Estadistica - Ejercicio 1

Cedric Prieels

Introduccion

Este ejercicio tiene como objetivos principales la familiarizacion con la herramienta estadística que es R y el estudio de unos distribuciones estadísticos de datos. Lo importante en este ejercicio consiste en estudiar las tres diferentes distribuciones que tenemos, y de calcular sus parámetros importantes (como la media, la mediana, la moda y la desviación estandar), para después dibujar sus histogramas de la mejor manera posible y intentar identificar y caracterizar cada una de estas distribuciones.

Metodología

Tenemos tres conjuntos de datos a nuestra disposición (datos1.dat, datos2.dat y datos3.dat). Lo primero que hay que hacer es manipular un poco de código R para abrir y leer estos conjuntos de puntos, para estudiar al ojo los datos a nuestra disposición. Una vez hecho, se usan directamente unas funciones básicas de R para medir de diferentes maneras los valores centrales (media, medianana y moda) y la dispersión (desviación estándar) de cada distribución.

Lo siguiente consiste en representar los datos que tenemos en histogramas, siguiendo las reglas vistas en clase para evitar todo problema estadístico a la hora de representarlos. Hay que elegir correctamente por ejemplo el numero de bines usados y sus valores centrales para representar de la mejor manera posible nuestros conjuntos de datos.

Por fin, se trata de interpretar los resultados precedientes para intentar encontrar el tipo de distribución en cada caso, y para caracterizarlas.

Resultados

Como explicado antes, para estudiar conjuntos de datos, lo primero que hay que hacer es abrir los 3 ficheros a nuestra disposición y que contienen tablas con todos los puntos experimentales medidos. Además se introducen variables que se definen como el numero de elementos de cada fichero (va a ser útil para pintar los histogramas).

```
setwd('/Users/ced2718/Documents/Universite/Estadistica/Ejercicio_1/')
data1 <- read.table("dat1.dat", header=FALSE)
data2 <- read.table("dat2.dat", header=FALSE)
data3 <- read.table("dat3.dat", header=FALSE)

ndata1 <- nrow(data1)
ndata2 <- nrow(data2)
ndata3 <- nrow(data3)
```

Se calcula primero la media, la mediana y la desviación estandar de la distribución del fichero 1. Para eso, se lee la primera (y única) columna del fichero, y se usan funciones de R directamente para calcular estos parámetros. Se pone el parámetro na.rm a true el las funciones para quitar directamente valores que no serián números y que podrían influir en los resultados finales.

```
data1 <- data1[,1]
mediaData1 <- mean(data1, na.rm = TRUE)
medianaData1 <- median(data1, na.rm = TRUE)
sdDesviacionData1 <- sd(data1, na.rm = TRUE)
```

Calcular la moda de la distribucion es un poco más difícil porque R no tiene función escrita para calcular directamente este parámetro. Después de una pequeña búsqueda, encontré en [Github] (https://gist.github.com/jmarhee/8530768) una función escrita por otro usuario que ayuda en este caso.

```
moda <- function(x) {
    if (is.numeric(x)) {
        x_table <- table(x)
        return(as.numeric(names(x_table)[which.max(x_table)]))
    }
}

modaData1 <- moda(data1)
```

Después, se vuelven a repetir los pasos precedientes para las distribuciones 2 y 3, que vienen respectivamente de los ficheros dat2.dat y dat3.dat, y que siguen el mismo formato que el fichero dat1.dat (una sola columna con los datos).

```
data2 <- data2[,1]
mediaData2 <- mean(data2, na.rm = TRUE)
medianaData2 <- median(data2, na.rm = TRUE)
sdDesviacionData2 <- sd(data2, na.rm = TRUE)
modaData2 <- moda(data2)
```

```
data3 <- data3[,1]
mediaData3 <- mean(data3, na.rm = TRUE)
medianaData3 <- median(data3, na.rm = TRUE)
sdDesviacionData3 <- sd(data3, na.rm = TRUE)
modaData3 <- moda(data3)
```

Ya se conocen todos los parámetros importantes de las distribuciones, así que se puede empezar a mirarlas un poco más en detalle antes de pintar sus histogramas.

```
resultados <- matrix(c(mediaData1, medianaData1, modaData1, sdDesviacionData1, mediaData2,
                       medianaData2, modaData2, sdDesviacionData2,mediaData3, medianaData3,
                       modaData3, sdDesviacionData3), ncol=4, byrow=TRUE)

colnames(resultados) <- c("Media","Mediana","Moda","Desviación")
rownames(resultados) <- c("data1","data2","data3")

rtab <- as.table(resultados)
head(rtab)
```

```
##           Media   Mediana      Moda Desviación
## data1 1.9863000 1.9759190 1.5133979  1.4147482
## data2 1.9863000 2.0000000 1.0000000  1.4147482
## data3 1.9863000 2.0019537 0.1635149  1.4147482
```

Lo primero que se puede ver es que las 3 distribuciones tienen exactamente el mismo valor medio y la misma desviación estándar. Tienen además todas un valor de mediana no exactamente igual, pero que tiene un valor casi identico en todos los casos. Las tres distribuciones se diferecian entonces solamente por la moda.

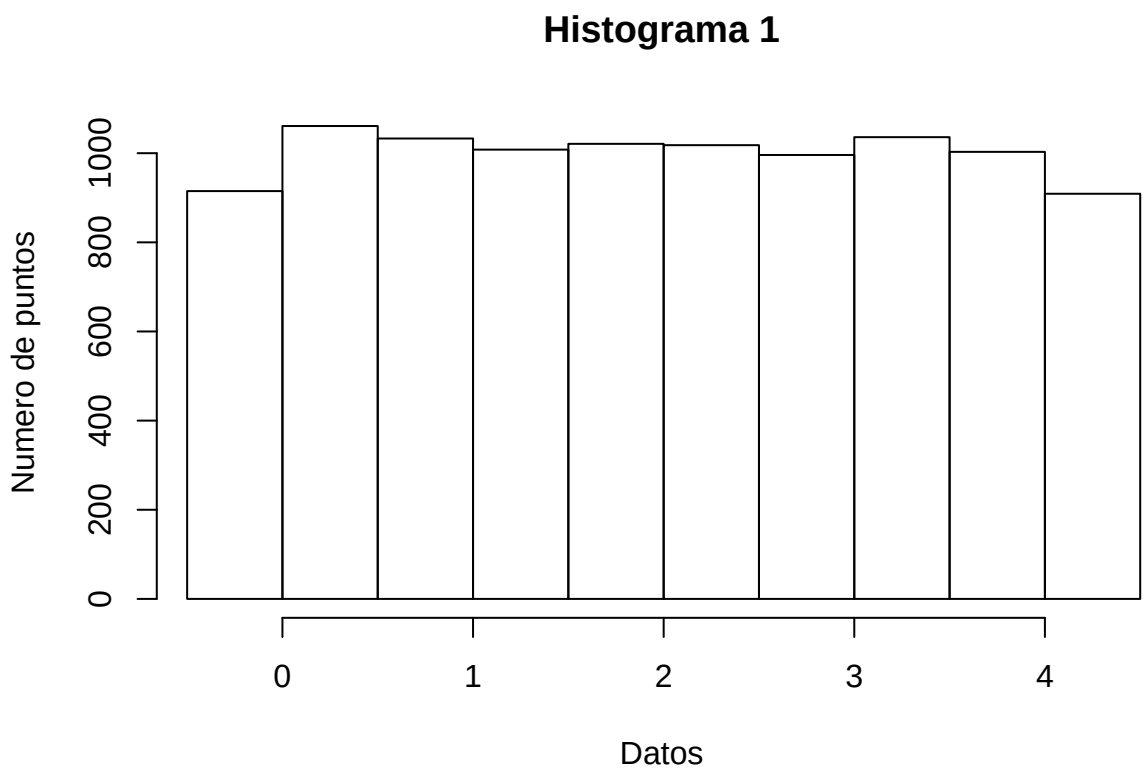
Como visto en clase, ninguna de estas 3 distribuciones es completamente simetrica (lo es únicamente cuando los valores de media, madiana y moda son iguales). Además, una diferencia significativa entre el valor de mediana y de media es generalmente el signo de una distribución que tiene una cola no-despreciable, pero no parece que sea el caso en este ejercicio.

Histogramas

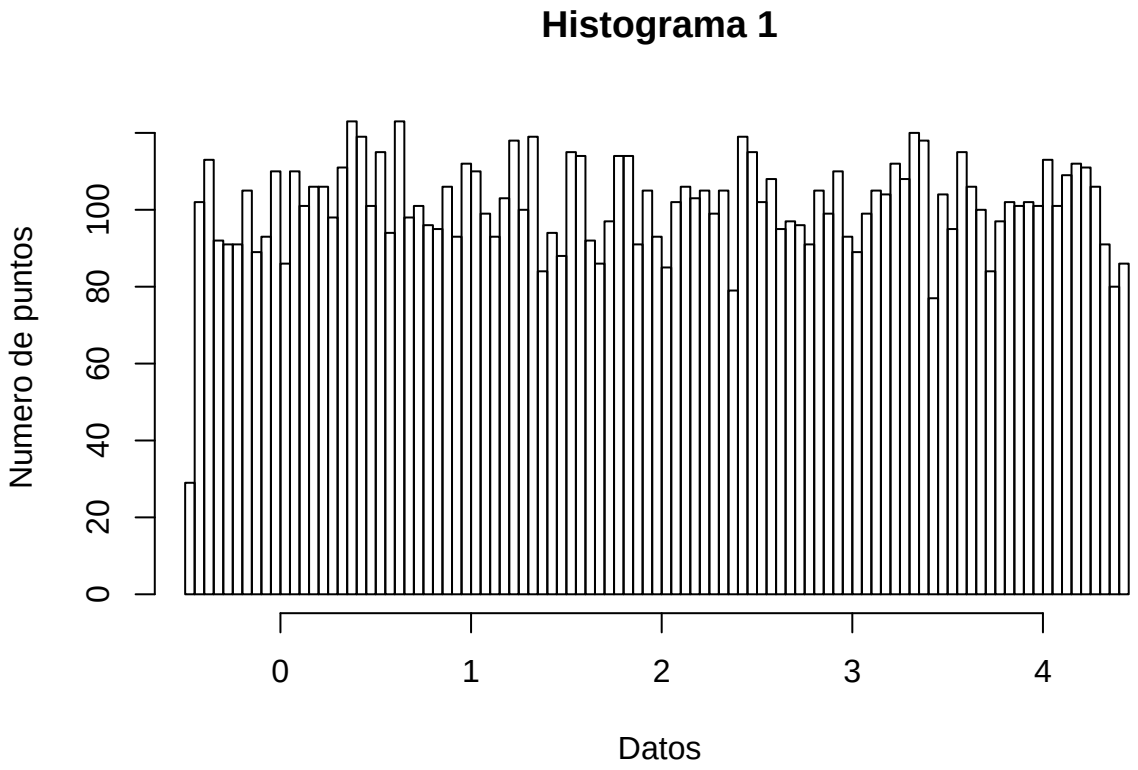
Histograma 1

Primero, se dibuja el histograma de la primera distribución de datos usando el comando hist, sin nada más. Después se usan además las opciones de la función hist de R para pintar correctamente el histograma, cambiando por ejemplo el números de bines (se pone igual a la raiz cuadrada del número de datos).

```
hData1 <- hist(data1,freq=TRUE, main="Histograma 1",
               xlab="Datos", ylab="Numero de puntos",
               xlim=c(min(data1),max(data1)))
```



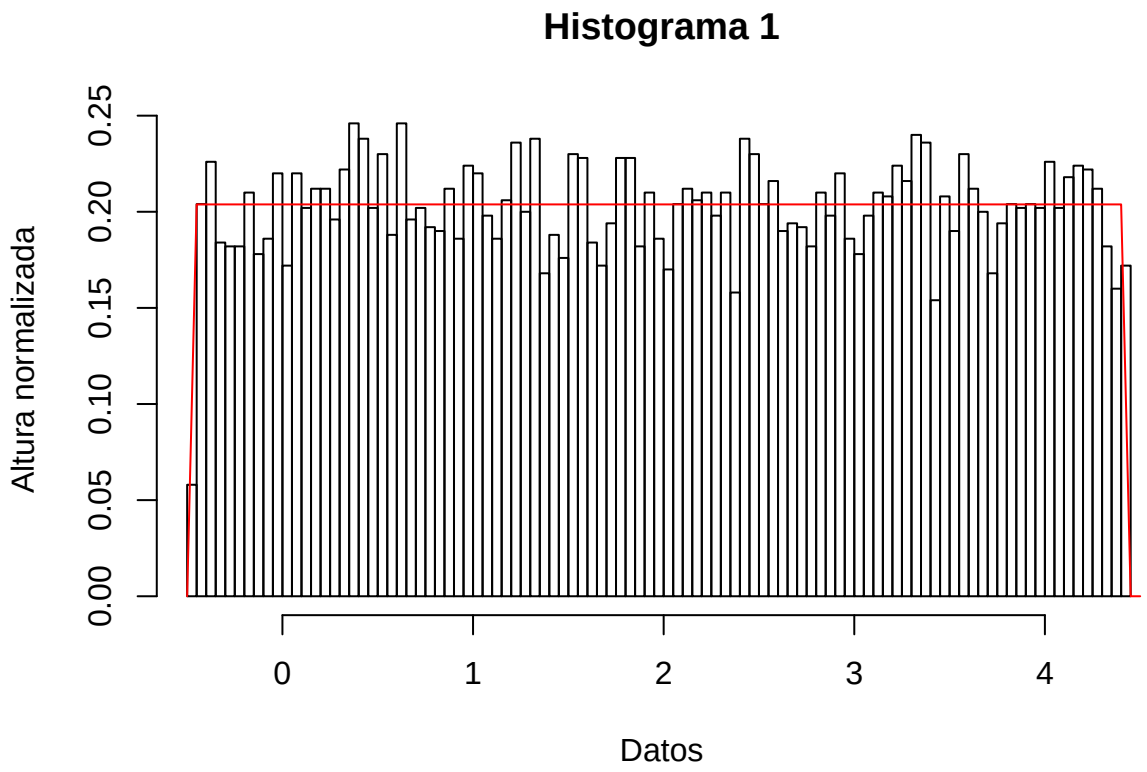

```
hData1bis <- hist(data1, breaks=sqrt(ndata1), freq=TRUE, main="Histograma 1",
                  xlab="Datos", ylab="Numero de puntos",
                  xlim=c(min(data1),max(data1)))
```



En los dos casos, se puede ver que la distribución que tenemos en el fichero dat1.dat parece ser una distribución uniforme, porque la altura es la misma en cada bin (considerando las fluctuaciones estadísticas). Como el fichero dat1.dat no está hecho de números enteros, no hay además que preocuparse mucho del valor central de cada bin.

También se puede representar el mismo histograma con una altura normalizada, y con una linea roja que corresponde a una distribución uniforme entre -0.5 y 4.5 (mínimo y máximo de la distribución, como se enseñara después).

```
hData1ter <- hist(data1, breaks=sqrt(ndata1), freq=FALSE, main="Histograma 1",
                  xlab="Datos", ylab="Altura normalizada",
                  xlim=c(min(data1),max(data1)))
curve(dunif(x, min(data1), max(data1)), from=-0.5, to=4.5, col="red", add=TRUE)
```

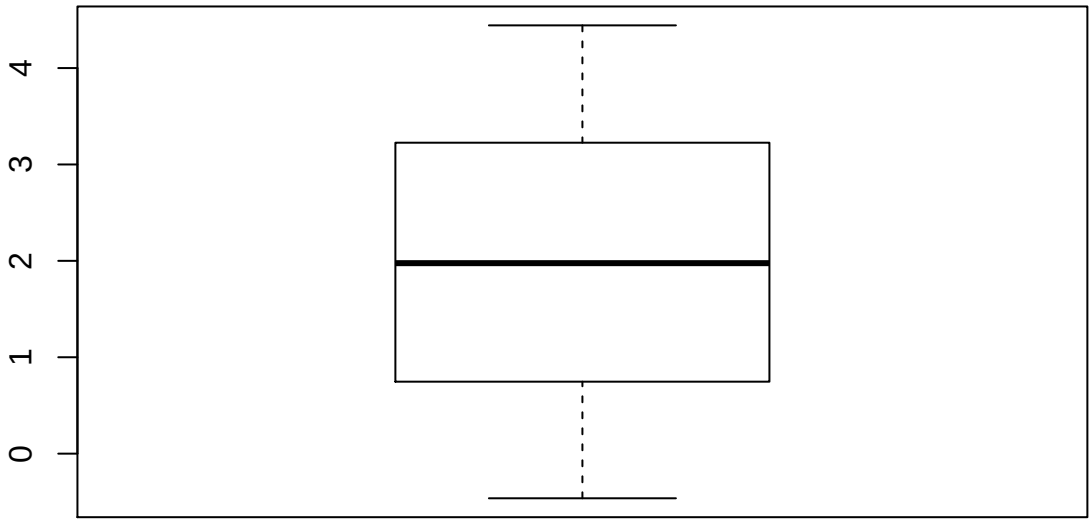


Por fin, se representa el "summary" y el "boxplot"" de esta distribución. Esta función y este plot nos dan directamente los valores de mínimo, de máximo y de los cuartiles más importantes, y los eventuales puntos rechazados (no parece que haya ninguno en este caso).

```
summary(data1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.4630  0.7467  1.9760  1.9860  3.2260  4.4430
```

```
boxplot(data1)
```



Este plot parece correcto porque volvemos a encontrar el valor del segundo cuartil (la mediana) a un valor cerca de 2, mientras el Q1 vale más o menos 1 y el Q3 vale más o menos 3 (como lo esperamos para una distribución uniforme).

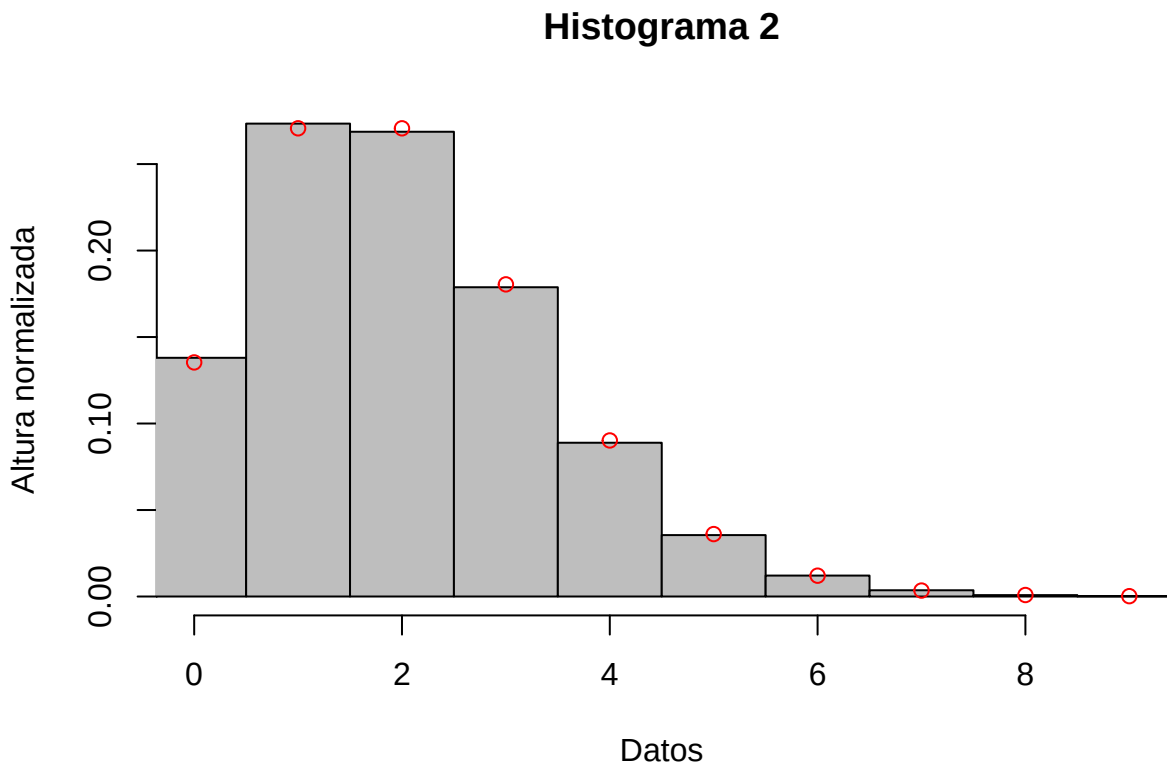
La única discusión que queda es saber porque parecía en la parte precediente que todas las distribuciones no eran simétricas, aunque una distribución uniforme lo es claramente. Es porque el valor de la moda no es un parámetro bien definido en este caso, porque todos los bines tienen en teória la misma altura, y porque el bin que tiene la altura más altura puede cambiar en función de número de bines (se ve en los plots precediente).

Histograma 2

Después, se repite exactamente el mismo proceso con los datos del fichero número 2. Como se ve en este fichero, los datos que tenemos son números enteros, así que como lo vimos en clase, hay que tener cuidado con la representación del histograma.

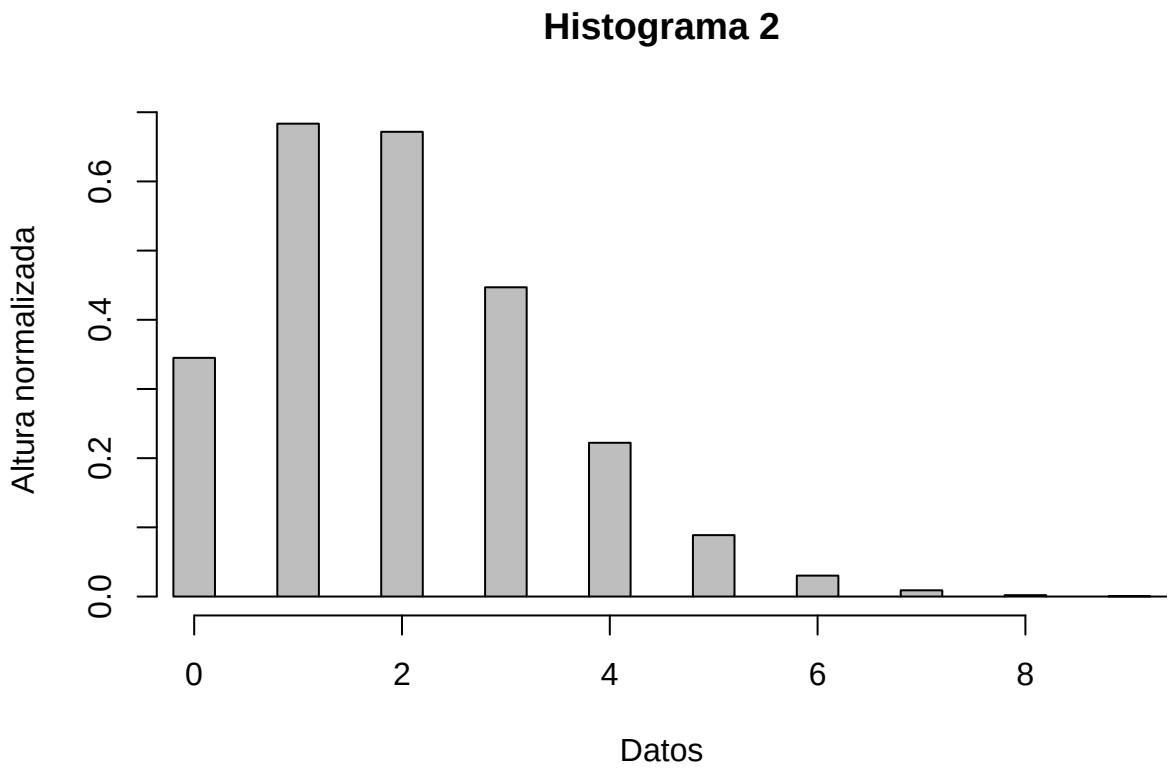
Por ejemplo, hay que poner en el centro de los bines los valores esperados enteros (para evitar problemas de redondeo del ordenador).

```
bines <- c(-0.5:9.5)
hData2 <- hist(data2, breaks=bines, freq=FALSE, main="Histograma 2",
               xlab="Datos", ylab="Altura normalizada",
               xlim=c(min(data2),max(data2)), col="grey")
points(0:9,dpois(0:9, lambda=2), col='red')
```



Está distribución es entonces claramente no simétrica, y se parece mucho a una distribución de Poisson de parámetro 2. Lo mejor sería también dejar espacios en blancos entre cada bin, como en la figura siguiente (usar un binning menor que 1) para insistir en que los números que tenemos son enteros.

```
bines <- c(-0.2, 0.2, 0.8, 1.2, 1.8, 2.2, 2.8, 3.2, 3.8, 4.2, 4.8,
           5.2, 5.8, 6.2, 6.8, 7.2, 7.8, 8.2, 8.8, 9.2, 9.8)
hData2 <- hist(data2, breaks=bines, freq=FALSE, main="Histograma 2",
               xlab="Datos", ylab="Altura normalizada",
               xlim=c(min(data2),max(data2)), col="grey")
```

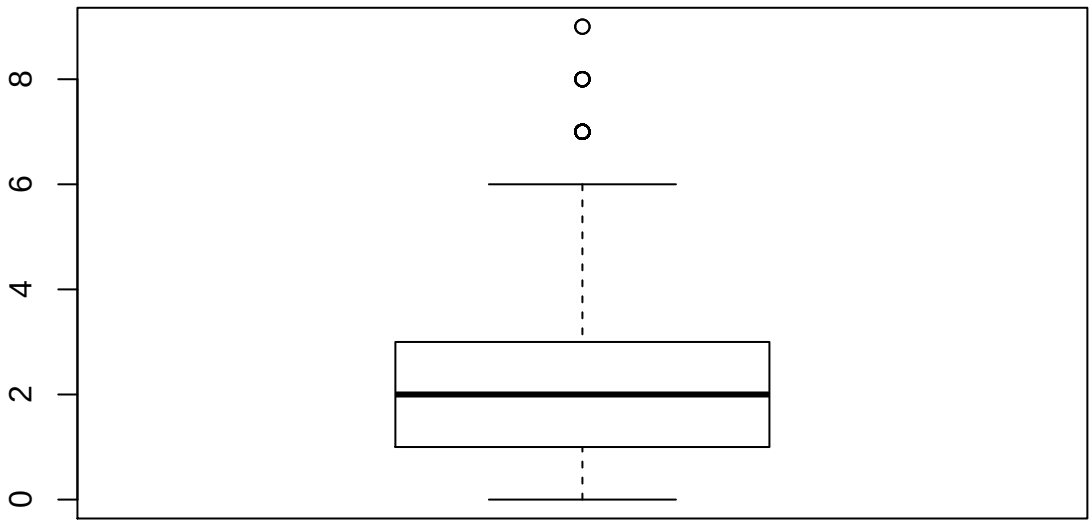


Se puede también estudiar el resumen y el histograma de tipo "boxplot" de este fichero. En este caso vemos que los cuartiles son números enteros también, y que el diagrama de "boxplot" nos da algunos puntos que se alejan bastante de los otros puntos.

```
summary(data2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   1.986   3.000   9.000
```

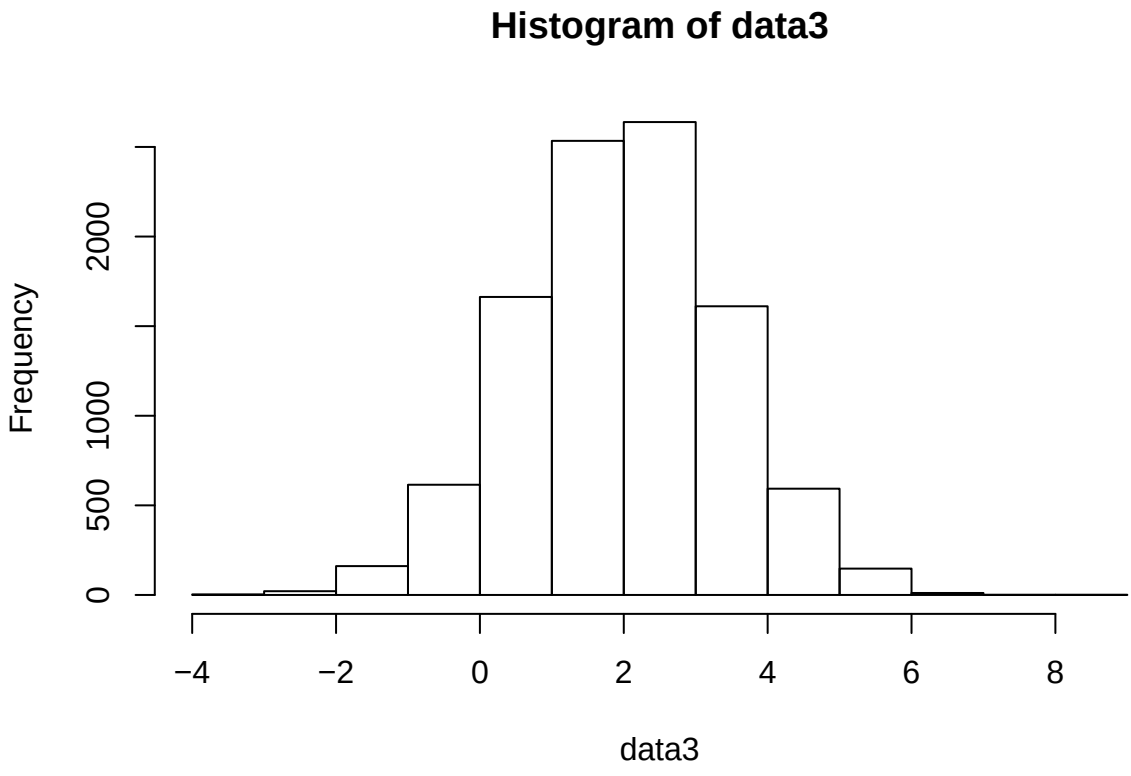
```
boxplot(data2)
```



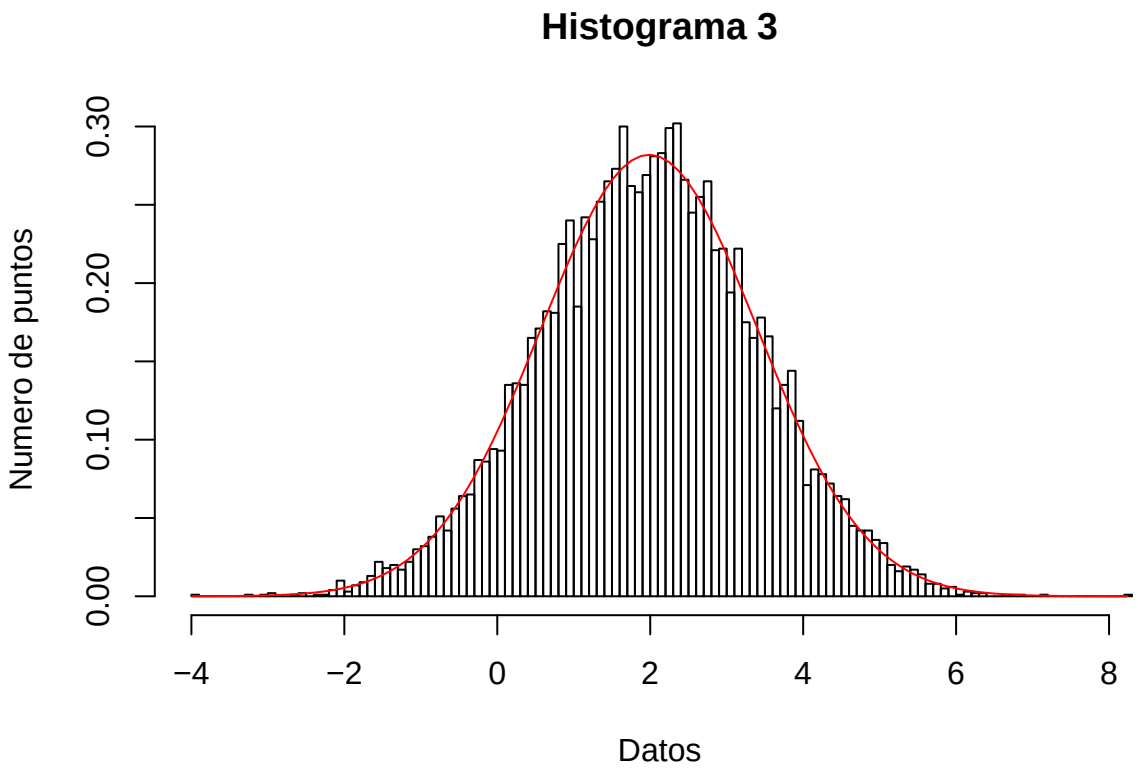
Histograma 3

El histograma 3 está hecho por números que no son enteros, así que se estudia de la misma manera que el primer fichero de puntos.

```
hData3 <- hist(data3)
```




```
hData3bis <- hist(data3, breaks=sqrt(ndata3), freq=FALSE, main="Histograma 3",
                  xlab="Datos", ylab="Numero de puntos",
                  xlim=c(min(data3),max(data3)))
curve(dnorm(x, mean=1.9863, sd=1.4148), col="red", add=TRUE)
```

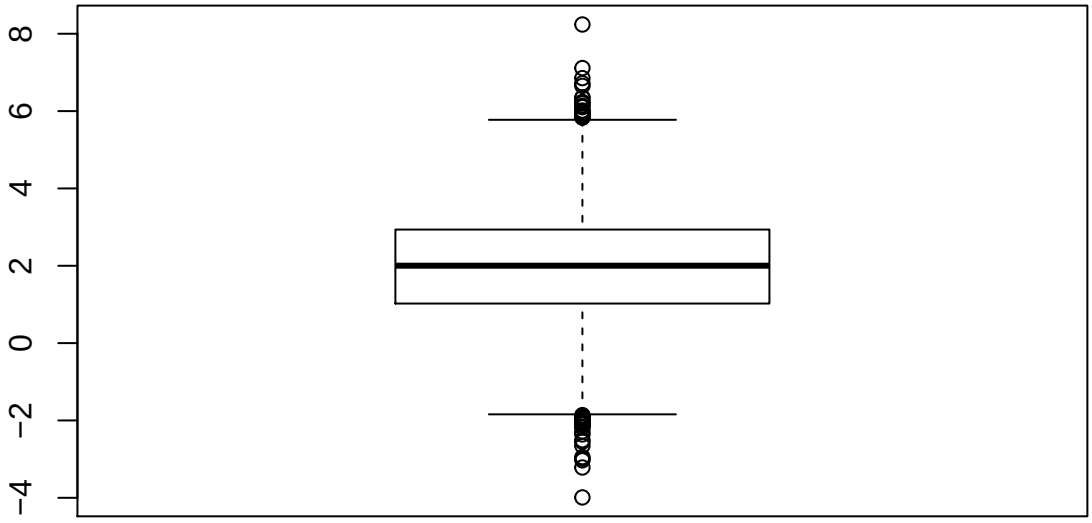


Y por fin se representan los valores minimo, maximo y de cuartiles de esta distribución. Esta distribución es simetrica, y parece mucho a una distribución normal (tiene colas que van rápidamente a 0).

```
summary(data3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -3.990   1.023   2.002   1.986   2.936   8.239
```

```
boxplot(data3)
```



Lo único extraño es este caso es el valor de la moda encontrado en este caso, que vale 0.1635 aunque claramente el pico de la distribución no está en este valor. Si miramos al fichero de datos, este valor aparece exactamente

dos veces mientras todos los otros aparecen solamente una vez. No es un parámetro muy relevante en este caso (la distribución normal viene definida por la media y la desviación estándar).

Conclusión

Hemos estudiado los valores centrales y dispersiones de tres distribuciones distintas, y hemos podido identificar cada distribución de puntos a una distribución típica. También hemos visto el impacto que puede tener el número de bines en la representación de un histograma, así que los parámetros relevantes en cada caso (por ejemplo, la moda no nos sirve a mucho en el último caso de la distribución normal).