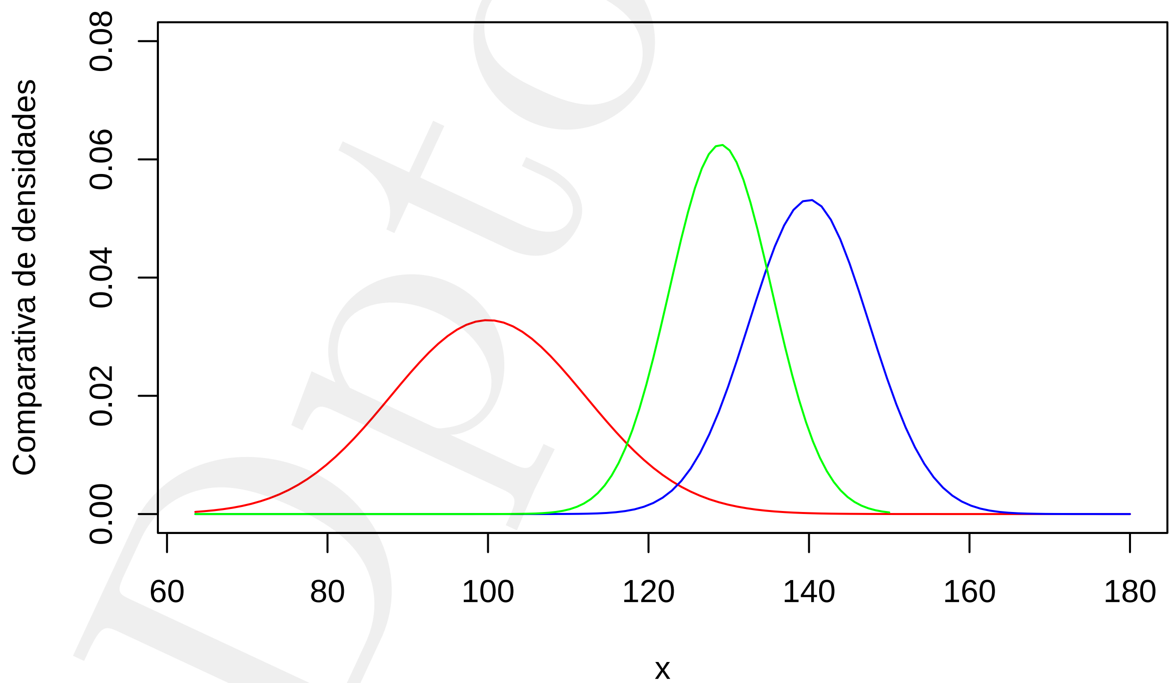
Caso 2

```
curve(dnorm(x,mean=mu,sd=tau),xlim=c(mu-3*tau,150),ylim=c(0,0.08),col="red",
      ylab="Comparativa de densidades")
curve(dnorm(x,mean=ybar[2],sd=se),xlim=c(mu-3*tau,150),ylim=c(0,0.05),col="blue",add=T)
curve(dnorm(x,mean=mu1[2],sd=tau1),xlim=c(mu-3*tau,150),ylim=c(0,0.05),col="green",add=T)
```



Caso 3

```
curve(dnorm(x,mean=mu,sd=tau),xlim=c(mu-3*tau,180),ylim=c(0,0.08),col="red",
      ylab="Comparativa de densidades")
curve(dnorm(x,mean=ybar[3],sd=se),xlim=c(mu-3*tau,180),ylim=c(0,0.05),col="blue",add=T)
curve(dnorm(x,mean=mu1[3],sd=tau1),xlim=c(mu-3*tau,150),ylim=c(0,0.05),col="green",add=T)
```



Vamos a considerar ahora una a priori alternativa para modelar nuestras creencias sobre θ. Podemos usar cualquier densidad simétrica que no sea la normal, como por ejemplo la t-student con parámetro de localización μ, escala *tscale* (denotado por τ) y dos grados de libertad. Ya que la mediana a priori es 100 vamos a poner en este caso $\mu = 100$. Para calcular el valor de τ vamos a usar que el percentil 95 de la t-student es de 120.

Fijémonos que

$$P(\theta < 120) = P(T < 20/\tau) = 0.95,$$

donde T es una t-student con dos grados de libertad. Por tanto, se obtiene que

$$tscale = 20/t(0.95, 2)$$

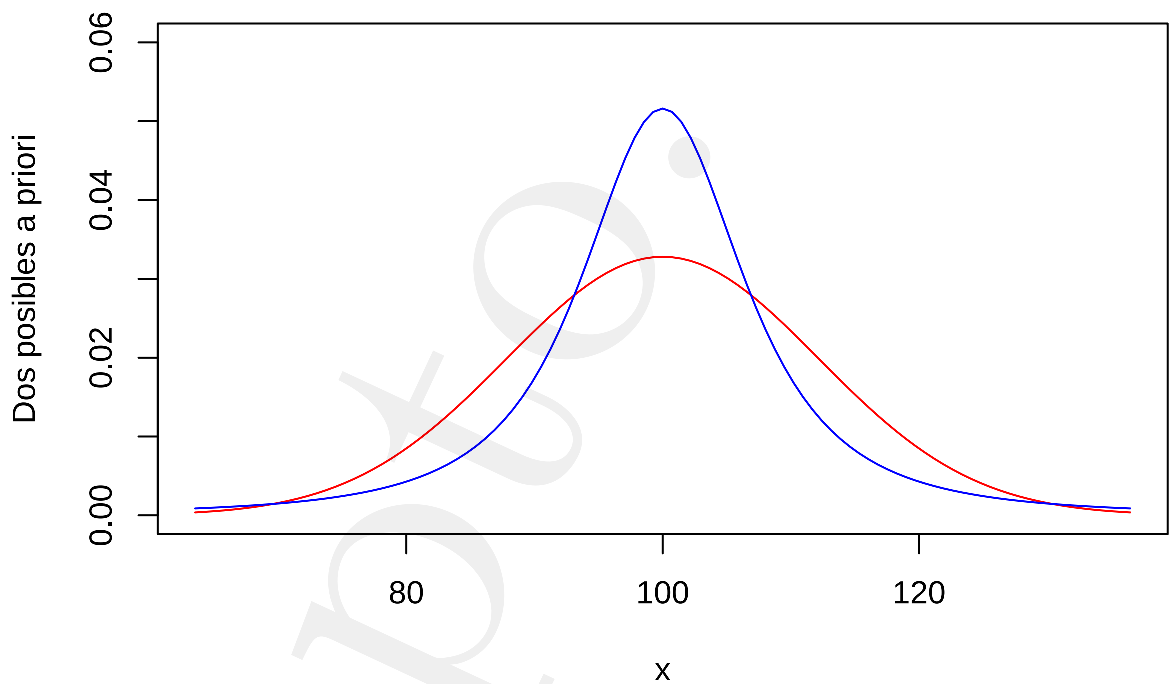
donde $t(p, n)$ es el percentil p de una t-student con n g.l. Recordemos que la función cuantil en R de la t-student es el comando `qt`.

```
tscale=20/qt(0.95,2)
tscale
```

```
## [1] 6.849349
```

Representemos ahora en un mismo gráfico las dos posibles a priori

```
curve(dnorm(x,mean=mu,sd=tau),xlim=c(mu-3*tau,mu+3*tau),ylim=c(0,0.06),col="red",
      ylab="Dos posibles a priori")
curve(1/tscale*dt((x-mu)/tscale,df=2),xlim=c(mu-3*tau,mu+3*tau),ylim=c(0,0.08),col="blue",add=T)
```



Ahora vamos a hacer los cálculos de la a posteriori usando una a priori t-Student. Fijémonos que en este caso la a posteriori para theta sería

$$g(\theta|datos) \propto f(\bar{y}|\theta, \sigma/\sqrt{(}n)) * T(\theta|v, \mu, \tau)$$

donde $f()$ es la densidad normal y $T()$ es la t-student con v g.l., centrada en μ y parámetro de escala τ.

Obviamente al tratar con t-student, la constante no se puede obtener a partir de una forma funciónal explícita luego tendremos que usar un método de aproximación.

```
summ2=c()
for (i in 1:3){
  theta=seq(60,180,length=500)
  like= dnorm((theta-ybar[i])/7.5)
  prior=dt((theta-mu)/tscale,2)
  post=like*prior
  post=post/sum(post)
  m= sum(theta*post)
  s=sqrt(sum(theta^2*post)-m^2)
  summ2=rbind(summ2,c(ybar[i],m,s))
  }
summ2
```

```
##      [,1]     [,2]     [,3]
## [1,]  110 105.2921 5.841676
## [2,]  125 118.0841 7.885174
## [3,]  140 135.4134 7.973498
```

Ahora vamos a comparar las dos a posterioris, normal frente a *t*-student.

```
cbind(summ1,summ2)
```

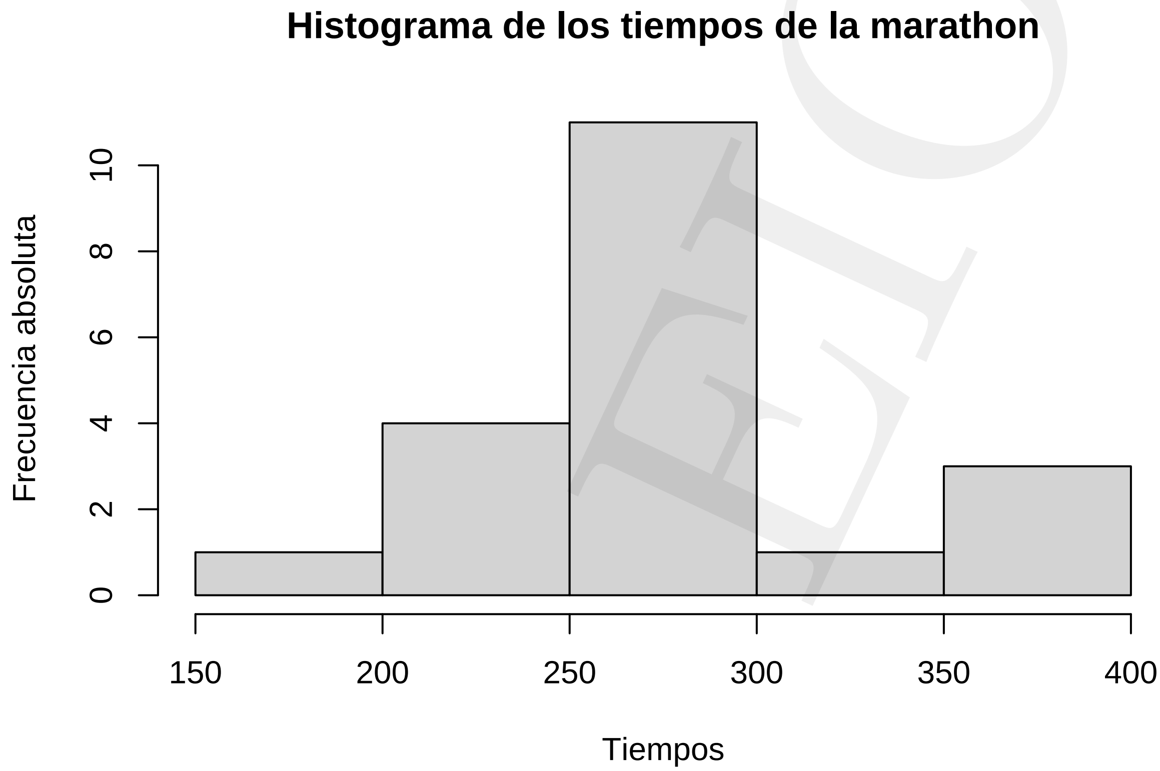
```
##      ybar      mu1     tau1
## [1,]  110 107.2442 6.383469 110 105.2921 5.841676
## [2,]  125 118.1105 6.383469 125 118.0841 7.885174
## [3,]  140 128.9768 6.383469 140 135.4134 7.973498
```

Como puede observarse si la media de los datos es consistente con las creencias a priori, la selección entre la normal o la *t*-student no influye demasiado en los resultados finales pero si el valor de la media de los datos corresponde a un “extremo’ ’ de la a priori entonces si puede haber diferencias significativas.

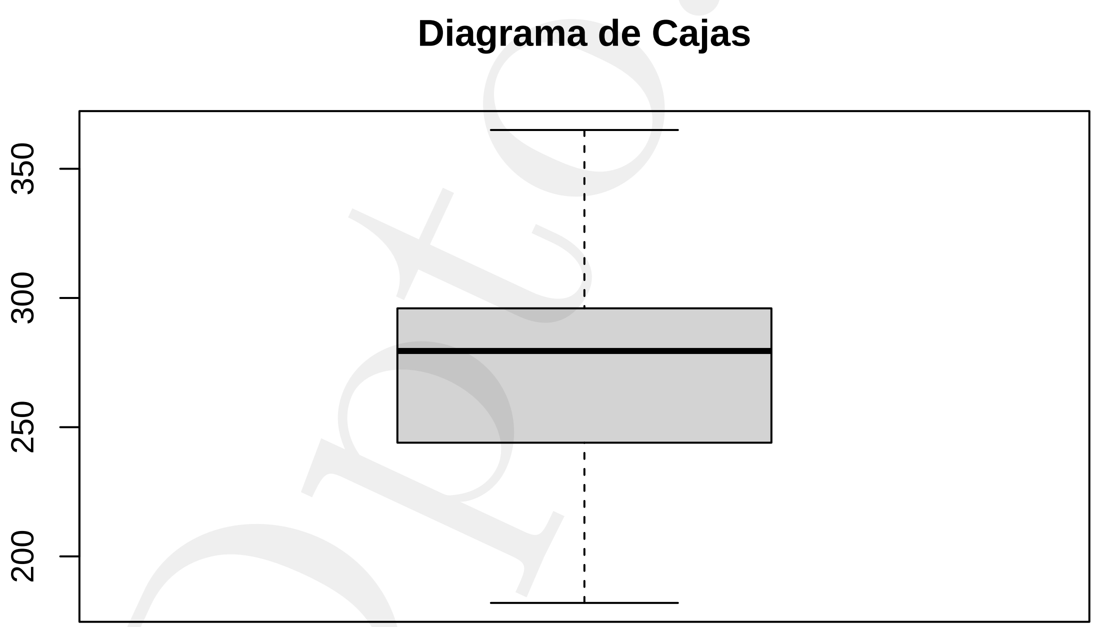
Problema 2. *LA MARATHON DE NUEVA YORK*

En el paquete de R denominado `LearnBayes` hay un conjunto de datos de los tiempos de la marathon de Nueva York en minutos de 20 atletas masculinos cuyas edades están comprendidas entre 20 y 29 años. Por experiencias previas se sabe que el percentil 10 va a ser aproximadamente 190 minutos y el percentil 90 es de 350 minutos. En primer lugar veamos si dichos datos proceden de una normal.

```
library(LearnBayes)
data("marathontimes")
View(marathontimes)
attach(marathontimes)
hist(time,ylab="Frecuencia absoluta",xlab="Tiempos",main="Histograma de los tiempos de la marathon")
```




```
boxplot(time,main="Diagrama de Cajas")
```




```
shapiro.test(time)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  time
## W = 0.97012, p-value = 0.7573
```

A continuación seleccionamos una normal a priori con la información de los percentiles que se dispone.

```
n=length(time)
media=mean(time)
q1=list(p=0.1,x=190)
q2=list(p=0.9,x=350)
nn=normal.select(q1,q2)
mu0=nn$mu
s0=nn$sigma
```

Por tanto se tiene que la media a priori y la desviación típica a priori son

$$\mu = 225 \text{ y } sd = 97.53$$

Supongamos que la desviación típica de la verosimilitud es conocida y vale $s = 75$

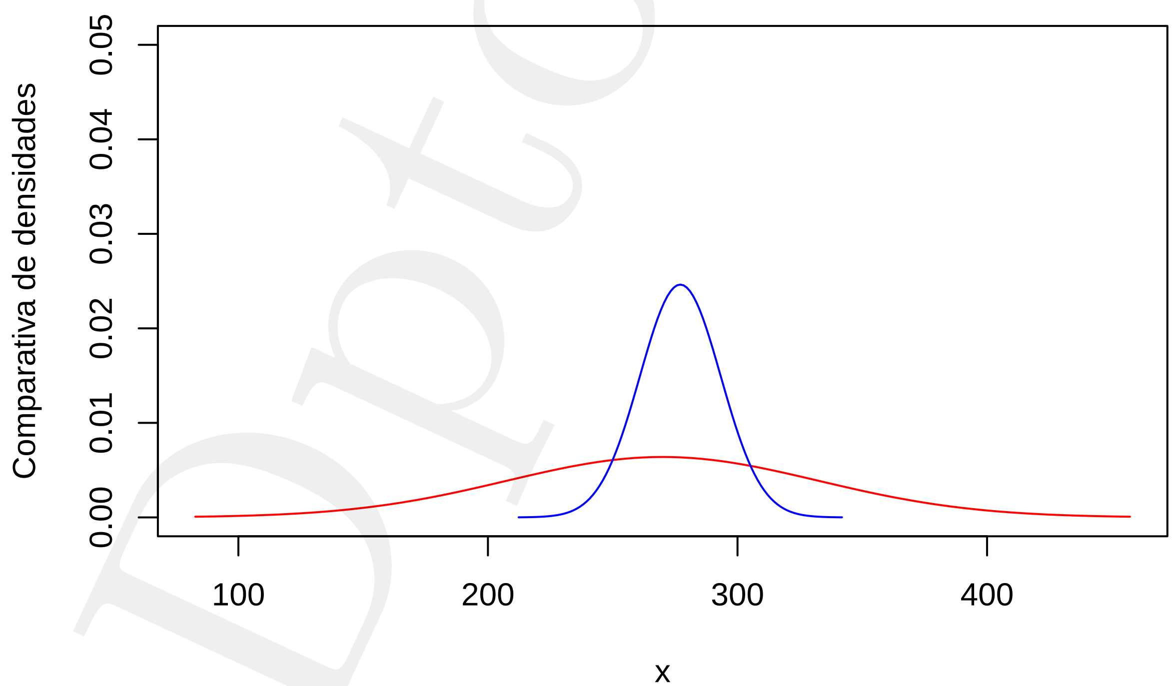
```
s=75
muposte= (n*s0^2*mean(time)+s^2*mu0)/(n*s0^2+s^2)

sposte=sqrt((s0^2*s^2)/(n*s0^2+s^2))

mm=cbind(muposte,sposte)
print(mm)
```

```
##       muposte   sposte
## [1,] 277.0884 16.19621
```

```
curve(dnorm(x,mean=mu0,sd=s0),xlim=c(mu0-3*s0,mu0+3*s0),ylim=c(0,0.05),col="red",
      ylab="Comparativa de densidades")
curve(dnorm(x,mean=muposte,sd=sposte),xlim=c(muposte-4*sposte,muposte+4*sposte),
      ylim=c(0,0.05),add=TRUE,col="blue")
```



Cálculo de la Región de Credibilidad

```
inferior=muposte-qnorm(0.975,0,1)*sposte
superior=muposte+qnorm(0.975,0,1)*sposte
```

```
rc=cbind(inferior,superior)
print(rc)
```

```
##      inferior superior
## [1,] 245.3444 308.8324
```

Factor Bayes para

$$H_0 : \mu \leq 310 \text{ frente a } H_1 : \mu > 310$$

```
f0=pnorm(310,mu0,s0)
f1=1-f0

f0/f1
```

```
## [1] 2.833853
```

```
f0poste=pnorm(310,muposte,sposte)
f0poste
```

```
## [1] 0.978926
```

```
f1poste=1-f0poste

f0poste/f1poste
```

```
## [1] 46.45178
```

```
BF01=(f0poste*f1)/(f1poste*f0)

print(BF01)
```

```
## [1] 16.39174
```

Por tanto, según los valores de la tabla de Jeffrey, hay una evidencia fuerte hacia que la hipótesis $\mu \leq 310$ con respecto a la hipótesis alternativa.