GLM: Examen

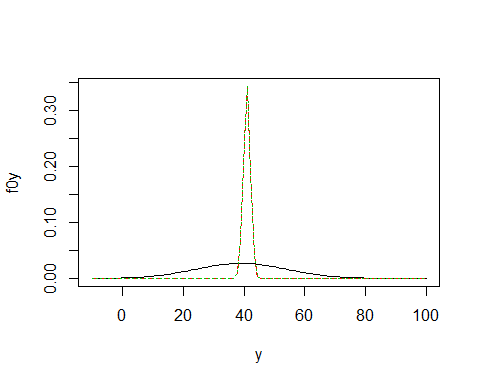
prob<-function(x){  
 out<-min(length(x[x>0])/length(x),length(x[x<0])/length(x))  
 out  
}

# Ilustración del Proceso de Inferencia

## Proceso Normal-Normal

Tenemos lo siguiente: es una m.a.

# Datos  
xbar<-40.9533  
sig2<-4  
n<-3  
# Distribución inicial del parámetro theta  
th0<-39  
sig20<-219.47  
# Area de graficación  
y<-seq(-10,100,,100)  
# Armo la normal de theta, centrada en theta0=39 y varianza=219.47  
f0y<-dnorm(y,th0,sqrt(sig20))  
# Armo la normal de los datos - verosimilitud -, en este caso son 3, centrada en xbar=40.9533 y varianza=4  
liky<-dnorm(y,xbar,sqrt(sig2/n))  
# Armo la distribución 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)



### Esta gráfica tiene en principio tres densidades:

Una de color negro: la densidad inicial, que está alrededor de 39, con una varianza de 219.47 (Una varianza muy grande indica que sé poco)

Una roja (verosimilitud - de los datos generados por mi proceso)

Una verde ( distribución final) encimadas, porque la final en este caso es la verosimilitud: LOS DATOS.

Y la razón por la cual la final solo dependió de la verosimilitud (Datos) y no de la inicial fué que la dispersiòn de los datos es muy chica (4) respecto de mi distribucion inicial (219.47). Entonces mi proceso geenerador de datos, dá más confianza a la dist final.

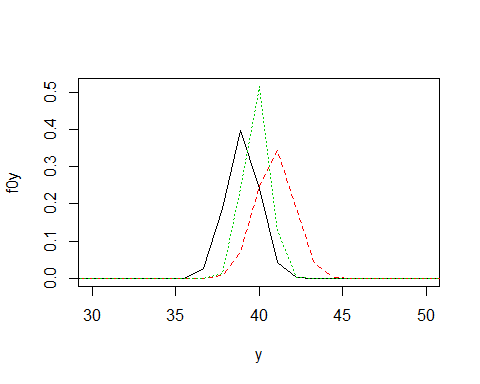
### Vamos a hacer algunos cambios

Cambio la dispersión de la distribución inicial de 219.47 a 1.

Y la escala de x a 30-50 para que nos muestre esa parte de interés Es decir cambio:

"plot(y,f0y,ylim=c(0,ymax),type="l")" por "plot(y,f0y,ylim=c(0,ymax),type="l",xlim=c(30,50))"

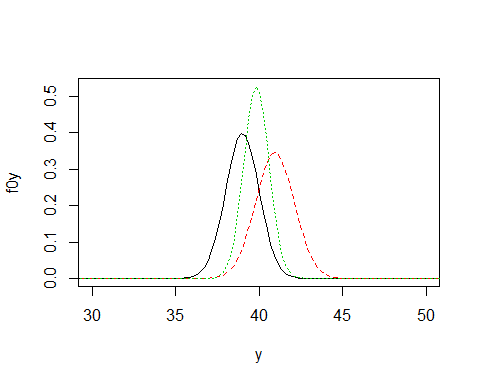
# Datos  
xbar<-40.9533  
sig2<-4  
n<-3  
# Distribución inicial del parámetro theta  
th0<-39  
sig20<-1  
# Area de graficación  
y<-seq(-10,100,,100)  
# Armo la normal de theta, centrada en theta0=39 y varianza=1  
f0y<-dnorm(y,th0,sqrt(sig20))  
# Armo la normal de los datos - verosimilitud -, en este caso son 3, centrada en xbar=40.9533 y varianza=4  
liky<-dnorm(y,xbar,sqrt(sig2/n))  
# Armo la distribución 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",xlim=c(30,50))  
lines(y,liky,lty=2,col=2)  
lines(y,f1y,lty=3,col=3)



Porque nos dan picos y no curvas suaves? porque la gráfica solo tiene 100 puntos en y<-seq(-10,100,,100), donde "y" es "la región de graficación" En la linea que tenemos, esta región es de 100 puntos en el intervalo (-10,100)

Si lo modificamos a 500 y<-seq(-10,100,,500) y limito las x en el plot entre 30 y 50.

# Datos  
xbar<-40.9533  
sig2<-4  
n<-3  
# Distribución inicial del parámetro theta  
th0<-39  
sig20<-1  
# Area de graficación  
y<-seq(-10,100,,500)  
# 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 3, centrada en xbar=40.9533 y varianza=4  
liky<-dnorm(y,xbar,sqrt(sig2/n))  
# Armo la distribución 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",xlim=c(30,50))  
lines(y,liky,lty=2,col=2)  
lines(y,f1y,lty=3,col=3)

 ### Esta gráfica tiene las tres densidades: Cambios-> Var inicial: 1 en vez de 219.47

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

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

* Un proceso generador con Varianza de 4/3 y la media es la (40.93)

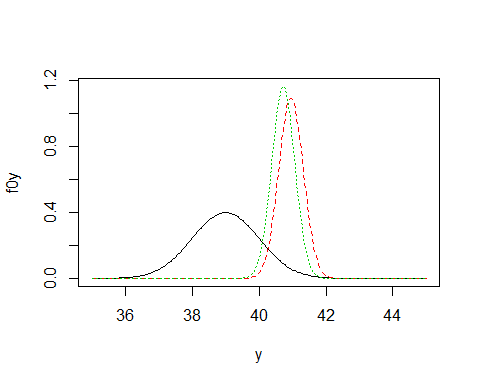
Aqui aparece esta información original: Tenemos lo siguiente: es una m.a.

Verde (distribución 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 sé bastante poco de como es esa distribución, y por lo tanto mis datos son los que dan mas confianza a la distribución 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 es la misma (40.9533), por lo tanto fué generada por 30 datos en vez de 3.
2. Como nuestros datos están centrados mas en 40, tambien modificamos la región de graficación de (-10,100) a (30,45) con 200 puntos ("y")

# Datos  
xbar<-40.9533  
sig2<-4  
n<-30  
# Distribución inicial del parámetro theta  
th0<-39  
sig20<-1  
# Area de graficación  
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 varianza=4  
liky<-dnorm(y,xbar,sqrt(sig2/n))  
# Armo la distribución 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)



¿Qué pasó?

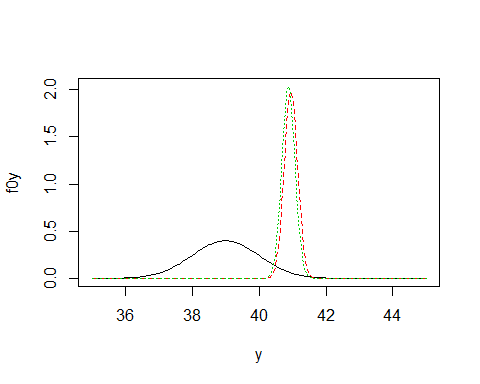
Tengo inicial con var=1 (negra)

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

¿Qué le pasa a la dist final?(verde) Se está 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.

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

# Datos  
xbar<-40.9533  
sig2<-4  
n<-100  
# Distribución inicial del parámetro theta  
th0<-39  
sig20<-1  
# Area de graficación  
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 varianza=4  
liky<-dnorm(y,xbar,sqrt(sig2/n))  
# Armo la distribución 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)



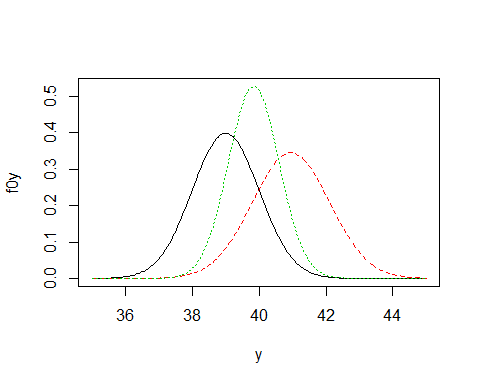
Mi distribuciòn final se mueve cada vez más hacia la verosimilitud.

ESTO SIGNIFICA QUE EN LA MEDIDA QUE TENGA MAS Y MAS DATOS, NO IMPORTA LA DISTRIBUCIÓN 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:

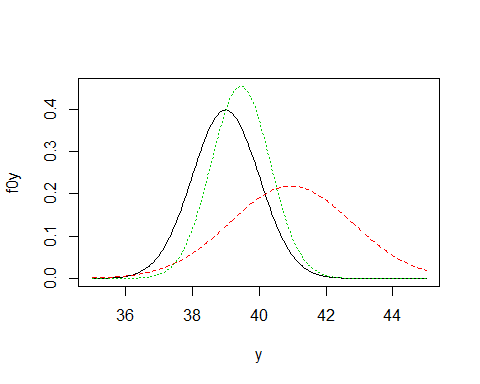
# Datos  
xbar<-40.9533  
sig2<-4  
n<-3  
# Distribución inicial del parámetro theta  
th0<-39  
sig20<-1  
# Area de graficación  
y<-seq(35,45,,200)  
# Armo la normal de theta, centrada en theta0=39 y varianza=1  
f0y<-dnorm(y,th0,sqrt(sig20))  
# Armo la normal de los datos - verosimilitud -, en este caso son 30, centrada en xbar=40.9533 y varianza=4  
liky<-dnorm(y,xbar,sqrt(sig2/n))  
# Armo la distribución 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)



Vamos ahora a suponer que nuestro proceso inicial de generación 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 parámetro theta.

# Datos  
xbar<-40.9533  
sig2<-10  
n<-3  
# Distribución inicial del parámetro theta  
th0<-39  
sig20<-1  
# Area de graficación  
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 varianza=4  
liky<-dnorm(y,xbar,sqrt(sig2/n))  
# Armo la distribución 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)



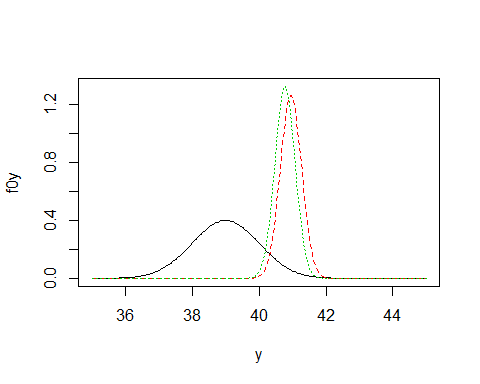
La verosimilitud (roja) se hace ancha, plana.

La inicial sigue igual.

¿qué 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).

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

# Datos  
xbar<-40.9533  
sig2<-10  
n<-100  
# Distribución inicial del parámetro theta  
th0<-39  
sig20<-1  
# Area de graficación  
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 varianza=4  
liky<-dnorm(y,xbar,sqrt(sig2/n))  
# Armo la distribución 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 están MIS DATOS, sin importar que la varianza sea de 10.

# Conclusión:

## El proceso de aprendizaje depende del tamaño de la muestra y de las varianzas de la inicial y la final.

En la práctica, yo voy a tener un tamaño 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 tamaño de muestra, n, si queda fija.

Y mi distribución inicial, reflejará aquello que yo conozco, pero eso no dependerá de los datos, es decir, no voy a ver los datos para ver como están, y asi poner mi inicial, NO.