Estadística descriptiva. Selección de actividades resueltas

2023-10-08

Actividad 1: cómputo del tiempo de CPU.

Medidas de tendencia central. Medidas de dispersión. Histogramas. Diagramas de caja.

El fichero TCPU.csv contiene los resultados de un test que consiste en ejecutar aleatoriamente diferentes programas en un ordenador y medir el tiempo de CPU consumido (en milisegundos) para cada programa (variable TCPU). También conocemos la longitud del código de cada uno de los programas ejecutados (variable LCODI). En este problema estudiaremos la variable TCPU.

a) Importad el fichero *TCPU.csv*.

Importamos los datos con la instrucción siguiente (indicando que las tres primeras filas contienen información sobre las variables, no datos; se ve abriendo el fichero con un procesador de textos o una hoja de cálculo cualquiera).

```
setwd("/home/xto/Documentos/01_UOC/Estadistica/Reto01/datosreto1/")
test1 <- read.table("ActR01TCPU.csv", skip = 3, sep =";", header = TRUE)
```

Siempre es conveniente después de la importación ver cómo han quedado las variables:

```
head(test1)
```

```
##   TCPU LCOD
## 1  127  146
## 2   83   80
## 3   85   60
## 4   93   90
## 5  103   58
## 6   80   88
```

b) Indicad el tipo de variable considerada.

La variable TCPU es una variable cuantitativa continua. Al tratarse de tiempo (milisegundos), podemos tomar un tiempo entre dos datos cualesquiera.

c) Calculad la media, la mediana, la desviación típica y los cuartiles, el máximo y el mínimo.

```
tmean <- mean(test1$TCPU)
tmean
```

```
## [1] 99.87
```

```
tmedian <- median(test1$TCPU)
tmedian
```

```
## [1] 101
```

```
tsd <- sd(test1$TCPU)
tsd
```

```
## [1] 21.55831
```

```r
tquantile1 <- quantile(test1$TCPU, 0.25)
tquantile1
```

```
## 25%
##  87
```

```r
tquantile3 <- quantile(test1$TCPU, 0.75)
tquantile3
```

```
##    75%
## 115.25
```

```r
tmax <- max(test1$TCPU)
tmax
```

```
## [1] 149
```

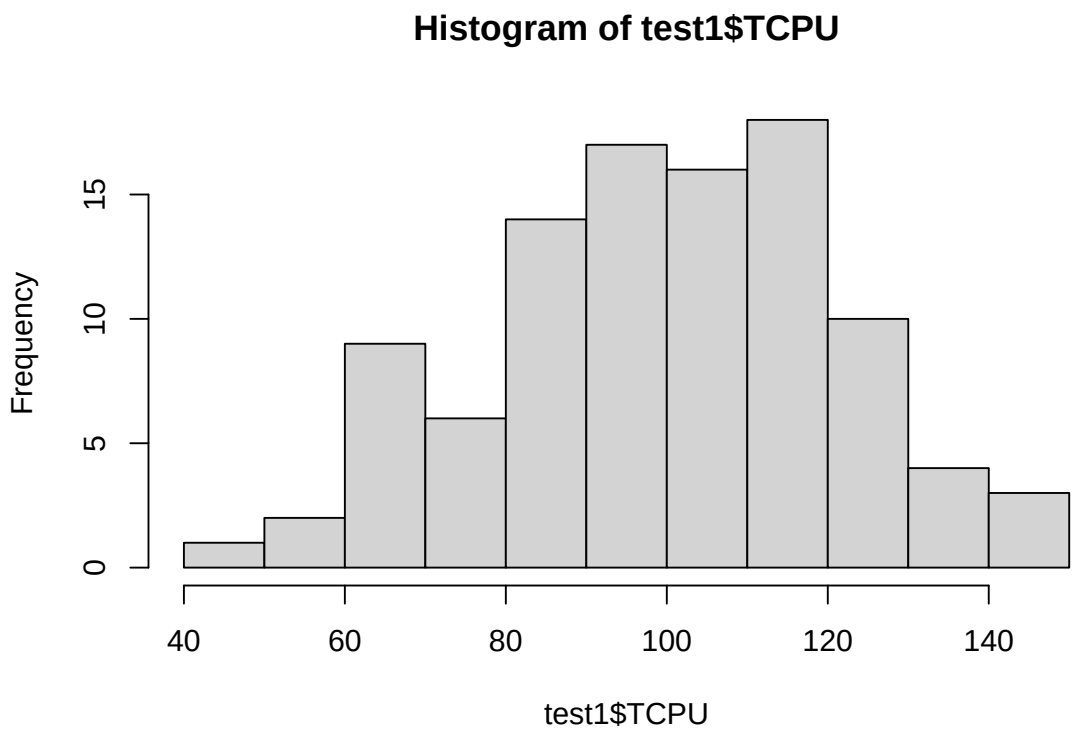
```r
tmin <- min(test1$TCPU)
tmin
```

```
## [1] 48
```

Por tanto, la media de la variable *TCPU* vale 99.87, la mediana 101, la desviación típica 21.55831. Los cuantiles primero y tercero valen 87 y 115.25 respectivamente, y el máximo y el mínimo, 149 y 38 respectivamente.

d) Dibujad un histograma de la variable y comentad su forma.

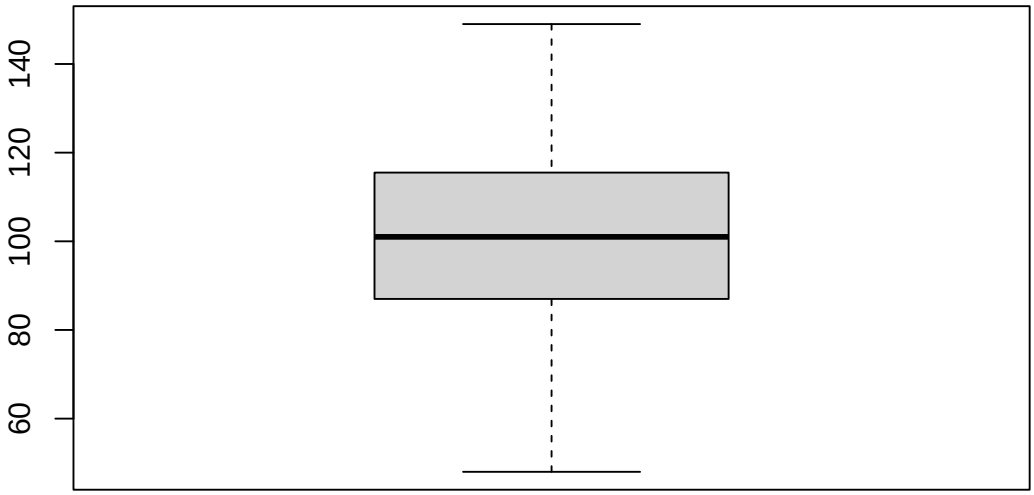
```r
hist(test1$TCPU)
```



Tiene una forma bastante simétrica sin ningún valor atípico.

e) Construid un diagrama de caja de la variable y comentad su forma. Indicad si hay datos anómalos o atípicos.

```
boxplot(test1$TCPU)
```



Comprobamos de nuevo la simetría de la variable, vemos cómo la caja es simétrica y no existen datos atípicos.

f) Comentad el estudio realizado.

Como conculusión, podemos afirmar que la variable *TCPU* es una variable continua con una distribución bastante simétrica y no hay datos atípicos.

Actividad 2: cómputo del tiempo de *CPU* agrupado.

Agrupamiento de datos estadísticos.

Con los datos de la actividad anterior, queremos tabular los datos para estudiar mejor la variable. Para hacerlo distribuiremos los tiempos de ejecución en tres categorías: "TC" (tiempo en el intervalo [48,81]), "T" (tiempo en el intervalo (81,114]), "TL" (tiempo en el intervalo (114,149]) creando la variable CLT. Para estudiar la variable CLT, se piden los resúmenes numéricos que ayuden a entender la distribución de la variable y un gráfico explicativo de la variable.

Como indica el enunciado de la actividad, agrupamos la variable *TCPU* usando la función *cut* de *R* e indicando los intervalos de agrupamiento de la forma siguiente:

```
CLT <- cut(test1$TCPU, breaks=c(48,81,114,149), labels=c('TC','T','TL'), include.lowest=TRUE)
```

Hemos creado una variable *CLT* que representa la variable *TCPU* agrupada y computamos los primeros valores:

```
head(CLT)
```

```
## [1] TL T  T  T  T  TC
## Levels: TC T TL
```

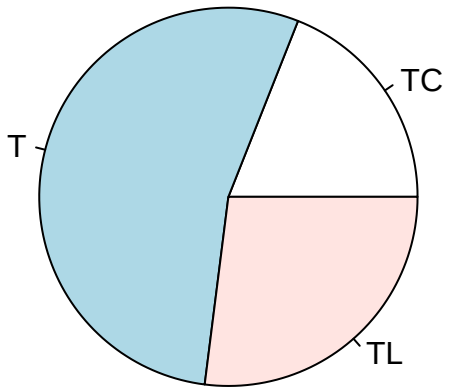
Los resúmenes numéricos para la variable *CLT* serán una tabla de frecuencias. Para esto, usamos la función *table* de *R*:

```
table(CLT)
```

```
## CLT
## TC  T TL
## 19 54 27
```

Vemos que los programas de media duración son los más abundantes y los de duración corta, los menos abundantes. Un gráfico explicativo de la variable podría ser un gráfico de sectores. Para ello, usamos la función *pie* de *R*:

```
pie(table(CLT))
```



Al observar el gŕafico obtenemos las mismas conclusiones que antes.

Actividad 3: inmersión de las tecnologías de la información y comunicación en los municipios.

Medidas de tendencia central. Medidas de dispersión. Histogramas. Diagramas de caja.

En el fichero *TICM.csv* se recogen los resultados de unas encuestas en diferentes municipios sobre el uso de las *TIC* el año 2007. De cada municipio tenemos cuatro valores: *PUORD* (proporción de hogares que tienen

ordenador), *PBA* (proporción de hogares que tienen banda ancha), *PUSUA* (proporción de habitantes que han utilizado el último mes el ordenador) y *PEMAIL* (proporción de habitantes que han usado el correo electrónico el último mes). En este problema estudiaremos y compararemos las variables *PUORD* y *PUSUA*.

a) Importad el fichero *TCIM.csv*.

Primero, nos situamos en la carpeta donde vamos a trabajar con *setwd*, luego importamos el archivo *csv* a la variable *Act3* y por último, comprobamos que la importación se haya hecho correctamente:

```
setwd("/home/xto/Documentos/01_UOC/Estadistica/Reto01/datosreto1/")
Act3 <- read.table("ActR01TICM.csv", sep = ";", dec = ",", skip = 4, header = TRUE,
                   fileEncoding = "UTF-8")
head(Act3)
```

```
##     PUORD     PBA   PUSUA  PEMAIL
## 1 64.6188 39.6013 58.9384 45.4938
## 2 70.6249 45.6342 64.3950 47.4064
## 3 64.2484 43.5412 60.4109 48.6097
## 4 71.2663 33.7878 52.4718 35.7512
## 5 57.6435 32.1987 50.8143 36.3483
## 6 64.7489 44.0721 61.5918 43.5046
```

b) Calculad la media, la mediana, la desviación típica y los cuartiles, el máximo y el mínimo de estas dos variables.

La media, la mediana, la desviación típica, los cuartiles, el máximo y el mínimo de la variable *PUORD* son:

```
mean(Act3$PUORD)
```

```
## [1] 64.79993
```

```
median(Act3$PUORD)
```

```
## [1] 64.2766
```

```
sd(Act3$PUORD)
```

```
## [1] 5.060672
```

```
quantile(Act3$PUORD, 0.25)
```

```
##     25%
## 61.2693
```

```
quantile(Act3$PUORD, 0.75)
```

```
##     75%
## 67.4023
```

```
max(Act3$PUORD)
```

```
## [1] 77.3762
```

```
min(Act3$PUORD)
```

```
## [1] 56.7604
```

Estos mismos datos para la variable *PUSUA* son:

```
mean(Act3$PUSUA)
```

```
## [1] 58.79159
```

```
median(Act3$PUSUA)
```

```
## [1] 60.1923
```

```
sd(Act3$PUSUA)
```

```
## [1] 5.453406
```

```
quantile(Act3$PUSUA, 0.25)
```

```
##     25%
## 56.1627
```

```
quantile(Act3$PUSUA, 0.75)
```

```
##   75%
## 62.53
```

```
max(Act3$PUSUA)
```

```
## [1] 67.8957
```

```
min(Act3$PUSUA)
```

```
## [1] 39.5549
```

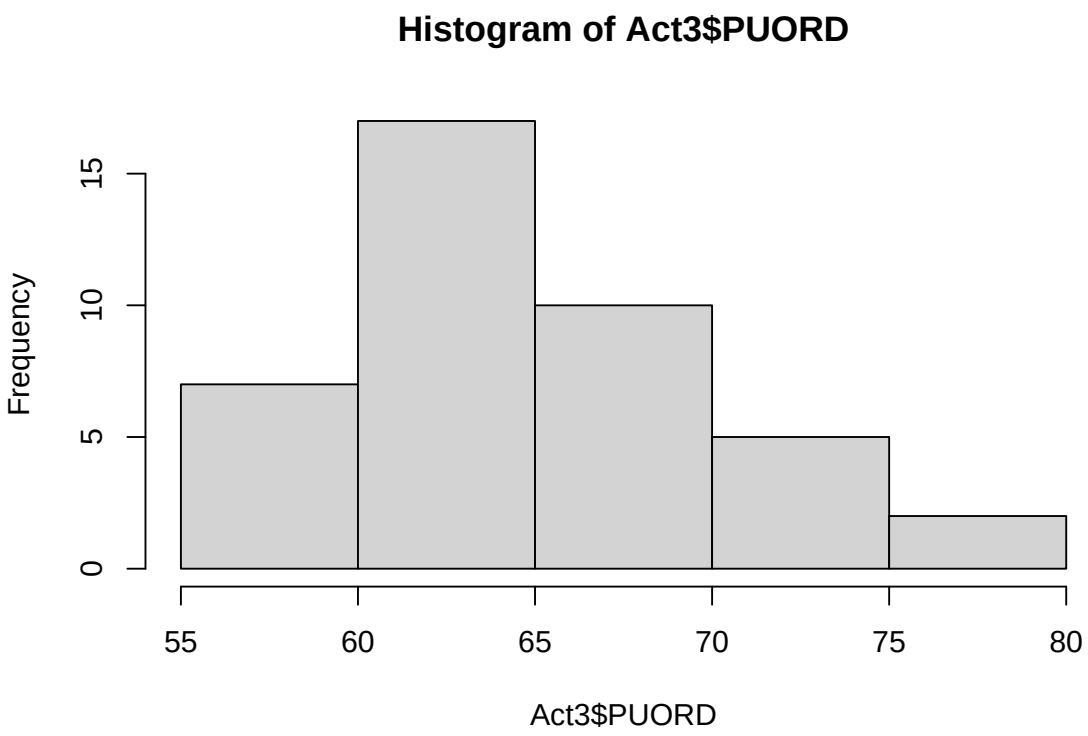
Escribimos los resultados anteriores en forma de tabla usando la función *data.frame* de *R*:

```
resul <- data.frame(c("Media", "Mediana", "Desviación típica", "Cuantil 25",
                      "Cuantil 75", "Máximo", "Mínimo"),
                    c(mean(Act3$PUORD), median(Act3$PUORD), sd(Act3$PUORD), quantile(
                      Act3$PUORD, 0.25), quantile(Act3$PUORD, 0.75), max(Act3$PUORD),
                      min(Act3$PUORD)),
                    c(mean(Act3$PUSUA), median(Act3$PUSUA), sd(Act3$PUSUA), quantile(
                      Act3$PUSUA, 0.25), quantile(Act3$PUSUA, 0.75), max(Act3$PUSUA),
                      min(Act3$PUSUA))
                    )
names(resul) <- c("Estadísticos", "PUORD", "PUSUA")
resul
```

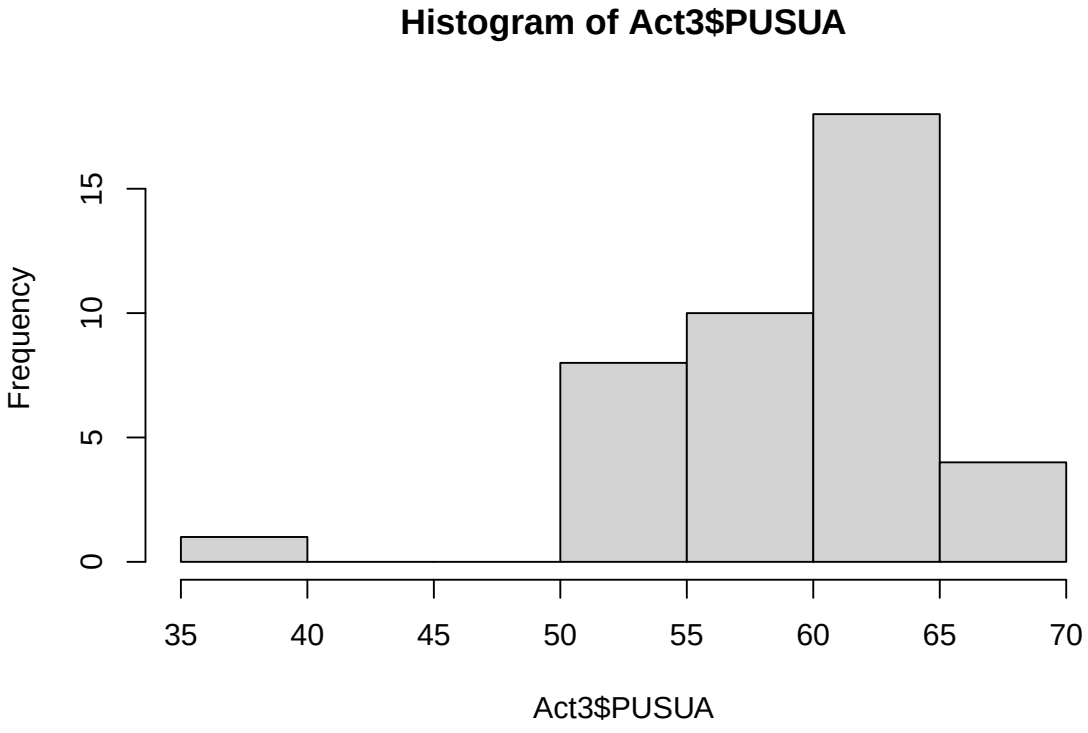
```
##        Estadísticos     PUORD     PUSUA
## 1             Media 64.799934 58.791588
## 2          Mediana 64.276600 60.192300
## 3 Desviación típica  5.060672  5.453406
## 4        Cuantil 25 61.269300 56.162700
## 5        Cuantil 75 67.402300 62.530000
## 6            Máximo 77.376200 67.895700
## 7            Mínimo 56.760400 39.554900
```

c) Dibujad un histograma de cada una de las dos variables.

```
hist(Act3$PUORD)
```



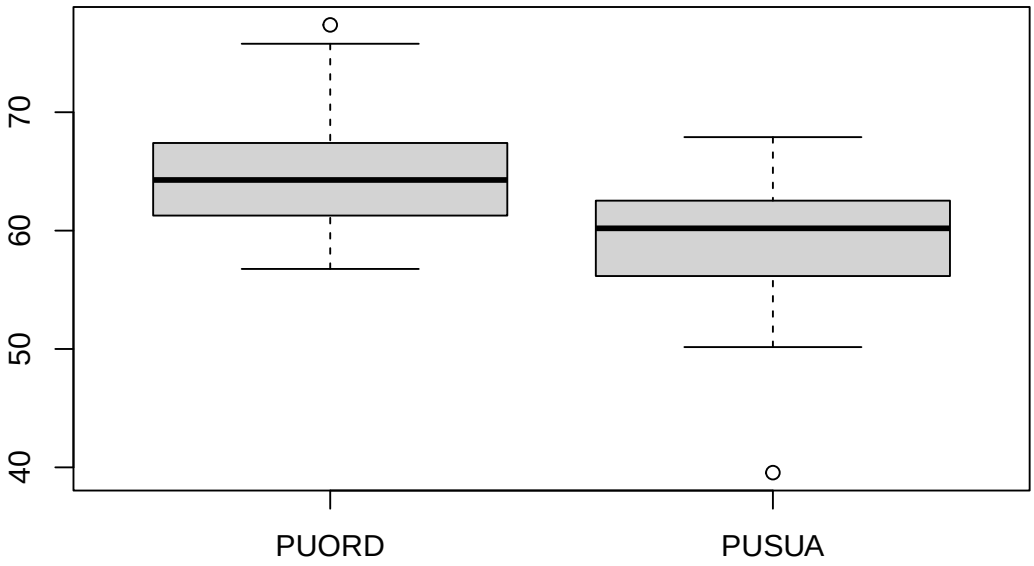

```
hist(Act3$PUSUA)
```



Vemos que el histograma de la variable *PUORD* tiene una cierta simetría, mientras que el histograma de la variable *PUSUA* es asimétrico con cola hacia la izquierda.

d) Construid diagramas de caja de las dos variables. Indicad si hay datos anómalos o atípicos.

```
boxplot(Act3$PUORD, Act3$PUSUA, names = c("PUORD", "PUSUA"))
```



En el gráfico comprobamos cierta simetría en la variable *PUORD* y la asimetría de la variable *PUSUA* con un dato atípico en cada variable.

e) Comentad los resultados comentando las diferencias-semejanzas entre las dos variables. Indicad qué gráfico o qué resumen numérico es más útil en este caso.

En los gráficos anteriores vemos que la proporción de hogares con ordenador (PUORD) es mayor que la proporción de usuarios que han usado el ordenador en el último mes (PUSUA), con una cierta simetría en el primer caso y una asimetría en el segundo. El gráfico más conveniente para realizar dicha comparación es el diagrama de caja.

Actividad 4: tiempo de computación de programas informáticos.

Agrupamiento de datos estadísticos. Medidas de tendencia central. Medidas de dispersión. Regla de Chebyshev.

El tiempo de computación (en segundos) de un determinado programa informático ejecutado de forma independiente cien veces en una misma máquina está recogido en el fichero *TCOMP.csv*.

a) Leed el fichero *TCOMP.csv* y agrupad la variable "tiempo de computación" en intervalos de amplitud 0.138, empezando con el valor 4.51. Calculad una tabla de frecuencias donde se indiquen las frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas.

Primero nos situamos en el directorio de trabajo y leemos el archivo *TCOMP.csv*, luego comprobamos que la importación ha sido correcta:

```
setwd("/home/xto/Documentos/01_UOC/Estadistica/Reto01/datosreto1/")
Act4 <- read.table("ActR01TCOMP.csv", header = TRUE, skip = 1, dec = ".",
                   fileEncoding = "UTF-8")
head(Act4)
```

```
##   TCOMP
## 1  4.67
## 2  4.94
## 3  5.09
## 4  4.74
## 5  4.63
## 6  4.62
```

Ahora hallamos el máximo de la columna anterior con la función *max* de *R* para saber hasta dónde tenemos que llegar:

```
max(Act4$TCOMP)
```

```
## [1] 5.2
```

Si el máximo de la variable "tiempo de computación" es 5.2, los intervalos serán los siguientes:

[4.51, 4.648], (4.648, 4.786], (4.786, 4.924], (4.924, 5.062], (5.062, 5.2]

Para agrupar la variable como se nos pide en el enunciado usaremos la función *cut* de *R*:

```
TCOMPAGRUP <- cut(Act4$TCOMP, breaks = seq(from = 4.51, to = 5.2, by = 0.138),
                  include.lowest = TRUE)
head(TCOMPAGRUP)
```

```
## [1] (4.65,4.79] (4.92,5.06] (5.06,5.2]  (4.65,4.79] [4.51,4.65] [4.51,4.65]
## Levels: [4.51,4.65] (4.65,4.79] (4.79,4.92] (4.92,5.06] (5.06,5.2]
```

Para construir la tabla de frecuencias que se nos pide, hallamos las frecuencias absolutas con la función *table*:

```
table(TCOMPAGRUP)
```

```
## TCOMPAGRUP
## [4.51,4.65] (4.65,4.79] (4.79,4.92] (4.92,5.06]  (5.06,5.2]
##          24          23          23          15          15
```

Para las frecuencias relativas usamos la función *prop.table*:

```
prop.table(table(TCOMPAGRUP))
```

```
## TCOMPAGRUP
## [4.51,4.65] (4.65,4.79] (4.79,4.92] (4.92,5.06]  (5.06,5.2]
##        0.24        0.23        0.23        0.15        0.15
```

Para las frecuencias absolutas acumuladas utilizamos la función *cumsum*:

```
cumsum(table(TCOMPAGRUP))
```

```
## [4.51,4.65] (4.65,4.79] (4.79,4.92] (4.92,5.06]  (5.06,5.2]
##          24          47          70          85         100
```

Para las frecuencias relativas acumuladas aplicaremos a la función *cumsum* el resultado de la función *prop.table*:

```
cumsum(prop.table(table(TCOMPAGRUP)))
```

```
## [4.51,4.65] (4.65,4.79] (4.79,4.92] (4.92,5.06]  (5.06,5.2]
##         0.24         0.47         0.70         0.85         1.00
```

Ahora, escribimos todos los resultados en forma de tabla resumen utilizando la función *cbind* de *R* como sigue:

```
resul2 <- cbind(table(TCOMPAGRUP), prop.table(table(TCOMPAGRUP)),
                cumsum(table(TCOMPAGRUP)), cumsum(prop.table(table(TCOMPAGRUP))))
colnames(resul2) = c("Frec. abs.", "Frec. rel.", "Frec. abs. acum.",
                     "Frec. rel. acum.")
resul2
```

```
##             Frec. abs. Frec. rel. Frec. abs. acum. Frec. rel. acum.
## [4.51,4.65]         24       0.24               24             0.24
## (4.65,4.79]         23       0.23               47             0.47
## (4.79,4.92]         23       0.23               70             0.70
## (4.92,5.06]         15       0.15               85             0.85
## (5.06,5.2]          15       0.15              100             1.00
```

b) Calculad la media y la mediana de la variable agrupada "tiempo de computación agrupado". Calculad también la desviación típica de la variable agrupada.

Para poder hallar la media, la mediana y la desviación típica de la variable agrupada anteriormente, necesitamos hallar la marca de clase para todos los intervalos anteriores. Primero hallaremos los extremos de la izquierda (EL) y derecha (ER) de los intervalos considerados, a continuación hallaremos las marcas de clase (MC) y por último, mostraremos la tabla de frecuencias anterior incluyendo las marcas de clase halladas:

```
EL = seq(from = 4.51, to = 5.2 - 0.138, by = 0.138)
ER = seq(from = 4.51 + 0.138, to = 5.2, by = 0.138)
MC = (EL + ER) / 2
MC
```

```
## [1] 4.579 4.717 4.855 4.993 5.131
```

```
resul3 <- cbind(MC, table(TCOMPAGRUP), prop.table(table(TCOMPAGRUP)),
                cumsum(table(TCOMPAGRUP)), cumsum(prop.table(table(TCOMPAGRUP))))
colnames(resul3) = c("Marcas de clase", "F. abs.", "F. rel.", "F. abs. acum.",
                     "F. rel. acum.")
resul3
```

```
##             Marcas de clase F. abs. F. rel. F. abs. acum. F. rel. acum.
## [4.51,4.65]           4.579      24    0.24            24          0.24
## (4.65,4.79]           4.717      23    0.23            47          0.47
## (4.79,4.92]           4.855      23    0.23            70          0.70
## (4.92,5.06]           4.993      15    0.15            85          0.85
## (5.06,5.2]            5.131      15    0.15           100          1.00
```

Para hallar la media, la mediana y la desviación típica, redefiniremos la variable *TCOMPAGRUP* usando como identificador de cada intervalo la marca de clase:

```
TCOMPAGRUP2 <- cut(Act4$TCOMP, breaks = seq(from = 4.51, to = 5.2, by = 0.138),
                   include.lowest = TRUE, labels = MC)
TCOMPAGRUP2 <- as.numeric(as.character(TCOMPAGRUP2))
```

El resultado de la función *cut* es una variable tipo *factor*. Por tanto, antes de hallar la media, la mediana y la varianza, hemos tenido que transformar la variable tipo *factor TCOMPAGRUP2* en una variable tipo numérico usando las funciones *as.character* y *as.numeric* de *R*. Por último, hallamos la media, la mediana y la desviación típica que se nos piden:

```r
media_TCOMPAGRUP2 <- mean(TCOMPAGRUP2)
mediana_TCOMPAGRUP2 <- median(TCOMPAGRUP2)
desviacion_tipica_TCOMPAGRUP2 <- sd(TCOMPAGRUP2)

media_TCOMPAGRUP2
```

```
## [1] 4.81912
```

```r
mediana_TCOMPAGRUP2
```

```
## [1] 4.855
```

```r
desviacion_tipica_TCOMPAGRUP2
```

```
## [1] 0.1897845
```

Por tanto, la media vale 4.81912, la mediana vale 4.855 y la desviación típica 0.1897845.