Practica 2

Alejandro Cáceres
UPC - Statistics 2019/2020

Estadística descriptiva

- describir, organizar, representar y resumir un conjunto de datos
- tablas de frecuencias, representarlas gráficamente, calcular algunos estadísticos importantes (media aritmética, varianza y moda)
- interpretar todos estos resultados en R

Datos

- Categóricos: NO necesitan números para expresarse, por ejemplo: sexo, color, etc
- Numéricos: SI necesitan números para expresarse, por ejemplo: edad, longitud, etc

Por cada variable tenemos una serie de observaciones que se pueden resumir en: Tablas, gráficos, medidas de tendencia central y variabilidad.

Tablas

- Matrices donde se guardan los datos que toma una determinada variable para cada objeto.
- Tablas de frecuencia

Tablas

- 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.

Tablas

- 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.

Tablas

Ejemplo1

- 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

Tablas

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

Tablas

```
# Se visualiza la tabla
> Tabla_Frec
 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
```

Tablas

Reactor	**Frecuencia absoluta** (n_i)	**Frecuencia relativa** (f_i)	**Frecuencia absoluta acumulada** (N_i)	**Frecuencia relativa acumulada** (F_i)
1	4	$4/20 = 0{,}2$	4	0.2
2	5	$5/20 = 0{,}25$	9	0.45
3	2	$2/20 = 0{,}1$	11	0.55
4	6	$6/20 = 0{,}3$	17	0.85
5	3	$3/20 = 0{,}15$	20	1

Tablas

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

Tablas

- En este caso conviene agrupar los datos en intervalos.
- Construimos las frecuencias correspondientes a cada intervalo.

Tablas

```
> 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)
```

Tablas

Se crea el vector que contiene los intervalos

```
> breaks = seq(70,250,by=20)
> breaks
[1] 70 90 110 130 150 170 190 210 230 250
```

Relaciona c/valor con su intervalo

```
> datos_2a = cut(datos_2, breaks, right=FALSE)
```

Tablas

```
> datos_2a
 [1] [90,110)  [210,230) [170,190) [170,190) [110,130)
 [8] [130,150) [150,170) [130,150) [90,110)  [150,170)
[15] [110,130) [150,170) [170,190) [110,130) [150,170)
[22] [210,230) [170,190) [190,210) [170,190) [150,170)
[29] [210,230) [130,150) [150,170) [130,150) [150,170)
[36] [190,210) [150,170) [150,170) [130,150) [150,170)
[43] [190,210) [190,210) [190,210) [130,150) [150,170)
[50] [130,150) [130,150) [170,190) [70,90)   [150,170)
[57] [210,230) [150,170) [90,110)  [170,190) [150,170)
[64] [150,170) [130,150) [150,170) [130,150) [170,190)
[71] [130,150) [70,90)   [150,170) [230,250) [190,210)
[78] [170,190) [150,170) [170,190)
9 Levels: [70,90) [90,110) [110,130) [130,150) [150,170
```

Por ejemplo el primer dato 105 pertenece al intervalo [90,110).

Tablas

Calculamos otra vez las frequencias para estos datos

```
# Frecuencia absoluta
> ni = table(datos_2a)
# Frecuencia relativa
> fi = table(datos_2a)/length(datos_2a)
# Frecuencia absoluta acumulada
> Ni = cumsum(ni)
# Frecuencia relativa acumulada
> Fi = cumsum(fi)
# Se crea una tabla con todas las frecuencias
> Tabla_Frec = cbind(ni,fi,Ni,Fi)
```

Tablas

```
# Se visualiza la tabla
> Tabla_Frec
          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
```

Tablas

Clase	Frecuencia absoluta (n_i)	Frecuencia relativa (f_i)	Frecuencia absoluta acumulada (N_i)	Frecuencia relativa acumulada (F_i)
$70 \leq x < 90$	2	0.0250	2	0.0250
$90 \leq x < 110$	3	0.0375	5	0.0625
$110 \leq x < 130$	6	0.0750	11	0.1375
$130 \leq x < 150$	14	0.1750	25	0.3125
$150 \leq x < 170$	22	0.2750	47	0.5875
$170 \leq x < 190$	17	0.2125	64	0.8000
$190 \leq x < 210$	10	0.1250	74	0.9250
$210 \leq x < 230$	4	0.0500	78	0.9750
$230 \leq x < 250$	2	0.0250	80	1

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.

Gráficos

Gráficos de tallos y hojas: Las datos se dividen en tallos y hojas, donde el tallo son las cifras númericas hasta un orden de magnitud dado. Las hojas son las cifras númericas menores a dicho orden de manitud.

Permite tener una idea de la distribución de los datos

Gráficos

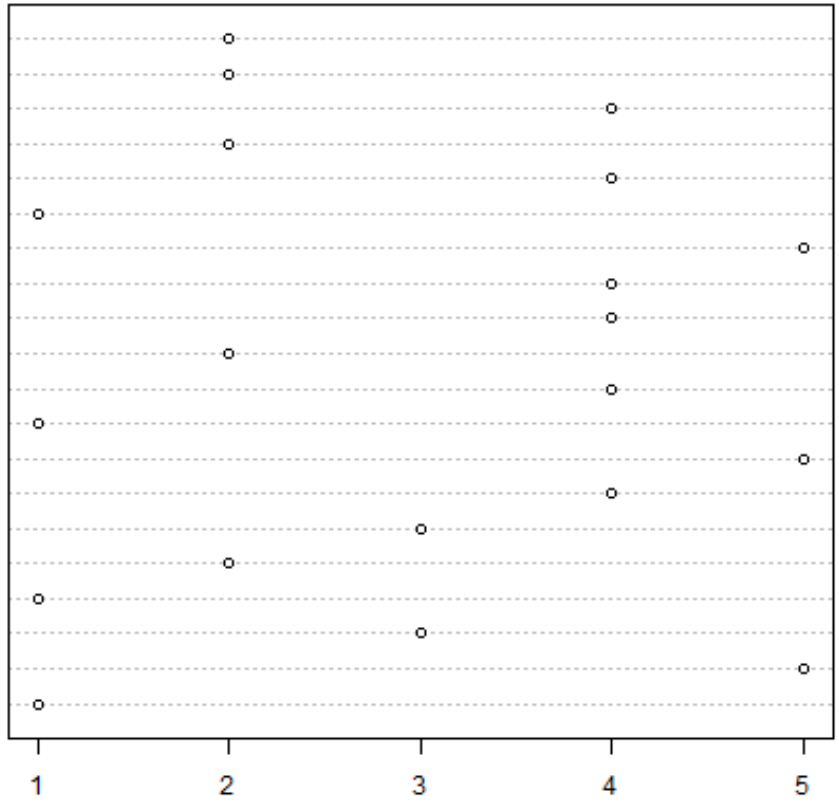
```
> stem(datos_2,scale=2)
  The decimal point is 1 digit(s) to
  the right of the |
   7 | 6
   8 | 7
   9 | 7
  10 | 15
  11 | 058
  12 | 013
  13 | 133455
  14 | 12356899
  15 | 001344678888
  16 | 0003357789
  17 | 0112445668
  18 | 0011346
  19 | 034699
  20 | 0178
```

Gráficos

Gráficos de puntos: Cuando el rango de los datos es pequeño la distribucion tambien de puede representar por medio de un gráfico donde el eje x representa el valor númerico de los datos

```
> dotchart(datos_1)
```

Gráficos



Gráficos

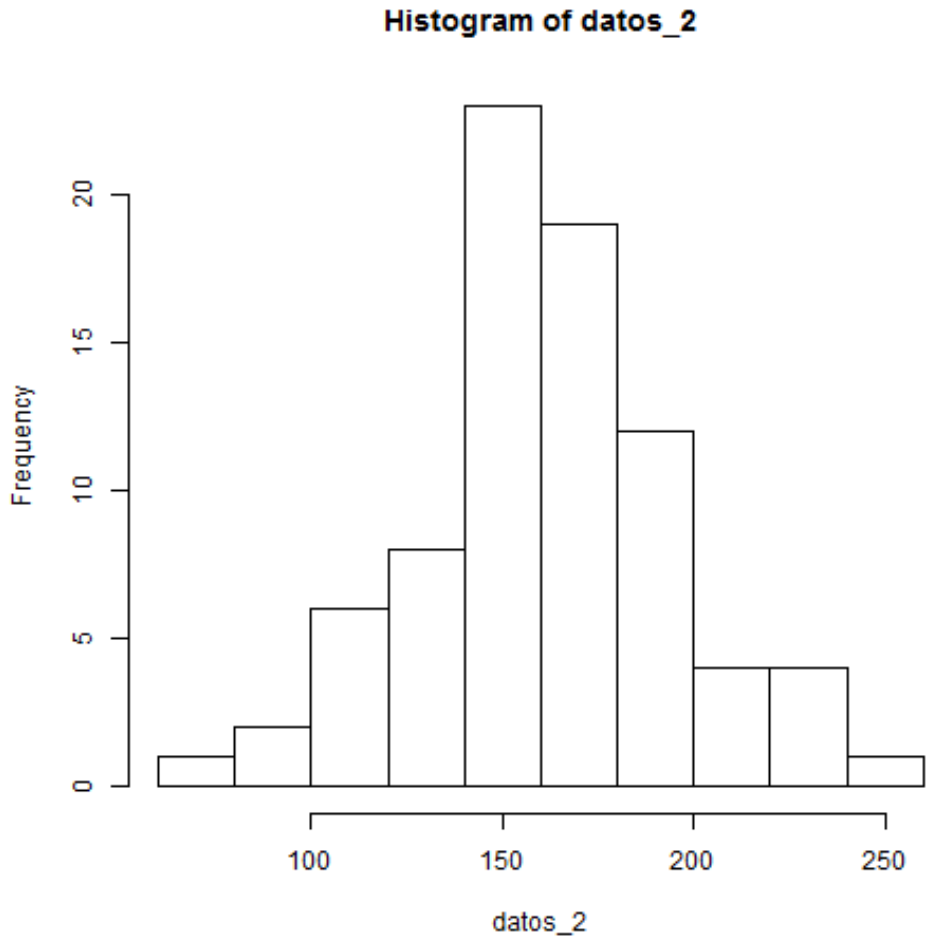
Histograma: Es el gráfico más usado para ver las frequencias de los datos en subintervalos (bins) determinados. A cada dato se le asigna un subintervalo y despué se cuentan cuantos de esos subintervalos se encuentran en la serie de datos.

```
> table(datos_2a)
datos_2a
  [70,90)  [90,110) [110,130) [130,150) [150,
        2         3         6        14
[190,210) [210,230) [230,250)
       10         4         2
```

Gráficos

R calcula automáticamente los subintervalos

```
> table(datos_2)
```



Gráficos

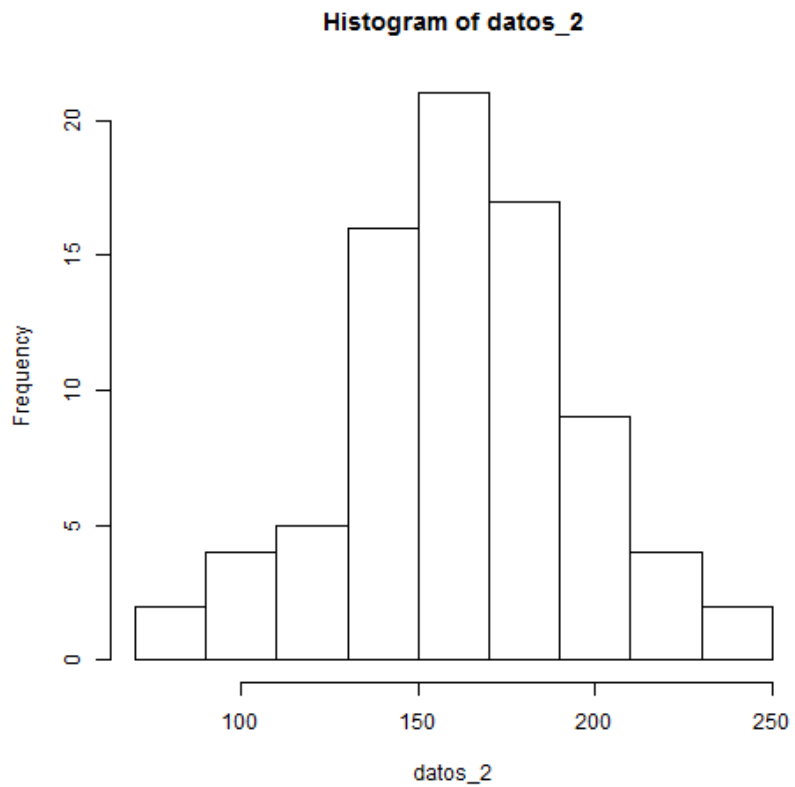
Para obtener los parámetros usados por R para calcular el histograma

```
> h$breaks # Lmites de los intervalos
[1] 60 80 100 120 140 160 180 200 220 240 260
> h$counts # Frecuencia de cada intervalo
[1] 1 2 6 8 23 19 12 4 4 1
> h$density # Densidad de cada intervalo
[1] 0.000625 0.001250 0.003750 0.005000 0.014
[10] 0.000625
> h$mids # Punto central de cada intervalo
[1] 70 90 110 130 150 170 190 210 230 250
```

Gráficos

Podemos variar el tamaño de los subintervalos

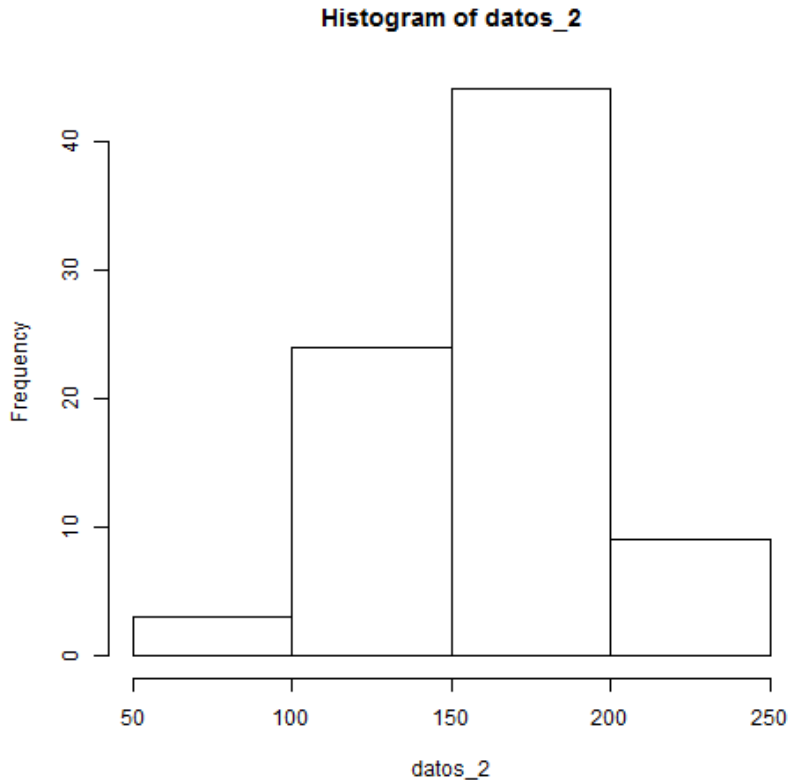
```
> new_breaks = seq(70,250,by=20)
> h1 = hist(datos_2,breaks=new_breaks)
```



Gráficos

Podemos variar el tamaño de los subintervalos

```
> h1 = hist(datos_2,breaks=3)
```



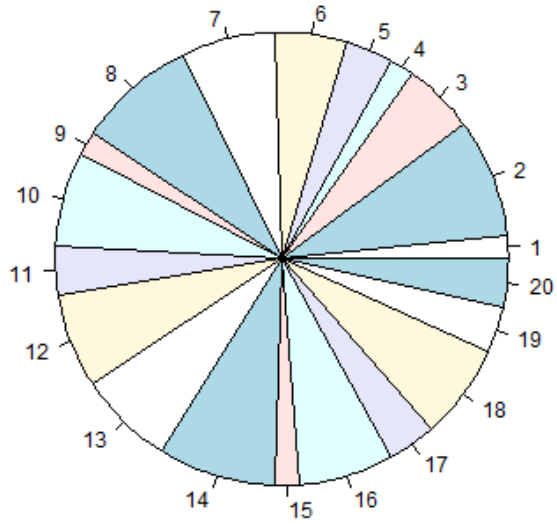
Gráficos

Gráficos de sectores: Illustra las frequencias relativas de los datos mediante las porciones de un circulo (pie/tarta)

```
> dotchart(datos_1)
```

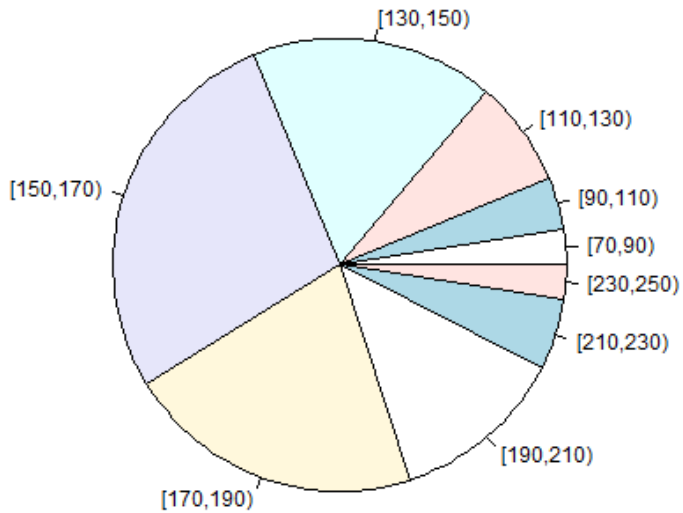
Gráficos

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



Gráficos

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



Medidas de posición y tendencia central

Las medidas de posición y tendencia central de los datos son valores numéricos importantes so solo para resumir los datos sino porque estan relacionados con los parámetros de la distribución de la población que generan los datos.

Medidas de posición y tendencia central

Media: es el promedio de los datos dado por

$$\bar{X} = \frac{1}{n}\Sigma_{i=1..n}x_i$$

es el centro de gravedad de los datos

Medidas de posición y tendencia central

Usando la definición:

```
> sum(datos_1)/length(datos_1)
[1] 2.95
> sum(datos_2)/length(datos_2)
[1] 162.6625
```

Usando la función mean

```
> mean(datos_1)
[1] 2.95
> mean(datos_2)
[1] 162.6625
```

Medidas de posición y tendencia central

Mediana: Es el valor que divide a los datos por la mitad.

La mitad de los datos esta por encima de la mediana y la otra mitad esta por debajo

Medidas de posición y tendencia central

Usando la definición:

```
> d1 = sort(datos_1); d1 # Organiza el vector
[1] 1 1 1 1 2 2 2 2 2 3 3 4 4 4 4 4 4 5 5 5
> (d1[10]+d1[11])/2
[1] 3
```

Usando la función median

```
> median(datos_1)
[1] 3
> median(datos_2)
[1] 161.5
```

Medidas de posición y tendencia central

Moda: Es el valor mas frequente en una serie de datos

Medidas de posición y tendencia central

Usando la definición:

```
> table(datos_1)
datos_1
1 2 3 4 5
4 5 2 6 3
```

Se organiza la tabla de frecuencias de mayor valor (el ms frecuente) a menor

Medidas de posición y tendencia central

```
> freq_ord=sort(table(datos_1),
              decreasing = TRUE)
> freq_ord
datos_1
4 2 1 5 3
6 5 4 3 2
```

Se toma el valor (o valores) que ms se repite (el primero de la tabla ordenada)

```
> moda = names(freq_ord[1]); moda
[1] "4"
```

Medidas de posición y tendencia central

Usando la función mode de la libreria modest

```
# carga la biblioteca modeest
> library(modeest)
> mfv(datos_1)
[1] 4
> mfv(datos_2)
[1] 158
```

Medidas de posición y tendencia central

Cuartiles: los tres valores que dividen en cuatro partes iguales una serie de datos
Deciles: los nueve valores que dividen en diez partes iguales una serie de datos
Percentil: los 99 valores que dividen en cien partes iguales una serie de datos

Medidas de posición y tendencia central

Cuantil δ: El dato que divide en una serie de datos en δ*100% y $(\delta - 1)$*100%

Por ejemplo: si $\delta = 0.95$ el cuantil 0.95 es el número que divide los datos en dos porciones: en la primera caen el 95% de los datos y en la segunda el 5%.

En las clases de teoría $\alpha = 1 - \delta$ que determina la significancia estadstica de intervalos de confianza

Medidas de posición y tendencia central

```
> quantile(datos_2,0.95) # Percentil de orden 95
95 %
221.35
> quantile(datos_2,seq(0.1,0.9,by=0.1)) # 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
> quantile(datos_2,seq(0.25,0.75,by=0.25)) # cuartiles
25 % 50 % 75 %
144.5 161.5 181.0
```

Medidas de posición y tendencia central

Rango intercuartílico: Es el rango entre el segundo y el tercer cuartil

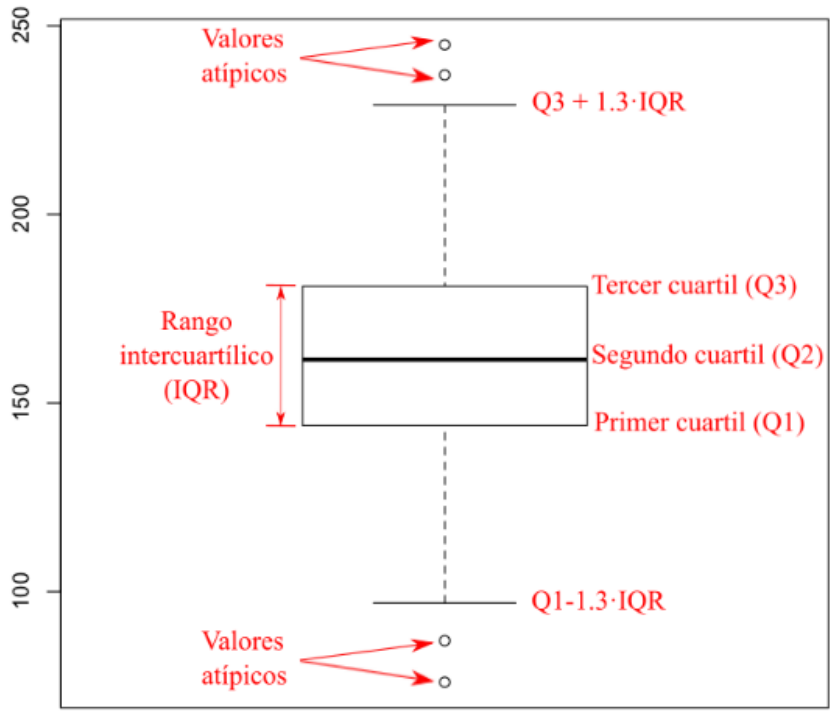
```
> IQR(datos_2)
[1] 36.5
```

La función summary da diferentes estadisticos de los datos

```
> summary(datos_2)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
76.0   144.5   161.5   162.7   181.0   245.0
```

Gráficos

Box plot: ahora que ya sabemos que es un cuartil entonces podemos hacer un boxplot que ilustra los cuartiles y los valores atípicos



Gráficos

Los box plots se consiguen con

```
> boxplot(datos_2)
```

Y para obetner los datos atípicos

```
> bp=boxplot(datos_2)
> bp$out
[1] 245 76 87 237
```