UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
Departamento de Matemáticas

ESTADÍSTICA
Facultad de Ciencias del Mar

PRÁCTICA 2
PROBABILIDAD Y VARIABLES ALEATORIAS **CURSO 2011-12**

R cuenta con numerosas funciones útiles para el cálculo de probabilidades, así como para la modelación y simulación de variables aleatorias. En esta práctica utilizaremos las siguientes:

Funciones de carácter general:

sum(x) Suma de los valores de `x`.

prod(x) Producto de todos los elementos de `x`.

cumsum(x) Sumas acumuladas de los valores de un vector `x`. El término i-ésimo es la suma de los valores `x[1]` hasta `x[i]`.

integrate(fn,lower,upper) Calcula numéricamente la integral de la función `fn` entre los valores `lower` y `upper`.

sample(x, size, replace=FALSE) Extrae sin reemplazamiento una muestra aleatoria de tamaño `n` de valores del vector `x`. La opción `replace=TRUE` realiza el muestreo con reemplazamiento.

replicate(n,expr) Replica `n` veces la expresión `expr`. Ésta puede ser cualquier objeto de R. Resulta particulamente interesante en simulación cuando `expr` es una función.

Combinatoria:

factorial(n) Factorial de `n`

choose(n, k) Calcula el número de combinaciones de n elementos tomados de k en k.

combinations(n, r, v=1:n, set=TRUE, repeats=FALSE) Genera todas las combinaciones de `n` elementos (especificados en el vector `v`, que puede ser de texto) tomados de `r` en `r`, con o sin repetición según se especifique en `repeats` (paquete `gtools`).

permutations(n, r, v=1:n, set=TRUE, repeats=FALSE) Genera todas las variaciones de `n` elementos (especificados en el vector `v`, que puede ser de texto) tomados de `r` en `r`, con o sin repetición según se especifique en `repeats` (paquete `gtools`). Cuando `n=r` se obtienen las permutaciones de `n` elementos.

Distribuciones de probabilidad (cálculo de probabilidades y simulación)

A continuación se muestran funciones para simular variables aleatorias con distintas distribuciones de probabilidad. Si la letra inicial `r` se sustituye por `d` se obtienen los valores de la función de densidad o de probabilidad; si se sustituye por `p` se obtienen los valores de la función de distribución acumulada; y si se sustituye por `q` se obtienen los percentiles que se especifiquen (consultar la ayuda sobre cada distribución para completar esta información).

Distribuciones discretas:

rbinom(n, size, prob) Binomial

rhyper(nn, m, n, k) Hipergeométrica

rgeom(n, prob) Geométrica

rnbinom(n, size, prob) Binomial negativa.

Distribuciones Continuas

rweibull(n, shape, scale) Weibull.

rnorm(n, mean=0, sd=1) Distribución normal (gaussiana).

---

1. En un estanque hay 10 peces, de los que 4 son machos y 6 son hembras.

   a) Genera un vector en **R** que represente los peces del estanque: M1, M2, M3, M4, H1, H2, ..., H6.

Solución:

Utilizamos las funciones `c()` (*concatenar*) y `paste()` (*pegar*, para pegar las letras M y H con los números correspondientes):

```
> peces = c(paste("M", 1:4, sep = ""), paste("H", 1:6, sep = ""))
> peces
 [1] "M1" "M2" "M3" "M4" "H1" "H2" "H3" "H4" "H5" "H6"
```

   b) Utilizando el comando `sample()` extrae dos muestras, una sin reemplazamiento y otra con reemplazamiento, de 6 peces del estanque.

Solución:

Sin reemplazamiento:

```
> sample(peces, 6, replace = FALSE)
[1] "H6" "H2" "H5" "M3" "M2" "H4"
```

Con reemplazamiento:

```
> sample(peces, 6, replace = TRUE)
[1] "M1" "H1" "H2" "H5" "H2" "H3"
```

c) Utilizando los comandos `sample()` y `replicate()` genera 1000 muestras con reemplazamiento de 6 peces del estanque.

Solución:

```
> muestras = t(replicate(1000, sort(sample(peces, 6, replace = FALSE))))
> head(muestras)
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,] "H2" "H3" "H6" "M2" "M3" "M4"
[2,] "H1" "H2" "H3" "H5" "H6" "M1"
[3,] "H1" "H2" "H4" "H6" "M2" "M3"
[4,] "H1" "H2" "H5" "H6" "M1" "M2"
[5,] "H1" "H2" "H4" "H5" "M1" "M4"
[6,] "H2" "H5" "H6" "M1" "M2" "M3"
```

Para hacer una tabla del número de veces que se ha observado cada posible resultado, convertimos la matriz anterior en un vector:

```
> muestras = apply(muestras, 1, paste, collapse = "-")
> head(table(muestras))
muestras
H1-H2-H3-H4-H5-H6 H1-H2-H3-H4-H5-M1 H1-H2-H3-H4-H5-M2 H1-H2-H3-H4-H5-M3
                5                 1                 3                 3
H1-H2-H3-H4-H5-M4 H1-H2-H3-H4-H6-M1
                1                 2
```

d) Escribe el espacio muestral asociado al experimento: *"Extraer sin reemplazamiento 4 peces del estanque"*. Cuenta el número de elementos de este espacio muestral. Calcula dicho número utilizando la fórmula combinatoria adecuada.

Solución:

Como no hay reemplazamiento podemos considerar que los 4 peces se extraen a la vez, por lo que el orden resulta irrelevante. El espacio muestral estará formado por todas las combinaciones de los 10 peces tomados de 4 en 4. Para generar estas combinaciones utilizamos la función `combinations()` del paquete `gtools()`:

```
> require(gtools)
> sinReemp = combinations(10, 4, v = peces)
> head(sinReemp)
     [,1] [,2] [,3] [,4]
[1,] "H1" "H2" "H3" "H4"
[2,] "H1" "H2" "H3" "H5"
[3,] "H1" "H2" "H3" "H6"
[4,] "H1" "H2" "H3" "M1"
[5,] "H1" "H2" "H3" "M2"
[6,] "H1" "H2" "H3" "M3"
```

(Hemos usado aquí el comando `head()` para que nos muestre solamente las primeras filas de la matriz `sinReemp`). Como puede verse, **R** nos devuelve el espacio muestral en forma de matriz, cada una de cuyas filas corresponde a un resultado posible. Podemos contar el número de elementos de este espacio contando el número total de filas de dicha matriz:

```
> nrow(sinReemp)
[1] 210
```

La fórmula combinatoria para obtener dicho número es:

$$C_{10}^{4} = \binom{10}{4} = \frac{10!}{6!4!} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2} = 210$$

que en **R** puede calcularse simplemente mediante:

```
> choose(10, 4)
[1] 210
```

e) Escribe el espacio muestral asociado al experimento "*Extraer, con reemplazamiento, 4 peces del estanque*". Cuenta el número de elementos de este espacio muestral. Calcula dicho número utilizando la fórmula combinatoria adecuada.

Solución:

Como ahora hay reemplazamiento, el orden de extracción resulta relevante. El espacio muestral estará formado por todas las formas posibles de extraer 4 peces entre 10 peces pudiendo repetirlos e importando el orden (variaciones con repetición de 10 objetos tomados de 4 en 4). Para generar este espacio muestral utilizamos la función `permutations()`, también del paquete `gtools()`:

```
> conReemp = permutations(10, 4, v = peces, repeats = TRUE)
> head(conReemp)
     [,1] [,2] [,3] [,4]
[1,] "H1" "H1" "H1" "H1"
[2,] "H1" "H1" "H1" "H2"
[3,] "H1" "H1" "H1" "H3"
[4,] "H1" "H1" "H1" "H4"
[5,] "H1" "H1" "H1" "H5"
[6,] "H1" "H1" "H1" "H6"
> nrow(conReemp)
[1] 10000
```

La fórmula combinatoria para obtener dicho número es:

$$VR_{10}^{4} = 10^{4} = 10000$$

f) Consideramos ahora la variable aleatoria X_M=*"Número de machos entre los cuatro peces extraídos del estanque".* Calcula el valor de dicha variable aleatoria para cada suceso del espacio muestral del apartado (b) (muestreo sin reemplazamiento). Utilizando dicha información calcula la distribución de probabilidad de X_M (esto es, calcula $P\left(X_M = k\right)$ para $k = 0, 1, 2, 3, 4$)

Solución:

La identificación del sexo de cada pez se realiza leyendo la primera letra de su denominación. Para leer dicha letra podemos utilizar la función `substr(x,1,1)` (extrae del *string* `x` los caracteres desde la posición 1 hasta la posición 1). Así, por ejemplo:

```
> suceso = c("H1", "H2", "M1", "M2")
> substr(suceso, 1, 1)
[1] "H" "H" "M" "M"
```

Ahora, para aplicar esta función a todas las filas de `sinReemp` podemos utilizar `apply()`, del siguiente modo:

```
> sex = t(apply(sinReemp, 1, substr, 1, 1))
```

Podemos leer esta sintaxis como "*aplicar a la matriz `sinReemp` por filas (`1`), la función `substr`, pasándole a esta última función los valores `1` y `1`*". Se ha aplicado también la función *trasponer* `t` para que el resultado de `apply()` vuelva a ser una matriz cuyas filas coincidan con los sucesos del espacio muestral. Vemos a continuación las primeras filas de la matriz `sex` que hemos generado:

```
> head(sex)
     [,1] [,2] [,3] [,4]
[1,] "H"  "H"  "H"  "H"
[2,] "H"  "H"  "H"  "H"
[3,] "H"  "H"  "H"  "H"
[4,] "H"  "H"  "H"  "M"
[5,] "H"  "H"  "H"  "M"
[6,] "H"  "H"  "H"  "M"
```

Por último, para contar machos y hembras, reconvertimos las "M" en 1 y las "H" en 0. Al sumar por filas, obtendremos de esta forma el número de machos en cada suceso elemental:

```
> sex = ifelse(sex == "M", 1, 0)
> head(sex)
     [,1] [,2] [,3] [,4]
[1,]    0    0    0    0
[2,]    0    0    0    0
[3,]    0    0    0    0
[4,]    0    0    0    1
[5,]    0    0    0    1
[6,]    0    0    0    1
> rowSums(sex)
  [1] 0 0 0 1 1 1 1 0 0 1 1 1 1 0 1 1 1 1 1 1 1 2 2 2 2 2 2 0 0 1 1
 [33] 1 1 0 1 1 1 1 1 1 1 1 2 2 2 2 2 2 0 1 1 1 1 1 1 1 2 2 2 2 2 2
 [65] 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 0 0 1 1 1 1 0 1 1 1 1 1
 [97] 1 1 1 2 2 2 2 2 2 0 1 1 1 1 1 1 1 2 2 2 2 2 2 1 1 1 1 2 2 2 2
[129] 2 2 2 2 2 2 2 2 3 3 3 3 0 1 1 1 1 1 1 1 1 2 2 2 2 2 2 1 1 1 1 2
[161] 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 3
[193] 3 3 3 2 2 2 2 2 2 3 3 3 3 3 3 3 3 4
```

La función `table` nos muestra el número de formas en que se puede producir cada posible valor de X_M:

```
> XM = table(rowSums(sex))
> XM
 0  1  2  3  4
15 80 90 24  1
```

La distribución de X_M se obtiene dividiendo el número de casos favorables a cada posible valor de X_M entre el número total de casos posibles (210):

```
> XM/210
        0         1         2         3         4
0.0714286 0.3809524 0.4285714 0.1142857 0.0047619
```

o también:

```
> prop.table(XM)
        0         1         2         3         4
0.0714286 0.3809524 0.4285714 0.1142857 0.0047619
```

Comprobamos que la suma es 1 (de lo contrario las probabilidades estarían mal calculadas):

```
> sum(prop.table(XM))
[1] 1
```

g) Repite el apartado anterior para el espacio muestral del apartado (c) (muestreo con reemplazamiento).

Solución:

```
> sex = t(apply(conReemp, 1, substr, 1, 1))
> sex = ifelse(sex == "M", 1, 0)
> XM = table(rowSums(sex))
> XM
   0    1    2    3    4
1296 3456 3456 1536  256
> prop.table(XM)
     0      1      2      3      4
0.1296 0.3456 0.3456 0.1536 0.0256
> sum(prop.table(XM))
[1] 1
```

h) Calcula de nuevo las distribuciones de probabilidad de los apartados anteriores utilizando `dhyper()` y `dbinom()` respectivamente.

Solución:

Cuando se dispone de una población de N objetos divididos en dos clases E y $\overline{E}$, de entre los cuales se extraen n objetos al azar, el número X_E de objetos de clase E entre los k extraidos es una variable aleatoria que:

- Si el muestreo se realiza **sin reeemplazamiento** recibe el nombre de **hipergeométrica** y:

$$P\left(X_E = k\right) = \frac{\binom{N_E}{k}\binom{N - N_E}{n-k}}{\binom{N}{n}} = \text{dhyper}(k, N_E, N - N_E, n)$$

  siendo N_E el número de objetos de clase E en la población (por tanto $N - N_E$ es el número de objetos de la otra clase).

- Si el muestreo se realiza **con reemplazamiento** recibe el nombre de **binomial** y:

$$P\left(X_E = k\right) = \binom{n}{k} p^k \left(1-p\right)^{n-k} = \texttt{dbinom(k,n,p)}$$

  siendo $p = \frac{N_E}{N}$ la probabilidad de que un objeto elegido al azar de esa población sea de clase E.

Así pues, la distribución de probabilidad del número de machos X_M entre 4 peces elegidos al azar de la población descrita puede calcularse del siguiente modo:

- Cuando **no hay reemplazamiento:**

  $P\left(X_M = 0\right)$=`dhyper(0,4,6,4)`=0.0714286
  $P\left(X_M = 1\right)$=`dhyper(1,4,6,4)`=0.380952
  $P\left(X_M = 2\right)$=`dhyper(2,4,6,4)`=0.428571
  $P\left(X_M = 3\right)$=`dhyper(3,4,6,4)`=0.114286
  $P\left(X_M = 4\right)$=`dhyper(4,4,6,4)`=0.0047619

  O, en una sóla instrucción:

```
> dhyper(0:4, 4, 6, 4)
[1] 0.0714286 0.3809524 0.4285714 0.1142857 0.0047619
```

- Cuando **hay reemplazamiento:**

  $P\left(X_M = 0\right)$=`dbinom(0,4,0.4)`=0.1296
  $P\left(X_M = 1\right)$=`dbinom(1,4,0.4)`=0.3456
  $P\left(X_M = 2\right)$=`dbinom(2,4,0.4)`=0.3456
  $P\left(X_M = 3\right)$=`dbinom(3,4,0.4)`=0.1536
  $P\left(X_M = 4\right)$=`dbinom(4,4,0.4)`=0.0256

O, en una sóla instrucción:

```
> dbinom(0:4, 4, 0.4)
[1] 0.1296 0.3456 0.3456 0.1536 0.0256
```

i) Se realizan sucesivas extracciones con reemplazamiento de peces del estanque hasta que sale el primer macho. Sea X_H el número de hembras extraídas hasta ese momento. Calcula $P(X_H = k)$ para $k = 0, 1, 2, 3, 4$ y 5.

Solución:

Cuando se realizan sucesivas observaciones independientes, cada una de las cuales puede ser E o su contrario $\overline{E}$, llamando $p = P(E)$, la probabilidad de que se observe k veces $\overline{E}$ antes de observar el primer E puede calcularse fácilmente como:

$$P(X = k) = P\left(\overline{E}\right) P\left(\overline{E}\right) \ldots P\left(\overline{E}\right) P(E) = \left(1 - P\left(\overline{E}\right)\right)^k P(E) = (1-p)^k p$$

Esta distribución de probabilidad se llama **geométrica.** En **R** :

$$P(X = k) = \texttt{dgeom(k,p)}$$

En este problema se cuenta el número de hembras antes del primer macho. Como las extracciones son con reemplazamiento, en cada extracción se tiene:

$$P(H) = \frac{6}{10} \qquad P(M) = \frac{4}{10}$$

Por tanto:

$P(X_H = 0) = P(M) = 0{,}4$

$P(X_H = 1) = P(H)\,P(M) = 0{,}6 \cdot 0{,}4 = 0{,}24$

P$(X_H = 2) = P(H)\,P(H)\,P(M) = 0{,}6^2 \cdot 0{,}4 = 0{,}144$

P$(X_H = 3) = P(H)\,P(H)\,P(H)\,P(M) = 0{,}6^3 \cdot 0{,}4 = 0{,}0864$

etc. En **R** :

```
> dgeom(0:5, 0.4)
[1] 0.400000 0.240000 0.144000 0.086400 0.051840 0.031104
```

j) Se realizan sucesivas extracciones con reemplazamiento de peces del estanque hasta que sale el cuarto macho. Sea X_H el número de hembras extraídas hasta ese momento. Calcula $P(X_H = k)$ para $k = 0, 1, 2, 3, 4$ y 5.

Solución:

Cuando se realizan sucesivas observaciones independientes, cada una de las cuales puede ser E o su contrario $\overline{E}$, llamando $p = P(E)$, la probabilidad de que se observe k veces $\overline{E}$ antes de observar el r-ésimo E puede calcularse teniendo en cuenta que, en tal caso, antes del r-ésimo E se habrá observado k veces $\overline{E}$ y $r-1$ veces E. Este suceso puede ocurrir de tantas formas como maneras hay de ordenar esos $k+r-1$ objetos teniendo en cuenta las repeticiones, esto es, $PR_{k+r-1}^{k,r-1} = \frac{(k+r-1)!}{k!(r-1)!} = \binom{k+r-1}{k}$. Cada uno de tales sucesos tiene como probabilidad de ocurrir $P\left(\overline{E}\right)^k P(E)^{r-1} P(E)$ (se multiplica de nuevo por $P(E)$ porque la última observación ha de ser necesariamente E). Por tanto, la probabilidad total de observar k veces $\overline{E}$ antes de la r-ésima observación de E es:

$$P(X=k) = \binom{k+r-1}{k} \left(1 - P\left(\overline{E}\right)\right)^k P(E)^r = \binom{k+r-1}{k} (1-p)^k p^r$$

Esta distribución de probabilidad se llama **binomial negativa.** En **R** :

$$P(X=k) = \texttt{dnbinom(k,r,p)}$$

En este problema se cuenta el número de hembras antes del cuarto macho. Como las extracciones son con reemplazamiento, en cada extracción se tiene:

$$P(H) = \frac{6}{10} \qquad P(M) = \frac{4}{10}$$

Por tanto:

$P(X_H = 0) = P(M)^4 = 0{,}4^4 =$0.0256

$P(X_H = 1) = \binom{4}{1} P(H) P(M)^4 = 4 \cdot 0{,}6 \cdot 0{,}4^4 =$0.06144

P$(X_H = 2) = \binom{5}{2} P(H)^2 P(M)^4 = 10 \cdot 0{,}6^2 \cdot 0{,}4^4 =$0.09216

P$(X_H = 3) = \binom{6}{3} P(H)^3 P(M)^4 = 20 \cdot 0{,}6^3 \cdot 0{,}4^4 =$0.110592

etc. En **R** :

```
> dnbinom(0:5, 4, 0.4)
[1] 0.025600 0.061440 0.092160 0.110592 0.116122 0.111477
```

k) Simula y cuenta el número de veces que ocurre cada valor de:

1) El número de machos en 10000 extracciones de 4 peces (sin reemplazamiento) del estanque.

Solución:

```
> simh = rhyper(10000, 4, 6, 4)
> table(simh)
simh
   0    1    2    3    4
 690 3815 4323 1127   45
```

2) El número de machos en 10000 extracciones de 4 peces (sin reemplazamiento) del estanque.

Solución:

```
> simb = rbinom(10000, 10, 0.4)
> table(simb)
simb
   0    1    2    3    4    5    6    7    8    9   10
  56  381 1206 2148 2461 2029 1147  455  102   14    1
```

3) El número de hembras antes del primer macho cuando se repite 10000 veces el experimento descrito en el apartado (g).

Solución:

```
> simg = rgeom(10000, 0.4)
> table(simg)
simg
   0    1    2    3    4    5    6    7    8    9   10   11   12   13
3942 2371 1479  912  549  278  167  116   76   46   29   11    8    6
  14   15   16   18
   3    3    3    1
```

4) El número de hembras antes del cuarto macho cuando se repite 10000 veces el experimento descrito en el apartado (h).

Solución:

```
> simbn = rnbinom(10000, 4, 0.4)
> table(simbn)
simbn
   0    1    2    3    4    5    6    7    8    9   10   11   12   13
 280  616  904 1110 1164 1193  996  874  652  584  418  315  280  179
  14   15   16   17   18   19   20   21   22   23   24   25   26   27
 135  103   64   36   35   22   16    8    5    3    1    2    4    1
```

l) Calcula las frecuencias relativas correspondientes a las distintas simulaciones realizadas en el apartado anterior y compáralas con las probabilidades teóricas.

Solución:

Distribución hipergeométrica:

- Frecuencias relativas observadas en la simulación:

```
> prop.table(table(simh))
simh
     0      1      2      3      4
0.0690 0.3815 0.4323 0.1127 0.0045
```

- Probabilidades teóricas:

```
> dhyper(0:4, 4, 6, 4)
[1] 0.0714286 0.3809524 0.4285714 0.1142857 0.0047619
```

Distribución binomial

- Frecuencias relativas observadas en la simulación:

```
> prop.table(table(simb))
simb
     0      1      2      3      4      5      6      7      8      9
0.0056 0.0381 0.1206 0.2148 0.2461 0.2029 0.1147 0.0455 0.0102 0.0014
    10
0.0001
```

- Probabilidades teóricas:

```
> dbinom(0:max(simb), 10, 0.4)
 [1] 0.006046618 0.040310784 0.120932352 0.214990848 0.250822656
 [6] 0.200658125 0.111476736 0.042467328 0.010616832 0.001572864
[11] 0.000104858
```

Distribución geométrica

- Frecuencias relativas observadas en la simulación:

```
> prop.table(table(simg))
simg
     0      1      2      3      4      5      6      7      8      9
0.3942 0.2371 0.1479 0.0912 0.0549 0.0278 0.0167 0.0116 0.0076 0.0046
    10     11     12     13     14     15     16     18
0.0029 0.0011 0.0008 0.0006 0.0003 0.0003 0.0003 0.0001
```

- Probabilidades teóricas:

```
> noquote(formatC(dgeom(0:max(simg), 0.4), format = "f",
     digits = 5))
 [1] 0.40000 0.24000 0.14400 0.08640 0.05184 0.03110 0.01866 0.01120
 [9] 0.00672 0.00403 0.00242 0.00145 0.00087 0.00052 0.00031 0.00019
[17] 0.00011 0.00007 0.00004
```

Distribución binomial negativa

- Frecuencias relativas observadas en la simulación:

```
> prop.table(table(simbn))
simbn
     0      1      2      3      4      5      6      7      8      9
0.0280 0.0616 0.0904 0.1110 0.1164 0.1193 0.0996 0.0874 0.0652 0.0584
    10     11     12     13     14     15     16     17     18     19
0.0418 0.0315 0.0280 0.0179 0.0135 0.0103 0.0064 0.0036 0.0035 0.0022
    20     21     22     23     24     25     26     27
0.0016 0.0008 0.0005 0.0003 0.0001 0.0002 0.0004 0.0001
```

- Probabilidades teóricas:

```
> noquote(formatC(dnbinom(0:max(simbn), 4, 0.4), format = "f",
     digits = 5))
 [1] 0.02560 0.06144 0.09216 0.11059 0.11612 0.11148 0.10033 0.08600
 [9] 0.07095 0.05676 0.04427 0.03381 0.02536 0.01872 0.01364 0.00982
[17] 0.00700 0.00494 0.00346 0.00240 0.00166 0.00114 0.00077 0.00053
[25] 0.00035 0.00024 0.00016 0.00011
```

Nota: Se ha empleado el comando `formatC()` para forzar a **R** a mostrar los resultados en coma flotante (sin notación científica); asimismo `noquote()` impide que los valores se muestren entre comillas.

2. Se dispone de 11 peceras, numeradas del 0 al 10. En cada pecera hay 10 peces, distribuidos por sexos de tal manera que en la pecera N (con N de 0 a 10) hay N machos y $10 - N$ hembras.

   a) Se elige una pecera al azar y se sacan 5 peces (sin reemplazamiento) ¿cuál es la probabilidad de que sean 2 machos y 3 hembras?

Solución:

Sea X_M el número de machos entre los 5 peces extraídos. Si se elige la pecera N, que contiene N machos y $10 - N$ hembras, la probabilidad de que $X_M = 2$ es:

$$P\left(X_M = 2 \,|\text{Pecera } N\right) = \frac{\binom{N}{2}\binom{10-N}{3}}{\binom{10}{5}} = \texttt{dhyper(2,N,10-N,5)}$$

De acuerdo con el teorema de la probabilidad total:

$$P\left(X_M = 2\right) = \sum_{N=0}^{10} P\left(X_M = 2 \,|\text{Pecera } N\right) P\left(\text{Pecera } N\right)$$

Dado que la pecera se elige al azar, la probabilidad de elegir cualquier pecera será la misma, esto es,

$$P\left(\text{Pecera } N\right) = \frac{1}{11}, \quad N = 0, 1, 2, \ldots, 10$$

Por tanto:

$$P\left(X_M = 2\right) = \sum_{N=0}^{10} \texttt{dhyper(2,N,10-N,5)} \cdot \frac{1}{11}$$

Para calcular esta suma cómodamente en **R** podemos definir la función:

```
> p_extraer2M = function(N) {
    dhyper(2, N, 10 - N, 5)
 }
```

y calcular sus valores de 0 a 10:

```
> p_extraer2M(0:10)
 [1] 0.0000000 0.0000000 0.2222222 0.4166667 0.4761905 0.3968254
 [7] 0.2380952 0.0833333 0.0000000 0.0000000 0.0000000
```

La suma del teorema de la probabilidad total puede calcularse entonces como:

```
> sum(p_extraer2M(0:10))/11
[1] 0.166667
```

1) Se elige una pecera al azar y se sacan 5 peces, que resultan ser 2 machos y 3 hembras ¿Cuál es la probabilidad de que la pecera sea la número 5?

Solución:

Aquí podemos aplicar el teorema de Bayes:

$$P\left(\text{Pecera } 5\,|X_M = 2\right) = \frac{P\left(X_M = 2\,|\text{Pecera } 5\right) P\left(\text{Pecera } 5\right)}{\sum_{N=0}^{10} P\left(X_M = 2\,|\text{Pecera } N\right) P\left(\text{Pecera } N\right)}$$

que se calcula fácilmente en **R** mediante:

```
> p_extraer2M(5) * (1/11)/sum(p_extraer2M(0:10))/11
[1] 0.00178884
```

2) Se elige una pecera al azar y se sacan 7 peces, que resultan ser 3 machos y 4 hembras. ¿Qué pecera es más probable que se haya elegido?

Solución:

Nuevamente aplicamos el teorema de Bayes:

$$\begin{aligned} P\left(\text{Pecera } K\,|X_M = 3 \text{ entre } 7\right) &= \frac{P\left(X_M = 3 \text{ entre } 7\,|\text{Pecera } K\right) P\left(\text{Pecera } K\right)}{\sum_{N=0}^{10} P\left(X_M = 3 \text{ entre } 7\,|\text{Pecera } N\right) P\left(\text{Pecera } N\right)} = \\ &= \frac{P\left(X_M = 3 \text{ entre } 7\,|\text{Pecera } K\right)}{\sum_{N=0}^{10} P\left(X_M = 3 \text{ entre } 7\,|\text{Pecera } N\right)} \end{aligned}$$

(hemos simplificado $P\left(\text{pecera } K\right)$ con $P\left(\text{pecera } N\right)$ ya que estas probabilidades valen siempre $\frac{1}{11}$ para cualesquiera K y N). Para calcular esta probabilidad para distintos valores de K en **R** definimos en primer lugar una función que calcule la probabilidad de que haya 3 machos entre 7 peces extraídos de la pecera N:

```
> p_extraer3M = function(N) {
    dhyper(3, N, 10 - N, 7)
 }
```

y a partir de esta función construimos otra función que calcule $P\left(\text{Pecera } K\,|X_M = 3 \text{ entre } 7\right)$ de acuerdo con el resultado anterior:

```
> probPecera = function(K) {
    p_extraer3M(K)/sum(p_extraer3M(0:10))
 }
```

Calculamos ahora el valor de esta función para $K = 0, 1, 2, \ldots, 10$:

```
> probPecera(0:10)
```

```
[1] 0.000000 0.000000 0.000000 0.212121 0.363636 0.303030 0.121212
[8] 0.000000 0.000000 0.000000 0.000000
```

de donde se sigue que lo más probable es que los peces hayan sido extraídos de la pecera 4.

3. La duración T (en horas) de una tormenta en el Mar del Norte es una variable aleatoria cuya función de distribución de probabilidad es de la forma:

$$F(t) = P(T \leq t) = 1 - \exp\left(-(t/\lambda)^{\kappa}\right) \quad : \quad t \geq 0$$

(*distribución de Weibull*). Se ha estimado que durante el invierno los parámetros κ y λ de dicha función de distribución valen, respectivamente,$\kappa = 2$ y $\lambda = 10$. Considerando estos valores:

a) Implementa la función $F(t)$ y dibújala en el intervalo $[0, 25]$.

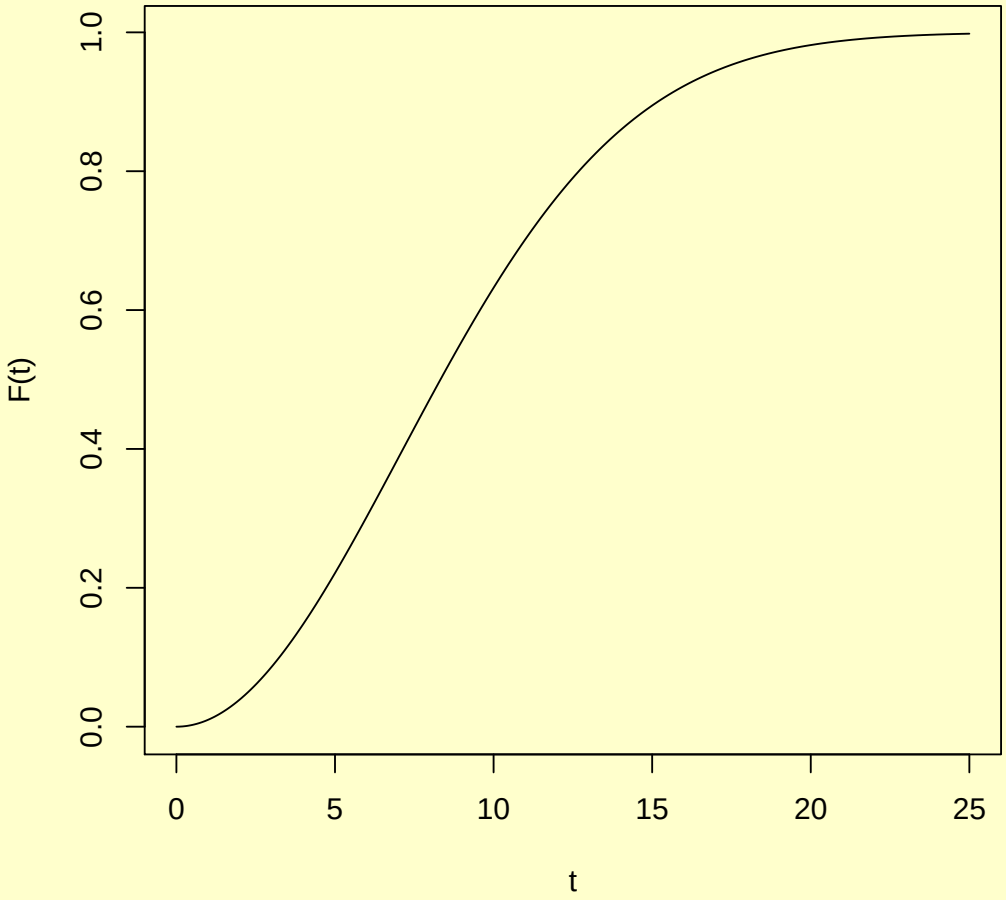
Solución:

Definimos primero la función:

```
> F = function(t, lambda = 10, k = 2) {
     1 - exp(-(t/lambda)^k)
 }
```

y ahora fijamos el soporte sobre el que se traza la curva, que dibujaremos utilizando `plot()`. El soporte se construye mediante el comando `seq()` que permite generar una colección de valores equiespaciados; en nuestro caso, para que la curva se vea bien, generamos 500 valores entre 0 y 25. En el comando `plot()` utilizamos la opción `type='l'` para que represente una linea continua.

```
> t = seq(0, 25, length = 500)
> plot(t, F(t), type = "l")
```



b) Calcula e implementa en **R** la función de densidad de la variable aleatoria anterior. Dibújala también en el intervalo $[0, 25]$.

Solución:

La función de densidad es la derivada de la función de distribución. En este caso:

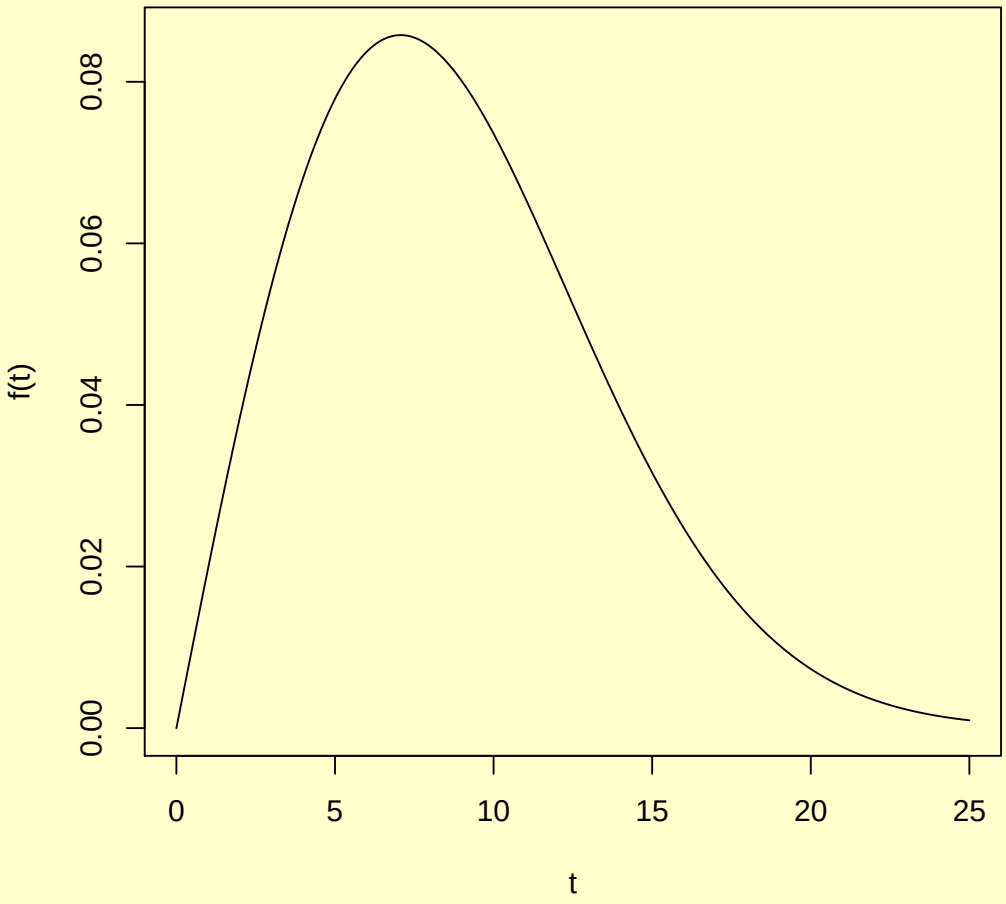
$$f(t) = \kappa \left(\frac{t}{\lambda}\right)^{\kappa-1} \frac{1}{\lambda} \exp\left(-\left(\frac{t}{\lambda}\right)^{\kappa}\right)$$

Definimos esta función en **R** :

```
> f = function(t, lambda = 10, k = 2) {
     (k/lambda) * ((t/lambda)^(k - 1)) * exp(-(t/lambda)^k)
 }
```

Podemos dibujar ahora la gráfica del mismo modo que antes:

```
> plot(t, f(t), type = "l")
```



c) Calcula, utilizando la función de distribución de probabilidad,

1) La probabilidad de que una tormenta dure menos de 10 horas.
2) La probabilidad de que una tormenta dure entre 2 y 8 horas.
3) La probabilidad de que una tormenta aún dure al menos tres horas más si hace ya dos horas que comenzó.

Solución:

Para calcular la primera probabilidad basta observar que es justamente el valor de la función de distribución en el 10, esto es, $P\left(T \leq 10\right) = F\left(10\right)$. Por tanto:

```
> F(10)
[1] 0.632121
```

Para resolver el segundo apartado hemos de tener en cuenta que $P(2 < T \leq 8) = P(T \leq 8) - P(T \leq 2) = F(8) - F(2)$. En **R** :

```
> F(8) - F(2)
[1] 0.433497
```

Por último, el tercer apartado corresponde a la probabilidad:

$$P(T \geq 5 \,|T \geq 2) = \frac{P(T \geq 5 \cap T \geq 2)}{P(T \geq 2)} = \frac{P(T \geq 5)}{P(T \geq 2)} = \frac{1 - F(5)}{1 - F(2)}$$

que podemos calcular fácilmente en **R** mediante:

```
> (1 - F(5))/(1 - F(2))
[1] 0.810584
```

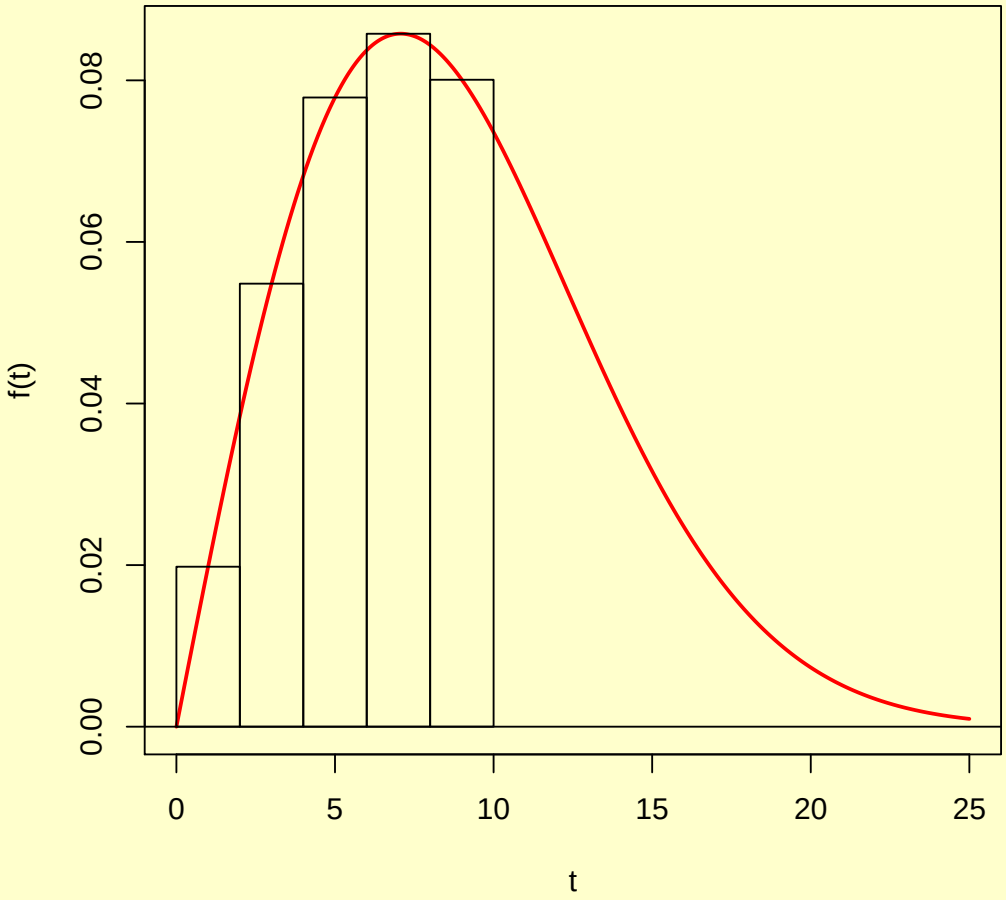
d) Repite el apartado anterior, pero utilizando ahora la función de densidad de probabilidad para realizar el cálculo.

Solución:

Para calcular la primera probabilidad, $P(X \leq 10)$, tenemos en cuenta que dicha probabilidad es justamente el area bajo la curva sobre el intervalo $[0, 10]$, esto es:

$$P(X \leq 10) = \int_0^{10} f(x)\, dx$$

Para calcular esta área de manera aproximada construimos rectángulos (cuya área es fácil de calcular) que rellenen el espacio bajo la curva y sumamos sus áreas. De esta forma habremos obtenido un valor aproximado de la probabilidad buscada. Obviamente, si construimos los rectángulos con una base más pequeña, rellenarán mejor el espacio bajo la curva y por tanto obtendremos una aproximación cada vez mejor al valor del área que buscamos.



La siguiente función permite construir y dibujar estos rectángulos utilizando **R** . Para ello definimos la base (`delta`) de cada rectángulo, así como los valores inicial (`ini`) y final (`fin`) de x entre los cuales dibujaremos nuestros rectángulos; como altura para cada rectángulo asignamos el valor de la función de densidad $f(x)$ en el centro de la base de cada rectángulo. Por último, la función calcula la suma de las áreas de todos los rectángulos dibujados:

```
> rectangulos = function(delta, ini, fin) {
     baseRect = seq(ini, fin, by = delta)
     n = length(baseRect)
     centroRect = seq(ini + delta/2, fin - delta/2, by = delta)
     hrect = f(centroRect)
     rect(baseRect[-n], 0, baseRect[-1], hrect)
     abline(h = 0)
     sum(hrect * delta)
 }
```

En la gráfica anterior se han construido rectángulos de base 2. La suma de las áreas de estos

rectángulos es:

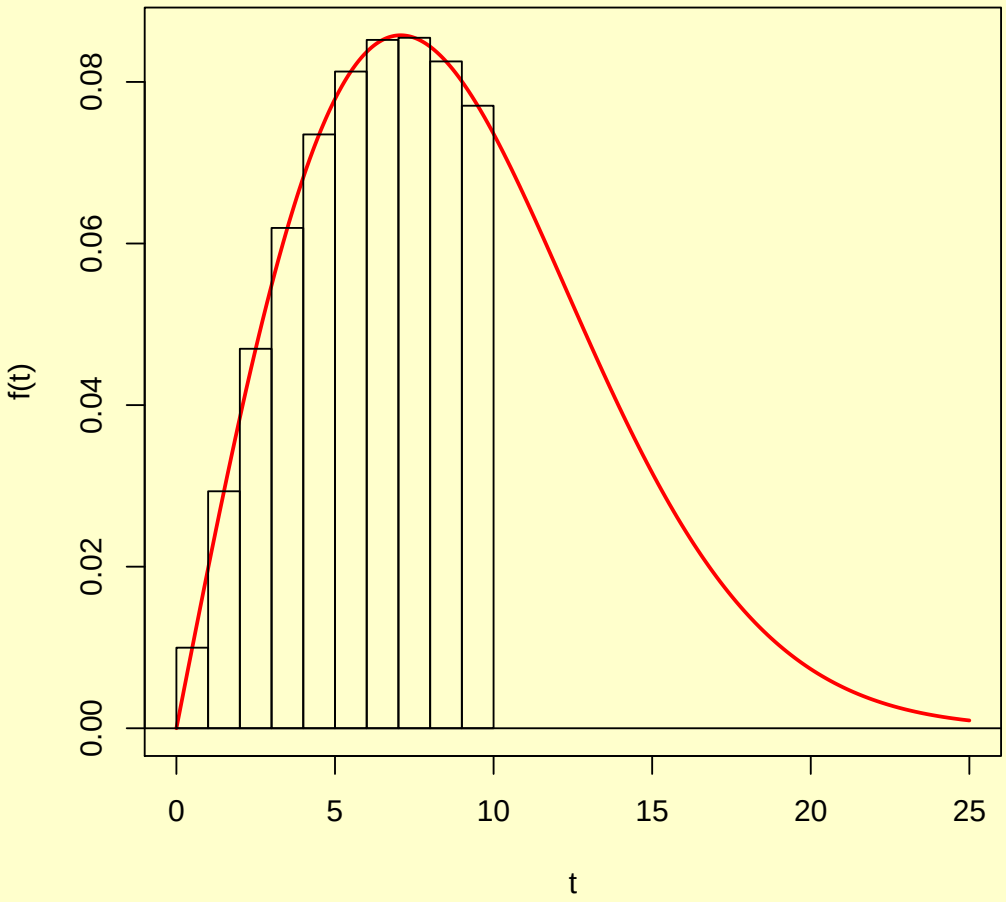
```
> plot(t, f(t), type = "l", col = "red", lwd = 2)
> rectangulos(2, 0, 10)
```

```
[1] 0.636718
```

Si hacemos los rectángulos con base 1 el área calculada es:

```
> plot(t, f(t), type = "l", col = "red", lwd = 2)
> rectangulos(1, 0, 10)
```

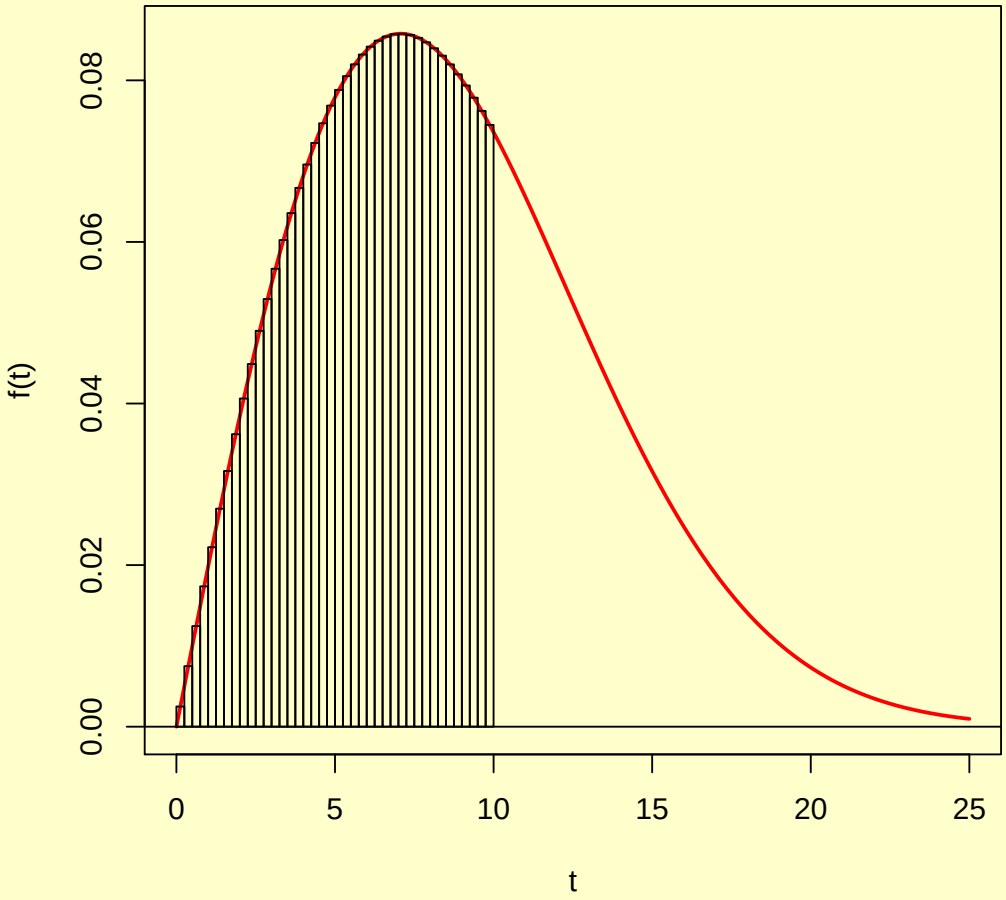
```
[1] 0.633263
```



Podemos hacer la base de los rectángulos aún más pequeña (anchura 0.25) y el área que se obtiene es:

```
> plot(t, f(t), type = "l", col = "red", lwd = 2)
> rectangulos(0.25, 0, 10)
```

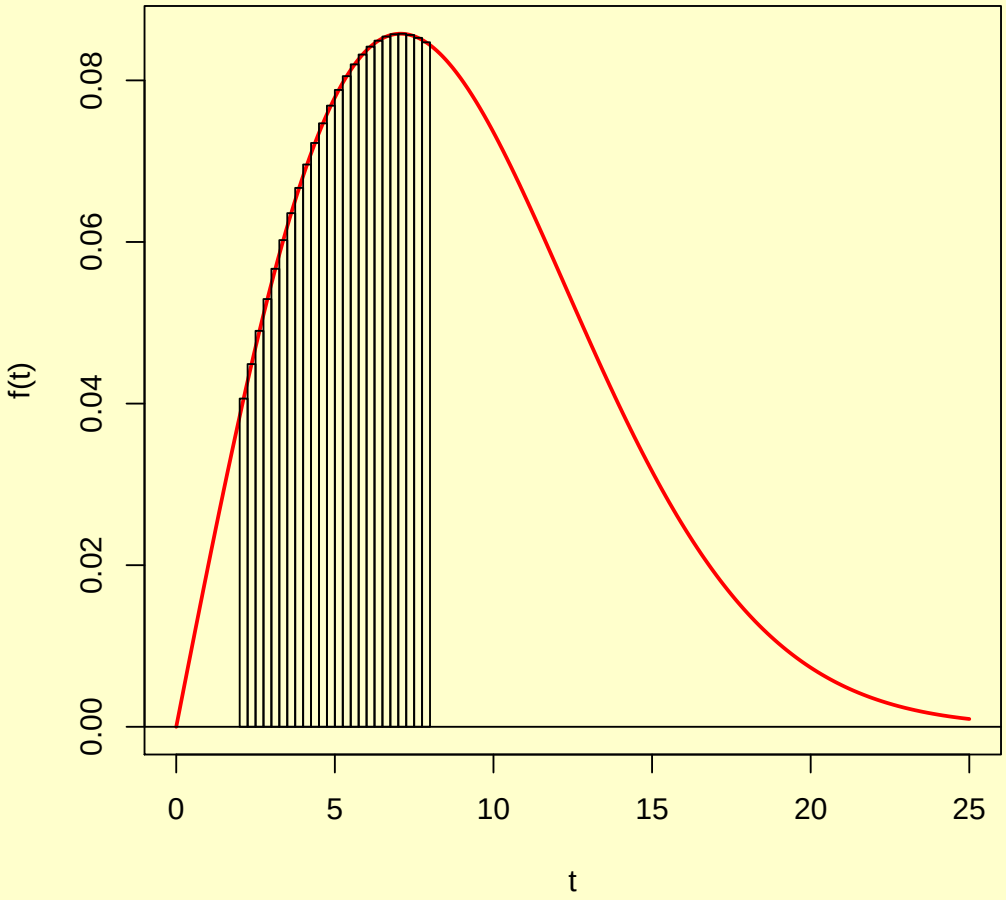
```
[1] 0.632192
```



que nos da un resultado muy similar al ya obtenido mediante la función de distribución; la diferencia se debe al error de aproximación. A medida que hacemos los rectángulos aún más estrechos, tal error se va reduciendo.

Ahora, para calcular la probabilidad pedida en el apartado (b) tenemos que calcular el área que deja la curva sobre el intervalo $[2, 8]$, esto es, $P\left(2 < X \leq 8\right) = \int_2^8 f\left(x\right) dx$. Para ello repetimos lo que hemos hecho anteriormente, pero ahora en este nuevo intervalo:

```
> plot(t, f(t), type = "l", col = "red", lwd = 2)
> rectangulos(0.25, 2, 8)
[1] 0.433551
```



Señalemos por último que **R** ya dispone de una función (`integrate()`) diseñada para calcular numéricamente el área bajo una curva (integral definida). Su aplicación en este caso es inmediata y muy sencilla:

```
> integrate(f, 0, 10)
0.632121 with absolute error < 7e-15
> integrate(f, 2, 8)
0.433497 with absolute error < 4.8e-15
```

La tercera pregunta del apartado anterior también puede resolverse haciendo uso de la función de densidad:

$$P\left(T \geq 5 \,|T \geq 2\right) = \frac{P\left(T \geq 5\right)}{P\left(T \geq 2\right)} = \frac{\int_5^\infty f\left(t\right) dt}{\int_5^\infty f\left(t\right) dt}$$

En **R** :

```
> integrate(f, 5, Inf)$value/integrate(f, 2, Inf)$value
[1] 0.810584
```

En este caso hemos tenido que añadir `$value` para que **R** utilice sólo el valor de la integral, ignorando el error de aproximación (si lo incluyera en el resultado, no podría evaluar el cociente anterior).

e) La distribución de Weibull se encuentra implementada en **R** . En particular, para los valores $\kappa = 2$ y $\lambda = 10$, se tiene $F(t) =$ `pweibull(t,2,10)`. Repite el apartado (c) utilizando esta función.

Solución:

$P(T \leq 10) = F(10) =$ `pweibull(10,2,10)` $= 0{,}632121$

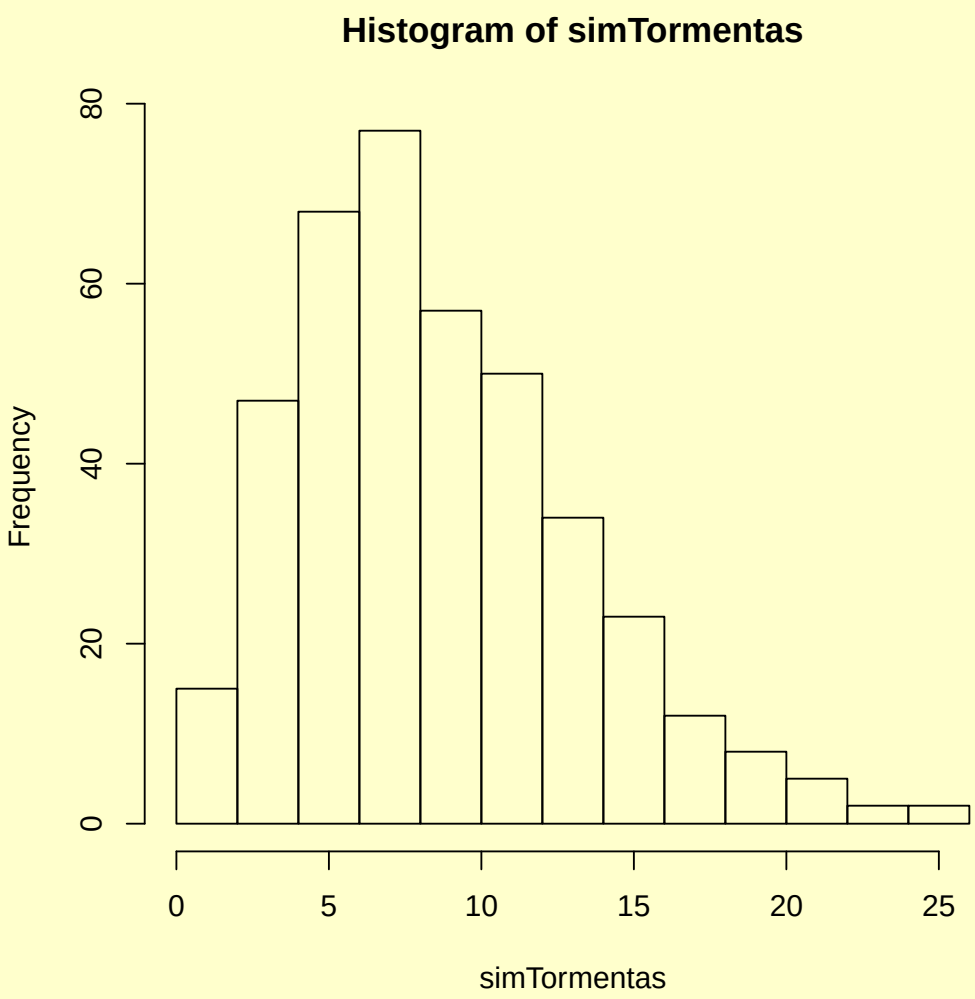
$P(2 < T \leq 8) = F(8) - F(2) =$ `pweibull(8,2,10)-pweibull(2,2,10)` $= 0{,}433497$

$P(T \geq 5 \,|T \geq 2) = \frac{1-F(5)}{1-F(2)} =$ `(1-pweibull(5,2,10))/(1-pweibull(2,2,10))` $= 0{,}810584$

f) Simula las duraciones de 400 tormentas en el Mar del Norte (utilizando la función `rweibull()` y construye el correspondiente histograma de frecuencias.

Solución:

```
> simTormentas = rweibull(400, 2, 10)
> head(simTormentas)
[1] 15.34205  5.57430  7.27309  9.31797  4.09077 15.93605
> hist(simTormentas)
```



4. Una variable aleatoria X sigue una *distribución Normal* de parámetros μ (media) y σ (desviación típica) si su función de densidad de probabilidad es de la forma:

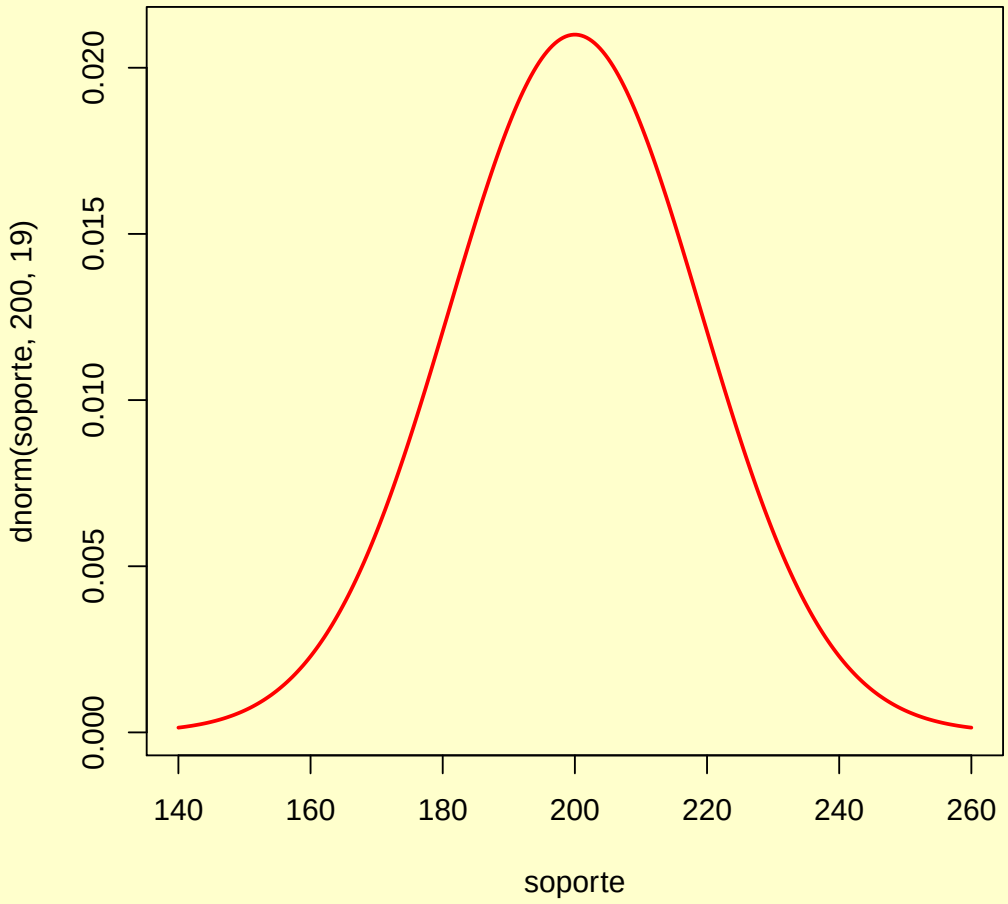
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \ x \in \mathbb{R}$$

que puede calcularse en **R** mediante `dnorm(x,mu,sigma)`. Una empresa de acuicultura dispone de 8 jaulas flotantes para la cría de doradas. La producción anual de cada jaula sigue una distribución normal de media 200 toneladas con desviación típica de 19 toneladas.

a) Representa gráficamente la función de densidad de la producción de una jaula (represéntala entre los valores 140 y 260).

Solución:

```
> soporte = seq(140, 260, length = 300)
> plot(soporte, dnorm(soporte, 200, 19), type = "l", col = "red",
     lwd = 2)
```



b) Calcula la probabilidad de que en una jaula se produzcan menos de 190 toneladas de doradas en un año.

Solución:

Si llamamos $X =$*"Producción de doradas en una jaula durante un año"*, utilizamos la función `pnorm()` para calcular el valor de la función de distribución de dicha variable. Como en el caso de la densidad, esta función lleva como argumentos adicionales los valores de la media μ y la desvaición típica σ:

$$P\left(X \leq 190\right) = \texttt{pnorm(190,200,19)} = 0{,}299334$$

c) Calcula la probabilidad de que en una jaula se produzcan más de 215 toneladas de doradas en un año.

Solución:
De manera similar al apartado anterior:

$$P(X > 215) = 1 - P(X \leq 215) = \texttt{1-pnorm(215,200,19)} = 0{,}214918$$

d) Con probabilidad 0.95 ¿Cuál es la producción total mínima de doradas por año en una jaula arbitraria?

Solución:
Buscamos el valor m tal que $P(X > m) = 0{,}95$, o lo que es lo mismo:

$$\begin{aligned} 1 - P(X > m) &= 1 - 0{,}95 \\ P(X \leq m) &= 0{,}05 \end{aligned}$$

Por tanto m es el percentil 5 de X. Para calcular los cuantiles de la distribución normal se utiliza `qnorm(q,mu,sigma)`:

$$m = \texttt{qnorm(0.05,200,19)} = 168{,}748$$

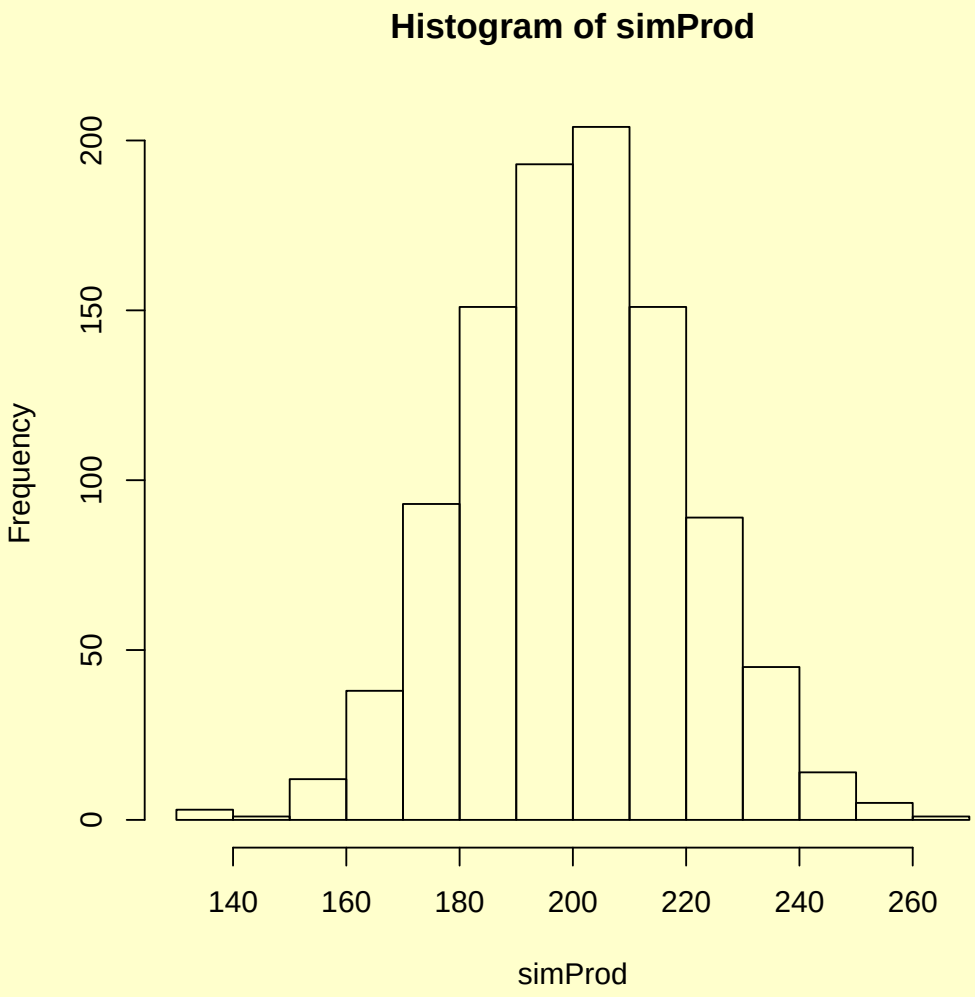
Por tanto, con probabilidad 0.95 la producción mínima anual de una jaula será de 168.748 toneladas de doradas.

e) Simula la producción anual de 1000 jaulas y representa los valores en un histograma. ¿En qué proporción de estas simulaciones se han producido menos de 190 toneladas? ¿En qué proporción de han producido más de 215 toneladas?. Calcula la producción media, así como la desviación típica de las 1000 simulaciones. Calcula los percentiles 5 y 95 de los valores simulados.

Solución:
Para simular valores de la distribución normal utilizamos `rnorm(n,mu,sigma)`, donde n es el número de valores a simular.

```
> simProd = rnorm(1000, 200, 19)
> hist(simProd)
> length(simProd[simProd < 190])/1000
```

```
[1] 0.298
> length(simProd[simProd > 215])/1000
[1] 0.221
> quantile(simProd, probs = c(0.05, 0.95))
     5%     95%
168.943 231.170
```



f) Simula la producción de la empresa durante un año (esto es, la suma de las producciones de 8 jaulas). Repite esta simulación 10000 veces y representa los valores en un histograma. Calcula media, desviación típica y percentiles 0.05 y 0.95 de estos valores. ¿Crees que la suma de la producción de 8 jaulas sigue una distribución normal?

Solución:

```
> simProd8 = sum(rnorm(8, 200, 19))
> simProd8
[1] 1686.53
> sp8 = replicate(10000, sum(rnorm(8, 200, 19)))
> hist(sp8)
> mean(sp8)
[1] 1600.59
> sd(sp8)
[1] 53.9669
> quantile(sp8, probs = c(0.05, 0.95))
     5%     95%
1511.71 1690.44
```