Técnicas de Simulación y Algoritmos de Muestreo

Ángel Ríos San Nicolás

30 de mayo de 2021

Práctica 1

Problema 1. *Considera los datos de la liga NBA de baloncesto profesional de la temporada 2014-2015 que figuran en el archivo* `players_stats.csv`.

i) Introduce los datos del archivo en el programa R.

ii) Ordena los datos de altura de menor a mayor y guarda los datos ordenados en un nuevo vector.

iii) Guarda el vector de los datos ordenados en un archivo de texto llamado datos_ordenados.txt.

iv) Calcula la suma total, la media, la varianza, la desviación típica, la mediana, los cuartiles, la skewness y la kurtosis de los datos de la muestra.

v) Calcula el máximo y el mínimo de los datos y el tamaño muestral.

vi) Calcula qué porcentaje tiene una altura igual o superior a 2 metros.

vii) Calcula la altura media por cada posición.

viii) Calcula los coeficientes de correlación de altura con porcentajes de acierto en tiros libres, tiros de campo y triples respectivamente.

Solución.

i) Cargamos los datos en *R*.

```
datos <- read.csv('players_stats.csv')
```

ii) Ordenamos las alturas y las guardamos en un nuevo vector.

```
alturas <- sort(datos$Height)  # Alturas ordenadas
```

iii) Guardamos el vector en un archivo llamado *datos_ordenados.txt*.

```
write(alturas, 'alturas_ordenadas.txt', ncolumns = 1)
```

iv) Calculamos y mostramos en orden la suma total, la media, la varianza, la desviación típica, la mediana, los cuartiles, la skewness y la kurtosis.

```
cat('Suma total:', sum(alturas), fill = T)
cat('Media:', mean(alturas), fill = T)
cat('Varianza:', var(alturas), fill = T)
cat('Desviación típica:', sd(alturas), fill = T)
cat('Mediana:', median(alturas), fill = T)
cat('Cuartiles:\n')
quantile(alturas)
# Importamos las funciones que calculan la skewness y la curtosis.
source('skewness.R')
```

```r
source('kurtosis.R')
cat('Skewness:', skewness(alturas), fill = T)
cat('Kurtosis:', kurtosis(alturas))
```

```
## Suma total: 83320
## Media: 197.4408
## Varianza: 76.39197
## Desviación típica: 8.74025
## Mediana: 197.5
## Cuartiles:
##    0%   25%   50%   75%  100%
## 172.5 190.0 197.5 205.0 222.5
## Skewness: -0.3063402
## Kurtosis: -0.5914659
```

v) Calculamos y mostramos el máximo, el mínimo y el tamaño muestral.

```r
cat('Máximo:', max(alturas), fill = T)
cat('Mínimo:', min(alturas), fill = T)
cat('Tamaño muestral:', length(alturas))
```

```
## Máximo: 222.5
## Mínimo: 172.5
## Tamaño muestral: 422
```

vi) Para calcular el porcentaje de jugadores con altura igual o superior a 2 metros, extraemos el subconjunto de alturas iguales o superiores a 200, y dividimos su longitud entre el total.

```r
mas200 <- subset(alturas, alturas >= 200)
length(mas200)/length(alturas)
```

```
## [1] 0.4881517
```

Un 48.81517% de los jugadores tiene una altura igual o superior a 2 metros.

vii) Extraemos los subconjuntos de alturas por cada posición y calculamos su media.

```r
alturasPG <- subset(datos$Height, datos$Pos == 'PG')
alturasSG <- subset(datos$Height, datos$Pos == 'SG')
alturasSF <- subset(datos$Height, datos$Pos == 'SF')
alturasC <- subset(datos$Height, datos$Pos == 'C')
cat('Altura media de PG:', mean(alturasPG), fill = T)
cat('Altura media de SG:', mean(alturasSG), fill = T)
cat('Altura media de SF:', mean(alturasSF), fill = T)
cat('Altura media de C:', mean(alturasC), fill = T)
```

```
## Altura media de PG: 185.3869
## Altura media de SG: 192.6
## Altura media de SF: 199.9671
## Altura media de C: 207.9514
```

viii) Consideramos los porcentajes de aciertos de cada uno de los tiros de los que la altura no sea NA para poder calcular los coeficientes de correlación.

```r
tiros_libres <- subset(datos$FT., !is.na(datos$Height))
tiros_campo <- subset(datos$FG., !is.na(datos$Height))
tiros_triples <- subset(datos$X3P., !is.na(datos$Height))
alturas <- subset(datos$Height, !is.na(datos$Height))
cat('Coeficiente de correlación de la altura con aciertos en tiros libres:',
```

```
    cor(alturas, tiros_libres), fill = T)
cat('Coeficiente de correlación de la altura con aciertos en tiros de campo:',
    cor(alturas, tiros_campo), fill = T)
cat('Coeficiente de correlación de la altura con aciertos en tiros triples:',
    cor(alturas, tiros_triples), fill = T)
```

```
## Coeficiente de correlación de la altura con aciertos en tiros libres:
## -0.1639388
## Coeficiente de correlación de la altura con aciertos en tiros de campo:
## 0.3665988
## Coeficiente de correlación de la altura con aciertos en tiros triples:
## -0.3557487
```

Problema 2. *Descarga los datos de extensión de hielo ártico del directorio:*

ftp://sidads.colorado.edu/DATASETS/NOAA/G02135/north/daily/data/

i) Carga los datos en R. Recomendación: Elimina la segunda línea del archivo utilizando los comandos indicados en README.R.

ii) Calcula la media y desviación típica de las extensiones para cada mes.

iii) Calcula los percentiles del 5%, 50% y 95% para cada mes.

Solución.

i) Cargamos los datos en *R*.

```
all_content <- readLines("N_seaice_extent_daily_v3.0.csv")
skip_second <- all_content[-2]
datos <- read.csv(textConnection(skip_second))
```

ii) Tomamos dos listas de 12 ceros que guardarán las medias y las desviaciones. Con un bucle for iteramos para cada mes y calculamos la media y la desviación.

```
medias <- rep(0, 12)
desviaciones <- rep(0, 12)
for(i in 1:12){
  medias[i] <- mean(subset(datos$Extent, datos$Month == i))
  desviaciones[i] <- sd(subset(datos$Extent, datos$Month == i))
}
cat("Media de cada mes:", medias, fill= T)
cat("Desviación de cada mes:", desviaciones, fill = T)
```

```
## Media de cada mes: 14.14464 15.0286 15.19501 14.46027 13.05441 11.46896
## 9.003051 6.692664 5.893771 7.804212 10.38216 12.59596
## Desviación de cada mes: 0.6835756 0.5966151 0.5467016 0.6179044 0.6375765
## 0.7920122 1.201316 1.126884 1.139729 1.453406 0.9582938 0.8182118
```

iii) Repetimos lo mismo para los percentiles.

```
percentil5 <- rep(0,12)
percentil50 <- rep(0,12)
percentil95 <- rep(0,12)
for (i in (1:12)){
  medias[i] <- mean(subset(datos$Extent, datos$Month == i))
  desviaciones[i] <- var(subset(datos$Extent, datos$Month == i))
  percentil5[i] <- quantile(subset(datos$Extent, datos$Month == i), 0.05)
  percentil50[i] <- quantile(subset(datos$Extent, datos$Month == i), 0.5)
  percentil95[i] <- quantile(subset(datos$Extent, datos$Month == i), 0.95)
```

```
}
cat("Percentil 5% de cada mes:", percentil5, fill = T)
cat("Percentil 50% de cada mes:", percentil50, fill = T)
cat("Percentil 95% de cada mes:", percentil95, fill = T)
```

```
## Percentil 5% de cada mes: 13.065 14.10945 14.3062 13.4652 11.9602 10.0092
## 6.9319 4.7595 4.147 5.18 8.7331 11.1616
## Percentil 50% de cada mes: 14.133 15.0955 15.243 14.443 13.06 11.571 9.045
## 6.7765 6.039 8.057 10.393 12.633
## Percentil 95% de cada mes: 15.299 15.9642 16.06 15.4884 14.11895 12.6426
## 10.9253 8.3806 7.54425 9.827 11.9641 13.8858
```

Práctica 2

Problema 1. *Considera los datos de la liga NBA de baloncesto profesional de la temporada 2014-2015 que figuran en el archivo* `players_stats.csv`.

i) Representa gráficamente las datos de la columna `Height` *de forma que se aprecie la distribución de los datos.*

ii) Calcula la media muestral y represéntala mediante una línea en la gráfica.

iii) Calcula la desviación típica muestral y representa el intervalo media muestral ± desviación típica muestral en la gráfica.

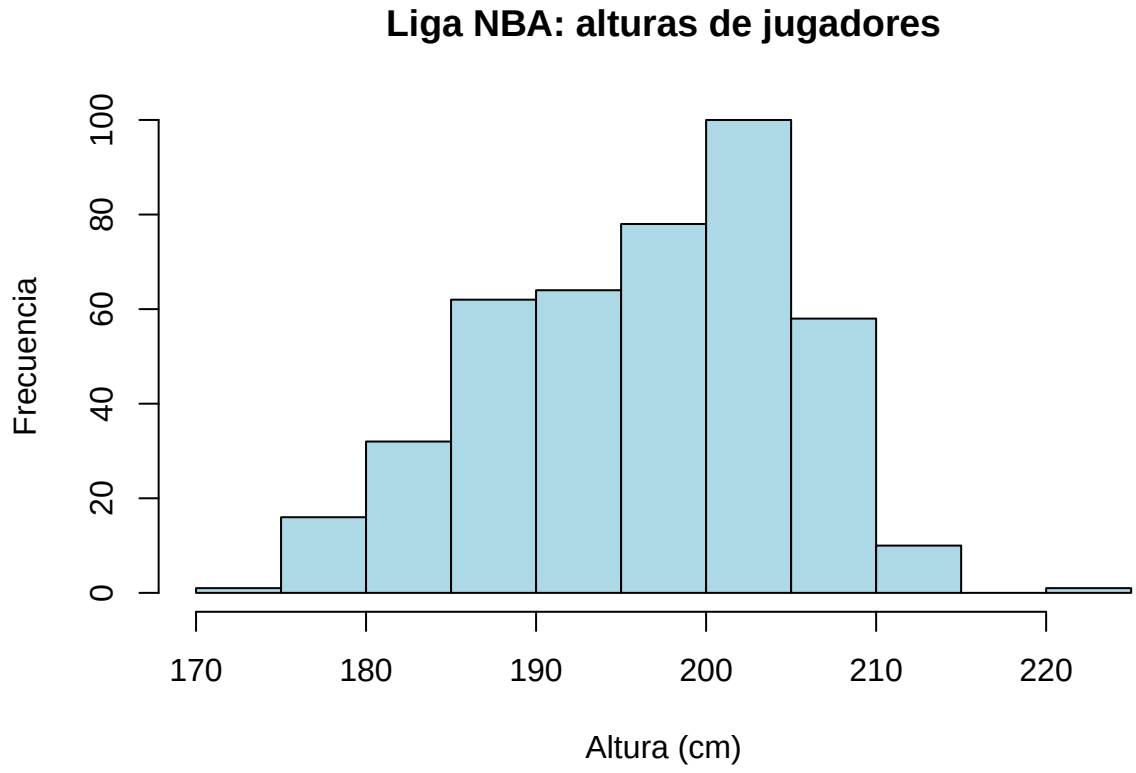
iv) Compara gráficamente las alturas según las diferentes posiciones de la columna `Pos`.

v) Comenta brevemente las gráficas.

Solución.

i) Cargamos los datos, extraemos las alturas y las representamos mediante un histograma.

```
datos <- read.csv("players_stats.csv")
alturas <- subset(datos$Height,!is.na(datos$Height))
hist(alturas, main = "Liga NBA: alturas de jugadores",
     xlab = "Altura (cm)", ylab = "Frecuencia", col = "lightblue")
```



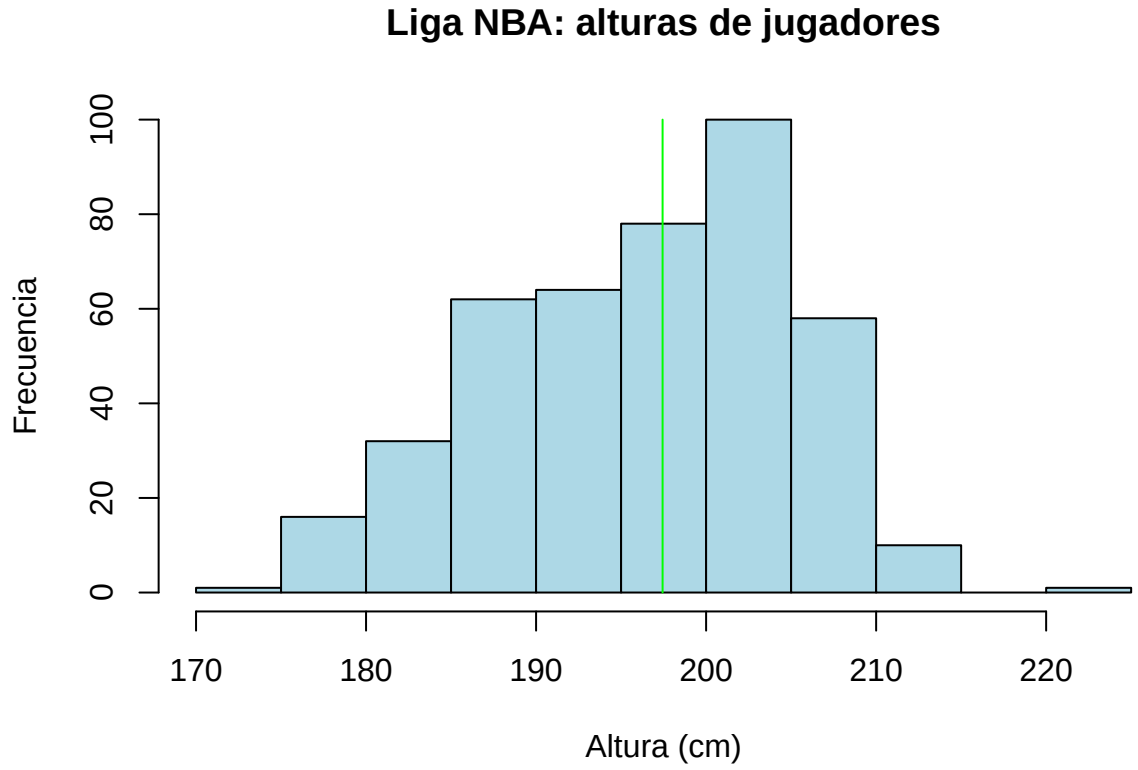
ii) Calculamos la media de las alturas.

```
media <- mean(alturas)
cat("Altura media:", media)
```

```
## Altura media: 197.4408
```

La representamos en el histograma mediante una línea vertical verde.

```r
hist(alturas, main = "Liga NBA: alturas de jugadores",
     xlab = "Altura (cm)", ylab = "Frecuencia", col = "lightblue")
lines(rep(media, 101), 0:100, col = "green")
```



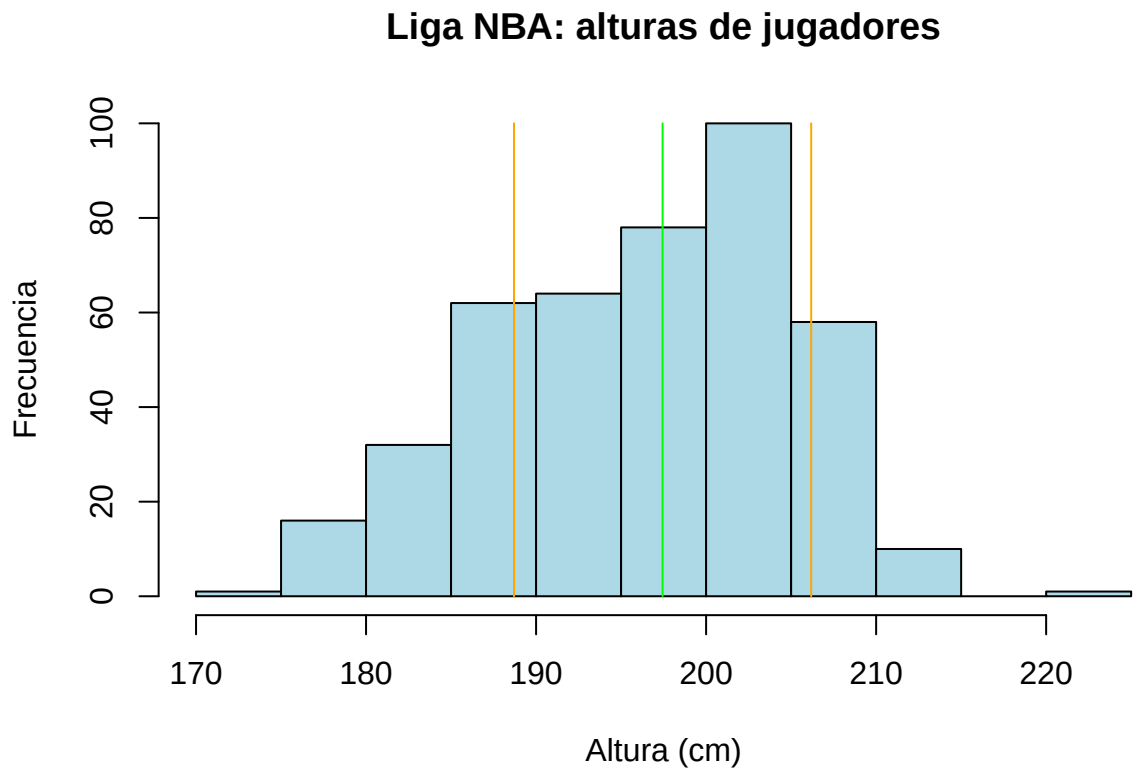
iii) Calculamos la desviación típica de las alturas.

```r
desviacion <- sd(alturas)
cat("Desviación típica:", desviacion)
```

```
## Desviación típica: 8.74025
```

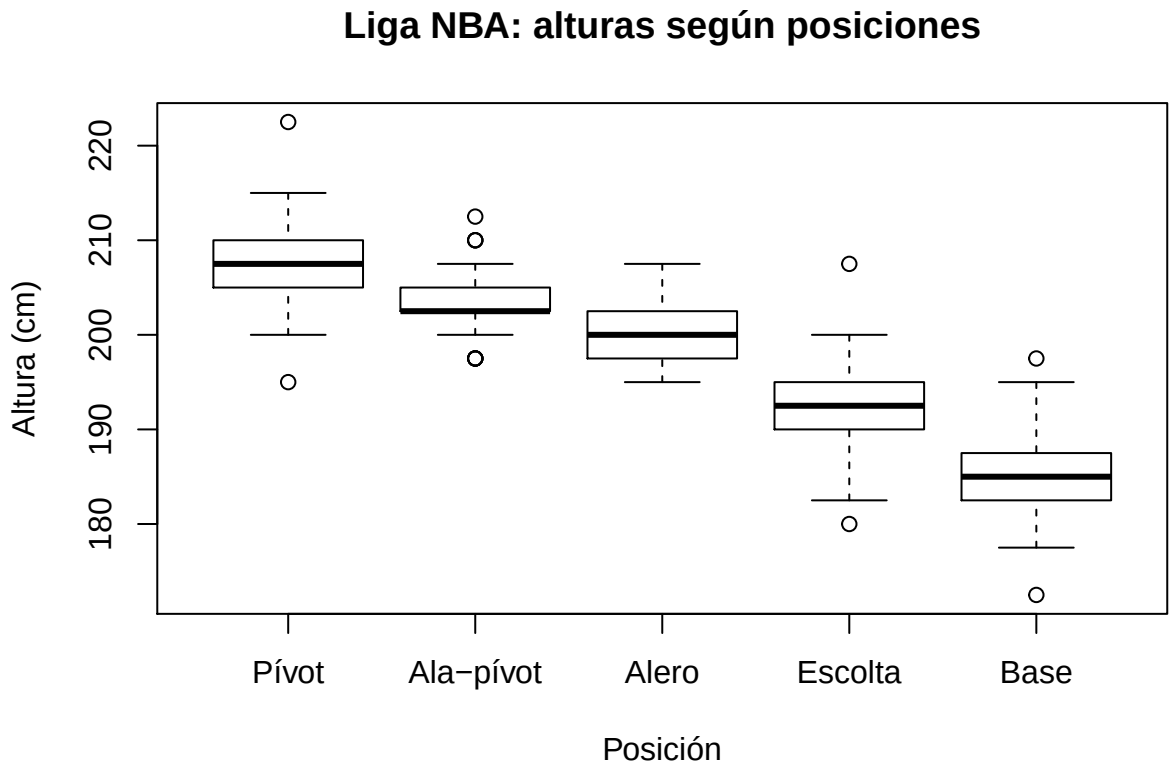
Representamos el intervalo media±desviación mediante dos líneas verticales naranjas.

```r
hist(alturas, main = "Liga NBA: alturas de jugadores",
     xlab = "Altura (cm)", ylab = "Frecuencia", col = "lightblue")
lines(rep(media, 101), 0:100, col = "green")
lines(rep(media - desviacion, 101), 0:100, col = "orange")
lines(rep(media + desviacion, 101), 0:100, col = "orange")
```



iv) Comparamos las alturas según las posiciones mediante un diagrama de cajas.

```
alturasC <- subset(datos$Height, !is.na(datos$Height) & datos$Pos == "C")
alturasPF <- subset(datos$Height, !is.na(datos$Height) & datos$Pos == "PF")
alturasPG <- subset(datos$Height, !is.na(datos$Height) & datos$Pos == "PG")
alturasSF <- subset(datos$Height, !is.na(datos$Height) & datos$Pos == "SF")
alturasSG <- subset(datos$Height, !is.na(datos$Height) & datos$Pos == "SG")
boxplot(alturasC, alturasPF, alturasSF, alturasSG, alturasPG, xlab = "Posición",
        ylab = "Altura (cm)", main = "Liga NBA: alturas según posiciones",
        names = c("Pívot", "Ala-pívot", "Alero", "Escolta", "Base"))
```



v) Observamos cierta asimetría negativa en el histograma de las alturas, con lo que la media es menor que la moda. Del diagrama de cajas, sabemos que, en general, los jugadores con mayor altura son los pívots, seguidos de los ala-pívots, los aleros, los escoltas y, por último, los bases.

Problema 2. *Descarga los datos de extensión de hielo ártico del directorio:*

ftp://sidads.colorado.edu/DATASETS/NOAA/G02135/north/daily/data/

i) Carga los datos en R. Recomendación: Elimina la segunda línea del archivo.

ii) Averigua en qué meses suelen ser las extensiones máximas y mínimas.

iii) Representa gráficamente los datos de extensión medianos de cada mes.

iv) Añade una región sombreada correspondiente a los percentiles del 5% y 95% de cada mes.

v) Añade una línea correspondiente a los datos de 2019 y otra para los datos de 2012.

vi) Comenta brevemente la gráfica.

Solución.

i) Cargamos los datos eliminando la segunda línea del archivo.

```
all_content <- readLines("N_seaice_extent_daily_v3.0.csv")
skip_second <- all_content[-2]
datos <- read.csv(textConnection(skip_second))
```

ii) Calculamos los meses en que se alcanzan las extensiones de hielo máximas y mínimas en el Ártico.

```
meses_maximo <- c()
meses_minimo <- c()
for(i in unique(datos$Year)){
  extension_year <- subset(datos$Extent, datos$Year == i)
  maximo <- max(extension_year)
  minimo <- min(extension_year)
```

```
  mes_maximo <- subset(datos$Month, datos$Year == i & datos$Extent == maximo)
  mes_minimo <- subset(datos$Month, datos$Year == i & datos$Extent == minimo)
  meses_maximo <- append(meses_maximo, mes_maximo)
  meses_minimo <- append(meses_minimo, mes_minimo)
}
```

Tenemos las siguientes frecuencias de meses con extensión de hielo máxima.

```
table(meses_maximo)
```

```
## meses_maximo
##  2  3 12
##  9 34  1
```

Observamos que marzo suele ser el mes con la extensión de hielo máxima en el Ártico con 34 veces frente a nueve de febrero, una de diciembre y ninguna del resto. Las frecuencias de meses con extensión de hielo mínima son las siguientes.

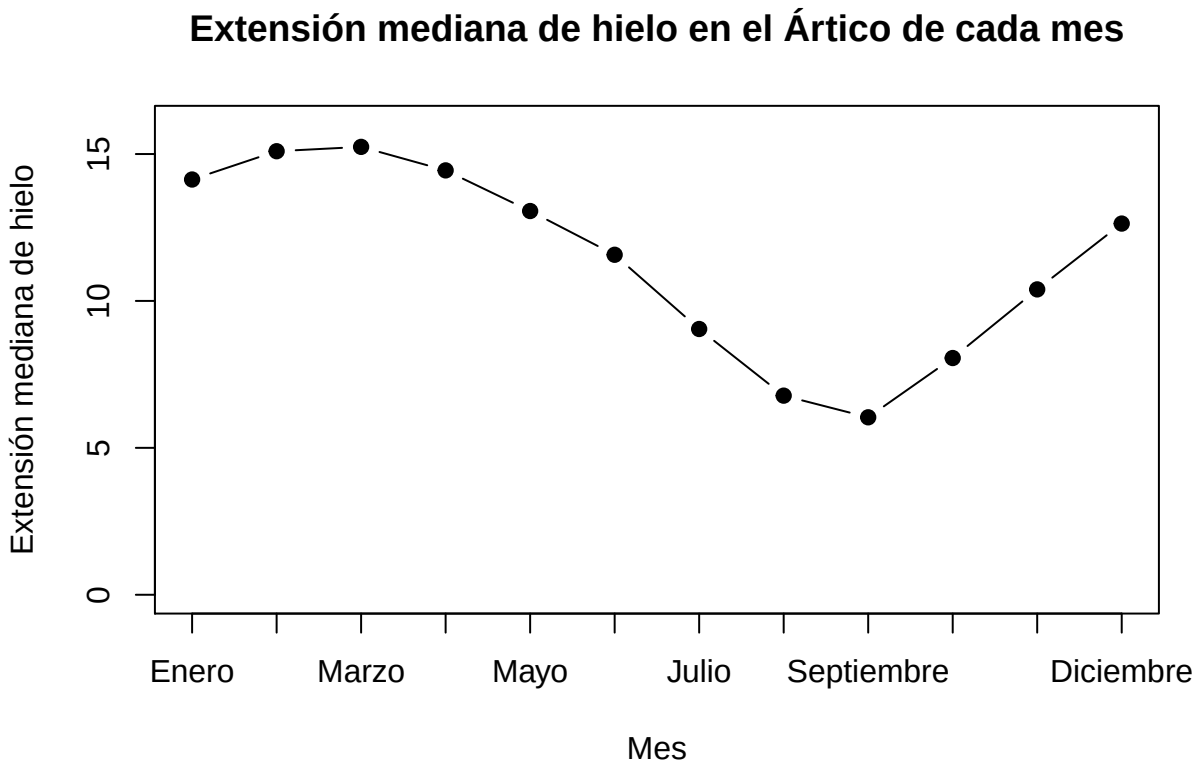
```
table(meses_minimo)
```

```
## meses_minimo
##  1  9 10
##  1 42  1
```

Septiembre suele ser el mes con la extensión de hielo mínima en el Ártico con 42 veces frente a una de enero y de octubre y ninguna del resto.

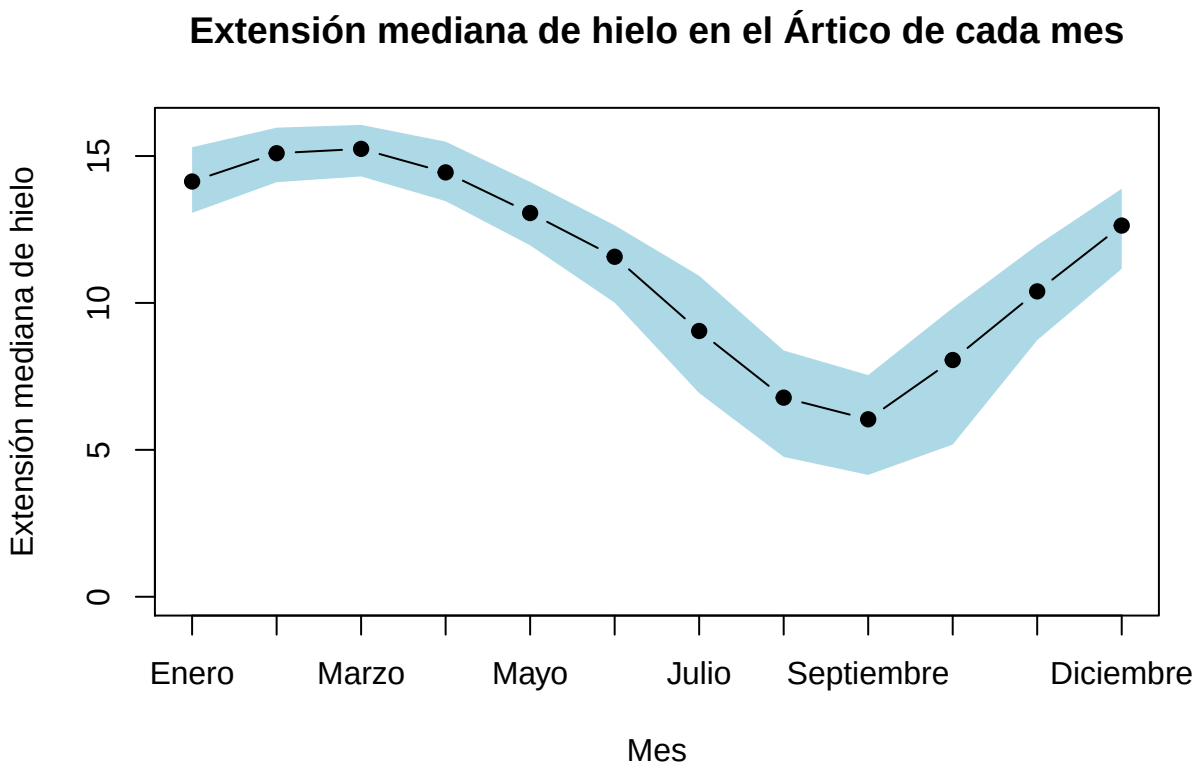
iii) Calculamos la mediana de cada mes y las representamos mediante un diagrama de puntos y lineas.

```
medianas <- c()
for(i in sort(unique(datos$Month))){
  extent.month <- subset(datos$Extent, datos$Month == i)
  medianas <- append(medianas, median(extent.month))
}
plot(medianas, xaxt = "n", ylab = "Extensión mediana de hielo", xlab = "Mes",
     main = "Extensión mediana de hielo en el Ártico de cada mes", type = "n",
     ylim = c(0,16))
axis(1, main = "Extensión mediana de hielo en el Ártico de cada mes",
     labels = c("Enero","Febrero","Marzo","Abril","Mayo","Junio","Julio","Agosto",
                "Septiembre","Octubre","Noviembre","Diciembre"),
     at=c(1,2,3,4,5,6,7,8,9,10,11,12))
lines(medianas, type = "b", pch = 19)
```



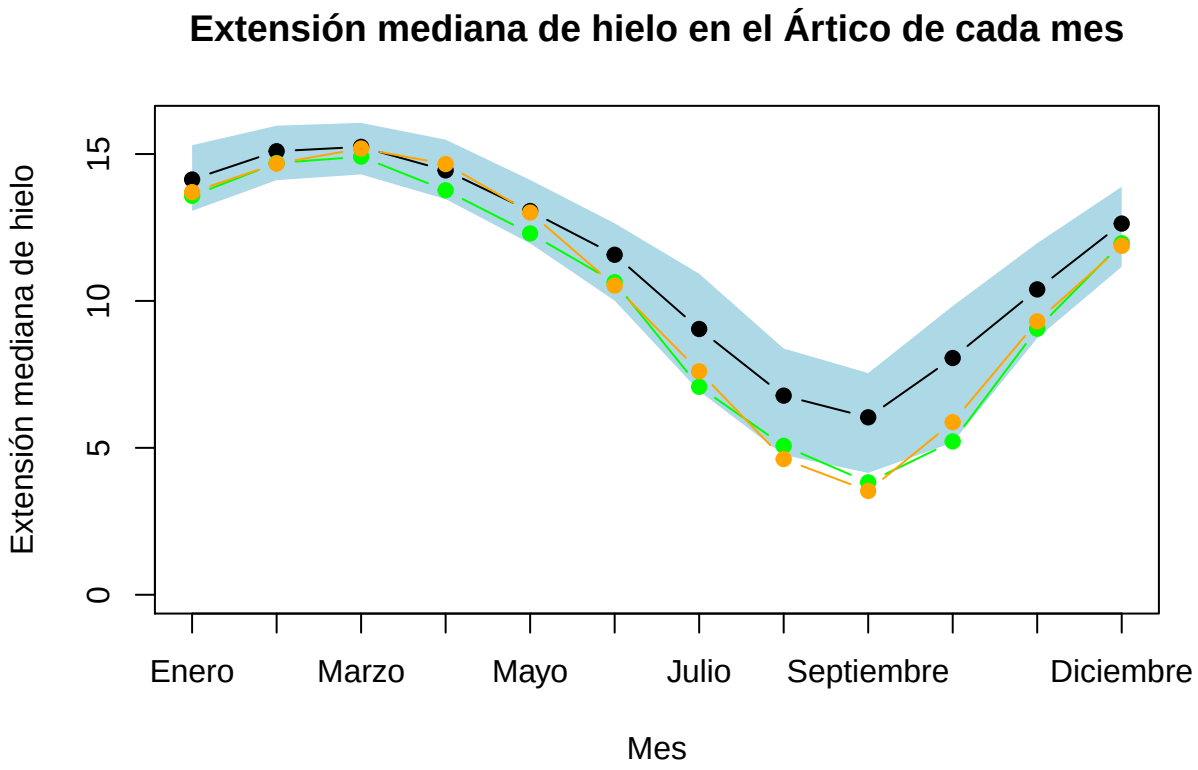
iv) Añadimos una región sombreada en azul entre los percentiles 5% y 95%.

```
medianas <- c()
percentil5 <- c()
percentil95 <- c()
for(i in sort(unique(datos$Month))){
  extent.month <- subset(datos$Extent, datos$Month == i)
  medianas <- append(medianas, median(extent.month))
  percentil5 <- append(percentil5, quantile(extent.month, 0.05))
  percentil95 <- append(percentil95, quantile(extent.month, 0.95))
}
plot(medianas, xaxt = "n", ylab = "Extensión mediana de hielo", xlab = "Mes",
     main = "Extensión mediana de hielo en el Ártico de cada mes", type = "n",
     ylim = c(0,16))
axis(1, main = "Extensión mediana de hielo en el Ártico de cada mes",
     labels = c("Enero","Febrero","Marzo","Abril","Mayo","Junio","Julio","Agosto",
                "Septiembre","Octubre","Noviembre","Diciembre"),
     at=c(1,2,3,4,5,6,7,8,9,10,11,12))
polygon(c(1:12, 12:1), c(percentil5, rev(percentil95)), col = "lightblue",
        border = "NA")
lines(medianas, type = "b", pch = 19)
```



v) Añadimos en verde las extensiones medianas en 2020 y en naranja las extensiones medianas en 2012.

```
medianas <- c(); percentil5 <- c(); percentil95 <- c();
extension2020 <- c(); extension2012 <- c()
for(i in sort(unique(datos$Month))){
  extent.month <- subset(datos$Extent, datos$Month == i)
  medianas <- append(medianas, median(extent.month))
  percentil5 <- append(percentil5, quantile(extent.month, 0.05))
  percentil95 <- append(percentil95, quantile(extent.month, 0.95))
  extension2020 <-
    append(extension2020,
           median(subset(datos$Extent, datos$Year == 2020 & datos$Month == i)))
  extension2012 <-
    append(extension2012,
           median(subset(datos$Extent, datos$Year == 2012 & datos$Month == i)))
}
plot(medianas, xaxt = "n", ylab = "Extensión mediana de hielo", xlab = "Mes",
     main = "Extensión mediana de hielo en el Ártico de cada mes", type = "n",
     ylim = c(0,16))
axis(1, main = "Extensión mediana de hielo en el Ártico de cada mes",
     labels = c("Enero","Febrero","Marzo","Abril","Mayo","Junio","Julio","Agosto",
                "Septiembre","Octubre","Noviembre","Diciembre"),
     at=c(1,2,3,4,5,6,7,8,9,10,11,12))
polygon(c(1:12, 12:1), c(percentil5, rev(percentil95)), col = "lightblue",
        border = "NA")
lines(medianas, type = "b", pch = 19)
lines(extension2020, type = "b", col = "green", pch = 19)
lines(extension2012, type="b", col = "orange", pch = 19)
```



vi) La gráfica concuerda con los cálculos anteriores ya que se observa que los meses en que se suelen alcanzar las extensiones de hielo máximas y mínimas son, efectivamente, marzo y septiembre respectivamente. Vemos claramente en la gráfica el mínimo histórico desde que hay registros, que sucedió en septiembre de 2012. En el año 2020 la mínima extensión de hielo también se alcanzó en septiembre y se acercó al mínimo histórico.

Práctica 3

Problema 1. *Considera un examen de* 10 *preguntas con* 5 *posibles respuestas para cada pregunta en el que se responde al azar. Considera la variable aleatoria* $X =$ número de preguntas acertadas en el examen.

i) ¿Cuánto valen $\mathbb{E}(X)$ *y* $\mathrm{SD}(X)$*?*

ii) Calcula la probabilidad de acertar $0, 1, 2, \ldots, 10$ *preguntas. ¿Cuánto vale la suma de las once probabilidades que acabas de calcular?*

iii) Calcula las probabilidades acumuladas, es decir, la probabilidad de obtener 0 *aciertos o menos,* 1 *acierto o menos,* 2 *aciertos o menos...* 10 *aciertos o menos.*

iv) Representa las probabilidades de los apartados anteriores en una gráfica con dos paneles. Guarda la figura en una archivo. ¿Cuál el número de aciertos más probable?

Solución.

i) X es una variable aleatoria con distribución binomial de parámetros $n = 10$ y $p = \frac{1}{5}$ porque hay 10 preguntas y 5 respuestas para cada una siendo solo una correcta.

La esperanza es

$$\mathbb{E}(X) = np = 10 \cdot \frac{1}{5} = 2.$$

La desviación típica es

$$\mathrm{SD}(X) = \sqrt{np(1-p)} = \sqrt{10\frac{1}{5}\left(1 - \frac{1}{5}\right)} = \sqrt{2\frac{4}{5}} = \frac{2}{5}\sqrt{10}.$$

ii) La función de probabilidad de la binomial está dada por $P(X = x) = \binom{n}{x} p^x (1-p)^x$. Calculamos la probabilidad $P(X = x)$ para cada $x \in \{1, \ldots, 10\}$ en R con `dbinom`.

```
prob <- dbinom(0:10, 10, 1/5)
prob
```

```
##  [1] 0.1073741824 0.2684354560 0.3019898880 0.2013265920 0.0880803840
##  [6] 0.0264241152 0.0055050240 0.0007864320 0.0000737280 0.0000040960
## [11] 0.0000001024
```

La suma de las probabilidades que acabamos de calcular es necesariamente 1.

```
sum(prob)
```

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

Efectivamente es 1 porque es la suma de probabilidades de sucesos disjuntos cuya unión es el total.

iii) Podemos hallar la probabilidad de acertar al menos una cierta cantidad de preguntas con la función de distribución de la binomial. Calculamos la probabilidad $P(X \leq x)$ para cada $x \in \{0, \ldots, 10\}$.

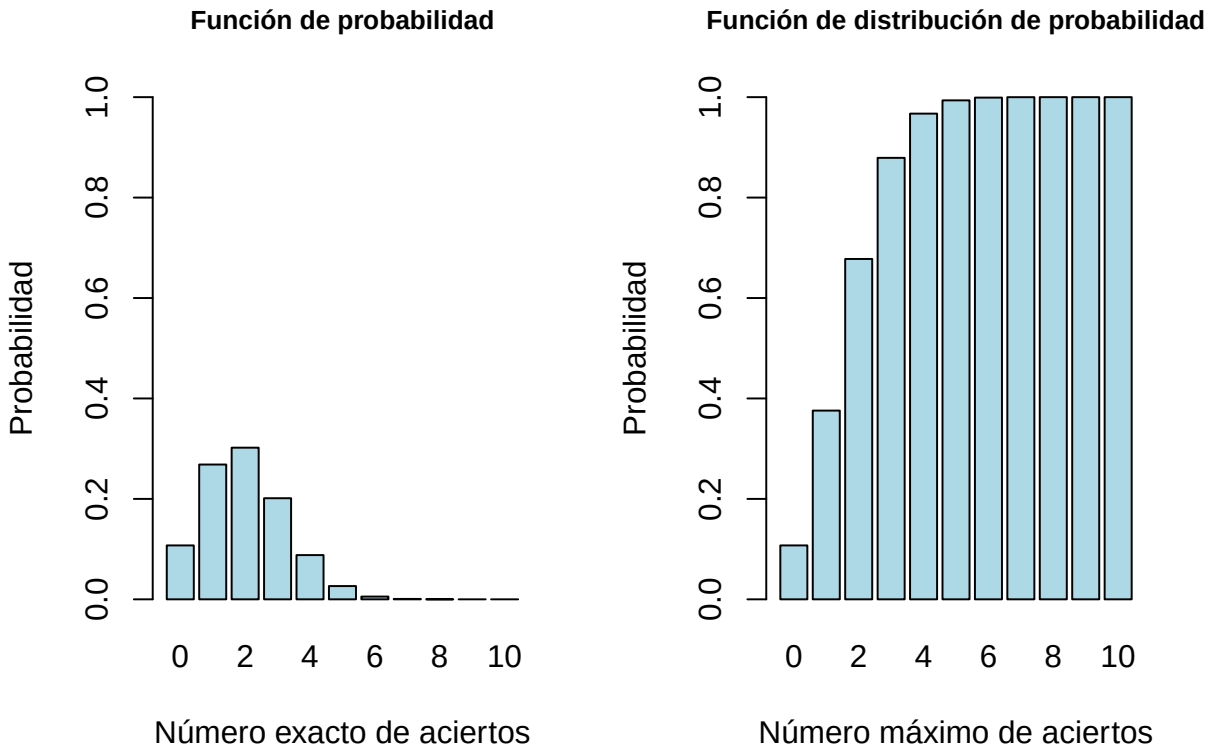
```
prob_acum <- pbinom(0:10, 10, 1/5)
prob_acum
```

```
##  [1] 0.1073742 0.3758096 0.6777995 0.8791261 0.9672065 0.9936306 0.9991356
##  [8] 0.9999221 0.9999958 0.9999999 1.0000000
```

iv) Representamos las probabilidades anteriores mediante diagramas de barras.

```
#pdf("figura.pdf")  # Guarda la figura en un archivo.
par(mfrow=c(1,2))
main <- "Función de probabilidad"; main2 <- "Función de distribución de probabilidad"
yetiq <- "Probabilidad"
```

```r
xetiq <- "Número exacto de aciertos"; xetiq2 <- "Número máximo de aciertos"
nombres <- c("0","1","2","3","4","5","6","7","8","9","10")
color <- "lightblue"
ylimite <- c(0,1)
barplot(prob, names = nombres, ylab = yetiq, xlab = xetiq, ylim = ylimite,
        main = main, col = color, cex.main = 0.8)
barplot(prob_acum, names = nombres, ylab = yetiq, xlab = xetiq2, ylim = ylimite,
        main = main2, col = color, cex.main = 0.8)
#dev.off()
```



Observamos que el número de aciertos más probable es 2 si cada pregunta se resuelve de manera completamente aleatoria.

Problema 2. *En el archivo* `results.csv` *figuran los resultados de* 40838 *partidos de fútbol internacionales. En la columna* `home_score` *figuran los goles marcados por los equipos locales en cada partido. En la columna* `home_team` *figura el nombre del equipo local en cada partido.*

i) Considera los datos de goles de España como equipo local. Razona qué distribución usarías para calcular probabilidades asociadas a esta variable. Representa gráficamente los datos de forma que te ayude a decidir qué distribución usar.

ii) Estima la probabilidad de que en el siguiente partido como local España marque más de 3 *goles. Explica qué suposiciones has hecho. ¿Podría fallar alguna de las suposiciones?*

iii) Calcula la probabilidad de que en el siguiente partido como local España marque exactamente $0, 1, 2, \ldots, 10$ *goles. Representa las probabilidades en una gráfica. ¿Cuál es el número de goles más probable? ¿Cuánto vale la suma de las once probabilidades? ¿Por qué?*

iv) ¿Cuánto valen $\mathbb{E}(X)$ *y* $\mathrm{SD}(X)$*? Explica su significado en este contexto.*

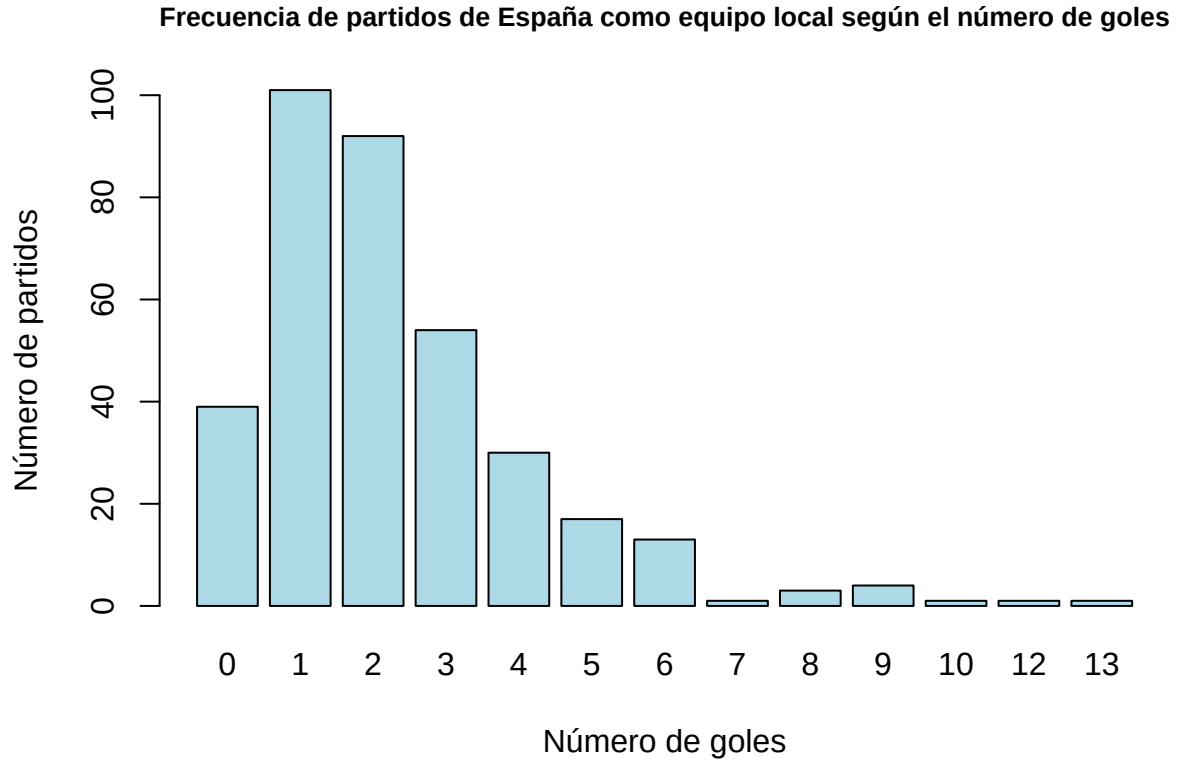
v) Una casa de apuestas paga 4 Euros por cada Euro apostado (ganancia neta = 3 *Euros) si España marca más de* 3 *goles en el próximo partido como local. En caso de que marcase* 3 *goles o menos perderíamos el*

Euro apostado. Razona matemáticamente si es rentable apostar.

Solución.

i) Cargamos los datos en R y representamos los partidos de España como equipo local según el número de goles.

```
datos <- read.csv("results.csv")
gol_esp_local <- subset(datos$home_score, datos$home_team == "Spain")
main <- "Frecuencia de partidos de España como equipo local según el número de goles"
barplot(table(gol_esp_local), ylim = c(0,100), ylab = "Número de partidos",
        xlab = "Número de goles", col = "lightblue", main = main, cex.main = 0.8)
```



Para calcular probabilidades asociadas a la variable X usaríamos la distribución de Poisson de media aproximada por la media muestral.

```
mean(gol_esp_local)
```

```
## [1] 2.330532
```

ii) Estimamos la probabilidad de que en el siguiente partido como local España marque más de 3 goles con la función de distribución de la Poisson de media 2.330532.

```
1 - ppois(3, 2.330532)
```

```
## [1] 0.2068887
```

Hemos supuesto que la variable se distribuye como una Poisson que, por hipótesis, supone que los sucesos son independientes entre sí. En este caso significaría que los goles marcados en un partido son independientes entre sí. Esta suposición podría fallar, por ejemplo, en un partido en el que tenga ventaja un equipo, que tendría mayor probabilidad de marcar más goles. También hemos supuesto que podemos aproximar la media de X a partir de la media muestral de los datos que tenemos, pero en general no se tiene la igualdad.

iii) Calculamos la probabilidad de que en el siguiente partido como local España marque exactamente $0, 1, \ldots, 10$ goles usando la función de probabilidad de la Poisson de media $\lambda = 2.330532$.

```
prob <- dpois(0:10, 2.330532)
prob
```

```
##  [1] 0.0972439995 0.2266302527 0.2640845280 0.2051524811 0.1195286055
##  [6] 0.0557130480 0.0216401735 0.0072047310 0.0020988570 0.0005434948
## [11] 0.0001266632
```

La suma de las probabilidades es

```
sum(prob)
```

```
## [1] 0.9999668
```

Coincide claramente con la función de distribución evaluada en 10, es decir, la probabilidad de no marcar ningún gol, o marcar un gol,... o marcar diez goles, que es una probabilidad muy alta. El complementario, marcar más de diez goles tiene una probabilidad positiva baja.

iv) Como estamos identificando la distribución de X como una Poisson de media $\lambda = 2.330532$, la esperanza $\mathbb{E}(X)$, que en general es desconocida, es la media muestral. Por lo tanto, la desviación típica teórica sería:

$$\mathrm{SD}(X) = \sqrt{\mathbb{E}\left(X - \mathbb{E}(X)\right)} = \lambda = 2.330532$$

En este caso la interpretación es que en promedio España marca 2.330532 goles por partido al jugar como local y que en promedio las diferencias con la media son de 2.330532 partidos.

v) Consideramos la variable aleatoria Y dada por la diferencia entre el dinero ganado y el dinero apostado, es decir, la ganancia neta. Es rentable apostar si y solo si la esperanza de Y es positiva, lo que significa que en promedio la ganancia neta es positiva y ganamos más de lo que gastamos.

$$\mathbb{E}(Y) = 3P(Y = 3) - P(Y = 4) = 3P(X > 3) - P(X \leq 3).$$

Calculamos $\mathbb{E}(Y)$.

```
pmenos3 <- ppois(3, 2.330532)
3 * (1 - pmenos3) - pmenos3
```

```
## [1] -0.172445
```

Como $\mathbb{E}(Y) < 0$, no es rentable apostar.

Problema 3. *Estamos esperando un tren que tiene programada su llegada en breves minutos. Definimos la variable aleatoria X como el tiempo de espera en minutos. La variable aleatoria X tiene función densidad de probabilidad $f(x) = hx$ si $x \in [0, 1)$, $f(x) = h$ si $x \in [1, 5)$, $f(x) = h(6 - x)$ si $x \in [5, 6]$, y $f(x) = 0$ en los demás casos.*

i) Representa gráficamente $f(x)$ y calcula la constante h.

ii) Calcula $P(X < 1)$, $P(X \leq 1)$, $P(0 < X \leq 3)$.

iii) ¿Cuánto vale $\mathbb{E}(X)$?

Solución.

i) Calculamos la constante K. Como $f : \mathbb{R} \longrightarrow \mathbb{R}$ es una función de densidad de probabilidad, su integral sobre $\mathbb{R}$ es 1.

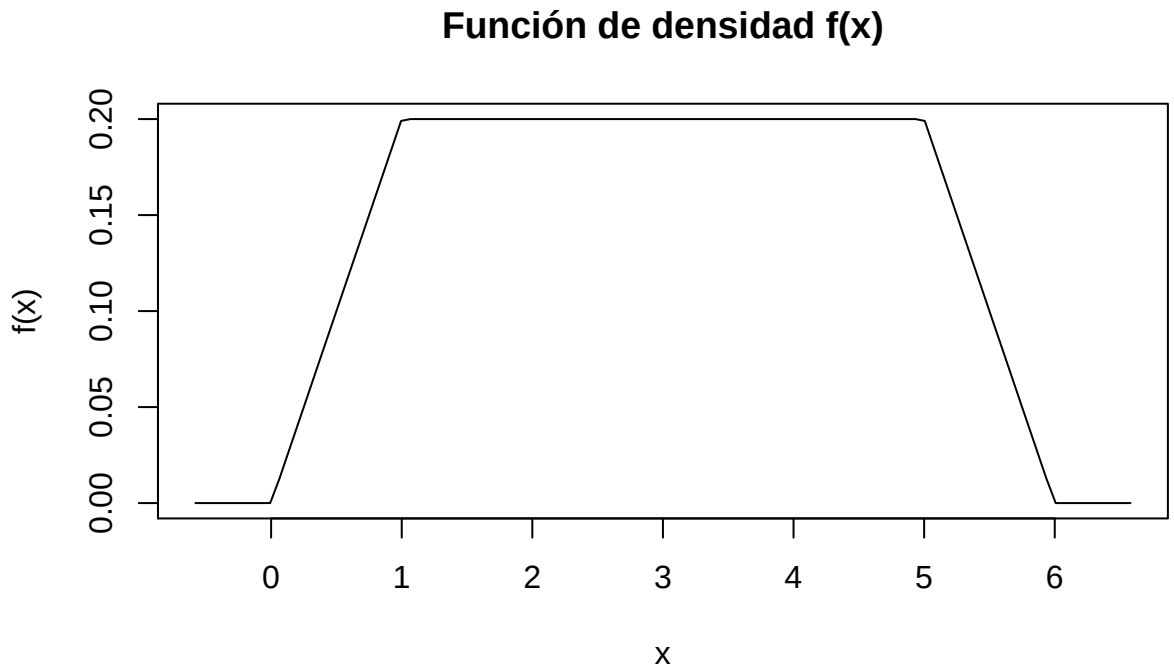
$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{1} hx\,dx + \int_{1}^{5} h\,dx + \int_{5}^{6} h(6-x)\,dx + \int_{6}^{\infty} 0\,dx =$$
$$= \frac{h}{2}\left[x^2\right]_0^1 + h[x]_1^5 + 6h[x]_5^6 - \frac{h}{2}\left[x^2\right]_5^6 = \frac{h}{2} + 4h + 6h - \frac{h}{2}(36 - 25) = 5h = 1 \longrightarrow h = \frac{1}{5}.$$

Por lo tanto, la función de densidad está dada por

$$f(x) = \begin{cases} \frac{1}{5}x & \text{si } x \in [0,1), \\ \frac{1}{5} & \text{si } x \in [1,5), \\ \frac{1}{5}(6-x) & \text{si } x \in [5,6], \\ 0 & \text{en otro caso.} \end{cases}$$

La representamos gráficamente.

```
# Definimos la función de densidad a trozos.
f <- function(x){
  if(0<=x & x<1){
    return((1/5)*x)
  } else if(1<= x & x<5){
    return(1/5)
  } else if(5<=x & x<=6){
    return((1/5)*(6-x))
  } else{
    return(0)
  }
}
# Vectorizamos f y la graficamos.
f <- Vectorize(f)
curve(f, -0.58, 6.58, main = "Función de densidad f(x)", ylab = "f(x)")
```



ii) Calculamos las probabilidades integrando la función de densidad.

$$P(X<1) = \int_{-\infty}^{1} f(x)\,dx = \int_{-\infty}^{0} 0\,dx = \int_{0}^{1} \frac{1}{5}x\,dx = \frac{1}{10}\left[x^2\right]_0^1 = \frac{1}{10}$$

Se cumple que $P(X<1) = P(X \leq 1)$ ya que P es una distribución de probabilidad continua y $P(X=1)=0$ porque es la probabilidad de un conjunto de medida nula.

$$P(0<x\leq 3) = \int_{0}^{3} f(x)\,dx = \int_{-\infty}^{0} 0\,dx = \int_{0}^{1} \frac{1}{5}x\,dx + \int_{1}^{3} \frac{1}{5}\,dx = \frac{1}{10} + \frac{1}{5}[x]_1^3 = \frac{1}{10} + \frac{2}{5} = \frac{1}{2}$$

iii) Calculamos la esperanza de X con la definición.

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{-\infty}^{0} 0x\,dx + \int_{0}^{1} \frac{1}{5}x^2\,dx + \int_{1}^{5} \frac{1}{5}x\,dx + \int_{5}^{6} \frac{1}{5}x(6-x)\,dx + \int_{6}^{\infty} 0x\,dx =$$
$$= \frac{1}{15}\left[x^3\right]_0^1 + \frac{1}{10}\left[x^2\right]_1^5 + \frac{3}{5}\left[x^2\right]_5^6 - \frac{1}{15}\left[x^3\right]_5^6 = \frac{1}{15} + \frac{1}{10}24 + \frac{3}{5}11 - \frac{1}{15}11 = 8,\overline{3}$$

Problema 4. *El colesterol se mide en una población con media* 180 *mg/dL y desviacón típica* 25 *mg/dL.*

i) Calcula la probabilidad de que una persona tenga más de 225 *mg/dL de colesterol en sangre.*

ii) Calcula el percentil del 95% *de colesterol en sangre en dicha población.*

iii) Representa la función de densidad y la función de distribución en una gráfica con dos paneles. Guarda la figura en un archivo.

iv) ¿Cuánto valen $\mathbb{E}(X)$ *y* $\mathrm{SD}(X)$*?*

Solución.

El nivel de colesterol en sangre X es una variable biológica cuya distribución de probabilidad podemos suponer que es una normal, en este caso, $X \sim N(180, 25)$.

i) Calculamos la probabilidad $P(X > 225) = 1 - P(X \leq 225)$.

```
1 - pnorm(225, 180, 25)
```

```
## [1] 0.03593032
```

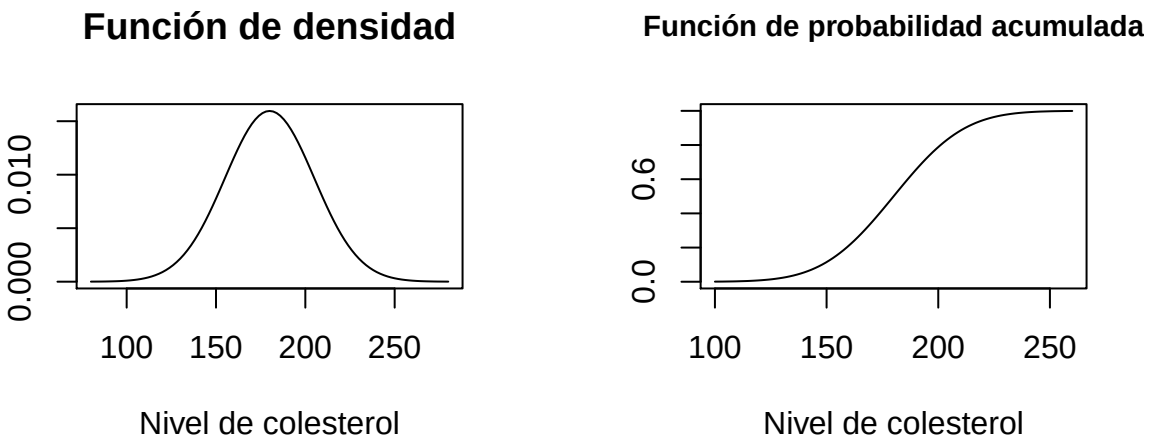
ii) Calculamos el percentil 95%.

```
qnorm(0.95, 180, 25)
```

```
## [1] 221.1213
```

iii) Representación de las funciones de densidad y distribución.

```
#pdf("figura2.pdf")   # Guarda la figura en un archivo.
par(mfrow = c(1, 2))
color <- "lightblue"; xl <- "Nivel de colesterol"
tit1  <- "Función de densidad"; tit2 <- "Función de probabilidad acumulada"
curve(dnorm(x, 180, 25), 80, 280, main = tit1, xlab = xl, ylab = NA)
curve(pnorm(x, 180, 25), 100, 260, main = tit2, xlab = xl, ylab = NA, cex.main = 0.9)
#dev.off()
```



iv) Como la variable se distribuye como una normal $N(180, 25)$, la esperanza es $\mathbb{E}(X) = 180$ y la desviación típica es $SD(X) = 25$.

Práctica 4

Problema 1. *Estamos esperando un tren que llegará en los próximos 100 minutos pero desconocemos el momento exacto. Definimos la variable aleatoria X = "tiempo de espera en minutos". Realiza las siguientes tareas:*

i) Considera que X es una variable aleatoria uniforme en el intervalo $[0, 100]$. ¿Cuánto vale $\mathbb{E}(X)$ y $\mathrm{Var}(X)$?

ii) Simula los tiempos de espera de una muestra aleatoria de personas con tamaño muestral $n = 30$ y calcula la media muestral, $\bar{X}$ y la varianza muestral S^2.

iii) ¿Cuanto valen $\mathbb{E}(\bar{X})$ y $\mathrm{Var}(\bar{X})$ para $n = 30$? ¿Y si aumentáramos a $n = 100$?

iv) Simula 500 muestras aleatorias de tamaño $n = 30$ de la variable X. Recomendación: Introduce las muestras en una matriz de 30 filas y 500 columnas.

v) Calcula las 500 medias muestrales, es decir, $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{500}$. Recomendación: Utiliza el comando `colMeans`.

vi) Calcula la media y varianza de las 500 medias muestrales y compáralas con los valores obtenidos en el tercer apartado.

vii) Repite los tres pasos anteriores para 500 muestras aleatorias de tamaño $n = 100$.

viii) Representa las 500 medias de tamaño $n = 30$ y las 500 medias de tamaño $n = 100$ mediante dos histogramas en una gráfica de dos paneles.

ix) Superpón a cada histograma la función de densidad normal con $\mu = \mathbb{E}(X)$ y $\sigma^2 = \mathrm{Var}(X)/n$.

x) Comenta el resultado.

Solución.

i) Si X es una distribución uniforme en el intervalo $[a, b]$, entonces

$$\mathbb{E}(X) = \frac{a+b}{2}, \qquad \mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

En este caso, tenemos que

$$\mathbb{E}(X) = \frac{100}{2} = 50, \qquad \mathrm{Var}(X) = \frac{100^2}{12} = 833,\bar{3}.$$

ii) Calculamos los tiempos de espera de una muestra aleatoria de personas con tamaño muestral $n = 30$ usando el comando `runif`.

```
n <- 30
tiempos30 <- runif(n, 0, 100)
mean(tiempos30)
```

```
## [1] 51.853
```

```
var(tiempos30)
```

```
## [1] 935.9177
```

iii) Sabemos que la media $\bar{X}$ es una variable aleatoria con esperanza $\mathbb{E}(\bar{X}) = \mathbb{E}(X)$ y, si la población es infinita, varianza $\mathrm{Var}(X) = \frac{\mathrm{Var}(X)}{n}$ porque es un muestreo de variables aleatorias independientes e igualmente distribuidas, por lo tanto, la esperanza y la aproximación de la varianza con $n = 30$ son

$$\mathbb{E}(\bar{X})) = \mathbb{E}(X) = 50, \qquad \mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{833,\bar{3}}{30} = 27.\bar{7}$$

Si aumentamos a $n = 100$, tenemos

$$\mathbb{E}(\bar{X}) = \mathbb{E}(X) = 50, \qquad \mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{\mathrm{Var}(X)}{100} = \frac{833,\bar{3}}{100} = 8,\bar{3}.$$

iv) Simulamos 500 muestras aleatorias de tamaño $n = 30$ de X.

```
datos30 <- matrix(nrow = 30, ncol = 500)
for(i in 1:30){
  datos30[i,] <- runif(500, 0, 100)
}
```

v) Calculamos las 500 medias muestrales, es decir, $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{500}$ utilizando el comando `colMeans`.

```
medias30 <- colMeans(datos30)
```

vi) Calculamos la media y la varianza de las medias muestrales

```
mean(medias30)
```

```
## [1] 50.19477
```

```
var(medias30)
```

```
## [1] 27.83071
```

Comparándolas con los valores del apartado *iii*) vemos que efectivamente se aproximan, como era de esperar.

vii) Repetimos lo mismo para 500 muestras de tamaño $n = 100$.

```
datos100 <- matrix(nrow = 100, ncol = 500)
for(i in 1:100){
  datos100[i,] <- runif(500, 0, 100)
}
medias100 <- colMeans(datos100)
mean(medias100)
```

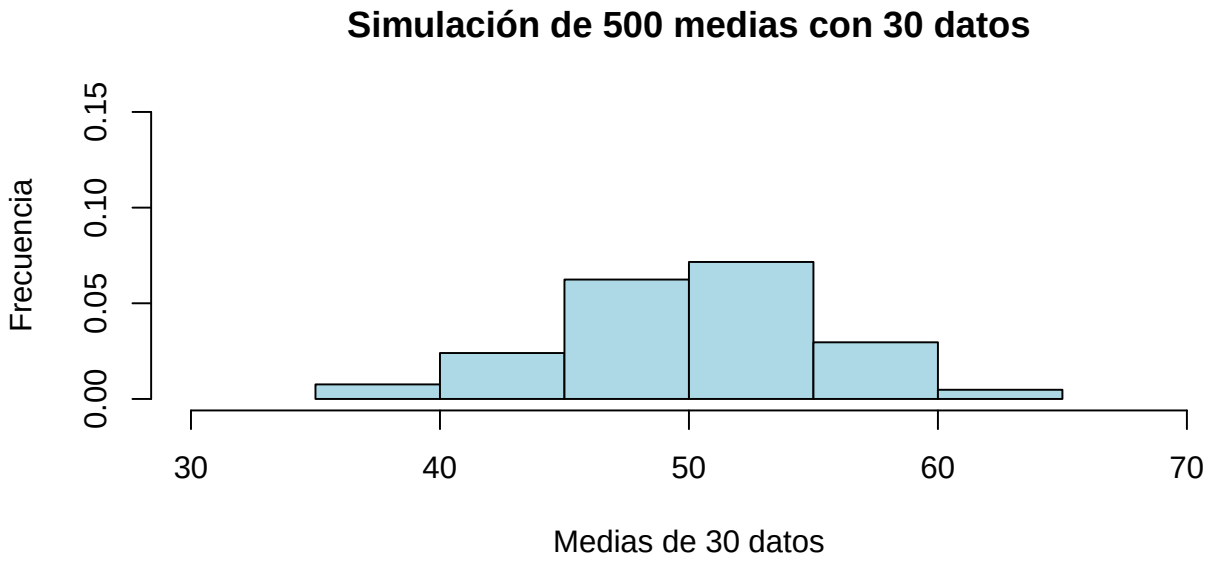
```
## [1] 49.91808
```

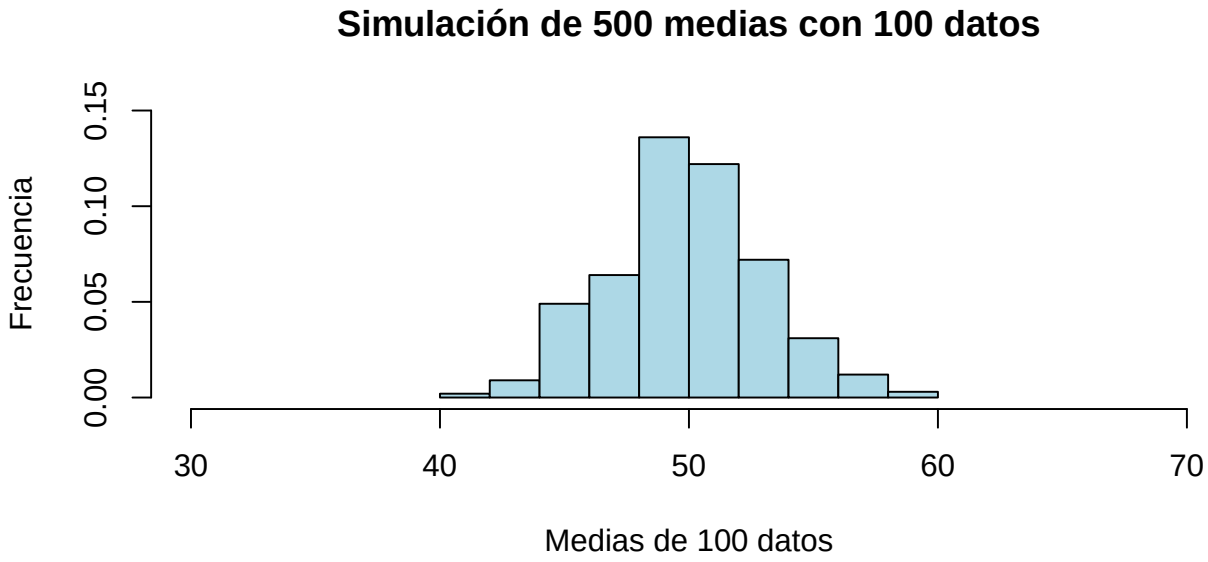
```
var(medias100)
```

```
## [1] 9.367999
```

viii) Representamos la simulación de las 500 medias con 30 y 100 datos en una gráfica con dos paneles.

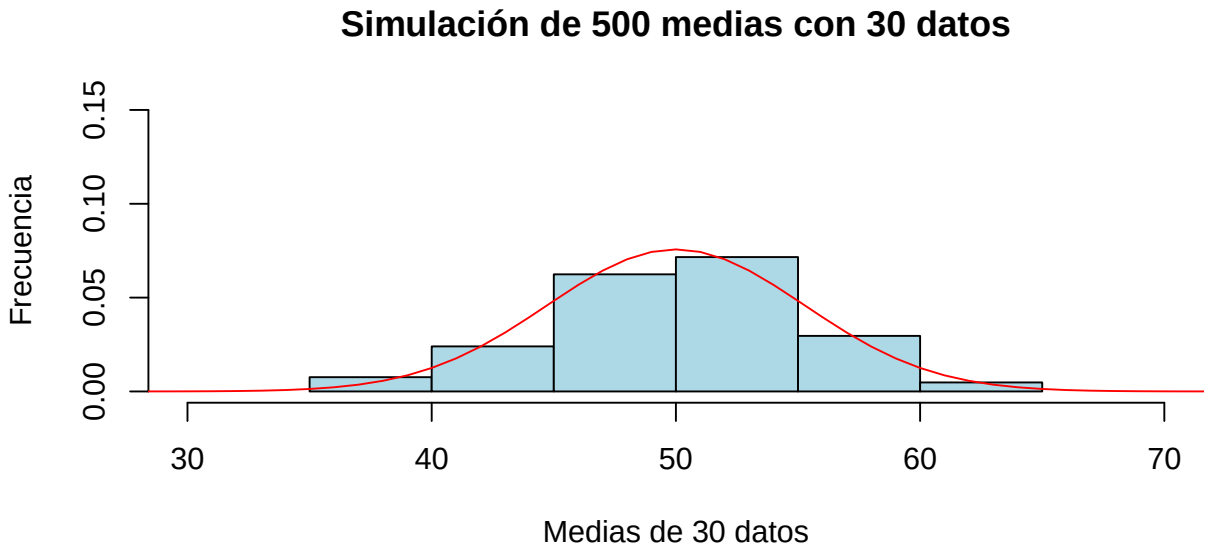
```
par(mfrow = c(2, 1))
hist(medias30, xlim = c(30,70), ylim=c(0, 0.15), xlab = "Medias de 30 datos",
     ylab = "Frecuencia", main = "Simulación de 500 medias con 30 datos",
     col = "lightblue", freq = FALSE)
hist(medias100, xlim = c(30,70), ylim = c(0, 0.15), xlab = "Medias de 100 datos",
     ylab = "Frecuencia", main = "Simulación de 500 medias con 100 datos",
     col = "lightblue", freq = FALSE)
```

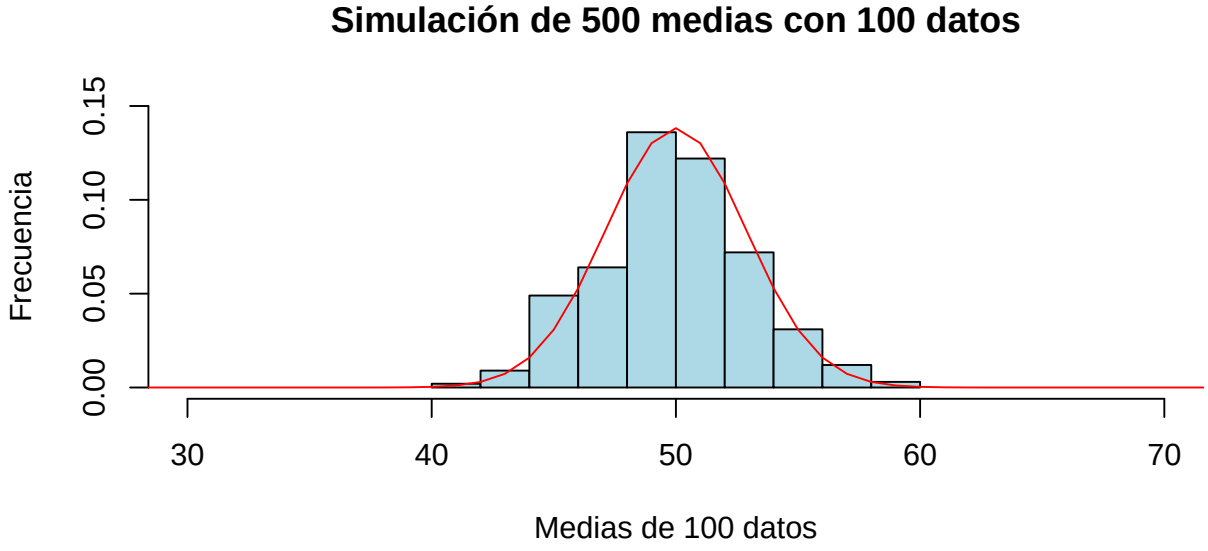




ix) Superponemos a cada histograma la función de densidad de la normal con $\mu = \mathbb{E}(X)$ y $\sigma^2 = \frac{\text{Var}(X)}{n}$.

```
par(mfrow = c(2, 1))
hist(medias30, xlim = c(30,70), ylim=c(0, 0.15), xlab = "Medias de 30 datos",
     ylab = "Frecuencia", main = "Simulación de 500 medias con 30 datos",
     col = "lightblue", freq = FALSE)
curve(dnorm(x, 50, sqrt(100^2/12/30)), 0, 100, col = "red", add = TRUE)
hist(medias100, xlim = c(30,70), ylim = c(0, 0.15), xlab = "Medias de 100 datos",
     ylab = "Frecuencia", main = "Simulación de 500 medias con 100 datos",
     col = "lightblue", freq = FALSE)
curve(dnorm(x, 50, sqrt(100^2/12/100)), 0, 100, col = "red", add = TRUE)
```





x) Observamos que hemos simulado una variable aleatoria normal de media 50 y varianza $100^2/12$ a partir de medias de un muestreo aleatorio uniforme en $[0, 100]$ porque la distribución obtenida se ajusta a la función de densidad de probabilidad de la normal.

Práctica 5

Problema 1. *Simula los siguientes experimentos aleatorios usando la función* `runif`*:*

i) Extraemos una carta al azar de una baraja española y vemos si es un basto o no.

ii) Repetimos n veces el experimento del apartado anterior y anotamos el número de bastos.

Solución.

i) Representamos el experimento mediante una variable aleatoria con distribución de Bernoulli de parámetro $p = \frac{1}{4}$ porque exactamente $1/4$ de las cartas son bastos y sacamos una. Tenemos que simular una distribución de Bernoulli a partir de una uniforme. Para ello, podemos simular números aleatorios uniformes $x \in (0, 1)$. Si $x \leq \frac{1}{4}$, entonces devolvemos `BASTO` y, en caso contrario, devolvemos `NO BASTO`.

```
carta <- runif(1,0,1)
basto <- {}
if (carta <= 1/4){
  basto <- "BASTO"
}else{
  basto <- "NO_BASTO"
}
basto
```

```
## [1] "NO_BASTO"
```

ii) Podemos repetir el el proceso anterior, por ejemplo, $N = 1000$ veces.

```
N = 10000
basto <- rep("NO_BASTO", N)
for(k in 1:N){
  carta <- runif(1,0,1)
  if(carta < 1/4){
    basto[k] <- "BASTO"
  }
}
table(basto)
```

```
## basto
##    BASTO NO_BASTO
##     2498     7502
```

Problema 2. *Considera una carrera ciclista con $\lambda = 0.01$ el número de caídas por kilómetro. Llamamos T al número de kilómetros recorridos hasta que se produce una caída y X al número de caídas en un intervalo t en kilómetros.*

i) Considerando que las caídas se producen de forma independiente a cuando ha sucedido la anterior caída, ¿qué distribución seguiría X?

ii) Calcula la probabilidad de que haya 0 caídas en t kilómetros.

iii) Calcula $P(T \leq t)$.

iv) ¿Cuál será la distribución de la variable T?

v) Simula 1000 distancias entre caídas.

vi) Compara gráficamente los números aleatorios con la distribución teórica.

Solución.

i) La variable aleatoria X sigue una distribución de probabilidad de Poisson de media λt ya que expresa la probabilidad de que se de un suceso con ocurrencia media λ en un intervalo de t kilómetros suponiendo

que cada suceso es independiente de los anteriores. Si X es una variable aleatoria con distribución de Poisson de media μ, entonces su función de probabilidad es

$$P(X = x) = \frac{e^{-\mu}\mu^k}{k!}.$$

En nuestro caso $\mu = \lambda t$, con lo que la función de probabilidad de X es

$$P(X = x) = \frac{e^{-\lambda t}\lambda^k t^k}{k!} = \frac{e^{-0.01t}0.01^k t^k}{k!}.$$

ii) Calculamos la probabilidad de que haya 0 caídas en t kilómetros aplicando la función de probabilidad anterior y obtenemos

$$P(T > t) = P(X = 0) = \frac{e^{-\lambda t}\lambda^0 t^0}{0!} = e^{-\lambda t} = e^{-0.01t}.$$

iii) Observamos que la probabilidad de que haya 0 caídas en t kilómetros coincide con la probabilidad de que el número de kilómetros T hasta que se produce una caída sea mayor que el intervalo de t kilómetros, con lo que

$$P(T > t) = P(X = 0) = e^{-\lambda t} = e^{-0.01t}.$$

Por lo tanto, $P(T \leq t)$ es la probabilidad del complementario y

$$P(T \leq t) = 1 - P(T > t) = 1 - e^{-\lambda t} = 1 - e^{-0.01t}.$$

Esta expresión es la función de distribución de T, que vemos que es una variable aleatoria con distribución exponencial de parámetro λ.

iv) Simulamos 1000 distancias entre caídas con el método de la función de distribución inversa. Tenemos que se cumple

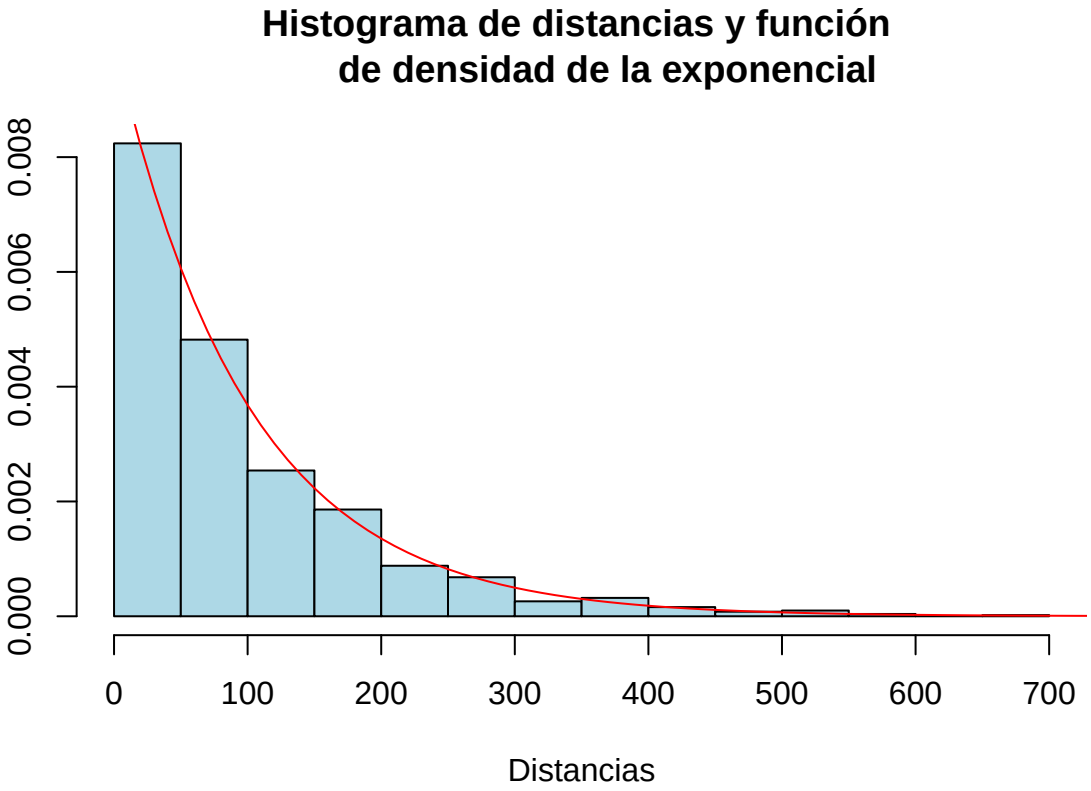
$$1 - e^{-\lambda t} = y \longrightarrow e^{-\lambda t} = 1 - y \longrightarrow -\lambda t = \ln(1 - y) \longrightarrow t = -\frac{1}{\lambda}\ln(1 - y),$$

con lo que la inversa de la función de distribución es $F^{-1}(x) = -\frac{1}{\lambda}\ln(1 - x) = -\frac{1}{0.01}\ln(1 - x)$.

```
N <- 1000
Finv <- function(p){
  Finv <- (-1/0.01) * log(1-p)
}
u <- runif(N, 0, 1)
distancias<- Finv(u)
```

v) Representamos el histograma de las distancias simuladas y superponemos la función de densidad de la exponencial de parámetro 0.01 calculada con el comando `dexp`.

```
hist(distancias, main = "Histograma de distancias y función
     de densidad de la exponencial", xlab = "Distancias",
     ylab = "", col = "lightblue", freq = F)
curve(dexp(x, 0.01), 0, 1000, col = "red", add = T)
```



Problema 3. *El tamaño angular en grados X de unos determinados objetos en el cielo es una variable aleatoria con función de densidad $f(x) = Kx^{-3}$ siendo $x > 1$ por razones físicas.*

i) Halla el valor de la constante $K > 0$.

ii) Genera 1000 números aleatorios siguiendo la distribución de X.

iii) Compara gráficamente los números aleatorios con la distribución teórica.

Solución.

i) Sabemos que la integral de la función de densidad en $\mathbb{R}$ debe ser 1 por lo que

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{1}^{\infty} Kx^{-3}\,dx = K\left[-\frac{1}{2}x^{-2}\right]_1^{\infty} = K\left(\lim_{t\to\infty} -\frac{1}{2}t^{-2} + \frac{1}{2}\right) = \frac{K}{2} = 1 \longrightarrow K = 2.$$

La función de densidad de X es

$$f(x) = \begin{cases} 2x^{-3} & \text{si x>1,} \\ 0 & \text{en otro caso.} \end{cases}$$

ii) Generamos 1000 números aleatorios de la distribución de X utilizando el método de la función de distribución inversa. Por definición, la función de distribución de X es

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt = \int_{1}^{x} 2t^{-3}\,dt = 2\left[-\frac{1}{2}t^{-2}\right]_1^{x} = 1 - \frac{1}{x^2}.$$

Observamos que se cumple

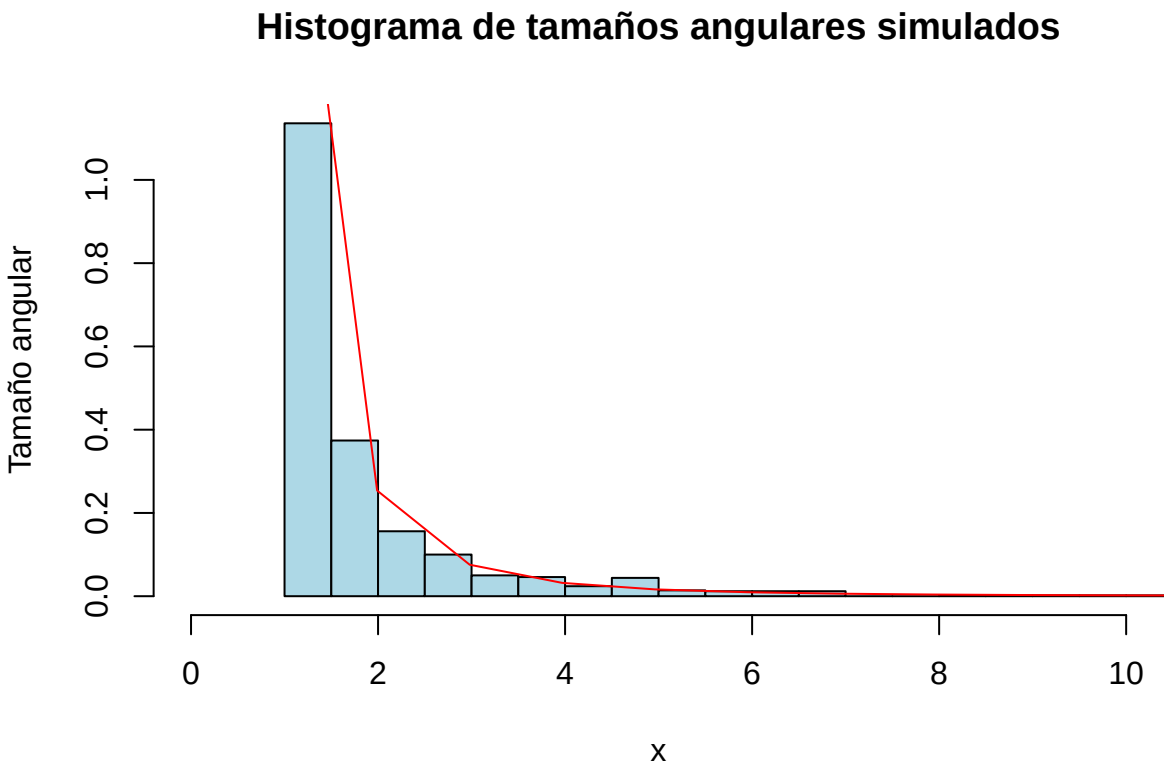
$$y = 1 - \frac{1}{x^2} \longrightarrow y = \frac{x^2-1}{x^2} \longrightarrow x^2 y = x^2 - 1 \longrightarrow 1 = x^2(1-y) \longrightarrow x^2 = \frac{1}{1-y} \longrightarrow x = \pm\frac{1}{\sqrt{1-y}},$$

con lo que inversa de la función de distribución es $F^{-1}(x) = \frac{1}{\sqrt{1-x}}$.

```r
Finv <- function(x){
  Finv <- 1/sqrt(1-x)
}
N <- 1000
u <- runif(N, 0, 1)
rand_num <- Finv(u)
```

iii) Representamos en un histograma los datos simulados de la distribución de X y los comparamos con la distribución teórica.

```r
hist(rand_num, breaks="Scott", main = "Histograma de tamaños angulares simulados",
     ylab = "Tamaño angular", xlab = "x", xlim = c(0,10), col = "lightblue", freq = F)
curve(2*x^(-3), 1, 100, col = "red", add = T)
```



Problema 4. *Lanzamos dardos a una diana y modelizamos la distancia a la diana como una variable aleatoria Z de coordenadas $Z = (Z_1, Z_2)$ siendo $Z_i \sim \mathcal{N}(0,1)$, $i = 1, 2$.*

i) Simula 1000 *vectores Z utilizando la función* **`rnorm`** *y represéntalos gráficamente.*

ii) ¿Qué distribución tendrán las coordenadas polares $Z = (r, \theta)$? Representa los histogramas de r y θ.

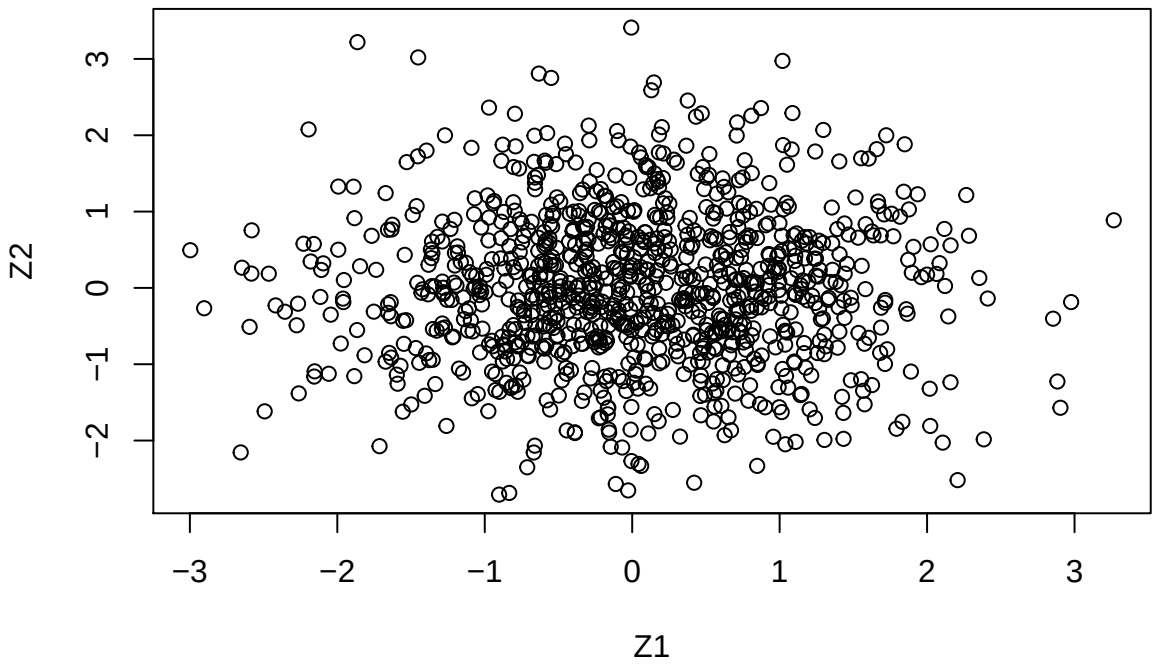
iii) Simula las coordenadas polares de Z a partir de números aleatorios uniformes.

iv) Transforma las coordenadas polares en Z_1 y Z_2 para obtener números aleatorios normales.

Solución.

i) Simulamos 1000 vectores Z con el comando `rnorm` y los representamos gráficamente.

```r
N <- 1000
Z1 <- rnorm(N, 0, 1)
Z2 <- rnorm(N, 0, 1)
plot(Z1, Z2, main = "Simulación de 1000 vectores con coordenadas normales N(0,1)")
```

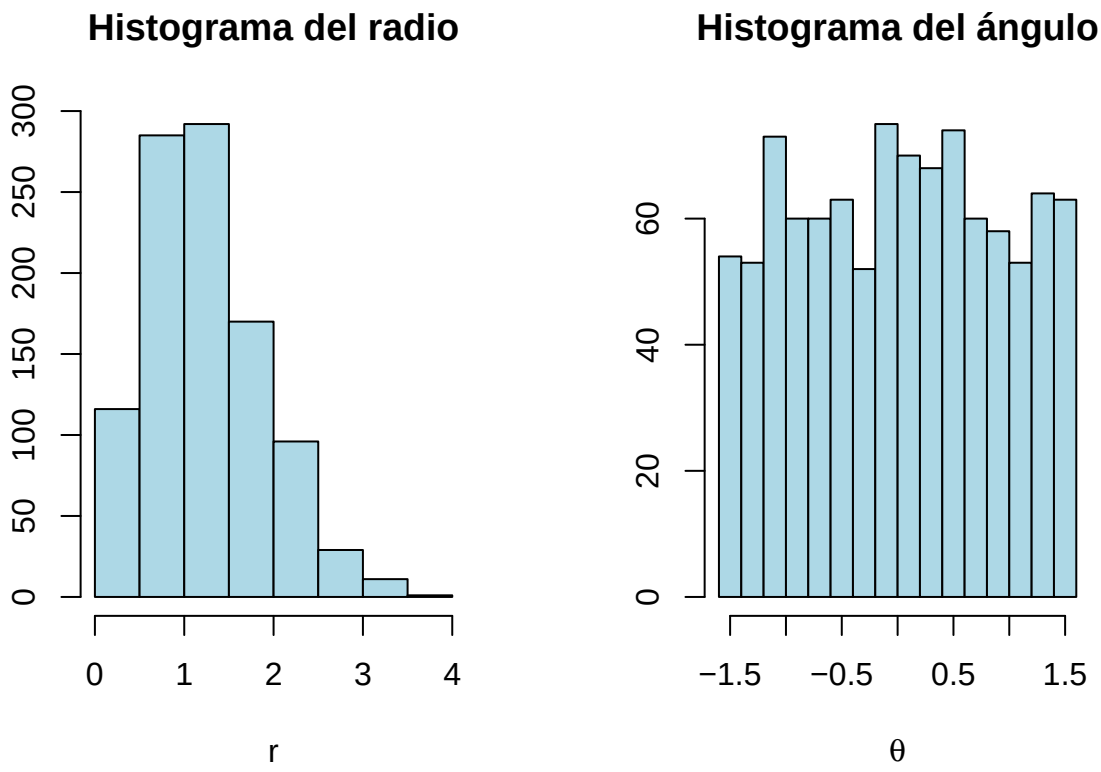


ii) Consideramos el cambio a coordenadas polares y tenemos

$$\begin{cases} Z_1 = \rho\cos(\theta) \\ Z_2 = \rho\sin(\theta) \end{cases} \longrightarrow \begin{cases} Z_1^2 + Z_2^2 = \rho^2 \\ \tan(\theta) = \dfrac{Z_1}{Z_2} \end{cases} \longrightarrow \begin{cases} \sqrt{Z_1^2 + Z_2^2} = \rho \\ \arctan\left(\dfrac{Z_1}{Z_2}\right) = \theta \end{cases}$$

Como ρ^2 es una suma de cuadrados de dos variables aleatorias independientes normales, se distribuye como una χ^2_2 y, entonces ρ sigue una distribución χ_2, la raíz cuadrada positiva de la χ^2_2. La coordenada θ sigue una distribución uniforme $U(-\pi, \pi)$. Representamos los histogramas de las simulaciones en una gráfica de dos paneles.

```
r <- sqrt(Z1^2 + Z2^2)
theta <- atan(Z2/Z1)
par(mfrow = c(1,2))
hist(r, main = "Histograma del radio", xlab = "r", ylab = NA, col = "lightblue")
hist(theta, breaks = 20, main = "Histograma del ángulo", xlab = expression(theta),
     ylab = NA, col ="lightblue")
```



iii) Simulamos 1000 coordenadas polares a partir de datos uniformes. La función de distribución de la χ^2_2 es $F(x) = 1 - e^{\frac{x}{2}}$. Observamos que se cumple

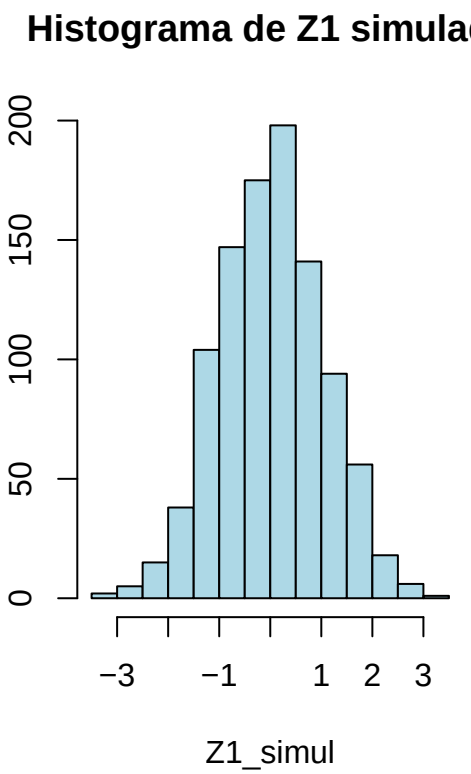
$$1 - e^{\frac{x}{2}} = y \longrightarrow e^{\frac{x}{2}} = 1 - y \longrightarrow \frac{x}{2} = \ln(1-y) \longrightarrow x = 2\ln(1-y),$$

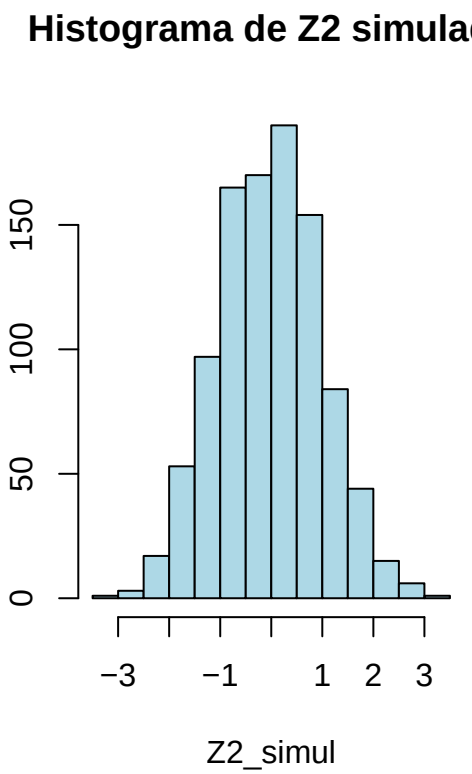
con lo que la inversa de la función de distribución de la χ^2_2 es $F(x) = 2\ln(1-x)$. Podemos simular ρ calculando simulaciones de ρ^2 con el método de la función de distribución inversa y tomando después la raíz cuadrada positiva. Las simulaciones de θ se obtienen directamente de la uniforme $U(-\pi, \pi)$.

```
u <- runif(N, 0, 1)
Finv3 <- function(x){
  Finv3 <- -2 * log(1 - x)
}
r_simul <- sqrt(Finv3(u))
theta_simul <- runif(N, -pi, pi)
```

iv) Pasamos de coordenadas polares a cartesianas para obtener coordenadas cartesianas con distribución normal.

```
Z1_simul <- r_simul * cos(theta_simul)
Z2_simul <- r_simul * sin(theta_simul)
par(mfrow = c(1,2))
hist(Z1_simul, main = "Histograma de Z1 simulado", xlab = "Z1_simul", ylab = NA,
     col = "lightblue")
hist(Z2_simul, main = "Histograma de Z2 simulado", xlab = "Z2_simul", ylab = NA,
     col = "lightblue")
```


Histograma de Z1 simulado
0
50
100
150
200
−3
−1
1
2
3
Z1_simul


Histograma de Z2 simulado
0
50
100
150
−3
−1
1
2
3
Z2_simul

Práctica 6

Problema 1. *Realiza las siguientes tareas:*

i) Simula 50 tiradas de un dado.

ii) Considera la variable aleatoria, X, definida como el número que sale al tirar el dado. ¿Cuánto valen $\bar{X}$, S^2, $\mathbb{E}(X)$, $\mathrm{Var}(X)$, $\mathbb{E}(\bar{X})$ y $\mathrm{Var}(\bar{X})$?

Solución.

i) Simulamos 50 tiradas de un dado con el comando `sample`.

```
x <- 1:6
N <- 50
dado <- sample(x, N, replace = TRUE)
dado
```

```
##  [1] 5 2 5 1 2 4 3 5 2 5 5 3 5 3 2 6 5 5 3 5 5 4 2 1 3 1 6 4 6 3 5 3 5 3 2 5 2 2
## [39] 5 2 5 1 3 1 3 6 3 2 4 2
```

ii) La media $\bar{X}$ y varianza S^2 muestrales de la simulación anterior son

```
cat("Media muestral: ", mean(dado), fill = T)
cat("Varianza muestral: ", var(dado), fill = T)
```

```
## Media muestral:  3.5
## Varianza muestral:  2.418367
```

Como la variable aleatoria X se distribuye como una uniforme discreta sobre $\{1, \ldots, 6\}$, la esperanza y la varianza de X son

$$\mathbb{E}(X) = \frac{1+6}{2} = \frac{7}{2} = 3.5, \qquad \mathrm{Var}(X) = \frac{6^2 - 1}{12} = \frac{35}{12} = 2.9\bar{6}$$

La esperanza y la varianza de la media $\bar{X}$ son

$$\mathbb{E}(\bar{X}) = \mathbb{E}(X) = \frac{7}{2} = 3.5, \qquad \mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{35}{12n} = \frac{2.9\bar{6}}{n}.$$

Problema 2. *Realiza las siguientes tareas:*

i) Simula 50 tiradas de dos dados.

ii) Considera la variable aleatoria, X, definida como la suma de los números que salen al tirar dos dados. ¿Cuánto valen $\bar{X}$, S^2, $\mathbb{E}(X)$, $\mathrm{Var}(X)$, $\mathbb{E}(\bar{X})$ y $\mathrm{Var}(\bar{X})$?

Solución.

i) Simulamos 50 tiradas de dos dados con el comando `sample`.

```
N <- 50
x <- 1:6
dado1 <- sample(x, N, replace = TRUE)
dado2 <- sample(x, N, replace = TRUE)
dado1
```

```
##  [1] 6 1 6 2 1 4 1 1 3 2 1 2 4 5 1 2 3 6 3 4 5 6 4 4 6 4 6 2 3 5 3 1 2 4 5 3 1 3
## [39] 3 4 3 1 3 6 4 2 5 3 1 6
```

```
dado2
```

```
##  [1] 1 6 2 1 4 3 6 3 6 2 6 2 5 4 2 5 4 1 5 3 1 3 4 2 4 1 1 4 6 3 1 3 6 1 2 6 6 4
## [39] 6 2 4 2 4 4 3 2 6 4 4 5
```

ii) Calculamos la simulación de la variable aleatoria X dada como la suma de los números que salen al tirar dos dados sumando las simulaciones del apartado anterior.

```
sumas <- dado1 + dado2
sumas
```

```
##  [1]  7  7  8  3  5  7  7  4  9  4  7  4  9  9  3  7  7  7  8  7  6  9  8  6 10
## [26]  5  7  6  9  8  4  4  8  5  7  9  7  7  9  6  7  3  7 10  7  4 11  7  5 11
```

La media $\bar{X}$ y la varianza S^2 muestrales de la simulación anterior son

```
cat("Media muestral: ", mean(sumas), fill = T)
cat("Varianza muestral: ", var(sumas), fill = T)
```

```
## Media muestral:  6.82
## Varianza muestral:  4.150612
```

Como la variable aleatoria es una suma $X = X_1 + X_2$ de variables aleatorias independientes igualmente distribuidas con distribución uniforme discreta sobre $\{1, \dots, 6\}$, tenemos que

$$\mathbb{E}(X) = \mathbb{E}(X_1 + X_2) = E(X_1) + E(X_2) = 2\mathbb{E}(X_1) = 2\frac{7}{2} = 7,$$

$$\mathrm{Var}(X) = \mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = 2\,\mathrm{Var}(X_1) = 2\frac{35}{12} = \frac{35}{6} = 5.8\bar{3}.$$

La esperanza y la varianza de la media $\bar{X}$ son

$$\mathbb{E}(\bar{X}) = \mathbb{E}(X) = 7, \qquad \mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{35}{6n} = \frac{5.8\bar{3}}{n}.$$

Problema 3. *Realiza las siguientes tareas:*

i) Simula alturas de 50 *personas de una población con altura media* 176 *y desviación típica* 11 *centímetros.*

ii) Considera la variable aleatoria, X, definida como la altura de una persona de la población descrita anteriormente. ¿Cuánto valen $\bar{X}$, S^2, $\mathbb{E}(X)$, $\mathrm{Var}(X)$, $\mathbb{E}(\bar{X})$ y $\mathrm{Var}(\bar{X})$?

Solución.

i) La altura de las personas es una variable aleatoria normal. En este caso, tiene media 176 y desviación típica 11. Simulamos las alturas de 50 personas con el comando `rnorm`.

```
alturas <- rnorm(50, 176, 11)
alturas
```

```
##  [1] 186.6968 175.9406 181.6727 187.4129 147.3708 183.7137 186.7524 162.3664
##  [9] 193.8084 178.1876 166.7683 178.6389 137.7075 156.4680 164.2225 165.2010
## [17] 182.2888 180.2856 167.6981 180.4712 169.9704 181.8998 181.7265 180.9935
## [25] 172.1833 163.6723 189.9824 177.1218 183.5196 173.4036 173.0444 193.5550
## [33] 191.6509 183.4096 165.0405 175.6594 151.3673 193.2071 170.6200 167.8179
## [41] 173.3521 180.7523 169.5860 182.6362 176.1168 169.7903 156.9256 161.7626
## [49] 160.7642 178.0882
```

ii) La media $\bar{X}$ y la varianza S^2 muestrales de la simulación anterior son

```
cat("Media muestral: ", mean(alturas), fill = T)
cat("Varianza muestral: ", var(alturas), fill = T)
```

```
## Media muestral:  174.2658
## Varianza muestral:  147.3224
```

Como la variable aleatoria X se distribuye como una normal $\mathcal{N}(176, 11)$, la esperanza y la varianza de X son

$$\mathbb{E}(X) = 176, \qquad \text{Var}(X) = 11^2 = 121.$$

La esperanza y la varianza de la media $\bar{X}$ son

$$\mathbb{E}(\bar{X}) = \mathbb{E}(X) = 176, \qquad \text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{11}{n}.$$

Problema 4. *Realiza las siguientes tareas:*

i) Simula el número de goles marcados en 10 *partidos siendo el promedio de goles por partido en la liga* 2.8*.*

ii) Considera la variable aleatoria, X, definida como el número de goles en un partido de la liga anteriormente mencionada. ¿Cuánto valen $\bar{X}$, S^2, $\mathbb{E}(X)$, $\text{Var}(X)$, $\mathbb{E}(\bar{X})$ y $\text{Var}(\bar{X})$?

Solución.

i) El número de goles marcados en un partido siendo la media de goles por partido 2.8 es una variable aleatoria con distribución de Poisson de media 2.8. Simulamos los goles marcados en 10 partidos con el comando `rpois`.

```
goles <- rpois(10, 2.8)
goles
```

```
##  [1] 1 0 4 2 1 1 2 5 3 3
```

ii) La media $\bar{X}$ y la varianza S^2 muestrales de la simulación anterior son

```
cat("Media muestral: ", mean(goles), fill = T)
cat("Varianza muestral: ", var(goles), fill = T)
```

```
## Media muestral:  2.2
## Varianza muestral:  2.4
```

Como la variable aleatoria X se distribuye como una Poisson de parámetro $\lambda = 2.8$, la esperanza y la varianza de X son

$$\mathbb{E}(X) = \lambda = 2.8, \qquad \text{Var}(X) = \lambda = 2.8.$$

La esperanza y la varianza de la media $\bar{X}$ son

$$\mathbb{E}(\bar{X}) = \mathbb{E}(X) = \lambda = 2.8, \qquad \text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{\lambda}{n}.$$

Problema 5. *Simula las posiciones de* 32 *fichas colocadas al azar en un tablero de* 8 *por* 8 *casillas teniendo en cuenta que solo puede haber una ficha por casilla.*

Solución.

Numeramos las 64 fichas del tablero de izquierda a derecha y de arriba a abajo. Podemos considerar una muestra ordenada $(x_1, \ldots, x_{32})$ extraída de manera uniforme de $\{1, \ldots, 64\}$ sin reemplazamiento. El elemento x_i indica la casilla en el tablero de la ficha i para todo $i \in \{1, \ldots, 32\}$.

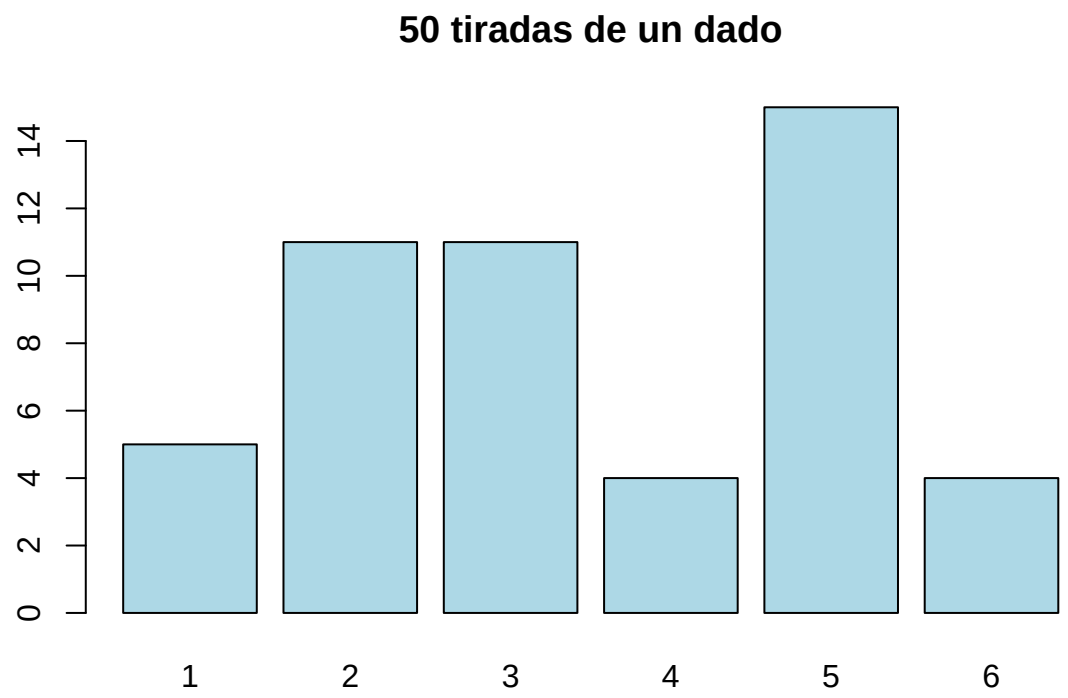
```
fichas <- sample(64, 32)  # Vector de casillas de cada ficha
tablero <- matrix(NA, 8, 8)  # Inicializamos un tablero vacío
for(aux in seq_along(fichas)){  # Recorremos las fichas rellenando el tablero
  tablero[fichas[aux]] <- aux
}
tablero
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]   NA   30   NA    7   NA   NA    3   NA
## [2,]   19   NA   26   NA   32   NA   NA   NA
## [3,]   22    8    2   NA   15    9   25   NA
## [4,]   NA   10    5   20    4   NA   NA   NA
## [5,]   NA   NA   NA   NA   28   23   17   NA
## [6,]   12   NA   NA   NA   NA   27   18   14
## [7,]   NA   21   16   NA   NA   NA   29   24
## [8,]   NA   13   31    6    1   NA   11   NA
```

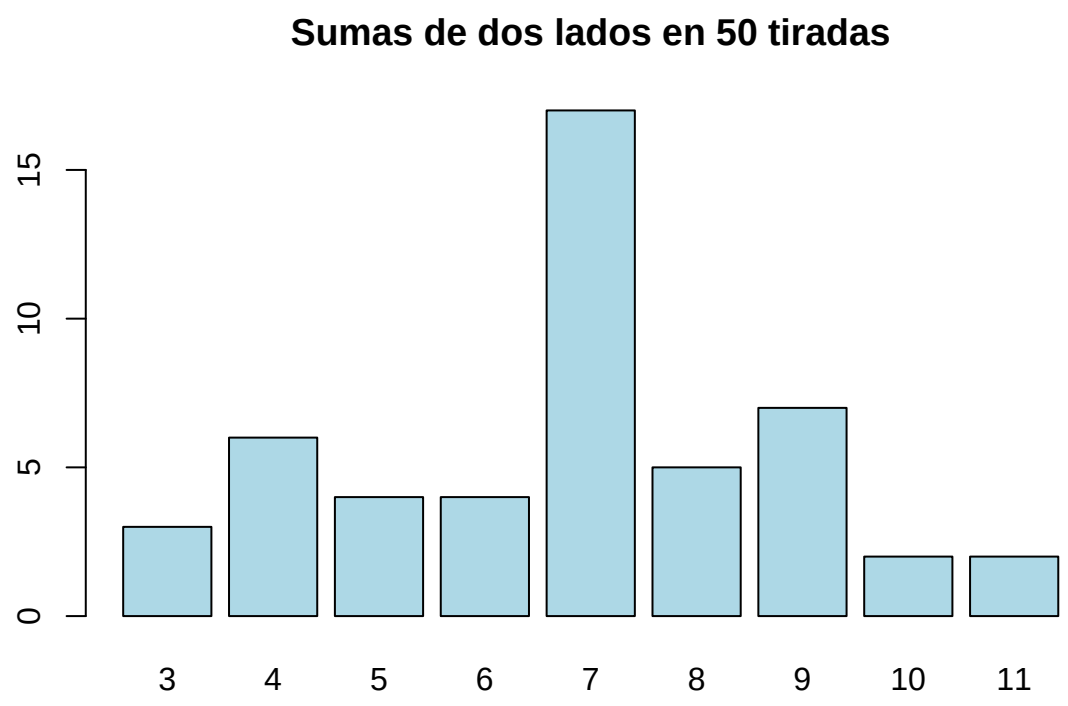
Problema 6. *Representa gráficamente cada una de las muestras simuladas y guarda las gráficas en un archivo.*

Solución.

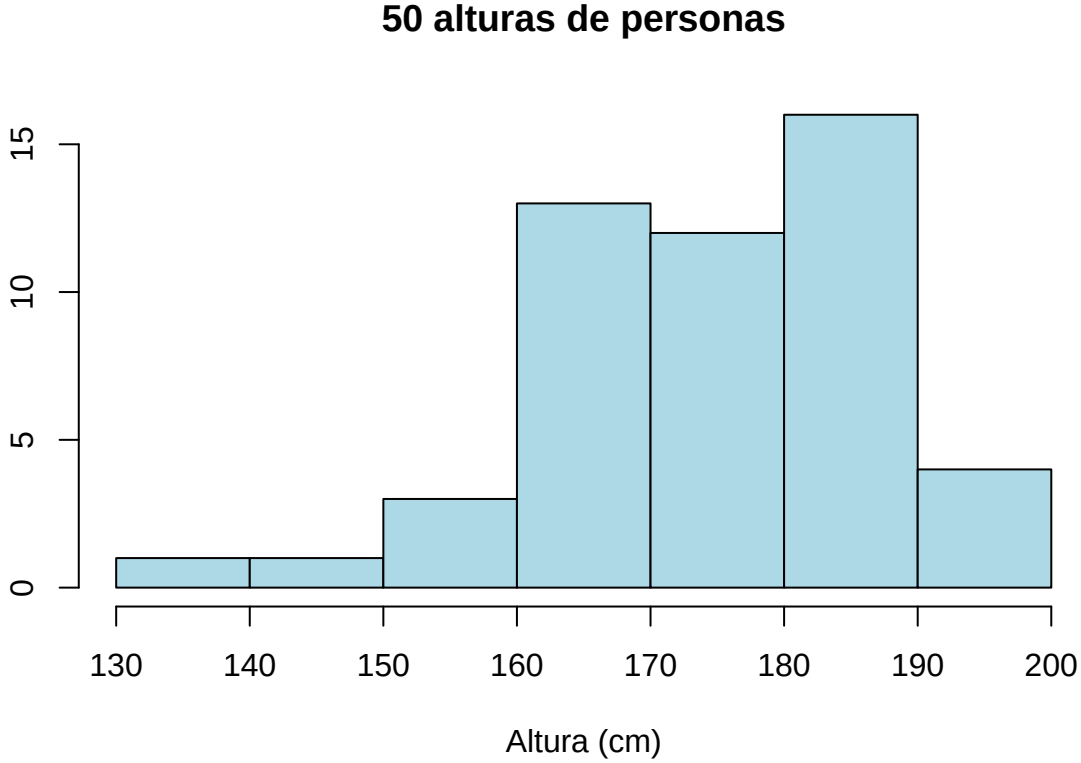
```
#pdf("figuras3.pdf")  # Guarda las gráficas en un archivo.
barplot(table(dado), col = "lightblue", main = "50 tiradas de un dado")
```



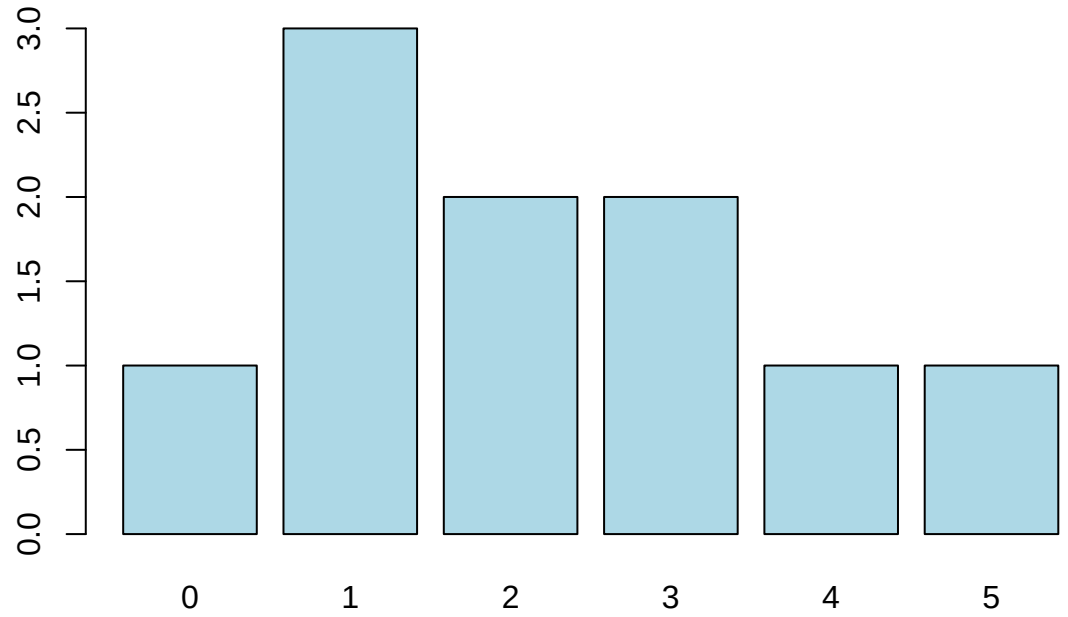

```
barplot(table(sumas), col = "lightblue", main = "Sumas de dos lados en 50 tiradas")
```



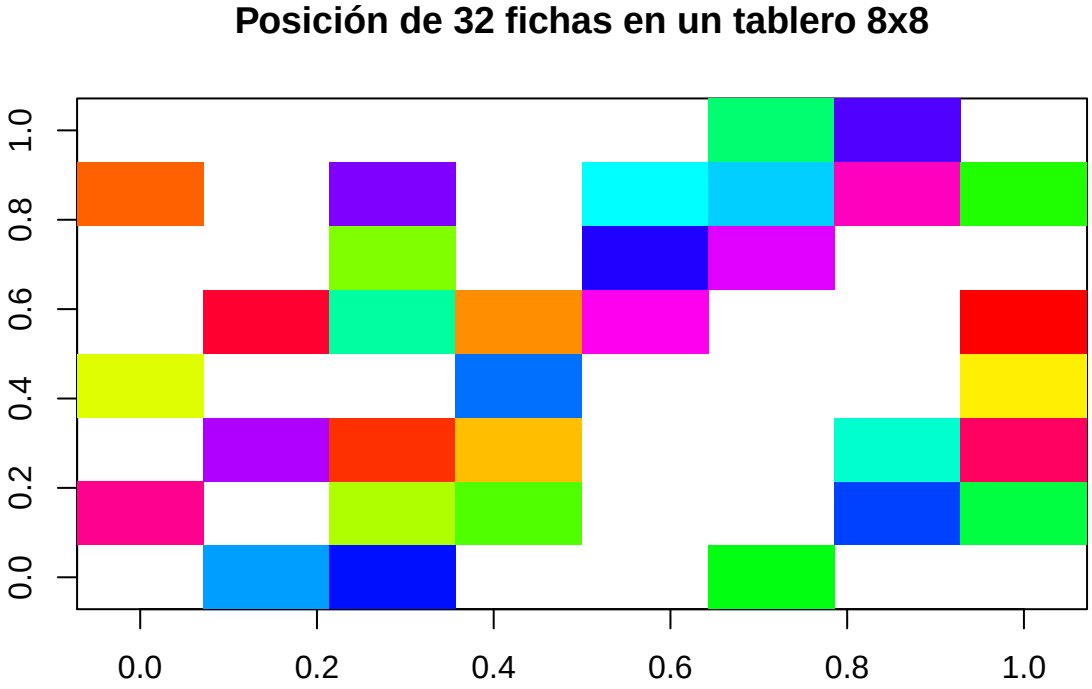

```
hist(alturas, col = "lightblue", xlab = "Altura (cm)",
     ylab = NA, main = "50 alturas de personas")
```




```
barplot(table(goles), col = "lightblue", main = "Goles en 10 partidos")
```




```
image(tablero, main = "Posición de 32 fichas en un tablero 8x8", col = rainbow(32))
```




```
#dev.off()
```

Práctica 7

Problema 1. *Considera la variable aleatoria* $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ *y la variable* $Y = a + bX$ *siendo* $a, b \in \mathbb{R}$.

i) ¿Cuánto valen $\mathbb{E}(Y)$ *y* $\mathrm{Var}(Y)$*? ¿Cuál será la distribución de* Y*?*

Solución.

i) Por las propiedades de la esperanza y la varianza

$$\mathbb{E}(Y) = \mathbb{E}(a + bX) = a + b\mathbb{E}(X) = a + b\mathbb{E}(\mathcal{N}(\mu_x, \sigma_x^2)) = a + b\mu_x$$

$$\mathrm{Var}(Y) = \mathrm{Var}(a + bX) = \mathrm{Var}(a) + b^2\,\mathrm{Var}(X) = 0 + b^2\,\mathrm{Var}(\mathcal{N}(\mu_x, \sigma_x^2)) = b^2\sigma_x^2.$$

Problema 2. *Considera los datos de extensión de hielo ártico utilizados en la práctica* 1.

i) Calcula $Y =$ *"Extensión media del mes de septiembre de cada año". Siendo* $X =$ *año.*

ii) Representa los datos en una gráfica.

iii) Calcula la covarianza, el coeficiente de determinación, la varianza de Y *y la varianza de* $Y|x$*.*

iv) ¿Conoces alguna fórmula que relacione las cantidades del apartado anterior?

v) ¿Qué es $\mathbb{E}(Y|x)$*?*

vi) Calcula la recta de regresión y dibújala en la gráfica.

vii) Calcula el intervalo de predicción al 95% *de confianza para el año 2031.*

viii) Simula 10 *realizaciones distintas de la evolución hasta 2031 y represéntalas gráficamente.*

Solución.

i) Cargamos primero los datos de la extensión de hielo en el Ártico.

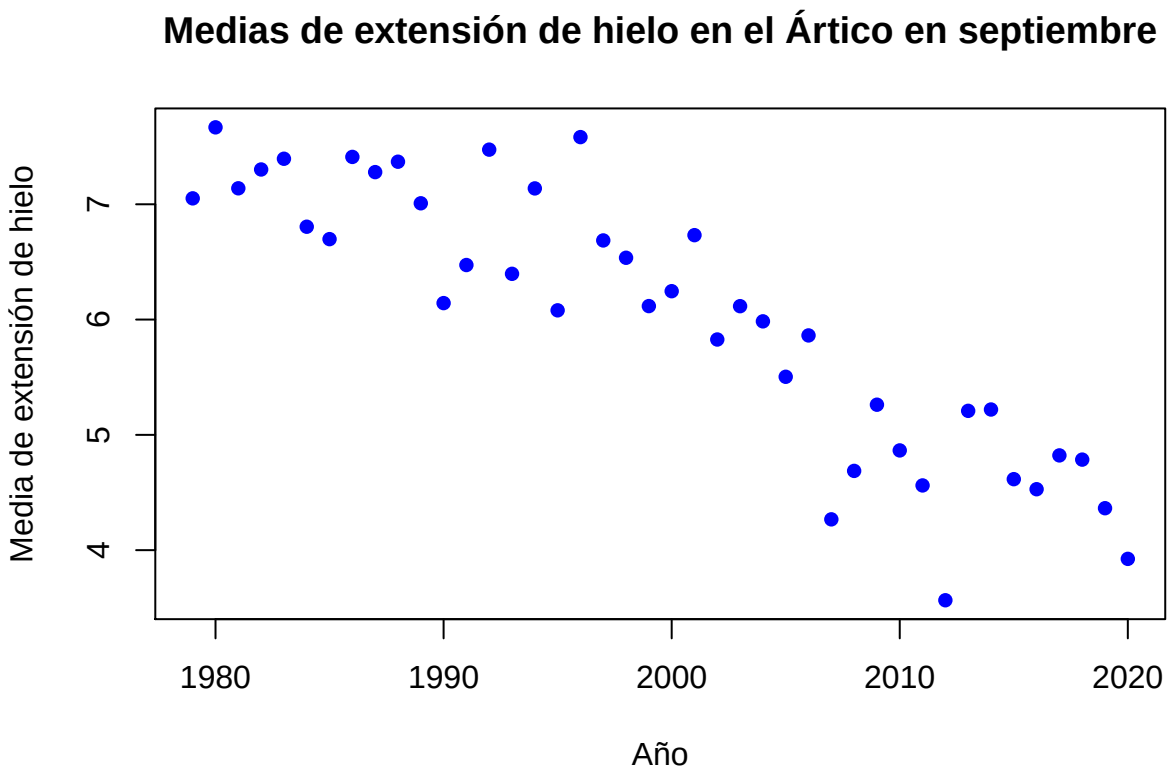
```
all_content <- readLines("N_seaice_extent_daily_v3.0.csv")
skip_second <- all_content[-2]
datos <- read.csv(textConnection(skip_second))
```

Calculamos las extensiones medias de hielo ártico en el mes de septiembre de cada año y representamos los datos.

```
ext_sept <- c()
medias <- c()
for(i in unique(datos$Year)){
  ext_sept <- subset(datos$Extent, datos$Year == i & datos$Month == 9)
  medias <- append(medias, mean(ext_sept))
}
medias <- medias[!is.na(medias)]
año = 1979:2020
media_sept <- data.frame(año, medias)
```

ii) Representamos los datos en una gráfica de puntos.

```
plot(media_sept, main = "Medias de extensión de hielo en el Ártico en septiembre",
     ylab = "Media de extensión de hielo", xlab = "Año", col = "blue", pch = 16)
```



iii) Calculamos la covarianza, el coeficiente de determinación, la varianza de Y y la varianza de $Y|x$.

```
modelo <- lm(medias~año)
S_X <- sd(año)
S_Y <- sd(medias)
S_XY <- cov(año, medias)
S_Yx <- sd(modelo$residuals)
r <- cor(año, medias)
cat("Covarianza: ", S_XY, fill = T)
cat("Coeficiente de determinación: ", r^2, fill = T)
cat("Varianza de Y: ", S_Y^2, fill = T)
cat("Varianza de Y|x: ", S_Yx^2)
```

```
## Covarianza:  -12.5961
## Coeficiente de determinación:  0.7969776
## Varianza de Y:  1.322786
## Varianza de Y|x:  0.2685552
```

iv) Una fórmula que relaciona las cantidades del apartado anterior es

$$S^2_{Y|x} = S^2_Y(1-r^2), \qquad r^2 = \left(\frac{S_{XY}}{S_X S_Y}\right)^2,$$

donde $S^2_{Y|x}$ denota la varianza de $Y|x$, S^2_Y la varianza de X, S^2_X la varianza de X, S_{XY} la covarianza de X e Y y r^2 el coeficiente de determinación. Lo comprobamos calculando ambos lados de la ecuación.

```
cat("Varianza de Y|x: ", S_Yx, fill = T)
cat("Varianza de Y multiplicado por (1-r^2): ", S_X^2*(1-r^2), fill = T)
cat("Coeficiente de determinación: ", r^2, fill = T)
cat("(S_XY/(S_XS_Y))^2: ", (S_XY/(S_X*S_Y))^2)
```

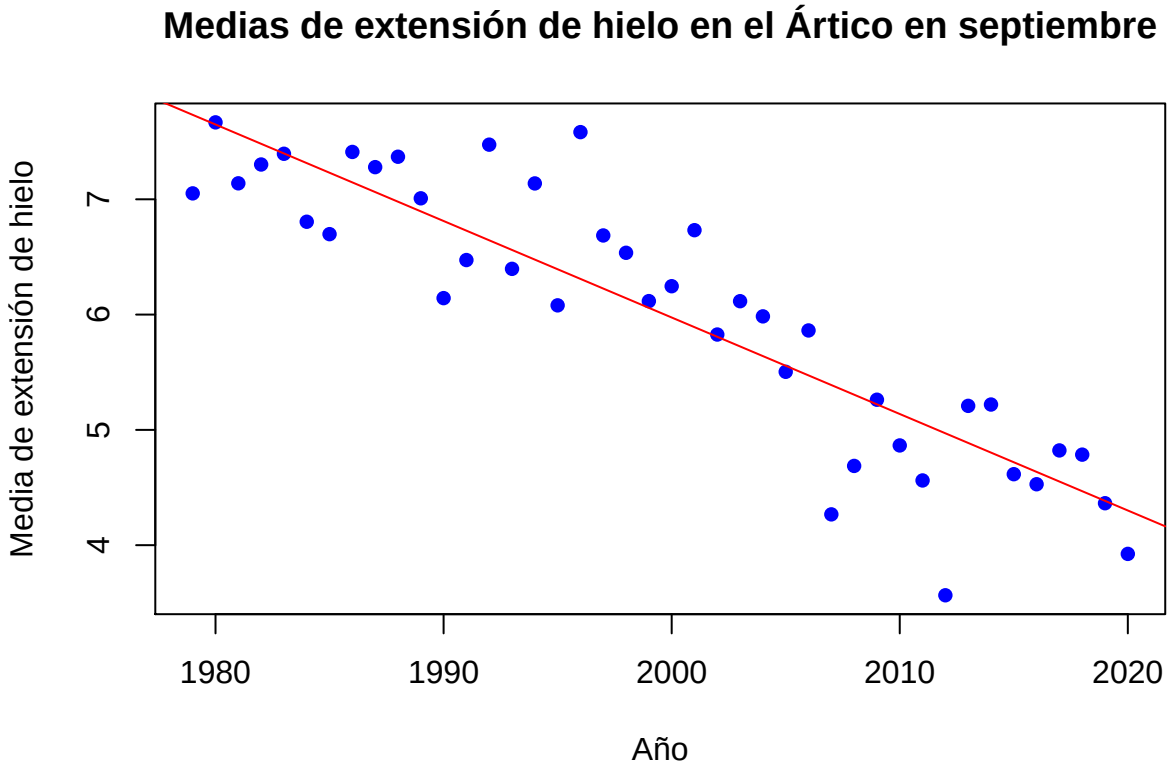
```
## Varianza de Y|x:  0.5182231
```

```
## Varianza de Y multiplicado por (1-r^2):  30.55488
## Coeficiente de determinación:  0.7969776
## (S_XY/(S_XS_Y))^2:  0.7969776
```

v) La esperanza $\mathbb{E}(Y|x)$ es el valor esperado de Y dado un valor x, es la recta de regresión.

vi) Representamos la recta de regresión con el comando `abline`.

```
plot(media_sept, main = "Medias de extensión de hielo en el Ártico en septiembre",
     ylab = "Media de extensión de hielo", xlab = "Año", col = "blue", pch = 16)
abline(modelo, col = "red")
```



vii) Calculamos un intervalo de confianza al nivel 95% y de predicción para el año 2031.

```
newdata <- data.frame(año = 2031)
confint(modelo)
```

```
##                  2.5 %       97.5 %
## (Intercept) 146.373052 200.35647843
## año          -0.097194  -0.07019603
```

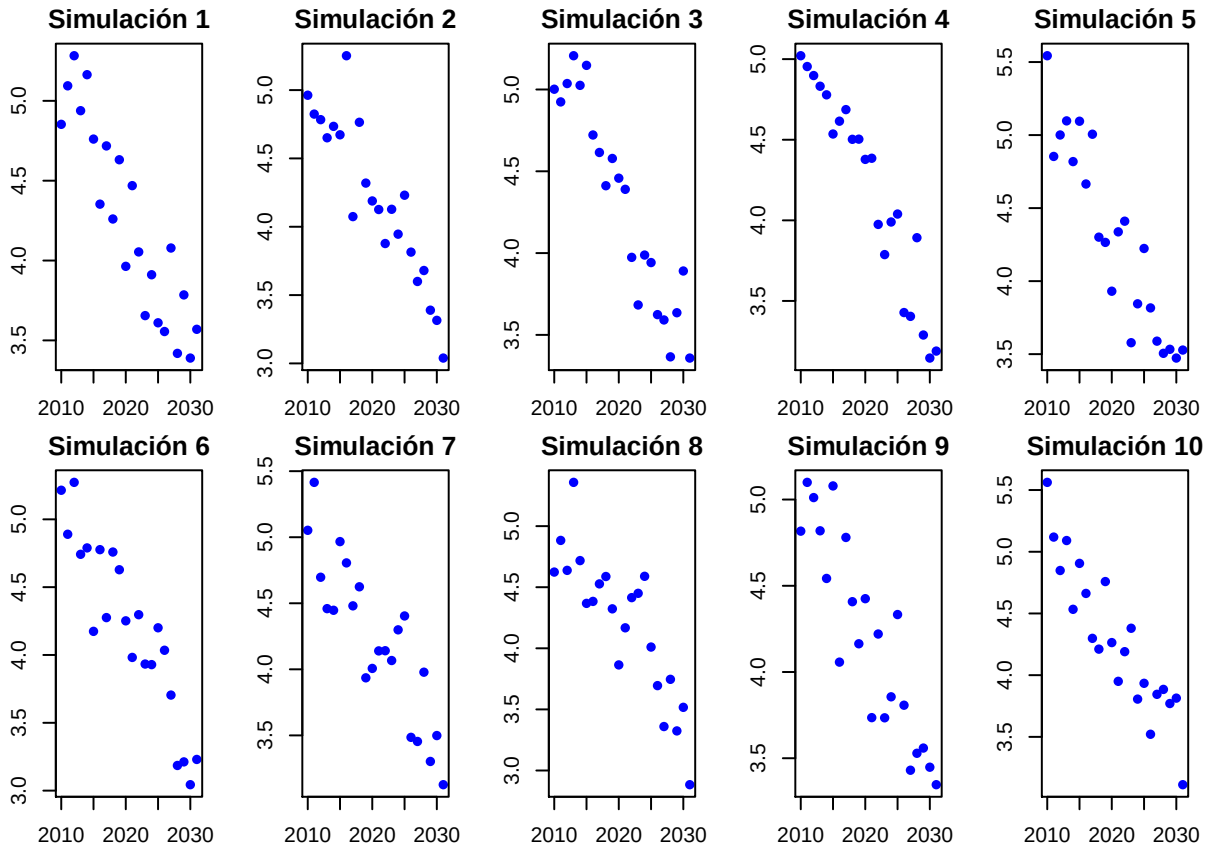
```
predict(modelo, newdata, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 3.380192 2.226075 4.534309
```

viii) Realizamos 10 simulaciones de la evolución hasta 2031. Para ello, dado un año $2010 \leq x \leq 2031$, simulamos una normal de media la imagen de x en la recta de regresión y varianza el valor del estimador insesgado para la varianza de los residuos dado por

$$s_2 = \frac{\sum_{i=1}^{n} Y|x_i}{n-2}.$$

```r
s_2 <- sum(residuals(modelo)^2)/(42-2)
x <- 2010:2031
media_recta <- 173.36476523-0.08369501*x
par(mfrow = c(2, 5), mar = rep(2, 4))
for(aux in 1:10){
  y <- rnorm(22, media_recta, s_2)
  plot(x, y, col = "blue", pch = 16,
       main = paste0("Simulación ", aux))
}
```



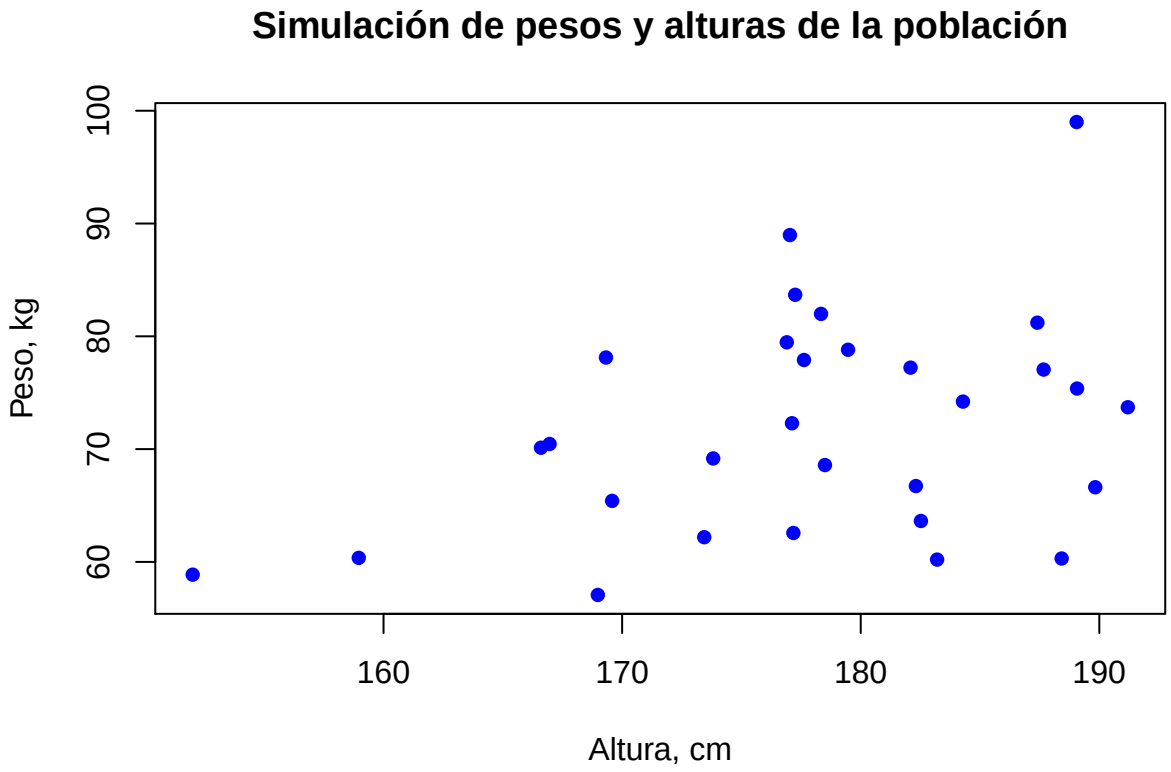
Problema 3. *Considera una población con pesos* $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ *y altura* $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ *con un coeficiente de correlación poblacional* $\rho = 0.3$. *Toma los valores* $\mu_x = 175$ *cm*, $\sigma_x = 10$ *cm*, $\mu_y = 70$ *kg*, $\sigma_y = 8$ *kg*.

i) Simula los pesos y alturas de 30 *personas de esta población.*

Solución.

i) Simulamos los pesos y alturas de 30 personas de esta población como una variable aleatoria bidimensional.

```r
muX <- 175; sigmaX <- 10; muY <- 70; sigmaY <- 8; rho <- 0.3
sigmaXY <- (sigmaX*sigmaY)*rho
x <- rnorm(30, muX, sigmaX)
y <- rnorm(30, muY+sigmaXY/sigmaX^2*(x-muX), sqrt(sigmaY^2*(1-rho^2)))
plot(x, y, main = "Simulación de pesos y alturas de la población",
     xlab="Altura, cm", ylab="Peso, kg", pch = 16, col = "blue")
```


Simulación de pesos y alturas de la población
Peso, kg
60
70
80
90
100
160
170
180
190
Altura, cm

Práctica 8

Problema 1. *En una ciudad concreta hay dos compañías de taxi: verdes y azules. Un* 85% *de los taxis son verdes y un* 15% *azules. En un accidente nocturno un taxi se fuga. Un testigo asegura que el taxi fugado era azul. Se sabe gracias a unas pruebas independientes que ese testigo es capaz de identificar correctamente el color de un taxi el* 80% *de las veces en las mismas condiciones que la noche del accidente. Calcula la probabilidad de que el taxi fugado fuese realmente azul.*

Solución.

Consideramos los siguientes sucesos

TaxiVerde: El taxi es verde.	DiceVerde: El testigo dice que el taxi es verde.
TaxiAzul: El taxi es azul.	DiceAzul: El testigo dice que el taxi es azul.

Sabemos que se cumple

$$\begin{aligned} P(\text{TaxiVerde}) &= 0.85, & P(\text{DiceVerde}|\text{TaxiVerde}) &= 0.8, \\ P(\text{TaxiAzul}) &= 0.15, & P(\text{DiceAzul}|\text{TaxiAzul}) &= 0.8. \end{aligned}$$

Además, tenemos que $\{\text{TaxiAzul}, \text{TaxiVerde}\}$ es una partición del espacio muestral de los taxis de la ciudad. Por el teorema de la probabilidad total,

$$\begin{aligned} P(\text{DiceAzul}) &= P(\text{DiceAzul}|\text{TaxiAzul})P(\text{TaxiAzul}) + P(\text{DiceAzul}|\text{TaxiVerde})P(\text{TaxiVerde}) \\ &= 0.8 \cdot 0.15 + (1 - 0.8) \cdot 0.85 \\ &= 0.29. \end{aligned}$$

La probabilidad de que el taxi fuese realmente azul es la probabilidad de que sea azul condicionado a que el testigo ha dicho que es azul. Aplicando el teorema de Bayes,

$$P(\text{TaxiAzul}|\text{DiceAzul}) = \frac{P(\text{DiceAzul}|\text{TaxiAzul})P(\text{TaxiAzul})}{P(\text{DiceAzul})} = \frac{0.8 \cdot 0.15}{0.29} \simeq 0.4138.$$

Problema 2. *Considera una moneda sesgada en la que la probabilidad de que salga cara es* θ*. Disponemos de dos modelos para* θ*.*

- M_0*: Solo puede haber tres valores posibles* $\theta = \{0.25, 0.5, 075\}$ *con probabilidades* $p(0.25) = p(0.75) = p(0.5) = 0.5$*.*
- M_1*: Hay* 63 *valores posibles de* θ*, listados junto a sus probabilidades en el archivo* Theta_pTheta.dat*.*

Una persona lanza una moneda 12 *veces obteniendo* 3 *caras y* 9 *cruces.*

i) Realiza la comparación de modelos bayesiana.

ii) Calcula el Likelihood *y* Posterior *que obtendrías para cada modelo. Represéntalos gráficamente junto con el* Prior *correspondiente.*

iii) Repite el ejercicio cambiando los datos por 1 *cara y* 11 *cruces.*

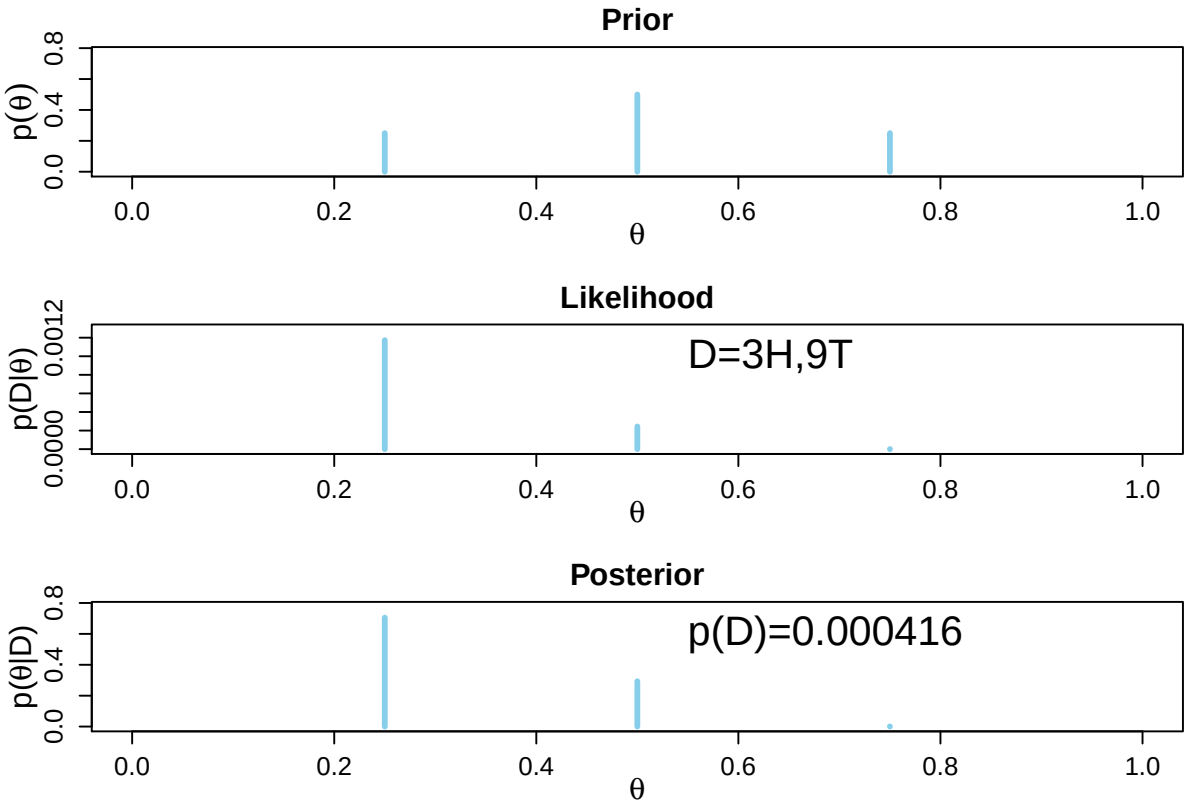
iv) Comenta los resultados

Solución.

i)+ii) Comparación de modelos bayesiana.

```
# Modelo simple, M0
##################
# theta = Probabilidad de que salga cara.
Theta <- c(0.25, 0.5, 0.75)  # Vector de valores posibles para theta.
nThetaVals <- length(Theta)  # Número de valores posibles para theta.
```

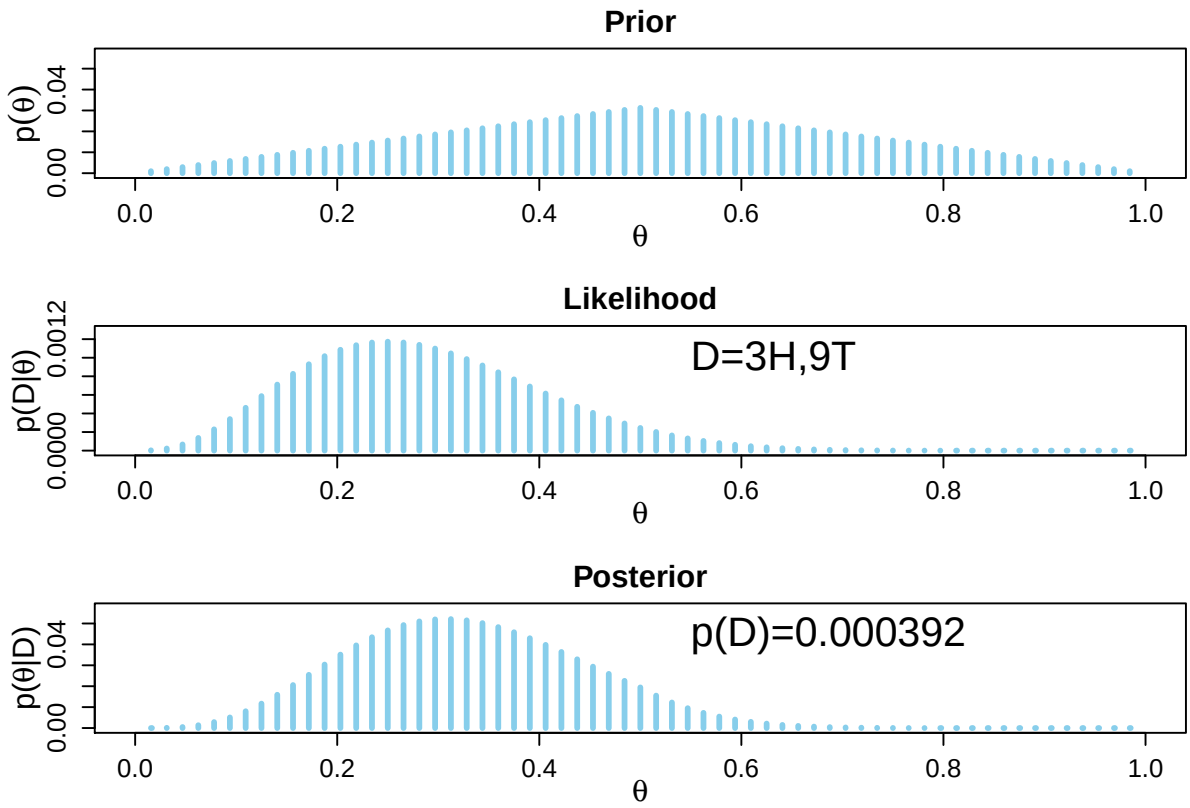
```r
#PRIOR
pTheta <- c(0.25, 0.5, 0.25)  # Vector de probabilidades a priori para los valores.
#DATA
# Cara = 1, Cruz = 0
Data <- c(rep(1, 3), rep(0, 9))
nHeads <- sum(Data == 1)
nTails <- sum(Data == 0)
#LIKELIHOOD
pDataGivenTheta <- Theta^nHeads * (1-Theta)^nTails
#EVIDENCE
pData <- sum(pDataGivenTheta * pTheta)
#POSTERIOR
pThetaGivenData = pDataGivenTheta * pTheta / pData
# Graficamos los resultados.
layout(matrix(c(1, 2, 3), nrow = 3 , ncol = 1, byrow = FALSE)) # 3x1 paneles
par(mar = c(3, 3, 1, 0))          # Márgenes
par(mgp = c(2, 1, 0))
par(mai = c(0.5, 0.5, 0.3, 0.1))
# Graficamos el Prior:
plot(Theta, pTheta, type = "h", lwd = 3, main = "Prior",
      xlim = c(0, 1), xlab=bquote(theta),
      ylim = c(0, 1.1*max(pThetaGivenData)), ylab = bquote(p(theta)),
      cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.5, col = "skyblue")
# Graficamos el Likelihood:
plot(Theta, pDataGivenTheta, type = "h", lwd = 3, main = "Likelihood",
      xlim = c(0, 1), xlab = bquote(theta),
      ylim = c(0, 1.1*max(pDataGivenTheta)),
      ylab = bquote(paste("p(D|", theta, ")")),
      cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.5, col = "skyblue")
text(.55, .85*max(pDataGivenTheta), cex = 2.0,
      bquote("D=" * .(nHeads) * "H," * .(nTails) * "T"), adj = c(0, .5))
# Graficamos el Posterior:
plot(Theta, pThetaGivenData, type = "h", lwd = 3, main = "Posterior",
      xlim = c(0, 1), xlab = bquote(theta),
      ylim = c(0, 1.1*max(pThetaGivenData)),
      ylab = bquote(paste("p(", theta, "|D)")),
      cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.5, col = "skyblue")
text(.55, .85*max(pThetaGivenData), cex = 2.0,
      bquote("p(D)=" * .(signif(pData, 3))), adj = c(0, .5))
```




```r
# Guardamos el Evidence para comparar
pData_0 <- pData

# Modelo complejo, M1
###################
# theta = probability of the coin coming up heads.
Theta_pTheta <- read.table("Theta_pTheta.txt", header = TRUE)
Theta <- Theta_pTheta$Theta  # Vector de valores posibles para theta.
nThetaVals <- length(Theta)  # Número de valores posibles para theta.
# PRIOR
pTheta <- Theta_pTheta$pTheta  # Vector de probabilidades a priori para los valores.
#DATA
# Head = 1, Tail = 0
Data <- c(rep(1, 3), rep(0, 9))
# Data <- c(rep(1, 1), rep(0, 11)).
nHeads <- sum(Data == 1)
nTails <- sum(Data == 0)
#LIKELIHOOD
pDataGivenTheta <- Theta^nHeads * (1-Theta)^nTails
#EVIDENCE
pData <- sum(pDataGivenTheta * pTheta)
#POSTERIOR
pThetaGivenData = pDataGivenTheta * pTheta / pData
# Graficamos los resultados.
layout(matrix(c(1, 2, 3), nrow = 3, ncol = 1, byrow = FALSE)) # 3x1 paneles
par(mar = c(3, 3, 1, 0))          # Márgenes
par(mgp = c(2, 1, 0))
par(mai = c(0.5, 0.5, 0.3, 0.1))
```

```r
# Graficamos el Prior:
plot(Theta, pTheta, type = "h", lwd = 3, main = "Prior",
      xlim = c(0, 1), xlab = bquote(theta),
      ylim = c(0, 1.1*max(pThetaGivenData)), ylab = bquote(p(theta)),
      cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.5, col = "skyblue")
# Graficamos el Likelihood:
plot(Theta, pDataGivenTheta, type = "h", lwd = 3, main = "Likelihood",
      xlim = c(0, 1), xlab = bquote(theta),
      ylim = c(0, 1.1*max(pDataGivenTheta)),
      ylab = bquote(paste("p(D|", theta, ")")),
      cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.5, col = "skyblue")
text(.55, .85*max(pDataGivenTheta), cex = 2.0,
      bquote("D=" * .(nHeads) * "H," * .(nTails) * "T"), adj = c(0, .5))
# Graficamos el Posterior:
plot(Theta, pThetaGivenData, type = "h", lwd = 3, main = "Posterior",
      xlim = c(0, 1), xlab = bquote(theta),
      ylim = c(0, 1.1*max(pThetaGivenData)),
      ylab = bquote(paste("p(", theta, "|D)")),
      cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.5, col = "skyblue")
text(.55, .85*max(pThetaGivenData), cex = 2.0,
      bquote("p(D)=" * .(signif(pData, 3))), adj = c(0, .5))
```

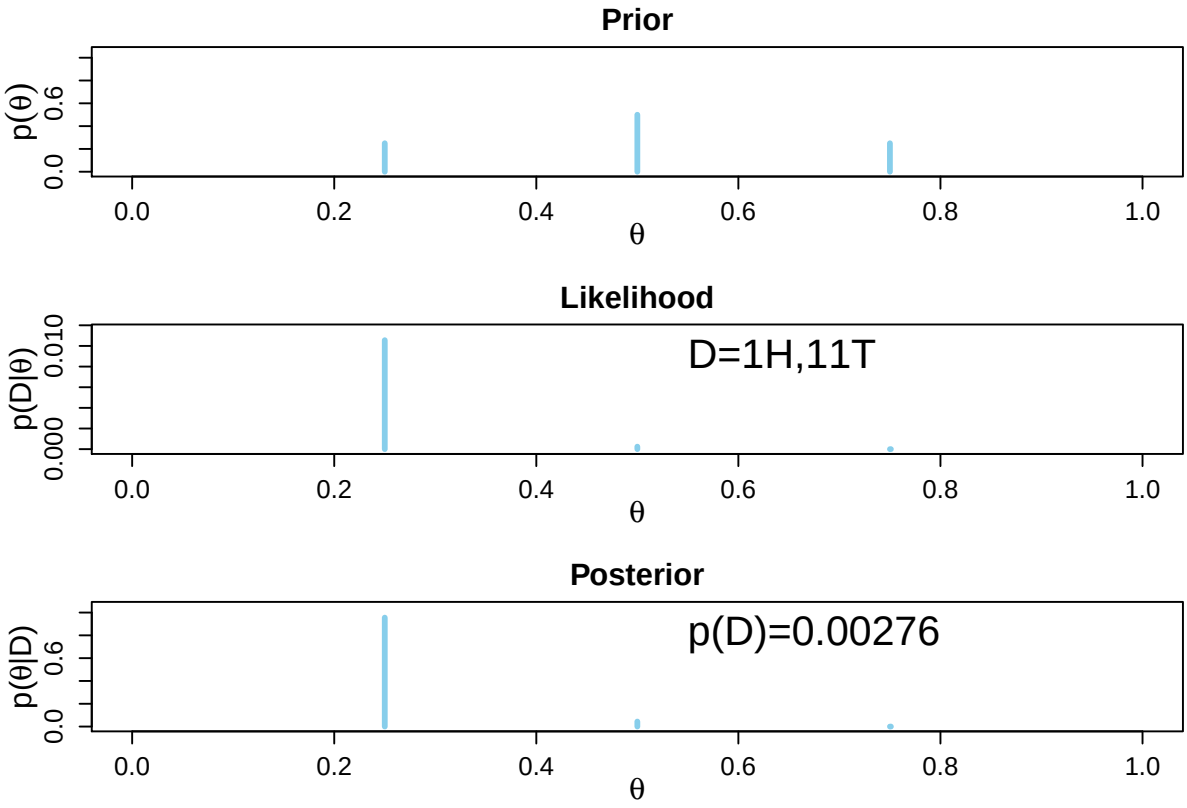



```r
# Guardamos el Evidence para comparar
pData_1 <- pData
# Comparación bayesiana de los modelos
pData_1/pData_0
```

```
## [1] 0.943836
```

iii) Repetimos lo anterior cambiando los datos a 1 cara y 11 cruces.

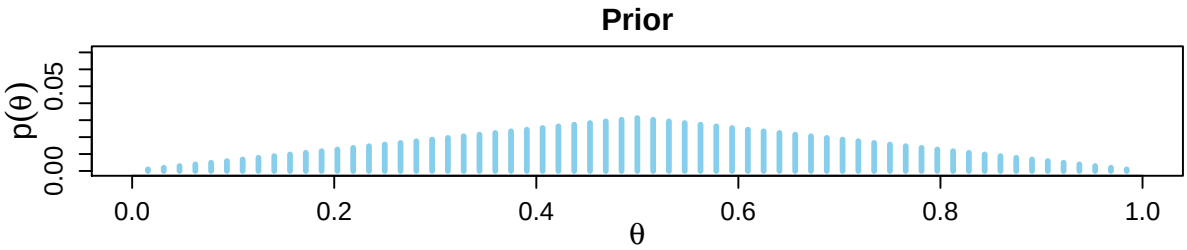
```
# Modelo simple, M0
#################
# theta = Probabilidad de que salga cara.
Theta <- c(0.25, 0.5, 0.75)  # Vector de valores posibles para theta.
nThetaVals <- length(Theta)  # Número de valores posibles para theta.
#PRIOR
pTheta <- c(0.25, 0.5, 0.25)  # Vector de probabilidades a priori para los valores.
#DATA
# Cara = 1, Cruz = 0
Data <- c(rep(1, 1), rep(0, 11))
nHeads <- sum(Data == 1)
nTails <- sum(Data == 0)
#LIKELIHOOD
pDataGivenTheta <- Theta^nHeads * (1-Theta)^nTails
#EVIDENCE
pData <- sum(pDataGivenTheta * pTheta)
#POSTERIOR
pThetaGivenData = pDataGivenTheta * pTheta / pData
# Graficamos los resultados.
layout(matrix(c(1, 2, 3), nrow = 3 , ncol = 1, byrow = FALSE)) # 3x1 paneles
par(mar = c(3, 3, 1, 0))          # Márgenes
par(mgp = c(2, 1, 0))
par(mai = c(0.5, 0.5, 0.3, 0.1))
# Graficamos el Prior:
plot(Theta, pTheta, type = "h", lwd = 3, main = "Prior",
      xlim = c(0, 1), xlab=bquote(theta),
      ylim = c(0, 1.1*max(pThetaGivenData)), ylab = bquote(p(theta)),
      cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.5, col = "skyblue")
# Graficamos el Likelihood:
plot(Theta, pDataGivenTheta, type = "h", lwd = 3, main = "Likelihood",
      xlim = c(0, 1), xlab = bquote(theta),
      ylim = c(0, 1.1*max(pDataGivenTheta)),
     ylab = bquote(paste("p(D|", theta, ")")),
      cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.5, col = "skyblue")
text(.55, .85*max(pDataGivenTheta), cex = 2.0,
      bquote("D=" * .(nHeads) * "H," * .(nTails) * "T"), adj = c(0, .5))
# Graficamos el Posterior:
plot(Theta, pThetaGivenData, type = "h", lwd = 3, main = "Posterior",
      xlim = c(0, 1), xlab = bquote(theta),
      ylim = c(0, 1.1*max(pThetaGivenData)),
     ylab = bquote(paste("p(", theta, "|D)")),
      cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.5, col = "skyblue")
text(.55, .85*max(pThetaGivenData), cex = 2.0,
      bquote("p(D)=" * .(signif(pData, 3))), adj = c(0, .5))
```

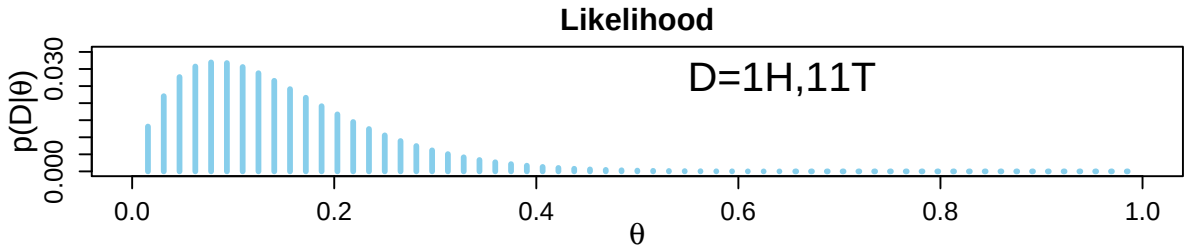


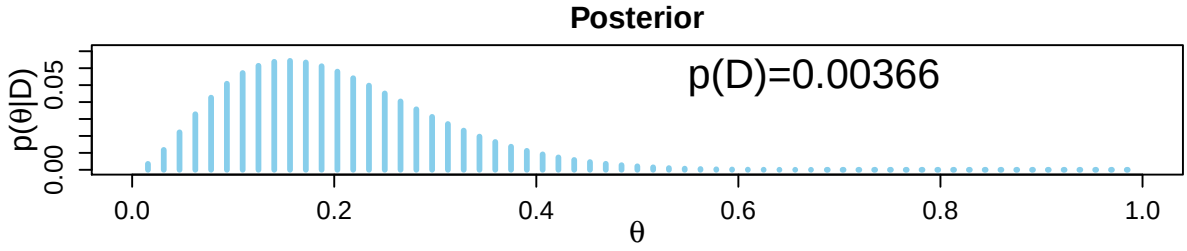

```r
# Guardamos el Evidence para comparar
pData_0 <- pData

# Modelo complejo, M1
###################
# theta = probability of the coin coming up heads.
Theta_pTheta <- read.table("Theta_pTheta.txt", header = TRUE)
Theta <- Theta_pTheta$Theta  # Vector de valores posibles para theta.
nThetaVals <- length(Theta)  # Número de valores posibles para theta.
# PRIOR
pTheta <- Theta_pTheta$pTheta  # Vector de probabilidades a priori para los valores.
#DATA
# Head = 1, Tail = 0
Data <- c(rep(1, 1), rep(0, 11))
nHeads <- sum(Data == 1)
nTails <- sum(Data == 0)
#LIKELIHOOD
pDataGivenTheta <- Theta^nHeads * (1-Theta)^nTails
#EVIDENCE
pData <- sum(pDataGivenTheta * pTheta)
#POSTERIOR
pThetaGivenData = pDataGivenTheta * pTheta / pData
# Graficamos los resultados.
layout(matrix(c(1, 2, 3), nrow = 3, ncol = 1, byrow = FALSE)) # 3x1 paneles
par(mar = c(3, 3, 1, 0))          # Márgenes
par(mgp = c(2, 1, 0))
par(mai = c(0.5, 0.5, 0.3, 0.1))
# Graficamos el Prior:
```

```r
plot(Theta, pTheta, type = "h", lwd = 3, main = "Prior",
      xlim = c(0, 1), xlab = bquote(theta),
      ylim = c(0, 1.1*max(pThetaGivenData)), ylab = bquote(p(theta)),
      cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.5, col = "skyblue")
# Graficamos el Likelihood:
plot(Theta, pDataGivenTheta, type = "h", lwd = 3, main = "Likelihood",
      xlim = c(0, 1), xlab = bquote(theta),
      ylim = c(0, 1.1*max(pDataGivenTheta)),
      ylab = bquote(paste("p(D|", theta, ")")),
      cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.5, col = "skyblue")
text(.55, .85*max(pDataGivenTheta), cex = 2.0,
      bquote("D=" * .(nHeads) * "H," * .(nTails) * "T"), adj = c(0, .5))
# Graficamos el Posterior:
plot(Theta, pThetaGivenData, type = "h", lwd = 3, main = "Posterior",
      xlim = c(0, 1), xlab = bquote(theta),
      ylim = c(0, 1.1*max(pThetaGivenData)),
      ylab = bquote(paste("p(", theta, "|D)")),
      cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.5, col = "skyblue")
text(.55, .85*max(pThetaGivenData), cex = 2.0,
      bquote("p(D)=" * .(signif(pData, 3))), adj = c(0, .5))
```








```r
# Guardamos el Evidence para comparar
pData_1 <- pData
# Comparación bayesiana de los modelos
pData_1/pData_0
```

```
## [1] 1.325016
```

iv) La comparación de modelos bayesiana está dada por

$$\frac{P(M_1|D)}{P(M_0|D)} = \underbrace{\frac{P(D|M_1)}{P(D|M_0)}}_{\text{Factor de Bayes}} \frac{P(M_1)}{P(M_0)}.$$

Observamos que el factor de Bayes en el primer caso, con 3 caras y 9 cruces, es menor que 1, con lo que el mejor modelo según la comparación bayesiana es el modelo simple. En el segundo caso, con 1 cara y 11 cruces el factor de Bayes es mayor que 1, con lo que el mejor modelo es el complejo.

El modelo simple tiene solo tres posibles valores para el parámetro frente a los 63 del modelo complejo, lo que implica que el modelo complejo tiene mayor capacidad de ajustar cualquier tipo de datos. Por otro lado, los valores del parámetro que coincidan en ambos modelos tienen mayor prior en el modelo simple porque el número de posibles valores en ese modelo es menor.

En el primer caso, tenemos 3 caras y 9 cruces con lo que son un 25% de caras lo que coincide exactamente con un valor del parámetro, por lo tanto, como tiene mayor prior en el modelo simple, este modelo se ve favorecido y es mejor que el modelo complejo.

El modelo complejo puede ser mejor cuando el modelo simple no ajuste bien los datos, que es precisamente lo que sucede en el segundo caso, con 1 cara y 11 cruces, en el que el porcentaje de caras es $8,\overline{3}\%$, que no se acerca a ninguno de los posibles valores del parámetro en el modelo simple. En este caso, el modelo complejo es mejor según la comparación bayesiana porque tiene valores del parámetro cercanos a la proporción observada y, aunque tienen baja probabilidad prior, consiguen ajustar mejor los datos.