Programación Estadística: Gráficas

Jocelyn Simmonds (jsimmond@dcc.uchile.cl)

Departamento de Ciencias de la Computación

Cambiar el directorio de trabajo

Pueden usar `getwd()` y `setwd()` para cambiar el directorio en el que están trabajando. En este ejemplo, asuman esta estructura de directorios:

```
R-workspace
 |--- clase01
 |--- clase02
 |--- clase03
 |--- clase04
 |--- ....
```

```
1  > getwd()
2  [1] "/home/jsimmond/R-workspace/clase04"
3  > setwd("../clase03")
4  > getwd()
5  [1] "/home/jsimmond/R-workspace/clase03"
6  > setwd("..")
7  > getwd()
8  [1] "/home/jsimmond/R-workspace"
9  > setwd("./clase04")
10 > getwd()
11 [1] "/home/jsimmond/R-workspace/clase04"
```

Dos nombres de directorios especiales:

- `.` corresponde al directorio actual
- `..` corresponde al directorio "padre" del directorio actual

Graficando datos cualitativos

Datos cualitativos

Hemos visto que podemos usar factores para categorizar datos:

```
> head(mtcars, 5)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
> mtcars$cyl
 [1] 6 6 4 6 8 6 8 4 4 6 6 8 8 8 8 8 8 4 4 4 4 8 8 8 8 4 4 4 8 6 8 4
>
> fact_cyl <- factor(mtcars$cyl)
> fact_cyl
 [1] 6 6 4 6 8 6 8 4 4 6 6 8 8 8 8 8 8 4 4 4 4 8 8 8 8 4 4 4 8 6 8 4
Levels: 4 6 8
```

Usen `help(nombre_data_set)` para ver la descripción de un dataset disponible en R.

Pruébenlo

El paquete MASS contiene varios datasets, como `painters` y `crabs`. Para usar estos datasets, primero deben cargar el paquete MASS:

```
1  > library(MASS)
```

Exploren los datasets `painters` y `crabs` del paquete MASS. ¿Cuáles columnas son factores, o candidatos a factores?

Distribución de frecuencia

Pueden usar la función `table()` para calcular la frecuencia de ocurrencia de cada categoría de un factor:

```
> table(painters$School)
 A  B  C  D  E  F  G  H
10  6  6 10  7  4  7  4
>
> table(fact_cyl)
fact_cyl
 4  6  8
11  7 14
> table(crabs$sex, crabs$sp)

     B  O
  F 50 50
  M 50 50
```

Estadisticas sobre categorias

`table()` solo cuenta ocurrencias, deben usar `tapply()` para calcular estadísticas sobre cruces de categorías:

```
> tapply(crabs$FL, list(crabs$sp, crabs$sex), mean)
      F      M
B 13.270 14.842
O 17.594 16.626
>
> tapply(painters$Composition, painters$School, mean)
        A         B         C         D         E         F         G         H
10.400000 12.166667 13.166667  9.100000 13.571429  7.250000 13.857143 14.000000
```

Distribución de frecuencia relativa

$$\text{Frecuencia relativa} = \frac{\text{Frecuencia}}{\text{Tamaño muestra}}$$

Calculamos la distribución de frecuencia relativa usando `table()` y `nrow()`:

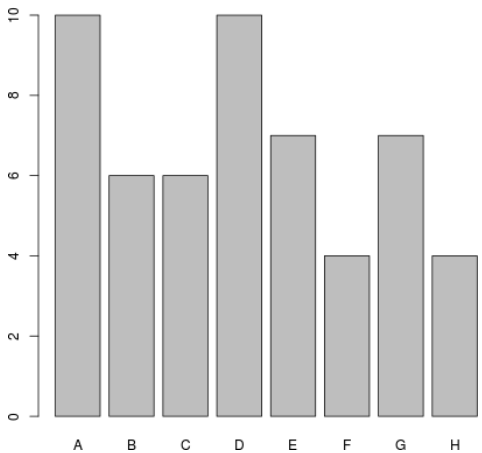
```
> freq_painters <- table(painters$School)
> freq_painters

 A  B  C  D  E  F  G  H
10  6  6 10  7  4  7  4
> rel_freq <- freq_painters/nrow(painters)
> rel_freq

          A           B           C           D           E           F           G
0.18518519 0.11111111 0.11111111 0.18518519 0.12962963 0.07407407 0.12962963
          H
0.07407407
```

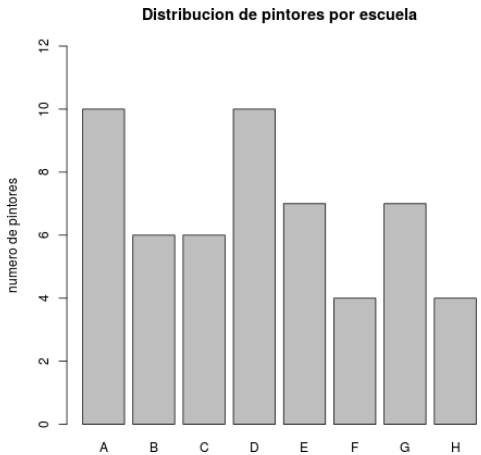
Cambiando las opciones de R

```
> rel_freq

          A           B           C           D           E           F           G
0.18518519 0.11111111 0.11111111 0.18518519 0.12962963 0.07407407 0.12962963
          H
0.07407407
> options(digits=1)
> rel_freq

   A    B    C    D    E    F    G    H
0.19 0.11 0.11 0.19 0.13 0.07 0.13 0.07
> options(digits=3)
> rel_freq

     A      B      C      D      E      F      G      H
0.1852 0.1111 0.1111 0.1852 0.1296 0.0741 0.1296 0.0741
```

Graficando la distribución de frecuencia

```
> freq_painters <- table(painters$School)
> barplot(freq_painters)          # grafica en pantalla
>
> png("mi_archivo.png")           # dejara el grafico en mi_archivo.png
> barplot(freq_painters)
> dev.off()                       # cierra el "device", copiando los
>                                 # contenidos al archivo
>
> x11()                           # o windows() si estan en Windows
> barplot(freq_painters, ylab = "numero de pintores", ylim = c(0,12),
+ main="Distribucion de pintores por escuela")
```

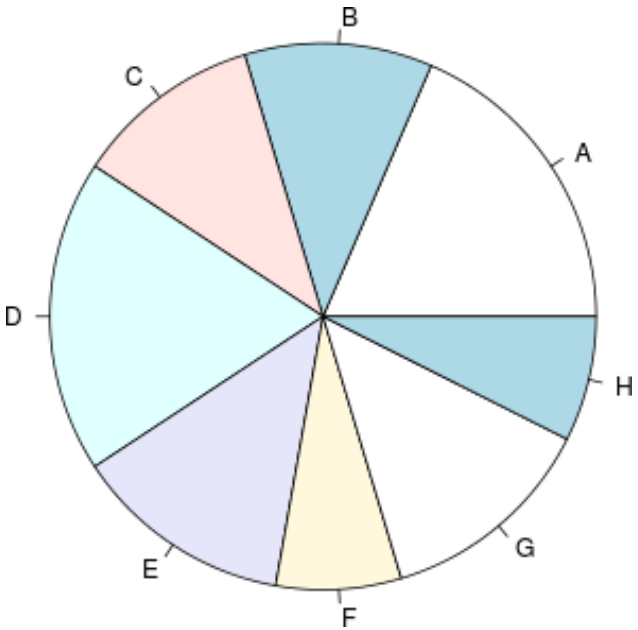


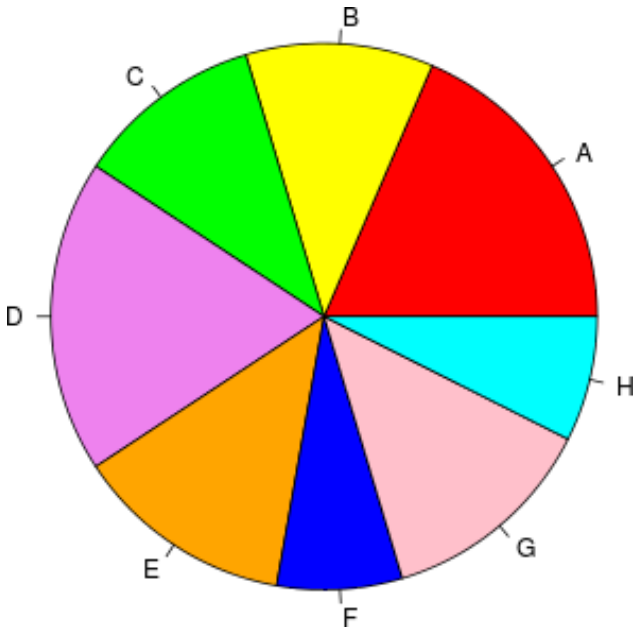


Graficando la distribución de frecuencia

También podemos generar gráficos de torta:

```
> pie(freq_painters)
>
> # con otros colores
> colores = c("red", "yellow", "green", "violet", "orange", "blue",
+ "pink", "cyan")
> pie(freq_painters, col = colores)
```





Otra forma de guardar gráficos

Pueden guardar un grafico generado en pantalla en forma directa, usando `dev.copy()`:

```
> pie(freq_painters, col = colores)
> dev.copy(png, "piechart.png")
png
  4
> dev.off()
png
  2
```

Al hacer `dev.copy()`, cambian el valor del `device` actual al indicado.

Graficando datos cuantitativos

Distribución de frecuencia

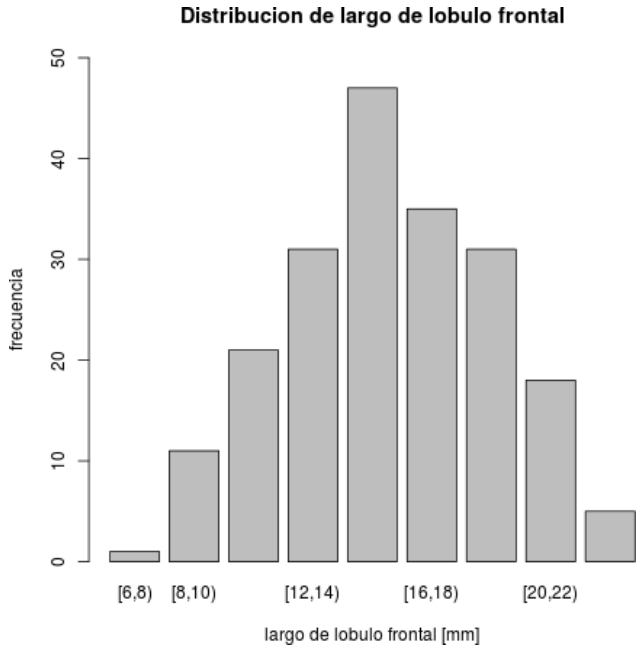
Primero debemos generar una partición del rango de los datos:

```
> head(crabs$FL, 20)    # tamanno del lobulo frontal [mm]
 [1]  8.1  8.8  9.2  9.6  9.8 10.8 11.1 11.6 11.8 11.8 12.2 12.3 12.6 12.8 12.8
[16] 12.9 13.1 13.1 13.3 13.9
>
> range(crabs$FL)
[1]  7.2 23.1
>
> cortes <- seq(6.0, 24.0, by=2)
> length(cortes)
[1] 10
> crabs.cut.FL <- cut(crabs$FL, cortes, right=FALSE)
> crabs.cut.FL
  [1] [8,10)  [8,10)  [8,10)  [8,10)  [8,10)  [10,12) [10,12) [10,12) [10,12)
 [10] [10,12) [12,14) [12,14) [12,14) [12,14) [12,14) [12,14) [12,14) [12,14)
 [19] [12,14) [12,14) [14,16) [14,16) [14,16) [14,16) [14,16) [14,16) [14,16)
...
[190] [20,22) [20,22) [20,22) [20,22) [20,22) [20,22) [20,22) [20,22) [20,22)
[199] [22,24) [22,24)
9 Levels: [6,8) [8,10) [10,12) [12,14) [14,16) [16,18) [18,20) ... [22,24)
```

Distribución de frecuencia

Ahora podemos usar `table()` y `barplot()`:

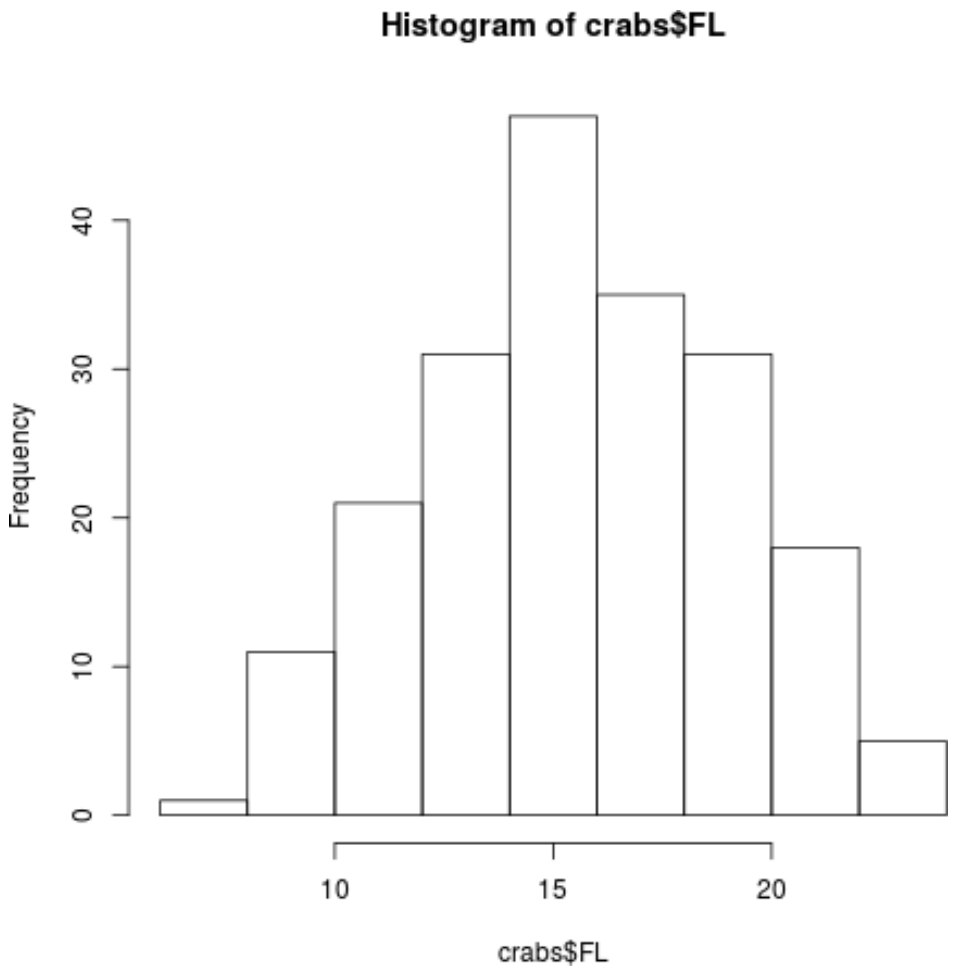
```
> freq.FL <- table(crabs.cut.FL)
> freq.FL
crabs.cut.FL
  [6,8)  [8,10) [10,12) [12,14) [14,16) [16,18) [18,20) [20,22) [22,24)
      1      11      21      31      47      35      31      18       5
> barplot(freq.FL, ylim=c(0, 50), main="Distribucion de largo de lobulo
+ frontal", ylab="frecuencia", xlab="largo de lobulo frontal [mm]")
```



Distribución de frecuencia

Tambien podemos usar la funcion `hist()` para generar histogramas:

```
1  > hist(crabs$FL, right=FALSE)
```



Distribución de frecuencia relativa y acumulada

Podemos reutilizar los intervalos que calculamos antes usando `cut()` para ahora calcular las distribuciones de frecuencia relativa y acumulada:

```
> freq.FL <- table(crabs.cut.FL)
>
> relfreq.FL <- freq.FL / nrow(crabs)
> relfreq.FL
crabs.cut.FL
  [6,8)  [8,10) [10,12) [12,14) [14,16) [16,18) [18,20) [20,22) [22,24)
  0.005   0.055   0.105   0.155   0.235   0.175   0.155   0.090   0.025
>
> acumfreq.FL <- cumsum(freq.FL)
> acumfreq.FL
  [6,8)  [8,10) [10,12) [12,14) [14,16) [16,18) [18,20) [20,22) [22,24)
      1      12      33      64     111     146     177     195     200
```

Generando un gráfico de la frecuencia acumulada

```
> acumfreq0.FL <- c(0, cumsum(freq.FL))
> plot(cortes, acumfreq0.FL,
+   main="Lobulo frontal del cangrejo Leptograpsus variegatus",
+   xlab="tamano [mm]", ylab="frecuencia acumulada")
> lines(cortes, acumfreq0.FL)
```

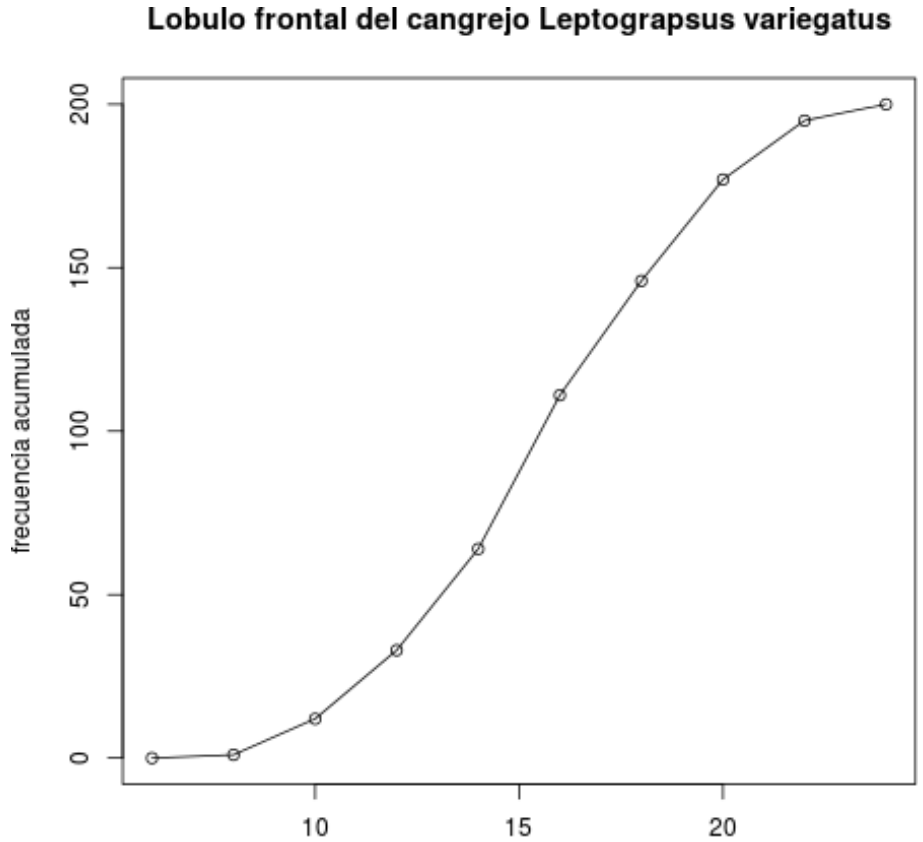


Gráfico de dispersión (scatter plot)

Para ver la relación entre dos variables numéricas, podemos usar un gráfico de dispersión:

```
1  > plot(crabs$FL, crabs$CL, xlab="tamano del lobulo frontal [mm]",
2  + ylab="tamano del caparazon [mm]", main="cangrejo Leptograpsus variegatus")
```

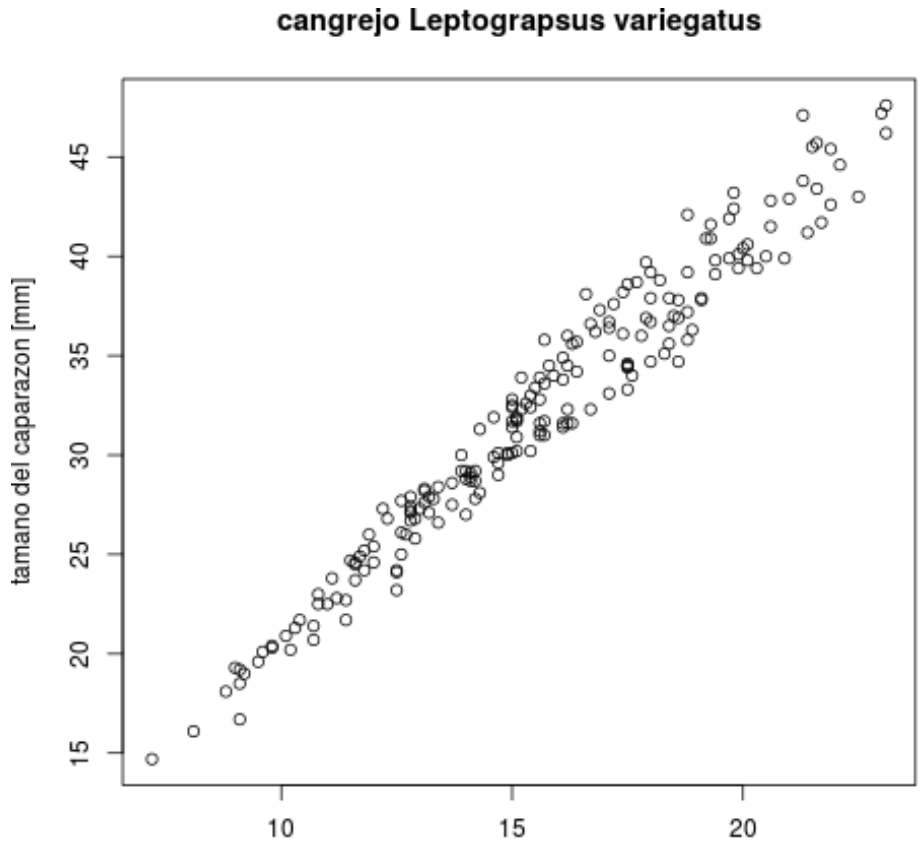
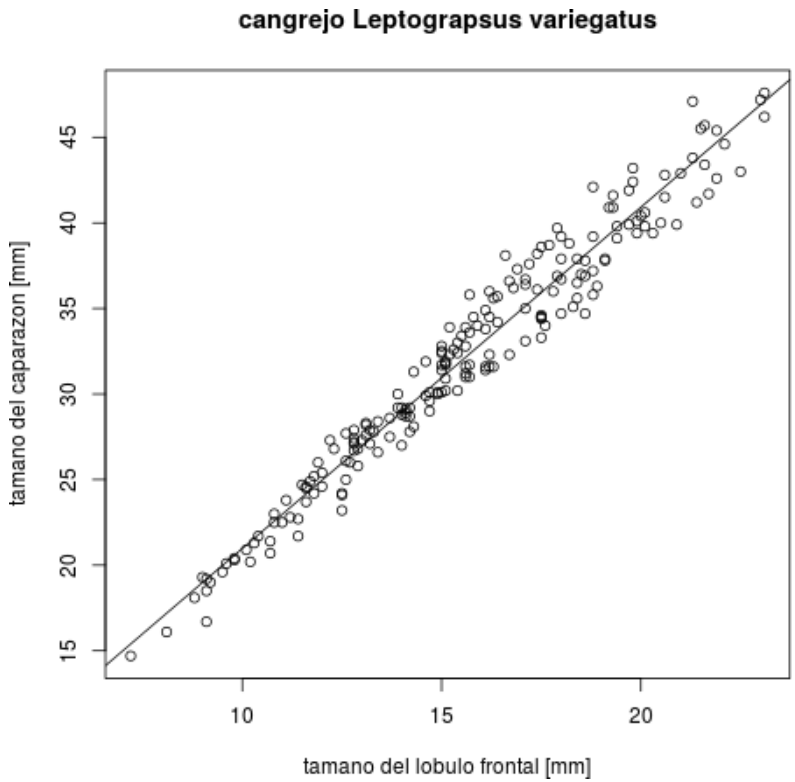


Gráfico de dispersión (scatter plot)

En este caso, hay una clara relación linear positiva entre las dos variables. Podemos usar `abline()` y `lm()` para agregar una linea de tendencia:

```
1  > plot(crabs$FL, crabs$CL, xlab="tamano del lobulo frontal [mm]",
2  + ylab="tamano del caparazon [mm]", main="cangrejo Leptograpsus variegatus")
3  > abline(lm(crabs$CL ~ crabs$FL))
```



Algunos estadísticas adicionales

Cuartiles

Podemos dividir los datos en cuartiles:

- primer cuartil (cuartil inferior o Q1): la mediana de la primera mitad de los datos
- segundo cuartil (el cuartil medio o Q2): es la mediana
- tercer cuartil (cuartil superior o Q3): la mediana de la segunda mitad de los datos

```
> quantile(crabs$FL)
    0%   25%   50%   75%  100%
 7.20 12.90 15.55 18.05 23.10
```

El rango intercuartil se define como Q3 - Q1:

```
> IQR(crabs$FL)
[1] 5.15
```

Percentiles

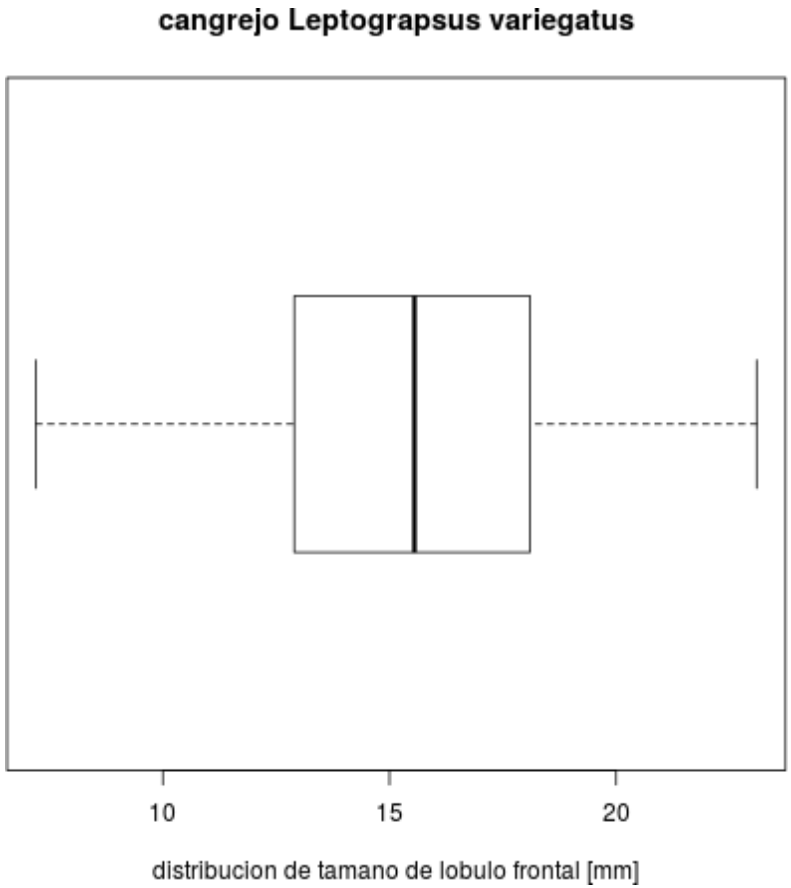
Los percentiles son los 99 valores que dividen la serie de datos en 100 partes iguales. Por ejemplo:

- P_{10} es el valor que separa el primer 10% de los valores de los datos, cuando se ordenan en forma ascendente
- P_{50} es la mediana

```
> quantile(crabs$FL, c(.10, .25, .67, .89))
   10%    25%    67%    89%
10.980 12.900 17.400 20.011
```

Gráfico de caja (box-and-whiskers)

```
1  > boxplot(crabs$FL, horizontal=TRUE, xlab="distribucion de tamano de
2  + lobulo frontal [mm]", main="cangrejo Leptograpsus variegatus")
```



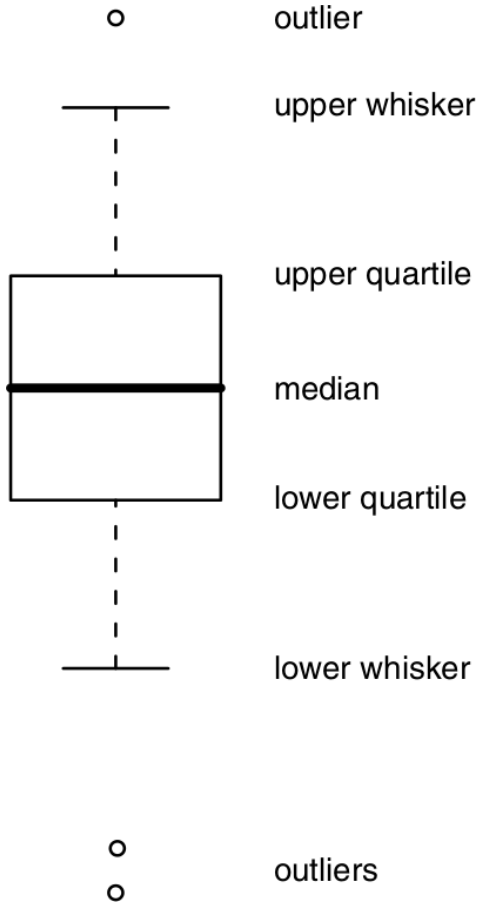
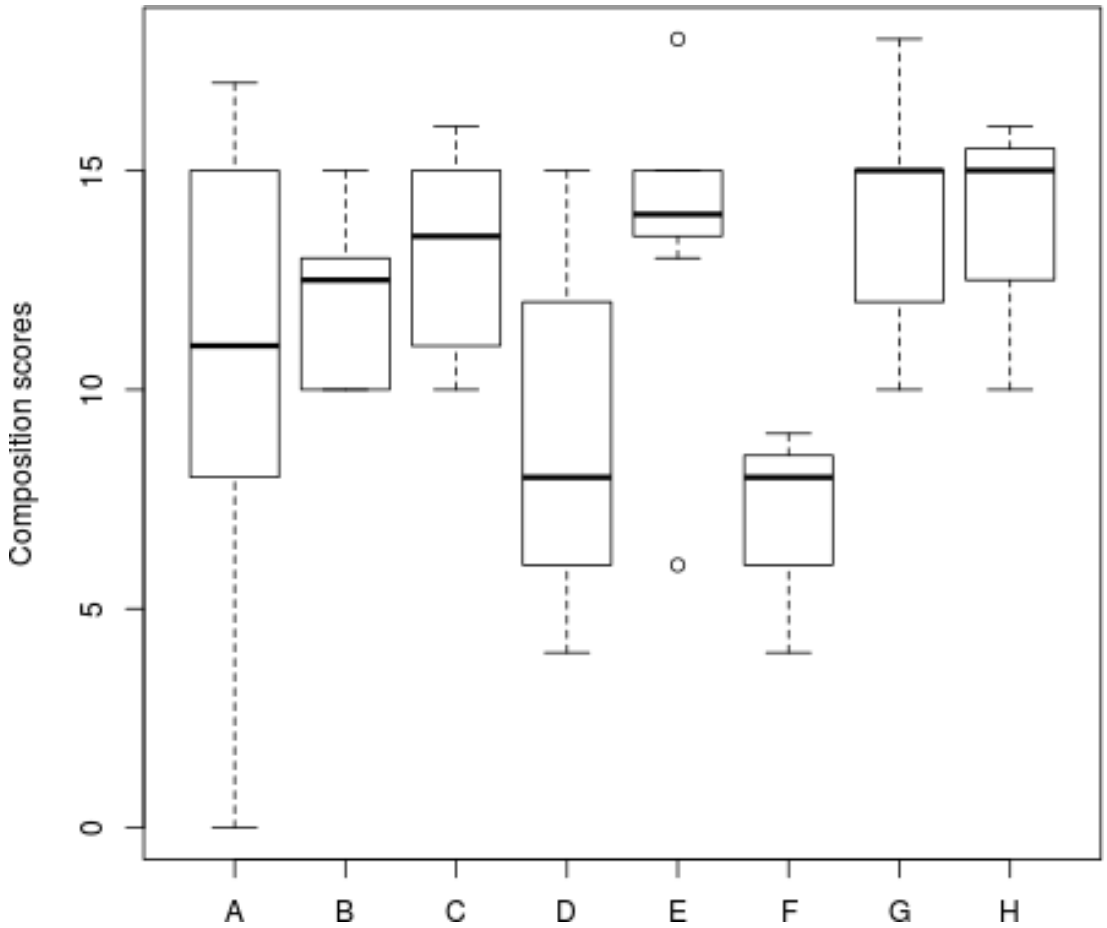


Gráfico de caja (box-and-whiskers)

También pueden mostrar distribuciones por alguna categoría:

```
1  > boxplot(Composition ~ School, data=painters, ylab="Composition scores")
```



Covarianza y el coeficiente de correlación

- La covarianza de dos variables da una idea de cuanto varían estas variables en forma conjunta.
- El coeficiente de correlación indica la magnitud de la correlación: positiva (valor cercano a 1), negativa (valor cercano a -1) o sin relación (valor cercano a 0).

```
1  > cov(crabs$FL, crabs$CL)
2  [1] 24.356677
3  >
4  > cor(crabs$FL, crabs$CL)
5  [1] 0.97884179
```

Podemos concluir que hay una correlación positiva entre el tamaño del caparazón y del lóbulo frontal en el cangrejo Leptograpsus variegatus.