Trabajo Práctico Probabilidades

Federico Brusa - Javier Garcia Skabar

5/09/2021

Ejercicio 1)

En una urna hay 4 bolas verdes, 3 amarillas y 3 rojas. Se extraen tres bolas al azar sin reposición. Sean X la cantidad de bolas verdes e Y la cantidad de bolas rojas extraídas.

Inciso a) Simular 1000 realizaciones del experimento que consiste en extraer 3 bolas y observar el color, guardando el resultado de la cantidad de verdes en el vector x, y la cantidad de rojas en el vector y.

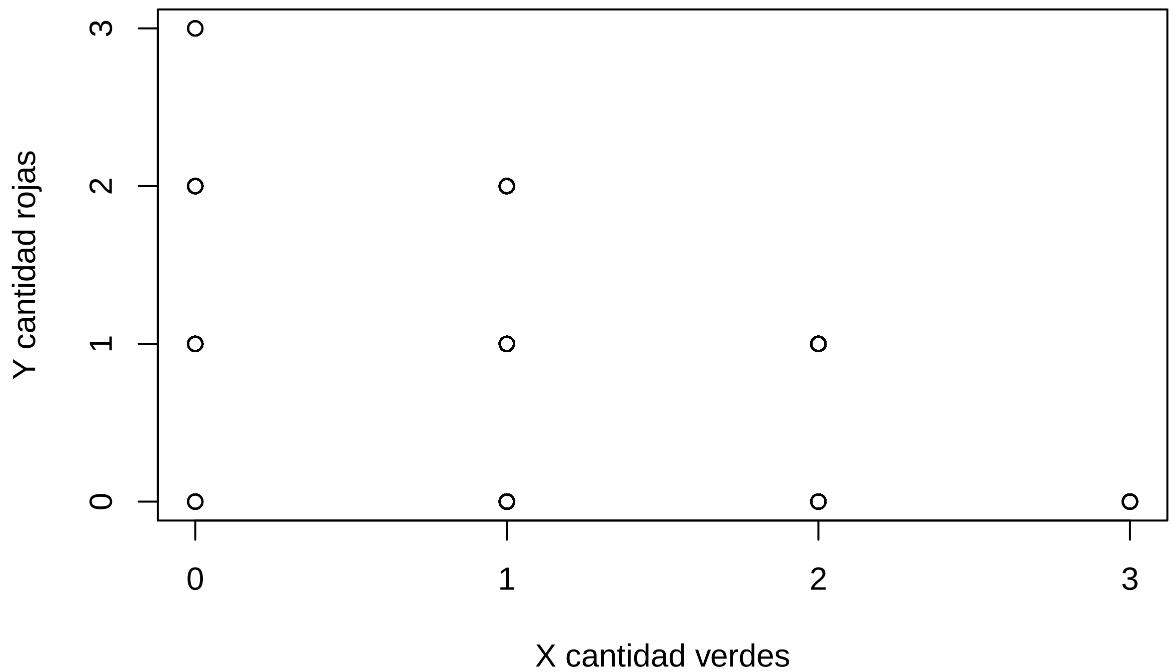
```
urna <- c("V","V","V","V","A","A","A","R","R","R")
resultados <- data.frame(ensayo = seq(1,1000),
                         X=rep(NA,1000),
                         Y=rep(NA,1000))
set.seed(17)
for(i in 1:nrow(resultados)){
  ensayo_i <- sample(1:10, 3, replace=F)
  bolas <- urna[ensayo_i]
  resultados$X[i]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

Inciso b) Realizar un gráfi

co de puntos de x vs. y. Qué se observa en este gráfi

co?

```
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)
```



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

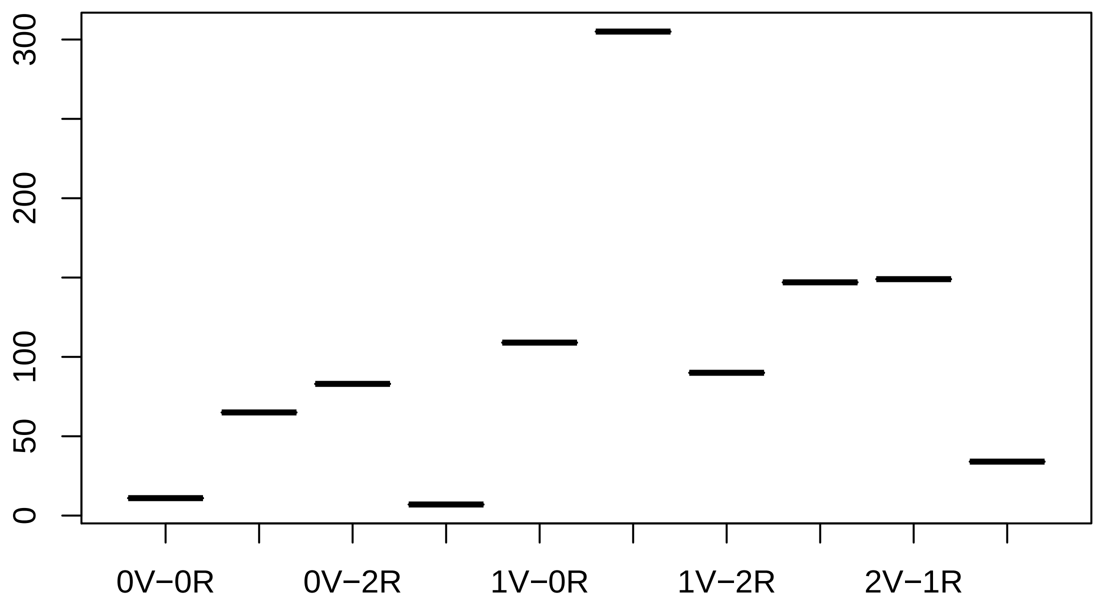
Inciso c) Hallar la tabla conjunta de frecuencias relativas para cada par (x; y). Interpretar.

```
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)
```

```
##    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
```

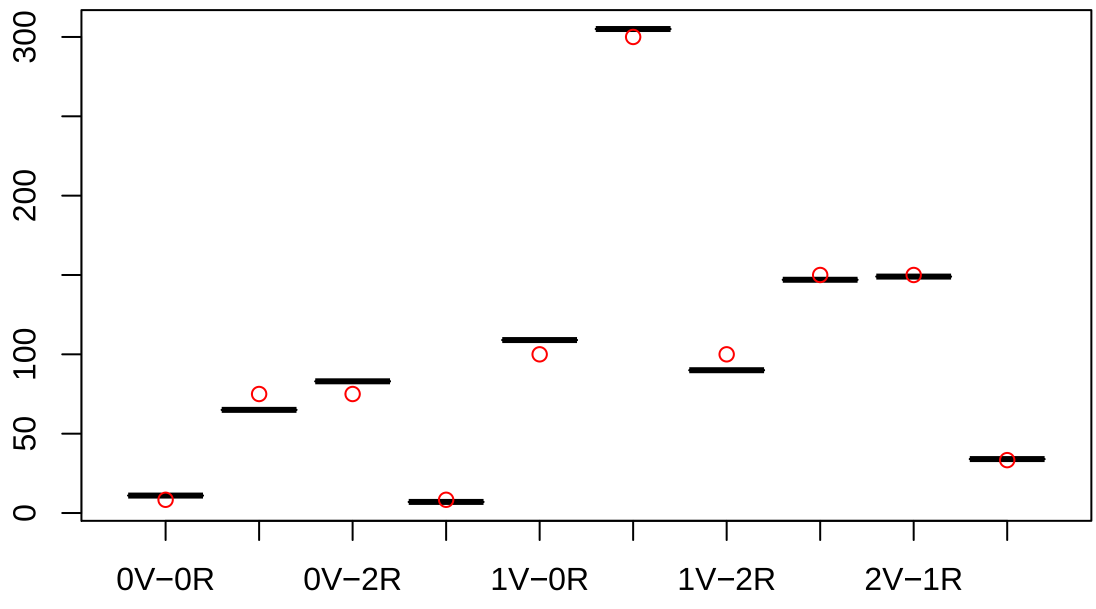


Lo que se observa en la tabla de frecuencias (y en el gráfico) es que el resultado más frecuente es extraer 1 bolilla verde y una roja entre las 3 bolillas extraídas.

Para hacer una mejor interpretación se calcularán las probabilidades de obtener cada resultado posible, y luego se multiplicará por las repeticiones (1000) para obtener una frecuencia teórica de resultados.

```
combinatorio <- function(n,r){
  if (r==0){
    resultado <- 1
  } else {
      resultado <- factorial(n) / ( factorial(n-r)*factorial(r) )
  }
  resultado
}


pxy <- rep(NA, 10)
for (i in 1:10){
```

```
  pxy[i] <- combinatorio(4, tabla_frecuencias$x[i]) * combinatorio(3, tabla_frecuencias$y[i]) * combina
}

fxy_teorica <- 1000*pxy
plot(tabla_frecuencias$resultado, tabla_frecuencias$frecuencia, xlab="resultado", ylab="frecuencia", ce
points(tabla_frecuencias$resultado, fxy_teorica, col="red")
```



Se observa que los resutados de la simulación y los resultados "teóricos" se acercan mucho.

Inciso d) Para cada valor observado x, calcular el promedio de los valores de y correspondientes.

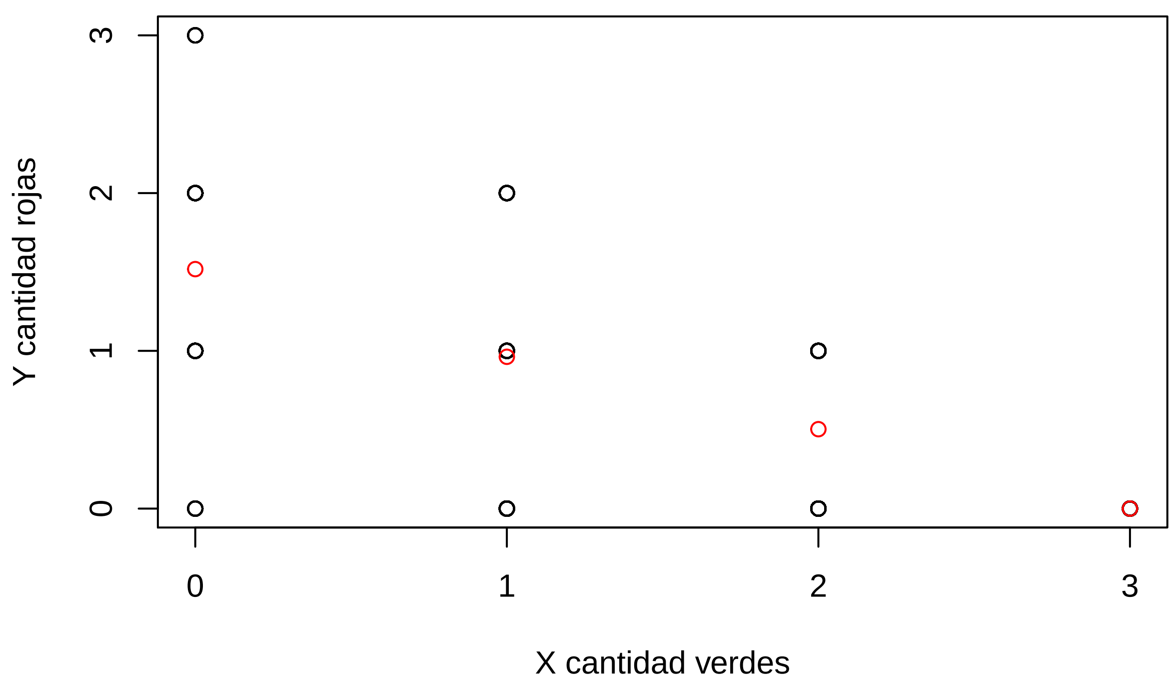
```
rango_x <- (0:3)
promedio_y <- c(mean(resultados[resultados$X==0 ,]$Y) , mean(resultados[resultados$X==1 ,]$Y) , mean(re
print(cbind(rango_x, promedio_y))
```

```
##      rango_x promedio_y
## [1,]       0  1.5180723
## [2,]       1  0.9623016
## [3,]       2  0.5033784
## [4,]       3  0.0000000
```

Inciso e) Gra

ficar los promedios de y en función de los valores observados x, sobre el gráfi co obtenido en el punto b.

```
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)
points(rango_x, promedio_y, col="red")
```



Inciso f) Superponer en el gráfi

co anterior la función de regresión $\phi(x) = \mathrm{E}[\mathrm{Y}|\mathrm{X} = \mathrm{x}]$. Concluir a partir de lo observado.

La función de probabilidad conjunta se obtiene por conteo de casos favorables sobre casos totales:

$P_{XY}(x, y) = \frac{C_{3,y}.C_{4,x}.C_{3,3-x-y}}{C_{10,3}}$

Cada una de las marginales tiene distribución hipergeométrica, por lo que:

$P_X(x) = \frac{C_{4,x}.C_{6,3-x}}{C_{10,3}}$

$P_Y(y) = \frac{C_{3,y}.C_{7,3-y}}{C_{10,3}}$

Por lo tanto, se obtiene la condicional:

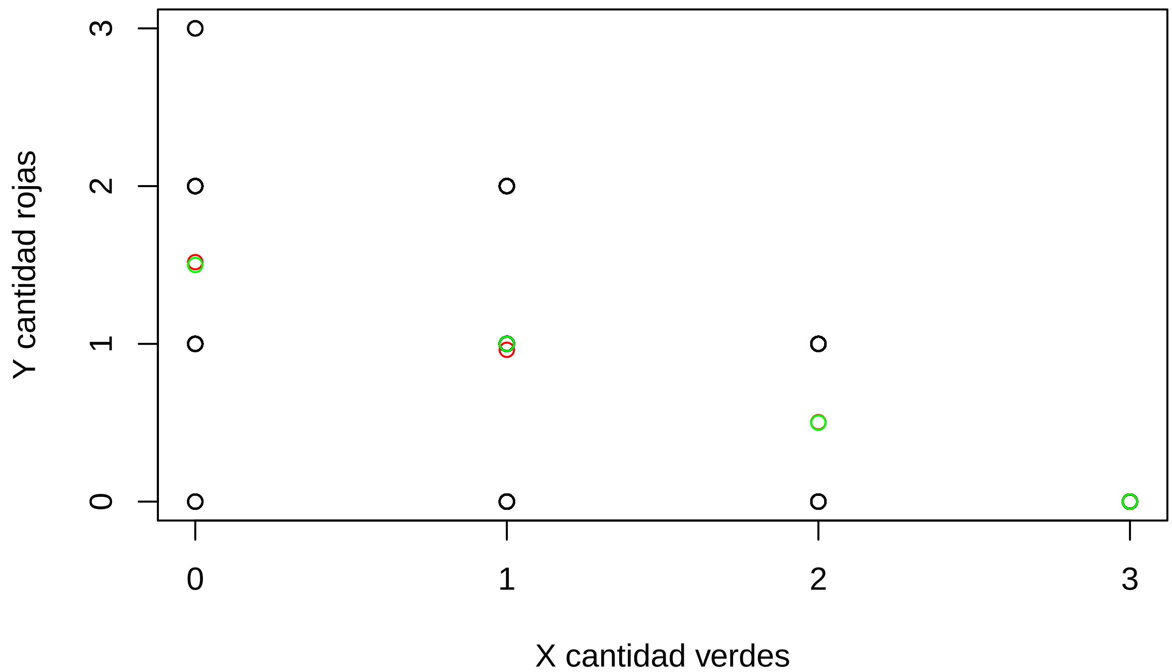
$P_{Y|X}(y) = \frac{P_{XY}(x,y)}{P_X(x)}$

$P_{Y|X}(y) = \frac{C_{3,y}.C_{3,3-y-x}}{C_{6,3-x}}$

Y a partir de la función de probabilidad conjunta, se obtiene la función de regresión. Para cada X del soporte:

$E_{[Y|X=x]} = \phi(x) = \Sigma y.P_{Y|X}(y)$

```
PYdadoX <- data.frame(X=rep(NA,10), Y=rep(NA,10), P_YdadoX=rep(NA,10))
contador <- 1
for (i in 0:3){
  for(j in 0:(3-i)){
    PYdadoX$X[contador] <- i
    PYdadoX$Y[contador] <- j
    PYdadoX$P_YdadoX[contador] <- combinatorio(3, j)*combinatorio(3,3-i-j) / combinatorio(6,3-i)
    contador <- contador+1
  }
}

EYdadoX <- data.frame(X=seq(0,3) , E_YdadoX=rep(0, 4))
contador <- 1

for (i in 0: 3 ){
  for(j in 0:(3-i)){
    EYdadoX$E_YdadoX[i+1] <- EYdadoX$E_YdadoX[i+1] + j*PYdadoX$P_YdadoX[contador]
    contador <- contador+1
  }
}

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)
points(rango_x, promedio_y, col="red")
points(EYdadoX$X, EYdadoX$E_YdadoX, col="green")
```



Se observa que los puntos obtenidos con la función de regresión coinciden casi exactamente con los promedios de Y de la simulación.

Ejercicio 2)

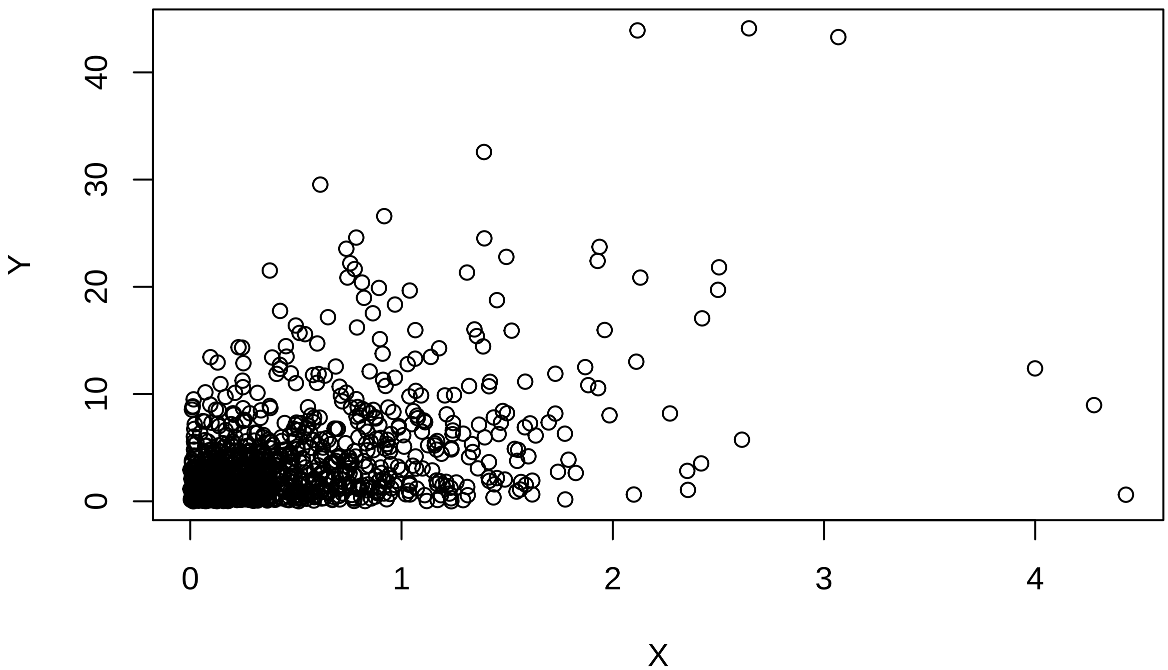
Realizar 1000 simulaciones del vector aleatorio (X; Y), cuya densidad conjunta es de la forma:

$f_{XY}(x, y) = \frac{1}{2x+1} e^{-2x-\frac{y}{4x+2}}.I(x > 0, y > 0)$

Inciso a) Gra

car x vs. y.

```
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)
```



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:

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

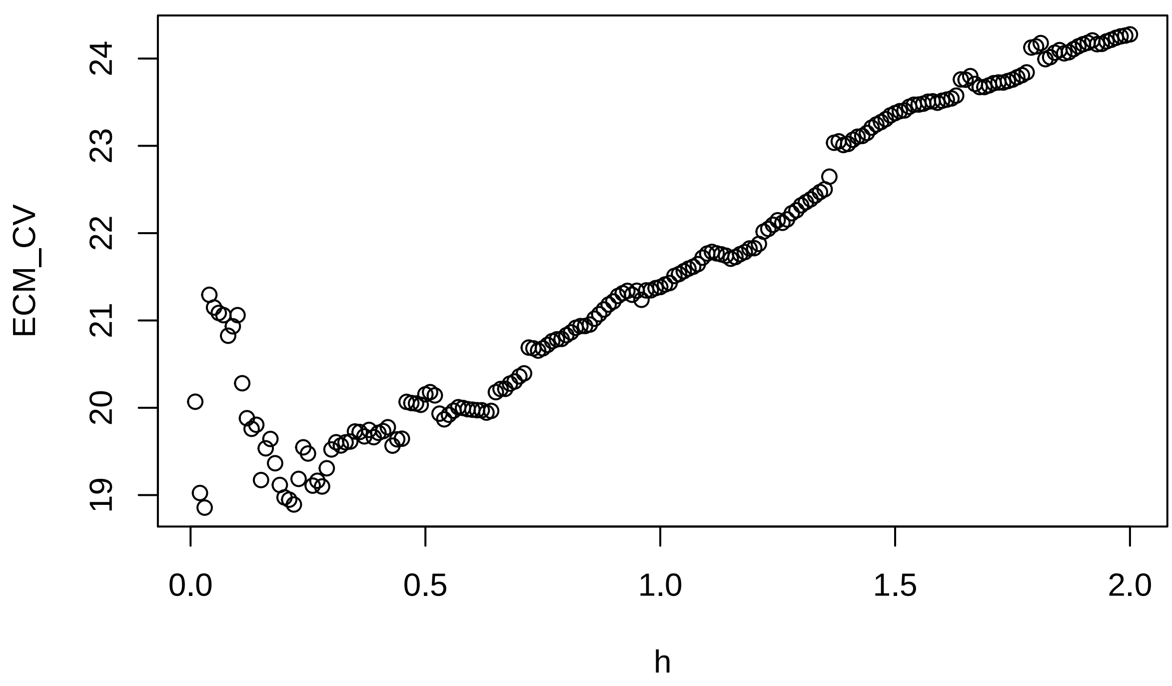
Para hacer la simulación, se simuló X con la distribución dada. Luego para cada X obtenido en la simulación, se simuló Y dado X.

Inciso b) Para cada valor observado x, tomar una ventana de (x-h; x+h), y calcular el promedio de los valores de Y para todas las observaciones que caen dentro de dicho intervalo. Elegir el valor de h que crea adecuado. Justi

car.

```
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)
```



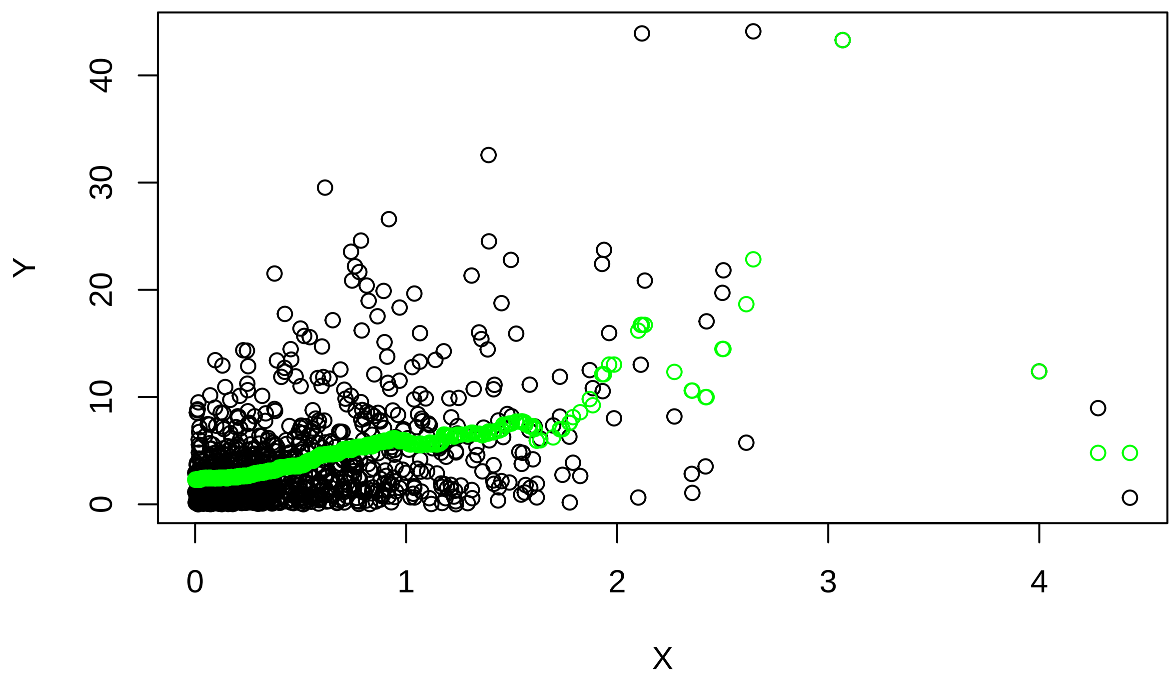
Para obtener la ventanan óptima se calculó el erros cuadrático medio para distintas h usando el método de convalidación cruzada. En cada X se obtuvo el promedio de los Y incluídos en la ventana h, dejando afuera el Y correspondiente al X del centro de la ventana (método "Leave one out").

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

Inciso c) Gra

ficar los promedios de y en función de los valores observados x, sobre el gráfico obtenido en el punto a.

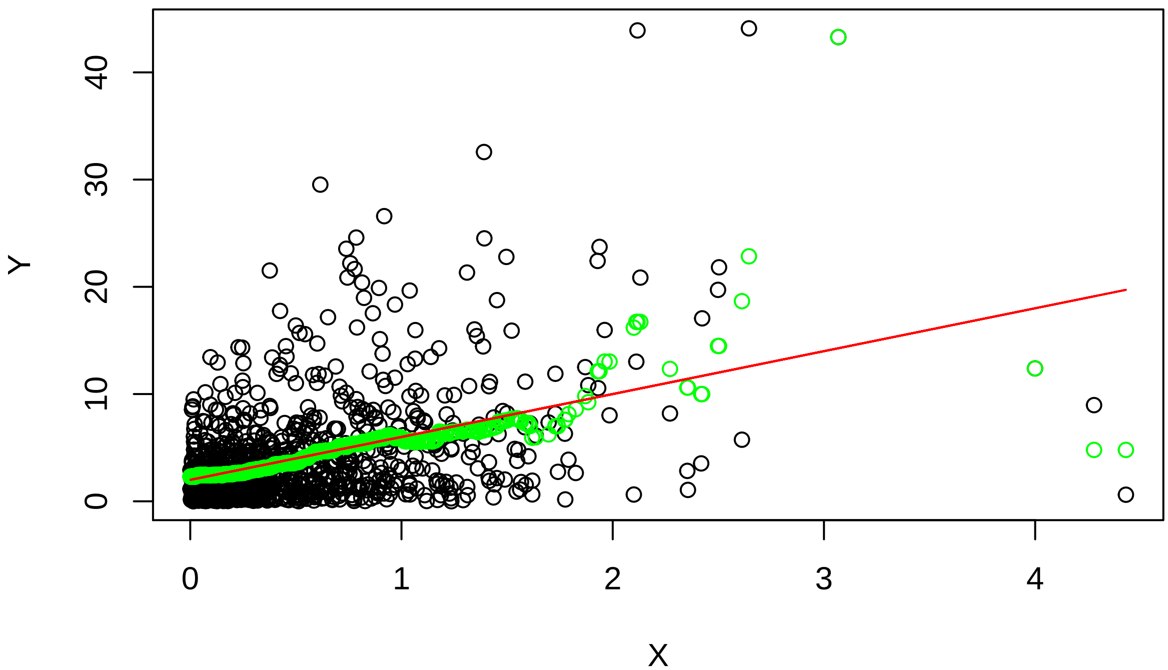
```
Y_promedio_optimo <- rep(NA, length(Y))
for(i in 1:length(Y)){
  Y_promedio_optimo[i] <- mean( XY[ abs(XY$X[i] - XY$X)<h_opt ,]$Y)
}

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



Inciso d) Superponer en el gráfi

co anterior la función de regresión $\phi(x) = \mathrm{E}[\mathrm{Y}|\mathrm{X} = \mathrm{x}]$.

```
E_YdadoX <- 4*X+2

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


Y
X
0
10
20
30
40
0
1
2
3
4