Práctica 4: Distribuciones poblacionales y en el muestreo

Introducción

En esta práctica vamos identificar visualmente distintas distribuciones poblacionales y a comprobar, mediante simulación, el alcance del Teorema Central del Límite.

Identificar visualmente distintas distribuciones poblacionales.

La forma del histograma de una muestra puede, si la muestra es suficientemente grande, ayudarnos a identificar la distribución poblacional de la que procede. Los parámetros que definen la distribución concreta pueden, en ciertas ocasiones, estimarse fácilmente a partir de los estadísticos muestrales.

Ejercicio 1 En el archivo *muestras_diversas_simuladas.csv* hay siete muestras, de tamaño 120:

```
muestras_diversas<- read.csv(file="./data/muestras_diversas_simuladas.txt", header=T,
sep = ";", dec = ",", fileEncoding = "WINDOWS-1252")
```

Todas ellas han sido simuladas a partir de poblaciones diferentes, cuya descripción es la siguiente:

- Muestra 1 Presencia/Ausencia de mutación genética.
- Muestra 2 Número de alelos dominantes en los gametos de un heterocigótico para 5 genes con dominancia completa.
- Muestra 3 Número de individuos que presentan una cierta patología en ciudades de 500000 habitantes.
- Muestra 4 Tiempos de reacción a un estímulo visual.
- Muestra 5 Nivel de colesterol.
- Muestra 6 Tiempo de recidiva[1] de erosiones corniales.
- Muestra 7 Porcentaje de macromoléculas (en tanto por uno).

Responde a las siguientes preguntas:

a. ¿Qué variables son discretas y cuáles continuas? ¿Qué valores toman las variables discretas?

Respuesta: `MUESTRA1`, `MUESTRA2` y `MUESTRA3` son variables discretas, el resto continuas. `MUESTRA1` toma los valores 1 (presencia) y 0 (ausencia). `MUESTRA2` toma los valores 0, 1, 2, 3, 4, 5 y `MUESTRA3` toma los valores 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14.

b. ¿Podrías identificar cuál de las muestras proviene de una distribución Bernoulli? ¿y de una Binomial? ¿Y de una Normal?

- *`MUESTRA1` sigue una distribución Bernoulli.*
- *`MUESTRA2` sigue una distribución Binomial.*
- *`MUESTRA3` podría seguir una distribución Binomial.*
- *`MUESTRA5` podría seguir una distribución Normal.*
- *`MUESTRA6` podría seguir una distribución Normal.*

c. Calcula estadísticos descriptivos para cada una de ellas.

```
sapply(muestras_diversas, FUN=summary)
```

```
##          MUESTRA1 MUESTRA2 MUESTRA3 MUESTRA4 MUESTRA5 MUESTRA6 MUESTRA7
## Min.    0.0000000    0.000     2.00   0.0000 227.3500   0.2800  0.46000
## 1st Qu. 0.0000000    2.000     4.75   0.0475 245.0625   1.2050  0.73000
```

[1] La recidiva consiste en la reaparición de una enfermedad tras la convalecencia y recuperación de la misma

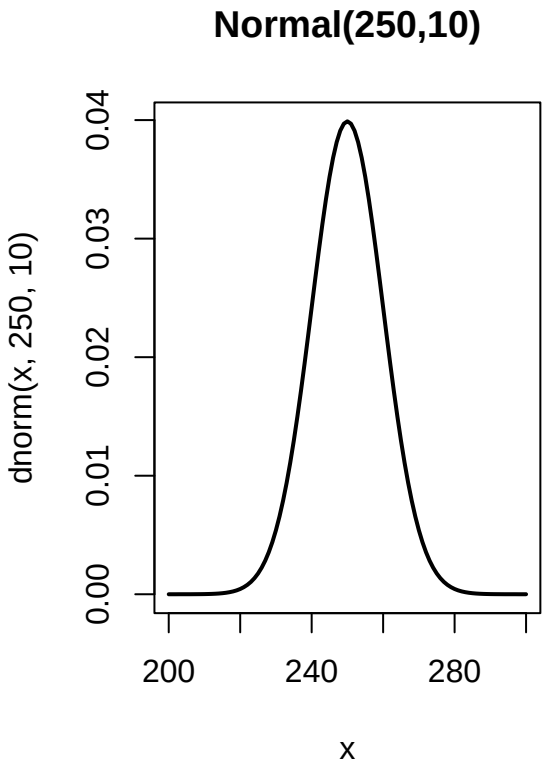
```
## Median  0.0000000    2.000     6.00   0.1300 252.7750   1.5950  0.82000
## Mean    0.3083333    2.425     6.40   0.2250 251.8553   1.7015  0.80075
## 3rd Qu. 1.0000000    3.000     8.00   0.3125 258.8300   2.2050  0.89000
## Max.    1.0000000    5.000    14.00   1.2700 272.9300   3.7500  0.98000
```

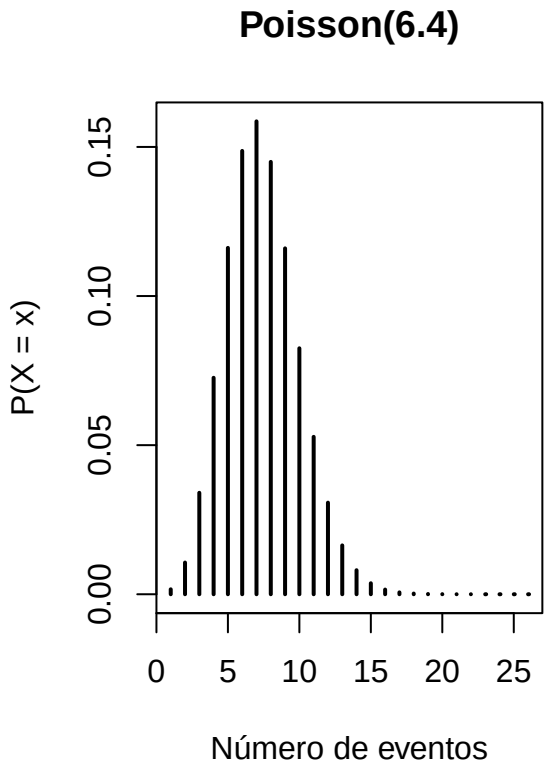
d. Si sabemos que las distribuciones utilizadas (en orden alfabético) son:

- $Bernoulli(\pi = 0.3)$,
- $Beta(\alpha = 9, \beta = 2)$,
- $Binomial(n = 5, \pi = 0.48)$,
- $Exponencial(\lambda = 4.4)$,
- $Gamma(\theta = 1.2, k = 1)$,
- $Normal(\mu = 250, \sigma = 10)$,
- $Poisson(\lambda = 6.4)$.

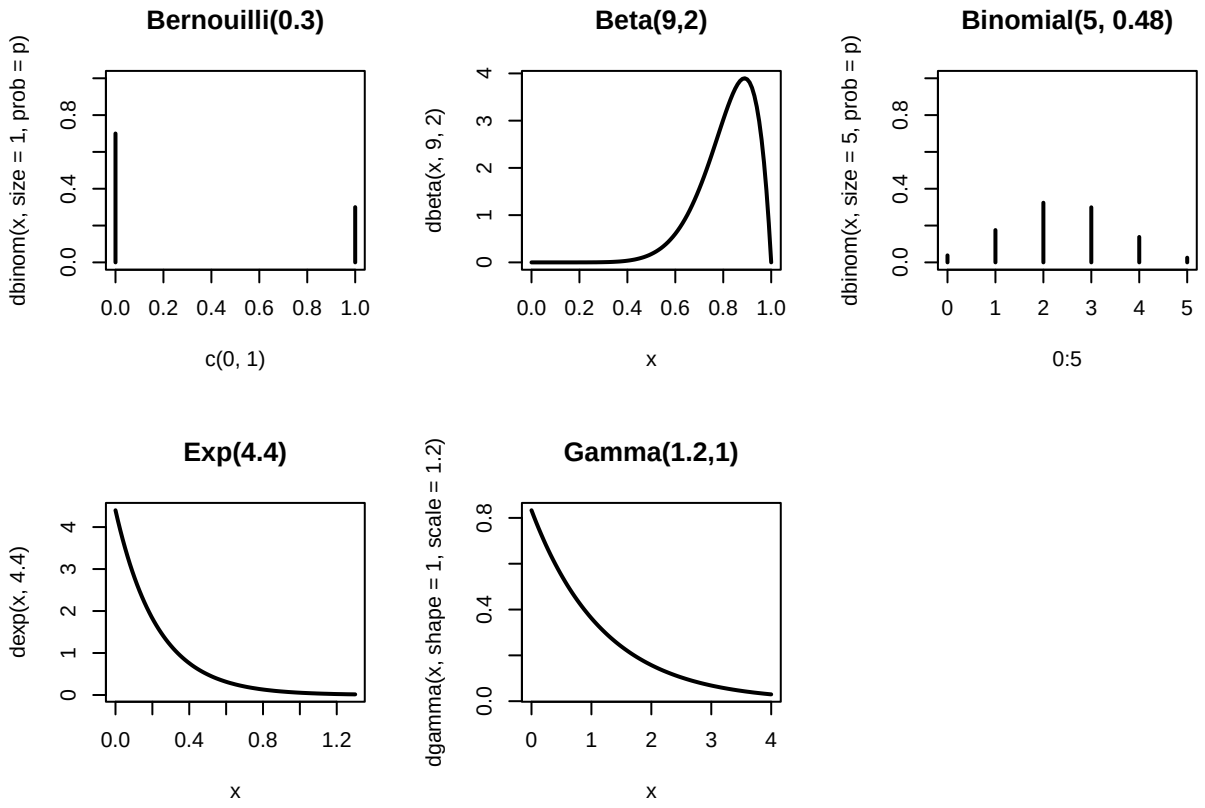
i. Dibuja cada una de estas distribuciones poblacionales. Para ello utiliza la función *curve*() y la función *plot*(), por ejemplo en el caso de la normal y de la poisson sería cómo sigue:

```
par(mfrow=c(1,2))
curve(dnorm(x,250,10), 200, 300, lwd = 2, main = "Normal(250,10)")
# Poisson
x <- 0:25
#-----------
# lambda: 6.4
#-----------
lambda <- 6.4
plot(dpois(x, lambda), type = "h", lwd = 2,
     main = "Poisson(6.4)",
     ylab = "P(X = x)", xlab = "Número de eventos")
```





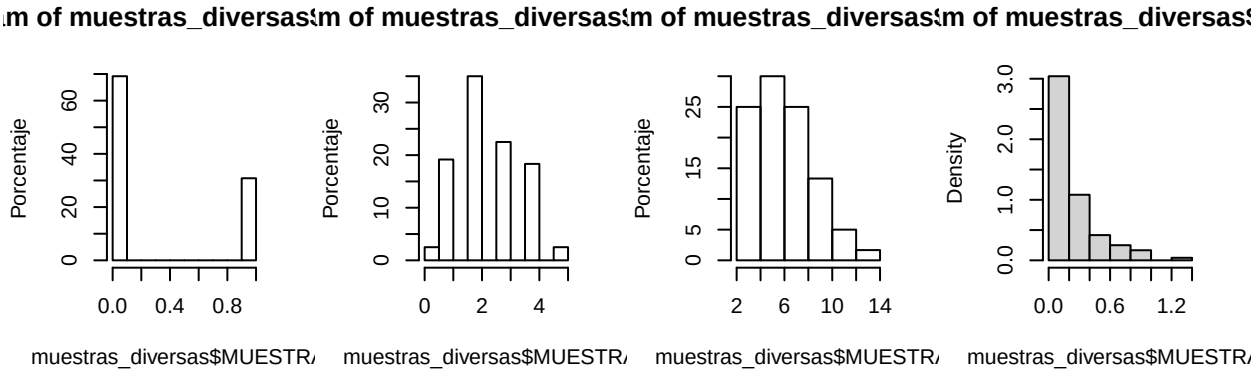

```r
par(mfrow=c(1,1))

par(mfrow=c(2,3))

# Bernouilli
x<-0:1
p<-0.3
plot(dbinom(x, size=1, prob=p),type="h", ylim=c(0,1), x=c(0,1),lwd=2, main="Bernouilli(0.3)")

# Beta
curve(dbeta(x,9,2), lwd = 2, main = "Beta(9,2)")

# Binomial
x<-0:5
p<-0.48
plot(dbinom(x, size=5, prob=p),type="h", ylim=c(0,1), x=0:5,lwd=2, main="Binomial(5, 0.48)")

# Exponencial
curve(dexp(x,4.4), 0,1.3,lwd = 2, main = "Exp(4.4)")

# Gamma
curve(dgamma(x,shape=1, scale=1.2),0,4,lwd = 2, main = "Gamma(1.2,1)")
```

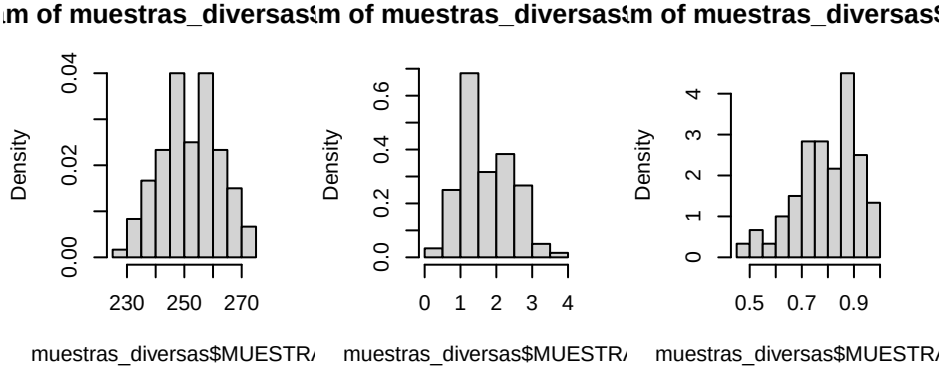


ii. Dibuja un histograma de cada muestra con la escala vertical en porcentaje (para las discretas) y densidad (para las continuas).

```
par(mfrow=c(2,4))

hist_info1 <- hist(muestras_diversas$MUESTRA1, plot = FALSE)# Store output of hist function
hist_info1$density<- hist_info1$counts/sum(hist_info1$counts)*100    # Compute density values
plot(hist_info1, freq = FALSE, ylab="Porcentaje")# Plot histogram with percentages

hist_info2 <- hist(muestras_diversas$MUESTRA2, plot = FALSE)# Store output of hist function
hist_info2$density<- hist_info2$counts/sum(hist_info2$counts)*100    # Compute density values
plot(hist_info2, freq = FALSE,ylab="Porcentaje")

hist_info3 <- hist(muestras_diversas$MUESTRA3, plot = FALSE)# Store output of hist function
hist_info3$density<- hist_info3$counts/sum(hist_info3$counts)*100    # Compute density values
plot(hist_info3, freq = FALSE,ylab="Porcentaje")

hist(x=muestras_diversas$MUESTRA4, freq=FALSE)


hist(x=muestras_diversas$MUESTRA5,freq=FALSE)


hist(x=muestras_diversas$MUESTRA6,freq=FALSE)


hist(x=muestras_diversas$MUESTRA7,freq=FALSE)
```

```
par(mfrow=c(1,1))
```

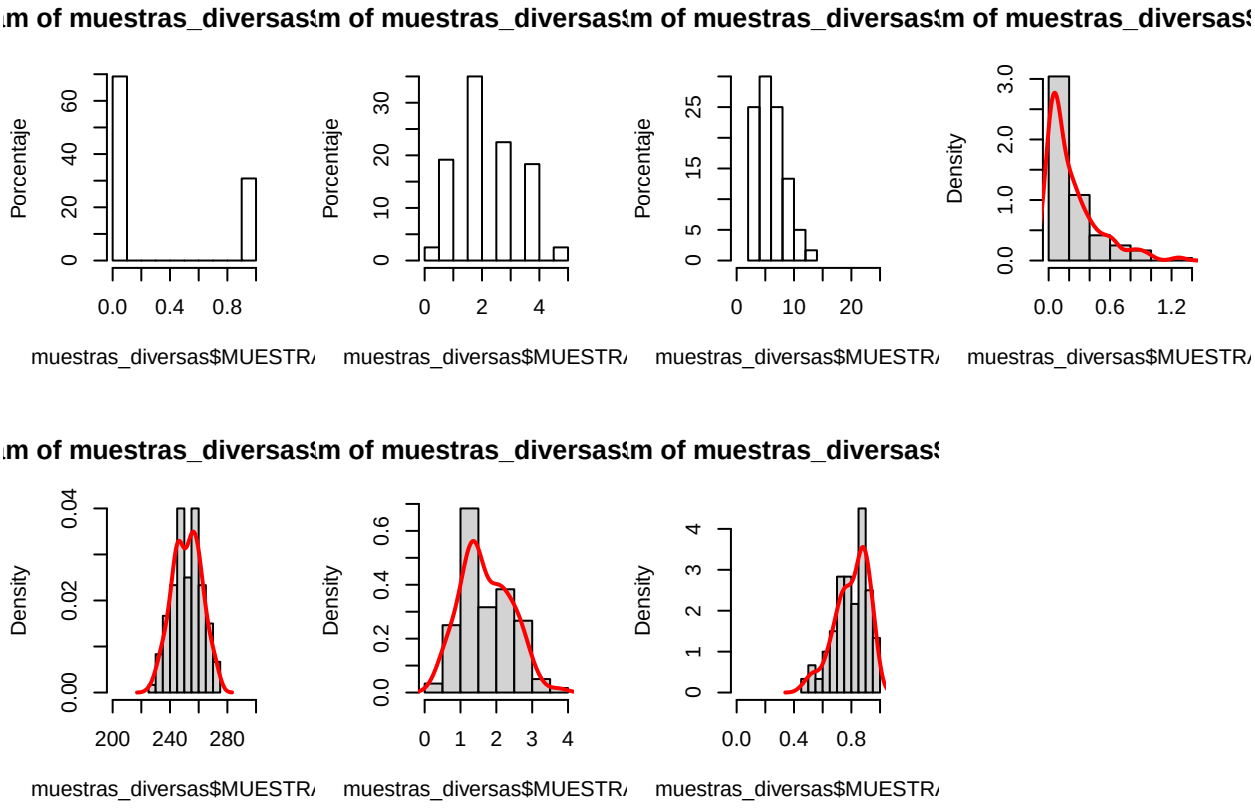




iii. Empareja cada muestra con la distribución poblacional teórica más adecuada. Para facilitar la comparación, redibuja los histogramas utilizando la misma escala horizontal que la correspondiente distribución poblacional. Para ello, si se quiere que se consideren valores de x entre a y b, tienes que añadir la opción $xlim = c(a, b)$.

```
par(mfrow=c(2,4))

hist_info1 <- hist(muestras_diversas$MUESTRA1, plot = FALSE)# Store output of hist function
hist_info1$density<- hist_info1$counts/sum(hist_info1$counts)*100    # Compute density values
plot(hist_info1, freq = FALSE, ylab="Porcentaje")# Plot histogram with percentages

hist_info2 <- hist(muestras_diversas$MUESTRA2, plot = FALSE)# Store output of hist function
hist_info2$density<- hist_info2$counts/sum(hist_info2$counts)*100    # Compute density values
plot(hist_info2, freq = FALSE,ylab="Porcentaje")

hist_info3 <- hist(muestras_diversas$MUESTRA3, plot = FALSE)# Store output of hist function
hist_info3$density<- hist_info3$counts/sum(hist_info3$counts)*100    # Compute density values
plot(hist_info3, freq = FALSE,ylab="Porcentaje", xlim=c(0,25))

hist(x=muestras_diversas$MUESTRA4, freq=FALSE)
lines(density(muestras_diversas$MUESTRA4),lwd=2, col="red")
hist(x=muestras_diversas$MUESTRA5,freq=FALSE,xlim=c(200, 300))
lines(density(muestras_diversas$MUESTRA5),lwd=2, col="red")
hist(x=muestras_diversas$MUESTRA6,freq=FALSE)
```

```
lines(density(muestras_diversas$MUESTRA6),lwd=2, col="red")
hist(x=muestras_diversas$MUESTRA7,freq=FALSE, xlim=c(0,1))
lines(density(muestras_diversas$MUESTRA7),lwd=2, col="red")
par(mfrow=c(1,1))
```



Respuesta:

- *MUESTRA1 sigue una distribución Bernoulli.*
- *MUESTRA2 sigue una distribución Binomial.*
- *MUESTRA3 podría seguir una distribución Poisson.*
- *MUESTRA4 podría seguir una distribución Exponencial.*
- *MUESTRA5 podría seguir una distribución Normal.*
- *MUESTRA6 podría seguir una distribución Gamma.*
- *MUESTRA7 podría seguir una distribución Beta.*

Comprobar, mediante simulación, el alcance del Teorema Central del Límite.

Recordemos que el Teorema Central del Límite (TCL) establece que:

Si $\{X_1, \ldots, X_n\}$ es una muestra aleatoria de una población con media μ y desviación típica σ. Entonces, para valores de n grandes, la distribución en el muestreo $\bar{X}$ es aproximadamente Normal con media μ y desviación típica $\sigma/\sqrt{n}$.

Para familiarizarnos con este resultado vamos a utilizar bancos de datos de poblaciones NO Normales y comprobaremos que, al tomar muestras cada vez mayores y calcular su medias muestrales, los valores de esas medias muestrales se comportan como los de una distribución Normal.

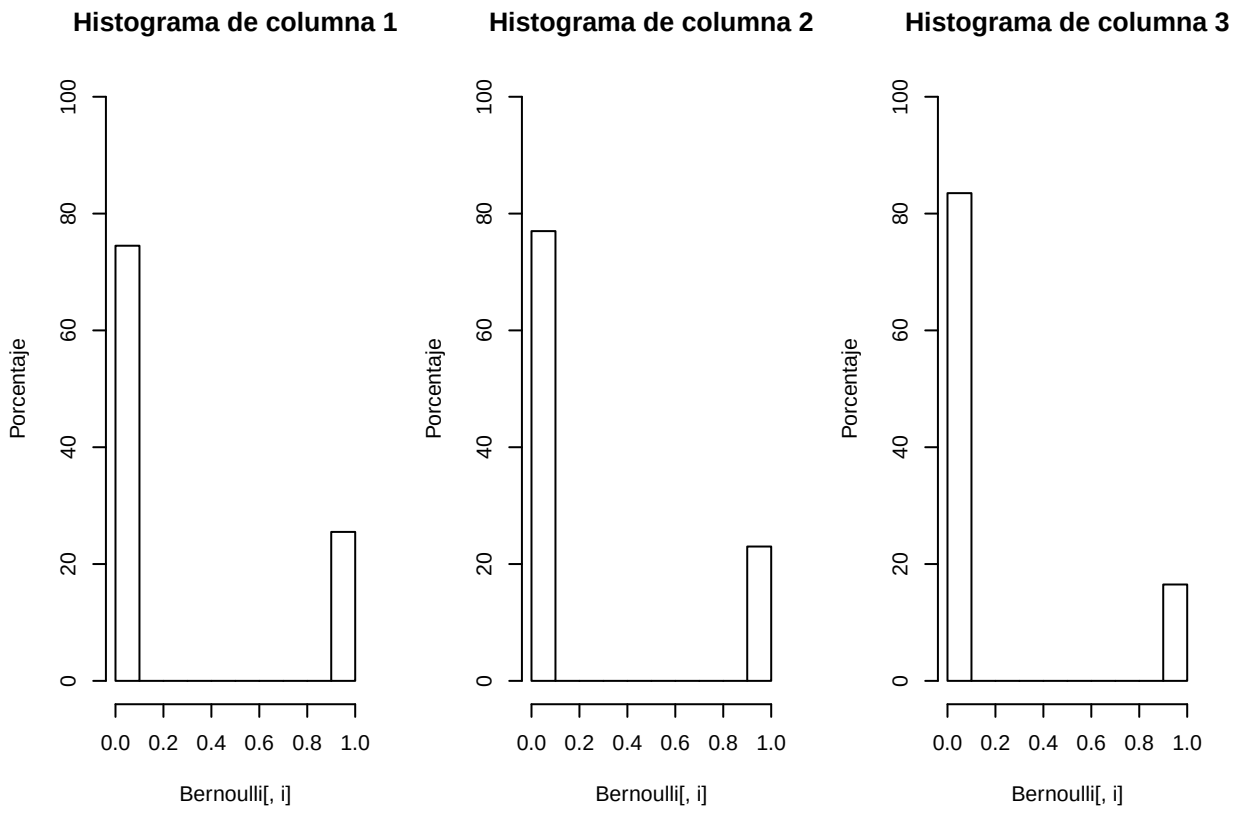
Ejercicio 2 En el archivo *cien_muestras_bernoulli.csv* encontrarás un banco de datos con 200 filas por 100

columnas (200×100), cuyos valores son "1" o "0". Se trata, de 100 muestras aleatorias, cada una de tamaño 200, de una población cuya característica es una variable aleatoria dicotómica (1=Éxito, 0=Fracaso), conocida como distribución Bernoulli y, por tanto, con distribución NO Normal. Piensa que podría ser la variable "género" (Hombre/Mujer), "mutación" (si/no) o cualquier característica que denotara presencia/ausencia (Enfermo/No Enfermo, Test+/Test-), etc. Pero, también podemos considerarlas como una muestra de tamaño 20000 (200×100), de la misma población, y utilizarlas de distintas formas. 1. Importa los datos del fichero *cien_muestras_bernoulli.xlsx* y llama Bernoulli al conjunto de datos activo.

```
Bernoulli<- read.csv("./data/cien_muestras_bernoulli.txt", sep=";")
```

2. Cada columna representa una muestra de la misma población: elige algunas (tres o cuatro) de esas muestras (columnas), las que quieras, y analiza si su histograma te permite suponer que provienen de una distribución Normal. Anota el valor de la media y la desviación típica de cada una de ellas. ¿Cuánto crees que vale la probabilidad de éxito π?

```
par(mfrow=c(1,3))
m<-numeric(3)
sd<-numeric(3)
for (i in 1:3){
  hist_info <- hist(Bernoulli[,i], plot = FALSE)# Store output of hist function
  hist_info$density<- hist_info$counts/sum(hist_info$counts)*100    # Compute density values
  plot(hist_info, freq = FALSE, ylab="Porcentaje", ylim=c(0,100), main=paste("Histograma de columna", i

  m[i]<-mean(Bernoulli[,i])
  sd[i]<-sd(Bernoulli[,i])
}
```



```
par(mfrow=c(1,1))

m
```

```
## [1] 0.255 0.230 0.165
```

```
sd
```

```
## [1] 0.4369550 0.4218886 0.3721120
```

Respuesta: la probabilidad de éxito de una Bernoulli podemos aproximarla por la media de los éxitos, en este caso, 0.255, 0.23, 0.165 respectivamente.

3. Vamos a calcular **MEDIAS MUESTRALES de muestras** de nuestra población y veremos cómo se comportan. Consideramos que **cada fila es una muestra** de la población (como hay 200 filas, tenemos **200 muestras**). Así, según el número de columnas que elijamos conseguiremos muestras de tamaño diferente. **Empezamos con n=10**:

   i. El siguiente código, que debes copiar en la Ventana de instrucciones, en cada muestra (fila) calculará la media de los 10 datos (valores que hay en las 10 primeras columnas de la fila correspondiente) y lo guardará en una variable llamada Media_n10. En ella tenemos 200 medias de muestras de tamaño 10:

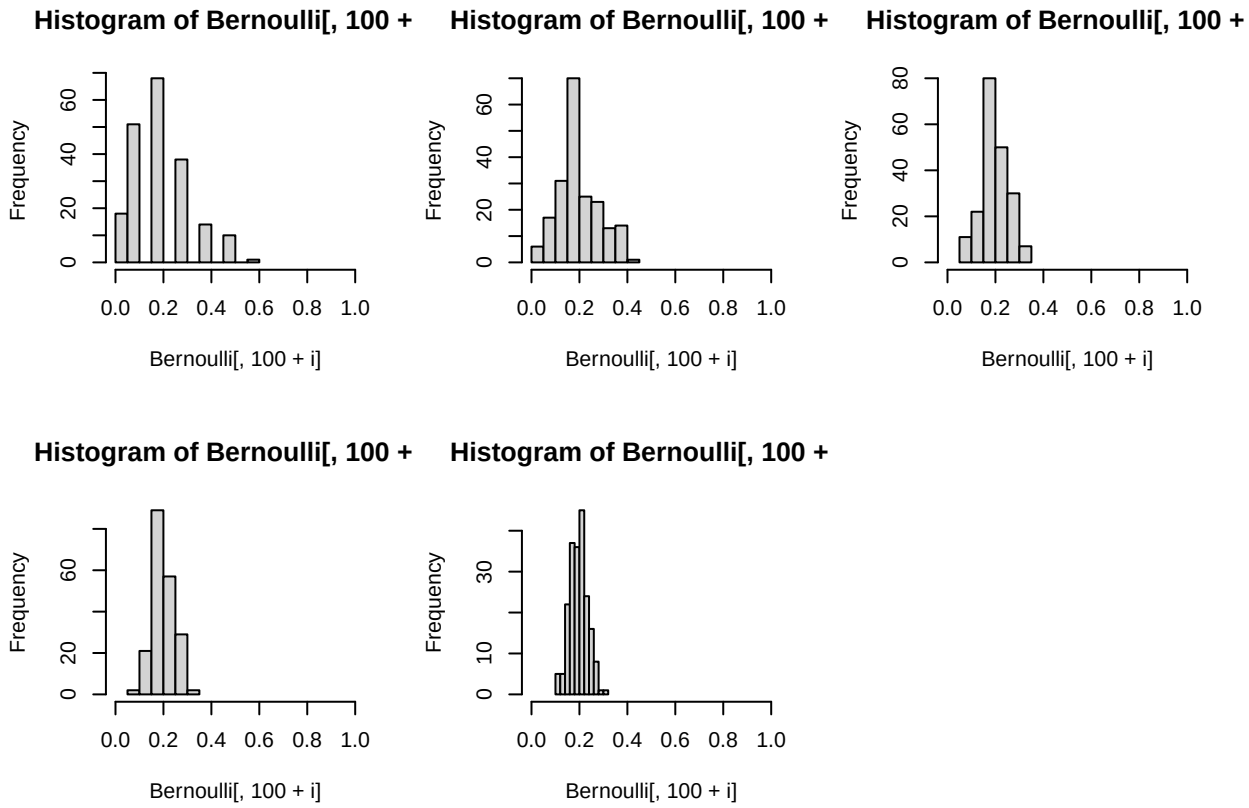
```
media_global<-mean(as.matrix(Bernoulli))
desv_global<-sd(as.matrix(Bernoulli))
for( i in 1:200 ) { Bernoulli$Media_n10[i] = sum( Bernoulli[i,1:10] )/10 }
```

   ii. Repite ese proceso para tamaños n=25, 50, 75 y 100, cambiando el valor 10 en la expresión anterior (aparece 3 veces). Así, para cada uno de ellos, obtendremos las medias de 200 muestras.

```
for( i in 1:200 ) { Bernoulli$Media_n25[i] = sum( Bernoulli[i,1:25] )/25}
for( i in 1:200 ) { Bernoulli$Media_n50[i] = sum( Bernoulli[i,1:50] )/50 }
for( i in 1:200 ) { Bernoulli$Media_n75[i] = sum( Bernoulli[i,1:75] )/75 }
for( i in 1:200 ) { Bernoulli$Media_n100[i] = sum( Bernoulli[i,1:100] )/100 }
```

   iii. Analiza la normalidad de Media_n10, . . ., Media_n100, a partir del correspondiente histograma (añadir la opción $xlim=c(0,1) para obtener todos en la misma escala horizontal).

```
par(mfrow=c(2,3))
for (i in 1:5){
  hist(Bernoulli[,100+i], xlim=c(0,1))
}
par(mfrow=c(1,1))
```



iv. Anota las medias y las desviaciones típicas de las variables Media_n10, ..., Media_n100, recordando que cada una representa la media y la desviación típica (estimadas) de las medias muestrales del tamaño correspondiente.

```
medias<-numeric(5)
desv<-numeric(5)
for (i in 1:5){
  medias[i]<-mean(Bernoulli[,100+i])
  desv[i]<-sd(Bernoulli[,100+i])
}
medias
```

```
## [1] 0.2065000 0.2012000 0.2030000 0.2019333 0.2018000
```

```
desv
```

```
## [1] 0.12723480 0.08848564 0.05641733 0.04383406 0.03662605
```

v. ¿Para qué tamaño muestral crees que ya se ha producido el resultado que "anuncia" el TCL?, es decir, ¿para qué n podemos afirmar que las medias de muestras de ese tamaño se comportan como una distribución Normal?

```
desv_teoricas<-numeric(5)
desv_teoricas[1]<-desv_global/sqrt(10)
desv_teoricas[2]<-desv_global/sqrt(25)
desv_teoricas[3]<-desv_global/sqrt(50)
desv_teoricas[4]<-desv_global/sqrt(75)
desv_teoricas[5]<-desv_global/sqrt(100)
```

```
media_global
```

```
## [1] 0.2018
```

```
desv_teoricas
```

```
## [1] 0.12691919 0.08027075 0.05675999 0.04634434 0.04013537
```

Respuesta: vemos cómo ya desde el tamaño 10 los datos parecen aproximarse a una normal. Además los parámetros de la muestra convergen rápidamente a los parámetros poblacionales teóricos.

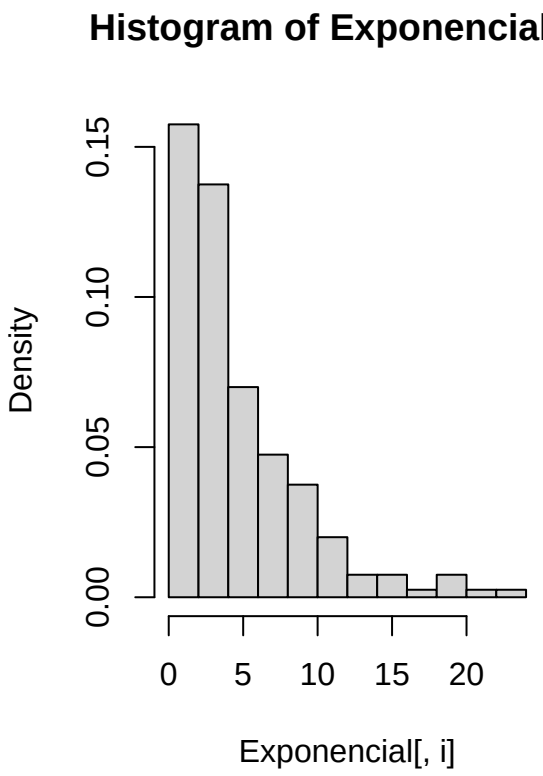
Ejercicio 3. En el archivo *25_muestras_exponencial.csv* encontrarás un banco de datos con 200 filas $\times$ 25 columnas, cuyos valores son datos de una población con distribución Exponencial ($\lambda = 1/4$) (serviría para describir, por ejemplo observaciones de tiempos de supervivencia, que tuvieran como media 4 unidades de tiempo). Vamos a repetir el estudio que hemos hecho con las Bernoulli:

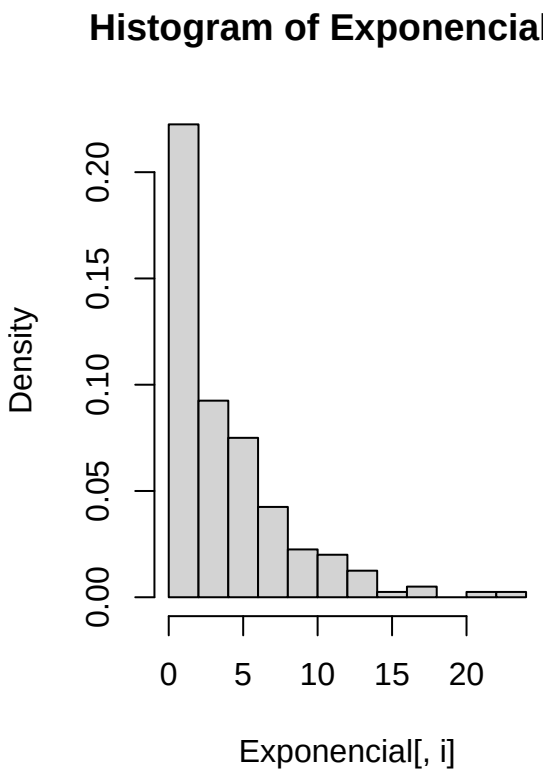
1. Importa los datos del fichero *25_muestras_exponencial.csv* y llama Exponencial al conjunto de datos activo.

```
Exponencial<- read.csv("./data/25_muestras_exponencial.txt", sep=";", dec=",")
media_global<-mean(as.matrix(Exponencial))
desv_global<-sd(as.matrix(Exponencial))
```

2. Cada columna representa una muestra de la misma población: Elige dos de esas muestras (columnas), las que quieras, y analiza si su histograma te permite suponer que provienen de una distribución Normal. Anota el valor de la media y la desviación típica de cada una de ellas.

```
par(mfrow=c(1,2))
m<-numeric(2)
sd<-numeric(2)
for (i in 1:2){
  hist(Exponencial[,i], plot = TRUE, freq=FALSE)# Store output of hist function
  m[i]<-mean(Exponencial[,i])
  sd[i]<-sd(Exponencial[,i])
}
```






```
par(mfrow=c(1,1))

m
```

```
## [1] 4.52945 4.00570
```

```
sd
```

```
## [1] 4.312859 4.041665
```

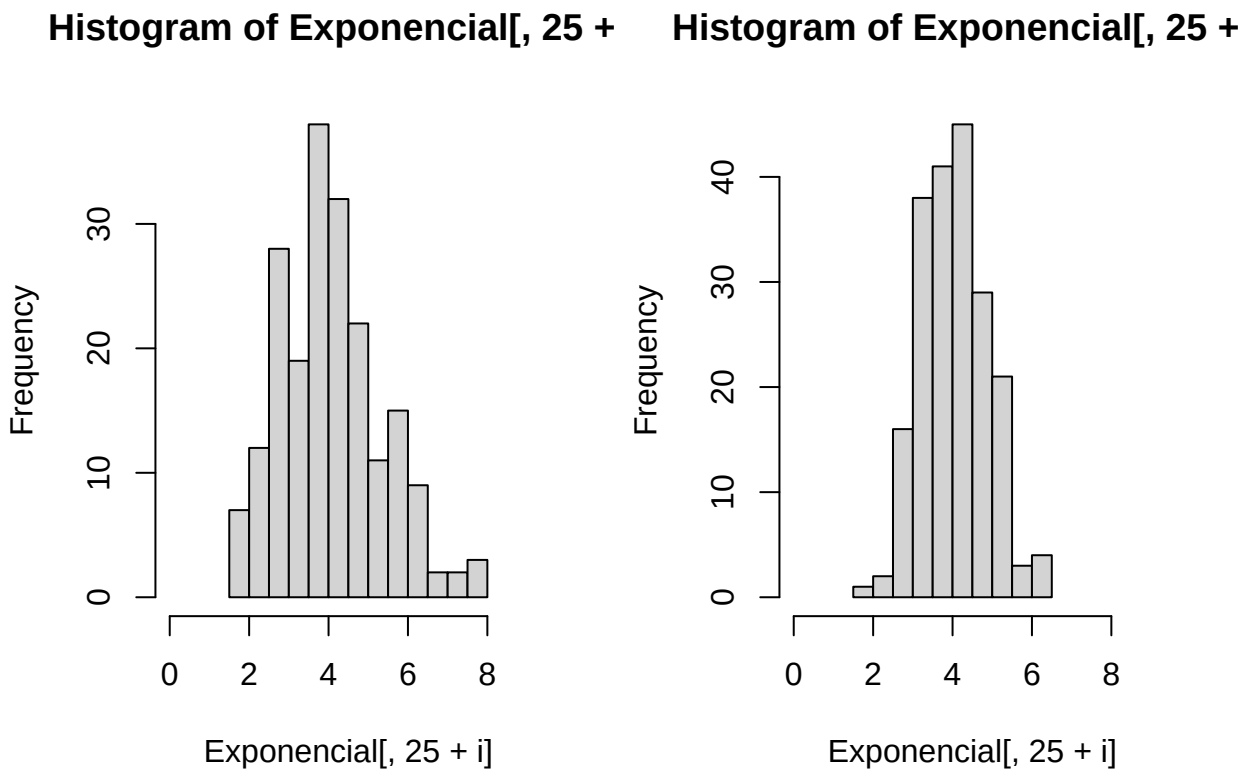
3. Vamos a calcular **MEDIAS MUESTRALES de muestras** de nuestra población y veremos cómo se comportan. Consideramos que cada fila es una muestra de la población (como hay 200 filas, tenemos 200 muestras). Así, según el número de columnas que elijamos conseguiremos muestras de tamaño diferente.

i. Crea las variables Media_n10 y Media_n25 de forma análoga al caso de una distribución Bernoulli.

```
for( i in 1:200 ) { Exponencial$Media_n10[i] = sum(Exponencial[i,1:10] )/10}
for( i in 1:200 ) { Exponencial$Media_n25[i] = sum( Exponencial[i,1:25] )/25}
```

ii. Analiza la normalidad de ambas variables a partir del correspondiente histograma.

```
par(mfrow=c(1,2))
for (i in 1:2){
  hist(Exponencial[,25+i], xlim=c(0,9))
}
```




```
par(mfrow=c(1,1))
```

iii. Anota las medias y las desviaciones típicas de ambas variables y comprueba que se parecen a los valores teóricos que nos indica el Teorema Central del Límite.

```
medias<-numeric(2)
medias[1]<-mean(Exponencial[,26])
medias[2]<-mean(Exponencial[,27])

desv<-numeric(2)
desv[1]<-sd(Exponencial[,26])
desv[2]<-sd(Exponencial[,27])

desv_teoricas<-numeric(2)
desv_teoricas[1]<-desv_global/sqrt(10)
desv_teoricas[2]<-desv_global/sqrt(25)

medias
```

```
## [1] 4.06457 4.05523
```

```
media_global
```

```
## [1] 4.05523
```

```
desv
```

```
## [1] 1.2792118 0.8432246
```

```
desv_teoricas
```

```
## [1] 1.2542187 0.7932375
```

iv. ¿Para qué tamaño muestral crees que ya se ha producido el resultado que "anuncia" el TCL?, es decir ¿para qué n podemos afirmar que las medias de muestras de ese tamaño se comportan como una distribución Normal?

Respuesta: los estadísticos muestrales convergen rápidamente a los parámetros poblacionales y podríamos ver que se aprecia la forma de campana gaussiana desde $n = 10$*, aunque más claramente desde* $n = 25$*.*

v. ¿A qué crees que se debe la diferencia que se observa en el valor de n con respecto al que se obtuvo en el ejercicio 2?

Respuesta: a las diferencias en el tamaño de las propias muestras, es decir, antes tomábamos 200 elementos para calcular cada media y en este caso la mitad, 100.