Estadística
Taller 2 de R: estadística descriptiva
Profesor: Edwin Santiago Alférez
MACC - Universidad del Rosario

Contents

1 Introducción y objetivos

Supongamos que se tienen los siguientes datos de 1384 pacientes con Gammapatía Monoclonal de Significado Incierto (MGUS, por sus siglas en inglés): ID del paciente (id), edad (age), sexo (sex), hemoglobina (hgb), creatinina (creat), tamaño del suero monoclonal (mspike), tiempo de progresión a la neoplasia maligna -PCM- (ptime), ocurrencia del PCM (pstat, 0=no, 1=si), tiempo hasta la muerte (futime) y ocurrencia de la muerte (death, 0=no, 1=si). A continuación se muestran los valores de las primeras 20 observaciones.

```
##    id age sex  hgb creat mspike ptime pstat futime death
## 1   1  88   F 13.1   1.3    0.5    30     0     30     1
## 2   2  78   F 11.5   1.2    2.0    25     0     25     1
## 3   3  94   M 10.5   1.5    2.6    46     0     46     1
## 4   4  68   M 15.2   1.2    1.2    92     0     92     1
## 5   5  90   F 10.7   0.8    1.0     8     0      8     1
## 6   6  90   M 12.9   1.0    0.5     4     0      4     1
## 7   7  89   F 10.5   0.9    1.3   151     0    151     1
## 8   8  87   F 12.3   1.2    1.6     2     0      2     1
## 9   9  86   F 14.5   0.9    2.4    57     0     57     0
## 10 10  79   F  9.4   1.1    2.3   136     0    136     1
## 11 11  86   M 11.8   1.0    2.3     2     0      2     1
## 12 12  89   F 11.3   1.3    1.2   108     0    108     1
## 13 13  87   M 11.2   1.1    1.3    10     0     10     1
## 14 14  80   F 13.1   1.0    1.3    14     0     14     1
## 15 15  85   M 13.0   1.1    1.0    18     0     18     1
## 16 16  90   F 14.1   1.2    0.5    43     0     43     1
## 17 17  94   F 11.0   1.1    0.7    34     0     34     1
## 18 18  86   M 16.0   1.5    1.9    67     0     67     1
## 19 19  82   M 14.6   1.0    1.6    16     0     16     1
## 20 20  73   M 13.1   1.2    0.5    62     0     62     1
```

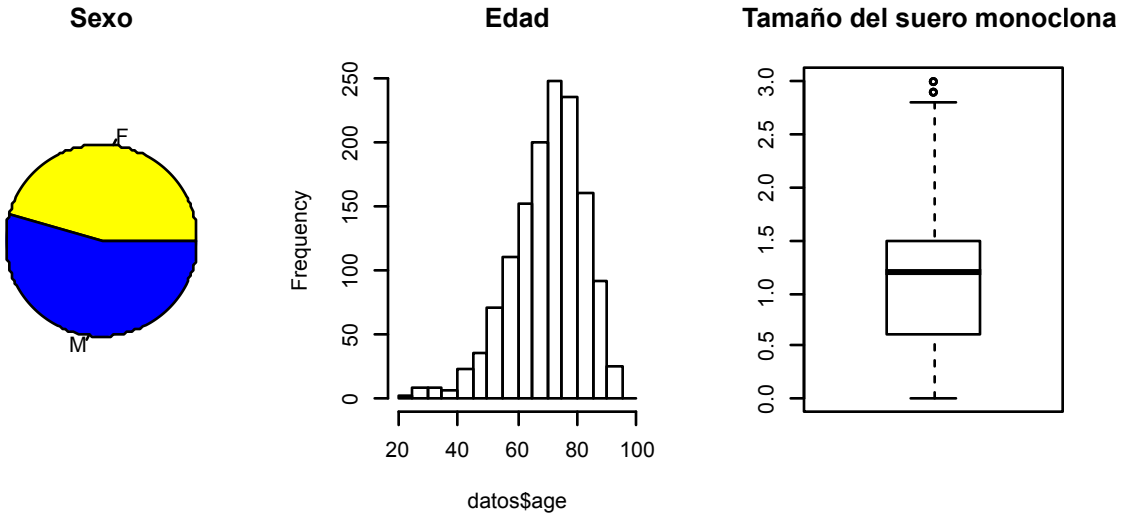


Figure 1: Información resumida y organizada de los resultados de 1341 pacientes con MGUS.

A simple vista, ¿qué información relevante podemos ver en el fichero?, ¿qué podemos concluir?. Es evidente que el primer paso cuando se trabajan con datos es la exploración inicial. Los datos se han de organizar, representar de alguna forma más amena y resumirlos. Por ejemplo, representar gráficamente (como se muestra en la figura) la proporción entre hombres y mujeres, analizar el rango de edad más común con la enfermedad, valor promedio de hemoglobina, etc.

En esta sesión se hace una introducción a las técnicas básicas para organizar, representar y resumir un conjunto de datos. En la estadística matemática, se conoce como **análisis exploratorio de datos** o **estadística descriptiva**. Además, se presentan las diferentes formas de aplicar la estadística descriptiva utilizando *R*, *Rstudio* y *R-Commander* en un conjunto de datos para su mejor interpretación. Al finalizar esta sesión, el alumno ha de ser capaz de:

- Identificar las principales maneras de describir, organizar, representar y resumir un conjunto de datos.
- Construir tablas de frecuencias, representarlas gráficamente, calcular algunos estadísticos importantes (media aritmética, varianza y moda) e interpretar todos estos resultados en *R*, *Rstudio* y *R-Commander*.

2 Estadística Descriptiva: Teoría y aplicación con *R-console* y *Rstudio*

La estadística descriptiva es la disciplina de la estadística que se encarga de organizar y resumir información cuantitativa para describir las características principales de un conjunto de datos. Frecuentemente, el conjunto de datos incluye diferentes **variables** (por ejemplo: velocidad, resistencia, elasticidad, etc). Por lo tanto, lo más usual es considerar las variables de una en una, sin tener posible relación entre ellas. Según las características de esas variables, se pueden encontrar **variables cualitativas** o **categóricas** (NO necesitan números para expresarse, por ejemplo: sexo, color, etc) y **variables cuantitativas** o **numéricas** (SI necesitan números para expresarse, por ejemplo: edad, longitud, etc). Por cada variable hay una serie de observaciones, las anotaciones sobre qué modalidad (cualitativas) o qué valor (cuantitativas) tiene cada observación se denominan **datos**. Estos datos se pueden organizar, resumir y representar mediante:

- **Tablas:** Matrices donde se guardan los datos que toma una determinada variable para cada objeto. Por ejemplo, tablas de frecuencia.
- **Gráficos:** Representaciones visuales de las tablas que otorgan una visión más general y completa de los datos. Por ejemplo, gráficos de barras, histogramas, gráficos sectoriales y polígonos de frecuencia.

• **Medidas de tendencia central:** Valores que pretenden dar información sobre el centro de la distribución de datos. Algunos ejemplos son la media, la mediana y la moda.

• **Medidas de variabilidad:** Valores que pretenden dar información sobre la homogeneidad de los valores entre sí. Algunos ejemplos son la desviación estándar, la varianza y los cuartiles.

2.1 Tabla de frecuencia

El modo más simple de presentar ordenadamente datos categóricos es por medio de una tabla de frecuencias. Esta tabla indica el número de repeticiones de cada una de las clases de la variable cualitativa. Se pueden distinguir los siguientes tipos de frecuencias:

• **Frecuencia absoluta** (n_i)**:** Es el número de repeticiones que presenta una observación.

• **Frecuencia relativa** (f_i)**:** Es la frecuencia absoluta dividida por el número total de datos.

• **Frecuencia absoluta acumulada** (N_i)**:** Es la suma de los distintos valores de la frecuencia absoluta tomando como referencia un individuo dado. La última frecuencia absoluta acumulada es igual al número de casos.

• **Frecuencia relativa acumulada** (F_i)**:** Es el resultado de dividir cada frecuencia absoluta acumulada por el número total de datos.

Ejemplo 1: El conjunto de datos para el control de calidad del agua de diferentes reactores es el siguiente, donde cada número representa el reactor que se eligió como el mejor:

$$1, 5, 3, 1, 2, 3, 4, 5, 1, 4, 2, 4, 4, 5, 1, 4, 2, 4, 2, 2$$

La tabla de frecuencias es:

Reactor	Frec. absoluta	Frec. relativa	Frec. abs. acumulada	Frec. rel. acumulada
1	4	0.20	4	0.20
2	5	0.25	9	0.45
3	2	0.10	11	0.55
4	6	0.30	17	0.85
5	3	0.15	20	1.00

En *R*, la tabla de frecuencias se puede calcular de la siguiente manera:

```
datos_1 = c(1,5,3,1,2,3,4,5,1,4,2,4,4,5,1,4,2,4,2,2)
ni = table(datos_1) # Frecuencia absoluta
fi = table(datos_1)/length(datos_1) # Frecuencia relativa
Ni = cumsum(ni) # Frecuencia absoluta acumulada
Fi = cumsum(fi) # Frecuencia relativa acumulada
Tabla_Frec = cbind(ni,fi,Ni,Fi) # Se crea una tabla con todas las frecuencias
Tabla_Frec # Se visualiza la tabla
```

```
##   ni   fi Ni   Fi
## 1  4 0.20  4 0.20
## 2  5 0.25  9 0.45
## 3  2 0.10 11 0.55
## 4  6 0.30 17 0.85
## 5  3 0.15 20 1.00
```

Ejemplo 2: Las resistencias a la compresión de la aleación en libras por pulgada cuadrada (psi) de 80 especímenes de una nueva aleación de aluminio-litio sometida a evaluación como material posible para

elementos estructurales de aeronaves son:

$$105, 221, 183, 186, 121, 181, 180, 143, 167, 141, 97, 154, 153, 174, 120, 168, 176, 110, 158, 133,$$
$$245, 228, 174, 199, 181, 158, 156, 123, 229, 146, 163, 131, 154, 115, 160, 208, 158, 169, 148, 158,$$
$$207, 180, 190, 193, 194, 133, 150, 135, 118, 149, 134, 178, 76, 167, 184, 135, 218, 157, 101, 171,$$
$$165, 172, 199, 151, 142, 163, 145, 171, 160, 175, 149, 87, 160, 237, 196, 201, 200, 176, 150, 170$$

Cuando los valores de la variable son muchos, conviene agrupar los datos en intervalos o clases para así realizar un mejor análisis e interpretación de ellos. Para construir una tabla de frecuencias con datos agrupados, conociendo los intervalos, se deben determinar la frecuencias correspondientes a cada intervalo.

	ni	fi	Ni	Fi
70 <= x < 90	2	0.0250	2	0.0250
90 <= x < 110	3	0.0375	5	0.0625
110 <= x < 130	6	0.0750	11	0.1375
130 <= x < 150	14	0.1750	25	0.3125
150 <= x < 170	22	0.2750	47	0.5875
170 <= x < 190	17	0.2125	64	0.8000
190 <= x < 210	10	0.1250	74	0.9250
210 <= x < 230	4	0.0500	78	0.9750
230 <= x < 250	2	0.0250	80	1.0000

En *R*, la tabla de frecuencias con datos agrupados se puede calcular de la siguiente manera:

```
datos_2=c(105,221,183,186,121,181,180,143,167,141,97,154,153,174,120,168,176,110,158,133,
          245,228,174,199,181,158,156,123,229,146,163,131,154,115,160,208,158,169,148,158,
          207,180,190,193,194,133,150,135,118,149,134,178,76,167,184,135,218,157,101,171,
          165,172,199,151,142,163,145,171,160,175,149,87,160,237,196,201,200,176,150,170)
breaks = seq(70,250,by=20); breaks # Se crea el vector que contiene los intervalos
```

```
##  [1]  70  90 110 130 150 170 190 210 230 250
```

```
datos_2a = cut(datos_2, breaks, right=FALSE); datos_2a # Asigna c/valor a un intervalo
```

```
##  [1] [90,110)  [210,230) [170,190) [170,190) [110,130) [170,190) [170,190)
##  [8] [130,150) [150,170) [130,150) [90,110)  [150,170) [150,170) [170,190)
## [15] [110,130) [150,170) [170,190) [110,130) [150,170) [130,150) [230,250)
## [22] [210,230) [170,190) [190,210) [170,190) [150,170) [150,170) [110,130)
## [29] [210,230) [130,150) [150,170) [130,150) [150,170) [110,130) [150,170)
## [36] [190,210) [150,170) [150,170) [130,150) [150,170) [190,210) [170,190)
## [43] [190,210) [190,210) [190,210) [130,150) [150,170) [130,150) [110,130)
## [50] [130,150) [130,150) [170,190) [70,90)   [150,170) [170,190) [130,150)
## [57] [210,230) [150,170) [90,110)  [170,190) [150,170) [170,190) [190,210)
## [64] [150,170) [130,150) [150,170) [130,150) [170,190) [150,170) [170,190)
## [71] [130,150) [70,90)   [150,170) [230,250) [190,210) [190,210) [190,210)
## [78] [170,190) [150,170) [170,190)
## 9 Levels: [70,90) [90,110) [110,130) [130,150) [150,170) ... [230,250)
```

```
ni = table(datos_2a) # Frecuencia absoluta
fi = table(datos_2a)/length(datos_2a) # Frecuencia relativa
Ni = cumsum(ni) # Frecuencia absoluta acumulada
Fi = cumsum(fi) # Frecuencia relativa acumulada
```

```r
Tabla_Frec = cbind(ni,fi,Ni,Fi) # Se crea una tabla con todas las frecuencias
Tabla_Frec # Se visualiza la tabla
```

```
##           ni     fi Ni     Fi
## [70,90)    2 0.0250  2 0.0250
## [90,110)   3 0.0375  5 0.0625
## [110,130)  6 0.0750 11 0.1375
## [130,150) 14 0.1750 25 0.3125
## [150,170) 22 0.2750 47 0.5875
## [170,190) 17 0.2125 64 0.8000
## [190,210) 10 0.1250 74 0.9250
## [210,230)  4 0.0500 78 0.9750
## [230,250)  2 0.0250 80 1.0000
```

Tips & Tricks !!!

- table() crea resultados tabulares de variables categóricas, o sea, determina la frecuencia absoluta de los datos.
- cumsum() calcula un vector cuyos elementos son la suma acumulada del vector de entrada.
- cbind() y rbind() combinan varios objetos de R en un solo objeto: por columnas y por filas respectivamente.
- cut() divide el rango del vector "datos" en los intervalos "breaks" y codifica los valores de los datos de acuerdo al intervalo que pertenecen.

2.2 Gráficas estadísticas

Las distribuciones de frecuencias se pueden presentar en tablas como las anteriores, o bien en gráficas. La representación gráfica se utiliza para facilitar la comprensión de los resultados, pero no añade ninguna información extra sobre la que contendría una tabla de frecuencias. Sin embargo, ya lo dice el dicho popular "Vale más una imagen que mil palabras". Existen diversos tipos de gráficas, cada una de ellas adecuada a cierto tipo de variables, a continuación se describen las más usadas y cómo se realizan en *R* utilizando los dos ejemplos anteriores.

2.2.1 Gráfico de tallos y hojas

Es una herramienta que presenta una tabla de datos en un formato gráfico para ayudar a visualizar la forma de la distribución. Es una tabla donde cada dato es dividido según su valor, en un tallo y una hoja. El último dígito del dato representa la hoja, y el resto de dígitos representan el tallo. Este tipo de gráficos otorgan información sobre la localización, dispersión y valores extremos de nuestros datos. El gráfico de tallos y hojas se calcula en R mediante la función stem(). La longitud del gráfico se puede modificar utilizando el atributo scale= , donde 1 es valor por defecto, 2 produce un gráfico aproximadamente el doble de largo, etc. El gráfico del ejemplo 2 se genera de la siguiente forma:

```r
stem(datos_2,scale=2)
```

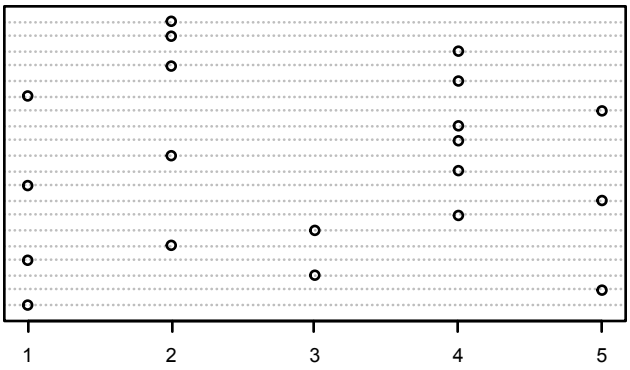
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    7 | 6
##    8 | 7
##    9 | 7
##   10 | 15
##   11 | 058
```

```
##   12 | 013
##   13 | 133455
##   14 | 12356899
##   15 | 001344678888
##   16 | 0003357789
##   17 | 0112445668
##   18 | 0011346
##   19 | 034699
##   20 | 0178
##   21 | 8
##   22 | 189
##   23 | 7
##   24 | 5
```

2.2.2 Gráfico de puntos

Cuando se tiene un tabla de frecuencias pequeña de variables categóricas y los valores no distan mucho entre sí, es una atractiva forma de representar los datos obtenidos. En este gráfico el eje horizontal representa los posibles valores de los datos y el eje vertical corresponde a la localización de cada dato dentro de la lista. Cada dato se representa con un punto y se coloca encima del valor que le corresponda y a una altura proporcional al orden que tiene en el conjunto. En ejemplo 1, el primer dato (1) se representa por un punto en la posición (1,1), el segundo dato (5), por un punto en la posición (5,2), el tercero que es 3, en la posición (3,3), el cuarto que es 1, en (1,4) y así respectivamente. El gráfico de puntos se genera con el comando dotchart() tal como se muestra a continuación.

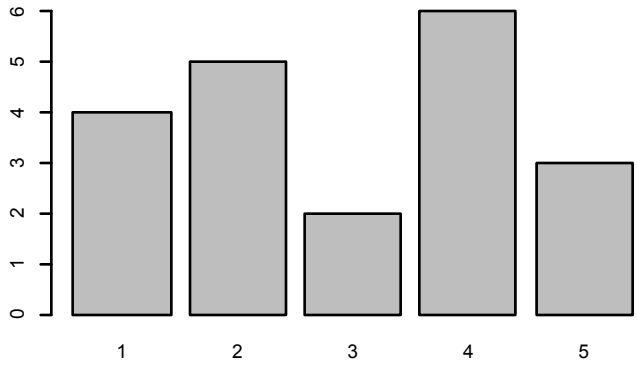
```
dotchart(datos_1)
```



2.2.3 Gráfico de barras

Este gráfico representa visualmente la frecuencia de variables categóricas mediante barras rectangulares de igual anchura. A cada categoría o clase de la variable se le asocia una barra cuya altura representa la frecuencia absoluta o la frecuencia relativa de esa clase. Para generar el gráfico de barras, se utiliza el comando 'barplot(), sin embargo, es necesario definir primero la tabla de frecuencias. Para el ejemplo 1, el gráfico de barras para las frecuencia absoluta es el siguiente:

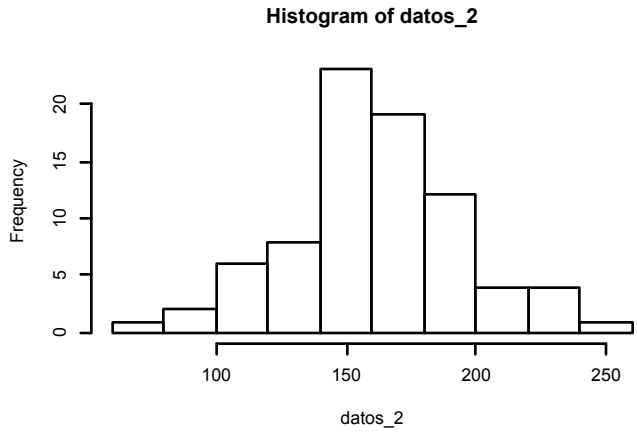
```
barplot(table(datos_1))
```



2.2.4 Histograma

Es la gráfica adecuada para representar variables cuantitativas con un gran número de valores distintos. Los datos se agrupan en intervalos y se representan gráficamente por rectángulos yuxtapuestos cuyas bases descansan sobre el eje horizontal y cuyas alturas son tales que el área de cada rectángulo sea proporcional a la frecuencia de cada intervalo. Si todos los intervalos tienen igual longitud, entonces la altura de cada rectángulo es proporcional a la frecuencia el intervalo. Para evitar confusiones, la principal diferencia con el gráfico de barras es la inexistencia de espacios entre rectángulos. La función hist() permite hacer el histograma de unos datos y además modificar la longitud de los intervalos si se desea. A diferencia del gráfico de barras, la función calcula automáticamente la frecuencia del intervalo. El histograma del ejemplo 2 se genera de la siguiente forma:

```
h=hist(datos_2)
```



Si el único argumento de la función es el vector de datos, el histograma se realiza con el número de intervalos (y por lo tanto su longitud) calculados de forma automática. Si el histograma se guarda en un objeto h = hist(), éste objeto contiene información como los límites de los intervalos, la frecuencia de cada intervalo, su densidad, el punto medio, etc.

```
h$breaks # Límites de los intervalos
```

```
##  [1]  60  80 100 120 140 160 180 200 220 240 260
```

```
h$counts # Frecuencia de cada intervalo
```

```
##  [1]  1  2  6  8 23 19 12  4  4  1
```

```
h$density # Densidad de cada intervalo
```

```
##  [1] 0.000625 0.001250 0.003750 0.005000 0.014375 0.011875 0.007500
```

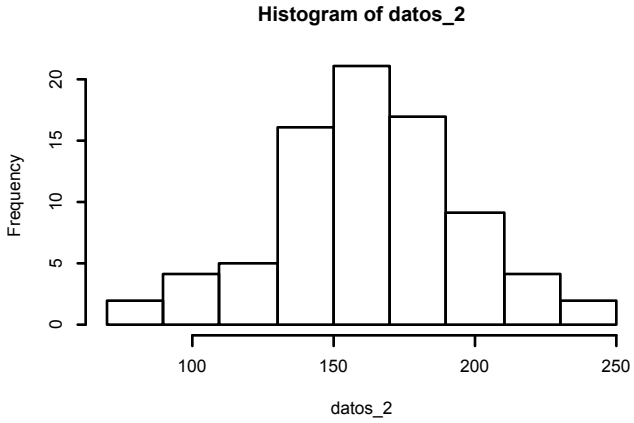
```
##  [8] 0.002500 0.002500 0.000625
```

```
h$mids # Punto central de cada intervalo
```

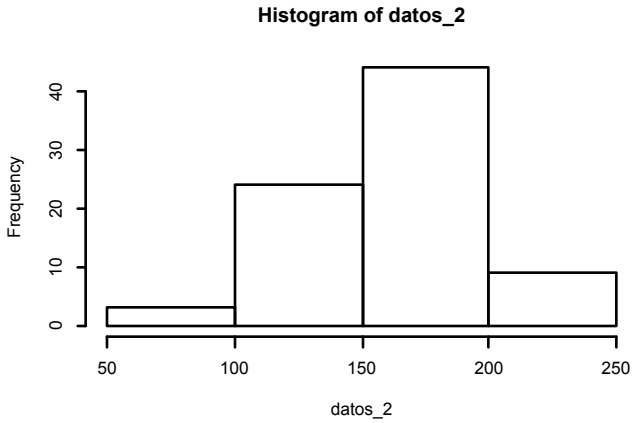
```
##  [1]  70  90 110 130 150 170 190 210 230 250
```

También se puede seleccionar los límites de los intervalos o el número de intervalos en que se quieren agrupar.

```
new_breaks = seq(70,250,by=20)
h1 = hist(datos_2,breaks=new_breaks)
```



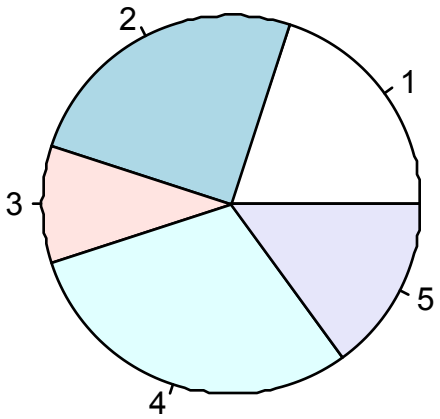

```
h2=hist(datos_2,breaks = 3)
```



2.2.5 Gráfico de sectores

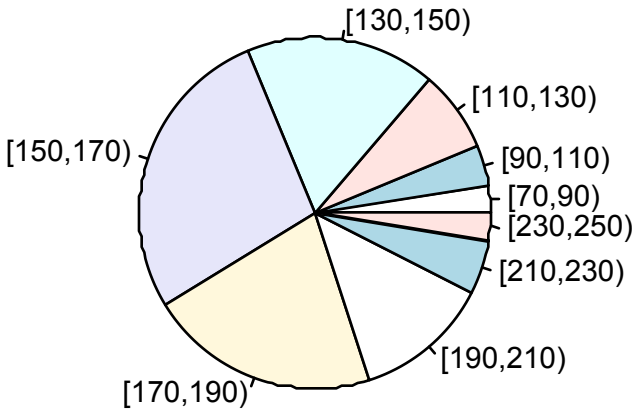
Este gráfico se representa como un círculo dividido en porciones, siendo éstas proporcionales a la frecuencia relativa de cada categoría. La función pie() permite realizar el gráfico de sectores. Tal como en el gráfico de barras, es necesario definir previamente la tabla de frecuencias. Para el ejemplo 1, el diagrama se realiza de la siguiente forma:

```
pie(table(datos_1))
```



Si se quiere hacer el gráfico de sectores para el ejemplo 2, los datos se tienen que agrupar para poder visualizar información relevante.

```
pie(table(datos_2a))
```



Tips & Tricks !!!

- Las funciones stem(), dotchart(), barplot(), hist() y pie() permiten resumir visualmente los datos.
- Estos gráficos se pueden mejorar definiendo algunos atributos, como por ejemplo: col , main , names.arg , etc. Utiliza la ayuda para conocer un poco más sobre ellos.

2.3 Medidas de posición y tendencia central

En ocasiones es conveniente resumir la información de un conjunto de datos numéricos en un solo valor para obtener indicadores del comportamiento de la variable y poder realizar comparaciones. Las medidas de tendencia central, también conocidas como medidas de posición o localización, describen un valor alrededor del cual se encuentran las observaciones.

2.3.1 Media

También conocida como valor promedio, se define como la suma de todos los valores de cada observación (x_i) dividido por el número total de observaciones del conjunto de datos (N).

$$\bar{X} = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Si se dispone de un conjunto de datos agrupados en los que se conoce el valor medio de cada intervalo ($\bar{x_i}$) y el número de datos de cada uno de ellos (n_i) , la media está dada por:

$$\bar{X} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4 + \cdots + x_N n_N}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i n_i$$

donde $n_1 + n_2 + n_3 + n_4 + \cdots + n_n = N$. Para los ejemplos anteriores, las medias se puede calcular de acuerdo a la definición de la siguiente manera:

```
sum(datos_1)/length(datos_1)
```

```
## [1] 2.95
```

```
sum(datos_2)/length(datos_2)
```

```
## [1] 162.6625
```

Sin embargo, la función mean() calcula la media directamente.

```
mean(datos_1)
```

```
## [1] 2.95
```

```
mean(datos_2)
```

```
## [1] 162.6625
```

2.3.2 Mediana

La mediana es el dato que ocupa la posición central en la muestra ordenada de menor a mayor, es un punto que divide la muestra ordenada en dos grupos iguales (deja el 50% de los valores por debajo y el otro 50% por encima). Para calcularla se ordenan los datos de menor a mayor, el dato central es el que ocupa la posición $\frac{N+1}{2}$ donde N es el número total de datos. Si N es impar, la mediana es el mismo dato central, si N es par, existen dos datos centrales, por lo tanto la mediana es el promedio de estos dos. Igualmente, existe una función que la calcula directamente: median().

```
median(datos_1)
```

```
## [1] 3
```

```
median(datos_2)
```

```
## [1] 161.5
```

2.3.3 Moda

La moda es el valor con mayor frecuencia absoluta en los datos obtenidos, indica cual es el valor más frecuente pero no cuántas veces se repite. Si existen más de dos valores que se repiten con mayor frecuencia, se dice que los datos son multimodales. Se puede calcular la mediante usando las siguientes instrucciones:

```
table(datos_1)
```

```
## datos_1
## 1 2 3 4 5
## 4 5 2 6 3
```

```r
#Se organiza la tabla de frecuencias de mayor valor (el más frecuente) a menor
freq_ord=sort(table(datos_1),decreasing = TRUE); freq_ord
```

```
## datos_1
## 4 2 1 5 3
## 6 5 4 3 2
```

```r
# Se toma el valor/es que más se repite/n (el primero de la tabla ordenada)
moda = names(freq_ord[1]); moda
```

```
## [1] "4"
```

En *R* básico (sin cargar librerías o biblioteca adicionales) no existe una función para calcular la moda, sin embargo, si se instala install.packages() y se carga library() la biblioteca modes, se puede calcular mediante la función modes().

```r
#install.packages("modes")
library(modes)
modes(datos_1)
```

```
##        [,1]
## Value     4
## Length    6
```

```r
modes(datos_2)
```

```
##        [,1]
## Value   158
## Length    4
```

2.3.4 Cuantiles

Los cuantiles son valores de la lista de datos que la dividen en partes iguales, es decir, en intervalos, que comprenden el mismo número de valores. Los más usados son los percentiles, los deciles y los cuartiles. Los percentiles son 99 valores que dividen en cien partes iguales el conjunto de datos ordenados. Por ejemplo, el percentil de orden 15 deja por debajo al 15% de las observaciones, y por encima queda el 85%. Los deciles son los nueve valores que dividen al conjunto de datos ordenados en diez partes iguales, son un caso particular de los percentiles. Los cuartiles son los tres valores que dividen al conjunto de datos ordenados en cuatro partes iguales, son también un caso particular de los percentiles. En R, cualquiera de estos se calcula con la función quantile(), donde adicionalmente se ha de especificar el cuantil o cuantiles deseados (como un valor entre 0 y 1) de la siguiente forma:

```r
quantile(datos_2,0.95) # Percentil de orden 95
```

```
##    95%
## 221.35
```

```r
quantile(datos_2,seq(0.1,0.9,by=0.1)) # Todos los deciles
```

```
##   10%   20%   30%   40%   50%   60%   70%   80%   90%
## 119.8 135.0 149.0 156.6 161.5 170.4 176.6 186.8 201.6
```

```r
quantile(datos_2,seq(0.25,0.75,by=0.25)) # Todos los cuartiles
```

```
##   25%   50%   75%
## 144.5 161.5 181.0
```

Finalmente, el rango intercuartílico es la extensión cubierta por la mitad central de los datos ordenados, excluyendo la cuarta parte inicial (los que son inferiores al primer cuartil) y la cuarta parte final (los que son superiores al tercer cuartil). La función IQR() calcula directamente el rango intercuartílico.

```
quantile(datos_2,0.75) - quantile(datos_2,0.25)
```

```
##  75%
## 36.5
```

```
IQR(datos_2)
```

```
## [1] 36.5
```

La media, mediana, mínimo, máximo y cuartiles se pueden calcular directamente mediante la función summary().

```
summary(datos_2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    76.0   144.5   161.5   162.7   181.0   245.0
```

2.4 Medidas de variabilidad y dispersión

Las medidas de posición dan una idea de dónde se encuentra el centro de la distribución, pero no nos dicen cuán disperso es el conjunto de datos. Las medidas de dispersión o variabilidad describen que tan cerca se encuentran los datos entre ellos, o de alguna medida de tendencia central.

2.4.1 Rango

Es el intervalo entre el valor máximo y el valor mínimo del conjunto de datos. Es altamente sensible a los valores extremos, es decir, es un parámetro estadístico débil. Con la función range() también se obtienen el mínimo y el máximo valor del conjunto de datos, por lo tanto, para calcular el rango, solo hace falta calcular su diferencia.

```
max(datos_1)-min(datos_1)
```

```
## [1] 4
```

```
max(datos_2)-min(datos_2)
```

```
## [1] 169
```

```
diff(range(datos_1))
```

```
## [1] 4
```

```
diff(range(datos_2))
```

```
## [1] 169
```

2.4.2 Varianza y desviación típica

Estas medidas miden cuán lejos difieren los datos de la media. Específicamente, expresan "el promedio de la distancia de cada punto respecto de la media". La varianza se calcula según:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

donde x_i es el valor de cada observación, $\overline{X}$ es la media y N es el número total de datos. Nótese que las unidades de la varianza están expresadas al cuadrado, por lo tanto, si se tienen datos de longitud (en mm), la varianza resulta con unidades de superficie (en mm^2), lo cual no tiene mucho sentido. Por lo tanto, se dispone de la desviación estándar o típica que no es más que la raíz cuadrada de la varianza, de esta forma, las unidades de la medida de dispersión son las mismas de los datos.

$$\sigma = \sqrt{\sigma^2} \sqrt{\frac{1}{N-1} \sum_{i=1}^{n} (x_i - \bar{X})^2}$$

Para el ejemplo 1, la varianza y desviación típica se puede calcular usando la definición de la siguiente forma:

```
sum((datos_1-mean(datos_1))^2)/length(datos_1) # Varianza
```

```
## [1] 1.9475
```

```
sqrt(sum((datos_1-mean(datos_1))^2)/length(datos_1)) # Desviación típica
```

```
## [1] 1.395529
```

En R la varianza y desviación estándar se pueden calcular mediante las funciones var () y sd() respectivamente, sin embargo, estas funciones utilizan $N-1$ (o $\sqrt{N-1}$) en el denominador en lugar de N (o $\sqrt{N}$) para poderlas usar como estimadores no sesgados en inferencia estadística. Estas medidas se conocen como varianza y desviación típica corregidas. Por lo tanto, para conocer la varianza y desviación típica sin corregir, se tienen que multiplicar por los factores $\frac{N-1}{N}$ y $\sqrt{\frac{N-1}{N}}$ respectivamente.

```
N = length(datos_1)
var(datos_1) # Varianza
```

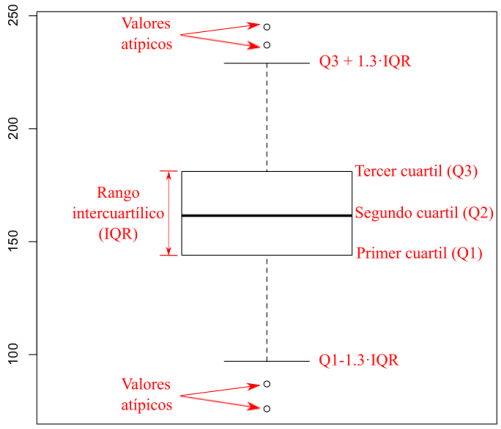
```
## [1] 2.05
```

```
sd(datos_1) # Desviación
```

```
## [1] 1.431782
```

2.5 Gráfico de caja

Los diagramas de caja son una presentación visual que describe varias características importantes al mismo tiempo, tales como la tendencia central, dispersión y simetría. Para su realización se representan los tres cuartiles y los valores mínimo y máximo de los datos sobre un rectángulo, alineado horizontal o verticalmente. Los valores con dispersión hasta 1.3 veces el rango intercuartílico se representan como unas líneas rectas o bigotes. Los valores fuera de ese intervalo se representan mediante puntos y se consideran valores extremos atípicos.



boxplot() es la función que se utiliza para la creación del gráfico. Tal cual como con el histograma, si se guarda el gráfico de caja en un objeto h = boxplot(), éste objeto contiene información como los límites para considerar los valores atípicos, cuáles son esos valores atípicos, los cuartiles, etc.

```
bp=boxplot(datos_2); bp
```

```
## $stats
##       [,1]
## [1,]  97.0
## [2,] 144.0
## [3,] 161.5
## [4,] 181.0
## [5,] 229.0
##
## $n
## [1] 80
##
## $conf
##         [,1]
## [1,] 154.964
## [2,] 168.036
##
## $out
## [1] 245  76  87 237
##
## $group
## [1] 1 1 1 1
##
## $names
## [1] "1"
```

Tips & Tricks !!!

- Las funciones mean(), median(), quantile(), IQR(), var(), sd() y boxplot() permiten nos dan información de tendencia central y variabilidad de los datos.
- Recuerde que puede consultar más información sobre cada función mediante la instrucción ?NombreDeLaFuncion, por ejemplo ?boxplot.

Ejercicios propuestos

Los siguientes datos se extrajeron de la revista Motor Trend 1974 de Estados Unidos, resume el consumo y 10 aspectos de diseño y rendimiento de 32 automóviles (modelos 1973-74). Este conjunto de datos, que se llama mtcars, contiene 11 variables con 32 observaciones y está almacenado en R . Para poder trabajar con ellos, solo hace falta adjudicarle un nombre al objeto, como por ejemplo:

```
a = mtcars
```

Las variables son las siguientes:

- mpg: Millas por galón de combustible
- cyl: Número de cilindros
- disp: Desplazamiento
- hp: Caballos de potencia
- drat: Relación del eje trasero
- wt: Peso (1000 lbs)
- qsec: Tiempo a 1/4 milla
- vs: V/S
- am: Transmisión (0 = automático, 1 = manual)
- gear: Número de marchas adelante
- carb: Número de carburadores

Una vez cargado el conjunto de datos proceda a la resolución del cuestionario.

1. Determine la media, la mediana, la moda y la desviación estándar de cada una de las variables. Se puede calcular a todas la variables? a cuales no? Justifique su respuesta
2. Determinar qué variable presenta valores atípicos, cómo los ha encontrado?
3. Hacer un gráfico de sectores para cada una de las variables. El gráfico de cuáles variables no cumple con el objetivo de "impactar" o "ser más clara" que la tabla de datos?
4. Hacer el histograma para cada una de las variable usando 5 intervalos. De nuevo, está gráfica es útil para todas las variables? justifique su respuesta.
5. Realice una gráfica que incluya el diagrama de cajas de todas las variables de tal manera de que se puedan comparar.