DOCTORADO EN

Investigación en Biomedicina

ULPGC

Análisis y Tratamiento de Datos: aplicaciones con R

Angelo Santana

20-30 de Junio de 2016

Índice general

1 Instalación de R, Rstudio y librerías adicionales.

1.1 Introducción.

R es un lenguaje y un entorno de trabajo orientado a la realización de estudios estadísticos. El manejo de R entraña sobre todo la dificultad de tener que conocer y manejar un lenguaje de programación que a muchos usuarios del mundo de la medicina, de la biología o de las ciencias sociales le puede resultar extraño y complejo. No obstante, para hacer un uso razonablemente eficiente de este lenguaje no es necesario convertirse en un programador avanzado; basta con conocer algunos elementos básicos y, sobre todo, ser capaz de trabajar de manera ordenada documentando adecuadamente la actividad realizada. De esta forma, en la realización de un nuevo trabajo o estudio será posible reutilizar con facilidad los procedimientos empleados en estudios anteriores.

Ello se consigue fundamentalmente a través de la utilización del programa Rstudio (http://www.rstudio.com) que integra diversas herramientas que facilitan la edición de informes que combinan texto, imágenes, tablas, gráficos y resultados del análisis de manera sencilla y fluida. Estas herramientas son básicamente:

- El lenguaje Markdown (https://es.wikipedia.org/wiki/Markdown). Este es un lenguaje formado por unos pocos comandos dedicados al formateo de texto y gráficos, que permite estructurar un documento con muy poco esfuerzo. En http://rmarkdown.rstudio.com/authoring_basics.html se puede encontrar una guía de referencia de los comandos markdown más comunes. Un guía muy similar en español puede encontrarse en http://joedicastro.com/pages/markdown.html.
- La función `knitr` (http://http://yihui.name/knitr/), que permite añadir gráficos, tablas y resultados estadísticos generados por R al texto.
- El conversor de formatos Pandoc (http://pandoc.org) permite convertir el texto escrito en markdown en un documento word, en una página web (con formato html) o en un documento pdf (para la generación automática de archivos pdf es necesario instalar además MiKTeX en windows (http://miktex.org/), o MacTeX (https://tug.org/mactex/) en Mac; para el usuario eventual puede resultar más sencillo simplemente generar los documentos en Word, y generar el pdf a partir del documento word).

Aunque el uso de estas herramientas pueda sonar complicado es en realidad muy sencillo, ya que se encuentran integradas en Rstudio de manera absolutamente transparente. Esto significa que el usuario puede concentrarse perfectamente en el texto que está redactando y en los resultados que obtiene sin tener que preocuparse de qué herramienta se utiliza en cada momento.

1.2 Instalacion de R y Rstudio

Para trabajar con R necesitaremos dos programas: el propio R, que nos proporciona el lenguaje y el entorno de trabajo (el motor de nuestras tareas estadísticas), y Rstudio que añade a R todas las herramientas citadas en la introducción.

La instalacion de ambos programas es muy sencilla. Basta con visitar sus respectivas paginas web (https://www.r-project.org/ y https://www.rstudio.com/), descargar los instaladores y ejecutarlos. Pueden encontrarse instrucciones detalladas para la instalación en https://dl.dropboxusercontent.com/u/7610774/cursoR4ULPGC/03-Instalacion.html

1.3 Librerías

La instalación básica de R viene equipada con múltiples funciones para la importación de datos, la realización de transformaciones, el ajuste y evaluación de modelos estadísticos, las representaciones gráficas, etc. Sin embargo, la enorme potencia de R deriva de su capacidad de incorporar en cualquier momento nuevas funciones capaces de realizar nuevas tareas.

Una librería o paquete (package) es una colección de nuevas funciones, datos y código que se añaden a R. En junio de 2016 hay más de 8600 paquetes disponibles para su descarga e instalación en la web de R: hay paquetes para el análisis de datos genéticos, para epidemiología, para farmacología, psicometría, econometría, datos espacio-temporales y un largo etcétera.

En la dirección https://dl.dropboxusercontent.com/u/7610774/cursoR4ULPGC/06-librerias.html se muestra con detalle como se pueden instalar nuevos paquetes de R en nuestro ordenador. De manera resumida, dicha tarea se puede llevar a cabo en Rstudio de dos formas:

- Accediendo a la pestaña "Packages" en la ventana inferior derecha de Rstudio, a continuación pinchando en el icono "Install" (en la parte superior izquierda de dicha ventana), y escribiendo el nombre del paquete que queremos descargar en el menú desplegable que aparece.
- Escribiendo `install.packages("nombre-del-paquete")` en la consola (ventana inferior izquierda).

Comenzaremos este curso instalando los siguientes paquetes:

- `openxlsx` (para la lectura de archivos excel)

- `ggplot2` (para la elaboración de gráficos)
- `pander` (para la elaboración de tablas)
- `devtools` (para la instalación de paquetes aún en desarrollo)
- `ULPGCmisc` (que contiene una miscelánea de funciones que facilitan la presentación de algunas tablas y gráficos). El paquete `ULPGCmisc` se encuentra aún en desarrollo y debe instalarse de manera ligeramente distinta al resto.

Para descargar e instalar todos estos paquetes podemos utilizar la siguiente sintaxis (deben ejecutarse todas las líneas en el orden especificado):

```
install.packages("openxlsx")
install.packages("ggplot2")
install.packages("digest")
install.packages("pander")
install.packages("devtools")
```

Por último, para instalar `ULPGCmisc`:

```
library(devtools)
install_github("angeloSdP/ULPGCmisc")
```

1.4 Antes de empezar a trabajar con R: cuestiones importantes a tener en cuenta.

- R es un lenguaje, y como tal, tiene sus reglas gramaticales y de sintaxis. Si estas reglas no se cumplen, el procesador de R será incapaz de entender nuestras instrucciones. Cosas muy sencillas que hay que vigilar son:
  - Las comillas: siempre van en pares, unas comillas de apertura y otras de cierre.
  - Los paréntesis: también van por pares, todo paréntesis que se abra debe cerrarse.
  - Los corchetes, las llaves, etc: también van por parejas.
  - Símbolos especiales: la barra inclinada de derecha a izquierda "\" tiene un significado especial en R, ya que sirve (según la letra que la acompañe) para especificar saltos de línea, retornos de carro, etc. Si se desea incluir esa barra en un texto, debe escribirse dos veces.
- Cuando en un documento se combina texto con código R, el código R ejecutable debe ir dentro de un "Chunk" (que presenta la apariencia de una cajita gris, precedida por tres tildes inversas y el símbolo {r}, y terminada también con tres tildes inversas (Ojo con no borrar las tres tildes, pues el código R no sería adecuadamente interpretado).
- Procurar dejar espacios y separar bien los párrafos y los "chunks" para que el texto quede lo más claro posible.

2 Importación de bases de datos

2.1 Bases de datos.

El primer paso para la realización de cualquier análisis estadístico es la elaboración de la base de datos que contenga la información que hemos de analizar. Para construir una base de datos existen muchas herramientas alternativas. Sin duda una de las más sencillas de utilizar (siempre y cuando la base de datos no sea excesivamente grande) es una simple hoja de cálculo, como las que ofrecen Microsoft Excel o LibreOfice Calc.

Las reglas para construir una hoja de cálculo que pueda servir para un análisis estadístico son muy simples:

- Cada fila debe ser un caso (un paciente o sujeto experimental)
- Cada columna debe ser una variable.
- Cada columna debe estar encabezada con el nombre de la variable que represente.
- Las casillas con valores perdidos deben dejarse en blanco, o al menos utilizar un código estándar que luego pueda interpretarse como tal sin posibilidad de confusion.
- Debe evitarse dejar filas vacías, columnas vacías, insertar figuras o introducir decoraciones de cualquier tipo.
- Hay que vigilar con especial cuidado en la introduccion de valores numéricos no mezclar puntos (.) con comas (,). Si en una misma columna escribimos el número 2,3 y más adelante 7.2, R interpretará todos los valores de la variable como caracteres no como números.
- De la misma forma, en una columna donde se registran valores numéricos no deben introducirse caracteres, pues en tal caso R interpretará la variable como de tipo carácter y no como numérica.

2.1.1 Importacion de datos desde Excel:

A modo de ejemplo, en el campus virtual de la asignatura, o pinchando aquí podemos descargar la hoja de cálculo `endocrino.xlsx` en la que se muestran parte de las variables medidas en un estudio transversal realizado, diseñado originalmente con el objetivo de estimar las prevalencias de diabetes mellitus (DM) en la población de Telde (Gran Canaria), así como para identificar los factores asociados con aquella. En el estudio se incluyeron 1030 sujetos con edades comprendidas entre los 30 y 82 años, que respondieron a un cuestionario diseñado para obtener información sobre la edad, sexo, historia personal y familiar de diabetes mellitus y estilos de vida. Se realizaron mediciones antropométricas y de presión arterial. Se tomaron asimismo muestras de sangre sobre las que se obtuvieron diversas mediciones (glucemia, lípidos, marcadores de inflamación, hemostasia, etc).

A continuación se muestran las primeras lineas de este archivo:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	ID	EDAD	SEXO	PESO	TALLA	SEDENTA	INSTRUC	TAS	TAD	HTA_cono	HTA_OMS	A_DIAB	ECV_B	TABACO	ALCOHOL	ST
2	22	51	1	86,3	174	0	Primer gr	125	80	1	1	1	1	0	1	
3	65	33	1	81,8	175	0	Segundo	120	80	0	0	0	0	1	1	
4	116	40	1	105	173	1	Segundo	150	100	0	1	1	0	1	1	
5	144	62	1	98	172	0	Primer gr	155	100	0	1	0	0	1	0	
6	204	33	1	124	181	0	Segundo	130	90	0	1	1	0	0	0	
7	337	33	1	93	177	1	Segundo	110	70	0	0	1	0	0	1	
8	458	68	1	81	174	0	Primer gr	120	70	0	0	1	0	0	0	
9	541	56	1	99	174	1	Tercer gra	110	70	0	0	0	0	0	1	
10	551	55	1	92	179	1	Primer gr	130	80	0	0	0	0	0	1	
11	557	46	1	94,3	169	0	Primer gr	120	80	0	0	1	0	1	1	
12	566	57	1	75	167	1	Primer gr	140	80	0	1	1	0	0	1	
13	576	46	1	93,3	170	1	Primer gr	140	100	0	1	0	0	1	1	
14	611	45	1	93,5	171	0	Tercer gra	110	70	1	1	1	0	0	0	
15	631	58	1	92	160	1	Tercer gra	120	70	0	0	1	0	0	1	
16	654	41	1	122	169	1	Segundo	125	80	0	0	0	0	1	1	
17	794	38	1	82	170	0	Primer gr	120	60	0	0	0	0	0	1	
18	805	55	1	101	167	0	Segundo	120	70	1	1	1	1	0	1	

Figura 1:

Para leer este archivo con R deberemos ejecutar los tres comandos siguientes:

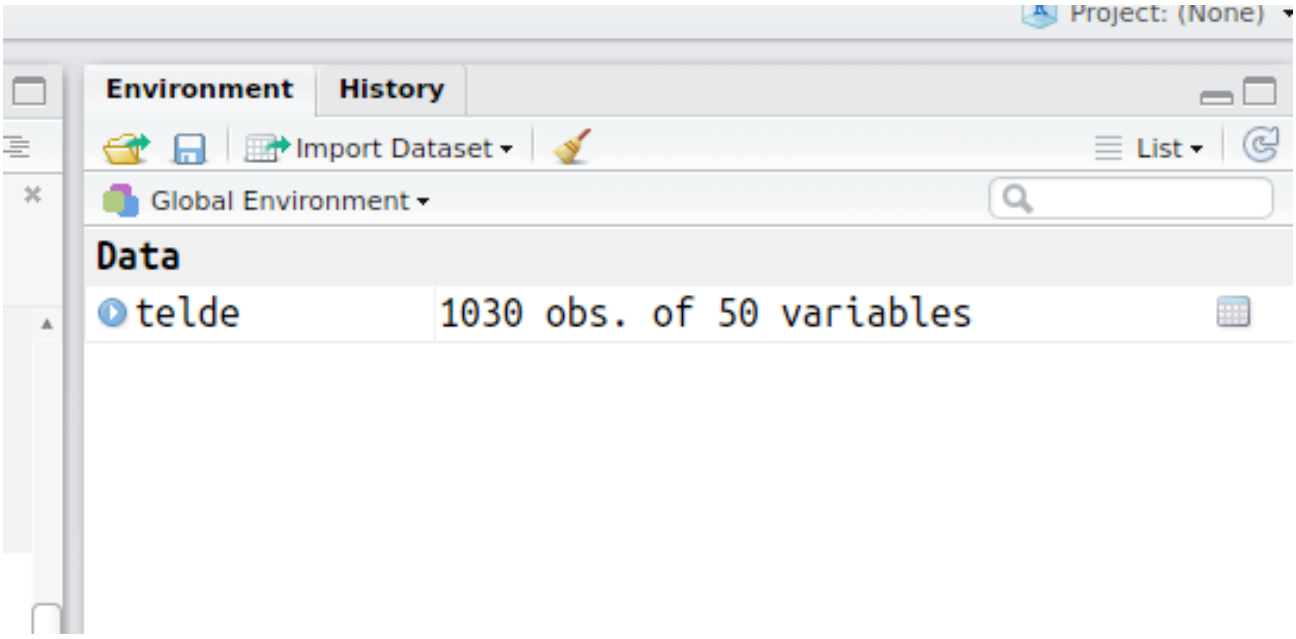
```
library(openxlsx)
setwd("c:/Users/aulas/Downloads/")
telde = read.xlsx("endocrino.xlsx")
```

- El primer comando, `library(openxlsx)`, carga en memoria la librería que permite leer archivos Excel.
- El segundo, `setwd("c:/Users/aulas/Downloads/")` indica a R la carpeta en la que se encuentra el archivo. Nótese que en R las carpetas se especifican mediante la barra "/", y no mediante la barra inversa "\" habitual en Windows.
- El tercero, `telde = read.xlsx("endocrino.xlsx")` lee los datos y los guarda dentro de un objeto llamado `telde`. Este objeto es la base de datos ya leída por R. En la terminología de R, `telde` es

un objeto de tipo `data.frame`.

2.2 Visualización de los datos

Si la base se ha leido correctamente, su nombre debe aparecer en la ventana superior derecha de Rstudio, bajo la pestaña `Environment`:



Si el nombre de la base de datos no aparece ahí es que no se ha leído correctamente.

Podemos ver las primeras líneas de esta base de datos simplemente tecleando `head(telde)` en la consola y pulsando `Return`. De esta forma podemos hacernos una primera idea de qué variables contiene y qué tipo de valores toma:

```
head(telde)
```

```
##    ID EDAD SEXO  PESO TALLA SEDENTARIO   INSTRUCCION TAS TAD HTA_conocida
## 1  22   51    1  86.3   174          0  Primer grado 125  80            1
## 2  65   33    1  81.8   175          0 Segundo grado 120  80            0
## 3 116   40    1 105.0   173          1 Segundo grado 150 100            0
## 4 144   62    1  98.0   172          0  Primer grado 155 100            0
## 5 204   33    1 124.0   181          0 Segundo grado 130  90            0
## 6 337   33    1  93.0   177          1 Segundo grado 110  70            0
```

```
##   HTA_OMS A_DIAB ECV_B TABACO ALCOHOL STATIN CINTURA CADERA OBCENT_ATP
## 1       1      1     1      0       1      1     104  101.5          1
## 2       0      0     0      1       1      0      86   89.0          0
## 3       1      1     0      1       1      0     116  107.0          1
## 4       1      0     0      1       0      0     104  103.0          1
## 5       1      1     0      0       0      0     130  122.0          1
## 6       0      1     0      0       1      0     108  106.0          1
##   COLESTEROL HDL LDL     LDL_C  TG CnoHDL  ApoA ApoB   LPA    A1C hba1
## 1        190  50 123 120 - 140  86    140 138.1   78  4.78 6.1730  7.9
## 2        217  42 144 140 - 160 155    175  93.1   94  2.39 5.5820  7.2
## 3        288  44  NA      <NA> 448    244 122.7  120 39.30 5.3850  7.1
## 4        225  45 153 140 - 160 132    180 115.8  101  7.63 5.3850  6.6
## 5        224  58 154 140 - 160  62    166 135.2   86 24.90 5.1880  6.3
## 6        163  40 111 100 - 120  59    123  98.5   71  2.41 4.8925  6.0
##   CREATININA GLUCB SOG Tol_Glucosa DM conocida SM  PCR INSULINEMIA PAI_1
## 1        0.9   101 122         IFG  0   Normal  1 0.34        16.2  47.8
## 2        1.0   103 100         IFG  0   Normal  0 0.32        19.8  57.2
## 3        1.0   103  90         IFG  0   Normal  1 0.58        11.5  38.5
## 4        0.9   109  96         IFG  0   Normal  1 0.32        17.8  32.4
## 5        0.8   107  69         IFG  0   Normal  1 0.32        11.6  32.0
## 6        0.9   103 117         IFG  0   Normal  0 0.34        22.7  71.7
##     fvw fibri   HMC     HOMA IR ecnos  ppr fibratos CETP PON_192
## 1 115.0  2.40 14.00 4.036687  1    ab   pp       No <NA>      QR
## 2  17.8  2.89 40.72 5.031426  1    bb   pp       No <NA>      QQ
## 3  91.0  2.71 15.34 2.922293  1    bb   pp       No <NA>      QR
## 4  96.1  2.45 15.33 4.786689  1    bb   pp       No B1B1      RR
## 5  75.0  3.78  8.64 3.062178  1  <NA> <NA>       No <NA>    <NA>
## 6 117.0  3.53 13.20 5.768352  1    bb   pp       No <NA>      QR
```

Podemos ver su estructura con más detalle mediante la función `str()`, lo que nos permite comprobar si las variables numéricas se han leído como numéricas y si las variables tipo carácter se han leido como caracteres:

```
str(telde)
```

```
## 'data.frame':    1030 obs. of  50 variables:
##  $ ID          : num  22 65 116 144 204 337 458 541 551 557 ...
##  $ EDAD        : num  51 33 40 62 33 33 68 56 55 46 ...
##  $ SEXO        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ PESO        : num  86.3 81.8 105 98 124 93 81 99 92 94.3 ...
##  $ TALLA       : num  174 175 173 172 181 177 174 174 179 169 ...
##  $ SEDENTARIO  : num  0 0 1 0 0 1 0 1 1 0 ...
##  $ INSTRUCCION : chr  "Primer grado" "Segundo grado" "Segundo grado" "Primer grado" ...
```

```
##  $ TAS          : num  125 120 150 155 130 110 120 110 130 120 ...
##  $ TAD          : num  80 80 100 100 90 70 70 70 80 80 ...
##  $ HTA_conocida: num  1 0 0 0 0 0 0 0 0 0 ...
##  $ HTA_OMS      : num  1 0 1 1 1 0 0 0 0 0 ...
##  $ A_DIAB       : num  1 0 1 0 1 1 1 0 0 1 ...
##  $ ECV_B        : num  1 0 0 0 0 0 0 0 0 0 ...
##  $ TABACO       : num  0 1 1 1 0 0 0 0 0 1 ...
##  $ ALCOHOL      : num  1 1 1 0 0 1 0 1 1 1 ...
##  $ STATIN       : num  1 0 0 0 0 0 0 1 0 0 ...
##  $ CINTURA      : num  104 86 116 104 130 108 100 114 105 109 ...
##  $ CADERA       : num  102 89 107 103 122 ...
##  $ OBCENT_ATP   : num  1 0 1 1 1 1 0 1 1 1 ...
##  $ COLESTEROL   : num  190 217 288 225 224 163 217 306 283 218 ...
##  $ HDL          : num  50 42 44 45 58 40 49 53 62 44 ...
##  $ LDL          : num  123 144 NA 153 154 111 138 205 186 120 ...
##  $ LDL_C        : chr  "120 - 140" "140 - 160" NA "140 - 160" ...
##  $ TG           : num  86 155 448 132 62 59 149 239 175 271 ...
##  $ CnoHDL       : num  140 175 244 180 166 123 168 253 221 174 ...
##  $ ApoA         : num  138.1 93.1 122.7 115.8 135.2 ...
##  $ ApoB         : num  78 94 120 101 86 71 NA 150 130 99 ...
##  $ LPA          : num  4.78 2.39 39.3 7.63 24.9 2.41 15.3 7.31 36.7 28.9 ...
##  $ A1C          : num  6.17 5.58 5.38 5.38 5.19 ...
##  $ hba1         : num  7.9 7.2 7.1 6.6 6.3 6 8.1 7.7 6.8 7.3 ...
##  $ CREATININA   : num  0.9 1 1 0.9 0.8 0.9 1 1 1.2 1 ...
##  $ GLUCB        : num  101 103 103 109 107 103 104 101 115 101 ...
##  $ SOG          : num  122 100 90 96 69 117 166 94 162 143 ...
##  $ Tol_Glucosa : chr  "IFG" "IFG" "IFG" "IFG" ...
##  $ DM           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ conocida     : chr  "Normal" "Normal" "Normal" "Normal" ...
##  $ SM           : num  1 0 1 1 1 0 0 1 1 1 ...
##  $ PCR          : num  0.34 0.32 0.58 0.32 0.32 0.34 0.34 0.34 0.34 0.49 ...
##  $ INSULINEMIA : num  16.2 19.8 11.5 17.8 11.6 22.7 14.7 17.2 14.6 15.6 ...
##  $ PAI_1        : num  47.8 57.2 38.5 32.4 32 71.7 46.3 56.8 71.1 31.6 ...
##  $ fvw          : num  115 17.8 91 96.1 75 117 111 75.4 150 126 ...
##  $ fibri        : num  2.4 2.89 2.71 2.45 3.78 3.53 3.85 3.18 3.2 2.64 ...
##  $ HMC          : num  14 40.72 15.34 15.33 8.64 ...
##  $ HOMA         : num  4.04 5.03 2.92 4.79 3.06 ...
##  $ IR           : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ ecnos        : chr  "ab" "bb" "bb" "bb" ...
##  $ ppr          : chr  "pp" "pp" "pp" "pp" ...
##  $ fibratos     : chr  "No" "No" "No" "No" ...
##  $ CETP         : chr  NA NA NA "B1B1" ...
##  $ PON_192      : chr  "QR" "QQ" "QR" "RR" ...
```

2.3 Posibles problemas que podemos encontrar.

Con la lectura de archivos Excel

Hay dos problemas muy frecuentes con los que se encuentra en usuario novel cuando lee archivos Excel con R:

- `Error: no se pudo encontrar la función "read.xlsx"`: esto significa que queremos ejecutar la función `read.xlsx()` sin haber cargado aún el paquete `openxlsx`. Lo único que hay que hacer es ejecutar primero `library(openxlsx)` y volver a hacer `telde = read.xlsx("endocrino.xlsx")`.
- `Error in library(openxlsx) : there is no package called ???openxlsx???`: esto significa que no tenemos el paquete `openxlxs` instalado en nuestro ordenador. Lo único que habrá que hacer es instalarlo mediante `install.packages("openxlsx")`, y a continuación volver a ejecutar `library(openxlsx)`.

Con el establecimiento del directorio de trabajo.

Si nos aparece el mensaje:

`Error in setwd("ruta/directorio") :  no es posible cambiar el directorio de trabajo`

ello significa que el directorio o carpeta que hemos especificado no existe. En las aulas de informática de la ULPGC, cuando un archivo se descarga de internet lo hace automáticamente a la carpeta `c:\Users\aulas\Downloads`. En ordenadores Mac, la carpeta de descargas suele ser `/Users/nombreUsuario/Downloads` (nótese que en Mac no hay que especificar la unidad `c:` al principio de la ruta).

Si desconocemos la ubicación de la carpeta, podemos ir a la pestaña "Files" en la ventana inferior derecha de Rstudio y navegar por las carpetas de nuestro ordenador hasta localizar el archivo. A continuación pinchamos en "More" en la misma ventana, y elegimos la opción "Set As Working Directory"; como resultado, R se situará en dicha carpeta y se mostrará su ruta en la consola de Rstudio.

Con la no asignación de un nombre al objeto resultante de la lectura de la base de datos.

Si nos limitamos a escribir:

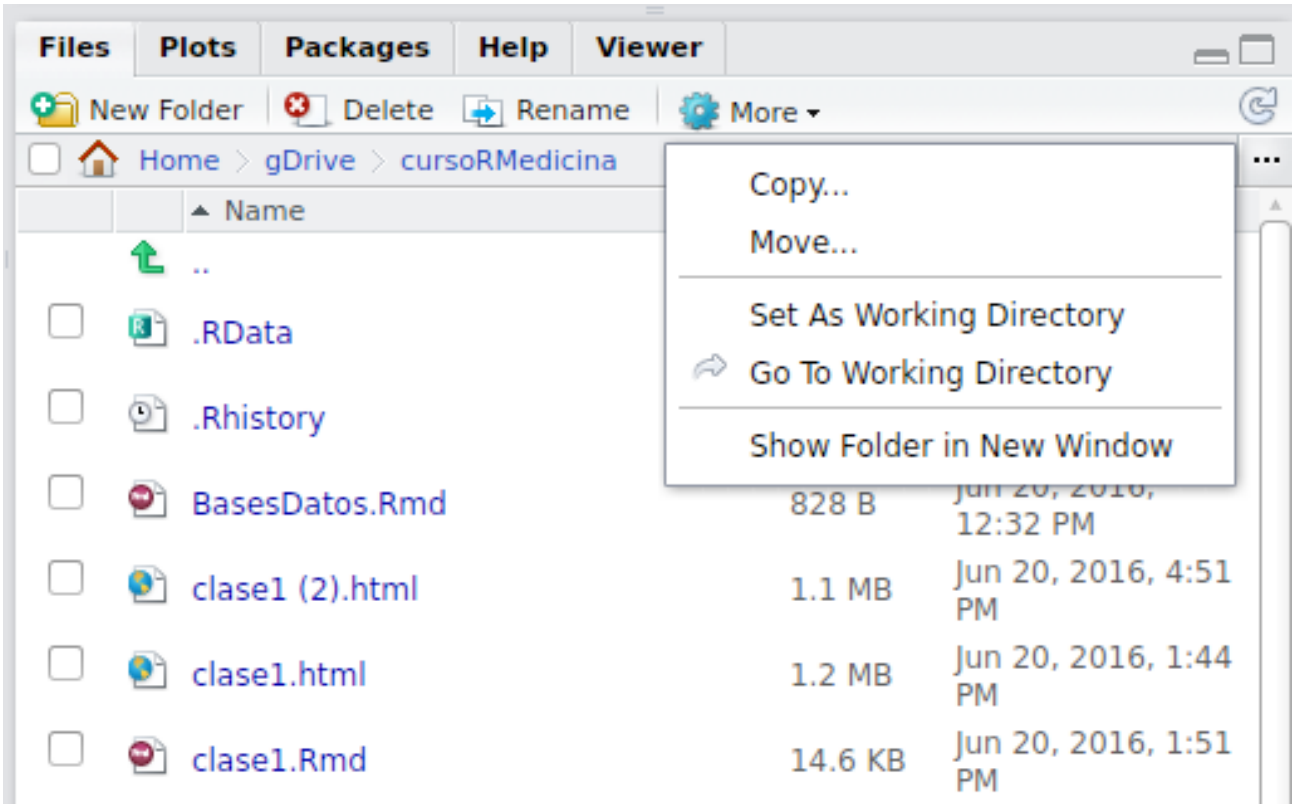


Figura 2:

```
read.xlsx("endocrino.xlsx")
```

el archivo se lee, y se muestra en pantalla, pero no se asigna a ningún `data.frame` por lo que no se conserva en la memoria del ordenador y R no puede acceder a él.

3 Descripción de datos

3.1 Introducción.

En esta sección veremos algunas de las herramientas disponibles en R para describir adecuadamente los datos objeto de nuestro estudio. En el caso de las variables categóricas la descripción más adecuada se consigue a través de tablas de frecuencias y diagramas de barras o sectores. En el caso de las variables continuas, las describiremos mediante medidas de tendencia central (media, mediana), dispersión (desviación típica), posición (percentiles), forma (asimetría, apuntamiento), y representaremos gráficamente su distribución a través de histogramas.

En primer lugar cargaremos los datos del estudio de Telde. Para ello deberemos ejecutar los tres comandos siguientes (téngase en cuenta que en `setwd()` debemos especificar la carpeta donde hemos guardado los datos, que puede ser diferente de la que figura aquí):

```
library(openxlsx)
setwd("c:/Users/aulas/Downloads/")
telde = read.xlsx("endocrino.xlsx")
```

3.2 Descripcion de datos: Tablas de frecuencias.

3.2.1 Tablas univariantes

Para construir una tabla de frecuencias absolutas utilizamos la funcion `table()`. Por ejemplo, para contar el número de diabéticos y no diabéticos en la muestra:

```
table(telde$DM)
```

```
##
##   0   1
## 902 128
```

Obsérvese que el nombre de la variable (`DM`) va precedido por el nombre del `data.frame` telde y el símbolo `$`. Ello permite utilizar simultáneamente varios `data.frames` que contengan información de las mismas variables (por ejemplo, podríamos tener un estudio similar cargado en otro `data.frame` llamado `arucas`, y accederíamos a la variable como `arucas$DM`).

De la misma forma, podemos contar el número de hombres y de mujeres:

```
table(telde$SEXO)
```

```
##
##   0   1
## 582 448
```

Esta tabla resulta poco informativa ya que a priori no sabemos quienes son los hombres y quienes las mujeres, ya que la variable `SEXO` toma los valores 0 y 1. Sabiendo que el 0 corresponde a los hombres y el 1 a las mujeres, podemos hacer la recodificación mediante:

```
telde$SEXO=factor(telde$SEXO,levels=c(0,1),labels=c("Hombre","Mujer"))
```

De esta forma, si repetimos la tabla, resultará mucho más clara:

```
table(telde$SEXO)
```

```
##
## Hombre  Mujer
##    582    448
```

Si en lugar de las frecuencias absolutas (número de hombres y mujeres, o número de diabéticos) queremos las frecuencias relativas (proporciones), utilizamos la función `prop.table()`:

```
prop.table(table(telde$DM))
```

```
##
##         0         1
## 0.8757282 0.1242718
```

```
prop.table(table(telde$SEXO))
```

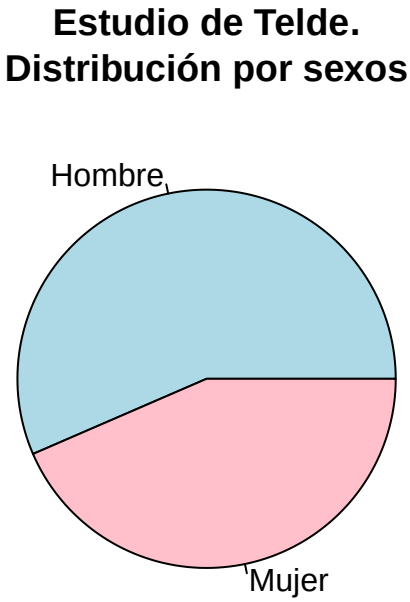
```
##
##    Hombre     Mujer
## 0.5650485 0.4349515
```

Podemos representar las tablas mediante diagramas de barras, en las que se pueden especificar colores, etiquetas para los ejes, títulos, etc:

```
barplot(table(telde$SEXO), col=c("lightblue","pink"),
        xlab="Sexo", ylab="Frecuencia",
        main="Estudio de Telde.\n Distribución por sexos.")
```

Otra representación habitual es mediante diagramas de sectores:

```
pie(table(telde$SEXO), col=c("lightblue","pink"),
        main="Estudio de Telde.\n Distribución por sexos.")
```

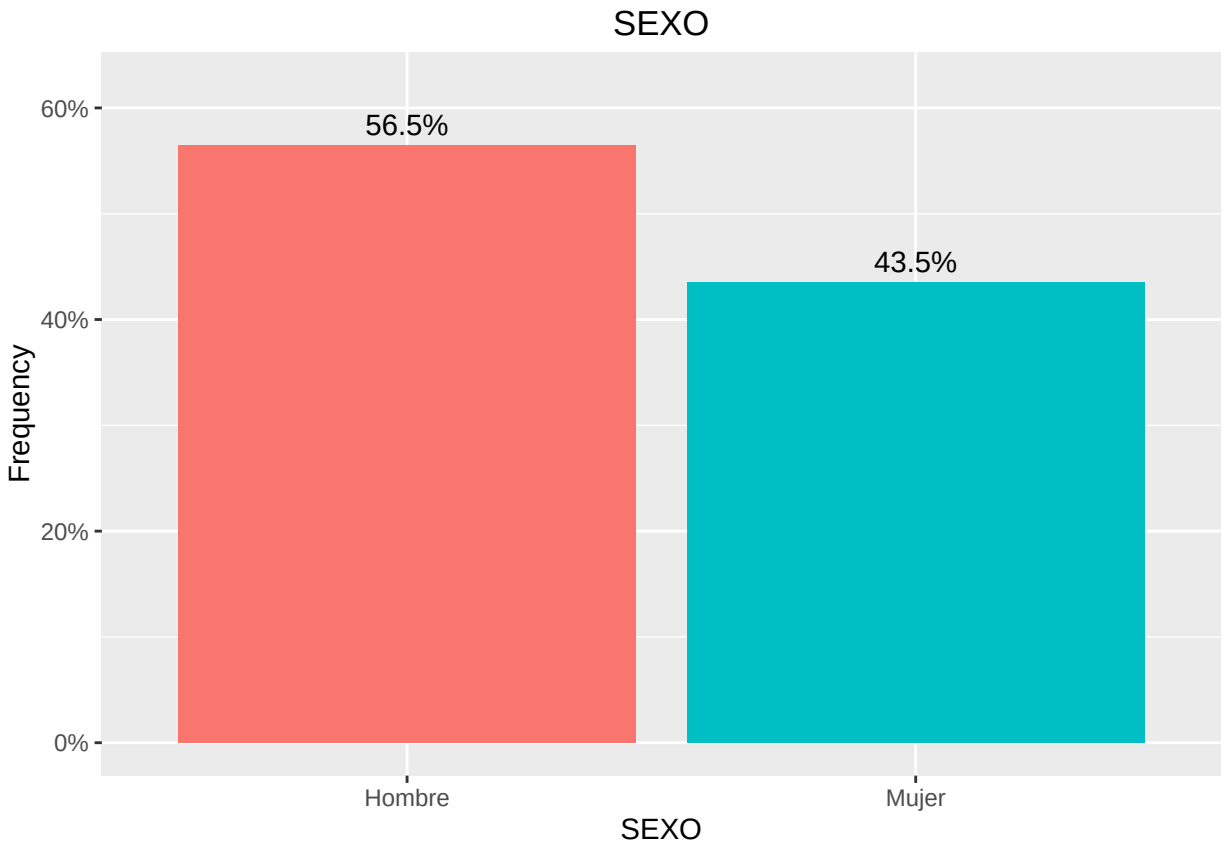


El paquete `ULPGCmisc` calcula tablas y gráficos conjuntamente con la función `freqTable()`:

```
library(ULPGCmisc) # Cargamos la librería ULPGCmisc
freqTable(telde$SEXO)
```

Tabla 1: Data are summarized in absolute frequencies and percentages, n(%)

Variable (levels)	All data (n=1030)
SEXO	
Hombre	582 (56.50)
Mujer	448 (43.50)



3.2.2 Tablas cruzadas

La función `table()` también permite calcular las frecuencias cruzadas de dos variables. Basta simplemente con indicar los dos nombres. Así, por ejemplo, para calcular el número de diabéticos y no diabéticos según el sexo podemos utilizar:

```
table(telde$SEXO,telde$DM)
```

```
##
##            0   1
##   Hombre 528  54
##   Mujer  374  74
```

o bien:

```
with(telde,table(SEXO,DM))
```

```
##         DM
## SEXO       0   1
##   Hombre 528  54
##   Mujer  374  74
```

Para calcular las frecuencias relativas puede resultar más cómodo asignar un nombre a la tabla cruzada:

```
tbSexDM <- with(telde,table(SEXO,DM))
```

Las frecuencias relativas pueden calcularse entonces:

- Para toda la tabla:

```
prop.table(tbSexDM)
```

```
##         DM
## SEXO              0          1
##   Hombre 0.51262136 0.05242718
##   Mujer  0.36310680 0.07184466
```

- Por filas: En la tabla cruzada, cada fila corresponde a uno de los dos sexos; si queremos calcular, para para cada sexo por separado, cuáles son las proporciones respectivas de diabéticos y no diabéticos ejecutamos la siguiente función:

```r
prop.table(tbSexDM,1)
```

```
##         DM
## SEXO               0          1
##   Hombre 0.90721649 0.09278351
##   Mujer  0.83482143 0.16517857
```

que nos indica que entre los hombres hay un 9.27% de diabéticos y entre las mujeres un 16.52% de diabéticas:

- Por columnas: La siguiente tabla nos muestra cuáles son las proporciones relativas de hombres y mujeres entre las personas diabéticas, así como entre las no diabéticas:

```r
prop.table(tbSexDM,2)
```

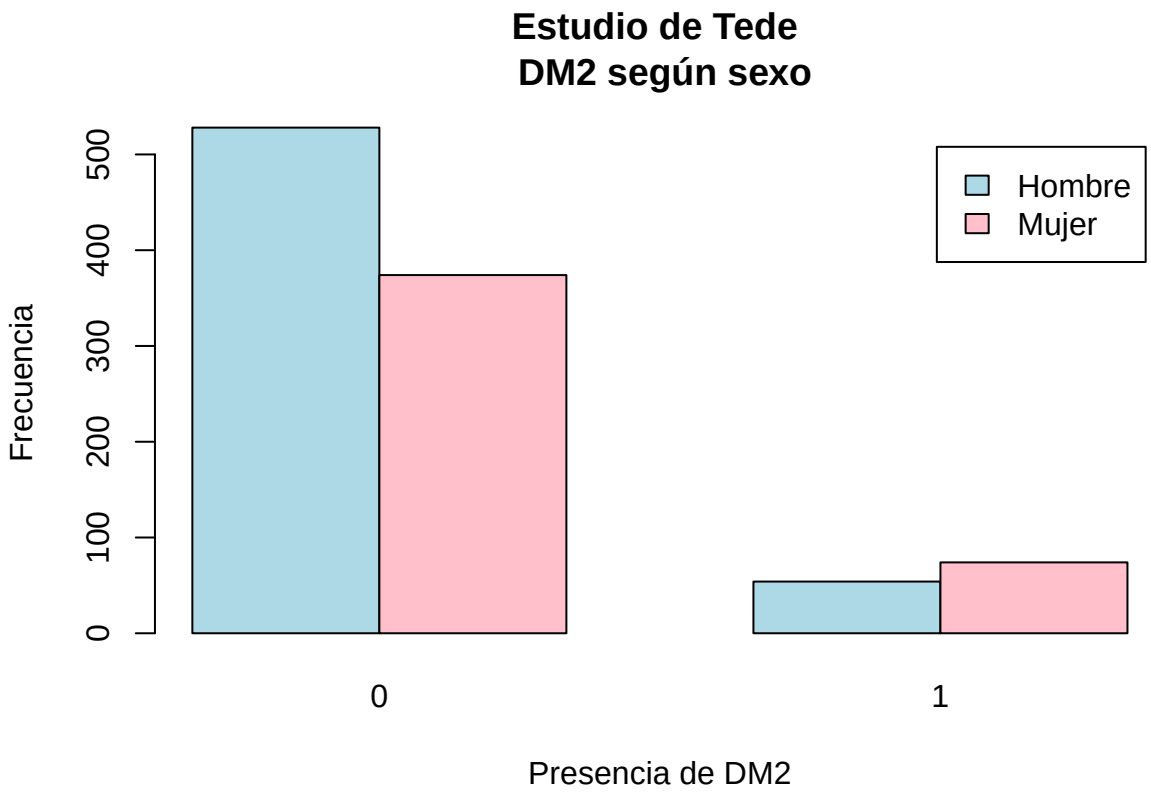
```
##         DM
## SEXO              0         1
##   Hombre 0.5853659 0.4218750
##   Mujer  0.4146341 0.5781250
```

Esta tabla nos indica que entre los diabéticos el 42.19% son hombres y el 57.81% mujeres; asimismo, entre los no diabéticos el 58.54% son hombres y el restante 41,46% mujeres.

Podemos representar gráficamente estas tablas mediante diagramas de barras:

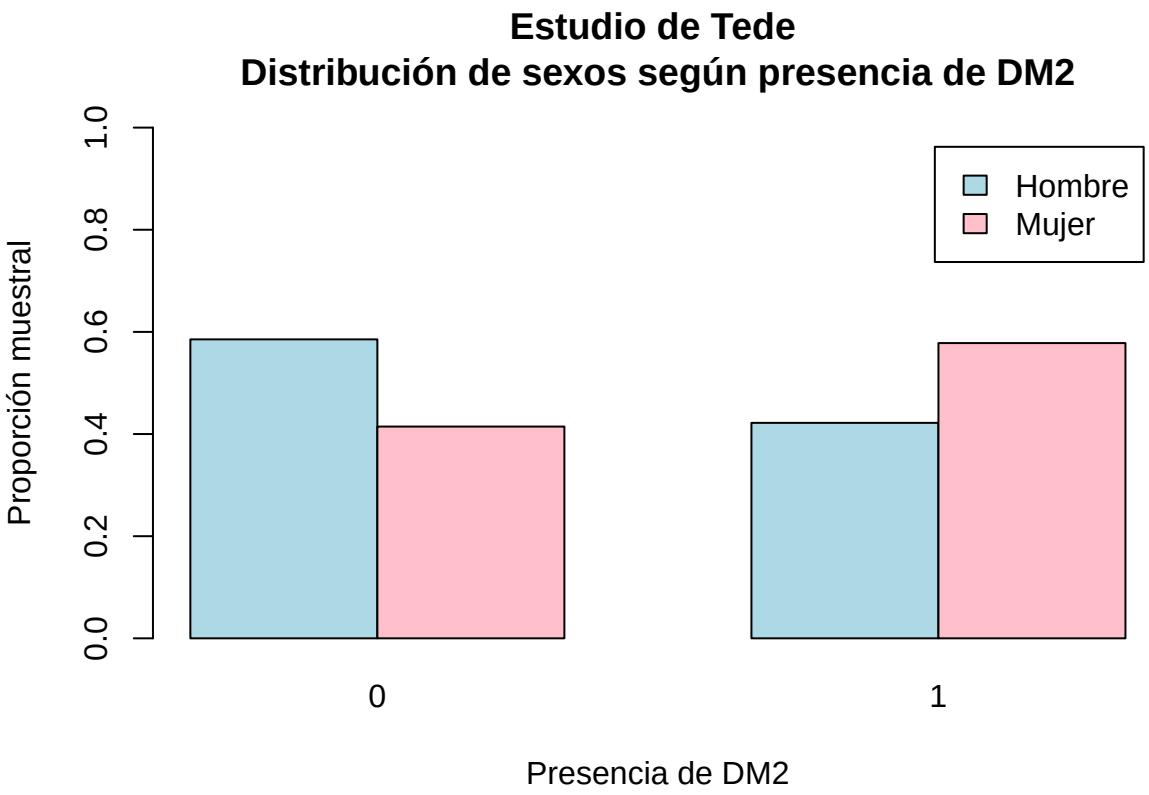
- Frecuencias absolutas totales:

```r
barplot(with(telde,table(SEXO,DM)),beside=TRUE,legend=TRUE,
        xlab="Presencia de DM2", ylab="Frecuencia",
        col=c("lightblue","pink"), main="Estudio de Tede \n DM2 según sexo")
```



- Frecuencias relativas por columnas:

```
tb2=with(telde,prop.table(table(SEXO,DM),2))
barplot(tb2,beside=TRUE,legend=TRUE,ylim=c(0,1),
        xlab="Presencia de DM2", ylab="Proporción muestral",
        col=c("lightblue","pink"),
        main="Estudio de Tede \nDistribución de sexos según presencia de DM2")
```

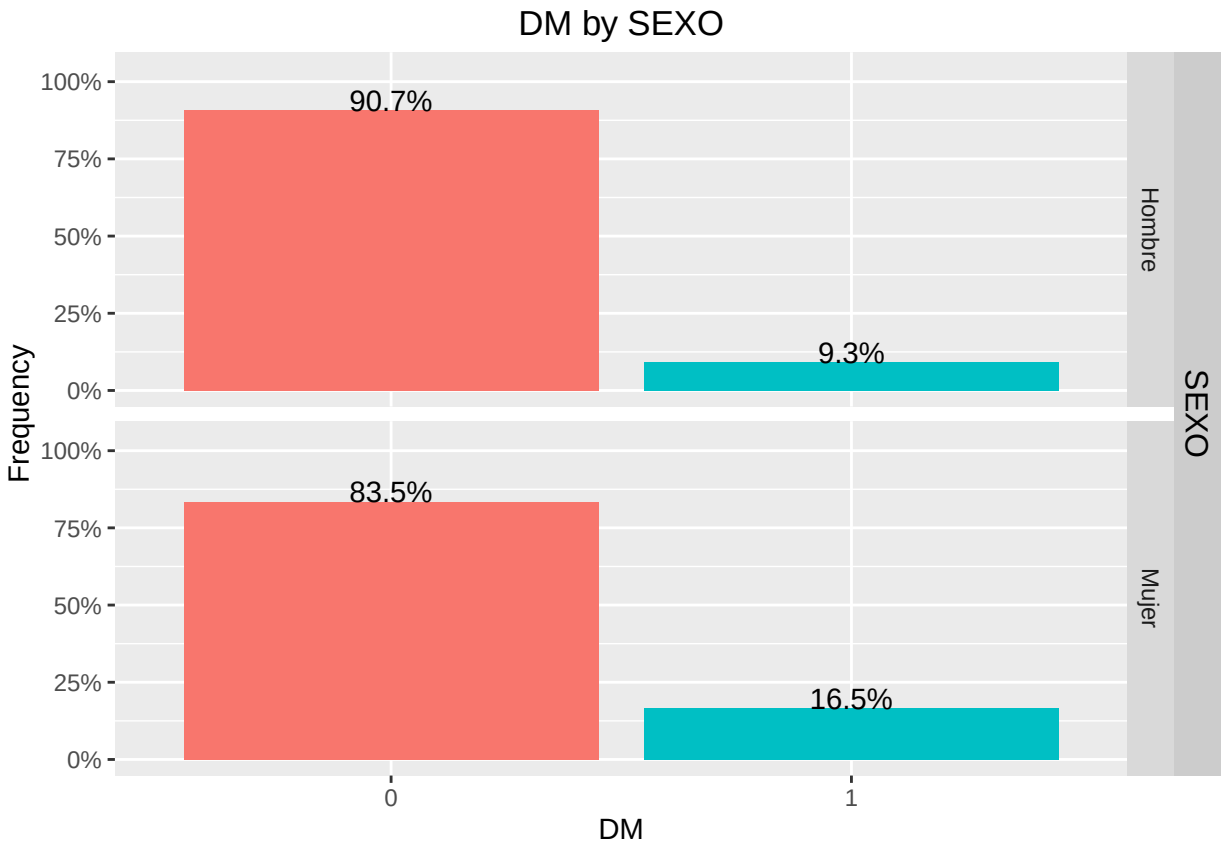


La función `freqTable()` del paquete `ULPGCmisc` construye tablas y gráficos conjuntamente:

```
with(telde, freqTable(DM, by=SEXO))
```

Tabla 2: Data are summarized in absolute frequencies and percentages, n(%)

Variable (levels)	All data (n=1030)	SEXO = Hombre (n=582)	SEXO = Mujer (n=448)	Chi-Squared test P
DM				0.0007
0	902 (87.57)	528 (90.72)	374 (83.48)	
1	128 (12.43)	54 (9.28)	74 (16.52)	



3.3 Ordenación de valores en variables categóricas.

Muchas veces nos encontramos con variables categóricas cuyos valores son ya ordinales (van de menor a mayor) o nos interesa que sus valores aparezcan siempre en determinado orden. Por ejemplo, en las tablas y gráficos de la presencia/ausencia de DM2 nos interesa que la categoría "enfermedad presente" (codificada como 1) aparezca en primer lugar y la categoría "enfermedad ausente" (codificada como 0) aparezca en segundo lugar. Podemos conseguir este objetivo redefiniendo la variable como:

```
telde$DM2=ordered(telde$DM,levels=c(1,0),c("Sí","No"))
with(telde,table(DM2))
```

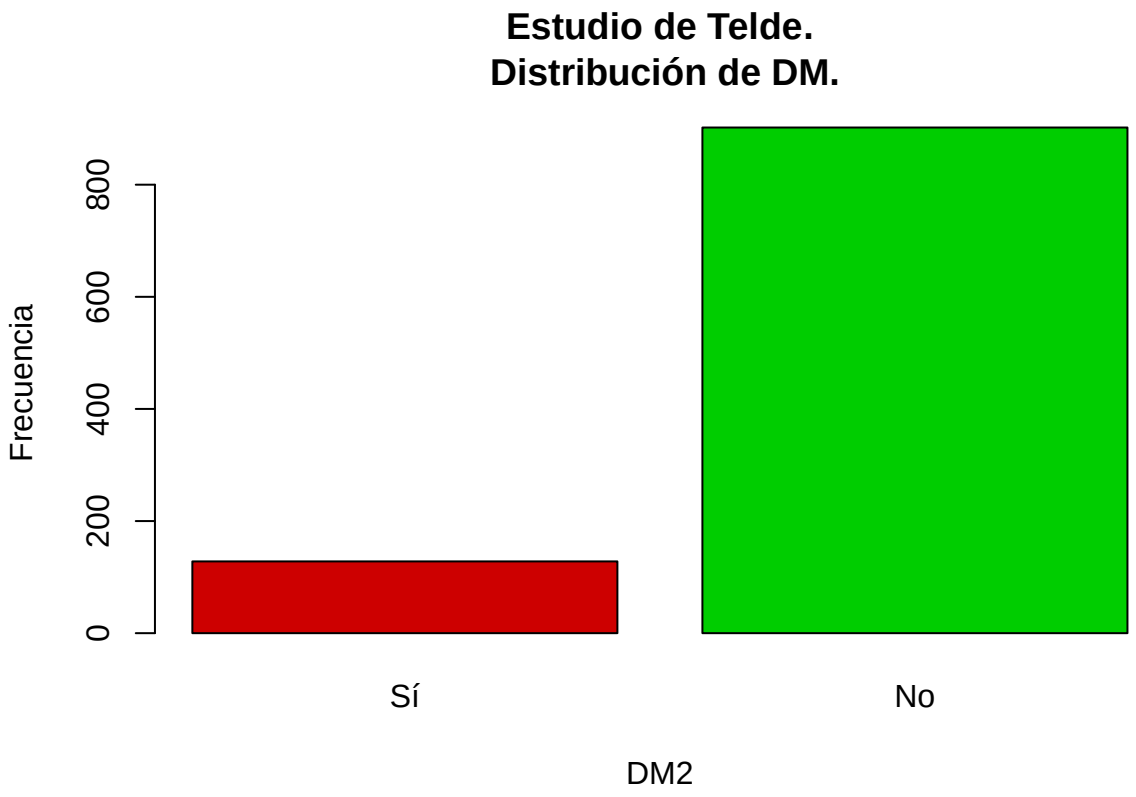
```
## DM2
##  Sí  No
## 128 902
```

```
with(telde,table(SEXO,DM2))
```

```
##         DM2
```

```
## SEXO      Sí  No
##   Hombre  54 528
##   Mujer   74 374
```

```r
barplot(table(telde$DM2), col=c("red3","green3"),
        xlab="DM2", ylab="Frecuencia",
        main="Estudio de Telde.\n Distribución de DM.")
```



3.3.1 Ejercicios

1. Cuántas personas con HTA hay en el estudio de Telde. ¿Cuál es la frecuencia relativa de hipertensos y no hipertensos?
2. ¿Cuántas personas tienen HTA y DM? ¿Cuántas no son HTA ni DM? Construye la tabla cruzada de ambas variables.
3. ¿Cuál es la proporción de personas hipertensas entre las que tienen DM?
4. ¿Cuál es la proporción de personas diabéticas entre las que tienen HTA?

4 Tasas ajustadas por edad

4.1 Introducción

A partir de los datos del estudio de Telde es posible estimar la prevalencia (cruda) de diabetes en la población adulta de dicha ciudad (las personas seleccionadas para el estudio son todas adultas con 30 o más años). Para ello contamos cuantos sujetos diabéticos y no diabéticos hay en la muestra de Telde, y calculamos la proporción que representa cada grupo respecto al total:

```
library(openxlsx)
telde=read.xlsx("endocrino.xlsx")
telde$DM=ordered(telde$DM, levels=c(1,0),labels=c("DM+","DM-"))
table(telde$DM)
```

```
##
## DM+ DM-
## 128 902
```

```
pDM=prop.table(table(telde$DM))
pDM
```

```
##
##       DM+       DM-
## 0.1242718 0.8757282
```

Este resultado nos indica que la prevalencia de DM en la muestra de Telde es 0.1242718, esto es, aproximadamente el 12,43%.

Ahora bien, si queremos comparar la prevalencia de esta enfermedad en Telde con la prevalencia en otra ciudad es preciso tomar ciertas precauciones. En concreto, se sabe que la DM se asocia con la edad. Por tanto, si en una ciudad arbitraria hay mayor prevalencia de DM que en Telde bien podría ser porque la población de Telde fuese más joven que la de la ciudad con la que se compara. Para que la comparación entre ciudades tuviese sentido deberían compararse ciudades con la misma composición por edades. Este es el objetivo del cálculo de la prevalencia ajustada por edad. La idea clave consiste entonces en calcular

cuál sería la prevalencia de la DM en Telde si la población de Telde tuviese una estructura de edades acorde a la de alguna población de referencia.

En esta tarea utilizaremos como referencia la población SEGI. En este estándar las edades de los individuos se agrupan en los siguientes grupos de edad:

- De 30 a 39 años
- De 40 a 49 años
- De 50 a 59 años
- De 60 a 69 años
- 70 o más años

siendo las proporciones respectivas de miembros de la población en cada clase las siguientes:

```
p.SEGI=c(0.2727,0.2727,0.2045,0.1591,0.0910)
names(p.SEGI)=c("30-39","40-49","50-59","60-69","70 o más")
p.SEGI
```

```
##    30-39    40-49    50-59    60-69 70 o más
##   0.2727   0.2727   0.2045   0.1591   0.0910
```

4.2 Cálculo de la prevalencia ajustada por edad.

En Telde (datos del censo del INE) el número total de personas dentro de cada uno de estos grupos de edad es:

```
n.TELDE=c(16739, 11878, 8445, 6085, 4918)
names(n.TELDE)=c("30-39","40-49","50-59","60-69","70 o más")
```

La proporción dentro de cada clase puede calcularse entonces como:

```
p.TELDE=prop.table(n.TELDE)
p.TELDE
```

```
##     30-39     40-49     50-59     60-69  70 o más
## 0.3482576 0.2471237 0.1756996 0.1265994 0.1023198
```

A continuación creamos una variable que a cada sujeto asigne su grupo de edad dentro de la muestra de 1030 personas de Telde:

```
grEdad=cut(telde$EDAD,breaks=c(30,40,50,60,70,100),right=FALSE)
```

Contamos cuantos sujetos hay en cada grupo de edad en la muestra de Telde, y su proporción:

```
n.Muestra <- table(grEdad)
n.Muestra
```

```
## grEdad
##  [30,40)  [40,50)  [50,60)  [60,70) [70,100)
##      311      288      236      142       53
```

```
p.Muestra <- prop.table(table(grEdad))
p.Muestra
```

```
## grEdad
##    [30,40)    [40,50)    [50,60)    [60,70)   [70,100)
## 0.30194175 0.27961165 0.22912621 0.13786408 0.05145631
```

Por último contamos cuantos diabéticos hay en cada grupo de edad y qué proporción representan los diabéticos en cada grupo:

```
table(telde$DM,grEdad)
```

```
##      grEdad
##       [30,40) [40,50) [50,60) [60,70) [70,100)
##   DM+       8      17      39      48       16
##   DM-     303     271     197      94       37
```

```
pdg <- prop.table(table(telde$DM,grEdad),2); # Proporción por columnas
pdg
```

```
##      grEdad
##          [30,40)    [40,50)    [50,60)    [60,70)   [70,100)
##   DM+ 0.02572347 0.05902778 0.16525424 0.33802817 0.30188679
##   DM- 0.97427653 0.94097222 0.83474576 0.66197183 0.69811321
```

La prevalencia de diabetes en cada grupo de edad es la primera fila de esta tabla:

```
preval.DMxEdad <- pdg[1,]
preval.DMxEdad
```

```
##     [30,40)     [40,50)     [50,60)     [60,70)    [70,100)
## 0.02572347 0.05902778 0.16525424 0.33802817 0.30188679
```

Podemos presentar todos los datos anteriores en una única tabla:

```
library(pander)
tablaDM=data.frame(n.TELDE,
                   p.TELDE,
                   n.Muestra=as.numeric(n.Muestra),
                   p.Muestra=as.numeric(p.Muestra),
                   preval.DMxEdad=as.numeric(preval.DMxEdad),
                   p.SEGI=p.SEGI)
rownames(tablaDM)=levels(grEdad)
pander(tablaDM)
```

	n.TELDE	p.TELDE	n.Muestra	p.Muestra	preval.DMxEdad	p.SEGI
[30,40)	16739	0.3483	311	0.3019	0.02572	0.2727
[40,50)	11878	0.2471	288	0.2796	0.05903	0.2727
[50,60)	8445	0.1757	236	0.2291	0.1653	0.2045
[60,70)	6085	0.1266	142	0.1379	0.338	0.1591
[70,100)	4918	0.1023	53	0.05146	0.3019	0.091

La prevalencia total se calcula como la suma de los productos de las prevalencias en cada grupo de edad por la proporción total de individuos en ese grupo:

$$P\left(D\right) = \sum_{i=1}^{5} P\left(D\,|E_i\right) P\left(E_i\right)$$

En esta ecuación:

- Los E_i son los 5 grupos de edad:

  E_1 = "30-39", E_2 = "40-49", E_3 = "50-59", E_4 = "60-69" y E_5 = " 70";
- Las $P\left(D\,|E_i\right)$ son las prevalencias de la enfermedad dentro de cada grupo de edad, que corresponden a la columna `preval.DMxEDAD` en la tabla anterior.
- Las $P\left(E_i\right)$ son las probabilidades de pertenencia a cada uno de estos grupos.

Así pues, la prevalencia total en la muestra se obtendría eligiendo como $p(E_i)$ las proporciones que representan los distintos grupos de edad E_i en la muestra, `p.Muestra`:

```
sum(preval.DMxEdad*p.Muestra)
```

```
## [1] 0.1242718
```

que, como vemos, coincide con la prevalencia muestral ya calculada en la introducción de este documento.

La prevalencia ajustada a la población total de Telde se calcularía eligiendo ahora como $p(E_i)$ las proporciones de los grupos de edad E_i en la población total de Telde, `p.TELDE`:

```
sum(preval.DMxEdad*p.TELDE)
```

```
## [1] 0.1262638
```

y por último la prevalencia ajustada a la población SEGI se obtiene eligiendo como $p(E_i)$ las proporciones de los grupos de edad E_i en la población de referencia SEGI, `p.SEGI`:

```
sum(preval.DMxEdad*p.SEGI)
```

```
## [1] 0.1381581
```

4.3 Ejercicio

Ajustar la prevalencia de DM en Telde a la población de referencia (WHO World Standard) de la OMS publicada en http://www.who.int/healthinfo/paper31.pdf (página 12) y que se reproduce a continuación:

```
edades=c(paste(seq(0,80,by=5),seq(4,84,by=5),sep="-"), "85+")
WWS=c(8.85,8.68,8.60,8.47,8.22,7.93,7.61,7.15,6.59,6.04,5.37,4.55,3.72,2.96,2.21,1.52,0.91,0.62)
names(WWS)=edades
library(pander)
pander(WWS)
```

Tabla 4: Table continues below

0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
8.85	8.68	8.6	8.47	8.22	7.93	7.61	7.15	6.59	6.04

50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+
5.37	4.55	3.72	2.96	2.21	1.52	0.91	0.62

Téngase en cuenta:

a. En la WHO World Standard los intervalos de edad están de 5 en 5 años, por lo que habrán de recalcularse nuevos intervalos para las edades en Telde.

b. Estamos interesados sólo en la población adulta (30 años o más), por lo que en las proporciones de la WHO World Standard habrán de eliminarse las correspondientes a edades menores de 30 años y reajustando el resto para que sume 1.

5 Riesgo Relativo y Odds Ratio

5.0.1 A Case-Control Study of Factors Associated with HIV Infection among Black Women

Data collection. Trained interviewers used standardized questionnaires to collect epidemiologic, behavioral, socio-economic and demographic data for the 12-month period preceding either the date of diagnosis for the cases and for their male partners, or the date of interview for the controls.

Los datos disponibles se muestran en la siguiente tabla:

	HIV-Possitive N = 31	HIV-Negative N = 101	P	OR (95%CI)
Unemployed, n (%)	22 (71)	38 (38)	< 0.01	4.0 (1.7 ; 10.0)
Receiving Public Assistance, n (%)	24 (77)	52 (51)	0.01	3.2 (1.3 ; 8.2)
Cocaine/crack, n (%)	5 (16)	5 (5)	0.05	3.7 (1.0 ; 13.7)
Receipt of money, drugs, gifts or shelter for sex, n (%)	11 (36)	15 (15)	0.01	3.2 (1.3 ; 7.9)
Male Partner Incarceration, n (%)	25 (81)	60 (59)	0.04	2.8 (1.1 ; 7.6)

Figura 3:

A partir de estos datos calcularemos el riesgo relativo (RR) de padecer HIV según estén presentes o no los distintos factores de riesgo considerados. Para ello tengamos en cuenta que el RR de HIV para el factor T se define como:

$$RR = \frac{P\left(HIV+|T+\right)}{P\left(HIV+|T-\right)}$$

donde $HIV+$ significa que el sujeto padece la enfermedad, $T+$ significa que el posible factor de riesgo está presente y $T-$ que está ausente.

Desempleo

- Cuando el factor de riesgo es el desempleo (Unemployed) la tabla nos indica que de las 31 mujeres con HIV, 22 estaban desempleadas, de donde se sigue que las otras 9 tenían empleo; asimismo de las 101 sin HIV, 38 estaban desempleadas, lo que significa que las restantes 101-38=63 tenían empleo. Podemos construir una tabla cruzada con estos datos del siguiente modo:

```
tabla1=array(c(22,9,38,63),dim=c(2,2))
colnames(tabla1)=c("HIV+","HIV-")
rownames(tabla1)=c("T+","T-")
tabla1
```

```
##    HIV+ HIV-
## T+   22   38
## T-    9   63
```

Puede mejorarse la presentación con `pander`:

```
library(pander)
pander(tabla1)
```

	HIV+	HIV-
T+	22	38
T-	9	63

Si calculamos las proporciones por columnas:

```
pc <- prop.table(tabla1,2)
pander(round(pc,2))  # Redondeamos a dos dígitos decimales
```

	HIV+	HIV-
T+	0.71	0.38
T-	0.29	0.62

confirmamos que, efectivamente el 71% de las mujeres HIV+ estaban desempleadas, y también lo estaban el 38% de las HIV-, tal como figura en la tabla original. Ahora bien, para calcular el riesgo relativo necesitamos la probabilidad $P\left(HIV+|T+\right)$, es decir, la probabilidad de padecer HIV cuando se está en desempleo, y $P\left(HIV+|T-\right)$, la probabilidad de padecer HIV cuando NO se está en desempleo. Por tanto necesitamos las probabilidades por filas de la tabla anterior:

```r
pf <- prop.table(tabla1,1)
pander(round(pf,4))
```

	HIV+	HIV-
T+	0.3667	0.6333
T-	0.125	0.875

El riesgo relativo es, por tanto:

$$RR = \frac{P\left(HIV+|T+\right)}{P\left(HIV+|T-\right)} = \frac{0.3667}{0.125}$$

que puede calcularse directamente con R mediante:

```r
RR <- pf[1,1]/pf[2,1]
RR
```

```
## [1] 2.933333
```

Por tanto, el riesgo de padecer HIV es casi el triple entre las mujeres negras desempleadas que entre las que tienen empleo.

5.0.2 ¿Es correcto calcular el Riesgo Relativo en este contexto (estudio caso-control)?

Para responder a esta pregunta observemos lo siguiente: en un estudio de casos y controles es el investigador el que decide cuántos casos y cuántos controles se seleccionan. Supongamos que hubiésemos decidido (o hubiésemos podido) tomar 310 casos (mujeres con HIV) en lugar de los 31 que tenemos, y 303 controles (en lugar de los 101 que tenemos). En otras palabras, multiplicamos por 10 el tamaño de la muestra de casos y por 3 el tamaño de la muestra de controles. Supongamos además que las proporciones de mujeres empleadas y desempleadas siguen siendo las mismas en ambos grupos. En tal caso, los datos disponibles serían:

```r
tabla2=array(c(220,90,114,189),dim=c(2,2))
colnames(tabla2)=c("HIV+","HIV-")
rownames(tabla2)=c("T+","T-")
pander(tabla2)
```

	HIV+	HIV-
T+	220	114
T-	90	189

Podemos comprobar que las proporciones por columnas coinciden con las mostradas en la tabla anterior:

```
pc2 <- prop.table(tabla2,2)
pander(round(pc2,2))
```

	HIV+	HIV-
T+	0.71	0.38
T-	0.29	0.62

Pero ahora las proporciones por filas son distintas:

```
pf2 <- prop.table(tabla2,1)
pander(round(pf2,4))
```

	HIV+	HIV-
T+	0.6587	0.3413
T-	0.3226	0.6774

y también lo es el riesgo relativo:

```
RR <- pf2[1,1]/pf2[2,1]
RR
```

```
## [1] 2.041916
```

que ha disminuído a 2.04. Por tanto el riesgo relativo ha disminuido en una tercera parte (de casi 3 a 2 y poco) simplemente por haber cambiado el número de casos y controles elegidos, aún cuando las prevalencias del factor de riesgo no han cambiado en cada grupo. Así pues, en los estudios de casos y controles, el riesgo relativo depende de los tamaños de muestra elegidos arbitrariamente por el investigador, y no es por tanto una característica intrínseca de la asociación entre la enfermedad y el factor de riesgo. Por tanto no puede ser un indicador útil del valor de dicha asociación, y no puede usarse con el objetivo de medir dicho valor.

5.0.3 ¿Es el riesgo relativo realmente una medida de asociación?

Señalemos además que el riesgo relativo no es propiamente una medida de asociación en sentido estricto. De una medida de asociación se espera que la asociación entre A y B sea la misma que entre B y A, lo que no ocurre con el riesgo relativo. Podemos comprobar esta afirmación de manera muy simple: hemos visto que en el estudio original, el riesgo relativo de padecer HIV (A) según que la mujer esté o no desempleada (B) es 2.9333; para calcular ahora el riesgo de que una mujer esté desempleada (B) según que tenga o no HIV (A), deberíamos partir de la tabla traspuesta de la original:

```
tabla3 <- t(tabla1)
tabla3
```

```
##      T+ T-
## HIV+ 22  9
## HIV- 38 63
```

y procediendo igual que antes, calcular las proporciones por filas, y a partir de ellas el RR:

```
pf3 <- prop.table(tabla3,1)
pander(round(pf3,4))
```

	T+	T-
HIV+	0.7097	0.2903
HIV-	0.3762	0.6238

```
RR <- pf3[1,1]/pf3[2,1]
RR
```

[1] 1.886248

Como vemos este riesgo no coincide con el calculado inicialmente: el riesgo de HIV condicionado por el desempleo (asociación de A con B) no coindice con el riesgo de desempleo condicionado por el HIV (asociación de B con A), lo que confirma que el riesgo relativo no es simétrico

5.1 Odds ratio

La odds-ratio que mide la asociación entre dos eventos A y B se define como:

$$OR(A,B) = \frac{P(A \mid B)/P(A^c \mid B)}{P(A \mid B^c)/P(A^c \mid B^c)}$$

Puede comprobarse que la odds-ratio sí que es una medida simétrica y que:

$$OR(A,B) = OR(B,A) = \frac{P(B \mid A)/P(B^c \mid A)}{P(B \mid A^c)/P(B^c \mid A^c)}$$

A modo de ejemplo, si volvemos a la relación entre HIV y desempleo del ejemplo anterior, si consideramos que el suceso A es el HIV+, entonces el contrario A^c es el HIV-. Asimismo, si el suceso B es estar desempleada ($T+$), el B^c sería tener empleo ($T-$). Entonces, podríamos calcular la OR mediante:

$$OR(HIV,T) = \frac{P(HIV+ \mid T+)/P(HIV- \mid T+)}{P(HIV+ \mid T-)/P(HIV- \mid T-)}$$

Para obtener este valor de OR partimos nuevamente de la tabla:

```
pander(tabla1)
```

	HIV+	HIV-
T+	22	38
T-	9	63

y de sus proporciones por fila:

```
pf=prop.table(tabla1,1)
pander(round(pf,4))
```

	HIV+	HIV-
T+	0.3667	0.6333
T-	0.125	0.875

Tenemos entonces:

$$OR = \frac{0.3667/0.6333}{0.125/0.875} = \frac{0.5789474}{0.1428571} = 4.0526$$

que en R se obtiene como:

```
(pf[1,1]/pf[1,2]) / (pf[2,1]/pf[2,2])
```

```
## [1] 4.052632
```

Si queremos calcular:

$$OR\,(T, HIV) = \frac{P(T+ \mid HIV+)/P(T- \mid HIV+)}{P(T+ \mid HIV-)/P(T- \mid HIV-)}$$

volvemos a partir de la tabla:

```
pander(tabla1)
```

	HIV+	HIV-
T+	22	38
T-	9	63

pero ahora habremos de calcular las proporciones por columnas:

```
pc=prop.table(tabla1,2)
pander(round(pc,4))
```

	HIV+	HIV-
T+	0.7097	0.3762
T-	0.2903	0.6238

Tenemos entonces:

$$OR = \frac{0.7097/0.0.2903}{0.3762/0.6238} = \frac{2.4444444}{0.6031746} = 4.0526$$

que en R se obtiene como:

```
(pc[1,1]/pc[1,2]) / (pc[2,1]/pc[2,2])
```

```
## [1] 4.052632
```

Así pues, este ejemplo confirma la simetría de la odds-ratio:

$$OR\,(HIV, T) = OR\,(T, HIV) = 4.0526316$$

Señalemos por último que la odds-ratio puede obtenerse también a partir de la tabla cruzada de frecuencias absolutas:

	B+	B-
A+	n[1,1]	n[1,2]
A-	n[2,1]	n[2,2]

sin necesidad de calcular proporciones por filas ni por columnas, mediante:

$$OR = \frac{n[1,1] \cdot n[2,2]}{n[1,2] \cdot n[2,1]}$$

NOTA: esto sólo es verdad si en la tabla la categoría positiva es la primera tanto en filas como en columnas.

Si calculamos la OR de esta manera en la `tabla1` anterior:

```
tabla1
```

```
##    HIV+ HIV-
## T+   22   38
## T-    9   63
```

```
(tabla1[1,1]*tabla1[2,2])/(tabla1[1,2]*tabla1[2,1])
```

```
## [1] 4.052632
```

volvemos a obtener el mismo resultado.

5.2 Cálculo de la Odds-Ratio con el paquete `epiR`

El paquete `epiR` contiene algunas funciones que permiten obtener fácilmente la OR y el RR a partir de una tabla cruzada. En primer lugar cargamos la librería:

```
library(epiR)
```

y podemos obtener inmediatamente la odds-ratio de HIV frente a desempleo mediante la función `epi.2by2` aplicada a la `tabla1`:

```
pander(tabla1)
```

	HIV+	HIV-
T+	22	38
T-	9	63

```
epi.2by2(tabla1)
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +           22           38         60              36.7
## Exposed -            9           63         72              12.5
## Total               31          101        132              23.5
##                   Odds
## Exposed +        0.579
## Exposed -        0.143
## Total            0.307
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 2.93 (1.46, 5.88)
## Odds ratio                                     4.05 (1.69, 9.71)
## Attrib risk *                                  24.17 (9.78, 38.56)
## Attrib risk in population *                    10.98 (0.47, 21.50)
## Attrib fraction in exposed (%)                 65.91 (31.64, 83.00)
## Attrib fraction in population (%)              46.77 (12.12, 67.76)
## -------------------------------------------------------------------
##  X2 test statistic: 10.637 p-value: 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
```

Notas importantes:

- Para que el cálculo sea correcto la tabla de partida (`tabla1` en este ejemplo) debe estar ordenada de forma que las categorías "positivas" (tener la enfermedad, tener el factor de riesgo, …) sean las primeras en cada variable.
- La función `epi.2by2`, por defecto devuelve también el riesgo relativo. Es responsabilidad del investigador tener en cuenta que el RR en este problema (diseño caso-control), carece de sentido.

5.3 Cálculo de la odds-ratio y riesgo relativo para variables incluídas en bases de datos (`data.frames`)

Cuando se desea calcular la odds-ratio entre dos variables que forman parte de un `data.frame` lo único que habremos de hacer es construir la tabla cruzada para dichas variables, con la precaución de que los valores positivos (presencia de la enfermedad, presencia del factor de riesgo, presencia de la exposición) deben situarse en primer lugar en las filas y columnas de la tabla, y los valores negativos (ausencia de enfermedad, factor de riesgo o exposición) en segundo lugar. Para ello utilizaremos la función `ordered`.

A modo de ejemplo, calculemos la odds-ratio entre diabetes e hipertensión para los datos de Telde. En primer lugar leemos la base de datos:

```
library(openxlsx)
telde=read.xlsx("endocrino.xlsx")
```

Ordenamos los valores de DM y HTA:

```
telde$DM <- ordered(telde$DM, levels=c(1,0), labels=c("DM+","DM-"))
telde$HTA_OMS <- ordered(telde$HTA_OMS, levels=c(1,0), labels=c("HTA+","HTA-"))
```

Construimos la tabla cruzada:

```
tadh <- with(telde, table(DM, HTA_OMS))
tadh
```

```
##      HTA_OMS
## DM    HTA+ HTA-
##   DM+   83   45
##   DM-  241  661
```

Y por último aplicamos la función `epi.2by2` a dicha tabla para calcular la odds-ratio:

```
epi.2by2(tadh)
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +           83           45        128              64.8
## Exposed -          241          661        902              26.7
## Total              324          706       1030              31.5
##                   Odds
## Exposed +        1.844
## Exposed -        0.365
## Total            0.459
##
## Point estimates and 95 % CIs:
```

```
## ------------------------------------------------------------------
## Inc risk ratio                                  2.43 (2.05, 2.87)
## Odds ratio                                      5.06 (3.42, 7.48)
## Attrib risk *                                   38.13 (29.36, 46.89)
## Attrib risk in population *                     4.74 (0.69, 8.79)
## Attrib fraction in exposed (%)                  58.80 (51.30, 65.14)
## Attrib fraction in population (%)               15.06 (10.77, 19.15)
## ------------------------------------------------------------------
##  X2 test statistic: 75.567 p-value: < 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
```

5.3.1 Ejercicios.

1. Calcular las odds-ratio para los restantes factores de riesgo considerados en el estudio original del HIV en mujeres negras (Recibir ayudas sociales, consumir cocaína/crack, prostitución o encarcelamiento de la pareja)
2. Utilizando la base de datos de Telde, calcular la odds-ratio entre las variables TABACO y ALCOHOL
3. Utilizando la base de datos de Telde, calcula el IMC del siguiente modo:

```
telde$IMC=telde$PESO/(telde$TALLA/100)^2
```

y construye la variable OBESIDAD con los valores `OB+` para los sujetos con IMC>30 y `OB-` cuando el IMC es menor o igual que 30:

```
telde$OBESIDAD=ifelse(telde$IMC>30,"OB+","OB-")
table(telde$OBESIDAD)
```

```
##
## OB- OB+
## 698 332
```

Calcula la OR entre obesidad y sedentarismo, y entre obesidad y DM.

6 Evaluación de pruebas diagnósticas y Curvas ROC

6.1 Odds-Ratio

Ya hemos visto en una tarea anterior como calcular la odds-ratio entre dos variables en una tabla 2 por 2. Por ejemplo, si queremos calcular la odds-ratio entre HTA y DM en la base de datos de Telde, procedemos del modo siguiente:

```
library(openxlsx)
telde <- read.xlsx("endocrino.xlsx",sheet=1)
telde$DM <- ordered(telde$DM, levels=c(1,0), labels=c("DM+", "DM-"))
telde$HTA_OMS <- ordered(telde$HTA_OMS, levels=c(1,0), labels=c("HTA+", "HTA-"))
thd=with(telde, table(HTA_OMS,DM))
thd
```

```
##         DM
## HTA_OMS DM+ DM-
##    HTA+  83 241
##    HTA-  45 661
```

Si calculamos las proporciones por columnas obtendremos las proporciones respectivas de hipertensos entre los DM+ y los DM-:

```
round(prop.table(thd,2),3)
```

```
##         DM
## HTA_OMS   DM+   DM-
##    HTA+ 0.648 0.267
##    HTA- 0.352 0.733
```

Vemos que la proporción de hipertensos en mucho mayor entre los diabéticos (64.8%) que entre los no diabéticos (26.7%). El test de la chi-cuadrado revela que esta asociación es significativa (no atribuíble al azar):

```r
chisq.test(thd, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  thd
## X-squared = 75.567, df = 1, p-value < 2.2e-16
```

Recordemos que en R la odds-ratio puede calcularse utilizando la función `epi.2by2` de la librería `epiR`:

```r
library(epiR)
epi.2by2(thd)
```

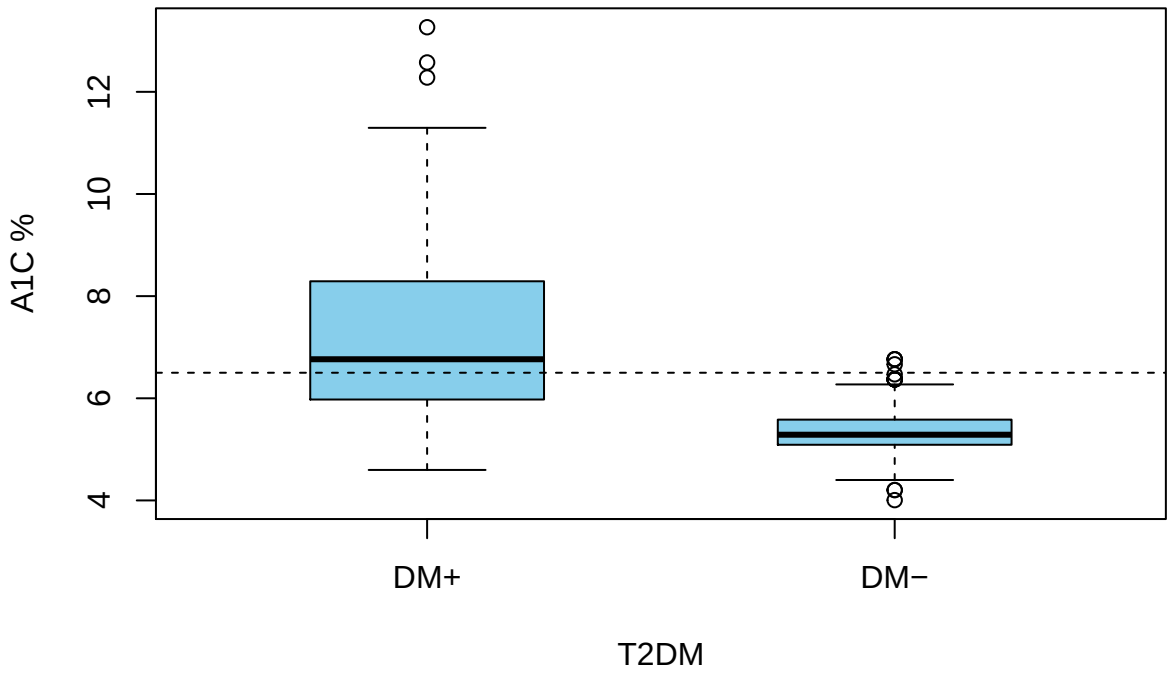
```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +           83          241        324             25.62
## Exposed -           45          661        706              6.37
## Total              128          902       1030             12.43
##                  Odds
## Exposed +      0.3444
## Exposed -      0.0681
## Total          0.1419
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 4.02 (2.87, 5.64)
## Odds ratio                                     5.06 (3.42, 7.48)
## Attrib risk *                                  19.24 (14.16, 24.33)
## Attrib risk in population *                    6.05 (3.35, 8.76)
## Attrib fraction in exposed (%)                 75.12 (65.11, 82.26)
## Attrib fraction in population (%)              48.71 (35.86, 58.99)
## -------------------------------------------------------------------
##  X2 test statistic: 75.567 p-value: < 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
```

6.2 Diagnóstico de la DM por A1C

Supondremos en esta sección que disponemos de un marcador, que es una variable continua, cuyos valores difieren significativamente entre el grupo de las personas enfermas y el grupo de las sanas. Por

ejemplo la hemoglobina glicosilada A1C difiere entre diabéticos y no diabéticos. Si representamos en un boxplot los valores de esta variable en la muestra de Telde, según que los sujetos tengan o no DM obtenemos la figura siguiente:

```
boxplot(A1C ~ DM, data=telde, xlab="T2DM",ylab="A1C %",boxwex=.5,col="skyblue")
abline(h=6.5,lty=2)  # Traza una linea horizontal en el punto de abcisa 6.5
```



donde hemos representado una línea horizontal para un valor de A1C del 6.5%. Tal como vemos, casi todos los sujetos con AIC menor que este valor están libres de DM, mientras que algo más de la mitad de los que tienen DM están por encima del mismo.

Podemos definir una nueva variable que llamaremos `DSC` (discriminante) que distinga si el sujeto tiene el valor de A1C por encima de 6.5 o no; si está por debajo de 6.5 entenderemos que el marcador es negativo (el gráfico indica que valores bajos se asocian con ausencia de DM), y por encima de 6.5 es positivo. En R la codificación sería la siguiente:

```
telde$DSC <- ifelse(telde$A1C>6.5, 1, 0)
telde$DSC <- ordered(telde$DSC, levels=c(1,0), labels=c("DSC+","DSC-"))
table(telde$DSC)
```

```
##
## DSC+ DSC-
##   79  951
```

Si cruzamos esta variable con la DM obtenemos:

```
tdd <- with(telde,table(DSC,DM))
tdd
```

```
##       DM
## DSC    DM+ DM-
##   DSC+  75   4
##   DSC-  53 898
```

y en proporciones:

```
round(prop.table(tdd,2),3)
```

```
##       DM
## DSC      DM+   DM-
##   DSC+ 0.586 0.004
##   DSC- 0.414 0.996
```

Por tanto medir el valor de A1c y comprobar si está por encima o por debajo de 6.5 constituye una prueba muy específica para la diabetes (la probabilidad estimada de que este marcador sea negativo si el sujeto es no diabético es del 99.6%), aunque poco sensible (la probabilidad estimada de que este marcador sea positivo si el sujeto es diabético es del 58.6%).

La función `BDtest` del paquete `bdpv` permite calcular de manera directa la especificidad y sensibilidad, así como otras características, de una prueba diagnóstica. Si no lo tenemos instalado, podemos instalarlo mediante:

```
install.packages("bdpv")
```

Un vez instalado, podemos calcular la sensibilidad y especificidad de la prueba mediante la sintaxis que se especifica a continuación. Nótese que hemos de convertir la tabla `tdd` en un objeto de clase `matrix` (ya que en caso contrario la función `BDtest` no procesa la tabla), y además debemos proporcionar una estimación de la prevalencia en la población. En este caso utilizamos como estimación de prevalencia la proporción de diabéticos observada en la muestra.

```
library(bdpv)  # Cargamos la librería
preval.Diab <- prop.table(table(telde$DM))[1] # Estimador de la prevalencia de diabetes en Telde
class(tdd) <- "matrix"  # se convierte la tabla tdd a la clase `matrix`
BDtest(tdd, pr=preval.Diab, conf.level = 0.95)
```

```
## Confidence intervals for binary diagnostic tests.
## Input data set with columns representing the true property of the compounds and rows representing the
##               True positive True negative
## Test positive            75             4
## Test negative            53           898
## Estimates and exact confidence limits for assay sensitivity and specificity.
##              Estimate Lower 95% limit Lower 97.5% limit Upper 97.5% limit
## Sensitivity 0.5859375       0.5094742         0.4955430         0.6722629
## Specificity 0.9955654       0.9898809         0.9886849         0.9987904
## Estimates and asymptotic confidence limits for predictive values. The prevalence is assumed to be 0.1
##      Estimate Lower 95% limit Lower 97.5% limit Upper 97.5% limit
## NPV 0.9442692       0.9344345         0.9323747         0.9541744
## PPV 0.9493671       0.8910529         0.8746377         0.9805408
```

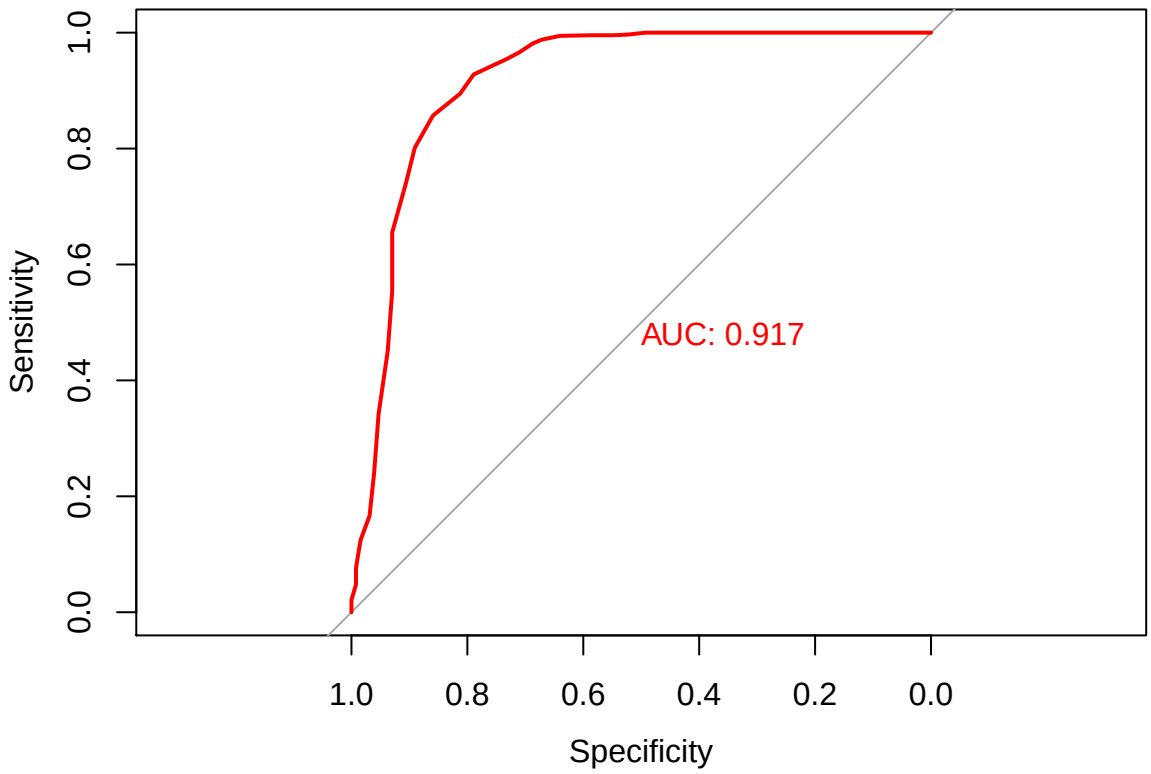
6.3 Curva ROC

Para construir la curva ROC necesitamos instalar el paquete `pROC` en caso de que aún no lo tengamos:

```
install.packages("pROC")
```

La curva ROC se obtiene mediante:

```
library(pROC)
rocA1C <- with(telde,roc(DM,A1C))
plot(rocA1C, col="red", print.auc=TRUE)
```




```
##
## Call:
## roc.default(response = DM, predictor = A1C)
##
## Data: A1C in 128 controls (DM DM+) > 902 cases (DM DM-).
## Area under the curve: 0.9171
```

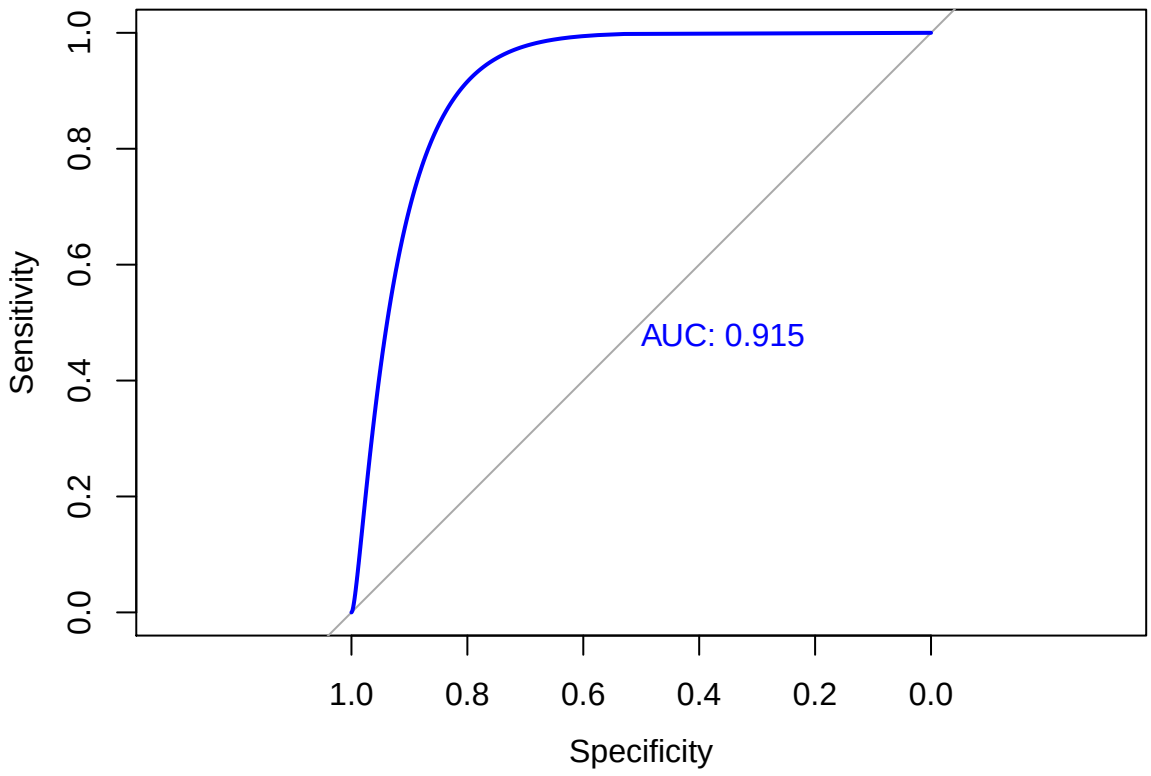
Un intervalo de confianza para el área bajo la curva se obtiene mediante:

```
ci.auc(rocA1C)
```

```
## 95% CI: 0.8821-0.9521 (DeLong)
```

y una gráfica "suavizada" de la curva ROC:

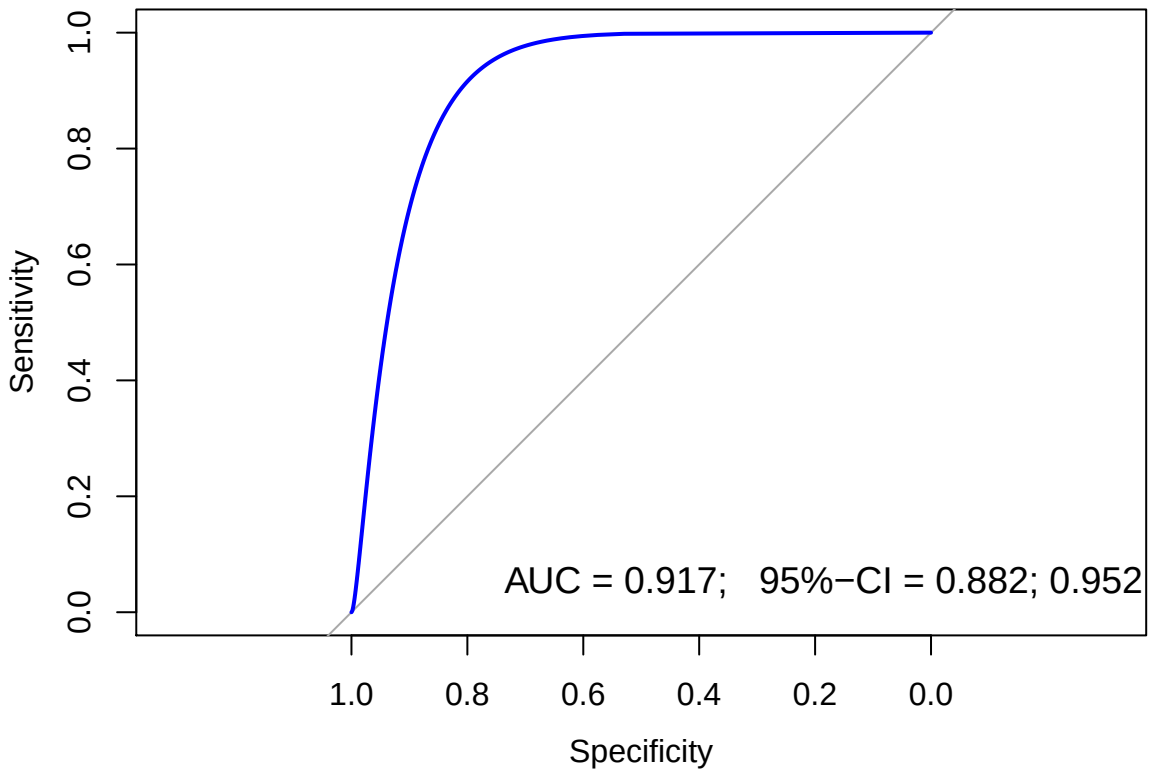
```r
plot.roc(smooth(rocA1C),col="blue", print.auc=TRUE)
```




```
##
## Call:
## smooth.roc(roc = rocA1C)
##
## Data: A1C in 128 controls (DM DM+) > 902 cases (DM DM-).
## Smoothing: binormal
## Area under the curve: 0.9154
```

Unas pocas líneas de código permiten añadir el área y el intervalo de confianza a la gráfica, dejándola lista para publicación:

```r
ci <- as.numeric(ci.auc(rocA1C)) # Valores numéricos del área y el intervalo de confianza
a<-plot.roc(smooth(rocA1C),col="blue")
legend(0.85,.15, sprintf("AUC = %.3f;   95%%-CI = %.3f; %.3f",ci[2],ci[1],ci[3]),bty="n",cex=1.15)
```



La función `coords` permite encontrar aquel valor de la variable que al ser usado como punto de corte (cutoff) para discriminar entre sanos y enfermos, maximiza cierta función de la sensibilidad y la especificidad. Por defecto la función a maximizar es:

$$(1 - sensitivity)^2 + (1 - specificity)^2$$

cuyo objetivo es conseguir que la sensibilidad y la especificidad se aproximen (simultáneamente) todo lo posible a 1. Al aplicar esta función a nuestros datos resulta:

```
closest <- coords(rocA1C,"b",ret=c("threshold","specificity","sensitivity","npv","ppv"),
       best.method="closest.topleft")
closest
```

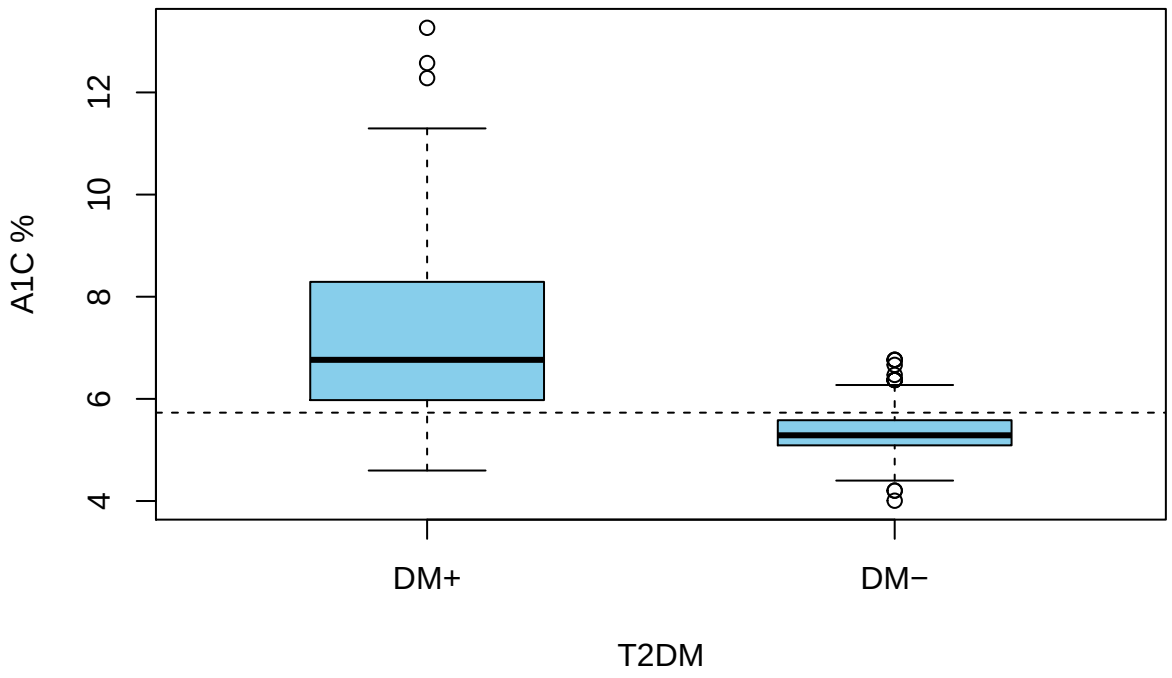
```
##   threshold specificity sensitivity         npv         ppv
##   5.7297500   0.8593750   0.8569845   0.4602510   0.9772440
```

Podemos definir la regla discriminante (prueba diagnóstica) en función de este cutoff (5.73%), igual que hicimos más arriba con el 6.5%, considerando que la prueba es positiva si A1C>5.73 y negativa en caso contrario:

```
telde$DSC <- ifelse(telde$A1C>closest[1], 1, 0)
telde$DSC <- ordered(telde$DSC, levels=c(1,0), labels=c("DSC+","DSC-"))
table(telde$DSC)
```

```
##
## DSC+ DSC-
##  239  791
```

```
boxplot(A1C ~ DM, data=telde, xlab="T2DM",ylab="A1C %",boxwex=.5,col="skyblue")
abline(h=closest[1],lty=2)
```



Por último, evaluamos la sensibilidad, especificidad, VPP, VPN de esta prueba diagnóstica mediante `BDtest`:

```
tdd <- with(telde,table(DSC,DM))
tdd
```

```
##       DM
## DSC    DM+ DM-
##   DSC+ 110 129
##   DSC-  18 773
```

```
class(tdd) <- "matrix"
BDtest(tdd, pr=preval.Diab, conf.level = 0.95)
```

```
## Confidence intervals for binary diagnostic tests.
```

```
## Input data set with columns representing the true property of the compounds and rows representing the
##               True positive True negative
## Test positive           110           129
## Test negative            18           773
## Estimates and exact confidence limits for assay sensitivity and specificity.
##              Estimate Lower 95% limit Lower 97.5% limit Upper 97.5% limit
## Sensitivity 0.8593750       0.7986299         0.7868852         0.9144768
## Specificity 0.8569845       0.8363907         0.8324189         0.8791902
## Estimates and asymptotic confidence limits for predictive values. The prevalence is assumed to be 0.1
##     Estimate Lower 95% limit Lower 97.5% limit Upper 97.5% limit
## NPV 0.977244       0.9676985         0.9654712         0.9850649
## PPV 0.460251       0.4241572         0.4173220         0.5037780
```

6.4 Ejercicios.

1. La siguiente tabla muestra los datos de un estudio sobre la utilización de la PSA (Prostate Specific Antigen) como prueba diagnóstica de la presencia de cáncer de próstata:

	D+	D-
4 ng/mL (T+)	443	855
$<$ 4 ng/mL (T-)	137	976

Utilizar R para calcular la sensibilidad, especificidad, valor predictivo negativo y valor predictivo positivo de esta prueba diagnóstica en los siguientes casos:

a. Cuando la prevalencia del cáncer de próstata en la población es del 5%
b. Cuando esta prevalencia es del 1%
c. cuando esta prevalencia es del 10%

2. HDL como predictor de la presencia del genotipo homocigoto B1/B1 en el polimorfismo Taq 1B. En el estudio de Telde se midió el polimorfismo Taq1B en el gen de la Cholesteryl ester transfer protein (CETP). Este polimorfismo tiene dos alelos, B1 y B2, por lo cual los posibles genotipos son

B1/B1, B1/B2 y B2/B2. Estudios previos indican que los individuos homocigotos B1/B1 tienden a presentar valores más bajos de HDL.

Podemos comprobar que en la base de datos de Telde, la variable CETP registra el genotipo de este polimorfismo en aquellas personas en que se pudo medir:

```
table(telde$CETP)
```

```
##
## B1B1 B1B2 B2B2
##  200  175   52
```

```
sum(table(telde$CETP))
```

```
## [1] 427
```

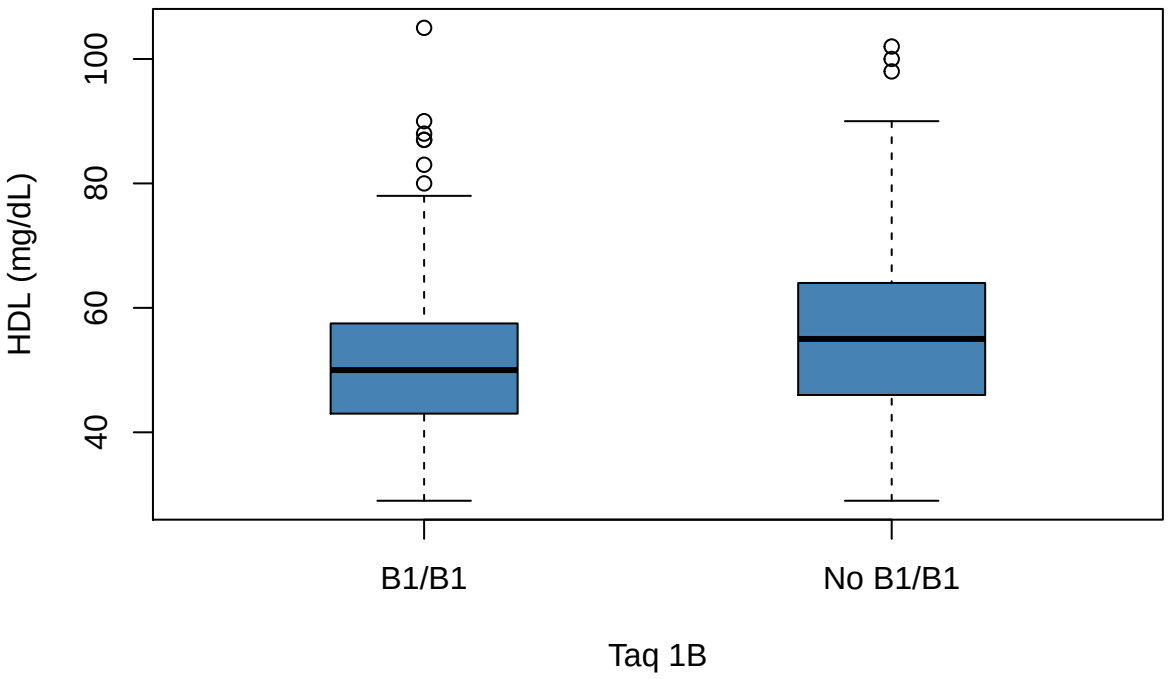
Construiremos un nuevo factor `Taq_1B` que nos indique si el individuo es portador o no del genotipo B1/B1:

```
telde$Taq_1B=factor(ifelse(telde$CETP=="B1B1","B1/B1","No B1/B1"))
table(telde$Taq_1B)
```

```
##
##    B1/B1 No B1/B1
##      200      227
```

Un sencillo `boxplot` nos muestra la posible asociación entre este genotipo y el valor del HDL, y confirma la presencia de valores ligeramente más bajos de HDL en los sujetos B1/B1:

```
boxplot(HDL ~ Taq_1B,data=telde, xlab="Taq 1B",ylab="HDL (mg/dL)",boxwex=.4,col="steelblue")
```



Construir la curva ROC para evaluar la capacidad discriminante del valor de HDL como predictor de la presencia del genotipo homocigoto B1/B1 en el polimorfismo Taq1B del gen CETP. Determinar el cutoff óptimo así como los valores de sensibilidad y especificidad alcanzados para dicho cutoff.

7 Distribucion binomial y regresion logistica

7.1 Concepto de variable aleatoria. Distribución binomial.

Una variable aleatoria es una cantidad cuyo valor depende del azar. A modo de ejemplo, si la prevalencia de cierta enfermedad en una población es $\pi = 0.30$, ello significa que la probabilidad de que una persona elegida al azar padezca esa enfermedad es 0.3. Si elegimos al azar 3 personas de esa población, podrá ocurrir:

- Que ninguna esté enferma. Este suceso tiene probabilidad $p_0 = (1-\pi)^3$
- Que solo haya una enferma: $p_1 = 3\pi\,(1-\pi)^2$.
- Que haya dos enfermas: $p_2 = 3\pi^2\,(1-\pi)$.
- Que las tres estén enfermas: $p_3 = \pi^3$

En este contexto, $X =$ "número de enfermos entre las tres personas elegidas al azar"" es una variable aleatoria en el sentido que se acaba de definir. Concretamente, si el hecho de que un sujeto esté enfermo es independiente de que el resto de sujetos esté enfermo o no, el reparto (o distribución) de probabilidades entre los distintos valores de la variable X recibe el nombre de distribución binomial, en este caso particular, de parámetros $n = 3$ y $\pi = 0.3$, y se suele denotar de la forma $X \approx b\,(n, \pi)$

Podemos calcular las probabilidades anteriores con R:

```
p0=(1-0.3)^3; p0
```

```
## [1] 0.343
```

```
p1=3*0.3*(1-0.3)^2; p1
```

```
## [1] 0.441
```

```
p2=3*(0.3^2)*(1-0.3); p2
```

```
## [1] 0.189
```

```
p3=0.3^3; p3
```

```
## [1] 0.027
```

Estas cuatro probabilidades suman 1:

```
p0+p1+p2+p3
```

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

7.1.1 Esperanza de una variable aleatoria (discreta):

Se define como:

$$E[X] = \sum_{k=0}^{n} k \cdot p(X = k)$$

En el caso de la variable aleatoria de nuestro ejemplo:

```
0*p0+1*p1+2*p2+3*p3
```

```
## [1] 0.9
```

Se puede demostrar que para la distribución binomial $b(n, \pi)$ la ecuación anterior puede simplificarse como:

$$E[X] = n \cdot \pi$$

En nuestro ejemplo $n \cdot \pi = 3 \cdot 0.3 = 0.9$ que coincide con el valor que se acaba de calcular.

Podemos interpretar intuitivamente el concepto de esperanza en este caso considerando que en lugar de una muetra de 3 personas tenemos una muestra de 300; si la probabilidad de que una persona elegida al azar esté enferma es del 30% (esto es, $\pi = 0.3$), cabe esperar que un 30% de las 300 personas (esto es, 90 personas) estén enfermas. Este valor esperado coincide precisamente con $E[X] = n \cdot \pi = 300 \cdot 0.3 = 90$.

Otra manera de interpretar la esperanza es como el valor medio de la variable en muchas muestras. R permite simular con facilidad valores de variables aleatorias; en particular, si $X \approx b(300, 0.3)$ (esto es, X es el número de enfermos entre 300 personas elegidas al azar cuando la probabilidad de estar enfermo es 0.3), puede simularse un valor de X como:

```r
rbinom(1,300,0.3)
```

```
## [1] 80
```

La siguiente sintaxis simula el número de enfermos en cada una de 10 muestras de 300 personas cada una:

```r
rbinom(10,300,0.3)
```

```
##  [1]  79  84 107  92  94 105  98  91 104  88
```

Y ahora generamos 1000 muestras de tamaño 300 y contamos el número de enfermos en cada una:

```r
enfermos=rbinom(1000,300,0.3)
```

Si calculamos el número medio de enfermos en estas muestras de tamaño 300:

```r
mean(enfermos)
```

```
## [1] 90.345
```

obtenemos un valor muy próximo a la esperanza calculada más arriba (0.9)

7.1.2 Varianza de una variable aleatoria (discreta)

Se define como:

$$Var\left(X\right)=\sum_{k=0}^{n}\left(k-E[X]\right)^{2}\cdot p\left(X=k\right)$$

La varianza es una medida de la variabilidad presente en una variable aleatoria. En el caso particular de la distribución binomial $b\left(n,\pi\right)$ la ecuación anterior puede simplificarse como:

$$Var\left(X\right)=n\cdot\pi\cdot\left(1-\pi\right)$$

La desviación típica es la raiz cuadrada de la varianza:

$$sd\left(X\right)=\sqrt{n\cdot\pi\cdot\left(1-\pi\right)}$$

Calculamos la varianza para una variable $b\left(300,0.3\right)$:

```r
300*0.3*(1-0.3)
```

```
## [1] 63
```

y la desviación típica:

```r
sqrt(300*0.3*(1-0.3))
```

```
## [1] 7.937254
```

Podemos calcular la varianza y desviación típica en las 1000 muestras anteriores y comprobar que son similares a estos valores teóricos:

```r
var(enfermos)
```

```
## [1] 60.13411
```

```r
sd(enfermos)
```

```
## [1] 7.754619
```

7.2 Estudio de Telde: prevalencia de HTA

Cargamos los datos del estudio de Telde:

```r
library(openxlsx)
setwd("c:/Users/aulas/Downloads/")
telde = read.xlsx("endocrino.xlsx")
```

En el estudio de Telde tenemos una muestra de $n = 1030$ personas. El número de personas con HTA entre estas 1030 es una variable aleatoria con distribución binomial $b\,(1030, \pi)$, donde π es la probabilidad de que una persona elegida al azar de esta población padezca HTA. El valor de π en la población adulta de Telde es desconocido, pero podemos estimarlo (obtener un valor aproximado) a partir de los datos de nuestra muestra, usnado como estimador la prevalencia observada de HTA. Dicha prevalencia puede calcularse mediante:

```
tb=table(telde$HTA_OMS)
tb
```

```
##
##   0   1
## 706 324
```

```
ptb=prop.table(tb)
ptb
```

```
##
##         0         1
## 0.6854369 0.3145631
```

Por tanto de acuerdo con nuestros datos, la prevalencia de HTA en Telde ronda un 31.46% (esto es, la probabilidad de que una persona elegida al azar en la población adulta de Telde tenga HTA es aproximadamente 0.3146).

Asimismo podemos estimar la prevalencia de HTA de acuerdo a la presencia/ausencia de T2DM:

```
tb2 <- with(telde,table(DM,HTA_OMS))
tb2
```

```
##    HTA_OMS
## DM    0   1
##   0 661 241
##   1  45  83
```

```
prop.table(tb2,1)
```

```
##    HTA_OMS
## DM          0         1
##   0 0.7328160 0.2671840
##   1 0.3515625 0.6484375
```

Esta tabla nos indica que entre los diabéticos hay un 64.8% de hipertensos (83 hipertensos de un total de 83+45=128 sujetos); asimismo entre los no diabéticos (241+661=902) hay 241 hipertensos, lo que da lugar a una prevalencia de HTA de un 26.7% entre los no diabéticos). Este resultado muestra bien a las claras que la probabilidad de que una persona tenga HTA depende de si dicha persona tiene o no DM: es más probable ser hipertenso cuando se es diabético que cuando no se es diabético.

7.3 Modelo de regresión logística con una única variable explicativa: HTA según DM

Preparamos un data.frame con los sujetos según tengan o no T2DM, indicando en cada grupo (con DM y sin DM), el número de sujetos con HTA, y el número total de sujetos (`Nt`).:

```
freqHTA=aggregate(HTA_OMS~DM, telde,
                  function(x){c(con.HTA=sum(x),Nt=length(x),pHTA=sum(x)/length(x))})
freqHTA=data.frame(DM=freqHTA[,1],freqHTA[,-1])
rownames(freqHTA)=c("DM-","DM+")
freqHTA
```

```
##     DM con.HTA  Nt      pHTA
## DM-  0     241 902 0.2671840
## DM+  1      83 128 0.6484375
```

La siguiente sintaxis permite construir el modelo de regresión logística para predecir la prevalencia de HTA según la presencia/ausencia de diabetes:

```
mlog <- glm(pHTA ~ DM, data=freqHTA ,weights=Nt, family=binomial)
summary(mlog)
```

```
##
## Call:
## glm(formula = pHTA ~ DM, family = binomial, data = freqHTA, weights = Nt)
##
## Deviance Residuals:
## [1]  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.00896    0.07525 -13.408  < 2e-16 ***
## DM           1.62114    0.19983   8.113 4.96e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.9666e+01  on 1  degrees of freedom
## Residual deviance: 5.4179e-14  on 0  degrees of freedom
## AIC: 16.228
##
## Number of Fisher Scoring iterations: 2
```

Las predicciones de este modelo se obtienen mediante:

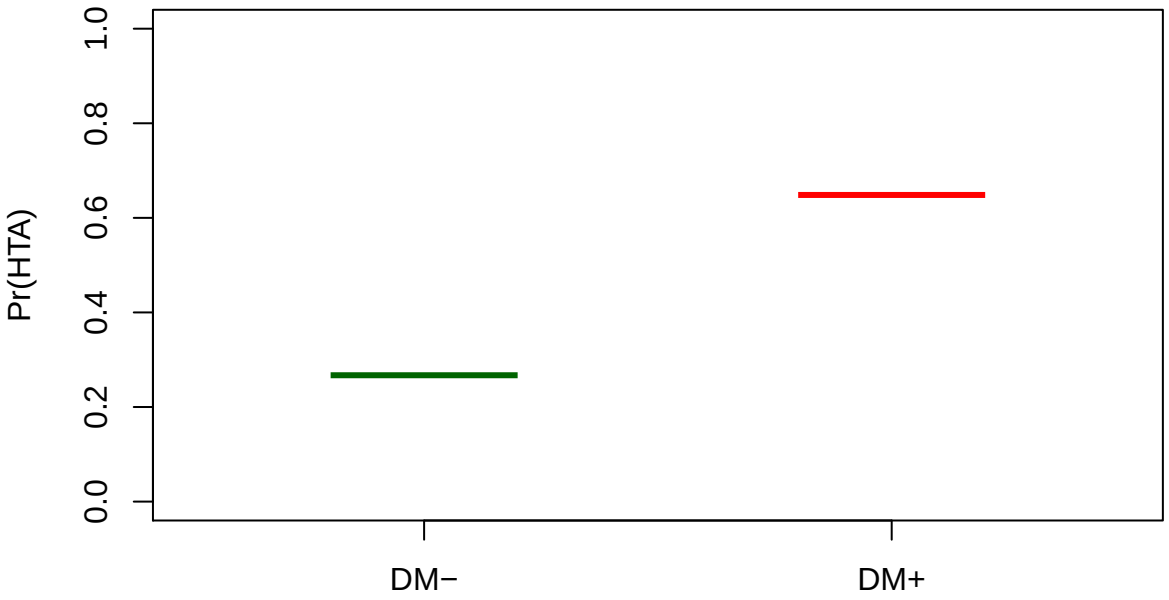
```
PrHTA=mlog$fit
PrHTA
```

```
##       DM-       DM+
## 0.2671840 0.6484375
```

Podemos observar que la predicción es exacta, esto es, se predicen exactamente los valores de las prevalencias observadas.

Gráficamente:

```
boxplot(PrHTA~names(PrHTA), ylim=c(0,1),boxwex=0.4,medcol=c("darkgreen","red"), ylab="Pr(HTA)",
        boxlty = 0,  whisklty = 0, staplelty = 0)
```



Para calcular la regresión logística anterior hemos construido una tabla resumen con las proporciones observadas de padecer HTA según se padezca o no DM. Podemos obtener el mismo resultado (ya que el modelo es el mismo) si se utiliza directamente la base de datos del estudio de Telde; en este caso la estimación del modelo se lleva a cabo de manera muy simple:

```
mlogB <- glm(HTA_OMS ~ DM, data=telde , family=binomial)
summary(mlogB)
```

```
##
## Call:
## glm(formula = HTA_OMS ~ DM, family = binomial, data = telde)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4459  -0.7885  -0.7885   0.9308   1.6247
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.00896    0.07525 -13.408  < 2e-16 ***
## DM           1.62114    0.19983   8.113 4.96e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1282.8  on 1029  degrees of freedom
## Residual deviance: 1213.1  on 1028  degrees of freedom
## AIC: 1217.1
##
## Number of Fisher Scoring iterations: 4
```

y ahora la predicción se realiza para cada sujeto por separado; para cada sujeto se predice la probabilidad de que padezca HTA en función de que tenga o no DM. Mostramos a continuación las predicciones para los sujetos 1, 2, 900, 901, 1001, 1002,1028,1029 (elegimos estos sujetos a modo de ejemplo, para no tener que mostrar las predicciones de los 1030 que componen la muestra completa):

```
mlogB$fit[c(1, 2, 900, 901, 1001, 1002,1028,1029)]
```

```
##         1         2       900       901      1001      1002      1028
## 0.2671840 0.2671840 0.2671840 0.2671840 0.6484375 0.6484375 0.6484375
##      1029
## 0.6484375
```

7.4 Modelo de regresión logística multivariante: HTA según IR y DM

Introduzcamos a continuación en el modelo anterior, además del efecto de la T2DM, el efecto de la resistencia a la insulina. Para ello construimos en primer lugar una tabla que nos da el número de hipertensos para cada una de las cuatro posibles combinaciones de T2DM e IR:

```
freqHTA=aggregate(HTA_OMS~IR+DM, telde,
                  function(x){c(con.HTA=sum(x),Nt=length(x),pHTA=sum(x)/length(x))})
freqHTA=data.frame(freqHTA[,(1:2)],freqHTA[,-(1:2)])
rownames(freqHTA)=c("IR-/DM-","IR+,DM-","IR-/DM+","IR+/DM+")
freqHTA
```

```
##          IR DM con.HTA  Nt      pHTA
## IR-/DM-   0  0     136 675 0.2014815
## IR+,DM-   1  0     105 227 0.4625551
## IR-/DM+   0  1      14  25 0.5600000
## IR+/DM+   1  1      69 103 0.6699029
```

El modelo de regresión logística para esta tabla se calcula mediante:

```
mlog2 <- glm(pHTA ~ DM  + IR, data=freqHTA, weights=Nt, family=binomial)
summary(mlog2)
```

```
##
## Call:
## glm(formula = pHTA ~ DM + IR, family = binomial, data = freqHTA,
##     weights = Nt)
##
## Deviance Residuals:
## IR-/DM-  IR+,DM-  IR-/DM+  IR+/DM+
## -0.3115   0.4353   1.3104  -0.6964
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.34726    0.09328 -14.443  < 2e-16 ***
## DM           1.06299    0.21625   4.915 8.86e-07 ***
## IR           1.13921    0.15430   7.383 1.55e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 126.0959  on 3  degrees of freedom
## Residual deviance:   2.4885  on 1  degrees of freedom
## AIC: 29.534
##
## Number of Fisher Scoring iterations: 3
```

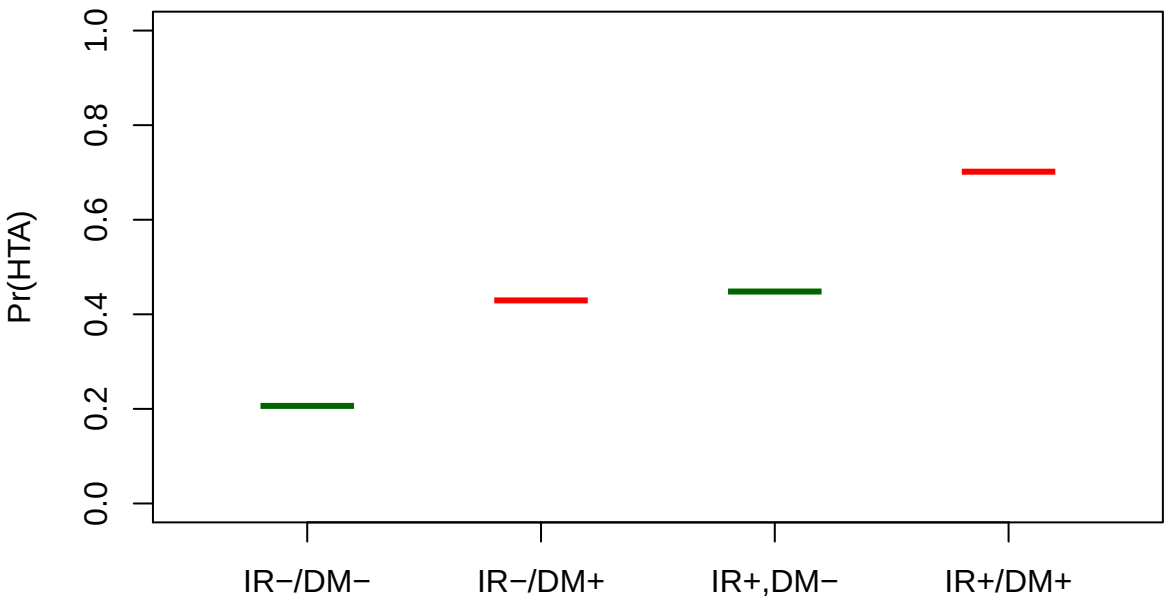
cuyas predicciones son las siguientes (obsérvese que ahora ya no coinciden con las proporciones observadas):

```
PrHTA2=mlog2$fit
PrHTA2
```

```
##   IR-/DM-   IR+,DM-   IR-/DM+   IR+/DM+
## 0.2063183 0.4481725 0.4294062 0.7016004
```

Gráficamente:

```
boxplot(PrHTA2~names(PrHTA2), ylim=c(0,1),boxwex=0.4,medcol=c("darkgreen","red"), ylab="Pr(HTA)",
        boxlty = 0,  whisklty = 0, staplelty = 0)
```



El modelo con interacciones sí que hace predicciones exactas:

```
mlog3 <- glm(pHTA ~ DM  * IR, data=freqHTA, weights=Nt, family=binomial)
summary(mlog3)
```

```
##
## Call:
## glm(formula = pHTA ~ DM * IR, family = binomial, data = freqHTA,
##     weights = Nt)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.37706    0.09596 -14.350  < 2e-16 ***
## DM           1.61822    0.41418   3.907 9.34e-05 ***
## IR           1.22700    0.16410   7.477 7.59e-14 ***
## DM:IR       -0.76042    0.48288  -1.575    0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:  1.2610e+02  on 3  degrees of freedom
## Residual deviance: -5.3513e-14  on 0  degrees of freedom
## AIC: 29.046
##
## Number of Fisher Scoring iterations: 3
```

```
mlog3$fit
```

```
##   IR-/DM-   IR+,DM-   IR-/DM+   IR+/DM+
## 0.2014815 0.4625551 0.5600000 0.6699029
```

Del mismo modo que antes, la regresión logística puede estimarse a partir de la base de datos original mediante:

```
mg <- glm(HTA_OMS ~ DM + IR , data =telde, family=binomial)
summary(mg)
```

```
##
## Call:
## glm(formula = HTA_OMS ~ DM + IR, family = binomial, data = telde)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5552  -0.6798  -0.6798   0.8419   1.7767
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.34726    0.09328 -14.443  < 2e-16 ***
## DM           1.06299    0.21625   4.915 8.86e-07 ***
## IR           1.13921    0.15430   7.383 1.55e-13 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1282.8  on 1029  degrees of freedom
## Residual deviance: 1159.2  on 1027  degrees of freedom
## AIC: 1165.2
##
## Number of Fisher Scoring iterations: 4
```

Ahora lo que se predice para cada sujeto es la probabilidad de que sea hipertenso según que tenga o no T2DM, y según que tenga o no IR; para los mismos sujetos de antes, las predicciones son:

```
mg$fit[c(1, 2, 900, 901, 1001, 1002,1028,1029)]
```

```
##         1         2       900       901      1001      1002      1028
## 0.4481725 0.4481725 0.2063183 0.2063183 0.7016004 0.7016004 0.4294062
##      1029
## 0.4294062
```

7.4.1 Regresión logística multivariante: Prevalencia de HTA según incluyendo DM y edad:

```
mg <- glm(HTA_OMS ~ DM + EDAD, data=telde, family=binomial)
summary(mg)
```

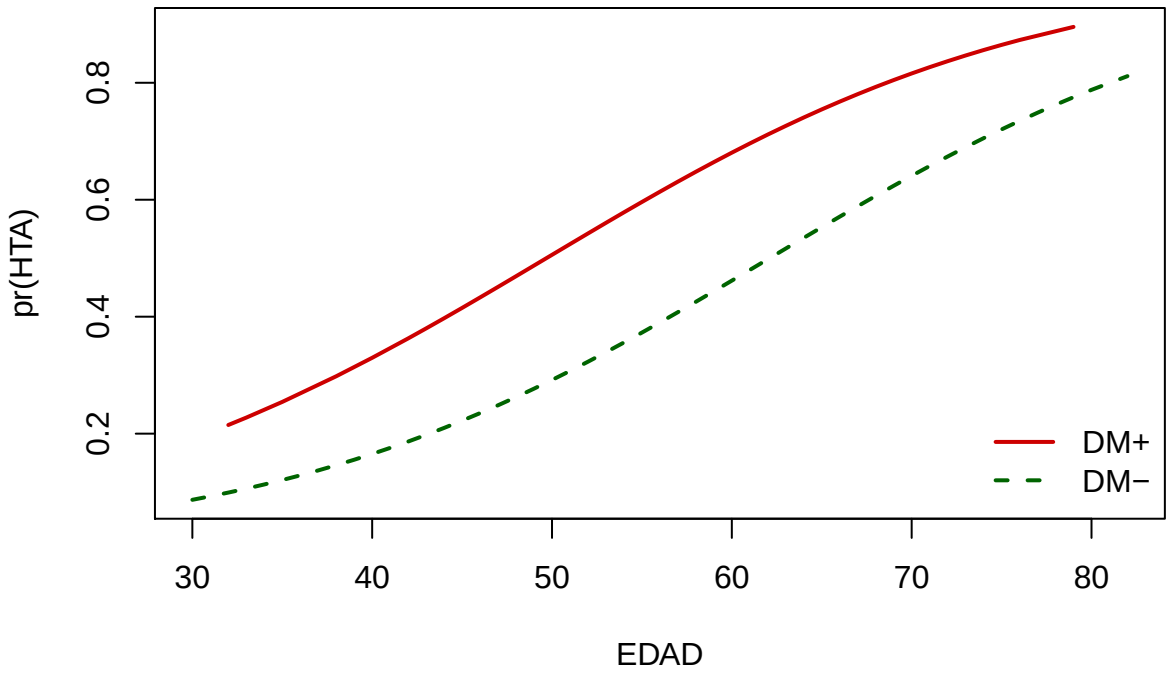
```
##
## Call:
## glm(formula = HTA_OMS ~ DM + EDAD, family = binomial, data = telde)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1255  -0.7804  -0.5432   0.9044   2.2104
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.550680   0.348603 -13.054  < 2e-16 ***
## DM           0.909660   0.219401   4.146 3.38e-05 ***
```

```
## EDAD         0.073285   0.006818  10.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1282.8  on 1029  degrees of freedom
## Residual deviance: 1081.5  on 1027  degrees of freedom
## AIC: 1087.5
##
## Number of Fisher Scoring iterations: 4
```

Podemos representar gráficamente el efecto de la edad en cada grupo (DM+ y DM-) del siguiente modo:

```
telde$pHTA=mg$fitted        # Añadimos las predicciones a la base de datos
telde=telde[order(telde$EDAD), ] # Ordenamos por edad
si.DM=subset(telde,DM==1)   # Subconjunto de diabéticos
no.DM=subset(telde,DM==0)   # Subconjunto de  no diabéticos

plot(pHTA~EDAD,telde, type="n", xlab="EDAD", ylab="pr(HTA)")  # Gráfico en blanco
lines(no.DM$EDAD,no.DM$pHTA,type="l",col="darkgreen",lwd=2,lty=2)
lines(si.DM$EDAD,si.DM$pHTA,type="l",col="red3",lwd=2,lty=1)
legend("bottomright", c("DM+","DM-"), lty=c(1,2), col=c("red3","darkgreen"), lwd=2, bty="n")
```



7.4.2 Regresión logística multivariante: Prevalencia de HTA según incluyendo DM, IR y edad:

```
mg2 <- glm(HTA_OMS ~ DM + IR + EDAD, data=telde, family=binomial)
summary(mg2)
```

```
##
## Call:
## glm(formula = HTA_OMS ~ DM + IR + EDAD, family = binomial, data = telde)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0154  -0.7342  -0.4717   0.8531   2.3655
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.965598   0.369248 -13.448  < 2e-16 ***
## DM           0.323218   0.236159   1.369    0.171
## IR           1.178958   0.166337   7.088 1.36e-12 ***
## EDAD         0.074358   0.007001  10.621  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1282.8  on 1029  degrees of freedom
## Residual deviance: 1031.2  on 1026  degrees of freedom
## AIC: 1039.2
##
## Number of Fisher Scoring iterations: 4
```

Gráficamente:

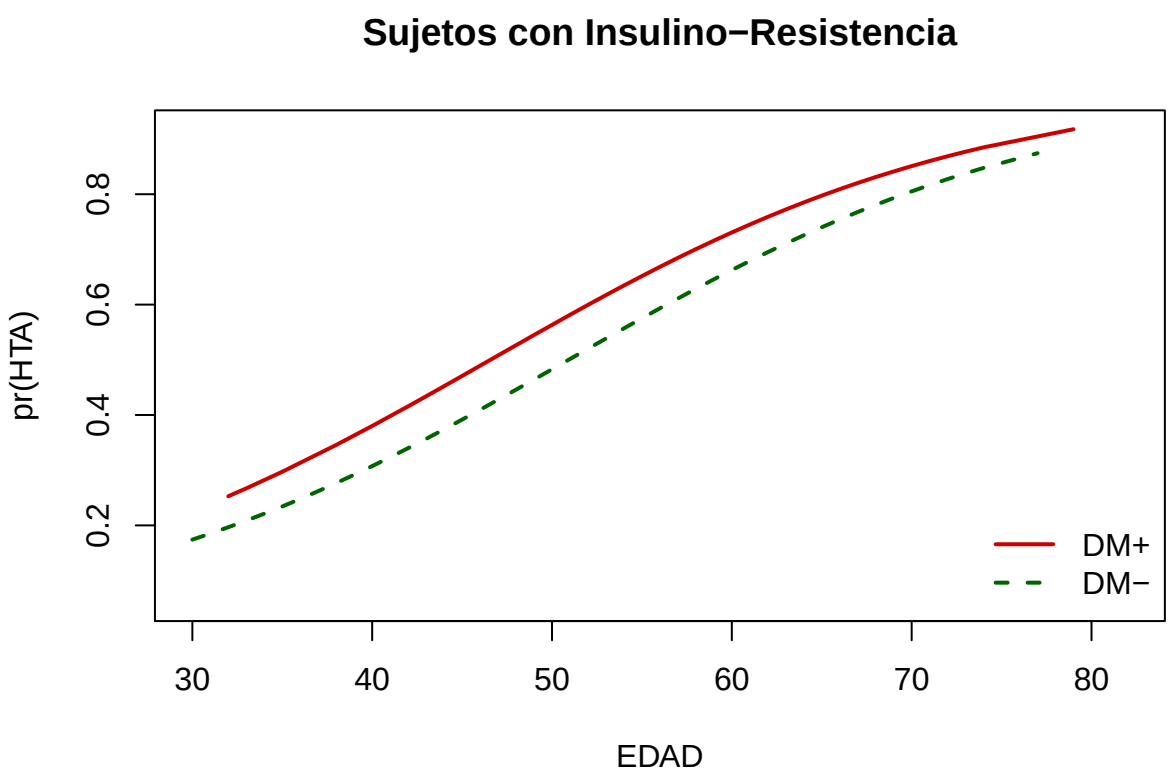
```
telde$pHTA=mg2$fitted        # Añadimos las predicciones a la base de datos
telde=telde[order(telde$EDAD), ] # Ordenamos por edad
```

- Representamos primero los sujetos insulino resistentes:

```
teldeIR=subset(telde,IR==1)
si.DM=subset(teldeIR,DM==1)   # Subconjunto de diabéticos e IR
no.DM=subset(teldeIR,DM==0)   # Subconjunto de no diabéticos e IR

plot(pHTA~EDAD,telde, type="n", xlab="EDAD", ylab="pr(HTA)",
```

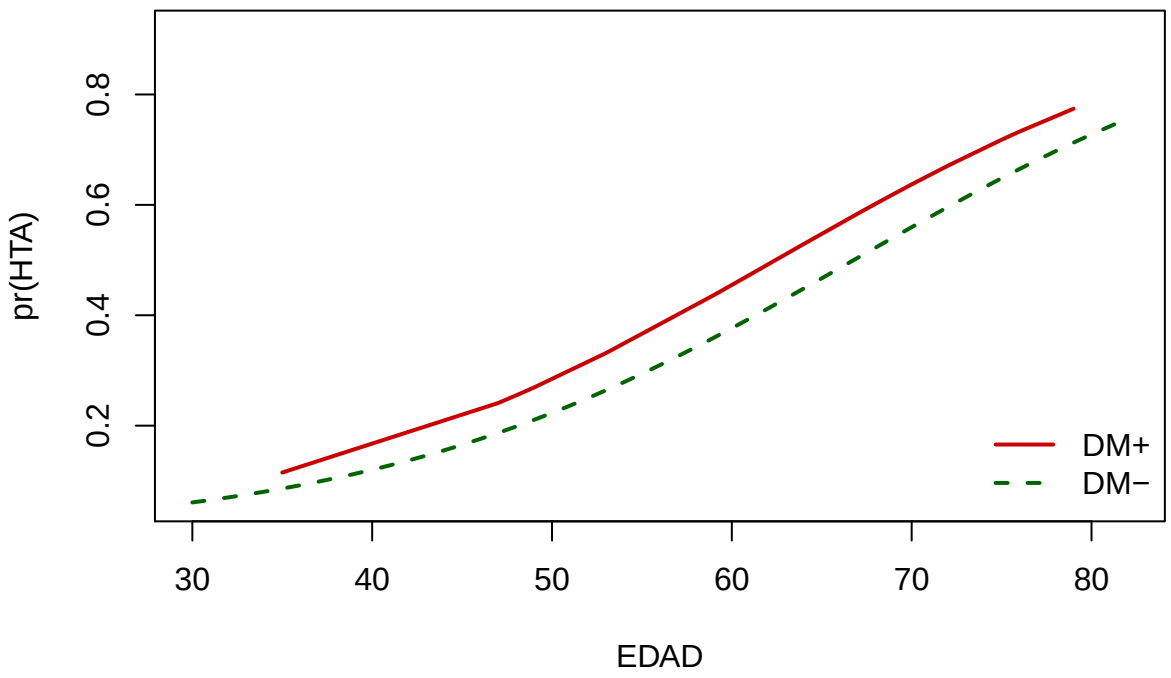
```
     main="Sujetos con Insulino-Resistencia")
lines(no.DM$EDAD,no.DM$pHTA,type="l",col="darkgreen",lwd=2,lty=2)
lines(si.DM$EDAD,si.DM$pHTA,type="l",col="red3",lwd=2,lty=1)
legend("bottomright", c("DM+","DM-"), lty=c(1,2), col=c("red3","darkgreen"), lwd=2, bty="n")
```



- Y ahora los no Insulino-Resistentes:

```
telde.noIR=subset(telde,IR==0)
si.DM=subset(telde.noIR,DM==1)    # Subconjunto de diabéticos e IR
no.DM=subset(telde.noIR,DM==0)    # Subconjunto de no diabéticos e IR

plot(pHTA~EDAD,telde, type="n", xlab="EDAD", ylab="pr(HTA)",
     main="Sujetos SIN Insulino-Resistencia")
lines(no.DM$EDAD,no.DM$pHTA,type="l",col="darkgreen",lwd=2,lty=2)
lines(si.DM$EDAD,si.DM$pHTA,type="l",col="red3",lwd=2,lty=1)
legend("bottomright", c("DM+","DM-"), lty=c(1,2), col=c("red3","darkgreen"), lwd=2, bty="n")
```



7.5 Cálculo de las odds-ratio a partir de la regresión logística.

Ya hemos visto como podemos usar `epiR` para calcular la odds-ratio entre HTA_OMS y DM:

```
library(epiR)
fDM=ordered(telde$DM,levels=c(1,0),labels=c("DM+","DM-"))
fHTA=ordered(telde$HTA_OMS,levels=c(1,0),labels=c("HTA+","HTA-"))
tb <- table(fDM,fHTA)
tb
```

```
##      fHTA
## fDM   HTA+ HTA-
##   DM+   83   45
##   DM-  241  661
```

```
epi.2by2(tb)
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +           83           45        128              64.8
## Exposed -          241          661        902              26.7
```

```
## Total              324          706       1030              31.5
##                   Odds
## Exposed +        1.844
## Exposed -        0.365
## Total            0.459
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                  2.43 (2.05, 2.87)
## Odds ratio                                      5.06 (3.42, 7.48)
## Attrib risk *                                   38.13 (29.36, 46.89)
## Attrib risk in population *                     4.74 (0.69, 8.79)
## Attrib fraction in exposed (%)                  58.80 (51.30, 65.14)
## Attrib fraction in population (%)               15.06 (10.77, 19.15)
## -------------------------------------------------------------------
##  X2 test statistic: 75.567 p-value: < 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
```

o directamente a partir de la tabla cruzada:

```
OR=tb[1,1]*tb[2,2]/(tb[1,2]*tb[2,1])
OR
```

```
## [1] 5.058829
```

La odds ratio es, como vemos, 5.058829. El modelo logístico para predecir la prevalencia de HTA_OMS a partir de DM es:

```
mg <- glm(HTA_OMS~DM, data=telde, family=binomial)
summary(mg)
```

```
##
## Call:
## glm(formula = HTA_OMS ~ DM, family = binomial, data = telde)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4459  -0.7885  -0.7885   0.9308   1.6247
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.00896    0.07525 -13.408  < 2e-16 ***
```

```
## DM           1.62114    0.19983   8.113 4.96e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1282.8  on 1029  degrees of freedom
## Residual deviance: 1213.1  on 1028  degrees of freedom
## AIC: 1217.1
##
## Number of Fisher Scoring iterations: 4
```

Los valores de las odds-ratio se obtienen como el resultado de calcular el valor del número e elevado a los coeficientes estimados del modelo:

```
exp(coef(mg))
```

```
## (Intercept)          DM
##   0.3645991   5.0588290
```

y podemos ver que el valor así obtenido coincide con el ya calculado a partir de la tabla anterior.

El empleo de la regresión logística nos permite obtener las odds-ratio ajustadas por otras variables:

```
mg2 <- glm(HTA_OMS ~ DM + IR + EDAD, data=telde, family=binomial)
exp(coef(mg2))
```

```
## (Intercept)          DM          IR        EDAD
## 0.006973777 1.381565898 3.250984231 1.077192451
```

Se pueden también obtener intervalos de confianza y contrastes de significación para las odds-ratio:

```
library(pander)
library(MASS)     # Para el cálculo de intervalos de confianza
smg<-summary(mg2)
pv=smg$coef[,4]   # p-valores de los coeficientes
pval=c(ifelse(pv<.001,"< .001",round(pv,3)))
b=smg$coef[,1]
sb=smg$coef[,2]
bs=array(sprintf("%.3f (%.3f)",b,sb),dim=c(length(b),1))
OR=exp(b)
OR.ci=exp(confint(mg))
```

```
IC95=array(sprintf("%.3f (%.3f,%.3f)",OR,OR.ci[,1],OR.ci[,2]),dim=c(length(OR),1))
ta5 <- data.frame(bs,pval,IC95)
names(ta5)=c("Coeficiente (SE)","P","OR (IC-95%)")
pander(ta5)
```

	Coeficiente (SE)	P	OR (IC-95%)
(Intercept)	-4.966 (0.369)	$< .001$	0.007 (0.314,0.422)
DM	0.323 (0.236)	0.171	1.382 (3.437,7.535)
IR	1.179 (0.166)	$< .001$	3.251 (0.314,0.422)
EDAD	0.074 (0.007)	$< .001$	1.077 (3.437,7.535)

8 Inferencia estadística. Tests de hipótesis e intervalos de confianza

En esta sesión veremos diversas herramientas disponibles en R para la realización de contrastes de hipótesis y el cálculo de intervalos de confianza.

Como en anteriores ocasiones, utilizaremos los datos del estudio de Telde:

```
library(openxlsx)
setwd("c:/Users/aulas/Downloads/")
telde = read.xlsx("endocrino.xlsx")
```

8.1 Test de la chi-cuadrado: ¿Existe asociación entre el sexo de un individuo y el padecer HTA?

Para determinar si el sexo de una persona se asocia con el hecho de que esa persona padezca o no HTA comenzamos por la construcción de una tabla cruzada de ambas variables:

```
telde$SEXO=ordered(telde$SEXO,levels=c(1,0),labels=c("Hombre","Mujer"))
telde$HTA=ordered(telde$HTA_OMS,levels=c(1,0),labels=c("HTA+","HTA-"))
```

```
tb=with(telde,table(HTA,SEXO))
tb
```

```
##       SEXO
## HTA    Hombre Mujer
##   HTA+    156   168
##   HTA-    292   414
```

```
prop.table(tb,2)
```

```
##       SEXO
## HTA       Hombre     Mujer
##   HTA+ 0.3482143 0.2886598
##   HTA- 0.6517857 0.7113402
```

Esta tabla indica que un 34.8% de los hombres son hipertensos, frente a un 28.9% de las mujeres, lo que podría interpretarse como que el ser hombre es de algún modo un factor de riesgo de hipertensión. Podemos evaluar la asociación entre sexo y HTA mediante la odds-ratio:

```
tb[1,1]*tb[2,2]/(tb[1,2]*tb[2,1])
```

```
## [1] 1.316536
```

cuyo valor (mayor que la unidad) está efectivamente acorde con la presunción de que ser hombre incrementa el riesgo de HTA.

En cualquier caso la diferencia entre un 34.8% y un 28.9% no parece excesivamente elevada. Dado que los datos con los que estamos trabajando constituyen una muestra elegida al azar entre la población adulta de Telde cabe hacerse la siguiente pregunta: si la muestra hubiese sido otra ¿se habría observado la misma relación entre estos porcentajes (esto es mayor porcentaje de hipertensos entre las mujeres que entre los hombres), o quizás estos porcentajes podrían llegar incluso a invertirse? En otras palabras, la pregunta que nos hacemos es: ¿la diferencia observada puede explicarse simplemente por la variabilidad aleatoria presente en el muestreo, o necesariamente se debe a que en realidad la hipertensión es más común entre los hombres que entre las mujeres de Telde?

Para responder a esta pregunta debe utilizarse el test de la chi-cuadrado:

- La hipótesis nula es que la hipertensión afecta por igual a hombres y mujeres y por tanto que la diferencia observada entre los porcentajes se debe simplemente al azar; si π_M es la probabilidad de que una mujer sea hipertensa y π_H es la probabilidad de que un hombre sea hipertenso, la hipótesis nula establece que:

$$\frac{\pi_M}{\pi_H} = 1$$

- La hipótesis alternativa es que la hipertensión afecta de manera diferente a hombres y mujeres, y por tanto que:

$$\frac{\pi_M}{\pi_H} \neq 1$$

La realización del test de la chi-cuadrado en R es muy simple: se utiliza la función `chisq.test` a la que se le pasa como argumento la tabla con los datos disponibles:

```
chisq.test(tb)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb
## X-squared = 3.8924, df = 1, p-value = 0.0485
```

Si deseamos calcular las frecuencias esperadas bajo la hipótesis nula (igual prevalencia de HTA en hombres y mujeres) podemos utilizar:

```
chisq.test(tb)$expected
```

```
##       SEXO
## HTA      Hombre    Mujer
##   HTA+ 140.9243 183.0757
##   HTA- 307.0757 398.9243
```

8.2 Test de la chi-cuadrado: ¿Existe asociación entre el nivel de instrucción de un individuo y el padecer HTA?

Para responder a esta pregunta podemos proceder del mismo modo que con el sexo:

```
telde$INSTRUCCION=ordered(telde$INSTRUCCION,levels=c("Sin estudios","Primer grado",
                                                     "Segundo grado","Tercer grado"))
tb=with(telde,table(HTA,INSTRUCCION))
tb
```

```
##       INSTRUCCION
## HTA    Sin estudios Primer grado Segundo grado Tercer grado
##   HTA+            6          167           130           21
##   HTA-            7          191           379          129
```

```
prop.table(tb,2)
```

```
##       INSTRUCCION
## HTA    Sin estudios Primer grado Segundo grado Tercer grado
##   HTA+    0.4615385    0.4664804     0.2554028    0.1400000
##   HTA-    0.5384615    0.5335196     0.7445972    0.8600000
```

Esta última tabla muestra que a medida que aumenta el nivel de instrucción aparentemente disminuye la probabilidad de padecer hipertensión.

Aplicamos el test de la chi-cuadrado:

```
chisq.test(tb)
```

```
## Warning in chisq.test(tb): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tb
## X-squared = 69.084, df = 3, p-value = 6.705e-15
```

Observemos que R nos muestra una advertencia relativa a que la aproximación chi-cuadrado puede ser incorrecta. Ello es debido a que las condiciones "técnicas" de realización del test requieren que ninguna frecuencia esperada sea menor que 1 y que al menos el 80 % de las frecuencias esperadas son mayores que 5. Las frecuencias esperadas en este caso son:

```
chisq.test(tb)$expected
```

```
## Warning in chisq.test(tb): Chi-squared approximation may be incorrect
```

```
##       INSTRUCCION
## HTA    Sin estudios Primer grado Segundo grado Tercer grado
##   HTA+      4.08932     112.6136      160.1126     47.18447
##   HTA-      8.91068     245.3864      348.8874    102.81553
```

En este caso hay una única frecuencia esperada menor que 5; en total hay 8 valores esperados de los que 7, esto es el 87.5%, son mayores que 5. Por tanto podemos confiar en el resultado del test y concluir que la asociación entre el nivel de instrucción y la HTA es significativa, esto es, no puede explicarse simplemente por efecto del azar.

En caso de que no se hubiesen cumplido las condiciones para la realización del test, se debería proceder a agrupar categorías; en este caso, como la categoría que puede causar problemas es la categoría "Sin estudios", ésta podría fusionarse con la categoría "Primer grado":

```
telde$INSTRUCCION2=ifelse((telde$INSTRUCCION=="Sin estudios")|(telde$INSTRUCCION=="Primer grado"),"
tb=with(telde,table(HTA,INSTRUCCION2))
tb
```

```
##       INSTRUCCION2
## HTA    Segundo grado Sin Estudios/Primer grado Tercer grado
##   HTA+           130                       173           21
##   HTA-           379                       198          129
```

```r
prop.table(tb,2)
```

```
##       INSTRUCCION2
## HTA    Segundo grado Sin Estudios/Primer grado Tercer grado
##   HTA+     0.2554028                 0.4663073    0.1400000
##   HTA-     0.7445972                 0.5336927    0.8600000
```

```r
chisq.test(tb)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tb
## X-squared = 69.082, df = 2, p-value = 9.976e-16
```

8.3 Test de la t de Student (t.test): ¿Existe asociación entre el índice de masa corporal y la hipertensión arterial?

El índice de masa corporal se calcula como el peso (en kg) partido por la talla (en m) elevada al cuadrado. Podemos calcular el IMC de las personas de la muestra de Telde mediante:

```r
telde$IMC=telde$PESO/(telde$TALLA/100)^2
```

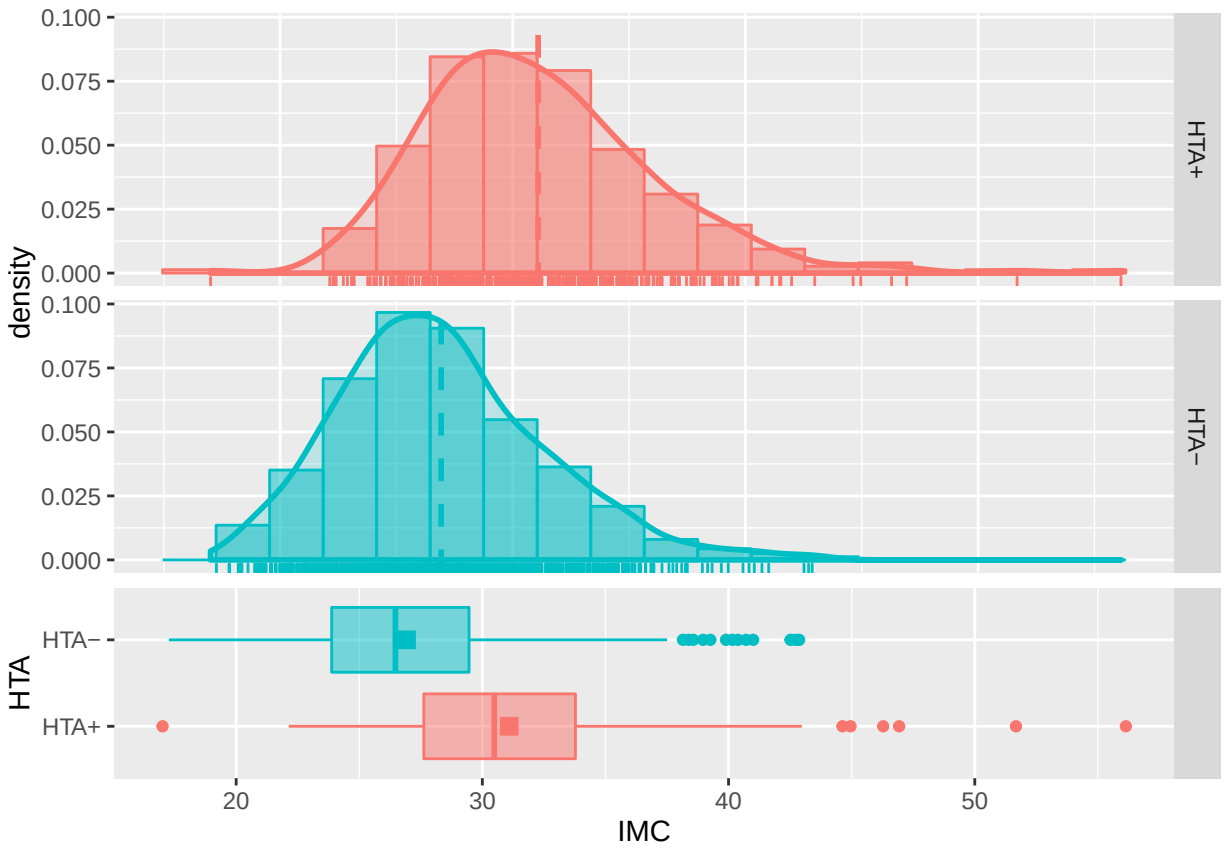
La siguiente tabla nos da el valor medio de IMC según las personas sean o no hipertensas:

```r
aggregate (IMC~HTA, telde, mean)
```

```
##    HTA      IMC
## 1 HTA+ 31.09140
## 2 HTA- 26.92119
```

Gráficamente:

```r
library(ULPGCmisc)
summarize(telde$IMC,by=telde$HTA, showSummary = FALSE)
```



Vemos, pues, que el IMC medio es mayor en los hipertensos (31.10) que en los normotensos (26.92). Igual que en el caso anterior: ¿podría esta diferencia ser efecto de la variabilidad aleatoria del muestreo o la diferencia es tan grande que solo puede explicarse porque en realidad un mayor IMC se asocia con riesgo de HTA?

Para responder a esta pregunta podemos utilizar el test de la de Student:

- La hipótesis nula es que la hipertensión no se asocia con el IMC; ello significaría que el valor medio de IMC sería el mismo en el grupo de hipertensos y en el grupo de normotensos. Si llamamos μ_H al valor medio de IMC en el primer grupo y μ_N en el segundo, la hipótesis nula especifica que:

$$\mu_H = \mu_N$$

- La hipótesis alternativa es que el valor medio de IMC difiere entre hipertensos y normotensos; podemos ser más específicos y considerar que la hipótesis alternativa es:

$$\mu_H > \mu_N$$

esto es que el valor medio de IMC es estrictamente mayor en hipertensos que en normotensos. Para resolver este contraste ejecutamos la siguiente función en R:

```r
t.test(IMC~HTA,data=telde,alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  IMC by HTA
## t = 12.909, df = 565.08, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.637958      Inf
## sample estimates:
## mean in group HTA+ mean in group HTA-
##           31.09140           26.92119
```

El p-valor es muy pequeño (<0.0001) lo que indica que hay evidencia suficiente para asegurar que el valor de IMC en el grupo de hipertensos es mayor que en el grupo de los normotensos; la diferencia observada no puede explicarse por el mero efecto del azar.

La salida de R nos informa además que con una confianza del 95% el valor medio de IMC en hipertensos es al menos 3.64 unidades mayor que en normotensos.

Si queremos obtener un intervalo de confianza al 95% para la diferencia en el valor medio de IMC entre hipertensos y normotensos utilizamos la sintaxis anterior sin especificar alternativa; asimismo podemos indicar fácilmente el nivel de confianza deseado:

```r
t.test(IMC~HTA,data=telde, conf=0.99)$conf.int
```

```
## [1] 3.335257 5.005155
## attr(,"conf.level")
## [1] 0.99
```

8.4 Test de Wilcoxon-Mann-Whitney (wilcox.test): ¿Existe asociación entre el nivel de glucosa en sangre y la hipertensión arterial?

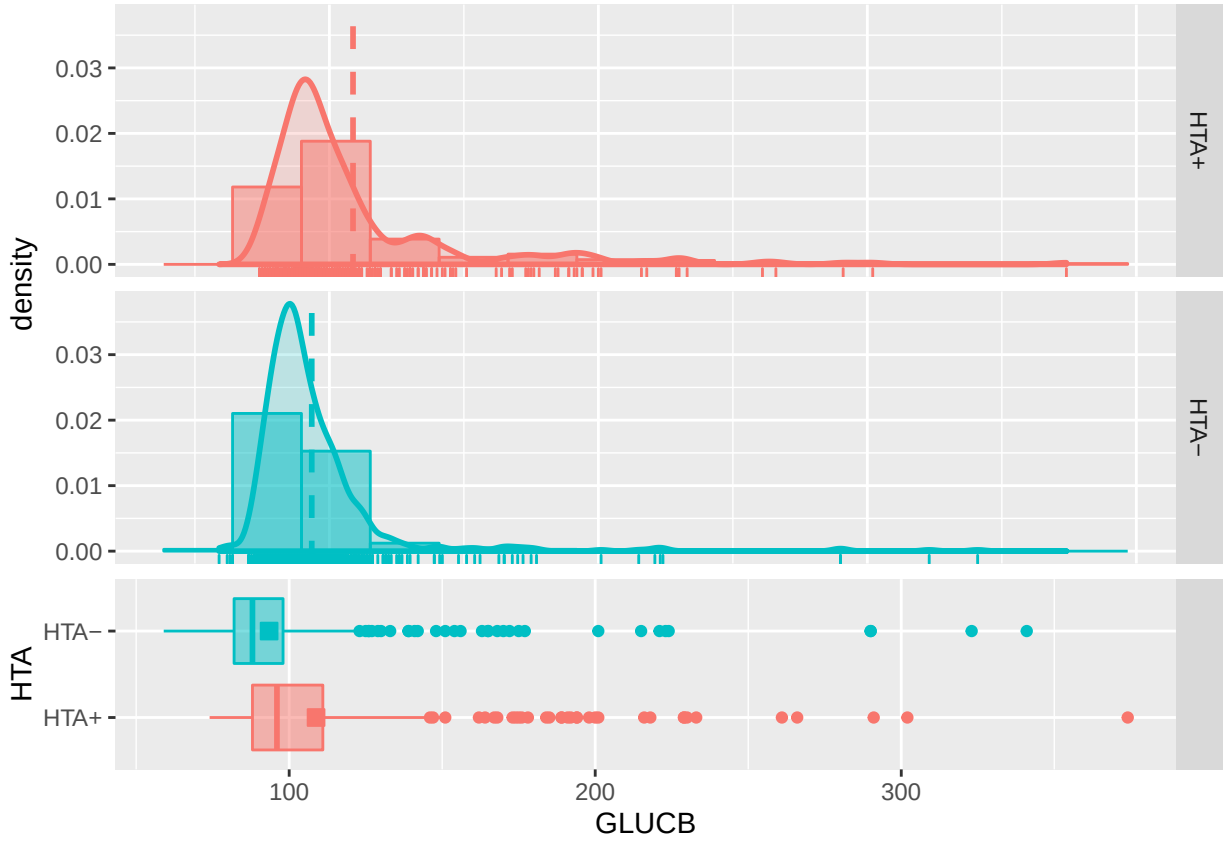
De igual modo que para el caso anterior podemos tratar de determinar la posible asociación entre el nivel de glucosa en sangre y la HTA. Los valores medios de glucosa en sangre en hipertensos y normotensos pueden obtenerse como:

```
aggregate(GLUCB~HTA, data=telde, mean)
```

```
##    HTA     GLUCB
## 1 HTA+ 108.78086
## 2 HTA-  93.43768
```

Gráficamente:

```
a<- summarize(telde$GLUCB,by=telde$HTA, showSummary = FALSE)
```



EStos gráficos indican que el nivel de glucosa en sangre tiene una distribución muy asimétrica en ambos grupos. Por tanto compararlos mediante sus valores medios puede ser muy poco adecuado; además la fuerte asimetría implica que la variable bajo estudio (GLUCB) no sigue una distribución normal por lo que el test de la t de Student puede no ser válido.

En este caso resulta más conveniente realizar la comparación mediante el test de Wilcoxon-Mann-Whitney; la hipótesis nula en este test es que los valores de la variable tienden a ser similares en ambos grupos; la hipótesis alternativa es que los valores en uno de los grupos tienden a ser mayores que en el otro. Concretamente, si sospechamos que el nivel de glucosa es mayor en hipertensos, el contraste a realizar es:

```r
wilcox.test(GLUCB~HTA, data=telde,alternative="greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  GLUCB by HTA
## W = 152680, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

El p-valor indica que hay evidencia suficiente para asegurar que efectivamente en el grupo HTA+ el nivel de glucosa es sangre es mayor que en el grupo HTA-.

8.5 Datos emparejados

La siguiente base de datos:

```r
library(foreign)
millac=read.spss("MILLAC.sav", to.data.frame=TRUE)
```

```
## Warning in read.spss("MILLAC.sav", to.data.frame = TRUE): MILLAC.sav:
## Unrecognized record type 7, subtype 18 encountered in system file
```

```r
head(millac)
```

```
##   ID          grupo sq  eLDL  vLDL eTG vTG eHDL vHDL eCT vCT elpa vlpa
## 1  1 Entera-Vegetal  1 101.2  89.0  44  90   56   55 166 162 15.0   16
## 2  2 Entera-Vegetal  1 147.4 121.0  73 140   46   45 208 194 19.0   14
## 3  3 Entera-Vegetal  1 118.0  94.8  90  56   46   55 182 161  8.7    7
## 4  4 Entera-Vegetal  1 105.4  91.8  78  96   68   67 189 178  0.6   NA
## 5  5 Entera-Vegetal  1 170.4 133.0  63  60   62   57 245 202  1.4   NA
## 6  6 Entera-Vegetal  1  98.0  70.8  70  61   48   48 160 131  9.6    7
##   eapoa vapoa eapob vapob
## 1   150   157    81    55
## 2   132   147   109    66
## 3   151   156   102    54
## 4   173   181    85    53
## 5   161   130   124    79
## 6   136   119    77    49
```

corresponde a un estudio sobre los niveles de HDL y LDL observados en niños dependiendo del tipo de leche que consumían. En concreto, el estudio se llevó a cabo mediante un diseño cruzado: los niños se dividieron aleatoriamente en dos grupos; en el primero (`sq=1`) los niños consumieron leche entera durante 6 meses y leche desnatada (con grasa vegetal) durante los 6 meses siguientes; en el segundo grupo (`sq=2`), durante los primeros 6 meses se consumió leche desnatada, y los siguientes 6 meses leche entera. Para cada niño se evaluaron los niveles de HDL y LDL al final de cada periodo de 6 meses. En lo que sigue denotamos por eLDL el valor de LDL tras el periodo de consumo de leche entera, y por vLDL el valor de LDL tras el consumo de leche desnatada enriquecida con grasa vegetal.

Los valores medios y desviaciones típicas de los valores de `eLDL` y `vLDL` observados han sido:

```
aggregate(eLDL~grupo,data=millac,function(x) sprintf("%.2f (%.2f)",mean(x),sd(x)))
```

```
##            grupo           eLDL
## 1 Entera-Vegetal 111.51 (23.50)
## 2 Vegetal-Entera  99.91 (31.09)
```

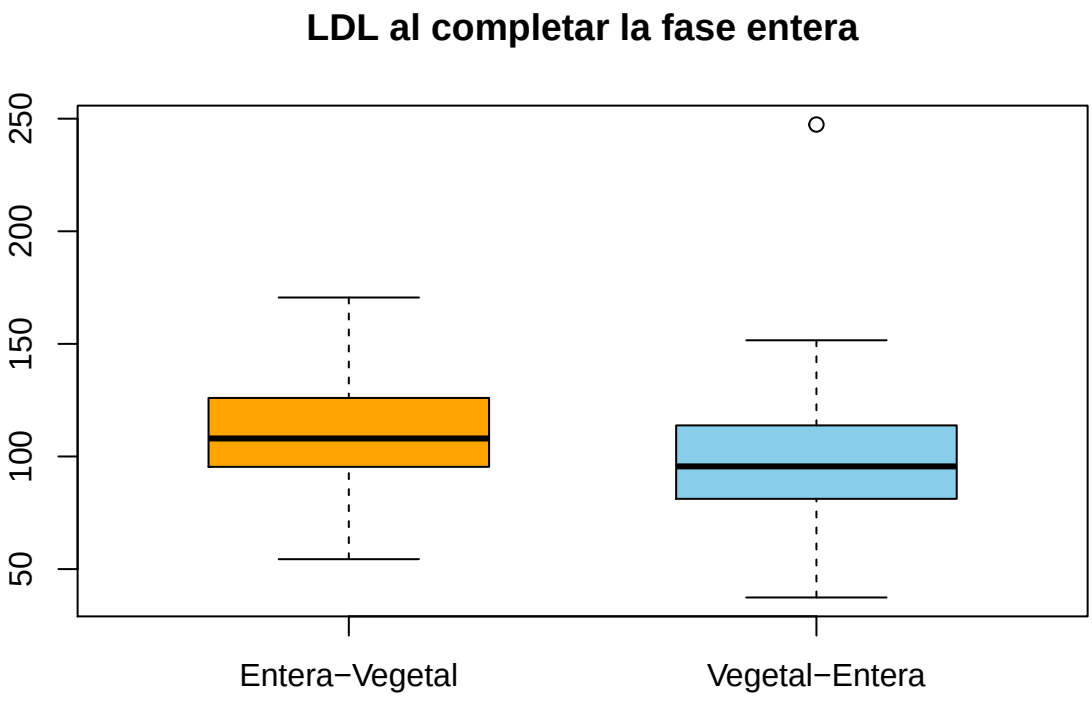
```
aggregate(vLDL~grupo,data=millac,function(x) sprintf("%.2f (%.2f)",mean(x),sd(x)))
```

```
##            grupo           vLDL
## 1 Entera-Vegetal 94.89 (19.95)
## 2 Vegetal-Entera 94.07 (22.83)
```

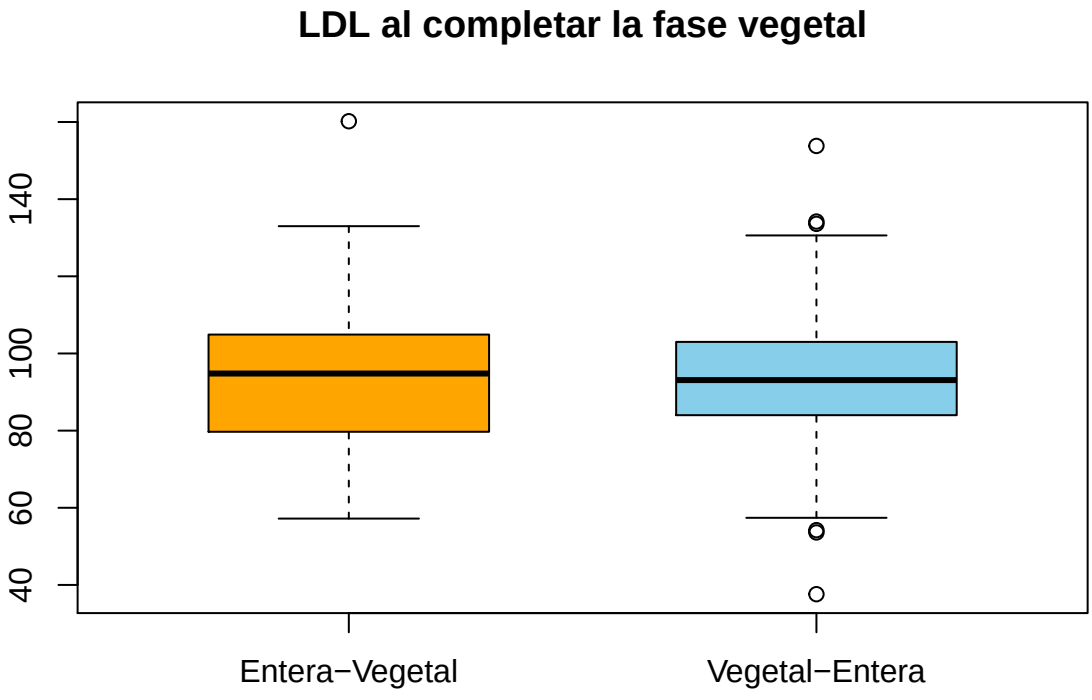
Así, los niveles de LDL tras la fase de consumo de leche entera (`eLDL`) fueron mayores en los niños que comenzaron con leche entera (LDL medio de 111.51) que en los que comenzaron con leche desnatada (LDL medio 99.91). Asimismo, los niveles de LDL tras la fase de consumo de leche desnatada (`vLDL`) fueron similares en los dos grupos (LDL medio de 94.89 entre los que hicieron la secuencia entera-vegetal, y 94.07 entre los que hicieron la secuencia vegetal-entera).

Gráficamente:

```
boxplot(eLDL~grupo, data=millac, main="LDL al completar la fase entera",
        boxwex=0.6,col=c("orange","skyblue"))
```




```
boxplot(vLDL~grupo, data=millac, main="LDL al completar la fase vegetal",
        boxwex=0.6,col=c("orange","skyblue"))
```



Los niños que han tomado leche en la secuencia "Entera-Vegetal" tuvieron un valor medio de LDL al terminar la fase entera de 111.51 unidades; esos mismos niños, al terminar la fase vegetal mostraron un valor medio de LDL de 94.89 unidades. ¿Es significativa esta diferencia? Es decir ¿puede explicarse por el mero efecto del azar o necesariamente debe atribuirse al efecto de haber cambiado el tipo de leche?

A diferencia de los casos anteriores para hacer esta comparación debe tenerse en cuenta que los datos están emparejados (cada niño se compara consigo mismo). Si suponemos que la variable LDL es aproximadamente normal, podemos aplicar un t.test:

```
eLDL=subset(millac,sq==1)$eLDL  # Datos del grupo 1 (secuencia entera-vegetal) al final de la fas
vLDL=subset(millac,sq==1)$vLDL  # Datos del grupo 1 al final de la fase vegetal
t.test(eLDL,vLDL, paired=TRUE, alternative="greater")
```

```
##
##  Paired t-test
##
## data:  eLDL and vLDL
## t = 5.2117, df = 46, p-value = 2.153e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  11.2359     Inf
## sample estimates:
## mean of the differences
##                16.57447
```

lo que indica que existe evidencia suficiente para asegurar que se ha producido una disminución del LDL. La magnitud de dicha dismimución puede estimarse mediante un intervalo de confianza:

```
t.test(eLDL,vLDL, paired=TRUE, alternative="greater")$conf.int
```

```
## [1] 11.2359     Inf
## attr(,"conf.level")
## [1] 0.95
```

lo que indica que, con un 95% de confianza, que se ha producido un incremento del LDL de al menos 11.236 unidades.

Por su parte, los niños que hicieron la secuencia inversa (primero leche desnadata-vegetal y luego leche entera), tenían un nivel medio de LDL al terminar la fase vegetal de 94.07, mientras que al terminar la fase entera el LDL había subido a 99.91. Nuevamente nos hacemos la pregunta: ¿es significativa esta diferencia?

Para responderla, al igual que antes, comparamos cada niño consigo mismo a través de un t-test para muestras emparejadas:

```r
eLDL=subset(millac,sq==2)$eLDL  # Datos del grupo 2 (secuencia vegetal-entera) al final de la fas
vLDL=subset(millac,sq==2)$vLDL  # Datos del grupo 2 al final de la fase vegetal
t.test(vLDL,eLDL, paired=TRUE, alternative="less")
```

```
##
##  Paired t-test
##
## data:  vLDL and eLDL
## t = -1.9392, df = 52, p-value = 0.02895
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.7803787
## sample estimates:
## mean of the differences
##               -5.720755
```

lo que indica que efectivamente se ha producido un incremento significativo del LDL

```r
t.test(eLDL,vLDL, paired=TRUE, alternative="less")$conf.int
```

```
## [1]     -Inf 10.66113
## attr(,"conf.level")
## [1] 0.95
```

Por tanto podemos decir, con un 95% de confianza, que el incremento en LDL ha sido inferior a 10.66 unidades.

9 Simulación

9.1 Introducción.

Consideremos la siguiente situación: queremos determinar la tasa de remisión que se consigue con cierto tratamiento para una enfermedad. No es factible aplicar dicho tratamiento a todos los posibles pacientes afectados, por lo que solo se aplica en una muestra de n pacientes elegidos al azar. Transcurrido el tiempo de aplicación del tratamiento, evaluamos a los pacientes tratados y determinamos para cada uno de ellos si la enfermedad ha remitido o no. Si la enfermedad ha remitido en n_R de los n pacientes iniciales, nuestra estimación (valor aproximado) de la tasa de remisión conseguida con el tratamiento será:

$$p = \frac{n_R}{n}$$

Surgen de inmediato varias preguntas: ¿este procedimiento proporciona realmente un valor aproximado de la tasa de remisión que se consiguiría con el tratamiento si se aplicara a todos los pacientes y no solo a una muestra? ¿Qué margen de error tiene esta aproximación? ¿Cuál es el tamaño adecuado de la muestra?

En este documento utilizaremos las herramientas de simulación que ofrece R para ayudarnos a entender las ideas subyacentes al problema de estimación de parámetros. A partir de estas ideas trataremos de dar respuesta a las preguntas que nos hemos hemos planteado.

9.2 Simulación de variables aleatorias.

R permite simular de manera muy sencilla el comportamiento de una variable aleatoria en el muestreo. Si la distribución de probabilidad de dicha variable aleatoria depende de un parámetro (desconocido) θ, esta simulación nos ayuda a entender qué información podemos obtener de dicho parámetro a partir de los datos muestrales.

En primer lugar aclaremos qué significa la frase “simular el comportamiento de una variable aleatoria en el muestreo”:

- Una variable aleatoria es una cantidad cuyo valor observado depende del azar. En el ejemplo anterior, la variable aleatoria es la tasa de remisión que se observa en la muestra de pacientes; a priori no podemos predecir en qué pacientes se producirá la remisión y en qué pacientes no, y por tanto no podemos conocer el valor exacto de dicha tasa hasta que termine el estudio.
- El muestreo es el proceso mediante el cual se observan valores de la variable aleatoria. En el ejemplo, es el procedimiento por el cual se seleccionan los pacientes que participan en el estudio, se les aplica el tratamiento y se determina en cuántos se ha conseguido la remisión.
- El comportamiento de la variable aleatoria en el muestreo se refiere a la caracterización de los valores que puede presentar dicha variable en el muestreo: cuales son los valores posibles, con qué frecuencia puede aparecer cada uno, si son valores muy parecidos entre sí, si son valores muy distintos, si tienden a estar agrupados alrededor de algún valor concreto … Debe quedar claro que aquí no nos referimos a un valor particular observado en una muestra concreta, sino a la colección de posibles valores susceptibles de ser observados en las distintas muestras que podrían obtenerse al llevar a cabo el muestreo. Volviendo a nuestro ejemplo, si decidimos tomar una muestra de 30 pacientes, hay muchísimas formas de elegir 30 pacientes de entre todos los afectados por la enfermedad, y por tanto hay muchas muestras posibles de 30 pacientes; en una de esas muestras la tasa de remision podria ser el 68%, en otra el 76%, en otra el 54%, …
- Simular la variable aleatoria significa utilizar R para generar valores de dicha variable con un comportamiento lo más parecido posible a lo que ocurriría en observaciones reales de la misma. Si, en nuestro ejemplo, suponemos que la tasa de remisión con el tratamiento empleado es del 70%, ello significa que para cada paciente individual, la probabilidad de que la enfermedad le remita es 0.7 y la probabilidad de que no le remita es 0.3. Lo que le ocurre a un individuo es muy fácil de simular en R. Una posible forma de hacerlo es la siguiente:

```
sample(c("remite","no remite"), size=1,prob=c(0.7,0.3))
```

```
## [1] "no remite"
```

Si queremos simular lo que le ocurre a 10 pacientes especificamos `size=10` e indicamos `replace=TRUE` lo que significa que cada valor ("remite" o "no remite") puede ocurrir varias veces en la muestra:

```
sample(c("remite","no remite"), size=10, prob=c(0.7,0.3), replace=TRUE)
```

```
##  [1] "remite"    "no remite" "remite"    "no remite" "no remite"
##  [6] "remite"    "remite"    "remite"    "remite"    "remite"
```

Si repetimos la simulación no obtendremos el mismo resultado, ya que representaría otra muestra de 10 pacientes distintos (que obviamente no producen los mismos valores que los 10 primeros):

```r
sample(c("remite","no remite"), size=10, prob=c(0.7,0.3), replace=TRUE)
```

```
##  [1] "remite"    "no remite" "no remite" "remite"    "remite"
##  [6] "remite"    "remite"    "remite"    "no remite" "no remite"
```

Podemos simplificar el código anterior, codificando como "1" la remisión y como "0" la no remisión:

```r
muestra=sample(c(1,0), size=10, prob=c(0.7,0.3), replace=TRUE)
muestra
```

```
##  [1] 1 0 0 0 0 1 0 1 1 1
```

De esta forma, la suma de los valores de la muestra nos proporcionaría justamente el número de personas de esta muestra en las que se ha producido la remisión:

```r
sum(muestra)
```

```
## [1] 5
```

y si dividimos por el tamaño muestral obtendríamos la tasa de remisión observada en esta muestra particular:

```r
sum(muestra)/10
```

```
## [1] 0.5
```

9.3 Simulación de la tasa de remisión

El procedimiento de simulación anterior, aunque es simple, se puede simplificar aún más si tenemos en cuenta que la variable en la que estamos interesados:

n_R = "Número de pacientes de una muestra de tamaño n en los que se produce la remisión."

es una variable aleatoria con distribución binomial de parámetros n y π, siendo π la tasa de remisión (en la práctica, desconocida) del tratamiento.

R dispone de una función específica para simular variables con distribución de probabilidad binomial; si la tasa de remisión real fuese del 70%, podemos simular una observación de n_R en una muestra de tamaño $n = 10$ como:

```r
nR=rbinom(1,10,0.7)
nR
```

```
## [1] 9
```

y la tasa de remisión en esta muestra sería:

```r
tasaMuestral=nR/10
tasaMuestral
```

```
## [1] 0.9
```

Si queremos que R simule el resultado de observar la tasa de remisión en 100 muestras, cada una de tamaño $n = 10$, con $\pi = 0.7$, bastará con ejecutar:

```r
n=10
pi=0.7
tasaMuestral=rbinom(100,n,pi)/n
tasaMuestral
```

```
##   [1] 0.6 0.7 0.9 0.7 0.8 0.6 0.5 0.7 0.7 0.7 0.7 0.6 0.6 1.0 0.9 0.8 0.6
##  [18] 0.8 0.6 0.6 0.4 0.4 0.8 0.7 0.5 0.8 0.7 0.7 0.4 0.5 0.6 0.7 0.5 0.9
##  [35] 0.9 0.8 0.7 0.7 0.8 0.9 1.0 0.7 0.7 0.6 0.7 0.7 0.6 0.6 0.8 0.7 0.5
##  [52] 0.5 0.7 0.5 0.8 0.7 0.8 0.7 0.8 0.8 0.6 0.6 0.6 0.6 0.7 0.7 1.0 0.4
##  [69] 0.9 0.7 0.9 0.4 0.6 0.7 0.7 0.7 0.9 0.7 0.9 0.9 0.7 0.6 0.7 0.8 0.3
##  [86] 0.8 0.7 0.6 0.7 0.6 0.6 0.7 0.8 0.7 0.7 0.6 0.9 0.5 0.7 0.6
```

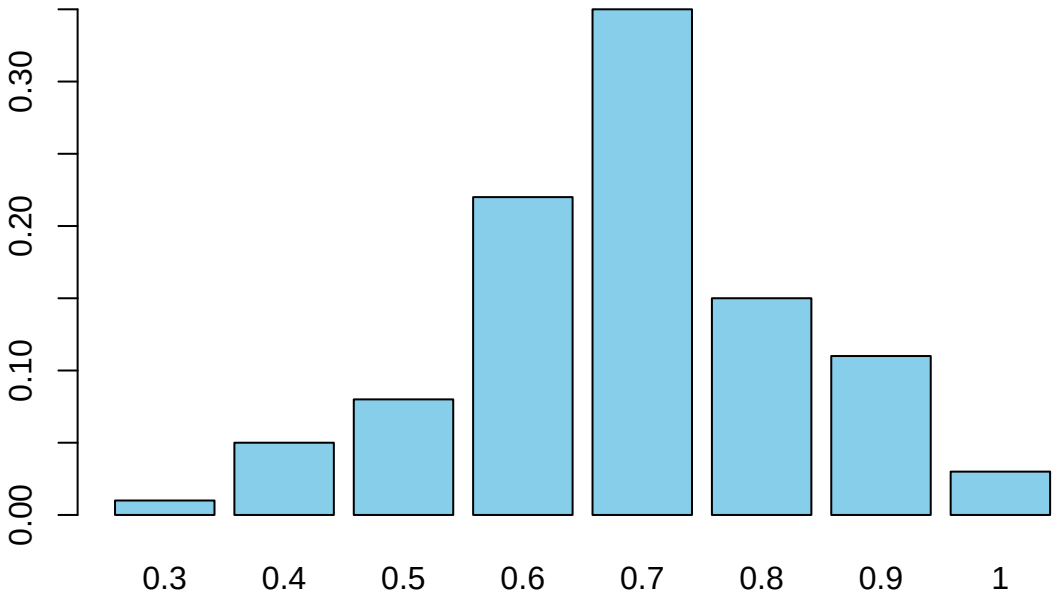
Podemos presentar estos resultados en una tabla:

```r
ttm=table(tasaMuestral)
library(pander)
pander(ttm)
```

0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
1	5	8	22	35	15	11	3

Es decir, de las 100 muestras, se ha observado una tasa de remisión de 0.3 en 1 de ellas, una tasa 0.4 en 5, 0.5 en 8, etc. En particular, la tasa 0.7 se ha observado en 35 muestras. Gráficamente:

```
barplot(prop.table(ttm),col="skyblue")
```

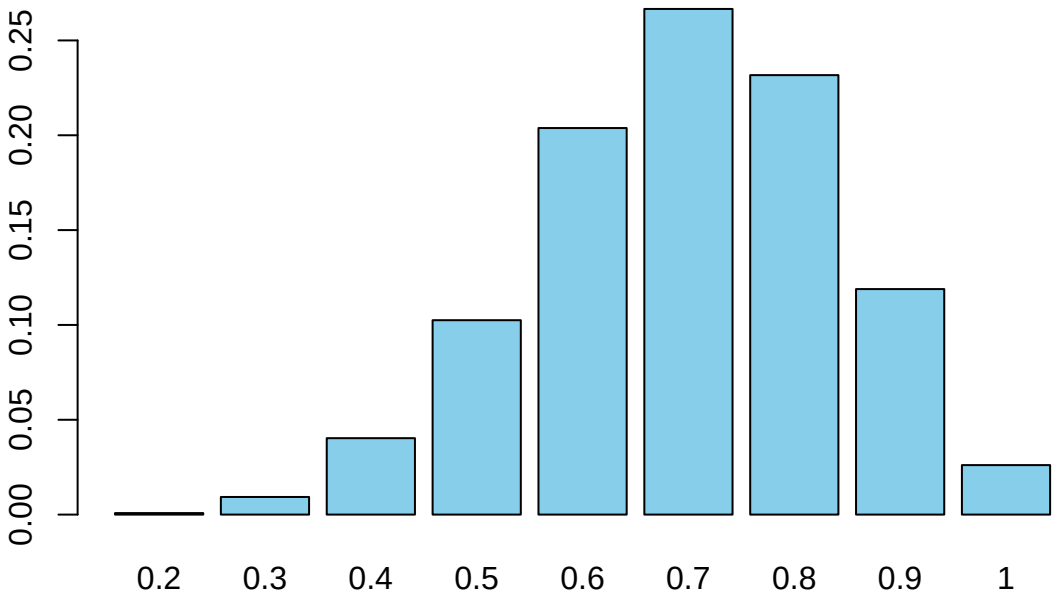


Podemos conseguir más información si en lugar de 100 muestras de tamaño 10, simulamos 10000 muestras de tamaño 10:

```
n=10
pi=0.7
tasaMuestral=rbinom(10000,n,pi)/n
pander(table(tasaMuestral))
```

0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
8	93	403	1025	2038	2666	2317	1189	261

```
barplot(prop.table(table(tasaMuestral)), col="skyblue")
```



¿Qué nos indican esta tabla y este gráfico? Básicamente que si la muestra disponible para estimar la prevalencia es de tamaño 10, aunque la mayor parte de las veces el valor estimado es 0.6, 0.7 ó 0.8, no resulta demasiado extraño que aparezcan estimaciones como 0.4, 0.9 o incluso 1. Si calculamos los percentiles 2.5 y 97.5, nos dan un intervalo dentro del cuál aparece el 95% de los valores de prevalencia estimados en las distintas muestras:

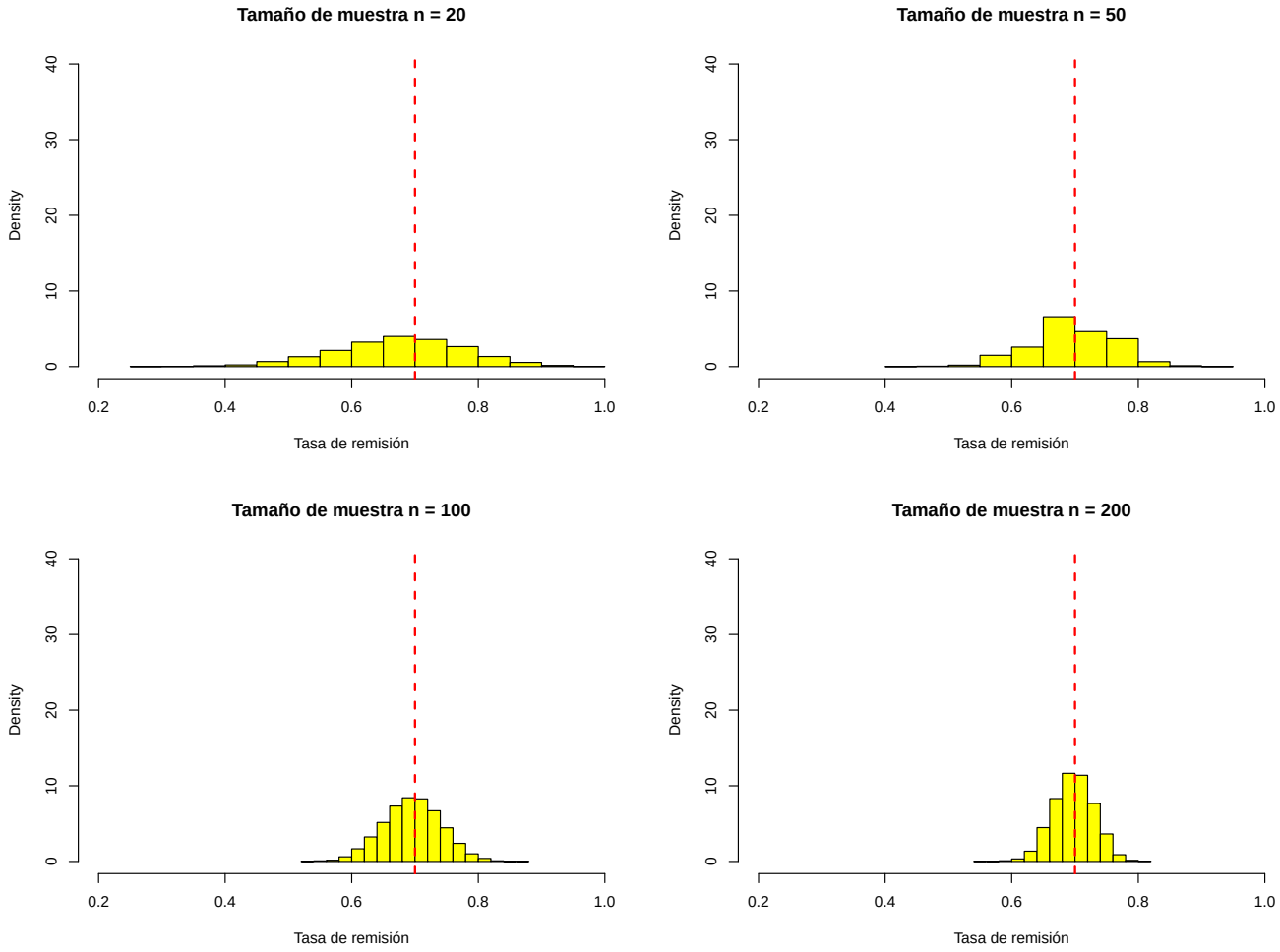
```
quantile(tasaMuestral, probs=c(0.025,0.975))
```

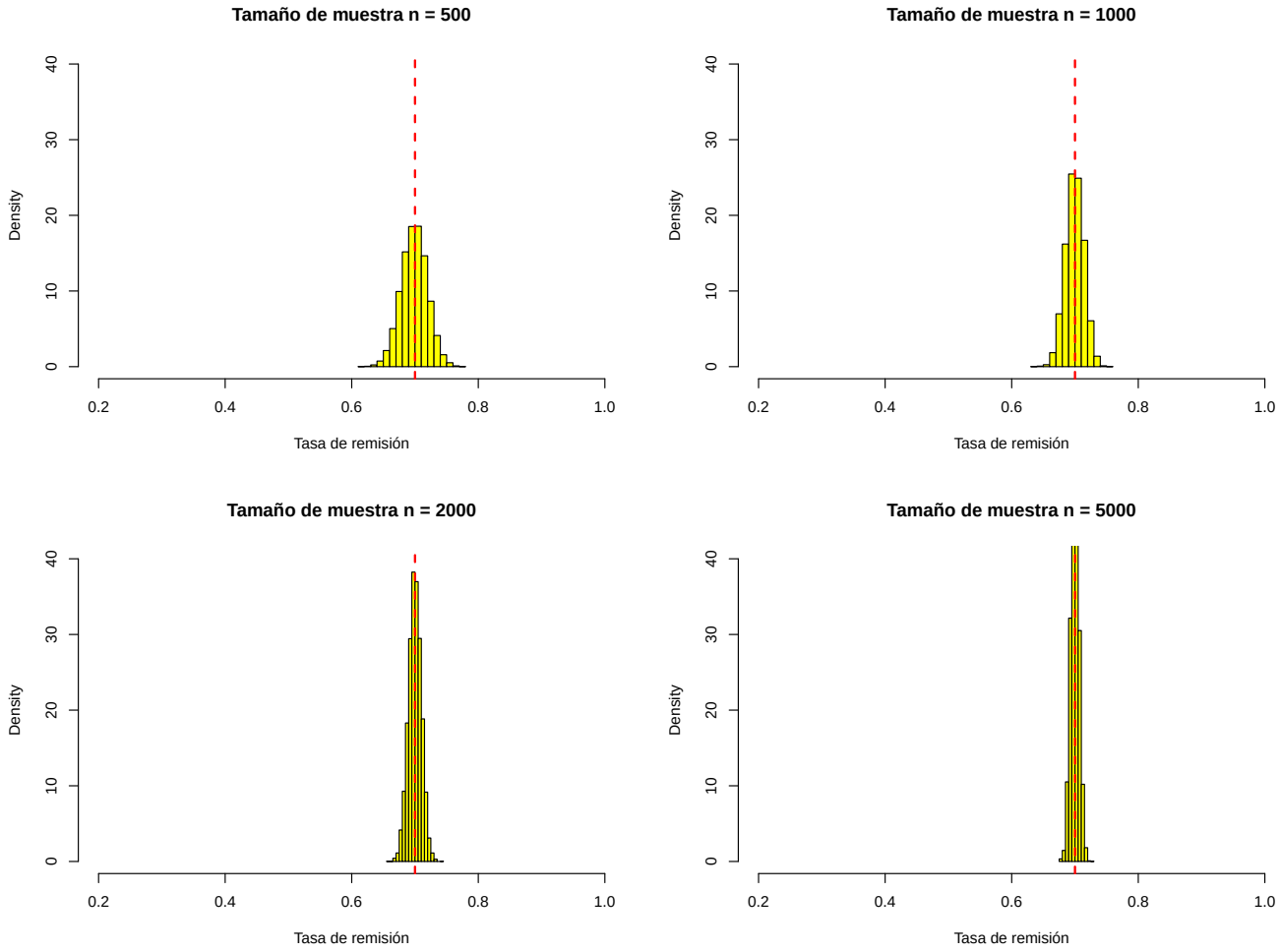
```
##  2.5% 97.5%
##   0.4   1.0
```

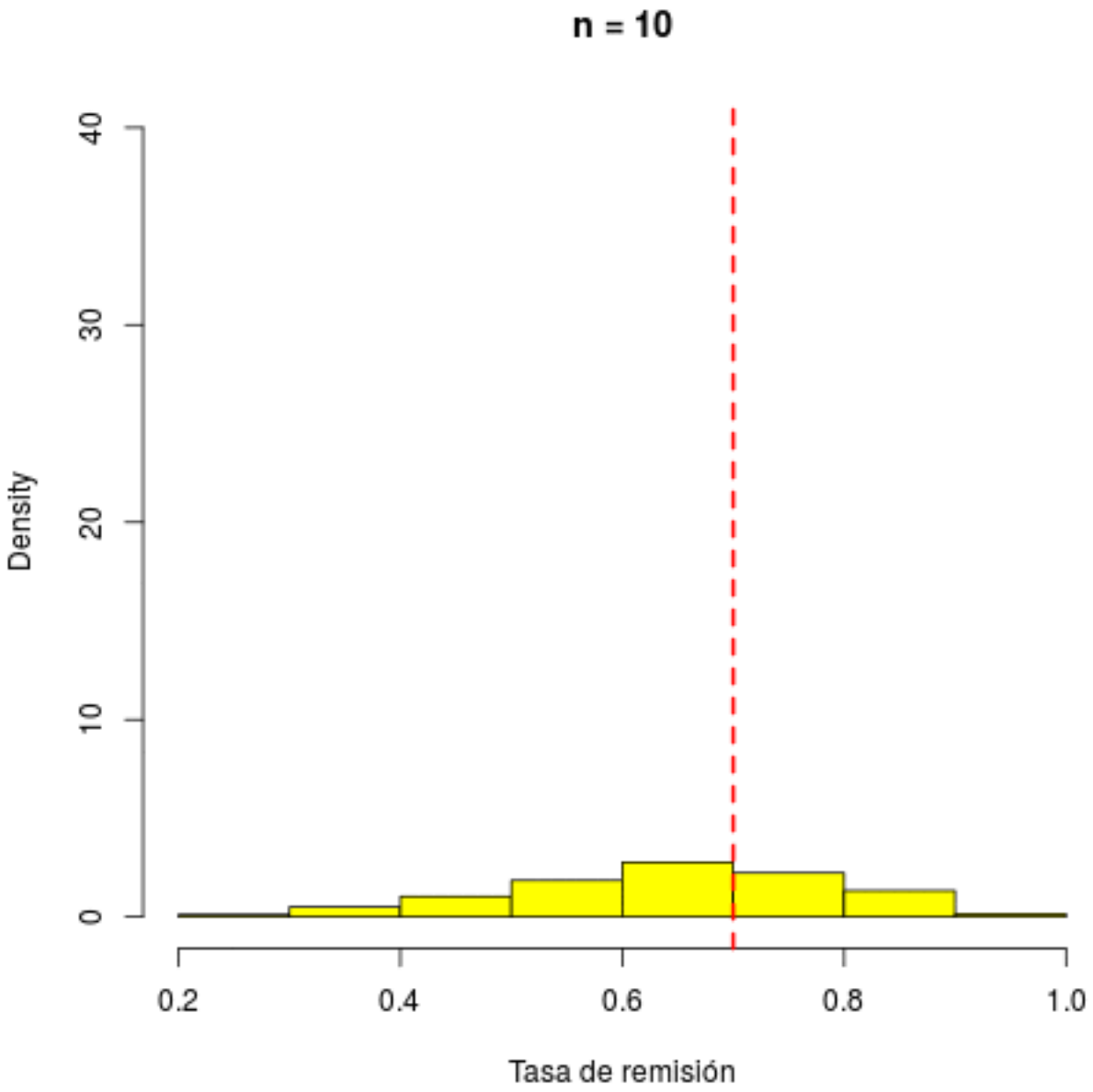
En otras palabras, esta simulación nos informa de que si la verdadera tasa de remisión fuese 0.7, una muestra de tamaño 10 el 95% de las veces produce valores estimados entre 0.4 y 1. Si pensamos que cuando tomamos una muestra no conocemos el verdadero valor de la tasa de remisión, estos resultados indican que si la tasa de remisión muestral que observamos fuese 0.4, sería perfectamente posible, incluso problable, que la tasa de remisión "real" fuese 0.7, lo que significa que nuestra estimación tendría un error de 0.3 unidades.

Veamos qué ocurre si tomamos muestras de tamaños 20, 50, 100, 200, 500, 1000, 2000 y 5000 (mostramos sólo los gráficos y una versión animada de éstos):

```
par(mfrow=c(2,2))
pi=0.7
for (n in c(20, 50, 100, 200, 500, 1000, 2000, 5000)){
  tasaMuestral=rbinom(10000,n,pi)/n
  hist(tasaMuestral, freq=FALSE, main=paste("Tamaño de muestra n =",n),
       xlim=c(0.2,1), ylim=c(0,40), col="yellow", xlab="Tasa de remisión")
  abline(v=0.7,col="red",lwd=2,lty=2)
}
```


Tamaño de muestra n = 20
Density
0
10
20
30
40
0.2
0.4
0.6
0.8
1.0
Tasa de remisión
Tamaño de muestra n = 50
Density
0
10
20
30
40
0.2
0.4
0.6
0.8
1.0
Tasa de remisión
Tamaño de muestra n = 100
Density
0
10
20
30
40
0.2
0.4
0.6
0.8
1.0
Tasa de remisión
Tamaño de muestra n = 200
Density
0
10
20
30
40
0.2
0.4
0.6
0.8
1.0
Tasa de remisión


Tamaño de muestra n = 500
Density
0
10
20
30
40
0.2
0.4
0.6
0.8
1.0
Tasa de remisión
Tamaño de muestra n = 1000
Density
0
10
20
30
40
0.2
0.4
0.6
0.8
1.0
Tasa de remisión
Tamaño de muestra n = 2000
Density
0
10
20
30
40
0.2
0.4
0.6
0.8
1.0
Tasa de remisión
Tamaño de muestra n = 5000
Density
0
10
20
30
40
0.2
0.4
0.6
0.8
1.0
Tasa de remisión



Estas figuras muestran que a medida que aumenta el tamaño de la muestra:

- Los valores muestrales de la tasa de remisión están siempre en torno al verdadero valor de la tasa de remisión (0.7).
- A medida que aumenta el tamaño de la muestra, la dispersión es menor, esto es, los valores de tasa de remisión muestrales aparecen cada vez más concentrados en torno al verdadero valor de la tasa de remisión. Ello nos indica que cuando por fin hagamos el muestreo real (no las simulaciones), cualquiera que sea la muestra particular que obtengamos finalmente, podemos estar bastante seguros de que, si la muestra es lo suficientemente grande, el valor estimado de la tasa de remisión será muy próximo al valor real de dicha tasa.

Para tener una idea de la proximidad entre las estimaciones y el verdadero valor de la tasa de remisión, podemos calcular los percentiles 2.5 y 97.5 para las estimaciones obtenidas con los distintos tamaños de muestra:

```
tab=NULL
for (n in c(20, 50, 100, 200, 500, 1000, 2000, 5000)){
  tasaMuestral=rbinom(10000,n,pi)/n
  tab=rbind(tab,c(n,quantile(tasaMuestral, probs=c(0.025,0.975))))
}
tab
```

```
##              2.5%  97.5%
## [1,]   20 0.5000 0.9000
## [2,]   50 0.5800 0.8200
## [3,]  100 0.6100 0.7900
## [4,]  200 0.6400 0.7600
## [5,]  500 0.6600 0.7400
## [6,] 1000 0.6710 0.7290
## [7,] 2000 0.6800 0.7195
## [8,] 5000 0.6872 0.7128
```

Así pues:

- Si la muestra es de tamaño 20, el 95% de las estimaciones muestrales están entre 0.5 y 0.9; dicho de otra forma, el 95% de las estimaciones realizadas con una muestra de tamaño 20 tendrán un error máximo de ± 0.2.
- Si la muestra es de tamaño 50, el 95% de las estimaciones muestrales están entre 0.58 y 0.82; dicho de otra forma, el 95% de las estimaciones realizadas con una muestra de tamaño 20 tendrán un error máximo de ± 0.12.
- Si la muestra es de tamaño 100, el 95% de las estimaciones muestrales están entre 0.61 y 0.79; dicho de otra forma, el 95% de las estimaciones realizadas con una muestra de tamaño 20 tendrán un error máximo de ± 0.09.
- Si la muestra es de tamaño 500, el 95% de las estimaciones muestrales están entre 0.66 y 0.74; dicho de otra forma, el 95% de las estimaciones realizadas con una muestra de tamaño 20 tendrán un error máximo de ± 0.04.
- Si la muestra es de tamaño 1000, el 95% de las estimaciones muestrales están entre 0.671 y 0.729; dicho de otra forma, el 95% de las estimaciones realizadas con una muestra de tamaño 20 tendrán un error máximo de ± 0.029.

Por tanto:

- Si, siendo la verdadera tasa de remisión 0.7, un error de 0.2 (un 20%) es admisible, una muestra de tamaño 20 es suficiente.
- Si queremos que el error máximo de estimación sea de 0.04 (un 4%) la muestra debe ser de tamaño 500.

Pueden ejecutarse simulaciones como las anteriores variando los valores de π y n para determinar el valor de n adecuado que permitiría estimar π con el error que deseemos.