Trabajo Práctico Probabilidades

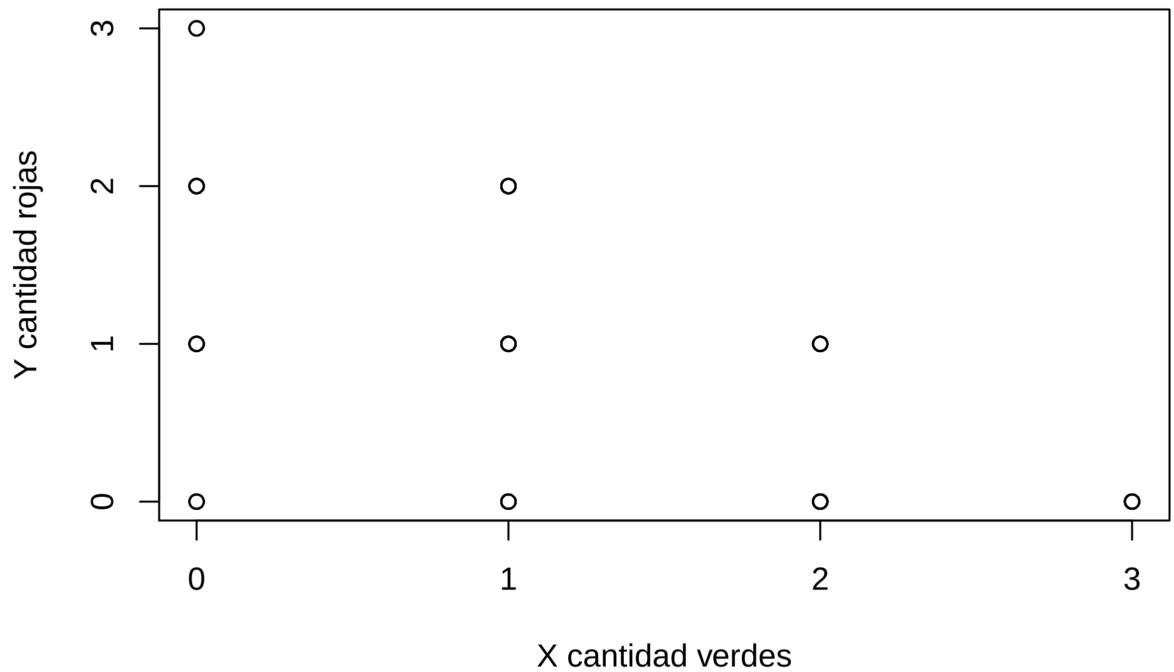
5/09/2021

Ejercicio 1)

```
urna <- c("V","V","V","V","A","A","A","R","R","R")
resultados <- data.frame(ensayo = seq(1,1000),
                         X=rep(NA,1000),
                         Y=rep(NA,1000))
set.seed(17)
for(i in 1:nrow(resultados)){
  ensayo_i <- sample(1:10, 3, replace=F)
  bolas <- urna[ensayo_i]
  resultados$X[i] <- sum(bolas=="V")
  resultados$Y[i] <- sum(bolas=="R")
}
```

Inciso a)

```
plot(resultados$X, resultados$Y, xlab="X cantidad verdes", ylab="Y cantidad rojas", xaxt="n", yaxt="n",
axis(1, at = c(0:3), cex.axis=1)
axis(2, at = c(0:3), cex.axis=1)
```



Inciso b)

Lo que se observa en el gráfico es que en 1000 repeticiones del ensayo, se obtienen todos los resultados posibles (todos los del soporte)

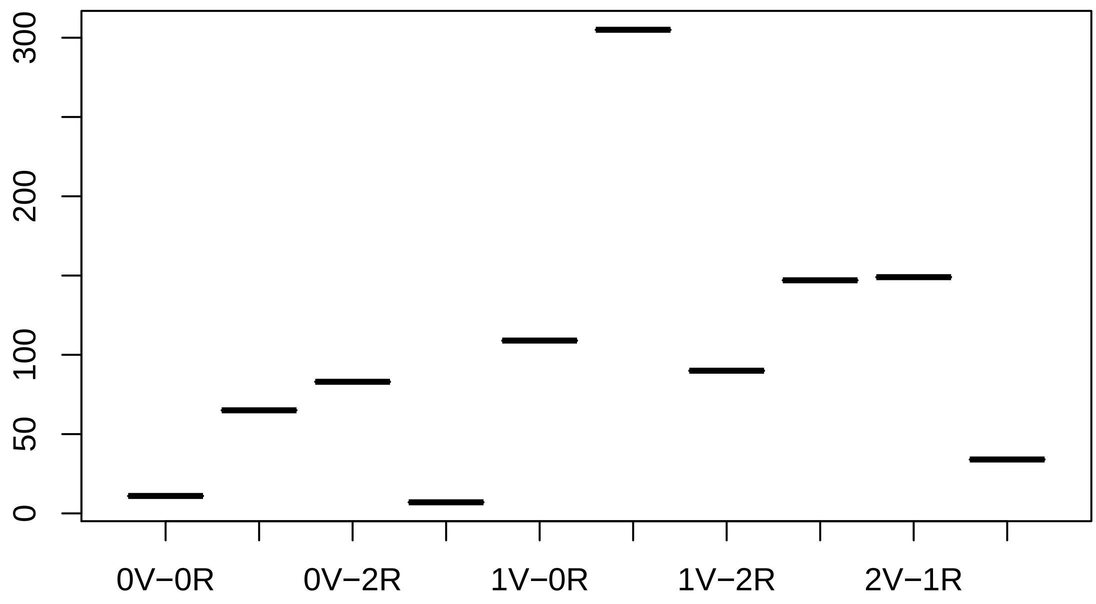
```
x <- c(rep(0,4), rep(1,3), rep(2,2), rep(3,1))
y <- c(seq(0,3), seq(0,2), seq(0,1), 0)
frecuencia <- rep(NA, length(x))
for(i in 1:length(x)){
frecuencia[i] <- nrow(resultados[resultados$X==x[i] & resultados$Y==y[i] , ])

}
resultado <- paste(paste(x,"V", sep=""), paste(y,"R", sep=""), sep="-")

tabla_frecuencias <- data.frame(x, y, resultado, frecuencia)

tabla_frecuencias$resultado <- as.factor(tabla_frecuencias$resultado)
print(tabla_frecuencias)
```

Inciso c)

```
##    x y resultado frecuencia
## 1  0 0     0V-0R         11
## 2  0 1     0V-1R         65
## 3  0 2     0V-2R         83
```

```
## 4  0 3     0V-3R          7
## 5  1 0     1V-0R        109
## 6  1 1     1V-1R        305
## 7  1 2     1V-2R         90
## 8  2 0     2V-0R        147
## 9  2 1     2V-1R        149
## 10 3 0     3V-0R         34
```

```
plot(tabla_frecuencias$resultado, tabla_frecuencias$frecuencia, xlab="resultado", ylab="frecuencia", ce
```



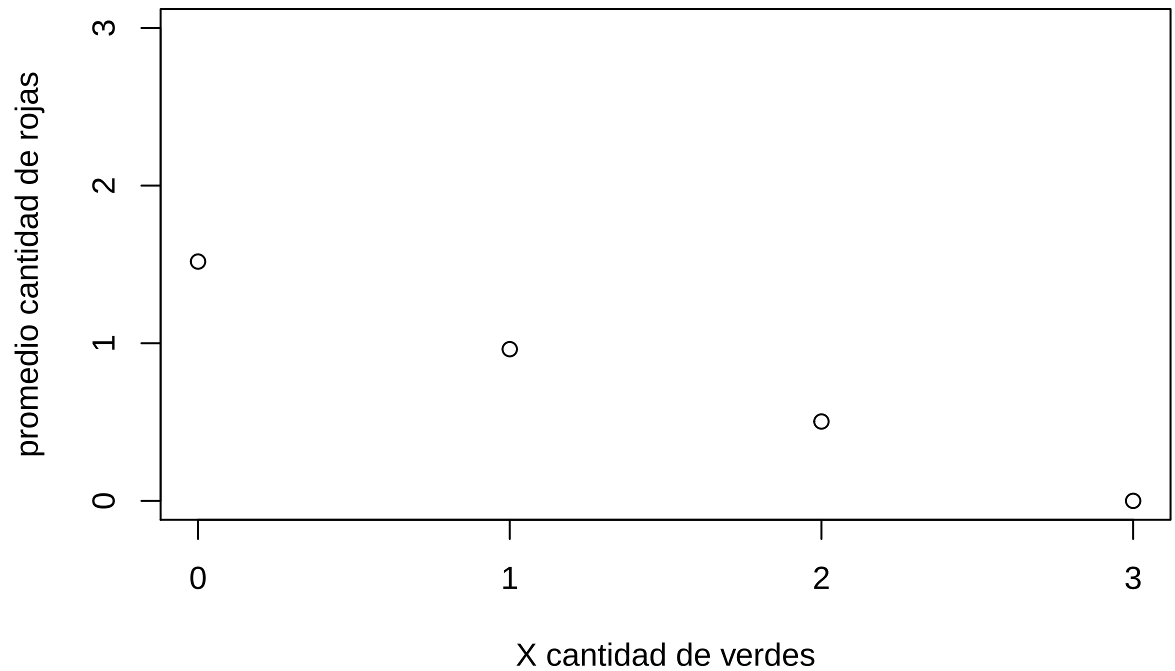

```
##histograma <- c()

##for(i in 1:10){
##histograma <- c(histograma, rep(as.character(tabla_frecuencias$resultado[i]), ##tabla_frecuencias$fre
##}
```

Lo que se observa en la tabla de frecuencias (y en el gráfico) es que el resultado más frecuente es extraer 1 bolilla verde y una roja entre las 3 bolillas extraídas.

```
rango_x <- (0:3)
promedio_y <- c(mean(resultados[resultados$X==0 ,]$Y) , mean(resultados[resultados$X==1 ,]$Y) , mean(re
```

Inciso d)

```
plot(rango_x, promedio_y, ylim=c(0,3), xaxt="n", yaxt="n", type="p", xlab="X cantidad de verdes" , ylab
axis(1, at = c(0:3), cex.axis=1)
axis(2, at = c(0:3), cex.axis=1)
```



Inciso e)

```
##prob_y_x0 =

##E_y_x0 <- sum()
##E_y_x1 <- 0
##E_y_x2 <- 0
##E_y_x3 <- 0

##for(i in 1:1000){
##  E_y_x0 <-
##}
```

Inciso f)

Ejercicio 2)

Inciso a) A partir de la función de distribución conjunta, obtenemos la función de distribución marginal

$f_{XY}(x,y) = \frac{1}{2x+1} e^{-2x - \frac{y}{4x+2}}$

$f_X(x) = 2e^{-2x}$

Que resulta una exponencial de parámetro:

$\lambda = 2$

Luego:

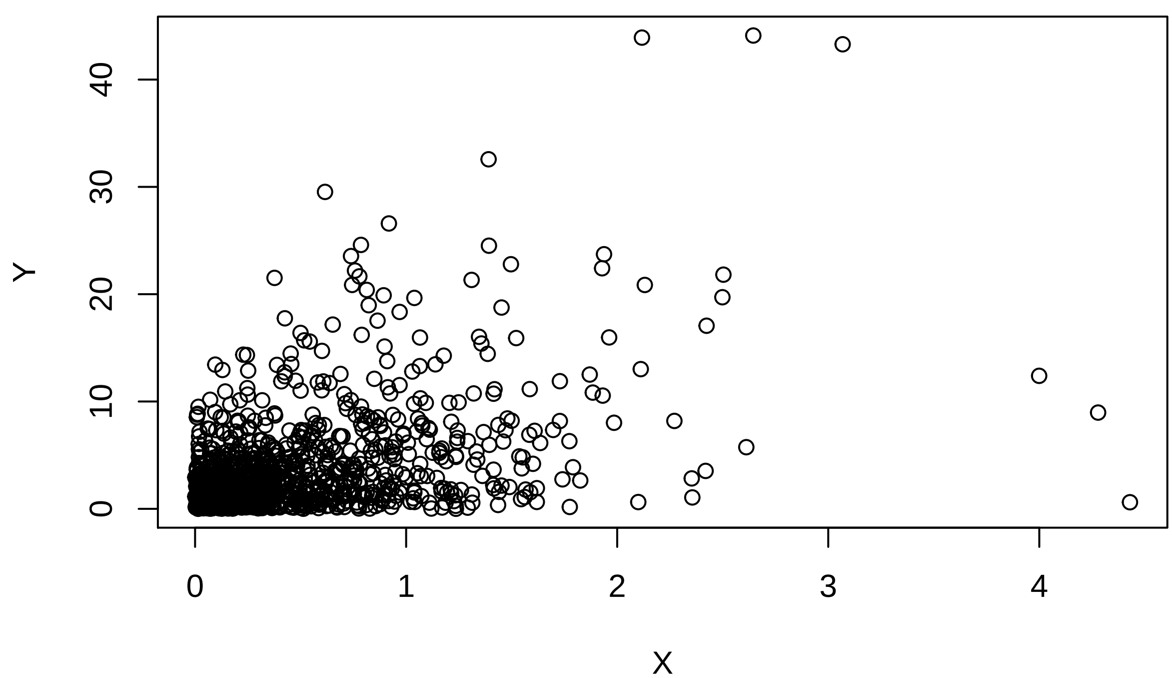
$f_{Y|X}(y) = \frac{1}{4x+2} e^{-\frac{y}{4x+2}}$

Que resulta una exponencial de parámetro:

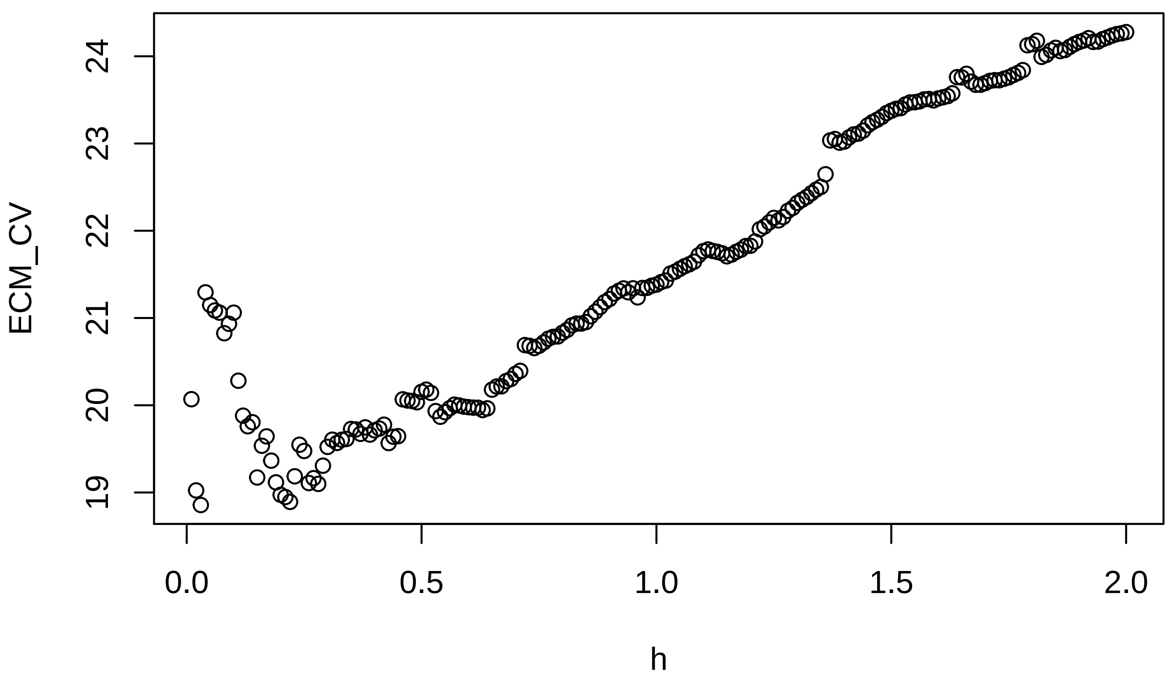
$\lambda = \frac{1}{4x+2}$

Para hacer la simulación, simularemos X con la distribución dada, luego y para cada x obtenido con la distribución dada.

```
set.seed(17)
repeticion <- seq(1,1000)
X <- rexp(1000, rate=2)
Y <- rep(NA, length(X))
for(i in 1:length(Y)){
  Y[i] <- rexp(1, rate= 1/(4*X[i]+2) )
}

plot(X,Y)
```



```
XY <- data.frame(X,Y)
h <- seq(0.01, 2 , 0.01)

for (j in 1:length(h)){
  Y_promedio <- rep(NA, nrow(XY))
  for(i in 1:nrow(XY)){
    if( nrow( XY[abs(XY$X[i] - XY$X)<h[j] ,]) >1 ){
      ##vector con el promedio de los valores cercanos, sin usar el Y en el punto
      Y_promedio[i] <-  ( sum( XY[ abs(XY$X[i] - XY$X)<h[j] ,]$Y) - XY$Y[i] ) / (nrow( XY[abs(XY$X[i] -
    } else {
      Y_promedio[i] <- XY$Y[i]
    }
  }
  XY <- cbind(XY, Y_promedio)
}
##Error cuadrático medio por convalidación cruzada
ECM_CV <- rep(NA, length(h))
for(i in 1:length(h)){
  ECM_CV[i] <- mean((XY$Y-XY[,i+2])**2)
}

h_opt <- h[which(ECM_CV==min(ECM_CV[-3]))]
plot(h, ECM_CV)
```

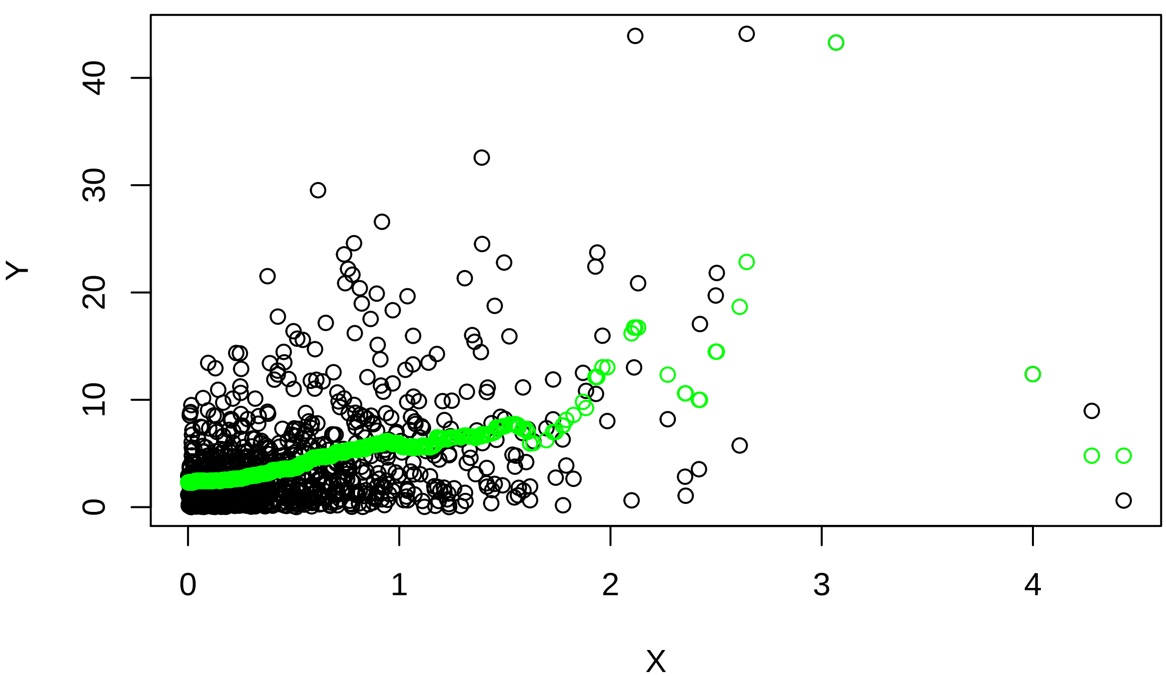


Inciso b)

Para obtener la ventanan óptima se calculó el erros cuadrático medio para distintas h usando el método de convalidación cruzada. En cada x se obtuvo el promedio de los Y incluídos en la ventana h, dejando afuera el Y correspondiente al x del centro de la ventana (método "Leave one out").

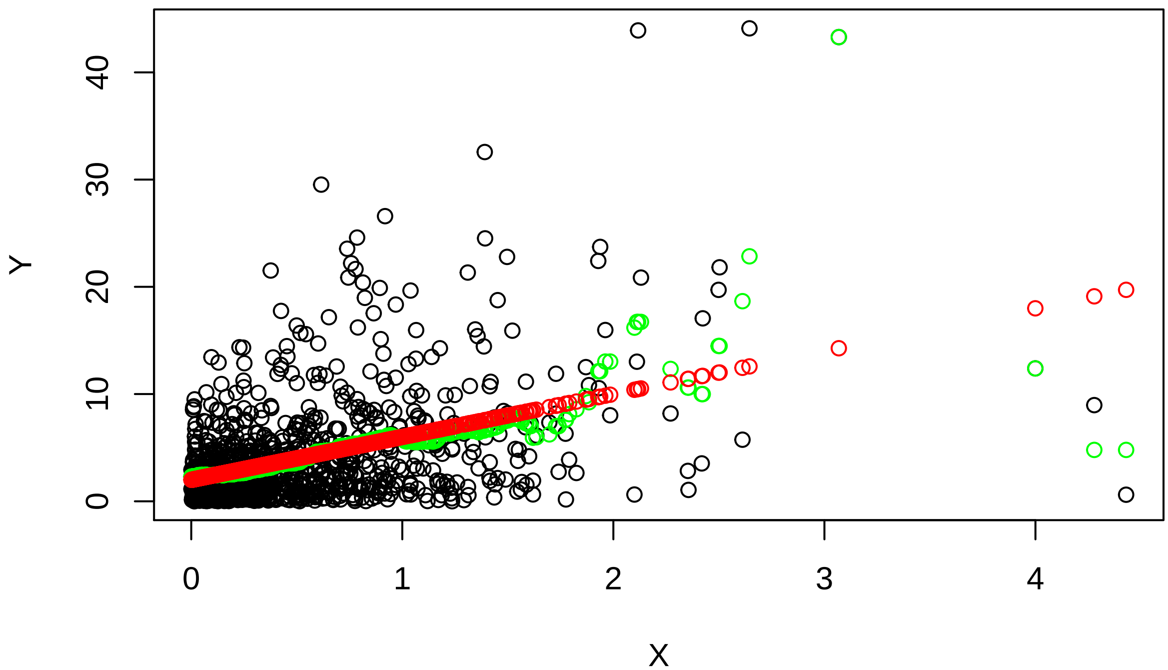
Si bien del gráfico se desprende que con los valores más chicos de h podría alcanzarse el mínimo, se decidió dejar afuera del análisis los h tan chicos por riesgo de "over-fitting", resultando el h óptimo = 0,22

```
Y_promedio_optimo <- rep(NA, length(Y))
for(i in 1:length(Y)){
  Y_promedio_optimo[i] <- mean( XY[ abs(XY$X[i] - XY$X)<h_opt ,]$Y)
}

plot(X,Y)
points(X, Y_promedio_optimo, col="green")
```



Inciso c)

```
E_YdadoX <- 4*X+2

plot(X,Y)
points(X, Y_promedio_optimo, col="green")
points(X, E_YdadoX, col="red")
```



Inciso d)