Estadística descriptiva

Selección de actividades resueltas

PID_00279894

Àngel J. Gil Estallo
Jose Fco. Martínez Boscá
Arnau Mir Torres
Lluís M. Pla Aragonés
Ángel A. Juan Pérez (editor)

Universitat Oberta de Catalunya

Àngel J. Gil Estallo | **Jose Fco. Martínez Boscá** | **Arnau Mir Torres**

Lluís M. Pla Aragonés | **Ángel A. Juan Pérez (editor)**

La revisión de este recurso de aprendizaje UOC ha sido coordinada por la profesora: Laura Calvet Liñán

Segunda edición: febrero 2021

Av. Tibidabo, 39-43, 08035 Barcelona
Autoría: Àngel J. Gil Estallo, Jose Fco. Martínez Boscá, Arnau Mir Torres, Lluís M. Pla Aragonés
Producción: FUOC

Índice

Introducción

Conocer cuántos mensajes *spam* pasan por un servidor de correo de una determinada empresa puede realizarse de dos maneras. En primer lugar, se puede intentar adivinar teniendo en cuenta el número de empleados de la empresa, el tipo de páginas web que visitan, cuántos de ellos chatean, etc. Otra forma de abordar el problema es recoger el número de mensajes *spam* que pasan por el servidor durante un conjunto de días. El primer método es muy complejo debido a la cantidad de variables que hay que tener en cuenta; en cambio, el segundo método es muy simple pero tenemos que aprender técnicas para poder alcanzar nuestro objetivo. La estadística es la disciplina que se dedica a resolver problemas como el anterior usando métodos como el que hemos mencionado de recogimiento de datos.

En todo estudio estadístico, existen dos fases bien diferenciadas:

- Fase 1: recogida de datos y,
- Fase 2: análisis de dichos datos.

En la fase 1, los datos se recogen, agrupan y se caracterizan. La parte de la estadística encargada de llevar a cabo la fase 1 se denomina estadística descriptiva. En la fase 2, se realiza el análisis de dichos datos con el fin de sacar conclusiones a partir de dicho análisis. La parte de la estadística encargada de llevar a cabo la fase 2 se denomina estadística inferencial. ¿Qué tipo de conclusiones esperamos obtener? Básicamente conocer información de toda la población a partir del estudio realizado de una muestra de datos. En el ejemplo anterior, una vez tengamos caracterizada la muestra correspondiente al número de mensajes *spam* que recibe el servidor durante treinta días, ¿qué podemos decir sobre el número de mensajes *spam* que recibe diariamente dicho servidor? Concretando un poco más, ¿podemos afirmar que la media de mensajes *spam* que recibe el servidor de correo de la empresa es representativa de la media de todos los mensajes *spam* que recibe diariamente dicho servidor?

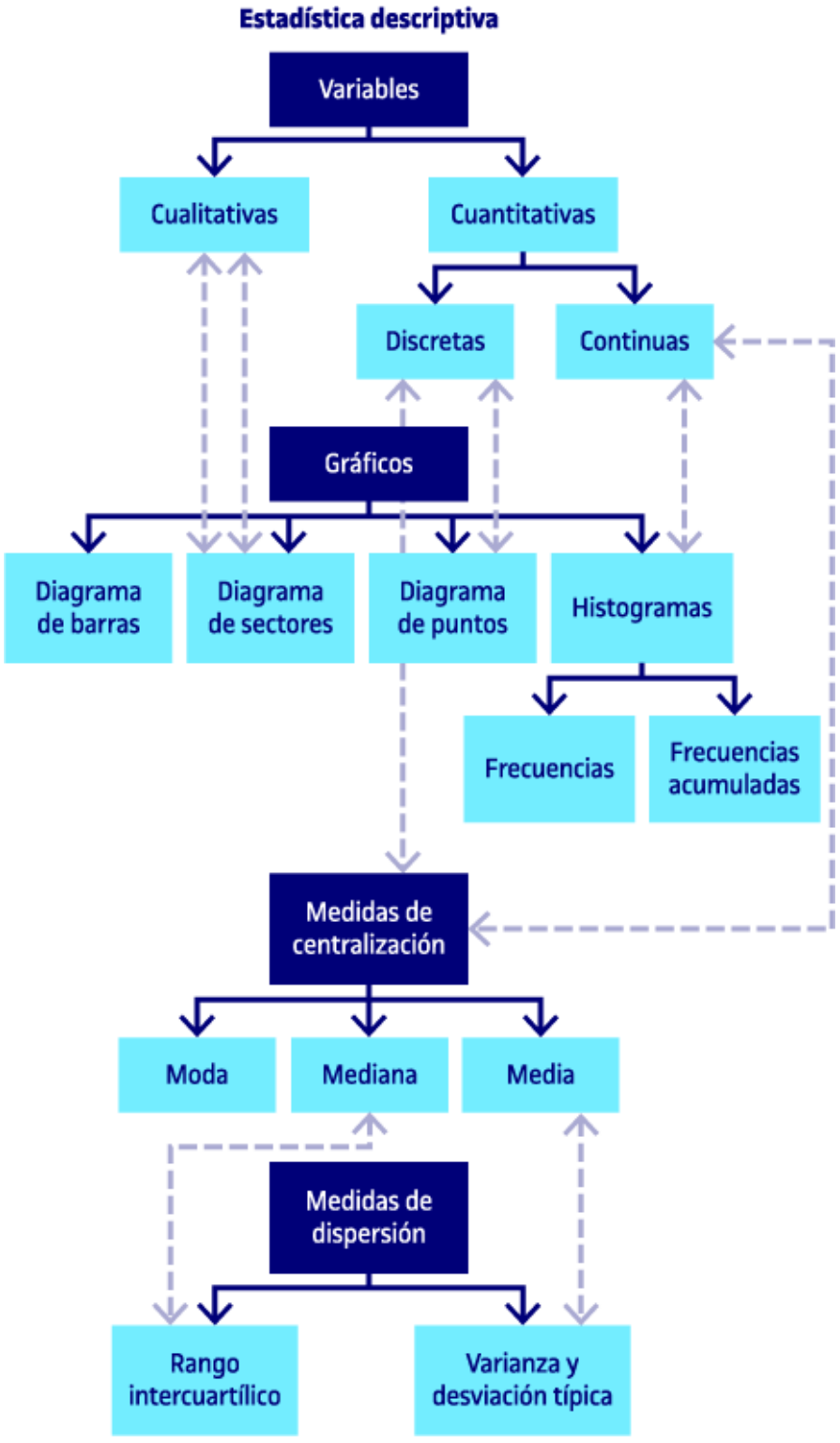
La estadística intenta resolver problemas más complejos. Por ejemplo, siguiendo con el ejemplo anterior, imaginemos que la empresa anterior no solo está interesada en conocer el número de mensajes *spam* que recibe cada día sino el tipo de mensajes *spam*. O sea, queremos estudiar el conjunto de mensajes *spam* que recibe el servidor de correo de dicha empresa. ¿Qué significa exactamente conocer dicho conjunto? La respuesta a esta pregunta puede ser muy amplia. Por ejemplo, clasificar los mensajes *spam* usando una serie de características comunes; o intentar estudiar a qué horas se recibe más *spam*, etc. Todo ello para intentar hacer una predicción del comporta-

miento de los mensajes *spam* de dicha empresa. Las técnicas anteriores son ejemplos de herramientas estadísticas denominadas *data mining*, herramientas muy importantes en ciencias de la computación.

En este módulo presentamos un conjunto de ejemplos dedicados a la recolección de datos e intentar caracterizar dichos datos. Los datos se pueden caracterizar hallando valores que los representen y hallando valores que nos indiquen lo dispersos que están. Los valores que representan a los datos reciben el nombre de medidas de centralización y los valores que indican la dispersión de los datos reciben el nombre de medidas de dispersión.

Las medidas de centralización más importantes son la media aritmética, la mediana y los percentiles, y las medidas de dispersión más importantes son la varianza y la desviación típica.

Mapa conceptual



Actividades

Ficheros necesarios para realizar las actividades:

- ActR01TCPU.csv
- ActR01TICM.csv
- ActR01TCOMP.csv
- ActR01LPROG.csv

En las primeras filas de los ficheros constan las descripciones de las variables y en el momento de importarlos a R hay que tener en cuenta que estas líneas no contienen datos (se aconseja usar *skip*).

En otras actividades en las que el número de observaciones es pequeño se sugiere copiar-pegar los datos o la instrucción que se usa para introducirlos en R.

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.

b) Indicad el tipo de variable considerada.

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

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

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

f) Comentad el estudio realizado.

Solución

a) 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).

```
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) La variable TCPU es una variable cuantitativa continua.

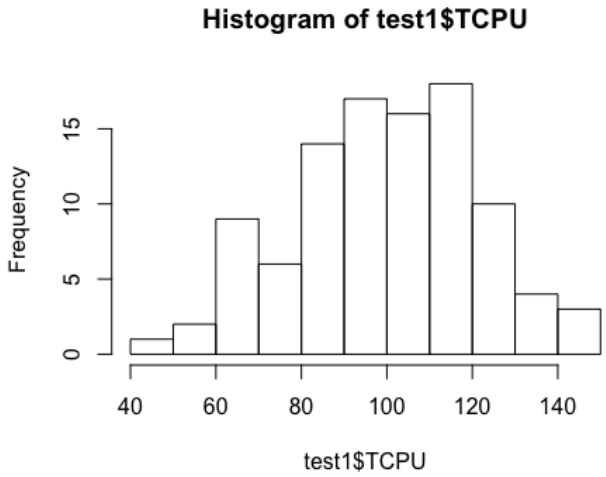
c) Calculad la media, la mediana, la desviación típica y los cuartiles, el máximo y el mínimo.

```
tm1<-mean(test1$TCPU)
tm1
## [1] 99.87
tm2<-median(test1$TCPU)
tm2
## [1] 101
tsd<-sd(test1$TCPU)
tsd
## [1] 21.55831
tq1<-quantile(test1$TCPU,0.25)
tq1
## 25%
##  87
tq3<-quantile(test1$TCPU,0.75)
tq3
##    75%
## 115.25
tmax<-max(test1$TCPU)
tmax
## [1] 149
tmin<-min(test1$TCPU)
tmin
## [1] 48
```

Por tanto, la media de la variable 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 48, respectivamente.

d) Para hacer el histograma de la variable, usamos la función *hist* de R:

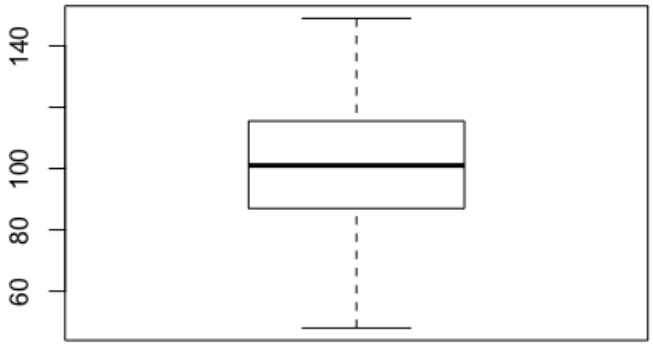
```
hist(test1$TCPU)
```



Vemos que tiene una forma bastante simétrica, con ningún valor atípico.

e) Para realizar un *boxplot*, usamos la función *boxplot* de R.

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



Comprobamos una vez más la simetría de la variable viendo cómo la caja del gráfico anterior es simétrica y la inexistencia de datos atípicos.

f) Como conclusión, 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.

Solución

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)
```

Observemos que hemos creado una variable nueva CLT que representa la variable agrupada del tiempo de CPU y computamos los primeros valores:

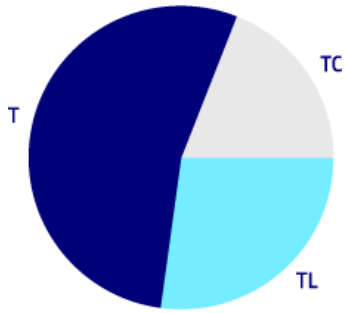
```
head(CLT)
## [1] TL T  T  T  T  TC
## Levels: TC T TL
```

Los resúmenes numéricos para la variable CLT serán una tabla de frecuencias. Para poder realizarla, 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 posible gráfico explicativo de la variable anterior podría ser un gráfico de sectores. Para ello, usamos la función *pie* de R:

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



Como podemos observar, las conclusiones son las mismas que hemos comentado 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 (propor-

ción 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.

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

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

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

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.

Solución

a) Primero importaremos el fichero y comprobaremos que la importación sea correcta:

```
Act3<-read.table("ActR01TICM.csv", sep=";" ,dec=",",skip=4, header=TRUE)
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) La media, la mediana, la desviación típica y los cuartiles, el máximo y el mínimo de la variable PUORD valen:

```
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
```

Y para la variable PUSUA:

```
mean(Act3$PUSUA)
## [1] 58.79159
median(Act3$PUSUA)
```

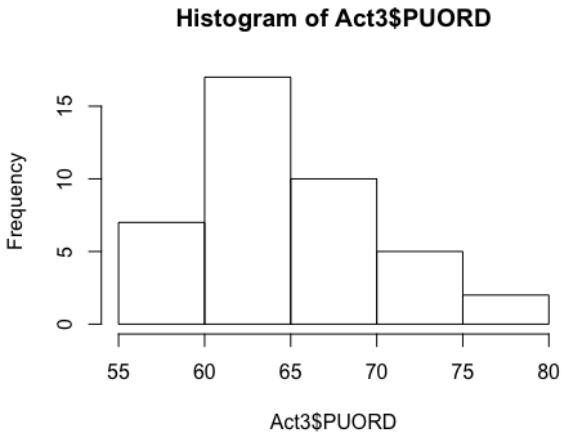
```
## [1] 60.1923
sd(Act3$PUSUA)
## [1] 5.453406
quantile(Act3$PUSUA,0.25)
##     25%
## 56.1627
quantile(Act3$PUSUA,0.75)
##   75%
## 62.53
max(Act3$PUSUA)
## [1] 67.8957
min(Act3$PUSUA)
## [1] 39.5549
```

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

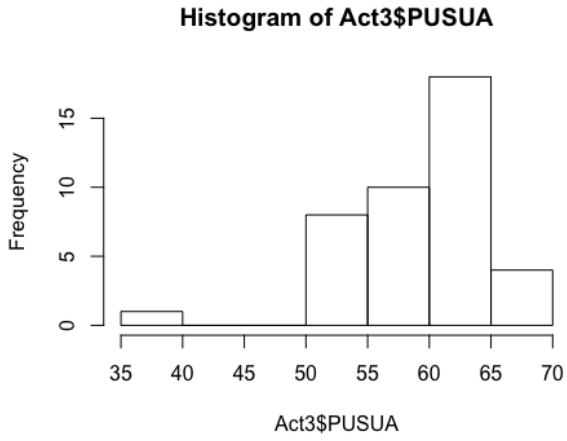
```
resul<-data.frame(c("media","mediana","desviación típica",
                    "cuantil 25","cuantil 75","máximo", "mínimo"),
                  c(mean(Act3$PUORD),median(Act3$PUORD),sd(Act3$PUORD),
                    quantile(Act3$PUORD,0.25),quantile(Act3$PUORD,0.75),
                    max(Act3$PUORD),
                    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) Los histogramas se realizan usando la función *hist* de R:

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



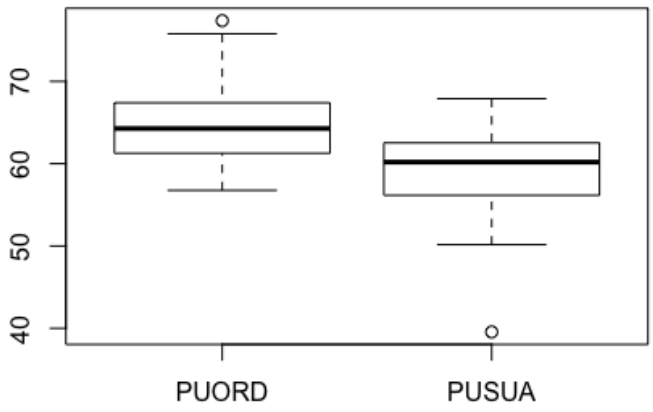

```
hist(Act3$PUSUA)
```



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) Los diagramas de caja se construyen usando la función *boxplot* de R:

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



En dicho gráfico, volvemos a comprobar cierta simetría en la variable PUORD y la asimetría de la variable PUSUA con un dato atípico en cada variable.

e) 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.
Se pide:
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.
b) Calculad la media y la mediana de la variable agrupada "tiempo de computación agrupado". Calculad también la desviación típica de la variable agrupada.
c) Suponemos ahora que ejecutamos de forma independiente mil veces el programa y obtenemos la misma media y la misma desviación típica que en los apartados anteriores. ¿Entre qué valores se encuentran como mínimo el 88.8% de los tiempos de computación?

Solución

a) Primero leemos el fichero con la variable tiempo de computación en R:

```
Act4<-read.table("ActR01TCOMP.csv", sep=";", skip=1, header=TRUE)
head(Act4)
##   TCOMP
## 1  4.67
## 2  4.94
## 3  5.09
## 4  4.74
## 5  4.63
## 6  4.62
```

Seguidamente vamos a hallar el máximo de la columna anterior. Para hacerlo, usamos la función *max* de R:

```
max(Act4$TCOMP)
## [1] 5.2
```

Por tanto, 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 usamos la función *cut* de R de la forma siguiente:

```
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 hallar la tabla de frecuencias pedida, hallamos las frecuencias absolutas usando 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
```

Las frecuencias relativas usando 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
```

Las frecuencias absolutas acumulados con 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
```

Y las frecuencias relativas acumuladas con la función *cumsum* aplicada al 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
```

Para escribir todos los resultados anteriores en forma de tabla resumen, podemos usar la función *cbind* de R de la forma siguiente:

```
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) Para poder hallar la media, la mediana y la desviación típica de la variable agrupada anterior, necesitamos hallar la marca de clase para todos los intervalos anteriores. Primero hallamos los extremos de la izquierda (EL) y derecha (ER) de los intervalos considerados, a continuación hallamos las marcas de clase (MC) y, por último, mostramos 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 clase","F. abs.","F. rel.",
                     "F. abs. acum.","F. rel. acum.")
resul3
##                     Marcas 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, redefinimos 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 pedidas:

```
mean(TCOMPAGRUP2)
## [1] 4.81912
median(TCOMPAGRUP2)
## [1] 4.855
sd(TCOMPAGRUP2)
## [1] 0.1897845
```

Por tanto, la media vale 4.8191, la mediana, 4.8550 y la desviación típica 0.1898.

c) Para realizar dicho apartado, hemos de usar la desigualdad de Chebyshev. El intervalo pedido es de la forma $(\bar{X} - ms_x, \bar{X} + ms_x)$, donde m cumple: $1 - \frac{1}{m^2} = 0.889$. Hallando m de la igualdad anterior, tenemos que $m = 3$. Seguidamente tenemos que calcular la media y la desviación típica de la variable no agrupada:

```
mean(Act4$TCOMP)
## [1] 4.8199
sd(Act4$TCOMP)
## [1] 0.1960442
```

La media será 4.8199 y la desviación típica, 0.1960. El intervalo será:
(4.8199 – 3 · 0.1960,4.8199 + 3 · 0.1960) = (4.2318,5.4080).
Por tanto, podemos afirmar que 889 de los 1000 tiempos de ejecución estarán entre 4.2318 y 5.4080.

Actividad 5: número de mensajes no deseados que recibe una empresa
Datos estadísticos discretos. Tabla de frecuencias. Diagrama de puntos. Medidas de tendencia central. Medidas de dispersión. Regla de Chebyshev

La siguiente tabla nos indica el número de mensajes *spam* que reciben en un día cualquiera los empleados de una determinada empresa:

Número de *spam*	0	1	2	3	4	5	6	7
Número de empleados	7	11	10	7	1	2	1	1

La tabla anterior se ha de interpretar así: 7 empleados no reciben ningún *spam* en el día considerado, 11 empleados reciben 1 *spam* en el día considerado, etc. Consideramos la variable X = "número de mensajes *spam* que recibe un empleado cualquiera de esta empresa por día". Se pide:

a) Calculad una tabla de frecuencias donde se indiquen las frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas.

b) Haced un diagrama de puntos de la variable X.

c) Con base en los apartados anteriores, comentad cómo es la variable X (forma, distribución…).

d) ¿Qué porcentaje de empleados reciben entre 2 y 6 mensajes *spam* cada día?

e) Calculad la media, la moda, la mediana y la desviación típica de X.

f) Si se duplica el número de mensajes *spam* que recibe cada empleado, ¿cuáles serán las nuevas media y variancia del número de mensajes *spam* que recibe un empleado cualquiera por día? Este apartado se debe hacer aplicando las propiedades de la media y la desviación típica sin volver a calcular la media y la desviación típica de la nueva variable.

Solución

a) En primer lugar, introducimos las variables que nos dan el número de mensajes *spam* y el número de empleados, respectivamente:

```
SPAM=0:7
EMPL=c(7,11,10,7,1,2,1,1)
```

A continuación, creamos la variable de estudio, "número de *spam* que recibe un empleado", la llamamos SPAM_EMPL, y la definimos de la forma siguiente: repetimos los valores de la columna *spam* tantas veces como indica la variable "EMPL", usando la función *rep*:

```
SPAMEMPL=rep(SPAM,EMPL)
SPAMEMPL[1:19]
##  [1] 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 2
```

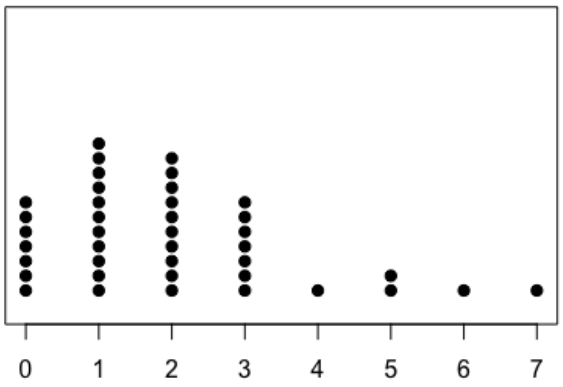
O sea, repetimos el valor 0, 7 veces, el valor 1, 11 veces y así sucesivamente. Para hallar la tabla de frecuencias, usamos el mismo procedimiento que hemos usado en el apartado a) de la actividad 4:

```
resul5<-cbind(table(SPAMEMPL),prop.table(table(SPAMEMPL)),
          cumsum(table(SPAMEMPL)),cumsum(prop.table(table(SPAMEMPL))))
colnames(resul5) = c("Frec. abs.","Frec. rel.","Frec. abs. acum.",
                     "Frec. rel. acum.")
resul5
##   Frec. abs. Frec. rel. Frec. abs. acum. Frec. rel. acum.
## 0          7      0.175                7            0.175
## 1         11      0.275               18            0.450
## 2         10      0.250               28            0.700
## 3          7      0.175               35            0.875
## 4          1      0.025               36            0.900
## 5          2      0.050               38            0.950
## 6          1      0.025               39            0.975
## 7          1      0.025               40            1.000
```

Las dos primeras columnas coinciden con las variables SPAM y EMPL que indican los valores de la variable y las frecuencias absolutas, respectivamente.

b) Para poder realizar un diagrama de puntos, podemos usar la función *stripchart* de R:

```
stripchart(SPAMEMPL, method = "stack", offset = .5, at = .15, pch = 19)
```



c) Vemos que la variable de estudio X tiene una distribución asimétrica con cola a la derecha donde los valores más repetidos son 0, 1, 2 y 3. Por tanto, la mayoría de los empleados reciben o ningún, o uno, o dos o tres *spams* diarios.

d) El número de empleados que reciben entre 2 y 6 *spams* son $21 = 39 - 18$. O sea, frecuencia absoluta acumulada de 6 menos frecuencia absoluta acumulada de 1. El porcentaje de EMPL será: $\frac{21}{40} \cdot 100 = 52{,}5\%$.

e) La media, la mediana y la desviación típica de X serán:

```
mean(SPAMEMPL)
## [1] 1.975
median(SPAMEMPL)
## [1] 2
sd(SPAMEMPL)
## [1] 1.671595
```

Por tanto, la media de mensajes *spam* recibidos por día es de 1.975, la mediana es de 2 y la desviación típica, de 1.672. La moda, como puede observarse en el diagrama de puntos anterior, vale 1.

f) Si se duplican el número de mensajes recibidos por día, también se duplicarán la media y la desviación típica. Por tanto, estas serán: ($\bar{X} = 1.975 \cdot 2 = 3.95$, $s_x = 1.672 \cdot 2 = 3.344$). Comprobamos dichos resultados en R definiendo una nueva variable que sea el doble de la variable SPAMEMPL:

```
SPAMEMPL2 = 2*SPAMEMPL
mean(SPAMEMPL2)
## [1] 3.95
sd(SPAMEMPL2)
## [1] 3.343191
```

Podemos comprobar que los valores de la media y la desviación típica coinciden con los valores calculados anteriormente.

Actividad 6: conocimiento de los lenguajes de programación actuales por parte de los estudiantes de ciencias de la computación

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

En el fichero LPROG.csv se encuentran datos sobre el conocimiento de los lenguajes de programación Java, Perl y Python por parte de veinticinco estudiantes de ciencias de la computación de 0 (ningún conocimiento) a 100 (máximo dominio del lenguaje):
Se pide:

a) Leed el fichero.

b) Agrupad las variables "conocimiento del lenguaje de programación Java, Perl y Python" en cinco intervalos de igual amplitud. Calculad una tabla de frecuencias donde se indiquen las frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas.

c) Calculad la media y la mediana de las variables agrupadas.

d) Calculad la desviación típica también de las variables agrupadas.

e) Haced un histograma de cada una de las variables.

f) Comentad los resultados indicando las diferencias-semejanzas entre las tres variables.

Solución

a) Primero leemos el fichero:

```
Act6<-read.table("ActR01LPROG.csv", sep=";", skip=1, header=TRUE)
head(Act6)
##    Java  Perl Python
## 1 50.27 75.85  20.30
## 2 50.51 74.04  19.68
## 3 49.58 74.32  20.41
## 4 50.09 75.04  20.00
## 5 50.27 74.68  20.43
## 6 50.72 74.93  19.78
```

b) En primer lugar, calculamos los intervalos de agrupamiento de cada variable:

```
INTJAVA<-seq(from=min(Act6$Java),to=max(Act6$Java),
             by=(max(Act6$Java)-min(Act6$Java))/5)
INTJAVA
## [1] 48.50 49.24 49.98 50.72 51.46 52.20
INTPERL<-seq(from=min(Act6$Perl),to=max(Act6$Perl),
             by=(max(Act6$Perl)-min(Act6$Perl))/5)
INTPERL
## [1] 73.220 73.746 74.272 74.798 75.324 75.850
INTPYTHON<-seq(from=min(Act6$Python),to=max(Act6$Python),
               by=(max(Act6$Python)-min(Act6$Python))/5)
INTPYTHON
## [1] 19.310 19.562 19.814 20.066 20.318 20.570
```

Por ejemplo, para la variable JAVA, los intervalos serán: $[48.5, 49.24)$, $[49.24, 49.98)$, $[49.98, 50.72)$, $[50.72, 51.46)$ y $[51.46, 52.20)$. Los demás se calculan de forma similar. A continuación calculamos las marcas de clase de los intervalos:

```
MCLASJAVA<-c()
for (i in 1:5) {MCLASJAVA<-c(MCLASJAVA,(INTJAVA[i+1]+INTJAVA[i])/2)}
MCLASJAVA
## [1] 48.87 49.61 50.35 51.09 51.83
MCLASPERL<-c()
for (i in 1:5) {MCLASPERL<-c(MCLASPERL,(INTPERL[i+1]+INTPERL[i])/2)}
MCLASPERL
## [1] 73.483 74.009 74.535 75.061 75.587
MCLASPYTHON<-c()
for (i in 1:5) {MCLASPYTHON<-c(MCLASPYTHON,(INTPYTHON[i+1]+INTPYTHON[i])/2)}
MCLASPYTHON
## [1] 19.436 19.688 19.940 20.192 20.444
```

A continuación, calculamos las variables agrupadas:

```
JAVAAGRUP<-cut(Act6$Java,breaks = INTJAVA,right = F,include.lowest = T)
JAVAAGRUP
##  [1] [50,50.7)   [50,50.7)   [49.2,50)   [50,50.7)   [50,50.7)   [50.7,51.5)
##  [7] [48.5,49.2) [50,50.7)   [50.7,51.5) [49.2,50)   [50,50.7)   [48.5,49.2)
## [13] [50,50.7)   [49.2,50)   [50,50.7)   [48.5,49.2) [48.5,49.2) [48.5,49.2)
## [19] [49.2,50)   [50.7,51.5) [48.5,49.2) [50.7,51.5) [49.2,50)   [51.5,52.2]
## [25] [50,50.7)
## Levels: [48.5,49.2) [49.2,50) [50,50.7) [50.7,51.5) [51.5,52.2]
PERLAGRUP<-cut(Act6$Perl,breaks = INTPERL,right = F,include.lowest = T)
PERLAGRUP
##  [1] [75.3,75.8] [73.7,74.3) [74.3,74.8) [74.8,75.3) [74.3,74.8) [74.8,75.3)
##  [7] [74.8,75.3) [74.3,74.8) [75.3,75.8] [75.3,75.8] [75.3,75.8] [74.8,75.3)
## [13] [74.8,75.3) [74.8,75.3) [74.3,74.8) [75.3,75.8] [73.2,73.7) [73.2,73.7)
## [19] [74.3,74.8) [75.3,75.8] [74.3,74.8) [75.3,75.8] [74.8,75.3) [75.3,75.8]
## [25] [73.7,74.3)
## Levels: [73.2,73.7) [73.7,74.3) [74.3,74.8) [74.8,75.3) [75.3,75.8]
PYTHONAGRUP<-cut(Act6$Python,breaks = INTPYTHON,right = F,include.lowest = T)
PYTHONAGRUP
##  [1] [20.1,20.3) [19.6,19.8) [20.3,20.6] [19.8,20.1) [20.3,20.6] [19.6,19.8)
##  [7] [20.1,20.3) [20.1,20.3) [20.1,20.3) [19.6,19.8) [20.3,20.6] [19.6,19.8)
## [13] [19.6,19.8) [19.3,19.6) [20.1,20.3) [19.6,19.8) [19.8,20.1) [20.3,20.6]
## [19] [19.6,19.8) [20.1,20.3) [19.8,20.1) [20.3,20.6] [20.1,20.3) [20.1,20.3)
## [25] [19.8,20.1)
## Levels: [19.3,19.6) [19.6,19.8) [19.8,20.1) [20.1,20.3) [20.3,20.6]
```

Las frecuencias absolutas se calculan con la instrucción *table*:

```
FREQABSJAVA<-table(JAVAAGRUP)
FREQABSJAVA
## JAVAAGRUP
## [48.5,49.2)   [49.2,50)   [50,50.7) [50.7,51.5) [51.5,52.2]
##           6           5           9           4           1
FREQABSPERL<-table(PERLAGRUP)
FREQABSPERL
## PERLAGRUP
## [73.2,73.7) [73.7,74.3) [74.3,74.8) [74.8,75.3) [75.3,75.8]
##           2           2           6           7           8
FREQABSPYTHON<-table(PYTHONAGRUP)
FREQABSPYTHON
```

```
## PYTHONAGRUP
## [19.3,19.6) [19.6,19.8) [19.8,20.1) [20.1,20.3) [20.3,20.6]
##           1           7           4           8           5
```

Las frecuencias relativas serán:

```
FREQRELJAVA<-table(JAVAAGRUP)/sum(table(JAVAAGRUP))
FREQRELJAVA
## JAVAAGRUP
## [48.5,49.2)   [49.2,50)   [50,50.7) [50.7,51.5) [51.5,52.2]
##        0.24        0.20        0.36        0.16        0.04
FREQRELPERL<-table(PERLAGRUP)/sum(table(PERLAGRUP))
FREQRELPERL
## PERLAGRUP
## [73.2,73.7) [73.7,74.3) [74.3,74.8) [74.8,75.3) [75.3,75.8]
##        0.08        0.08        0.24        0.28        0.32
FREQRELPYTHON<-table(PYTHONAGRUP)/sum(table(PYTHONAGRUP))
FREQRELPYTHON
## PYTHONAGRUP
## [19.3,19.6) [19.6,19.8) [19.8,20.1) [20.1,20.3) [20.3,20.6]
##        0.04        0.28        0.16        0.32        0.20
```

Las frecuencias absolutas acumuladas serán:

```
FREQABSACUMJAVA<-cumsum(FREQABSJAVA)
FREQABSACUMJAVA
## [48.5,49.2)   [49.2,50)   [50,50.7) [50.7,51.5) [51.5,52.2]
##           6          11          20          24          25
FREQABSACUMPERL<-cumsum(FREQABSPERL)
FREQABSACUMPERL
## [73.2,73.7) [73.7,74.3) [74.3,74.8) [74.8,75.3) [75.3,75.8]
##           2           4          10          17          25
FREQABSACUMPYTHON<-cumsum(FREQABSPYTHON)
FREQABSACUMPYTHON
## [19.3,19.6) [19.6,19.8) [19.8,20.1) [20.1,20.3) [20.3,20.6]
##           1           8          12          20          25
```

Las frecuencias relativas acumuladas serán:

```
FREQRELACUMJAVA<-cumsum(FREQRELJAVA)
FREQRELACUMJAVA
## [48.5,49.2)   [49.2,50)   [50,50.7) [50.7,51.5) [51.5,52.2]
##        0.24        0.44        0.80        0.96        1.00
FREQRELACUMPERL<-cumsum(FREQRELPERL)
FREQRELACUMPERL
## [73.2,73.7) [73.7,74.3) [74.3,74.8) [74.8,75.3) [75.3,75.8]
##        0.08        0.16        0.40        0.68        1.00
FREQRELACUMPYTHON<-cumsum(FREQRELPYTHON)
FREQRELACUMPYTHON
## [19.3,19.6) [19.6,19.8) [19.8,20.1) [20.1,20.3) [20.3,20.6]
##        0.04        0.32        0.48        0.80        1.00
```

Por último, calculamos la tabla de frecuencias para cada variable:

```
TABJAVA<-
cbind(MCLASJAVA,FREQABSJAVA,FREQRELJAVA,FREQABSACUMJAVA,FREQRELACUMJAVA)
TABJAVA
##             MCLASJAVA FREQABSJAVA FREQRELJAVA FREQABSACUMJAVA FREQRELACUMJAVA
## [48.5,49.2)     48.87           6        0.24               6            0.24
## [49.2,50)       49.61           5        0.20              11            0.44
```

```
## [50,50.7)       50.35           9        0.36              20            0.80
## [50.7,51.5)     51.09           4        0.16              24            0.96
## [51.5,52.2]     51.83           1        0.04              25            1.00
TABPERL<-
cbind(MCLASPERL,FREQABSPERL,FREQRELPERL,FREQABSACUMPERL,FREQRELACUMPERL)
TABPERL
##             MCLASPERL FREQABSPERL FREQRELPERL FREQABSACUMPERL FREQRELACUMPERL
## [73.2,73.7)    73.483           2        0.08               2            0.08
## [73.7,74.3)    74.009           2        0.08               4            0.16
## [74.3,74.8)    74.535           6        0.24              10            0.40
## [74.8,75.3)    75.061           7        0.28              17            0.68
## [75.3,75.8]    75.587           8        0.32              25            1.00
TABPYTHON<-cbind(MCLASPYTHON,FREQABSPYTHON,FREQRELPYTHON,
             FREQABSACUMPYTHON,FREQRELACUMPYTHON)
TABPYTHON
##             MCLASPYTHON FREQABSPYTHON FREQRELPYTHON FREQABSACUMPYTHON
## [19.3,19.6)      19.436             1          0.04                 1
## [19.6,19.8)      19.688             7          0.28                 8
## [19.8,20.1)      19.940             4          0.16                12
## [20.1,20.3)      20.192             8          0.32                20
## [20.3,20.6]      20.444             5          0.20                25
##             FREQRELACUMPYTHON
## [19.3,19.6)              0.04
## [19.6,19.8)              0.32
## [19.8,20.1)              0.48
## [20.1,20.3)              0.80
## [20.3,20.6]              1.00
```

c) Para calcular la media y la mediana de la variable agrupada, volvemos a calcular las variables agrupadas repitiendo las marcas de clase tantas veces como indican las correspondientes frecuencias absolutas:

```
JAVAAGRUP2<-rep(MCLASJAVA,FREQABSJAVA)
JAVAAGRUP2
##  [1] 48.87 48.87 48.87 48.87 48.87 48.87 49.61 49.61 49.61 49.61 49.61 50.35
## [13] 50.35 50.35 50.35 50.35 50.35 50.35 50.35 51.09 51.09 51.09 51.09
## [25] 51.83
PERLAGRUP2<-rep(MCLASPERL,FREQABSPERL)
PERLAGRUP2
##  [1] 73.483 73.483 74.009 74.009 74.535 74.535 74.535 74.535 74.535 74.535
## [11] 75.061 75.061 75.061 75.061 75.061 75.061 75.061 75.587 75.587 75.587
## [21] 75.587 75.587 75.587 75.587 75.587
PYTHONAGRUP2<-rep(MCLASPYTHON,FREQABSPYTHON)
PYTHONAGRUP2
##  [1] 19.436 19.688 19.688 19.688 19.688 19.688 19.688 19.688 19.940 19.940
## [11] 19.940 19.940 20.192 20.192 20.192 20.192 20.192 20.192 20.192 20.192
## [21] 20.444 20.444 20.444 20.444 20.444
```

La instrucción *summary* nos da la media y la mediana de las variables agrupadas:

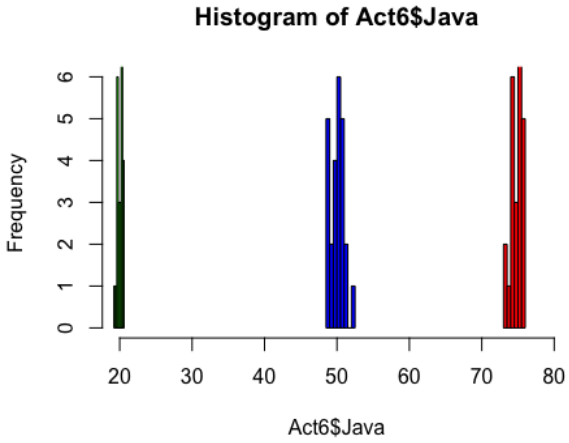
```
summary(JAVAAGRUP2)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.87   49.61   50.35   50.02   50.35   51.83
summary(PERLAGRUP2)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   73.48   74.53   75.06   74.89   75.59   75.59
summary(PYTHONAGRUP2)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.44   19.69   20.19   20.03   20.19   20.44
```

d) Para calcular la desviación típica usamos la instrucción *sd*:

```
sd(JAVAAGRUP2)
## [1] 0.856611
sd(PERLAGRUP2)
## [1] 0.6569738
sd(PYTHONAGRUP2)
## [1] 0.3076052
```

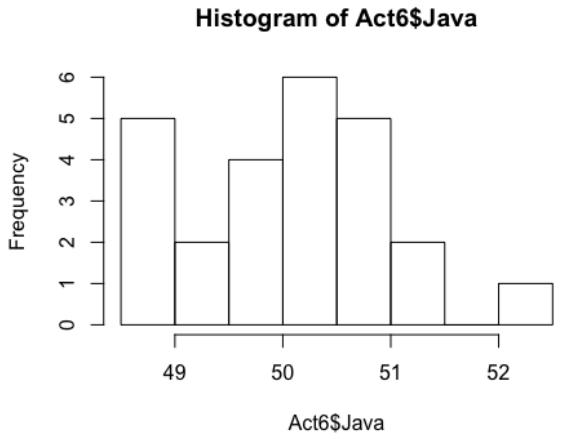
e) Los histogramas se muestran usando la instrucción *hist*. Primero las tres variables juntas para comparar:

```
hist(Act6$Java,xlim=c(20,80),col='blue')
hist(Act6$Perl,xlim=c(20,80),col='red',add=T)
hist(Act6$Python,xlim=c(20,80),col='green',add=T,main="Comparación lenguajes")
```

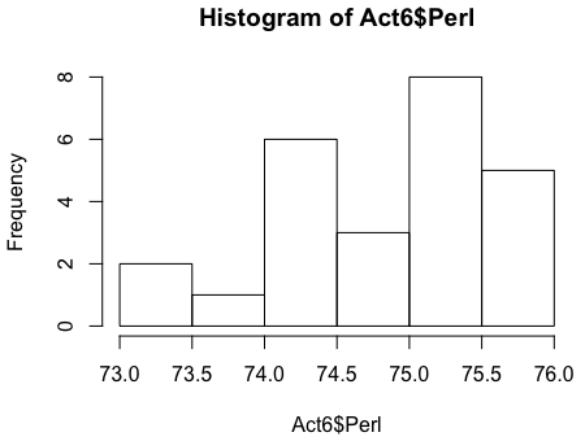


Y después por separado:

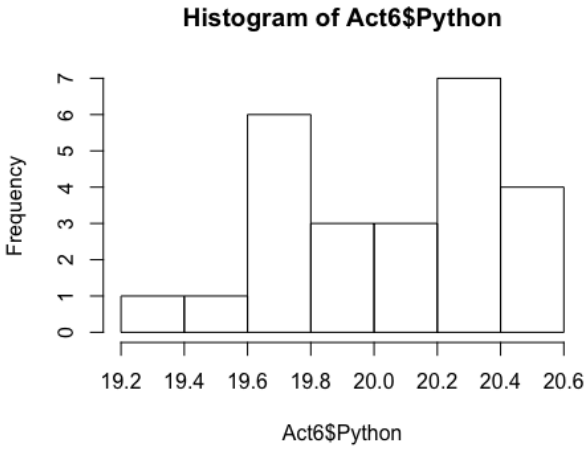
```
hist(Act6$Java)
```




```
hist(Act6$Perl)
```




```
hist(Act6$Python)
```



f) La distribución relativa de Java es la más simétrica y es bimodal, con agrupación en los valores alrededor del 50; la distribución de Perl es más asimétrica a la derecha y muestra que en el momento de realizar la encuesta era el más conocido (valores entre 70 y 80); en cambio, Python era el menos conocido con valores muy concentrados alrededor del 20.

Actividad 7: número de cortes en la red de una empresa de servicios de Internet

Datos estadísticos discretos. Tabla de frecuencias. Diagrama de puntos. Medidas de tendencia central. Medidas de dispersión

Una pequeña empresa que se dedica a dar servicio de Internet tiene durante cincuenta días el número de cortes siguientes en la red: 2, 1, 0, 0, 1, 1, 1, 2, 0, 1, 0, 0, 0, 0, 0, 1, 2, 0, 1, 2, 0, 0, 0, 2, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0. Se pide:

a) Calculad una tabla de frecuencias del número de cortes en la red por día donde se indiquen las frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas.

b) Haced un diagrama de puntos de la variable anterior.

c) Con base en los apartados anteriores comentad cómo es la variable que estamos estudiando (forma, distribución...).
d) Calculad la media, la moda y la mediana del número de cortes diarios de Internet.
e) Calculad también la desviación típica.
f) Comentad el estudio realizado.

Solución

a) En primer lugar, introducimos la variable CTS que nos da el número de CTS durante los cincuenta días:

```
CTS<-c(2,1,0,0,1,1,1,2,0,1,0,0,0,0,0,1,2,0,1,
       2,0,0,0,2,0,1,0,1,0,1,0,0,0,0,0,2,0,0,2,0,0,0,0,0,0,0,0,0,1,0)
```

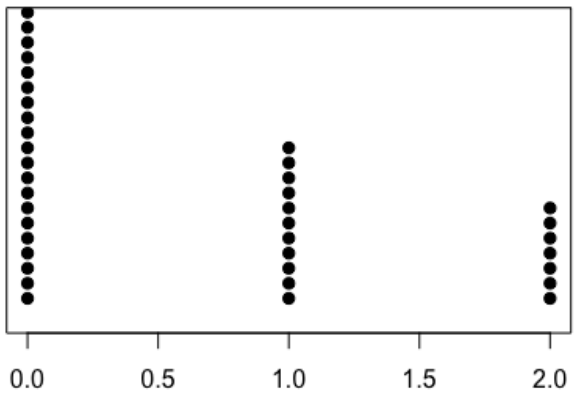
Para hallar las tablas de frecuencias, usamos el mismo procedimiento utilizado en las actividades 4 y 5:

```
resul7<-cbind(table(CTS),prop.table(table(CTS)),
          cumsum(table(CTS)),cumsum(prop.table(table(CTS))))
colnames(resul7) = c("Frec. abs.","Frec. rel.","Frec. abs. acum.","Frec. rel.
acum.")
resul7
##   Frec. abs. Frec. rel. Frec. abs. acum. Frec. rel. acum.
## 0         32       0.64               32             0.64
## 1         11       0.22               43             0.86
## 2          7       0.14               50             1.00
```

Vemos en la tabla anterior que en 32 días no ha habido cortes, en 11 días ha habido un corte y en 7 días ha habido 2 cortes.

b) Para poder realizar un diagrama de puntos, podemos usar la función *stripchart* de R:

```
stripchart(CTS, method = "stack", offset = .5, at = .15, pch = 19)
```



c) La variable anterior tiene una forma asimétrica con una asimetría con cola hacia la derecha.

d) La media y la mediana de la variable CTS valen:

```
mean(CTS)
## [1] 0.5
median(CTS)
## [1] 0
```

e) La desviación típica vale:

```
sd(CTS)
## [1] 0.7354022
```

Vemos que la media vale 0.5, la mediana 0 y la desviación típica vale 0.735. La moda vale claramente 0, como puede observarse en el diagrama de puntos.

f) Concluimos que la mayoría de los días no hay cortes en Internet, lo que se manifiesta en los valores de la mediana y la moda, aunque hay que comentar que la distribución del número de cortes de Internet tiene una desviación típica bastante elevada.

Actividad 8: tiempo de infección de un ordenador por parte de un virus
Agrupamiento de datos estadísticos. Medidas de tendencia central. Medidas de dispersión. Histogramas

El tiempo de infección de una muestra de veinticinco ordenadores con el sistema operativo VENT por parte del virus Malasombra es el siguiente (en segundos): 33.19, 31.59, 30.48, 30.35, 29.44, 29.73, 28.94, 28.57, 32.90, 30.64, 29.32, 30.43, 28.66, 29.02, 29.56, 27.70, 30.93, 30.26, 32.03, 29.28, 31.36, 29.58, 29.74, 28.92, 28.97. Se pide:
a) Agrupad la variable T: "tiempo de infección del ordenador" en cinco intervalos de igual amplitud. Calculad una tabla de frecuencias donde se indiquen las frecuencias absolutas, relativas, absolutas acumuladas y relativas acumuladas.
b) Calculad la media y la mediana de la variable agrupada.
c) Calculad la desviación típica también de la variable agrupada.
d) Haced un histograma de la variable.
e) Comentad los resultados obtenidos.

Solución

a) En primer lugar, metemos los datos en una variable a la que llamamos TINFEC:

```
TINFEC<-c(33.19, 31.59, 30.48, 30.35, 29.44, 29.73,
      28.94, 28.57, 32.90, 30.64, 29.32, 30.43, 28.66, 29.02,
      29.56, 27.70, 30.93, 30.26,
      32.03, 29.28, 31.36, 29.58, 29.74, 28.92, 28.97)
```

A continuación, vamos a agrupar la variable anterior en cinco intervalos.
Primeramente hallamos los intervalos a agrupar:

```
interval<-seq(from=min(TINFEC),to=max(TINFEC),by=(max(TINFEC)-min(TINFEC))/5)
```

Los intervalos serán:

```
interval
## [1] 27.700 28.798 29.896 30.994 32.092 33.190
```

O sea [27.7,28.798), [28.798,29.896), [29.896,30.994), [30.994,32.092) y [32.092,33.190). Para agrupar la variable, usamos la instrucción *cut*:

```
TINFECAGRUP<-cut(TINFEC,breaks<-interval,right=F,include.lowest = T)
TINFECAGRUP
##  [1] [32.1,33.2] [31,32.1)   [29.9,31)   [29.9,31)   [28.8,29.9) [28.8,29.9)
##  [7] [28.8,29.9) [27.7,28.8) [32.1,33.2] [29.9,31)   [28.8,29.9) [29.9,31)
## [13] [27.7,28.8) [28.8,29.9) [28.8,29.9) [27.7,28.8) [29.9,31)   [29.9,31)
## [19] [31,32.1)   [28.8,29.9) [31,32.1)   [28.8,29.9) [28.8,29.9) [28.8,29.9)
## [25] [28.8,29.9)
## Levels: [27.7,28.8) [28.8,29.9) [29.9,31) [31,32.1) [32.1,33.2]
```

Las frecuencias absolutas se hallan usando la instrucción *table*:

```
FREQABS<-table(TINFECAGRUP)
FREQABS
## TINFECAGRUP
## [27.7,28.8) [28.8,29.9)   [29.9,31)   [31,32.1) [32.1,33.2]
##           3          11           6           3           2
```

Aquí aparecen las frecuencias relativas:

```
FREQREL<-table(TINFECAGRUP)/sum(table(TINFECAGRUP))
FREQREL
## TINFECAGRUP
## [27.7,28.8) [28.8,29.9)   [29.9,31)   [31,32.1) [32.1,33.2]
##        0.12        0.44        0.24        0.12        0.08
```

A continuación aparecen las frecuencias absolutas acumuladas:

```
FREQABSACUM<-cumsum(table(TINFECAGRUP))
FREQABSACUM
## [27.7,28.8) [28.8,29.9)   [29.9,31)   [31,32.1) [32.1,33.2]
##           3          14          20          23          25
```

Y por último, las frecuencias relativas acumuladas:

```
FREQRELACUM<-cumsum(table(TINFECAGRUP))/sum(table(TINFECAGRUP))
FREQRELACUM
## [27.7,28.8) [28.8,29.9)   [29.9,31)   [31,32.1) [32.1,33.2]
##        0.12        0.56        0.80        0.92        1.00
```

A continuación, se muestra la tabla de frecuencias:

```
cbind(FREQABS,FREQREL,FREQABSACUM,FREQRELACUM)
##             FREQABS FREQREL FREQABSACUM FREQRELACUM
## [27.7,28.8)       3    0.12           3        0.12
```

```
## [28.8,29.9)      11    0.44          14        0.56
## [29.9,31)         6    0.24          20        0.80
## [31,32.1)         3    0.12          23        0.92
## [32.1,33.2]       2    0.08          25        1.00
```

b) Para hallar los estadísticos pedidos de la variable agrupada, necesitamos hallar las marcas de clase de los intervalos:

```
MCLAS<-c()
for (i in 1:length(interval)-1) {MCLAS<-c(MCLAS,(interval[i+1]+interval[i])/2)}
MCLAS
## [1] 28.249 29.347 30.445 31.543 32.641
```

A continuación, creamos la variable TAGRUP repitiendo cada marca de clase según su frecuencia absoluta:

```
TAGRUP<-rep(MCLAS,FREQABS)
TAGRUP
##  [1] 28.249 28.249 28.249 29.347 29.347 29.347 29.347 29.347 29.347 29.347
## [11] 29.347 29.347 29.347 29.347 30.445 30.445 30.445 30.445 30.445 30.445
## [21] 31.543 31.543 31.543 32.641 32.641
```

La media de la variable agrupada valdrá:

```
mean(TAGRUP)
## [1] 30.0058
```

La mediana valdrá:

```
median(TAGRUP)
## [1] 29.347
```

c) La varianza valdrá:

```
VARI<-var(TAGRUP)*(length(TAGRUP)-1)/length(TAGRUP)
VARI
## [1] 1.446725
```

Hemos multiplicado por $(n-1)/n$, donde n es el número de datos porque R calcula la cuasivarianza (varianza mostral) en lugar de la varianza donde la cuasivarianza se define como: $\tilde{s}^2 = \frac{\sum_{I=1}^{N}(x_i-\bar{x})^2}{n-1}$.

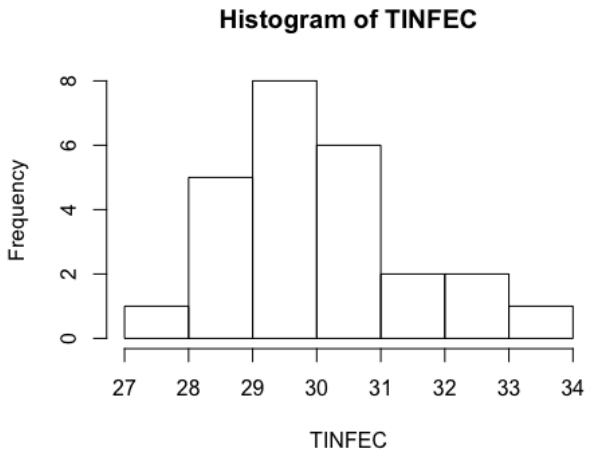
La desviación típica será la raíz cuadrada de la varianza:

```
DESVST<-sqrt(VARI)
DESVST
## [1] 1.202799
```

d) El histograma se realiza con la instrucción *hist* (tiempo):

```
hist(TINFEC)
```

e) Es una variable ligeramente asimétrica a la derecha, con la moda entre 29 y 30 y valores muy concentrados.


Histogram of TINFEC
Frequency
0
2
4
6
8
27
28
29
30
31
32
33
34
TINFEC