GLM: Examen 2016

Ejercicio 7

Tasas de mortalidad. Una compañía de seguros quiere lanzar un nuevo seguro médico para mineros. Para ello desea estimar la probabilidad de muerte (πi), con base en el tiempo de exposición al mineral (xi en horas). Se cuenta con información de las muertes registradas entre 1950 y 1959, junto con el tiempo de exposición al mineral y el número de mineros expuestos. Realiza un ánalisis bayesiano completo de los datos y obtén la distribución predictiva del número de muertes suponiendo que hay 100 mineros con un tiempo de exposición de 200 horas. El modelo es el siguiente:

$Yi|\pi i \sim Bin(ni, \pi i)$

$logit(\pi)i{=}\beta 0 + \beta 1xi$

Con $\beta_0 \sim N(0, 0.001)$ y $\beta_1 \sim N(0, 0.001)$

```
#Cargamos los datos y graficamos
datos<-read.table("http://allman.rhon.itam.mx/~lnieto/index_archivos/mortality.txt",header=TRUE)
n<-nrow(datos)
datos
```

```
##     x  y   n
## 1   0 13 391
## 2   5  5 205
## 3  30  5 156
## 4  75  3  50
## 5 150  4  35
## 6 250 18  51
```

```
#Dedinimos las variables a predecir de muertos dado  numero de expuestos futuros igual a 100 y el númer
m<-1
nef<-c(100) #numero de expuestos futuros
xf<-c(200)  #numero de horas
```

De acuerdo al nuestros datos la variable xi corresponde al tiempo de exposición al mineral en horas.

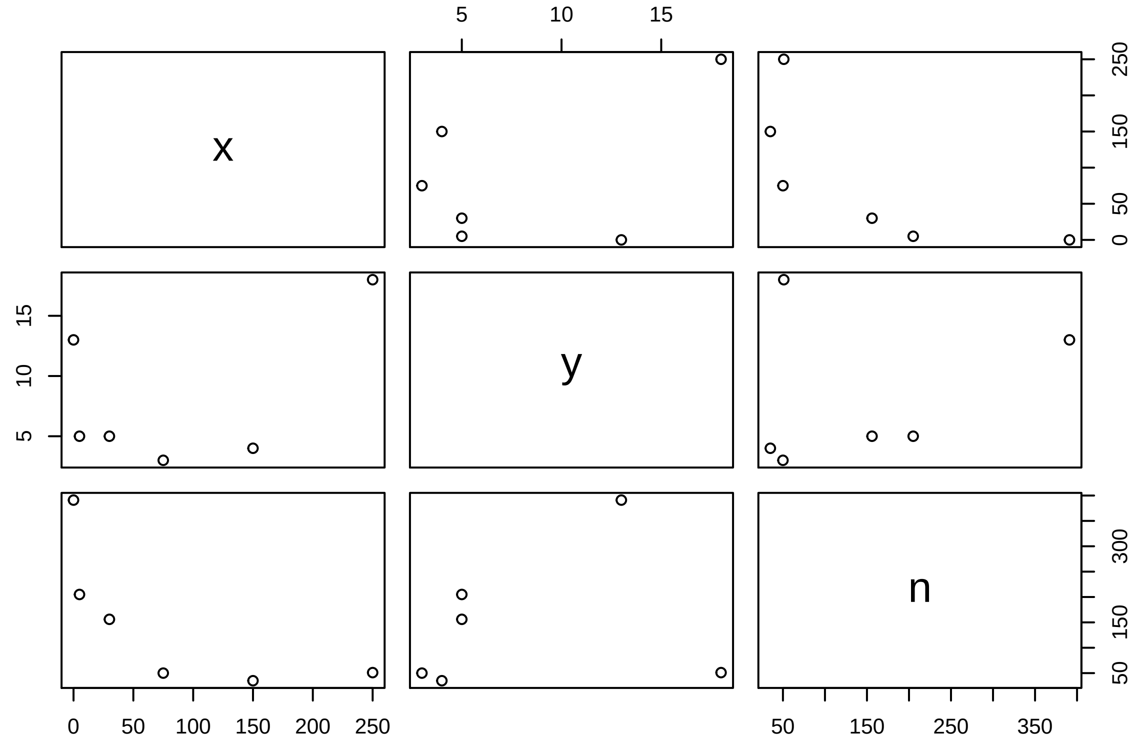
La columna n corresponde al número de mineros expuestos

La columna y es el número de muertos dado el numero de horas expuestas xi

```
datos
```

```
##     x  y   n
## 1   0 13 391
## 2   5  5 205
## 3  30  5 156
## 4  75  3  50
## 5 150  4  35
## 6 250 18  51
```

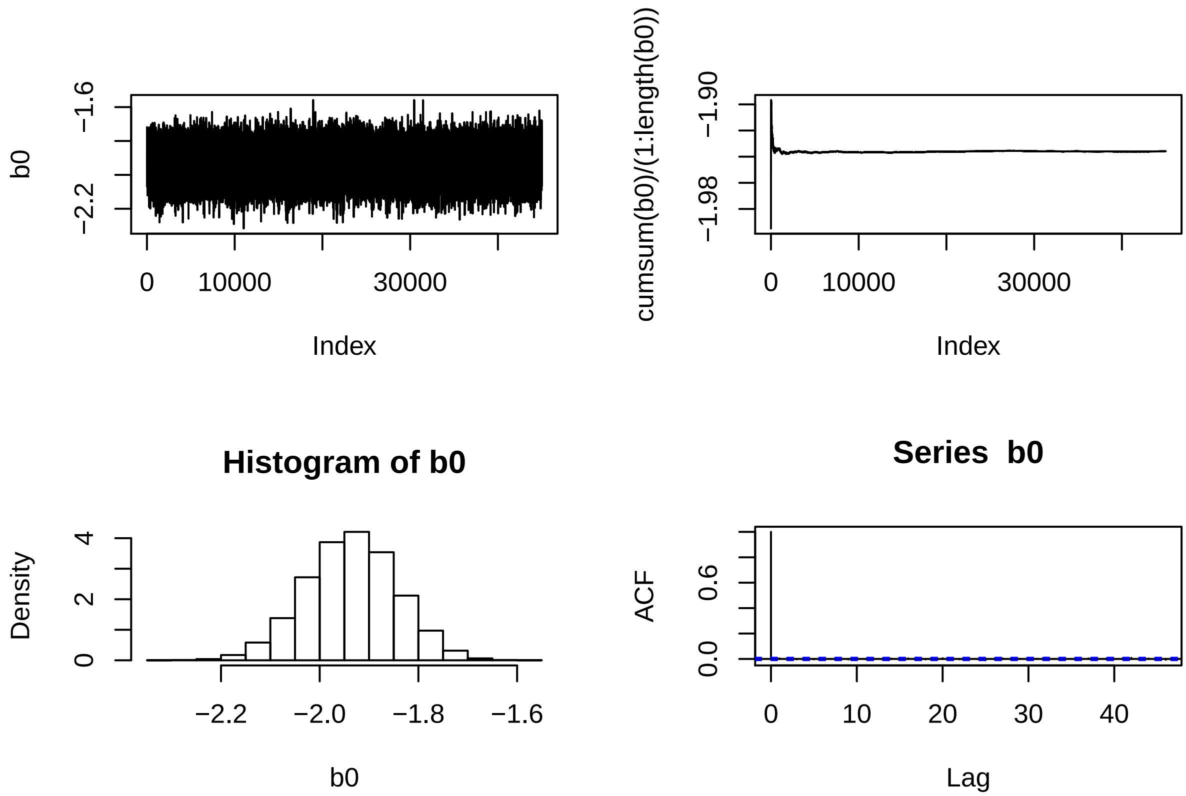
```
plot(datos)
```

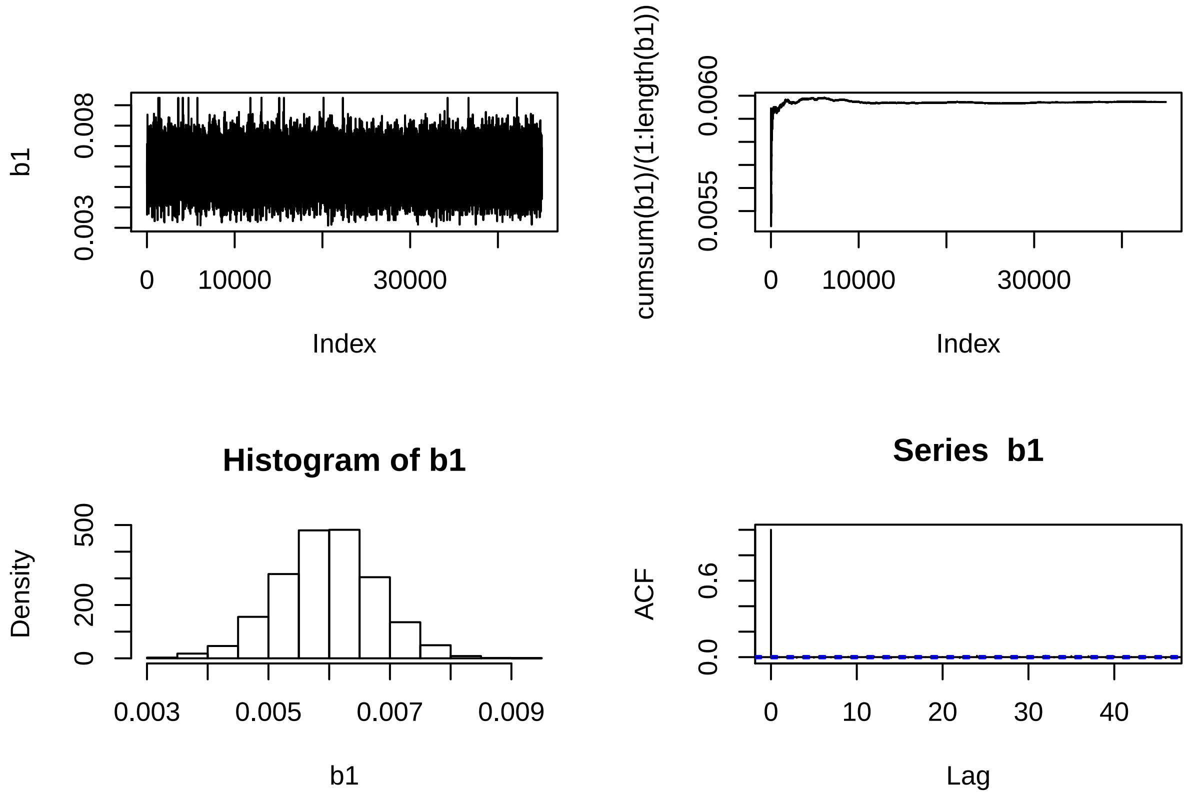



```
#Establecemos los imputs necesarios para bugs del modelo Binomial

#-Defining data-
data<-list("n"=n,"ne"=datos$n,"y"=datos$y,"x"=datos$x,"m"=m,"nef"=nef,"xf"=xf)
#-Defining inits-
inits<-function(){list(beta=rep(0,2),yf1=rep(1,n),yf2=1)}

#-Selecting parameters to monitor-
parameters<-c("beta","yf1","yf2")
  #-Running code-
```

```
#OpenBUGS

#Modelo Binomial
#Liga logistica
exa1<-bugs(data,inits,parameters,model.file="ModBinomial.txt",            n.iter=50000,n.chains=1,n.burn

#Liga probit
exa2<-bugs(data,inits,parameters,model.file="ModBinomial.txt",
             n.iter=50000,n.chains=1,n.burnin=5000)

#Modelo Poisson
exa3<-bugs(data,inits,parameters,model.file="ModPoisson.txt",
               n.iter=50000,n.chains=1,n.burnin=5000)
```

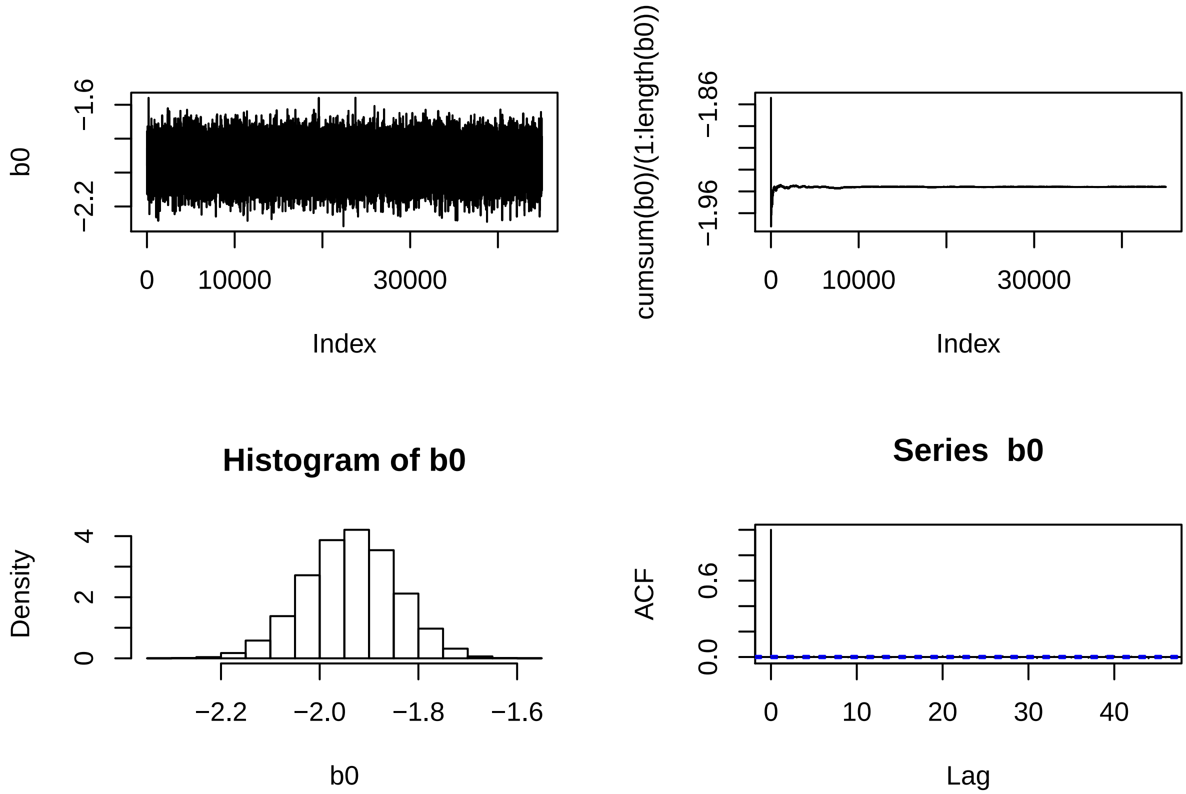
```r
#Análisis de los parámetros
#OpenBUGS

out1<-Result(exa1)
```

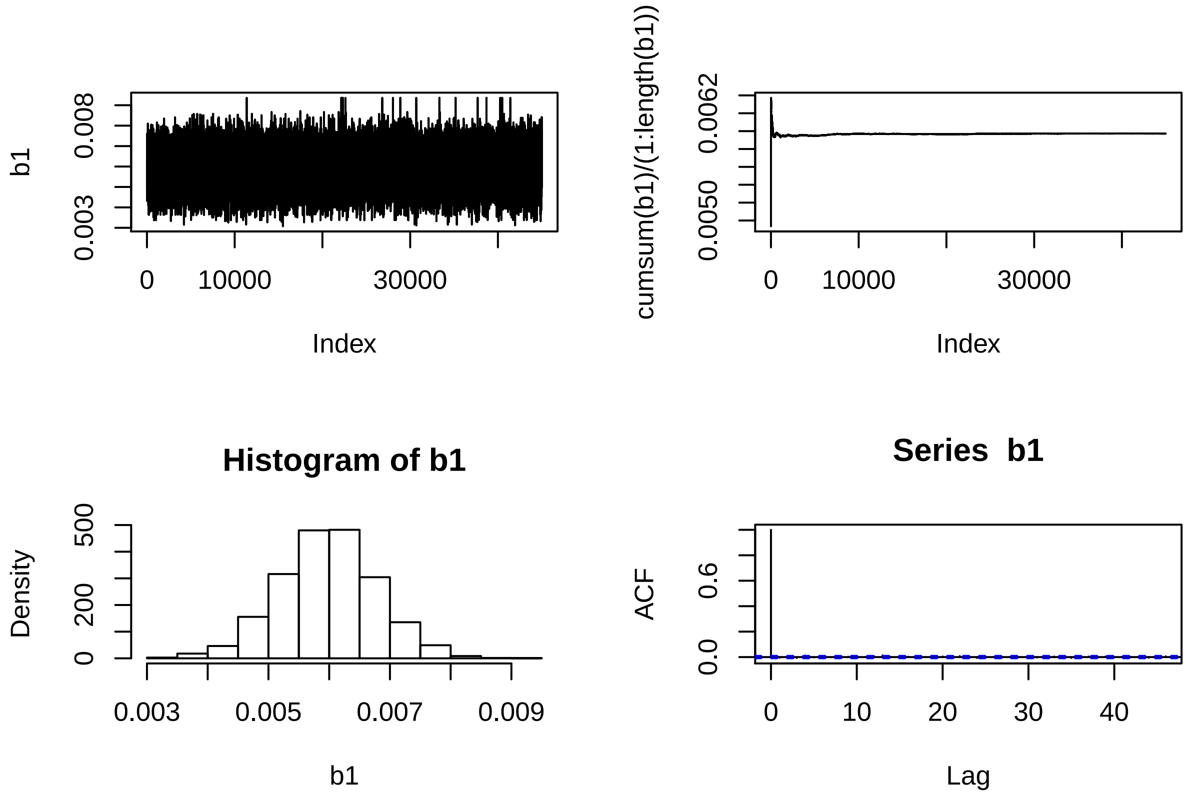





```
## [1] 0
## [1] 0
```

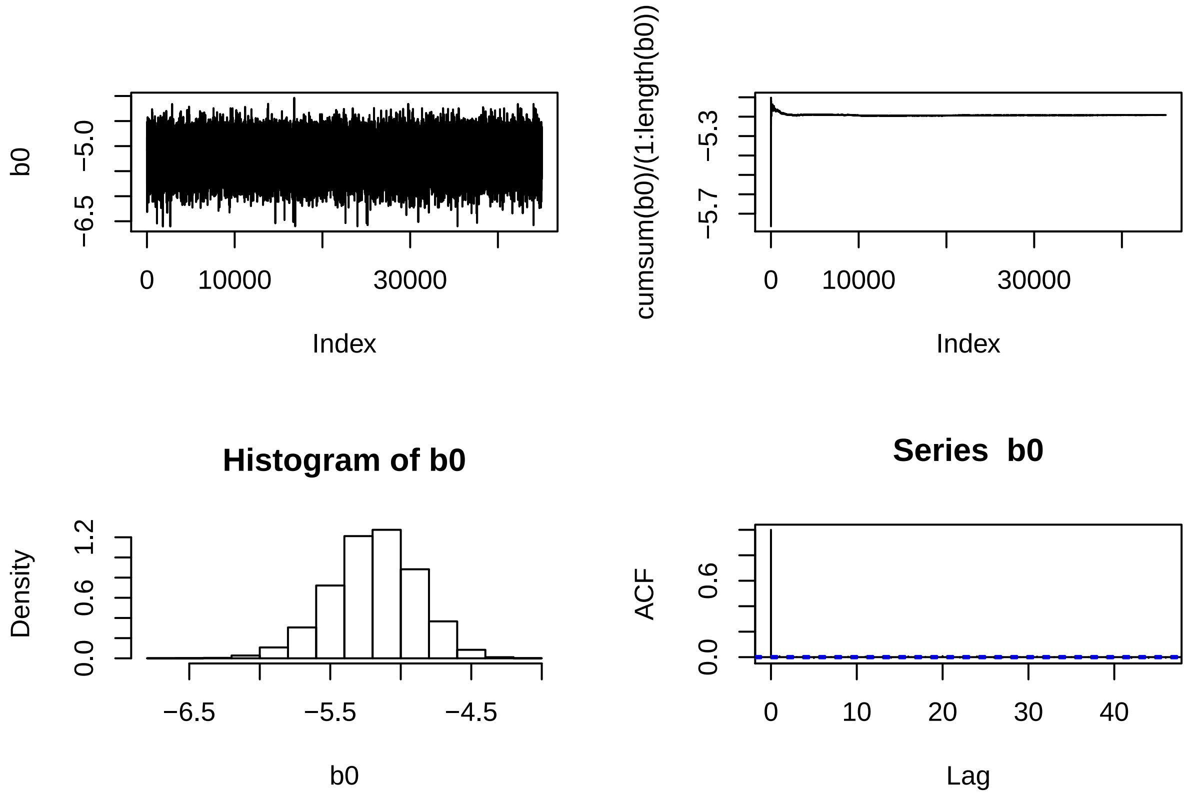
```
out2<-Result(exa2)
```

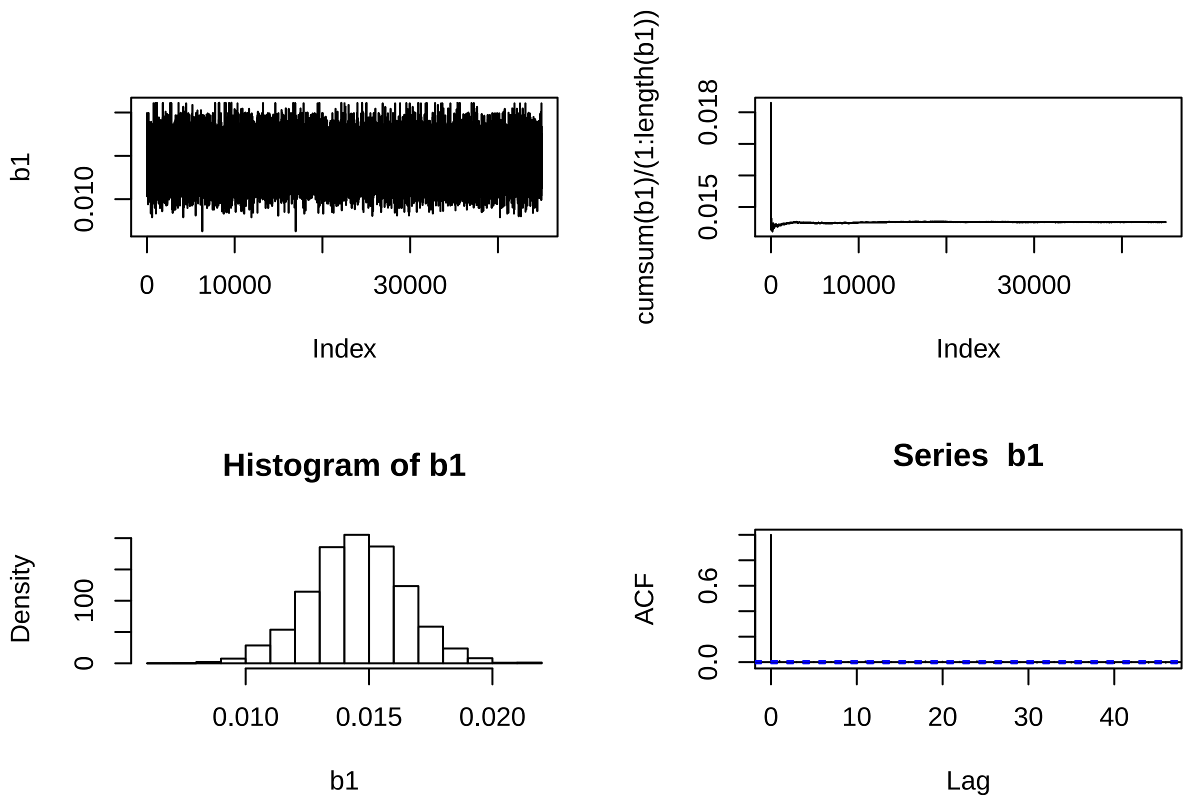

b0
−1.6
−2.2
0
10000
30000
Index
cumsum(b0)/(1:length(b0))
−1.86
−1.96
0
10000
30000
Index
Histogram of b0
Density
4
2
0
−2.2
−2.0
−1.8
−1.6
b0
Series b0
ACF
0.6
0.0
0
10
20
30
40
Lag




```
## [1] 0
## [1] 0
```

```
out3<-Result(exa3)
```


b0
−6.5
−5.0
0
10000
30000
Index
cumsum(b0)/(1:length(b0))
−5.7
−5.3
0
10000
30000
Index
Histogram of b0
Density
0.0
0.6
1.2
−6.5
−5.5
−4.5
b0
Series b0
ACF
0.0
0.6
0
10
20
30
40
Lag




```
## [1] 0
## [1] 0
```

```
#Resumen (estimadores)
#OpenBUGS
out1.sum<-exa1$summary
#print(out1.sum)
head(out1.sum)
```

```
##                  mean           sd      2.5%       25%       50%       75%
## beta[1] -1.93583820 0.0919123146 -2.119000 -1.998000 -1.935000 -1.872000
## beta[2]  0.00597254 0.0008000086  0.004339  0.005455  0.005979  0.006499
## yf1[1]  10.53122222 3.9021457331  4.000000  8.000000 10.000000 13.000000
## yf1[2]   5.91737778 2.6811688962  1.000000  4.000000  6.000000  8.000000
## yf1[3]   6.23971111 2.6694250820  2.000000  4.000000  6.000000  8.000000
## yf1[4]   3.43822222 1.8531156456  0.000000  2.000000  3.000000  5.000000
##                 97.5%
## beta[1] -1.757000000
## beta[2]  0.007556025
## yf1[1]  19.000000000
## yf1[2]  12.000000000
## yf1[3]  12.000000000
## yf1[4]   7.000000000
```

```
#DIC
out1.dic<-exa1$DIC
print(out1.dic)
```

```
## [1] 27.21
```

```r
#Resumen (estimadores)
#OpenBUGS
out2.sum<-exa2$summary
#print(out1.sum)
head(out2.sum)
```

```
##                mean           sd      2.5%       25%       50%       75%
## beta[1] -1.93583820 0.0919123146 -2.119000 -1.998000 -1.935000 -1.872000
## beta[2]  0.00597254 0.0008000086  0.004339  0.005455  0.005979  0.006499
## yf1[1]  10.53122222 3.9021457331  4.000000  8.000000 10.000000 13.000000
## yf1[2]   5.91737778 2.6811688962  1.000000  4.000000  6.000000  8.000000
## yf1[3]   6.23971111 2.6694250820  2.000000  4.000000  6.000000  8.000000
## yf1[4]   3.43822222 1.8531156456  0.000000  2.000000  3.000000  5.000000
##                97.5%
## beta[1] -1.757000000
## beta[2]  0.007556025
## yf1[1]  19.000000000
## yf1[2]  12.000000000
## yf1[3]  12.000000000
## yf1[4]   7.000000000
```

```r
#DIC
out2.dic<-exa2$DIC
print(out2.dic)
```

```
## [1] 27.21
```

```r
#Resumen (estimadores)
#OpenBUGS

out3.sum<-exa3$summary
print(out2.sum)
```

```
##                 mean           sd      2.5%       25%       50%       75%
## beta[1]  -1.93583820 0.0919123146 -2.119000 -1.998000 -1.935000 -1.872000
## beta[2]   0.00597254 0.0008000086  0.004339  0.005455  0.005979  0.006499
## yf1[1]   10.53122222 3.9021457331  4.000000  8.000000 10.000000 13.000000
## yf1[2]    5.91737778 2.6811688962  1.000000  4.000000  6.000000  8.000000
## yf1[3]    6.23971111 2.6694250820  2.000000  4.000000  6.000000  8.000000
## yf1[4]    3.43822222 1.8531156456  0.000000  2.000000  3.000000  5.000000
## yf1[5]    5.26264444 2.2688614523  1.000000  4.000000  5.000000  7.000000
## yf1[6]   16.89386667 4.4912192808  9.000000 14.000000 17.000000 20.000000
## yf2      23.14100000 5.7766305528 13.000000 19.000000 23.000000 27.000000
## deviance 25.22058600 2.0205812805 23.270000 23.800000 24.580000 25.970000
##                 97.5%
## beta[1]  -1.757000000
## beta[2]   0.007556025
## yf1[1]   19.000000000
## yf1[2]   12.000000000
## yf1[3]   12.000000000
## yf1[4]    7.000000000
## yf1[5]   10.000000000
## yf1[6]   26.000000000
## yf2      35.000000000
```

```
## deviance 30.740249960
```

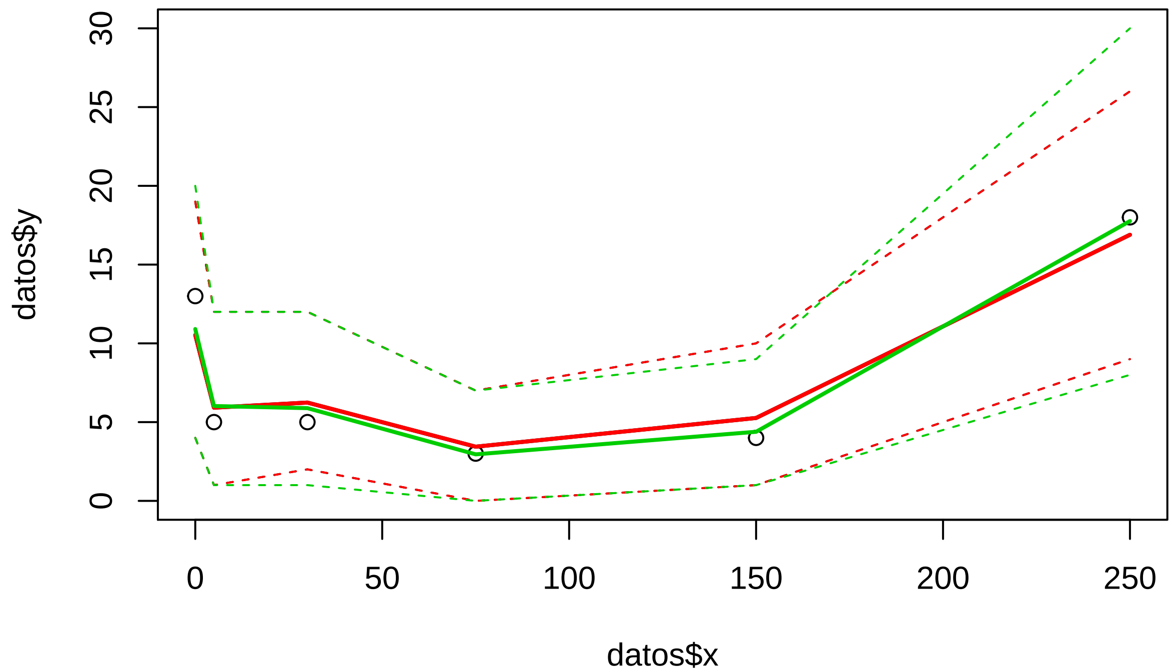
```r
head(out2.sum)
```

```
##                 mean           sd      2.5%       25%       50%       75%
## beta[1] -1.93583820 0.0919123146 -2.119000 -1.998000 -1.935000 -1.872000
## beta[2]  0.00597254 0.0008000086  0.004339  0.005455  0.005979  0.006499
## yf1[1]  10.53122222 3.9021457331  4.000000  8.000000 10.000000 13.000000
## yf1[2]   5.91737778 2.6811688962  1.000000  4.000000  6.000000  8.000000
## yf1[3]   6.23971111 2.6694250820  2.000000  4.000000  6.000000  8.000000
## yf1[4]   3.43822222 1.8531156456  0.000000  2.000000  3.000000  5.000000
##                97.5%
## beta[1] -1.757000000
## beta[2]  0.007556025
## yf1[1]  19.000000000
## yf1[2]  12.000000000
## yf1[3]  12.000000000
## yf1[4]   7.000000000
```

```r
#DIC
out3.dic<-exa3$DIC
print(out3.dic)
```

```
## [1] 27.1
```

```r
#Predictions
out1.yf<-out1.sum[grep("yf1",rownames(out1.sum)),]
out2.yf<-out2.sum[grep("yf1",rownames(out2.sum)),]
out3.yf<-out3.sum[grep("yf1",rownames(out3.sum)),]

or<-order(datos$x)
ymin<-min(datos$y,out1.yf[,c(1,3,7)],out2.yf[,c(1,3,7)],out3.yf[,c(1,3,7)])
ymax<-max(datos$y,out1.yf[,c(1,3,7)],out2.yf[,c(1,3,7)],out3.yf[,c(1,3,7)])

par(mfrow=c(1,1))
plot(datos$x,datos$y,ylim=c(ymin,ymax))


#Modelo 1
lines(datos$x[or],out1.yf[or,1],lwd=2,col=1)
lines(datos$x[or],out1.yf[or,3],lty=2,col=1)
lines(datos$x[or],out1.yf[or,7],lty=2,col=1)

#Modelo 2
lines(datos$x[or],out2.yf[or,1],lwd=2,col=2)
lines(datos$x[or],out2.yf[or,3],lty=2,col=2)
lines(datos$x[or],out2.yf[or,7],lty=2,col=2)

#Modelo 3
lines(datos$x[or],out3.yf[or,1],lwd=2,col=3)
lines(datos$x[or],out3.yf[or,3],lty=2,col=3)
lines(datos$x[or],out3.yf[or,7],lty=2,col=3)
```



Esta grÃ¡fica tiene las tres densidades: Cambios-> Var inicial: 1 en vez de 219.47

Negra: la densidad o distribuciÃ³n inicial -> media 39 con var 1

Roja (verosimilitud - de los datos o muestra) - Tengo tres datos - n=3

- Un proceso generador con Varianza de $Var(\bar{X}) = (\varrho^2/n)$ 4/3 y la media es la $\bar{X}$ (40.93)

Aqui aparece esta informaciÃ³n original: Tenemos lo siguiente: $x1, ..., xn$ es una m.a.

$\bar{X} = (1/n)\Sigma xi$

$E(\bar{X}) = \mu$

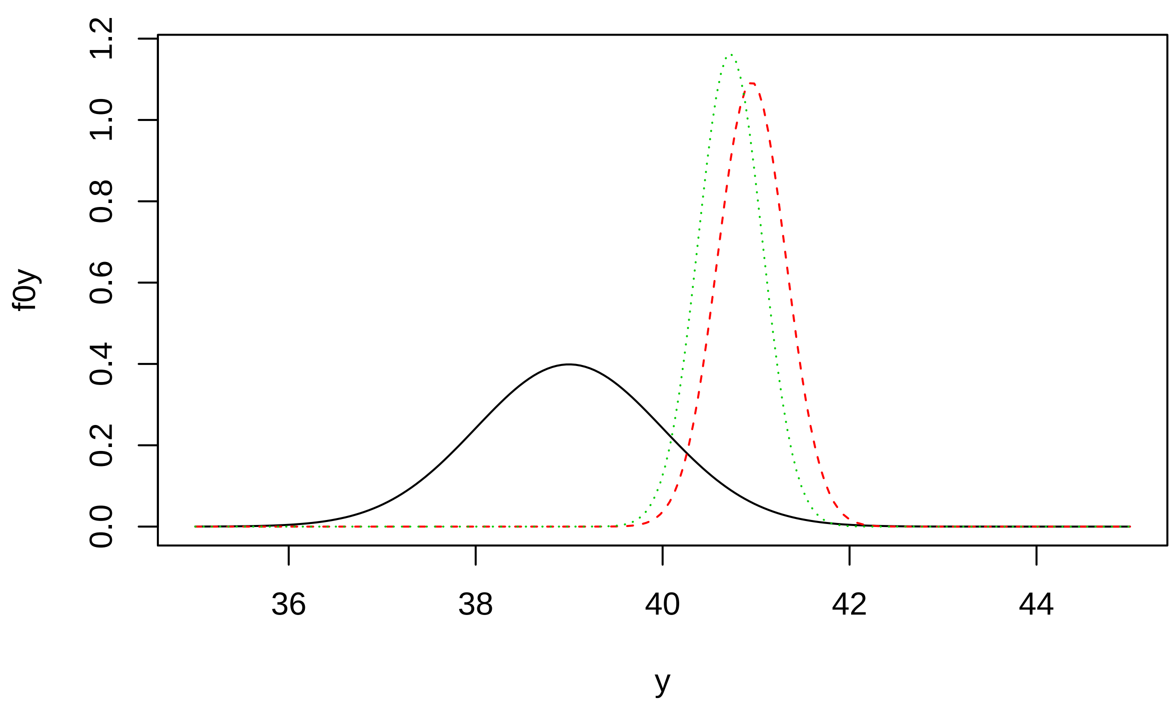
$Var(\bar{X}) = (\varrho^2/n)$

Verde (distribuciÃ³n final) Al combinar la inicial y la verosimilitud, la final queda en medio de las dos, porque las varianzas de ambas (inicial y veros.) son chiquitas. En el primer ejemplo, la varianza de la inicial era inmensa - 219.47 - lo cual indica que sÃ© bastante poco de como es esa distribuciÃ³n, y por lo tanto mis datos son los que dan mas confianza a la distribuciÃ³n final, por asi decirlo.

Vamos a hacer varios cambios

1. Aumento n de 3 a 30. Es decir, la dist inicial es la misma, pero supongo que en vez de tener 3 datos, tengo 30. De un proceso que genera datos con varianza = 4 y la $\bar{X}$ es la misma (40.9533), por lo tanto fuÃ© generada por 30 datos en vez de 3.
2. Como nuestros datos estÃ¡n centrados mas en 40, tambien modificamos la regiÃ³n de graficaciÃ³n de (-10,100) a (30,45) con 200 puntos ("y")

```
# Datos
xbar<-40.9533
sig2<-4
n<-30
# DistribuciÃ³n inicial del parÃ¡metro theta
th0<-39
sig20<-1
# Area de graficaciÃ³n
y<-seq(35,45,,200)
# Armo la normal de theta, centrada en theta0=39 y varianza=1
f0y<-dnorm(y,th0,sqrt(sig20))
# Armo la normal de los datos - verosimilitud -, en este caso son 30, centrada en xbar=40.9533 y varian
liky<-dnorm(y,xbar,sqrt(sig2/n))
# Armo la distribuciÃ³n final conjugada tomando la base de la inicial con los datos
sig21<-1/(n/sig2+1/sig20)
th1<-sig21*(n/sig2*xbar+th0/sig20)
f1y<-dnorm(y,th1,sqrt(sig21))

# Grafico las 3 juntas
ymax<-max(f0y,liky,f1y)
plot(y,f0y,ylim=c(0,ymax),type="l")
lines(y,liky,lty=2,col=2)
lines(y,f1y,lty=3,col=3)
```



Â¿QuÃ© pasÃ³?

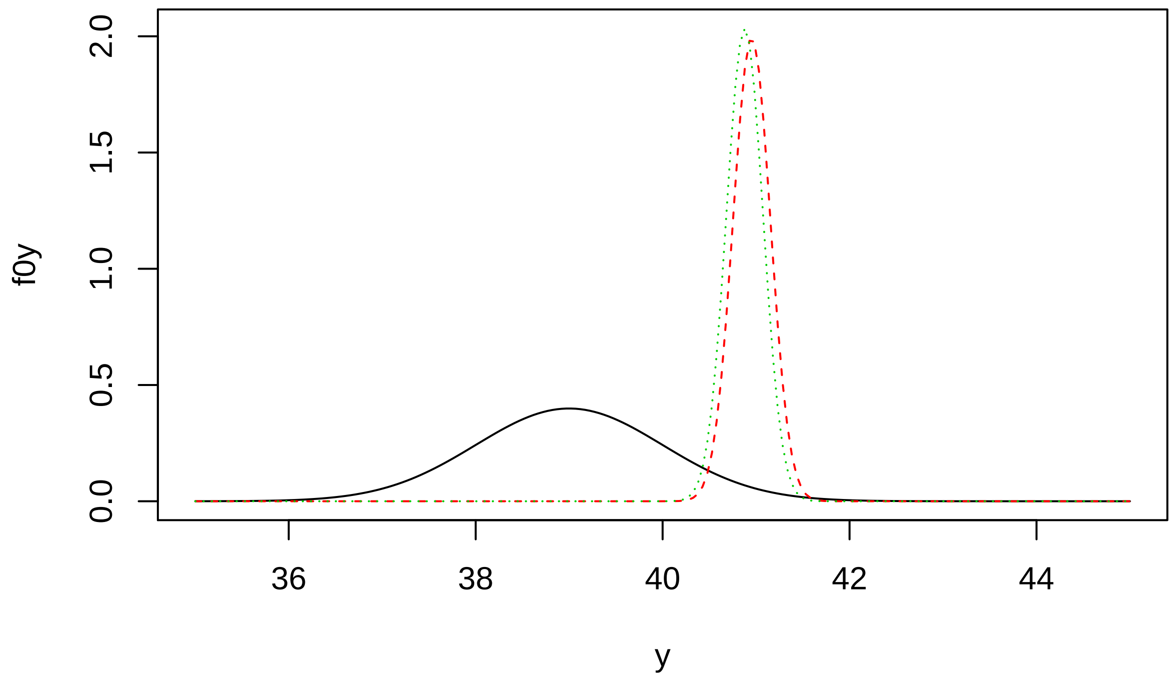
Tengo inicial con var=1 (negra)

Tengo la distribuciÃ³n de mis datos (veros) (roja) que provienen de un proceso con var=4, pero ahora tengo 30 datos en lugar de 3.

Â¿QuÃ© le pasa a la dist final?(verde) Se estÃ¡ cargando hacia donde estan los datos. En el caso anterior la dis final estaba en medio de la nega y la roja, es decir, que podriamos decir que pesaban igual la inicial y la final.

Â¿Que pasa si en vez de 30 datos ahora tengo 100?

```
# Datos
xbar<-40.9533
sig2<-4
n<-100
# DistribuciÃ³n inicial del parÃ¡metro theta
th0<-39
sig20<-1
# Area de graficaciÃ³n
y<-seq(35,45,,200)
# Armo la normal de theta, centrada en theta0=39 y varianza=1
f0y<-dnorm(y,th0,sqrt(sig20))
# Armo la normal de los datos - verosimilitud -, en este caso son 30, centrada en xbar=40.9533 y varian
liky<-dnorm(y,xbar,sqrt(sig2/n))
# Armo la distribuciÃ³n final conjugada tomando la base de la inicial con los datos
sig21<-1/(n/sig2+1/sig20)
th1<-sig21*(n/sig2*xbar+th0/sig20)
f1y<-dnorm(y,th1,sqrt(sig21))

# Grafico las 3 juntas
ymax<-max(f0y,liky,f1y)
plot(y,f0y,ylim=c(0,ymax),type="l")
lines(y,liky,lty=2,col=2)
lines(y,f1y,lty=3,col=3)
```



Mi distribuciÃ²n final se mueve cada vez mÃ¡s hacia la verosimilitud.

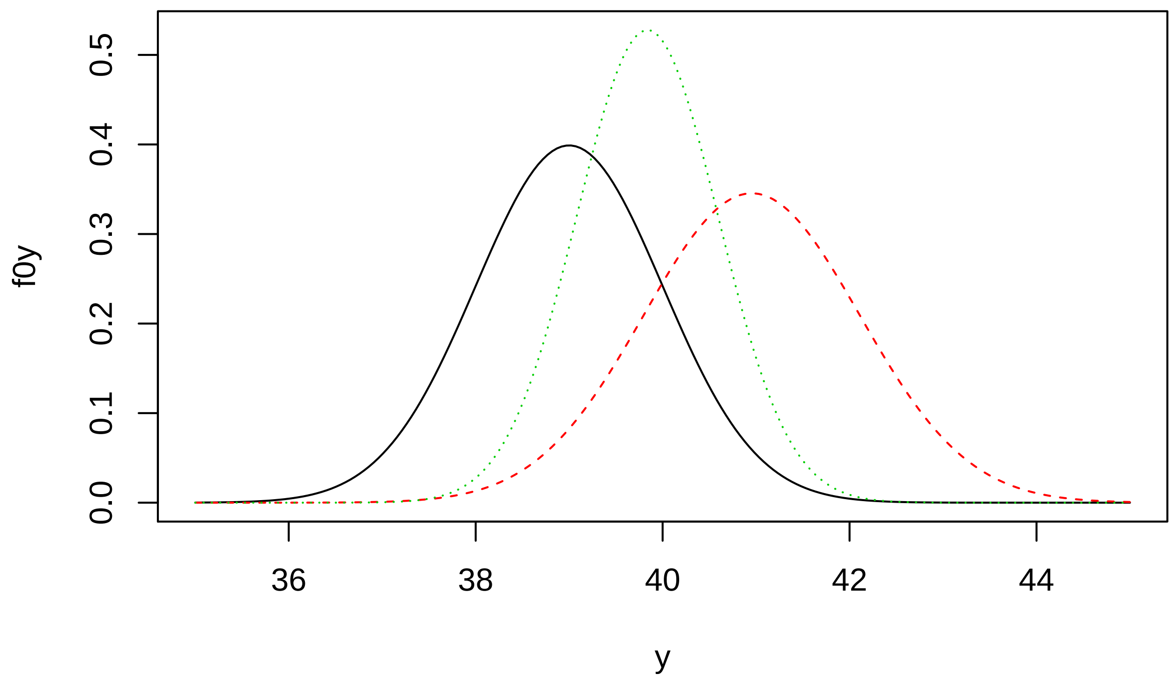
ESTO SIGNIFICA QUE EN LA MEDIDA QUE TENGA MAS Y MAS DATOS, NO IMPORTA LA DISTRIBUCIÃ“N INICIAL QUE LE HAYA DADO, LA FINAL SE VA A CARGAR O MOVERSE, MAS SIEMPRE HACIA LA VEROSIMILITUD.

QUE PASA SI CAMBIAMOS LA VARIANZA?

Regresamos al ejemplo original con 3 datos:

```
# Datos
xbar<-40.9533
sig2<-4
n<-3
# DistribuciÃ³n inicial del parÃ¡metro theta
th0<-39
sig20<-1
# Area de graficaciÃ³n
y<-seq(35,45,,200)
# Armo la normal de theta, centrada en theta0=39 y varianza=1
f0y<-dnorm(y,th0,sqrt(sig20))
# Armo la normal de los datos - verosimilitud -, en este caso son 30, centrada en xbar=40.9533 y varian
liky<-dnorm(y,xbar,sqrt(sig2/n))
# Armo la distribuciÃ³n final conjugada tomando la base de la inicial con los datos
sig21<-1/(n/sig2+1/sig20)
th1<-sig21*(n/sig2*xbar+th0/sig20)
```

```r
f1y<-dnorm(y,th1,sqrt(sig21))

# Grafico las 3 juntas
ymax<-max(f0y,liky,f1y)
plot(y,f0y,ylim=c(0,ymax),type="l")
lines(y,liky,lty=2,col=2)
lines(y,f1y,lty=3,col=3)
```

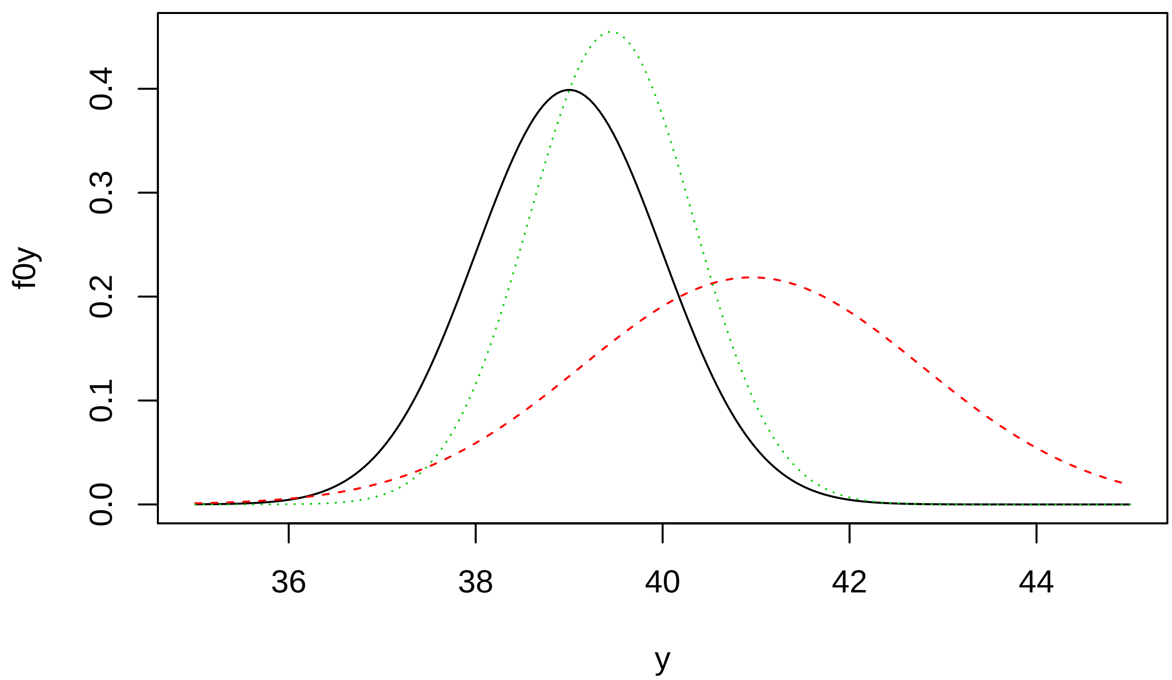


Vamos ahora a suponer que nuestro proceso inicial de generaciÃ³n de datos no tiene var=4 sino var=10.

Entonces, a mis 3 datos no le voy a creer nada o muy poco. Le voy a creer mas a la inicial del parÃ¡metro theta.

```r
# Datos
xbar<-40.9533
sig2<-10
n<-3
# DistribuciÃ³n inicial del parÃ¡metro theta
th0<-39
sig20<-1
# Area de graficaciÃ³n
y<-seq(35,45,,200)
# Armo la normal de theta, centrada en theta0=39 y varianza=1
f0y<-dnorm(y,th0,sqrt(sig20))
# Armo la normal de los datos - verosimilitud -, en este caso son 30, centrada en xbar=40.9533 y varian
liky<-dnorm(y,xbar,sqrt(sig2/n))
# Armo la distribuciÃ³n final conjugada tomando la base de la inicial con los datos
```

```
sig21<-1/(n/sig2+1/sig20)
th1<-sig21*(n/sig2*xbar+th0/sig20)
f1y<-dnorm(y,th1,sqrt(sig21))

# Grafico las 3 juntas
ymax<-max(f0y,liky,f1y)
plot(y,f0y,ylim=c(0,ymax),type="l")
lines(y,liky,lty=2,col=2)
lines(y,f1y,lty=3,col=3)
```



La verosimilitud (roja) se hace ancha, plana.

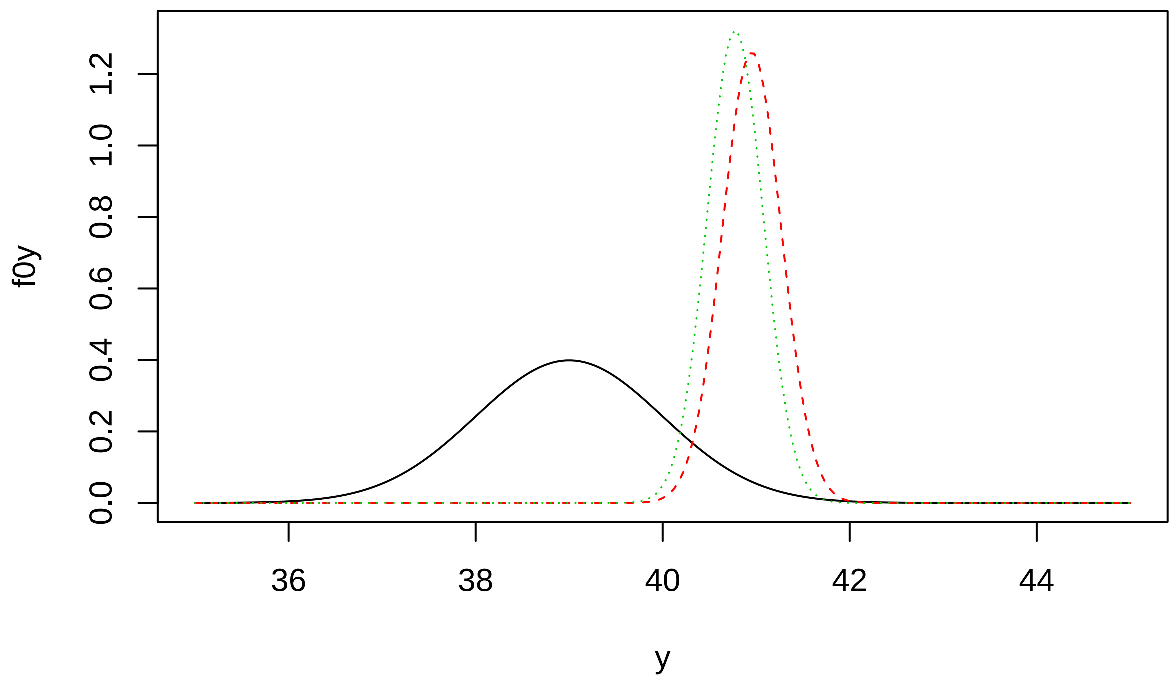
La inicial sigue igual.

Â¿quÃ© le pasa a la dist final? La final casi no se mueve respecto de la inicial. Esto sucede porque mi proceso generador de datos (roja) tiene una varianza muy grande comparada a la varianza de la dist inicial (negra).

Â¿QuÃ© sucede si conservando esta misma varianza=10, genero mas datos 100 en vez de 3.

```
# Datos
xbar<-40.9533
sig2<-10
n<-100
# DistribuciÃ³n inicial del parÃ¡metro theta
th0<-39
sig20<-1
# Area de graficaciÃ³n
y<-seq(35,45,,200)
```

```
# Armo la normal de theta, centrada en theta0=39 y varianza=1
f0y<-dnorm(y,th0,sqrt(sig20))
# Armo la normal de los datos - verosimilitud -, en este caso son 30, centrada en xbar=40.9533 y varian
liky<-dnorm(y,xbar,sqrt(sig2/n))
# Armo la distribuciÃ³n final conjugada tomando la base de la inicial con los datos
sig21<-1/(n/sig2+1/sig20)
th1<-sig21*(n/sig2*xbar+th0/sig20)
f1y<-dnorm(y,th1,sqrt(sig21))

# Grafico las 3 juntas
ymax<-max(f0y,liky,f1y)
plot(y,f0y,ylim=c(0,ymax),type="l")
lines(y,liky,lty=2,col=2)
lines(y,f1y,lty=3,col=3)
```



La final se va hacia donde estÃ¡n MIS DATOS, sin importar que la varianza sea de 10.

ConclusiÃ³n:

El proceso de aprendizaje depende del tamaÃ±o de la muestra y de las varianzas de la inicial y la final.

En la prÃ¡ctica, yo voy a tener un tamaÃ±o de muestra fijo (n) y no lo voy a poder cambiar. Y la varianza de mi proceso generador, o esta fija como en este caso de ejemplo, o la tengo que estimar dentro del proceso.

Podria jugar un poco con la varianza porque la tengo que estimar, pero mi tamaÃ±o de muestra, n, si queda fija.

Y mi distribuciÃ³n inicial, reflejarÃ¡ aquello que yo conozco, pero eso no dependerÃ¡ de los datos, es decir, no voy a ver los datos para ver como estÃ¡n, y asi poner mi inicial, NO.