Capitulo 2: Estadistica Descriptiva Aplicaciones

Dereck Amesquita

6/4/2021

Aplicaciones

La funcion summary

Summary nos da los principales estadisticos descriptivos. Pero estos deben aplicarse a un vector numerico, es decir si lo aplicamos a un dataframe podremos estos todos los estadisticos.

```
cars=mtcars
cars=cars[,1:5] #Cortamos el dataframe
str(cars)
```

```
## 'data.frame':    32 obs. of  5 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
```

```
summary(cars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat      
##  Min.   :2.760  
##  1st Qu.:3.080  
##  Median :3.695  
##  Mean   :3.597  
##  3rd Qu.:3.920  
##  Max.   :4.930
```

```
subcars=subset(cars, cyl==4, c("disp","hp")) #Asi obtenemos determinadas columnas gracias a subset
summary(subcars)
```

```
##       disp             hp       
##  Min.   : 71.10   Min.   : 52.00  
##  1st Qu.: 78.85   1st Qu.: 65.50  
##  Median :108.00   Median : 91.00  
##  Mean   :105.14   Mean   : 82.64  
##  3rd Qu.:120.65   3rd Qu.: 96.00  
##  Max.   :146.70   Max.   :113.00
```

La funcion by

Con by podemos obtener el resultado de una poblacion aplicado a un subconjunto de datos

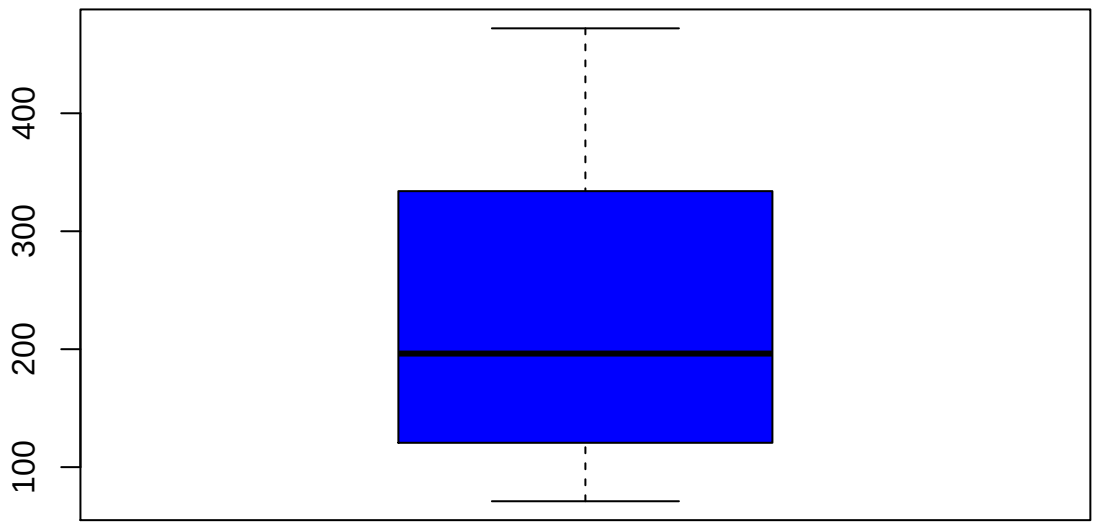
```
by(cars, cars$cyl, FUN = summary)
```

```
## cars$cyl: 4
##       mpg             cyl         disp             hp             drat      
##  Min.   :21.40   Min.   :4   Min.   : 71.10   Min.   : 52.00   Min.   :3.690  
##  1st Qu.:22.80   1st Qu.:4   1st Qu.: 78.85   1st Qu.: 65.50   1st Qu.:3.810  
##  Median :26.00   Median :4   Median :108.00   Median : 91.00   Median :4.080  
##  Mean   :26.66   Mean   :4   Mean   :105.14   Mean   : 82.64   Mean   :4.071  
##  3rd Qu.:30.40   3rd Qu.:4   3rd Qu.:120.65   3rd Qu.: 96.00   3rd Qu.:4.165  
##  Max.   :33.90   Max.   :4   Max.   :146.70   Max.   :113.00   Max.   :4.930  
## ------------------------------------------------------------ 
## cars$cyl: 6
##       mpg             cyl         disp            hp             drat      
##  Min.   :17.80   Min.   :6   Min.   :145.0   Min.   :105.0   Min.   :2.760  
##  1st Qu.:18.65   1st Qu.:6   1st Qu.:160.0   1st Qu.:110.0   1st Qu.:3.350  
##  Median :19.70   Median :6   Median :167.6   Median :110.0   Median :3.900  
##  Mean   :19.74   Mean   :6   Mean   :183.3   Mean   :122.3   Mean   :3.586  
##  3rd Qu.:21.00   3rd Qu.:6   3rd Qu.:196.3   3rd Qu.:123.0   3rd Qu.:3.910  
##  Max.   :21.40   Max.   :6   Max.   :258.0   Max.   :175.0   Max.   :3.920  
## ------------------------------------------------------------ 
## cars$cyl: 8
##       mpg             cyl         disp            hp             drat      
##  Min.   :10.40   Min.   :8   Min.   :275.8   Min.   :150.0   Min.   :2.760  
##  1st Qu.:14.40   1st Qu.:8   1st Qu.:301.8   1st Qu.:176.2   1st Qu.:3.070  
##  Median :15.20   Median :8   Median :350.5   Median :192.5   Median :3.115  
##  Mean   :15.10   Mean   :8   Mean   :353.1   Mean   :209.2   Mean   :3.229  
##  3rd Qu.:16.25   3rd Qu.:8   3rd Qu.:390.0   3rd Qu.:241.2   3rd Qu.:3.225  
##  Max.   :19.20   Max.   :8   Max.   :472.0   Max.   :335.0   Max.   :4.220
```

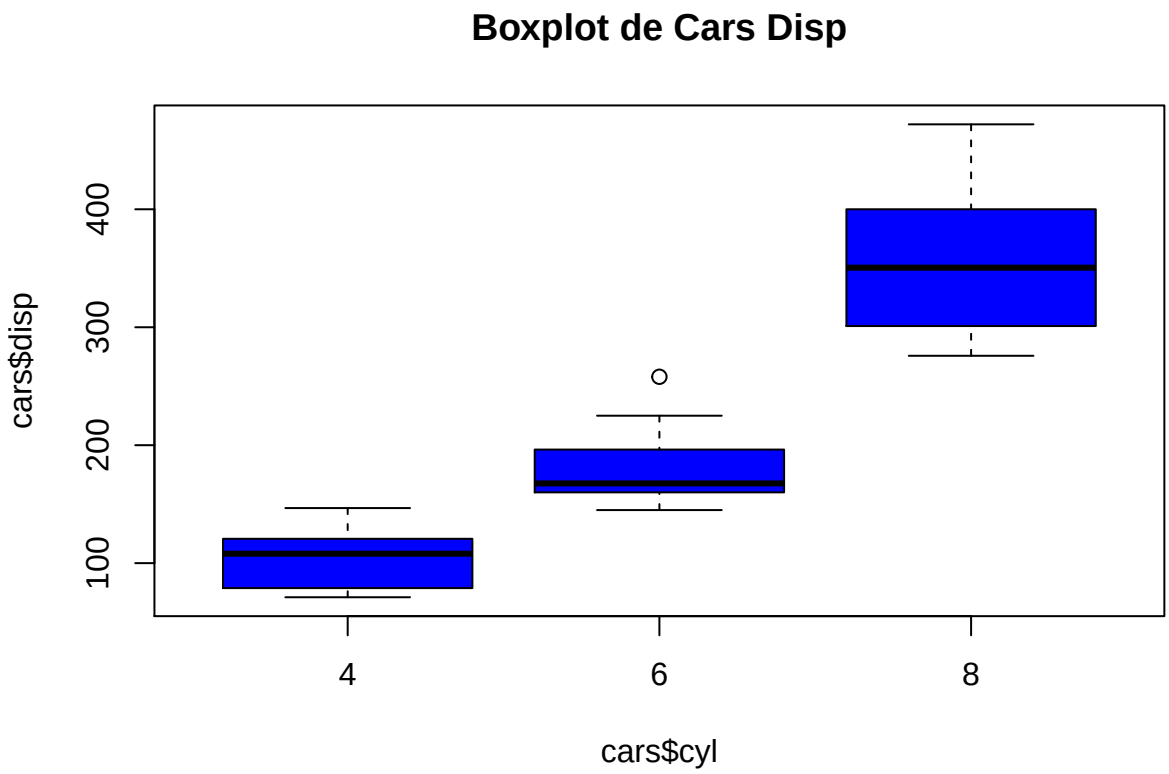
Diagramas de caja

El diagrama de caja divide los datos en cuartiles, la base de la caja nos da el primer cuartil el techo de la cada nos da el tercer cuartil, los bigotes son los maximos y los minimos.

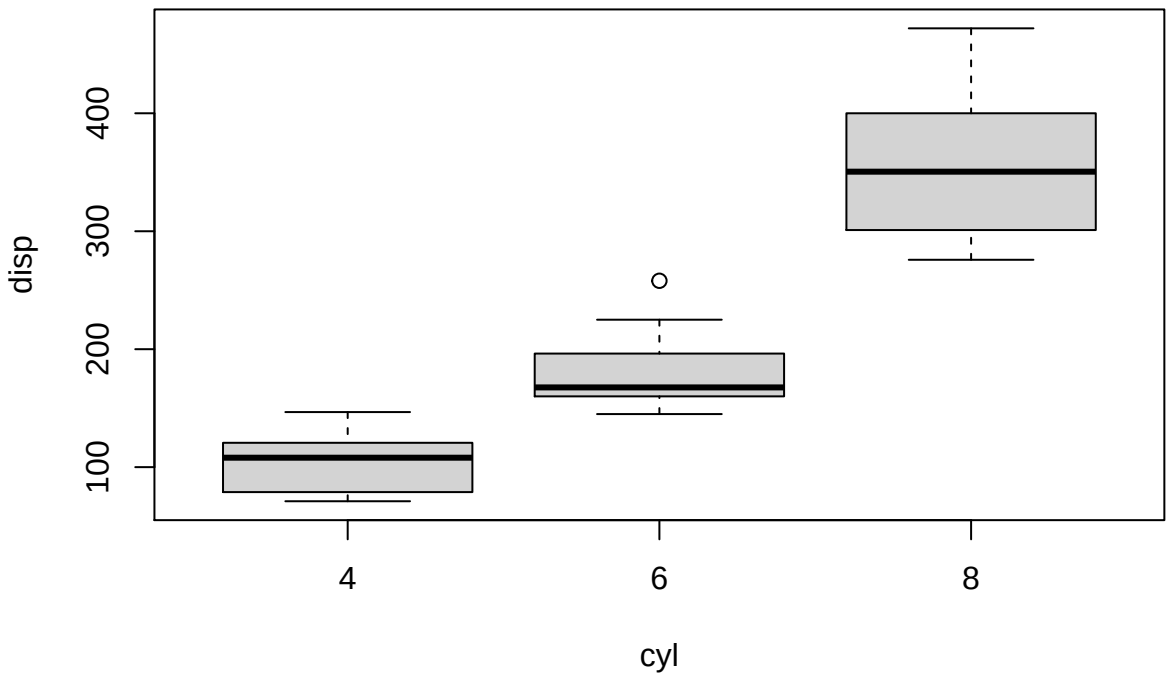
```
boxplot(cars$disp, main="Boxplot de Cars Disp",
        col = "blue")
```




```
boxplot(cars$disp~cars$cyl, main="Boxplot de Cars Disp",
        col = "blue")
```




```
boxplot(disp~cyl, data=cars) #En este caso no usamos $ ya que la data la ponemos aparte.
```



Podemos que ver que los de 6 cilindros tienen mayor mediana de disp en comparacion con la de 4 cilindros.

La funcion stats nos devuelve los valores con los que se forman el boxplot

Ejericio de eficacia de insectisida

Debemos analizar individualmente la eficacia de los insecticidas.

```
insect=InsectSprays
str(insect)
```

```
## 'data.frame':    72 obs. of  2 variables:
##  $ count: num  10 7 20 14 14 12 10 23 17 20 ...
##  $ spray: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
by(insect$count, insect$spray, FUN=summary )
```

```
## insect$spray: A
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   11.50   14.00   14.50   17.75   23.00
## ------------------------------------------------------------
## insect$spray: B
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   12.50   16.50   15.33   17.50   21.00
## ------------------------------------------------------------
```

```
## insect$spray: C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.500   2.083   3.000   7.000
## ------------------------------------------------------------
## insect$spray: D
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.750   5.000   4.917   5.000  12.000
## ------------------------------------------------------------
## insect$spray: E
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.75    3.00    3.50    5.00    6.00
## ------------------------------------------------------------
## insect$spray: F
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   12.50   15.00   16.67   22.50   26.00
```

```
#obtenemos los estadisticos divididos por spry.
```

Tenemos la sospecha que existe un rango alto. Por lo cual procedemos a calcular la desviacion estandar.

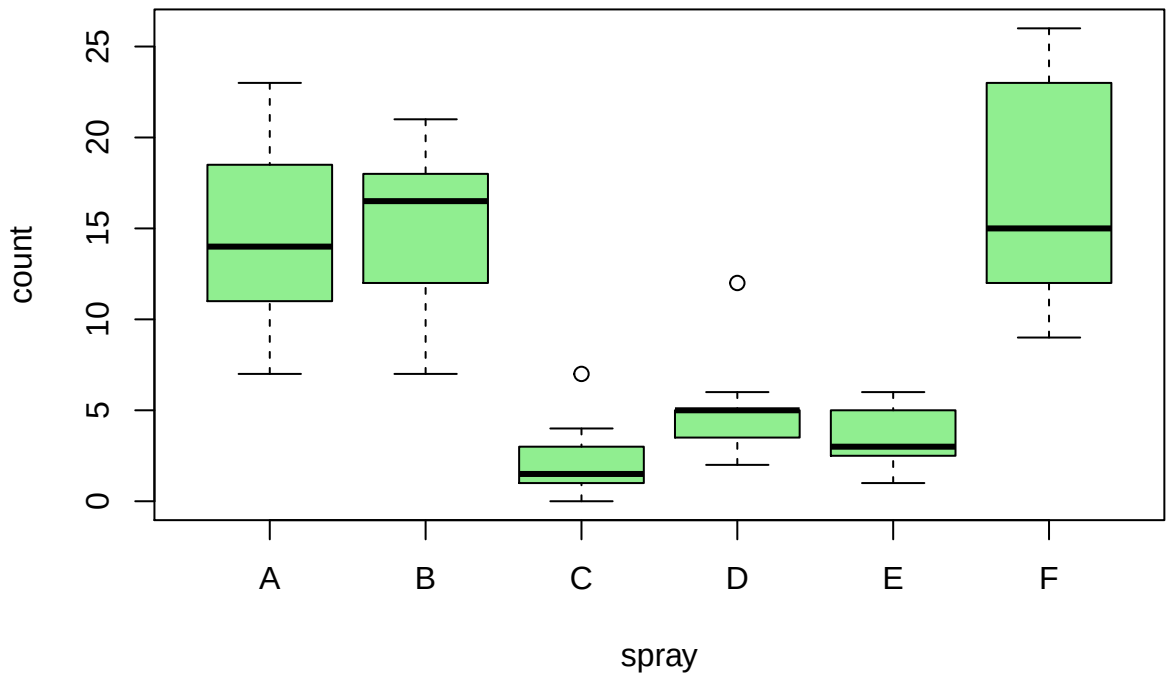
```
aggregate(count~spray, data=insect, FUN=sd)
```

```
##   spray    count
## 1     A 4.719399
## 2     B 4.271115
## 3     C 1.975225
## 4     D 2.503028
## 5     E 1.732051
## 6     F 6.213378
```

```
#Agregamos count por tipo de spray < Esta es la forma de leer el codigo
```

Ahora podemos ver que los botes que tenian un maximo mas grande, tienen mayor desviacion. En el caso de A esta disperso en 4.71 en promedio de su media.

```
boxplot(count~spray, data=insect, col="lightgreen")
```

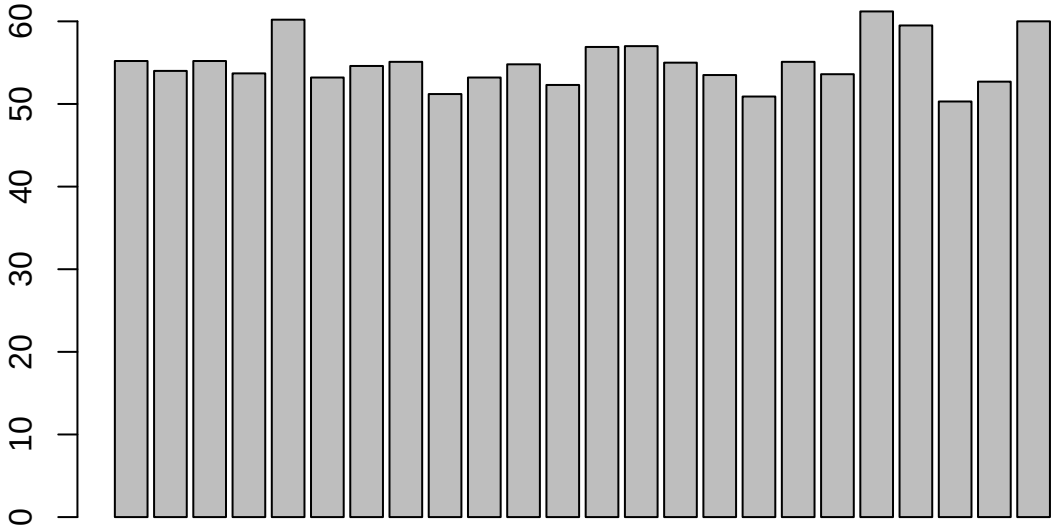


Ahora podemos ver que A B F matan a mas insector pero tienen mayor diferencia en su cuartil 1 y 3. En diferencia C D E tienen menor diferencia en sus cuartil, tecnicamente son mas faciles de predecir.

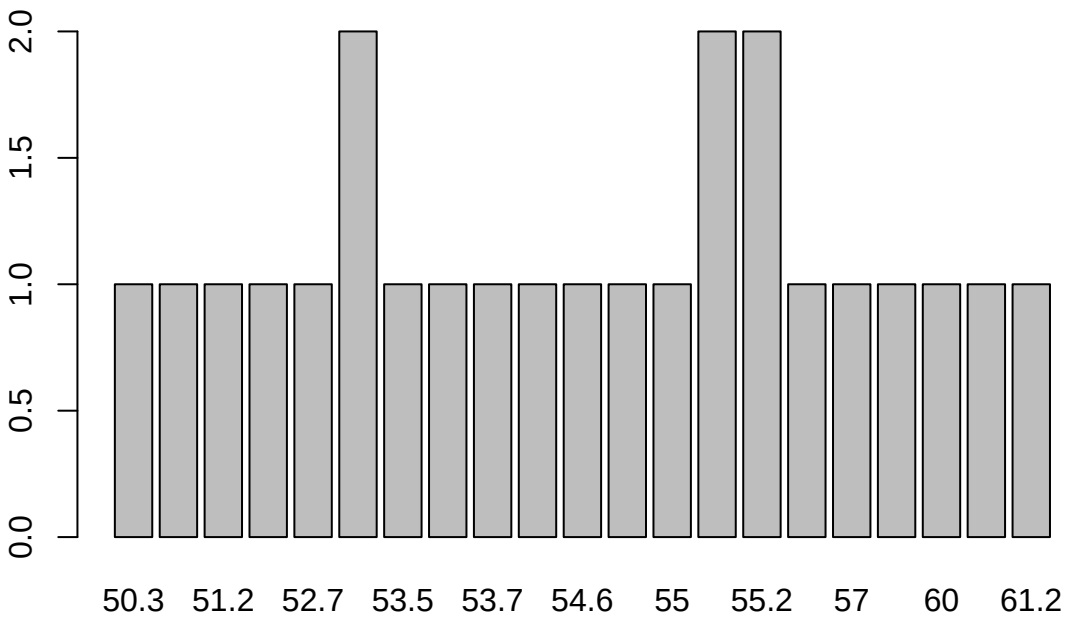
Agrupamiento de datos

No agrupar los datos puede limitarnos al momento de querer graficarlos. Por ejemplo cuando hablamos de la edad de una persona, nos referimos a una edad entera, decimos 40 años, desde que cumple 40 hasta que cumple 41, no decirmos 40.1 o 40.5.

```
pesos = c(55.2,54.0,55.2,53.7,60.2,53.2,54.6,55.1,51.2,53.2,54.8,52.3,56.9,57.0,55.0,
          53.5,50.9,55.1,53.6,61.2,59.5,50.3,52.7,60.0)
barplot(pesos) #Divide demasiado los datos
```




```
barplot(table(pesos))#Su frecuencia absoluta incluso es dificil de visualizar
```



Para dar solución a este problema es que podemos hacer una división por intervalos.