PRÁCTICA 03: Distribuciones unidimensionales

DESARROLLO DE LA PRÁCTICA

Para el desarrollo de la práctica se pueden usar las librerías de la instalación básica, así como sus gráficos más sencillos. No obstante, tenemos una sensible mejora si utilizamos alguna librería especial para estos menesteres. Podemos citar a Desctools, Hmisc, sjPlot, stat,.., entre otras, para la elaboración de una tabla estadística y posteriores estadísticos descriptivos de una variable. Nosotros utilizaremos **Desctools**, una librería muy completa para estos objetivos que nos hemos marcado en esta práctica. Para las representaciones gráficas podemos usar las librerías **graphics**, **ggplot2**, **lattice**, …

Conjunto de datos: Se usan los datos del fichero HIPER200.RData para el desarrollo básico de la práctica. Posteriormente se le pedirá al alumno que realice una tarea similar con los datos almacenados en Auto.RData o de otros data frame.

Tablas de frecuencias.

El software R permite la realización de tablas de frecuencias mediante varios paquetes para variables de tipo nominal u ordinal (factor) y para intervalo, razón, discreta o continua (variable numérica).
Si tomamos una variable cualitativa ordinal como la actividad física (act_fisi), podemos obtener su tabla mediante los comandos

```
library(DescTools)
Freq(act_fisi)
```

que produce la siguiente tabla estadística

```
      level  freq   perc  cumfreq  cumperc
1    escasa    74  37.0%       74    37.0%
2  moderada    93  46.5%      167    83.5%
3   intensa    33  16.5%      200   100.0%
```

Si tenemos una variable cualitativa sin orden, como género, tenemos respectivamente (comando y resultados)

```
Freq(genero)
```

```
       level  freq   perc  cumfreq  cumperc
1  masculino   100  50.0%      100    50.0%
2   femenino   100  50.0%      200   100.0%
```

Debemos tener algo de cuidado, pues las dos últimas columnas no tienen sentido, al no haber orden entre las modalidades de la variable.

Cuando se tiene una variable cuantitativa continua, lo usual es agrupar los valores en intervalos de clase y realizar la tabla estadística sobre dichos intervalos. El comando

```
Freq(edad)
```

nos produce la tabla estadística siguiente

	level	freq	perc	cumfreq	cumperc
1	[10,20]	16	8.0%	16	8.0%
2	(20,30]	38	19.0%	54	27.0%
3	(30,40]	27	13.5%	81	40.5%
4	(40,50]	25	12.5%	106	53.0%
5	(50,60]	32	16.0%	138	69.0%
6	(60,70]	28	14.0%	166	83.0%
7	(70,80]	30	15.0%	196	98.0%
8	(80,90]	4	2.0%	200	100.0%

donde el programa ha elegido una amplitud de intervalo y un número de intervalos de acuerdo a unos criterios generales del propio programa. Si observamos la tabla, podemos pensar en reunir los dos últimos intervalos y cambiar la forma de elegir los intervalos (igual que en clase). Para ello escribimos los comandos

```
c1<-c(15,20,30,40,50,60,70,90)
Freq(edad, breaks=c1, right=FALSE)
```

y nos resulta la siguiente tabla estadística

	level	freq	perc	cumfreq	cumperc
1	[15,20)	12	6.0%	12	6.0%
2	[20,30)	38	19.0%	50	25.0%
3	[30,40)	28	14.0%	78	39.0%
4	[40,50)	25	12.5%	103	51.5%
5	[50,60)	34	17.0%	137	68.5%
6	[60,70)	26	13.0%	163	81.5%
7	[70,90]	37	18.5%	200	100.0%

Estadísticos descriptivos

Además, de realizar las tablas estadísticas de las variables seleccionadas en el apartado anterior, podemos solicitar los estadísticos descriptivos más comunes con el comando

```
Desc(edad, plotit=FALSE)
```

```
edad (numeric)

  length       n     NAs  unique      0s    mean  meanCI
     200     200       0      67       0   47.73   44.99
          100.0%   0.0%           0.0%           50.47

     .05     .10     .25  median     .75     .90     .95
   18.00   21.00   29.75   48.00   65.00   74.10   78.00

   range      sd   vcoef     mad     IQR    skew    kurt
   71.00   19.68    0.41   26.69   35.25    0.07   -1.23

lowest : 17.0 (7), 18.0 (4), 19.0, 20.0 (4), 21.0 (6)
highest: 80.0 (4), 81.0, 82.0, 84.0, 88.0
```

que nos muestra casi todos los estadísticos descriptivos comunes de una variable estadística cuantitativa.

Este análisis descriptivo se puede realizar para todas las variables del conjunto de datos (data frame) mediante el comando

Desc(HIPER200)

Si se desea un cuantil determinado para una variable, podemos ejecutar el comando

quantile(edad,0.35)

que nos permite calcular, entre otros, la mediana, cuartiles, deciles y percentiles de la distribución de una variable cuantitativa. Incluso se puede realizar para una variable cualitativa ordinal.

Gráficos: Diagramas de barras, diagramas de sectores e histogramas

Si deseamos acompañar, a la tabla de frecuencias y estadísticos descriptivos, un gráfico de los datos disponemos de los paquetes graphics, ggplot2 y lattice, entre otros disponibles.

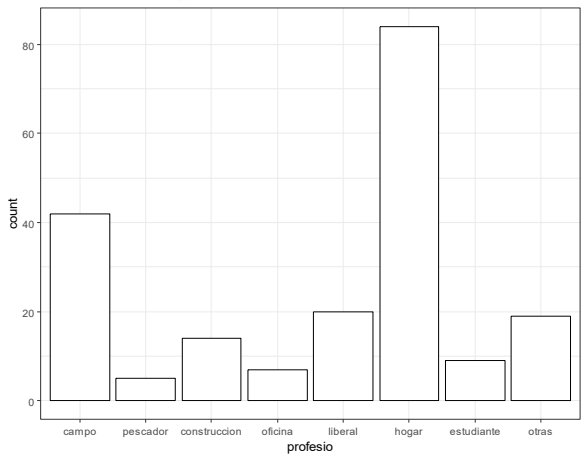
Mediante la librería graphics, de la instalación básica, las siguientes instrucciones nos permiten realizar un diagrama de barras, un diagrama de sectores y un histograma para las variables indicadas.

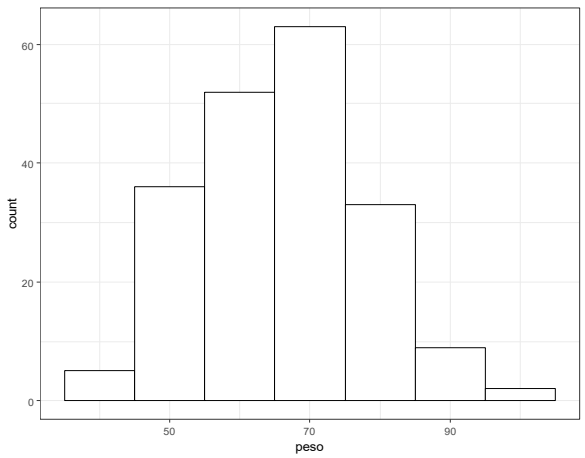
```
barplot(table(estudios), main="Diagrama de barras",sub="Nivel de estudios", space=4, beside=FALSE)
pie(table(profesio), main="Grafico de sectores\nProfesion")
hist(sist_ini, freq=TRUE, breaks=7, col="darkgray", xlab="peso", ylab="frec", main="histograma")
```

Si utilizamos la librería ggplot2, supuestamente mejora al programa básico de graphics, las instrucciones

```
library(ggplot2)
ggplot(HIPER, aes(x=profesio)) + geom_bar(fill="white", colour="black") + theme_bw()
ggplot(HIPER, aes(x=peso)) + geom_histogram(binwidth=10, fill="white", colour="black")
+ theme_bw()
```

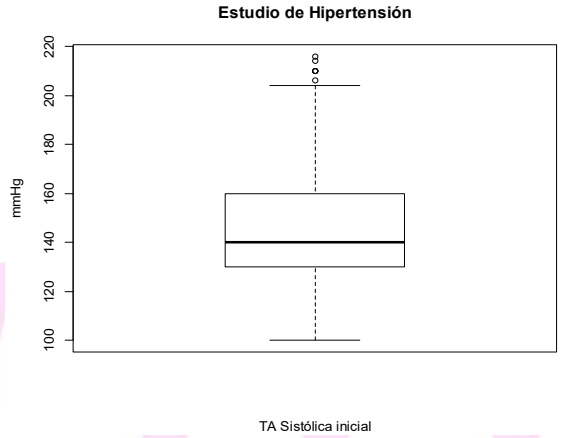
nos permiten realizar las siguientes graficas





Con el paquete “graphics”, se puede crear un diagrama de cajas con la siguiente instrucción.

boxplot(TAsist0, range=1.5,ylab="mmHg", main="Estudio de Hipertensión", sub="TA Sistólica inicial")



Con los tres paquetes de gráficos indicados, se pueden realizar otros tipos de gráficos. A medida que se va siendo más exigente con el gráfico final más compleja se torna la instrucción o conjunto de instrucciones a realizar, todo depende de nuestra exigencia en la calidad del gráfico final.

Tener presente que todo se va a realizar mediante ventanas, donde vamos seleccionando opciones con el puntero del ratón. Lo más usual es que se utilicen otros comandos, que podemos ver en la consola en color rojo. No preocuparse de los comandos que se utilicen, sino de la selección de la ventana y opciones adecuadas. Otra cosa bien distinta es con el RStudio donde ejecutaremos preferentemente los comandos que van en el fichero de instrucciones (script) que acompaña a cada práctica. Como comprenderás existen otras librerías que utilizan otros comandos similares. Como ya he dicho no deben obsesionarse con los comandos. No hay persona que sepa el uso de todos los comandos de R, aparte de ser una tarea imposible de realizar, es al mismo tiempo infructuosa, pues como podrán apreciar con el paso del tiempo (semestres o años) algunos comandos se modifican o algunas librerías desaparecen, otras se incorporan. Esto es un cuerpo de conocimiento estadístico en constante evolución. En mi opinión, es más importante saber manejar la ayuda de RStudio.

EJERCICIOS:

1. En 200 intervalos de tiempo de 5 segundos, se cuenta el número de mensajes, infectados de algún virus, que llegan a un servidor:

1	1	1	2	0	2	0	0	2	2	0	1	0	2	2	0	0	1	2	2	1	2	3	0	0
1	4	0	1	0	0	3	3	2	0	1	0	0	0	3	2	0	3	3	1	1	3	2	0	1
2	1	1	0	0	0	1	0	1	1	0	3	2	0	0	2	1	0	3	1	0	2	1	3	2
1	1	0	3	1	0	0	0	2	0	1	1	3	1	0	0	2	3	2	1	1	1	0	4	1
4	0	1	1	2	0	3	3	1	2	3	2	1	0	1	1	1	2	3	0	2	2	1	3	6
2	0	2	2	0	1	0	2	0	1	3	0	2	0	0	1	0	3	1	1	1	1	3	1	1
2	0	1	4	4	1	2	0	3	0	1	0	0	1	2	3	0	2	3	2	1	2	3	2	0
1	1	1	1	1	2	1	2	2	1	1	1	0	2	0	0	3	0	2	2	0	1	2	3	0

a) Determinar la correspondiente tabla de frecuencias.
b) Hallar media, mediana, moda, varianza y coeficiente de variación.
c) Dibujar un diagrama de barras de frecuencias absolutas y un polígono de frecuencias relativas acumuladas.

2. Los siguientes datos se refieren a la duración (en horas) de las lámparas de distintas unidades de un determinado modelo de cañón retroproyector:

1256	1345	1254	1410	1412
1256	1345	1254	1410	1412
1256	1345	1254	1410	1412
1256	1345	1254	1410	1412
1256	1345	1254	1410	1412
1256	1345	1254	1410	1412
1256	1345	1254	1410	1412

a) Hallar media, moda y mediana.
b) Hallar el percentil 32, el decil 7 y el cuantil 0,8.
c) Agrupando en intervalos de amplitud 25 (partiendo de 1250) construir un histograma de frecuencias absolutas, un histograma de frecuencias relativas y un polígono de frecuencias relativas acumuladas.

3. Los siguientes datos están referidos al tiempo (en centésimas de segundo) de transmisión de 50 señales sobre una red de fibra óptica:

0.25	0.26	0.22	0.31	0.27	0.42	0.45	0.35	0.23	0.47
0.42	0.48	0.19	0.15	0.26	0.33	0.12	0.17	0.38	0.24
0.15	0.17	0.23	0.43	0.35	0.43	0.34	0.22	0.21	0.34
0.16	0.32	0.43	0.48	0.36	0.35	0.37	0.43	0.24	0.32
0.32	0.35	0.35	0.19	0.25	0.45	0.27	0.28	0.32	0.13

a) Agrupar los datos en intervalos de amplitud 0.05 centésima de segundo, empezando por 0.1 .
b) Hallar media, mediana, moda, cuartiles, varianza y coeficiente de variación.
c) Construir histogramas de frecuencias relativas y relativas acumuladas. Construir los correspondientes polígonos de frecuencias.

4. Carga el fichero "iris" con la función data(iris)

a) Con los datos de la primera variable numérica construye una matriz de 15 filas y 10 columnas. Extrae de ella la fila 7, la columna 9 y el elemento (7,9).

b) Construye la tabla de frecuencias de la variable Species.

c) Construye el vector de medias de las variables numéricas.

d) Construye un histograma de la variable Petal.Length.

e) Realiza un estudio detallado de cada variable