Práctica 1. Exploración de datos I. Soluciones.

Universidad de Alcalá. Curso 2023-24.

Estadística (650008). Grado en Biología sanitaria.

Actualizado: 2023-10-05

Introducción

Trabajaremos con una tabla de datos procedente de una muestra de 1000 mujeres que participaron en un estudio sobre osteoporosis. El fichero contiene algunas variables auxiliares en las columnas iniciales, pero nosotros nos vamos a fijar en estas.

- edad (en años).
- peso (en kg).
- talla (altura en cm).
- imc (índice de masa corporal)
- bua (resultado de la exploración densitométrica)
- clasific (normal / osteopenia / osteoporosis)
- menarqui (edad primera menstruación, en años)
- edad_menop (edad inicio menopausia, en años)
- menopausia (sí, no)
- tipo de menopausia
- nivel educativo

Organiza tu entorno de trabajo

- Para leer los datos, copia, pega en tu script y ejecuta este código

```
mi_url = "https://marcos-marva.web.uah.es/CursoSanitaria/practicas/datos/osteoporosis.csv"
osteoporosis = read.table(file = mi_url,
                          sep = "\t", dec = ",", header = TRUE)
```

- En la próxima práctica aprenderemos cómo leer datos de ficheros.

En lo que sigue, recuerda que para acceder a los datos en las columnas del data.frame que has creado en el paso anterior puedes utilizar la notación `nombre_tabla$nombre_variable`.

Ejercicio 1

Para la variable `tipo_men`:

1. **¿De qué tipo es la variable?** El resumen

```
summary(osteoporosis$tipo_men)
```

```
##    Length     Class      Mode
##      1000 character character
```

muestra que es cualitativa nomial, no hay orden alguno en los niveles del factor.

2. **Usa la función `unique()` para determinar cuántos valores distintos toma**

```r
length(unique(osteoporosis$tipo_men))
```

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

3. **Construye las tablas de frecuencias absolutas y relativas.**

```r
table(osteoporosis$tipo_men)
```

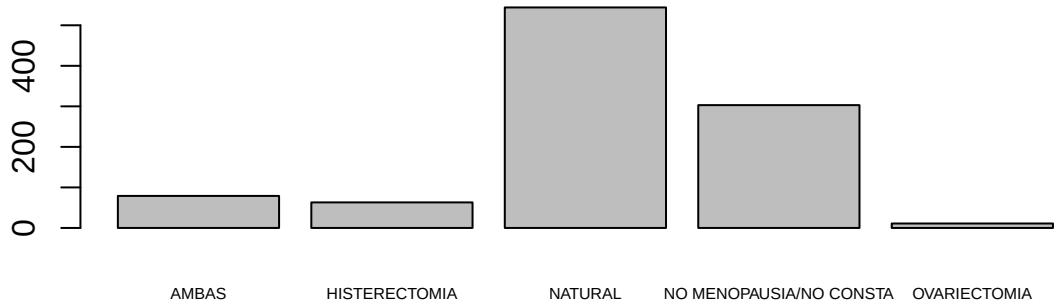
```
##
##                   AMBAS           HISTERECTOMIA                 NATURAL
##                      79                      63                     544
## NO MENOPAUSIA/NO CONSTA            OVARIECTOMIA
##                     303                      11
```

```r
table(osteoporosis$tipo_men)/nrow(osteoporosis)
```

```
##
##                   AMBAS           HISTERECTOMIA                 NATURAL
##                   0.079                   0.063                   0.544
## NO MENOPAUSIA/NO CONSTA            OVARIECTOMIA
##                   0.303                   0.011
```

4. **Para esta variable, ¿tienen sentido las tablas de frecuencias acumuladas? En caso afirmativo, construyelas.** Aunque técnicamente se puede calcular, no aporta mucha información porque los valores que se acumulan (suman) no se refieren a valores de la variable ordenados.
5. **Representa las frecuencias absolutas en un diagrama de barras.**

```r
barplot(table(osteoporosis$tipo_men), cex.names = .5)
```



6. **¿Qué medida de centralización usarías? Calcula su valor.** La moda, se puede obtener a partir de la tabla de frecuencias absolutas

Ejercicio 2

Para la variable `nivel_ed`:

1. **¿De qué tipo es la variable?** Vamos a explorar la variable

```r
summary(osteoporosis$nivel_ed)
```

```
##    Length     Class      Mode
##      1000 character character
```

y observamos que es cualitativa ordenada. 2. **¿Cuántos valores distintos toma?**

```r
length(unique(osteoporosis$nivel_ed))
```

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

3. **Construye las tablas de frecuencias absolutas y relativas.**

```r
# frecuencias absolutas
table(osteoporosis$nivel_ed)
```

```
##
##               PRIMARIOS PRIMARIOS SIN FINALIZAR             SECUNDARIOS
##                     467                     212                     150
##            SIN ESTUDIOS              SUPERIORES
##                     122                      49
```

Dividir cada frecuencia entre el número de filas. La función `nrow` las calcula por ti y, si luego cambias las dimensiones de la tabla (imagina que añades/eliminas datos) el cálculo se actualiza automáticamente

```r
# frecuencias relativas
table(osteoporosis$nivel_ed)/nrow(osteoporosis)
```

```
##
##               PRIMARIOS PRIMARIOS SIN FINALIZAR             SECUNDARIOS
##                   0.467                   0.212                   0.150
##            SIN ESTUDIOS              SUPERIORES
##                   0.122                   0.049
```

4. **Construye ahora las tablas de frecuencias absolutas y relativas acumuladas.** La función `cumsum` calcula la suma acumulada de cada valor de la variable

```r
# frecuencias absolutas acumuladas
cumsum(table(osteoporosis$nivel_ed))
```

```
##               PRIMARIOS PRIMARIOS SIN FINALIZAR             SECUNDARIOS
##                     467                     679                     829
##            SIN ESTUDIOS              SUPERIORES
##                     951                    1000
```

```r
# frecuencias relativas acumuladas
cumsum(table(osteoporosis$nivel_ed))/nrow(osteoporosis)
```

```
##               PRIMARIOS PRIMARIOS SIN FINALIZAR             SECUNDARIOS
##                   0.467                   0.679                   0.829
##            SIN ESTUDIOS              SUPERIORES
##                   0.951                   1.000
```

¿Tienen sentido esas tablas? Las frecuencias acumuladas tienen más sentido cuando los valores de la variable están ordenados de menor a mayor (fíjate en que ahora están ordenados alfabéticamente, es lo que hace `r` por defecto). Ahora podrías decir, por ejemplo (segunda tabla) que el 82.9% de las mujeres de la muestra tienen estudios secundarios, primarios, o primarios sin finalizar, pero NO puedes afirmar que ese número de mujeres tenga, como mucho estudios secundarios, porque no incluye a los individuos sin estudios.

Ejercicio 3

Para la variable `menarqui`:

1. **¿De qué tipo es la variable?** Vamos a explorar la variable

```
summary(osteoporosis$menarqui)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   12.00   13.00   12.71   14.00   17.00
```

y observamos que es cuantitativa. Se trata de tiempo (en años), que en principio es continua. Sin embargo,

```
unique(osteoporosis$menarqui)
```

```
##  [1] 12 13 14 10 11 15  9 16 17  8
```

resulta que toma pocos valores y todos son números enteros, por lo que optamos por tratarla como una variable discreta.

2. **1. Determina cuántos valores diferentes toma** Igual que hicimos antes,

```
length(unique(osteoporosis$menarqui))
```

```
## [1] 10
```

3. **Calcula el recorrido.** La variable variable varía entre los valores

```
range(osteoporosis$menarqui)
```

```
## [1]  8 17
```

y su recorrido es

```
max(osteoporosis$menarqui) - min(osteoporosis$menarqui)
```

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

4. ** Calcula la media y la desviación típica muestral.** Podemos hacer

```
mean((osteoporosis$menarqui))
```

```
## [1] 12.707
```

y, para la desviación típica muestral

```
sd(osteoporosis$menarqui)
```

```
## [1] 1.552921
```

ten en cuenta que R da, por defecto, la desviación `típica muestral`; de hecho, estás trabajando con una muestra. Si quieres obtener la desviación típica poblacional:

```
media_menarq = mean(osteoporosis$menarqui)
sqrt(sum((osteoporosis$menarqui - media_menarq)^2)/nrow(osteoporosis))
```

```
## [1] 1.552144
```

verás que en este caso la diferencia es pequeña, porque se divide o bien entre 1000 o bien entre 999. Para tamaños muestrales más pequeños, la diferencia sí es apreciable.

5. **Tabla de frecuencias absolutas, relativas, acumuladas y relativas acumuladas.** Podemos reciclar el código de la anterior pregunta

```
table(osteoporosis$menarqui)
```

```
##
##   8   9  10  11  12  13  14  15  16  17
##   1  14  56 159 214 241 211  65  29  10
```

La primera fila indica los valores de la variable y la segunda su frecuencia (en este caso, absoluta). En cuanto a la tabla de frecuencias relativas

```
table(osteoporosis$menarqui)/nrow(osteoporosis)
```

```
##
##     8     9    10    11    12    13    14    15    16    17
## 0.001 0.014 0.056 0.159 0.214 0.241 0.211 0.065 0.029 0.010
```

nos indica el peso (tanto por uno) que tiene cada uno de los valores en la muestra. Por ejemplo, la edad de 13 años tiene una frecuencia relativa de 0.241, es decir, se corresponde con un 24.1% de la muestra.

Las tablas de frecuencias acumuladas indican cuántos individuos (frec absolutas) o qué proporción de ellos (frec relativas) toman valores menores o iguales que uno dado.

```
cumsum(table(osteoporosis$menarqui))
```

```
##    8    9   10   11   12   13   14   15   16   17
##    1   15   71  230  444  685  896  961  990 1000
```

```
cumsum(table(osteoporosis$menarqui))/nrow(osteoporosis)
```

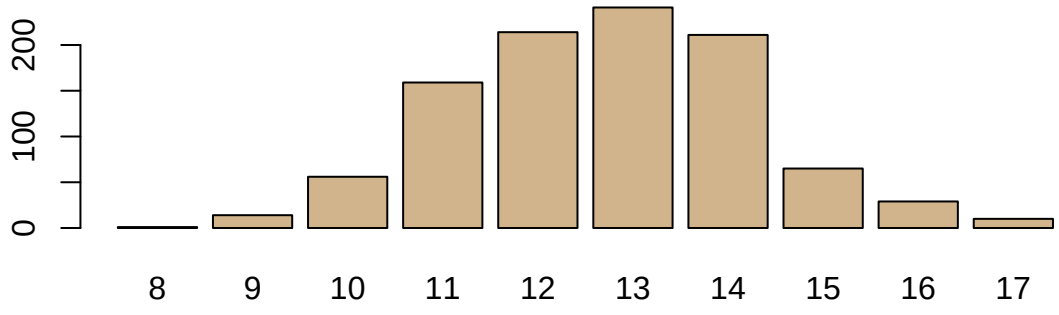
```
##     8     9    10    11    12    13    14    15    16    17
## 0.001 0.015 0.071 0.230 0.444 0.685 0.896 0.961 0.990 1.000
```

Por ejemplo:

- 444 mujeres tenían 12 o menos años.
- Un 44.4% de las mujeres tenían 12 o menos años (esa "coincidencia" se debe a que el tamaño muestral es 1000, la frecuencia absoluta.

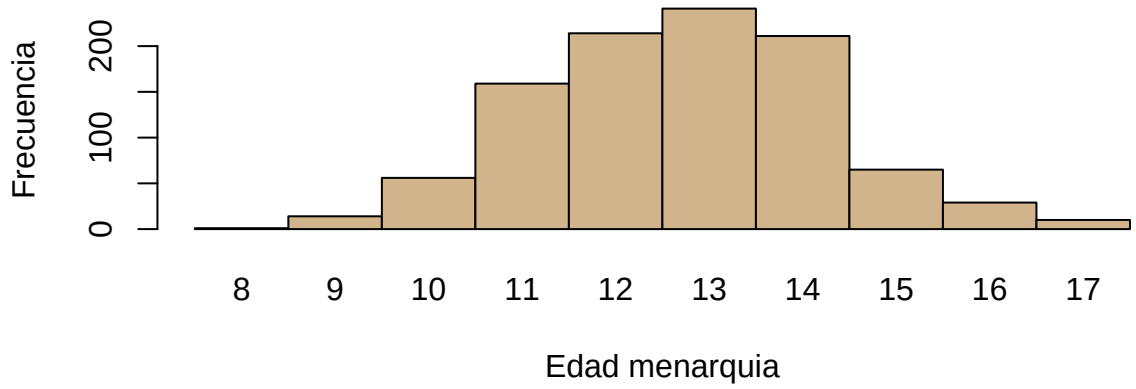
6. **Representa las frecuencias absolutas con un gráfico.** Al tratarla como una variable discreta, lo indicado es un diagrama de barras (bar plot, en inglés)

```
barplot(table(osteoporosis$menarqui), col = "tan")
```



Ten en cuenta que, como se trata de una variable continua, podríamos eliminar los huecos que hay entre las barras:

```
barplot(table(osteoporosis$menarqui), col = "tan", space = 0,
        xlab = "Edad menarquia", ylab = "Frecuencia")
```



Ejercicio 4

Para la variable `imc`:

Casi todo se puede resolver reciclando las órdenes usadas antes

1. **¿De qué tipo es la variable?** Las unidades (kg/m^2) sugieren que es continua. Observar los 10 primeros valores que toma la variable

```
head(osteoporosis$imc, 10)
```

```
##  [1] 24.80 22.94 25.64 30.09 22.72 21.97 33.18 22.87 24.23 28.83
```

corrobora que se trata de una variable continua.

2. **¿Cuántos valores distintos toma?**

```
length(unique(osteoporosis$imc))
```

```
## [1] 643
```

3. **Recorrido.**

```
range(osteoporosis$imc)
```

```
## [1] 17.21 48.39
```

El recorrido abarca

```
max(osteoporosis$imc) - min(osteoporosis$imc)
```

```
## [1] 31.18
```

unidades.

4. **Calcula el `imc` medio y la desviación típica muestral de las edades.**

Usamos ahora las funciones `mean` y `sd`

```
mean(osteoporosis$imc)
```

```
## [1] 28.10776
```

```
sd(osteoporosis$imc)
```

```
## [1] 4.717925
```

5. **Tabla de frecuencias absolutas.**

```
table(osteoporosis$imc)
```

```
##
## 17.21 17.58 18.16 18.82 19.13 19.15  19.2 19.72    20 20.07 20.09 20.31 20.32
##     1     1     1     1     1     1     1     1     1     1     1     1     1
##  20.4 20.43 20.53 20.54 20.58  20.6 20.69  20.7 20.72 20.78 20.83 20.89 20.96
##     1     1     1     1     1     1     1     1     1     1     1     1     2
## 21.05 21.08 21.09 21.17 21.23 21.25 21.27 21.29  21.3 21.34 21.36 21.41 21.44
##     1     2     1     1     1     1     1     1     3     2     1     2     2
## 21.45 21.48 21.51 21.52 21.55 21.56  21.6 21.61 21.63 21.67 21.71 21.75 21.77
##     2     1     1     1     1     1     1     1     1     1     1     1     1
## 21.78 21.88 21.91 21.93 21.94 21.97 22.03 22.04 22.06 22.07 22.15 22.17 22.21
##     2     1     2     1     2     2     3     2     5     2     1     1     1
## 22.23 22.31 22.35 22.37 22.38 22.39  22.4 22.41 22.43 22.44 22.48 22.49  22.5
##     2     1     1     1     2     1     1     1     4     1     1     2     1
## 22.51 22.59  22.6 22.63 22.66 22.67 22.68 22.72 22.74 22.76 22.83 22.85 22.86
##     1     1     2     1     3     1     1     2     1     1     1     1     3
## 22.87 22.89 22.94 22.95 23.01 23.03 23.05 23.11 23.15 23.19 23.22 23.24 23.25
##     1     1     5     2     1     1     2     1     2     3     1     1     1
## 23.28 23.29 23.31 23.32 23.33 23.34 23.37 23.42 23.43 23.44 23.47  23.5 23.51
##     1     1     4     1     1     3     1     4     4     2     1     1     2
## 23.53 23.59 23.61 23.62 23.63 23.68 23.73  23.8 23.81 23.83 23.87  23.9 23.92
##     2     2     1     2     1     1     5     1     1     3     1     1     1
## 23.94 23.98    24 24.01 24.02 24.03 24.06 24.09 24.12 24.13 24.15 24.16 24.17
##     3     1     1     4     1     3     2     2     1     1     1     1     1
## 24.23 24.24 24.25  24.3 24.31 24.34 24.35 24.38 24.39  24.4 24.44 24.45 24.46
##     2     4     1     1     1     1     3     1     1     1     1     2     2
## 24.51 24.52 24.53 24.54 24.65 24.67 24.74 24.75 24.76 24.77 24.78 24.79  24.8
##     1     1     1     1     1     2     1     1     1     1     2     1     3
## 24.84 24.86 24.88 24.95 24.96 24.97 24.99    25 25.03 25.07 25.08 25.09  25.1
##     3     1     3     1     1     2     2     3     1     2     2     2     3
## 25.11 25.21 25.22 25.24 25.26 25.27 25.28  25.3 25.31 25.32 25.34 25.35 25.36
##     1     1     1     2     1     1     1     1     1     2     1     2     1
## 25.39  25.4 25.44 25.49 25.53 25.56 25.57 25.58 25.59 25.63 25.64 25.65 25.67
##     2     1     4     1     2     2     1     1     5     6     5     2     2
## 25.71 25.72 25.75 25.78  25.8 25.81 25.83 25.84 25.85 25.88 25.89 25.91 25.92
##     2     1     1     5     1     1     1     2     5     2     2     1     1
## 25.95 25.96 25.97 25.98 26.01 26.02 26.03 26.04 26.05 26.06 26.11 26.14 26.17
##     1     2     7     1     1     1     2     1     1     1     1     1     4
## 26.19  26.2 26.22 26.24 26.29  26.3 26.31 26.32 26.35 26.37  26.4 26.42 26.43
##     2     1     5     1     2     2     3     1     2     3     2     1     1
## 26.44 26.48 26.49  26.5 26.57 26.58 26.64 26.66 26.67  26.7 26.71 26.72 26.74
##     1     1     2     4     2     1     3     1     3     1     2     1     1
## 26.75 26.78 26.81 26.84 26.85  26.9 26.91 26.95 26.98 26.99 27.03 27.04 27.05
##     4     1     1     1     1     3     3     1     1     1     1     2     3
```

```
## 27.06 27.07  27.1 27.11 27.12 27.13 27.14 27.16 27.18 27.19  27.2 27.21 27.24
##     2     1     2     1     1     1     1     1     2     1     1     1     2
## 27.26 27.27 27.29 27.31 27.32 27.33 27.34 27.36 27.38 27.39 27.41 27.43 27.44
##     1     1     2     1     1     1     6     1     1     1     4     4     1
## 27.46 27.47 27.48 27.51 27.53 27.54 27.55 27.56 27.58 27.59 27.62 27.63 27.64
##     1     3     2     4     1     1     1     1     3     2     1     1     2
## 27.68 27.69  27.7 27.73 27.74 27.75 27.76 27.77 27.78 27.79 27.82 27.83 27.84
##     3     2     2     2     1     1     1     2     2     1     1     2     1
## 27.85 27.89 27.93 27.94 27.96 27.97 27.99    28 28.01 28.04 28.07 28.08 28.12
##     3     2     2     1     1     1     2     1     3     3     1     1     2
## 28.13 28.15 28.19  28.2 28.23 28.25  28.3 28.31 28.35 28.36 28.38  28.4 28.41
##     2     1     1     2     1     2     3     1     3     1     1     1     2
## 28.42 28.44 28.46 28.48 28.51 28.52 28.55 28.57 28.58 28.62 28.63 28.64 28.65
##     1     3     2     1     1     1     2     1     1     1     1     1     1
## 28.67 28.69 28.72 28.73 28.76 28.77  28.8 28.83 28.84 28.86 28.88 28.89 28.91
##     3     1     1     2     5     1     2     3     1     1     1     1     1
## 28.93 28.96    29 29.01 29.03 29.04 29.05 29.06 29.07 29.09 29.14 29.22 29.24
##     3     2     1     3     1     1     3     1     2     3     4     2     4
## 29.27  29.3 29.34 29.38 29.41 29.43 29.44 29.49 29.52 29.53 29.55 29.56 29.59
##     1     3     1     2     3     3     1     1     2     1     3     1     1
##  29.6 29.62 29.63 29.64 29.67 29.69 29.74 29.82 29.86  29.9 29.91 29.92 29.94
##     1     4     1     1     2     1     2     3     3     4     2     1     1
## 29.97    30 30.08 30.09 30.11 30.16 30.18 30.24 30.26 30.27  30.3 30.38 30.39
##     2     3     1     1     1     1     1     1     2     1     2     1     2
##  30.4 30.41 30.42 30.43 30.44 30.46 30.47 30.48 30.49 30.57 30.58 30.59 30.61
##     1     2     1     2     2     1     1     2     1     1     1     1     2
## 30.63 30.66 30.67  30.7 30.73 30.76 30.77 30.78 30.79 30.82 30.84 30.85 30.86
##     3     1     1     1     1     2     1     2     1     2     1     1     2
##  30.9 30.99 31.02 31.05 31.08 31.12 31.14 31.16 31.18  31.2 31.22 31.23 31.24
##     1     1     1     1     2     1     2     1     1     2     1     3     3
## 31.25  31.3 31.33 31.35 31.38 31.48 31.53 31.56 31.58 31.59 31.61 31.62 31.63
##     2     1     1     1     1     1     1     1     1     1     1     2     5
## 31.64 31.65 31.81 31.82 31.83 31.84 31.85 31.88 31.93 31.97 31.98    32 32.01
##     4     1     1     1     1     2     2     1     1     1     1     1     2
## 32.02 32.03 32.05 32.16 32.19 32.21 32.24 32.26  32.3 32.31 32.32 32.37 32.39
##     1     2     5     1     1     2     1     1     2     1     2     1     1
## 32.41 32.45 32.46 32.47 32.68 32.69 32.77 32.78 32.86 32.89 32.98 33.03 33.05
##     1     2     3     1     1     1     1     1     1     2     1     1     1
## 33.11 33.18  33.2 33.24 33.25 33.29  33.3 33.31 33.33 33.46 33.47  33.5 33.51
##     1     1     2     1     2     1     1     1     4     1     1     1     2
## 33.56 33.59 33.65 33.67 33.69 33.73 33.74 33.76 33.77 33.89  33.9 33.98    34
##     1     1     2     1     2     1     1     1     1     1     1     1     1
## 34.11 34.13 34.21 34.23 34.24 34.25 34.28 34.38 34.47 34.48  34.6 34.61 34.63
##     2     1     2     1     2     1     1     2     1     3     1     1     1
## 34.65 34.68 34.72 34.82 34.84 34.87 34.93 34.96 34.97    35 35.16 35.21 35.31
##     1     1     1     1     1     1     2     3     1     1     1     1     1
## 35.34 35.38 35.42 35.44  35.5 35.52 35.56 35.57 35.59 35.67 35.69  35.7  35.8
##     1     2     1     2     1     1     2     1     1     1     1     1     1
## 35.85    36 36.03  36.1 36.11 36.14  36.2 36.26 36.33  36.4 36.52 36.65 36.73
##     1     1     2     1     2     1     1     1     1     1     4     1     1
## 36.92 36.98 37.11 37.22 37.25 37.39 37.44  37.5 37.52 37.53 37.61 37.66 37.72
##     1     1     1     1     1     1     1     1     1     1     1     1     2
## 38.05 38.34 38.46 38.57 38.62 38.71 38.76 38.82 38.96 39.35 39.76 39.91 40.03
##     1     1     1     1     1     1     1     1     1     1     1     1     1
```

```
## 40.04 40.08 40.25 40.31 40.37  40.8  41.1  41.4  41.5 42.02    43 43.15 43.29
##     1     1     1     1     1     1     1     1     1     1     1     1     1
## 43.62 44.79 45.45 46.47 46.77 48.39
##     1     1     1     1     1     1
```

Como los valores apenas se repiten, esta tabla no aporta información. Por eso hay que agrupar los datos en clases.

6. **Agrupa la variable en cinco clases de la misma longitud.** Para definir 5 clases (intervalos) necesitamos conocer 6 puntos (dibújalo y te convencerás). La siguiente función divide el recorrido de la variable en 5 clases de igual longitud y calcula a qué clase corresponde cada observación

```
clases.imc = cut(osteoporosis$imc, breaks = 5, include.lowest = TRUE)
```

El último argumento se usa para incluir el valor más pequeño (la clase más pequeña es un intervalo cerrado por ambos lados). Compara los primeros valores de `imc` con los primeros valores de `clases.imc` y verás la correspondecia

```
head(osteoporosis$imc)
```

```
## [1] 24.80 22.94 25.64 30.09 22.72 21.97
```

```
head(clases.imc)
```

```
## [1] (23.4,29.7] [17.2,23.4] (23.4,29.7] (29.7,35.9] [17.2,23.4] [17.2,23.4]
## Levels: [17.2,23.4] (23.4,29.7] (29.7,35.9] (35.9,42.2] (42.2,48.4]
```

Ahora vasta hacer una tabla con `clases.imc`

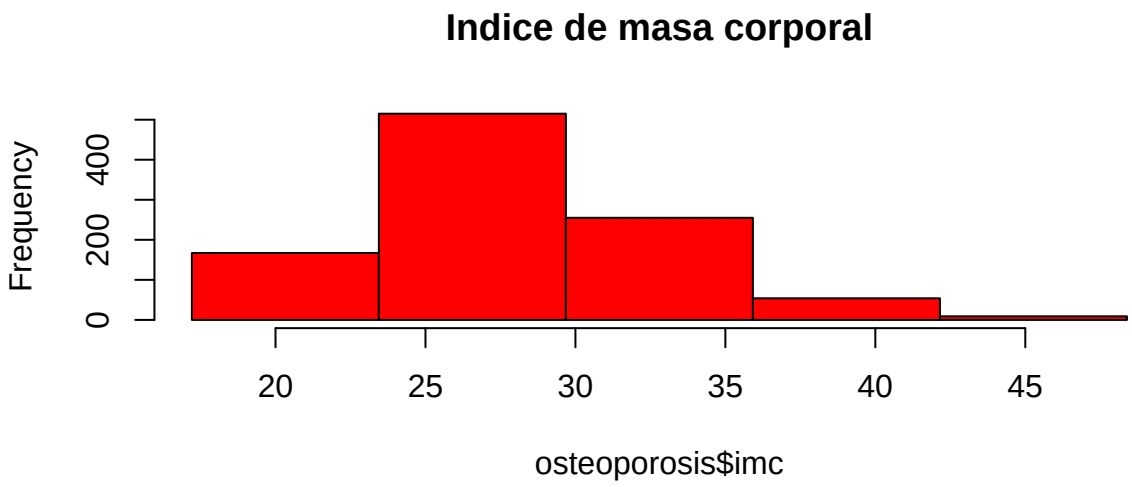
```
table(clases.imc)
```

```
## clases.imc
## [17.2,23.4] (23.4,29.7] (29.7,35.9] (35.9,42.2] (42.2,48.4]
##         167         515         255          54           9
```

Esto tiene más sentido que con los valores sin agrupar

7. **Representa esas clases mediante un histograma.** Fíjate en que tenemos que indicar los puntos de corte de las clases. En este caso, el significado del argumento `break` es diferente del de que tiene en la función `cut` (ver ejercicio 2, apartado 6). Para indicar lospuntos decortegeneramos una sucesión con `seq()` que consta de 6 puntos, empezando y terminando en el mínimo y el máximo de `imc`

```
hist(osteoporosis$imc,
     breaks = seq(from = min(osteoporosis$imc), to = max(osteoporosis$imc), length.out = 6),
     col = "red", main = "Indice de masa corporal")
```



8. **¿Qué porcentaje de los datos pertenece a cada clase?** Se trata de calcular la frecuencia relativa y multiplicar por 100

```
table(clases.imc)/nrow(osteoporosis)*100
```

```
## clases.imc
## [17.2,23.4] (23.4,29.7] (29.7,35.9] (35.9,42.2] (42.2,48.4]
##        16.7        51.5        25.5         5.4         0.9
```

9. **¿Entre qué valores se mueve el 80% central de la muestra?** La pregunta es equivalente a determinar los percentiles 10 y 90:

```
quantile(osteoporosis$imc, probs = c(.1, .9))
```

```
##    10%    90%
## 22.476 34.480
```

10. **Calcula los cuartiles y el boxplot de esta variable e interpretalo** EL valor de los cuartiles lo obtenemos con

```
quantile(osteoporosis$imc)
```

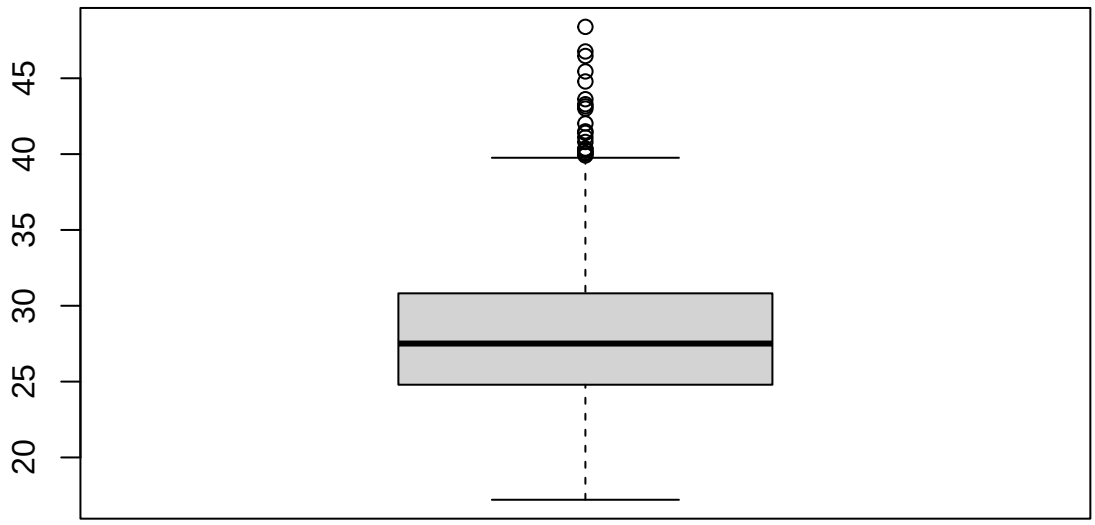
```
##      0%     25%     50%     75%    100%
## 17.2100 24.7975 27.5100 30.8200 48.3900
```

Así, el 25% de los individuos tiene un `imc` inferior a 24.8 o superior a 30.82 (por encima del tercer cuartil); la mitad de ellos está por debajo de 27.51 y la otra mitad por encima.

Se aprecia que la dispersión de la variable es menor en la mitad de individuos con menor `imc` (por debajo de la mediana) que en la mitad que más `imc` tiene (por encima de la mediana)

En cuanto al boxplot

```r
boxplot(osteoporosis$imc)
```



además de permitir visualizar lo ya comentado, añade la presencia de bastantes valores atípicos por encima del tercer cuartil.