Tarea 2 - Estadística para Data Science

Nicole Lastra Quiroz Ari Romero Garrido Karen Romero Garrido

Teorema del Límite Central

1. Considere la siguiente densidad uniforme:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a < x < b \\ 0 & \text{EOC} \end{cases}$$

Elija dos valores cualquiera para a y b y luego genere 200 muestras de tamaño 30, 50 y 100. Para lo anterior, se pide:

a) Las medias muestrales referidas a los distintos tamaños muestrales.

Usaremos una población de tamaño `m`, con valores distribuidos uniformemente entre `a` y `b`.

```
n <- 200
m <- 1000
a <- 10
b <- 100

set.seed(2021)
pob<-runif(m, a, b)
```

Luego se obtienen `n`=200 muestras de tamaños 30, 50 y 100 y se les calula el valor medio.

```
m30  <-vector("numeric", length = n)
m50  <-vector("numeric", length = n)
m100 <-vector("numeric", length = n)

for (i in 1:n ) {
  m30[i]  = mean(sample(pob,  30, replace = TRUE))
  m50[i]  = mean(sample(pob,  50, replace = TRUE))
  m100[i] = mean(sample(pob, 100, replace = TRUE))
}
```

Finalmente, la media para cada tamaño muestral es

```
mean(m30)
```

```
## [1] 54.31003
```

```
mean(m50)
```

```
## [1] 54.31064
```

```
mean(m100)
```

```
## [1] 54.44885
```

b) Una distribución de frecuencias de las medias muestrales para cada uno de los tamaños señalados.

Se utiliza la biblioteca `fdth` para obtener la distribución de frecuencias para los valores dados, con sus intervalos, sus frecuencias y otros datos adicionales.

```
d30  <- fdt(m30,  breaks="Sturges")
d50  <- fdt(m50,  breaks="Sturges")
d100 <- fdt(m100, breaks="Sturges")
```

Al observar los valores ya podemos tener una primera idea de que las medias se distribuyen de forma normal.

```
d30
```

```
##     Class limits  f   rf rf(%)  cf cf(%)
##  [43.233,45.811)  4 0.02   2.0   4   2.0
##  [45.811,48.389) 14 0.07   7.0  18   9.0
##  [48.389,50.967) 20 0.10  10.0  38  19.0
##  [50.967,53.545) 50 0.25  25.0  88  44.0
##  [53.545,56.123) 47 0.23  23.5 135  67.5
##  [56.123,58.701) 31 0.16  15.5 166  83.0
##  [58.701,61.279) 23 0.12  11.5 189  94.5
##  [61.279,63.857)  9 0.04   4.5 198  99.0
##  [63.857,66.435)  2 0.01   1.0 200 100.0
```

```
d50
```

```
##     Class limits  f   rf rf(%)  cf cf(%)
##  [43.594,45.977)  4 0.02   2.0   4   2.0
##   [45.977,48.36)  7 0.04   3.5  11   5.5
##   [48.36,50.743) 23 0.12  11.5  34  17.0
##  [50.743,53.127) 33 0.16  16.5  67  33.5
##   [53.127,55.51) 61 0.30  30.5 128  64.0
##   [55.51,57.893) 42 0.21  21.0 170  85.0
##  [57.893,60.276) 18 0.09   9.0 188  94.0
##  [60.276,62.659) 11 0.06   5.5 199  99.5
##  [62.659,65.042)  1 0.00   0.5 200 100.0
```

```
d100
```

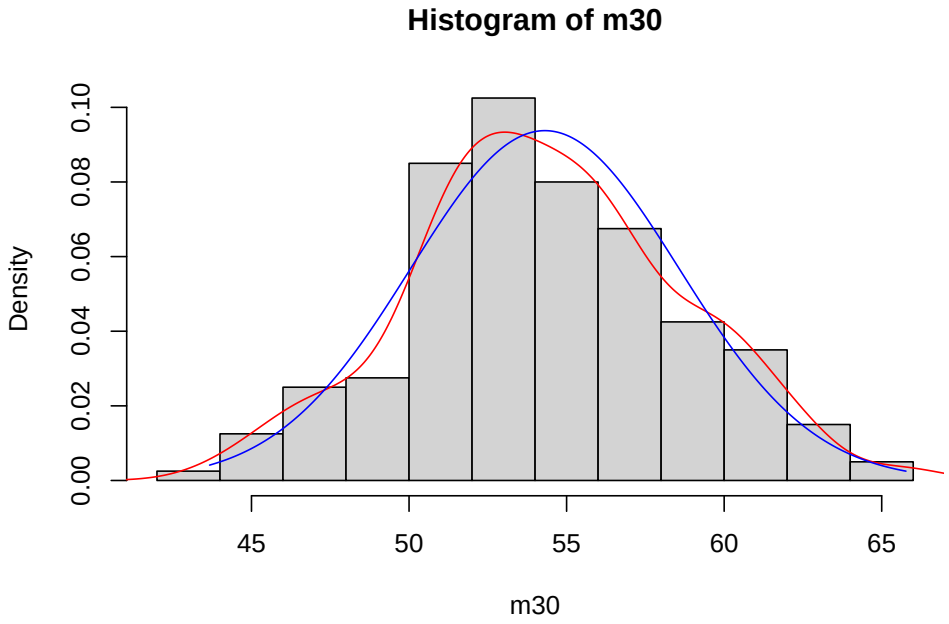
```
##     Class limits  f   rf rf(%)  cf cf(%)
##   [46.76,48.324)  3 0.01   1.5   3   1.5
##  [48.324,49.888)  7 0.04   3.5  10   5.0
##  [49.888,51.451) 18 0.09   9.0  28  14.0
##  [51.451,53.015) 28 0.14  14.0  56  28.0
##  [53.015,54.579) 45 0.22  22.5 101  50.5
##  [54.579,56.143) 42 0.21  21.0 143  71.5
##  [56.143,57.706) 37 0.18  18.5 180  90.0
##   [57.706,59.27) 13 0.06   6.5 193  96.5
##   [59.27,60.834)  7 0.04   3.5 200 100.0
```

c) Un histograma para las medias muestrales con sus respectivos ajustes de curvas.

Se ajustan los histogramas a densidad y se comparan las gráficas con las distribuciones normales dadas por las medias y las desviaciones estándar de cada vector de 200 medias.

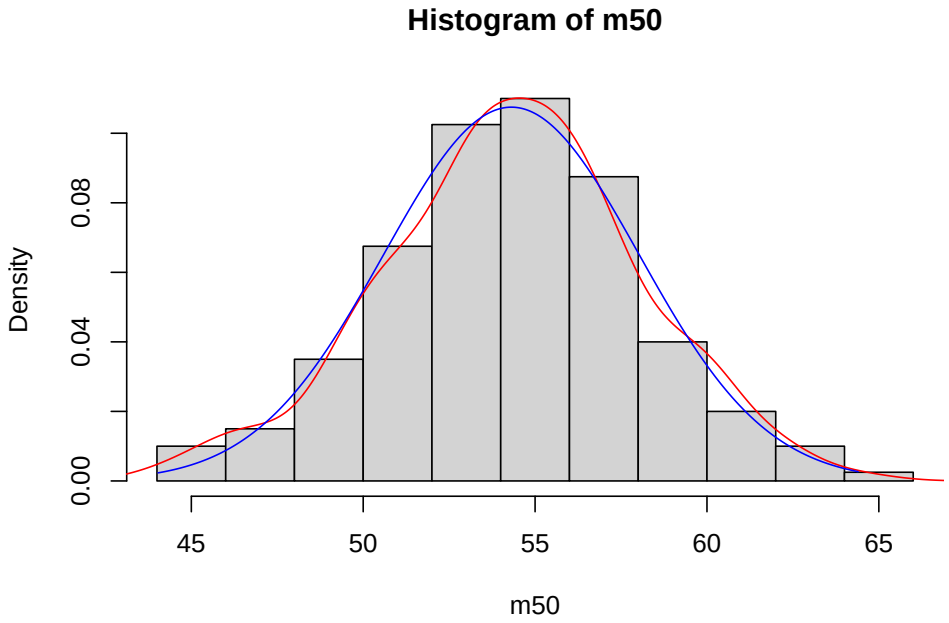
Para las muestras de n=30.

```
dz <- seq(min(m30), max(m30), 0.01)
hist(m30, freq=F)
lines(density(m30),col="Red")
lines(dnorm(dz, mean(m30), sd(m30))~dz, type="l",col="blue")
```



Para las muestras de n=50.

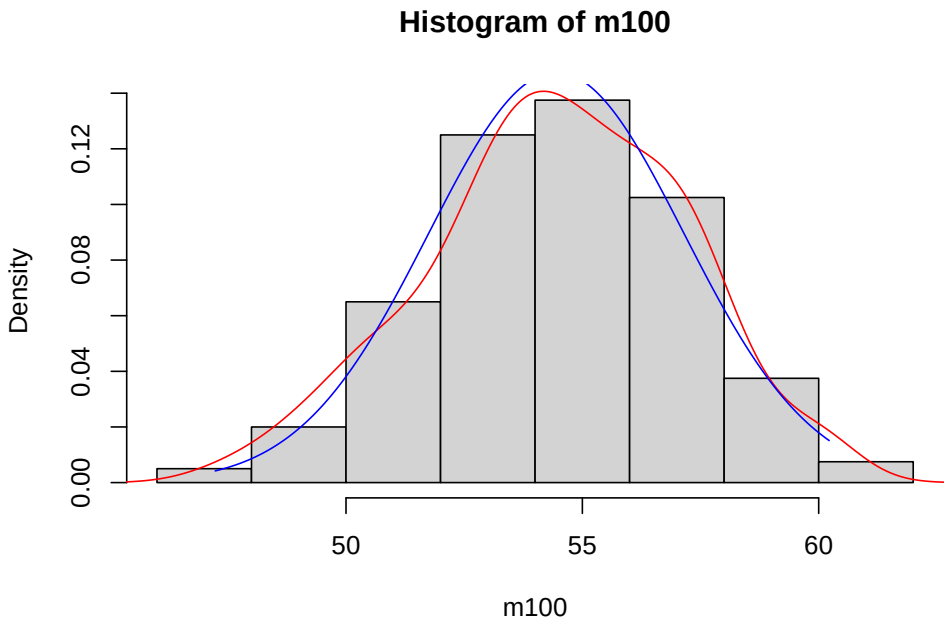
```
dz <- seq(min(m50), max(m50), 0.01)
hist(m50, freq=F)
lines(density(m50),col="Red")
lines(dnorm(dz, mean(m50), sd(m50))~dz, type="l",col="blue")
```



Para las muestras de n=100.

```
dz <- seq(min(m100), max(m100), 0.01)
```

```r
hist(m100, freq=F)
lines(density(m100),col="Red")
lines(dnorm(dz, mean(m100), sd(m100))~dz, type="l",col="blue")
```



d) Efectúe una comparación entre las medias de las medias muestrales para los distintos tamaños con respecto a la verdadera media de la distribucion dada.

Anteriormente, se obtuvo la media para cada tamaño muestral:

```r
mean(m30)
```

```
## [1] 54.31003
```

```r
mean(m50)
```

```
## [1] 54.31064
```

```r
mean(m100)
```

```
## [1] 54.44885
```

La media real de los datos generados es:

```r
mean(pob)
```

```
## [1] 54.46718
```

Que es el resultado esperable, a mayor n, menor diferencia entre las medias.

e) Conclusiones relativas a la aplicación del teorema del límite central.

Al observar los resultados anteriores, podemos concluir que para la distribución uniforme es aplicable el teorema del límite central, ya que pudimos observar en sus gráficas de densidad que sus medias ajustadas tienen una distribución normal, que es progresivamente mejor a medida que aumenta el tamaño muestral.

Distribuciones Conjuntas

Sparragowsky Associates realizó un estudio sobre el tiempo en minutos para atender a un cliente en la ventanilla de su automóvil en cierto restaurante de comida rápida X, además, el tiempo que tarda un cliente en cancelar su compra y retirarse edl local Y. El estudio mostraba que estas dos variables están relacionadas conjuntamente de acuerdo a la siguiente función:

$$f_{XY}(x,y) = \frac{\tau}{2} e^{-\frac{1}{2}((x-2)^2+y)} \quad -\infty < x < \infty, y > 0$$

con

$$\tau = \frac{1}{\sqrt{2\pi}}$$

1. Muestre que la distribución marginal de X e Y son:

$$f_X(x) = \tau e^{-\frac{1}{2}(x-2)^2} \quad -\infty < x < \infty$$

$$f_Y(y) = \frac{1}{2} e^{-\frac{1}{2}y} \quad y > 0$$

En el caso de la distribución marginal f_X, se integra f_{XY} respecto de y y se evalúa en el intervalo $]0,\infty[$, ya que se cumple que y>0:

$$f_X(x) = \int_0^\infty f_{XY}(x,y)dy$$

De acuerdo a la definición:

$$f_X(x) = \int_0^\infty \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2} e^{-\frac{1}{2}y} dy$$

Separando la parte constante respecto de y:

$$f_X(x) = \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2} \int_0^\infty e^{-\frac{1}{2}y} dy$$

Integrando

$$f_X(x) = \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2} \left[-2e^{-\frac{1}{2}y}\right]_0^\infty$$

Evaluando en los límites de integración

$$f_X(x) = \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2} \left[-2e^{-\frac{1}{2}\infty}\right]_{=0} - \left[-2e^{-\frac{1}{2}0}\right]_{=-2}$$

Por lo que finalmente se demuestra que:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}, \quad -\infty < x < \infty$$

Análogamente, para f_Y:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx$$

De acuerdo a la definición

$$f_Y(y) = \int_{-\infty}^{\infty} \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2} e^{-\frac{1}{2}y} dx$$

Separando la parte constante respecto de x:

$$f_Y(y) = \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-2)^2} dx$$

Se resolverá la integral respecto de x:

$$(*) = \int e^{-\frac{1}{2}(x-2)^2} dx$$

Usando la sustitución $u = x - 2$ se obtiene

$$(*) = \int e^{-\frac{1}{2}u^2} du$$

Usando la sustitución $v = u/\sqrt{2}$:

$$(*) = \sqrt{2}\int e^{-v^2} dv = \sqrt{2}\frac{\sqrt{pi}}{2} Erf(v)$$

Podemos volver a escribir la variable original utilizando las sustituciones $v = u/\sqrt{2}$ y $u = x - 2$ y evaluar en sus límites de integración, utilizando el hecho de que $Erf(-\infty) = -1$ y $Erf(\infty) = 1$

$$(*) = \sqrt{2}\frac{\sqrt{\pi}}{2}\left[Erf\left(\frac{x-2}{\sqrt{2}}\right)\right]_{-\infty}^{\infty} = \sqrt{2\pi}$$

Volviendo a la expresión original para f_Y se demuestra que:

$$f_Y(y) = \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y}\sqrt{2\pi}$$

$$f_Y(y) = \frac{1}{2} e^{-\frac{1}{2}y} \quad , y > 0$$

2. Muestre que X e Y variables aleatorias independientes

Para demostrarlo, se debe cumplir que:

$$f_{XY}(x,y) = f_X(x) \cdot f_Y(y) \quad , \forall (x,y)$$

Desarrollaremos utilizando las expresiones obtenidas en el paso anterior:

$$f_X(x) \cdot f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2} \cdot \frac{1}{2} e^{-\frac{1}{2}y} = \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2}((x-2)^2+y)}$$

Utilizando la sustición

$$\tau = \frac{1}{\sqrt{2\pi}}$$

Reemplazamos en la expresión anterior:

$$\frac{\tau}{2} e^{-\frac{1}{2}((x-2)^2+y)} = f_{XY}(x,y) \quad -\infty < x < \infty, y > 0$$

Por lo que se demuestra que las variables aleatorias son independientes.

3. De acuerdo al estudio, el tiempo de eficiencia (en minutos) de los trabajadores de la ventanilla del restaurante esta dado por la expresión $W = 2X - 0.25Y$. Encuentre $E(W)$ y $Var(W)$.

Para resolver, se debe identificar el valor de la esperanza y varianza de X e Y a partir de sus expresiones de f_X y f_Y.

En el caso de f_X podemos observar que:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-2)^2}{1^2}} \quad , -\infty < x < \infty$$

Por lo que podemos identificar una distribución normal con $\mu = 2$ y varianza $\sigma^2 = 1$. Análogamente, para F_Y

$$f_Y(y) = \frac{1}{2} e^{-\frac{1}{2}y} \quad , y > 0$$

Podemos identificar una distribución exponencial con $\lambda = \frac{1}{2}$, lo que nos permite conocer expresiones para su esperanza y varianza:

$$E(y) = \frac{1}{\lambda} = 2$$

$$Var(y) = \frac{1}{\lambda^2} = 4$$

Dado que anteriormente se demostró que X e Y son independientes, podemos establecer que:

$$E(W) = E(2X - 0.25Y) = 2E(X) - 0.25E(Y) = 3.5$$

$$Var(W) = Var(2X - 0.25Y) = 2^2 Var(X) - 0.25^2 Var(Y) = 3.75$$

Estimadores Puntuales

Recordemos que si $X_1, X_2, ..., X_n$ es una muestra aleatoria simlpe extraída de una población X con distribución de probabilidad dada por $f(x|\theta)$, siendo θ uno (o más) parámetro poblacional desconocido, entonces la función de verosimilitud de la muestra se define mediante la expresión:

$$L(\theta) = f(x_1, x_2, ...x_n|\theta)$$

Ahora bien, como cada x_i es una realización de la v.a. X_i y estas son independientes e idénticamente distribuidas (iid), la función de verosimilitud puede escribirse como sigue:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Ahora para encontrar el estimador máximo verosímil (EMV) para el parámetro poblacional desconocido θ, debemos resolver el siguiente problema de maximización:

$$Max_{\hat{\theta}} L(\theta) \quad \text{o bien} \quad Max_{\hat{\theta}} ln(L(\theta))$$

Sin embargo, computacionalmente hablando, el problema que realmente se resuelve es minimizar el negativo de la función de log-verosimilitud.

$$-ln(L(\theta)) = -ln(\prod_{i=1}^{n} f(x_i|\theta)) = \sum_{i=1}^{n} lnf(x_i|\theta)$$

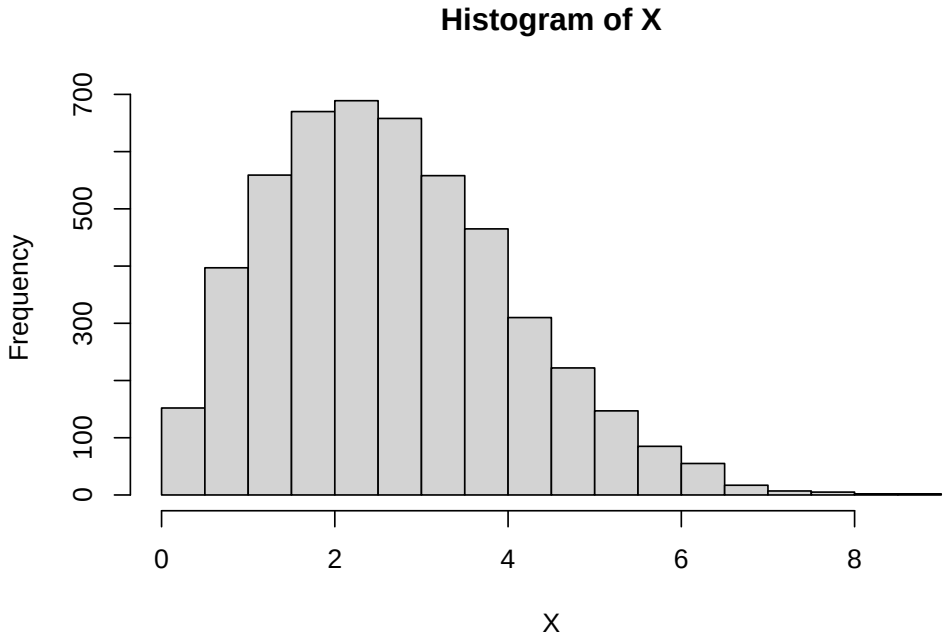
Lo anterior es posible gracias a que $ max(Z) = min(-Z)$, siendo Z una función objetivo cualquiera.

1. Genere una v.a. X relativa a una distribución Weibull tal que $X $\sim Weibull(\alpha = 3, \beta = 2)$ de tamaño 5000 para esto use el comando `rweibull()`.

```
m <- 5000
a <- 3
b <- 2
set.seed(2021)
X<-rweibull(m, b, a)
```

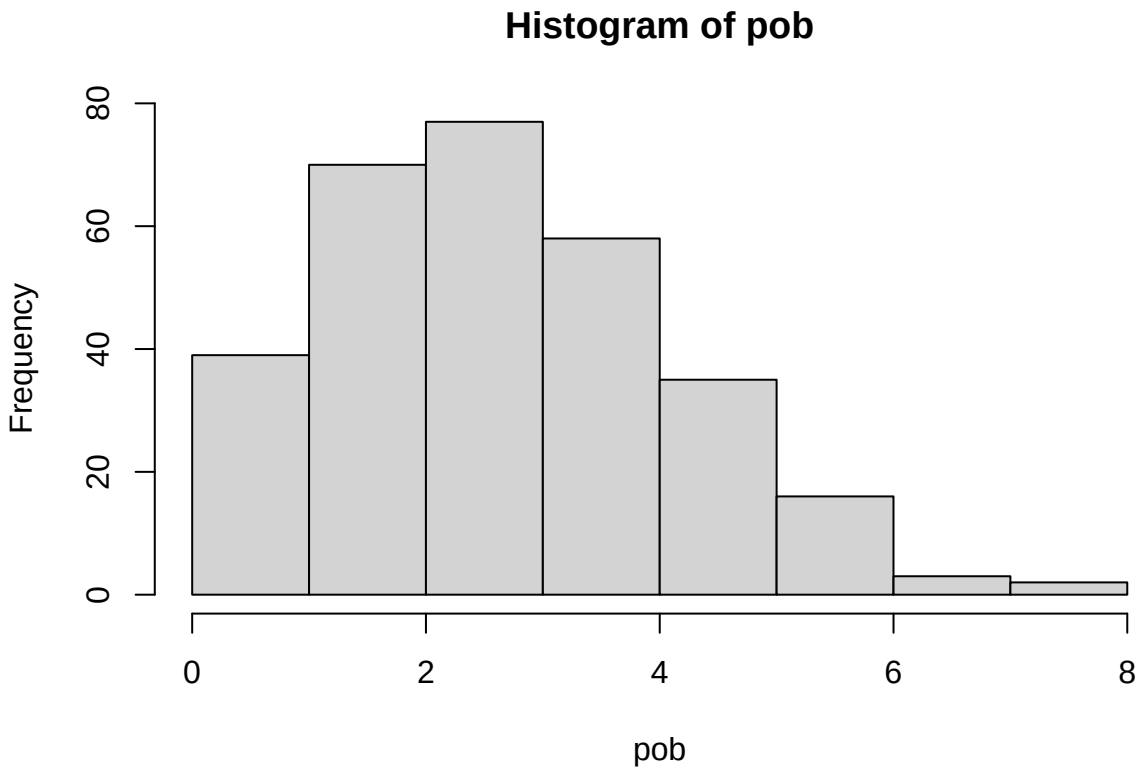
2. Grafique dicha v.a. con el uso de un histograma.

```
hist(X)
```



3. Extraiga de la población anterior una muestra de tamaño $n = 300$.

```
pob <- sample(X, 300, replace=TRUE )
hist(pob)
```



4. Construya un EMV de la muestra y encuentre los parámetros α y β.

Se utilizará la función negativa de máxima verosimilitud, dada en el enunciado para luego utilizar la función de estimación `nlm` para minimizar dicha función.

```
emv_neg <- function ( argv ) {
  -sum(log(dweibull( pob ,argv[1], argv[2] )))
}
p<- nlm(emv_neg, c(1,1))
p
```

```
## $minimum
## [1] 521.9561
##
## $estimate
## [1] 1.893762 2.998049
##
## $gradient
## [1] 2.281226e-06 7.508214e-06
##
## $code
## [1] 1
##
## $iterations
## [1] 13
```

Como podemos ver, los valores para los parámetros obtenidos son muy cercanos a los establecidos originalmente:

para α

```
p$estimate[2]
```

```
## [1] 2.998049
```

para β

```
p$estimate[1]
```

```
## [1] 1.893762
```