Estadística Descriptiva en R: Gráficos

José Enrique Martín García
Universidad Politécnica de Gimialcón

Diagrama de cajas

Un diagrama de caja, John Tukey (1977), es un gráfico, basado en cuartiles, mediante el cual se visualiza un conjunto de datos. Está compuesto por un rectángulo (la caja) y dos brazos (los bigotes).

En dicho diagrama se representan además de los cuartiles los valores máximos y mínimos, así como los valores atípicos si los hubiera

También llamados 'diagramas de cajas y bigotes'.

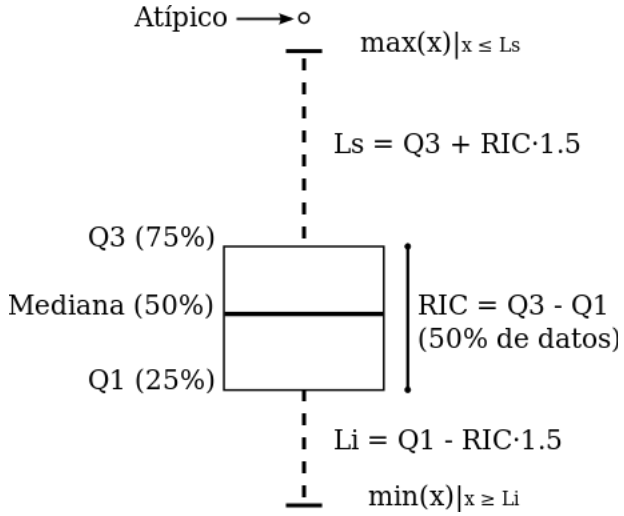


Diagrama de cajas

```
# Boxplot de MPG
>boxplot(mpg~cyl,data=mtcars, main="Datos de coches Milage",
        + xlab="Numero de cilindros", ylab="Millas por Galon")
```

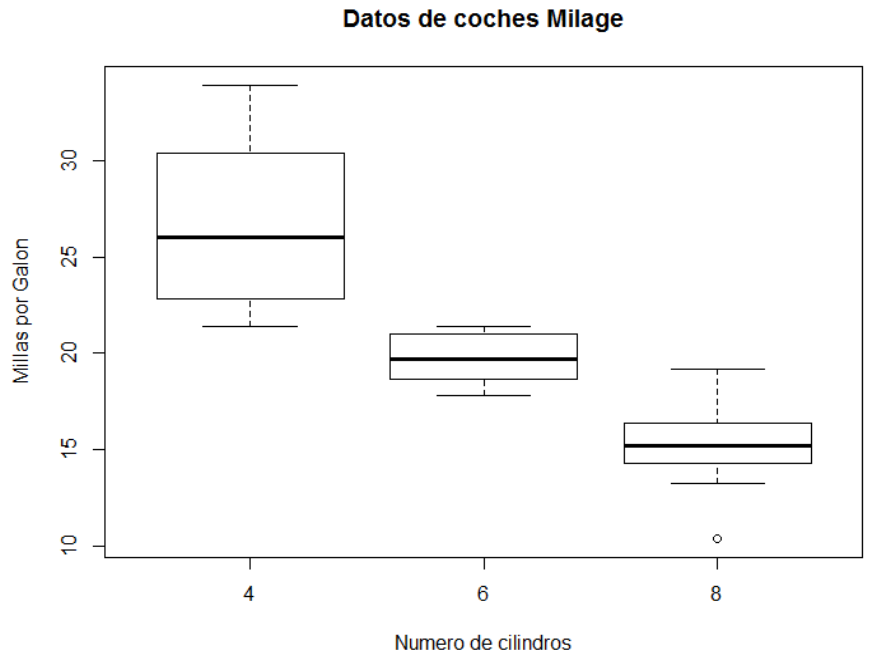


Diagrama de cajas

```
# Boxplot de Tooth Growth con dos factores
# cajas coloreadas para mejor interpretación
> boxplot(len~supp*dose, data=ToothGrowth, notch=TRUE,
+ col=(c("gold","darkgreen")),
+ main="Crecimiento de dientes", xlab="Suplemento y dosis")
```

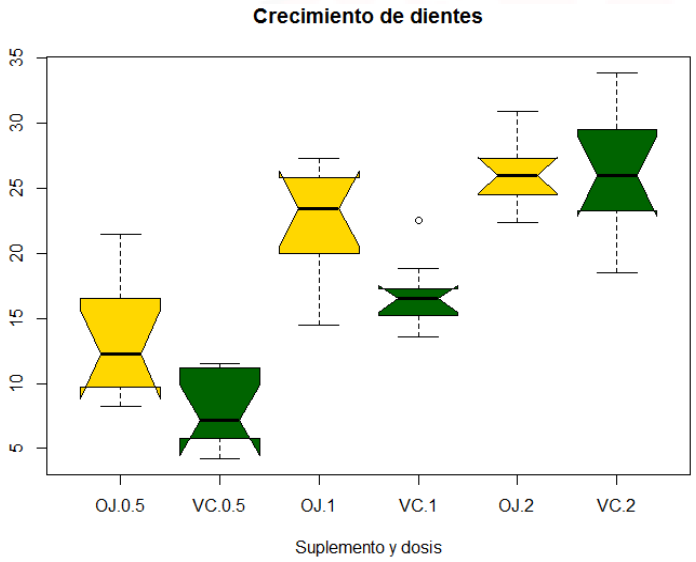


Diagrama de cajas (gráficos violín)

```
# Gráficos de tipo Violin
> library(vioplot)
> x1 <- mtcars$mpg[mtcars$cyl==4]
> x2 <- mtcars$mpg[mtcars$cyl==6]
> x3 <- mtcars$mpg[mtcars$cyl==8]
> vioplot(x1, x2, x3, names=c("4 cyl", "6 cyl", "8 cyl"),
+ col="gold")
> title("Graficos de Violin de Miles Per Gallon")
```

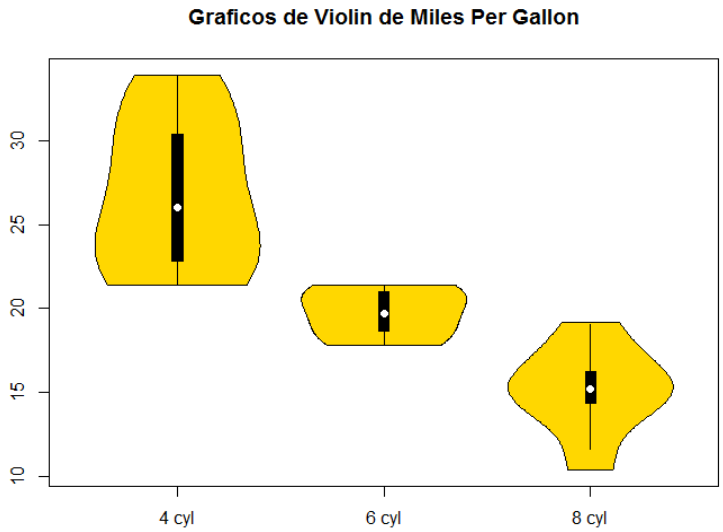
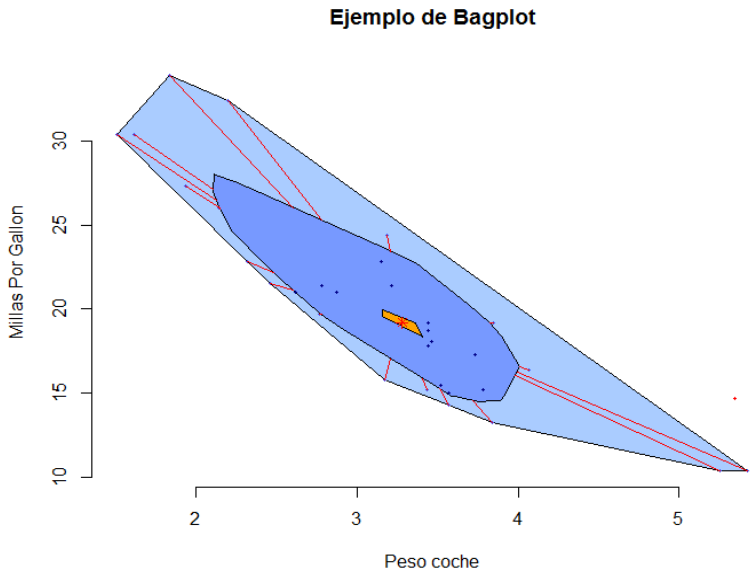


Diagrama de cajas (gráficos tipo bolsa)

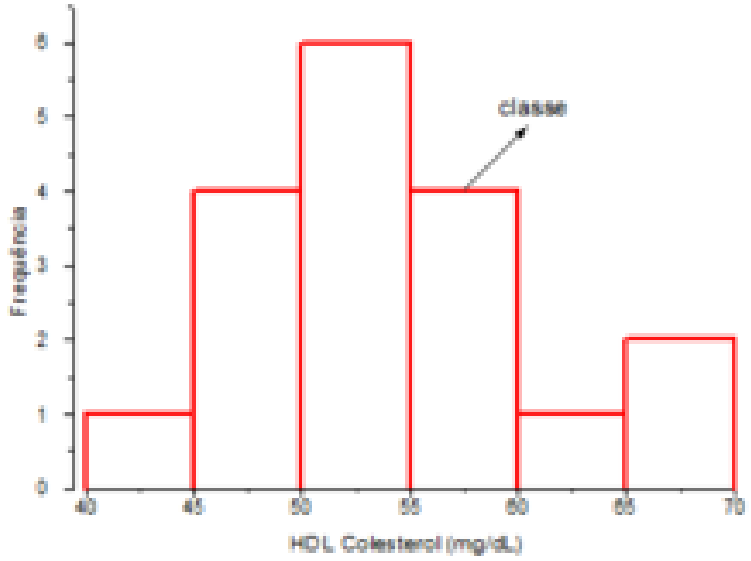
```
# Ejemplo de Bagplot
> library(aplpack)
> attach(mtcars)
> bagplot(wt,mpg, xlab="Peso coche", ylab="Millas Por Gallon",
+ main="Ejemplo de Bagplot")
```



Histograma

Un histograma es una representación gráfica de una variable en forma de barras, donde la altura de cada barra es proporcional a la frecuencia de los valores representados, ya sea en forma diferencial o acumulada. Sirven para obtener una "primera vista" general, o panorama, de la distribución de la población, o la muestra, respecto a una característica, cuantitativa y continua, de la misma y que es de interés para el observador.

En el eje vertical se representan las frecuencias, es decir, la cantidad de población o la muestra, según sea el caso, que se ubica en un determinado valor o subrango de valores de la característica que toma la característica de interés

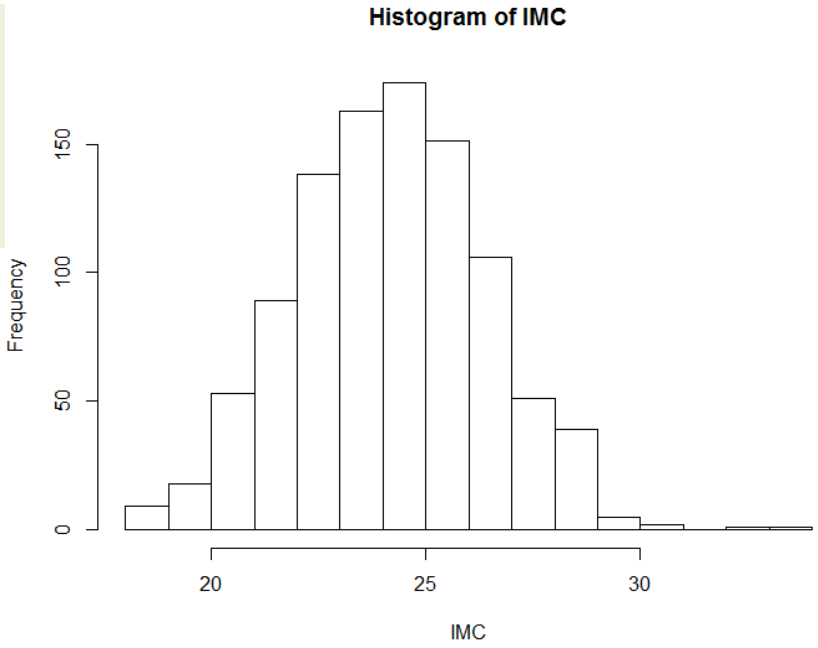


Histograma

Representar histogramas en R es tan sencillo como crear un objeto hist, con la función **hist().**

Ejemplo. Vamos a representar el Indice de Mas Corporal (IMC) que se comporta como una distribución normal de media 24.2 y desviación típica de 2.2

```
## Representación de Histogramas
## Ejemplo Indice de Masa Corporal
> IMC<-rnorm(n=1000, m=24.2, sd=2.2)
> hist(IMC)
```



Histograma

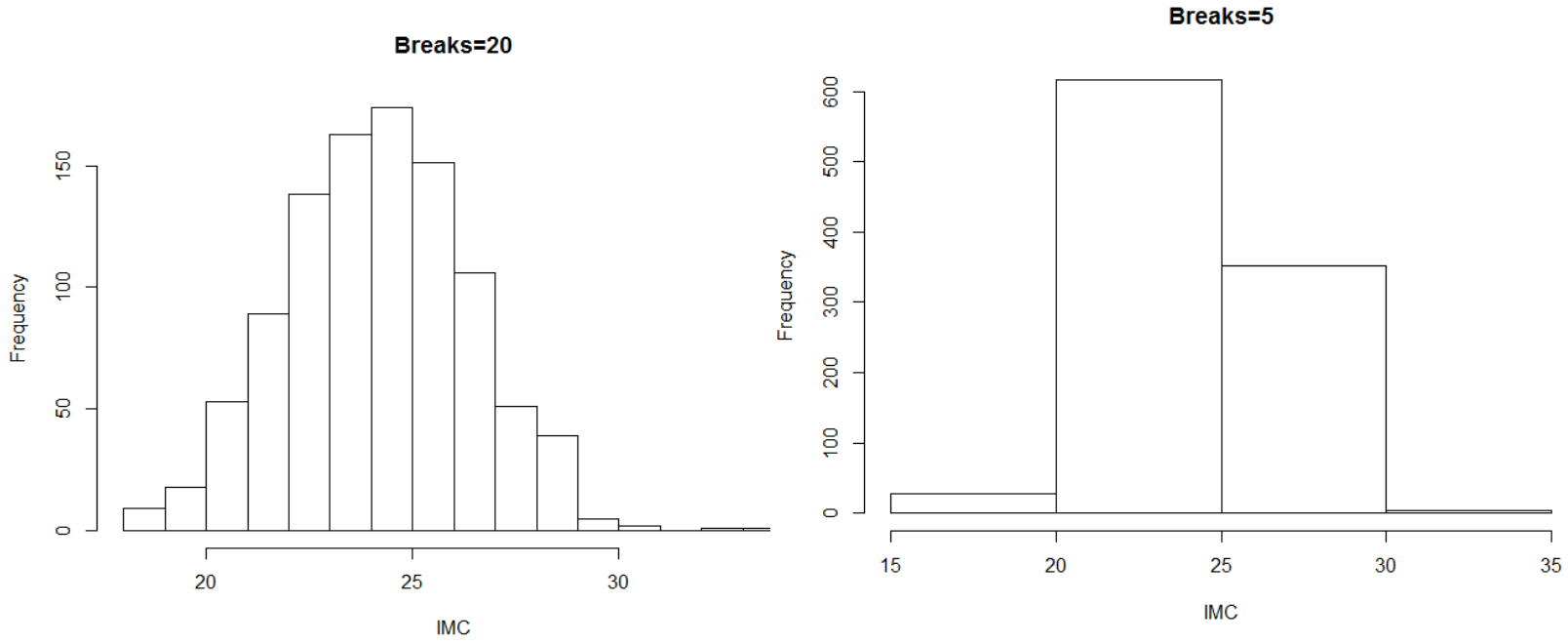
Si queremos obtener información sobre cualquier histograma, basta con poner lo siguiente:

```
#Información del Histograma
> histinfo<-hist(IMC)
> histinfo
$breaks
 [1] 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
$counts
 [1]   9  18  53  89 138 163 174 151 106  51  39   5   2   0   1   1
$density
 [1] 0.009 0.018 0.053 0.089 0.138 0.163 0.174 0.151 0.106 0.051 0.039 0.005
[13] 0.002 0.000 0.001 0.001
$mids
 [1] 18.5 19.5 20.5 21.5 22.5 23.5 24.5 25.5 26.5 27.5 28.5 29.5 30.5 31.5 32.5
[16] 33.5
$xname
[1] "IMC"
$equidist
[1] TRUE
attr(,"class")
[1] "histogram"
```

Histograma

Podemos cambiar el numero de clases con el parámetro **break()**

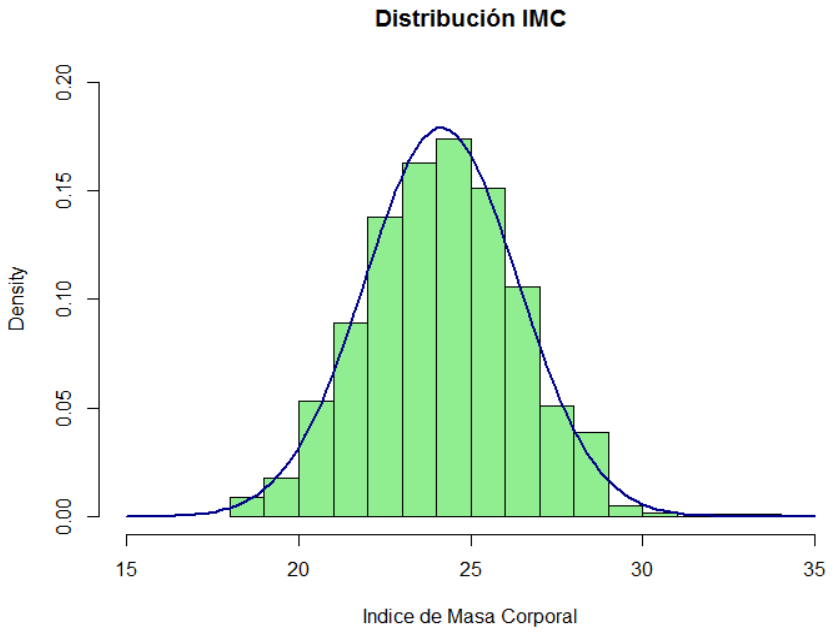
```
>hist(IMC, breaks=20, main="Breaks=20")
>hist(IMC, breaks=5, main="Breaks=5")
```



Histograma

Podemos añadir titulo, nombre de ejes y curva de distribución:

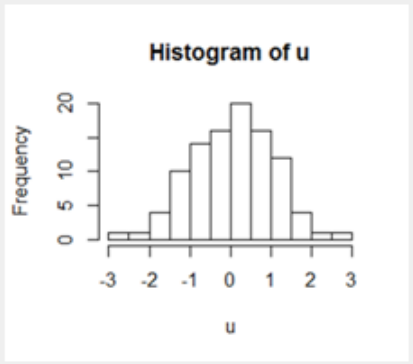
```
> hist(IMC, freq=FALSE, xlab="Indice de Masa Corporal",
+ main="Distribución IMC", col="lightgreen", xlim=c(15,35),  ylim=c(0, .20))
> curve(dnorm(x, mean=mean(IMC), sd=sd(IMC)), add=TRUE, col="darkblue", lwd=2)
```



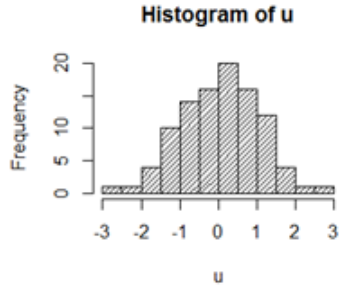
Histograma

Cambios en la forma de los histogramas:

```
u <- rnorm(100)
```




```
#histograma por defecto
hist(u)
```

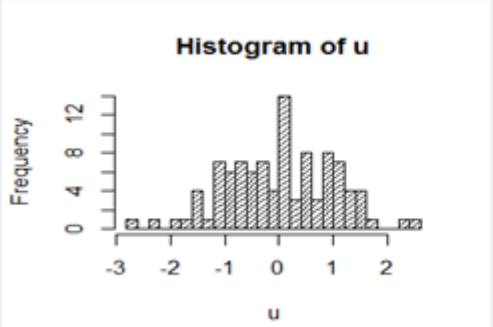



```
#cambiando en relleno
hist(u, density=20)
```

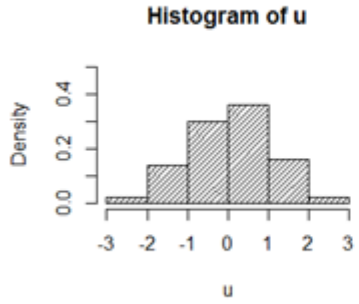
Histograma

Cambios en la forma de los histogramas:

```
u <- rnorm(100)
```



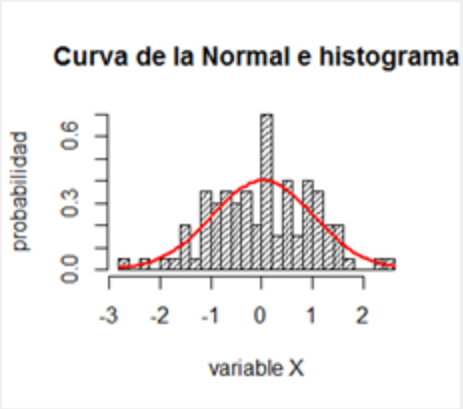

```
#Identificando el nº de columnas
hist(u, density=20, breaks=20)
```



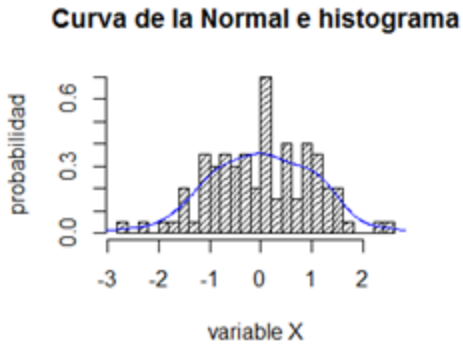

```
hist(u, density=20, breaks=-3:3,
     ylim=c(0,.5), prob=TRUE)
```

Histograma

Cambios en la forma de los histogramas: `u <- rnorm(100)`




```
m<-mean(u)
std<-sqrt(var(u))
hist(u, density=20, breaks=20,
prob=TRUE, xlab="variable X",
ylab="probabilidad",ylim=c(0, 0.7),
main="Curva de la Normal e
histograma")
curve(dnorm(x, mean=m,sd=std),
col="red", lwd=2, add=TRUE)
```

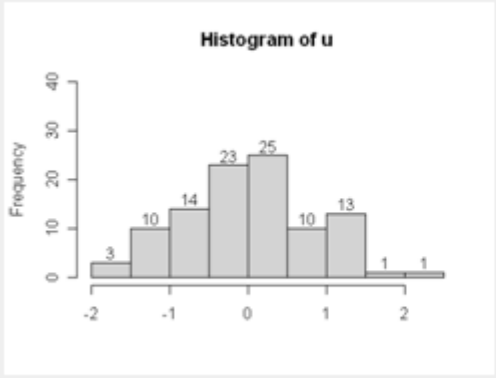



```
lines(density(u), col = "blue")
```

Histograma

Cambios en la forma de los histogramas:

```
u <- rnorm(100)
```




```
# en este caso escribe el n° de
# datos que hay en cada intervalo
aa<-hist(u)
plot(aa, ylim=c(0, 40),
col="lightgray",
    xlab="", main="Histogram of u")
text(aa$mids, aa$counts+2,
label=c(aa$counts))
```

Gráfica de tallos y hojas

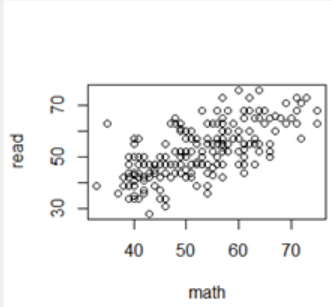
Permite la descripción de los datos agrupados en filas y columnas donde recuenta la frecuencia hasta la fila donde se encuentra la mediana, señalada por medio de paréntesis ():

```
> stem.leaf(AAA)
1 | 2: represents 12
 leaf unit: 1
            n: 52
   1    2. | 6
   4    3* | 122
  10    3. | 899999
  18    4* | 01112344
  (9)   4. | 566667789
  25    5* | 011333344
  16    5. | 55788
  11    6* | 000011134
   2    6. | 59
```

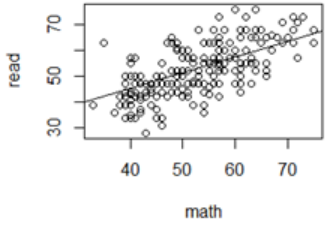
Diagramas de dispersión

Muestra conjuntamente datos de dos variables (en X y en Y) para ver su correlación, y permite considerar grupos (niveles de un factor)

```
hsb2 <- read.table('http://www.ats.ucla.edu/stat/r/modules/hsb2.csv',
header=T, sep=",")

attach(hsb2)
```

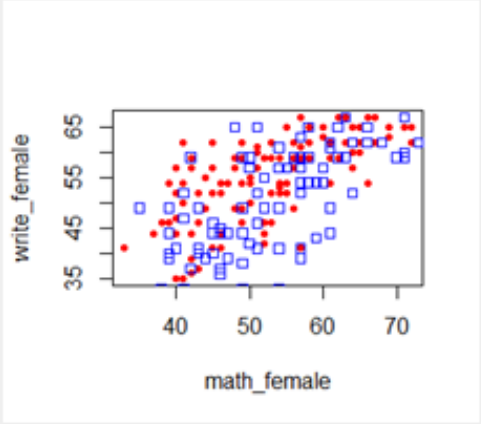



```
plot(math, read)
```

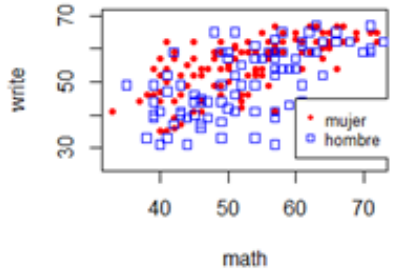



```
plot(math, read)

abline(lsfit(read, math))
```

Diagramas de dispersión



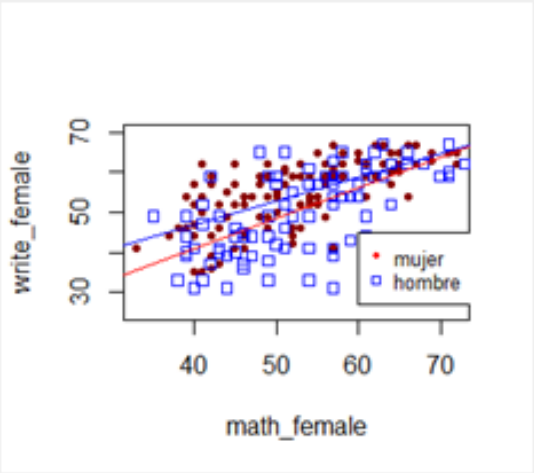

```
math_male<-hsb2$math[female==0]
write_male<-hsb2$write[female==0]
math_female<-hsb2$math[female==1]
write_female<-hsb2$write[female==1]

plot(math_female, write_female,
type="p",
     pch=20, col="red")
points(math_male, write_male,
pch=22, col="blue")
```



Añadiendo leyenda

```
hsb2_female<-hsb2[female==1,]
> hsb2_male<-hsb2[female==0,]
> with(hsb2_female, plot(math, write,
pch=20,
+ col="red", ylim=c(25, 70)))
> with(hsb2_male, points(math, write,
pch=22,
+ col="blue"))
> legend(60, 45, c("mujer", "hombre"),
 pch=c(20, 22),
+ cex=.8, col=c("red", "blue"))
```

Diagramas de dispersión




```
math_male<-hsb2$math[female==0]
> write_male<-hsb2$write[female==0]
> math_female<-hsb2$math[female==1]
> write_female<-hsb2$write[female==1]
> plot(math_female, write_female, type
="p", pch=20,
+ col="darkred", ylim=c(25, 70))
> points(math_male, write_male, pch=22
, col="blue")
> abline(lsfit(write_female, math_fema
le), col="red")
> abline(lsfit(write_male, math_male),
 col="blue")
> legend(60, 45, c("mujer", "hombre"),
 pch=c(20, 22),
+ cex=.8, col=c("red", "blue"))
```

Gráfica de sectores

Se trata de un diagrama en forma de círculo dividido en tantos sectores como datos distintos haya, en el que el ángulo de cada sector es proporcional a la frecuencia relativa del correspondiente dato.

Esta representación gráfica se denomina diagrama de sectores o diagrama de tarta. También puede usarse para datos cuantitativos agrupados en clases, y en tales casos, cada sector corresponde a una clase.

La función para diseñar diagramas de sectores en R es: **pie()**

Gráfica de sectores

Por ejemplo, la encuesta de población activa elaborada por el Instituto Nacional de Estadística referente al cuarto trimestre de 1970, presenta para el número de empleados por rama de actividad los siguientes datos:

Rama de Actividad	Miles de Empleados
Agricultura, caza y pesca	3706.3
Fabriles	3437.8
Construcción	1096.3
Comercio	1388.3
Transporte	648.7
Otros servicios	2454.8

Para almacenarlos en R:

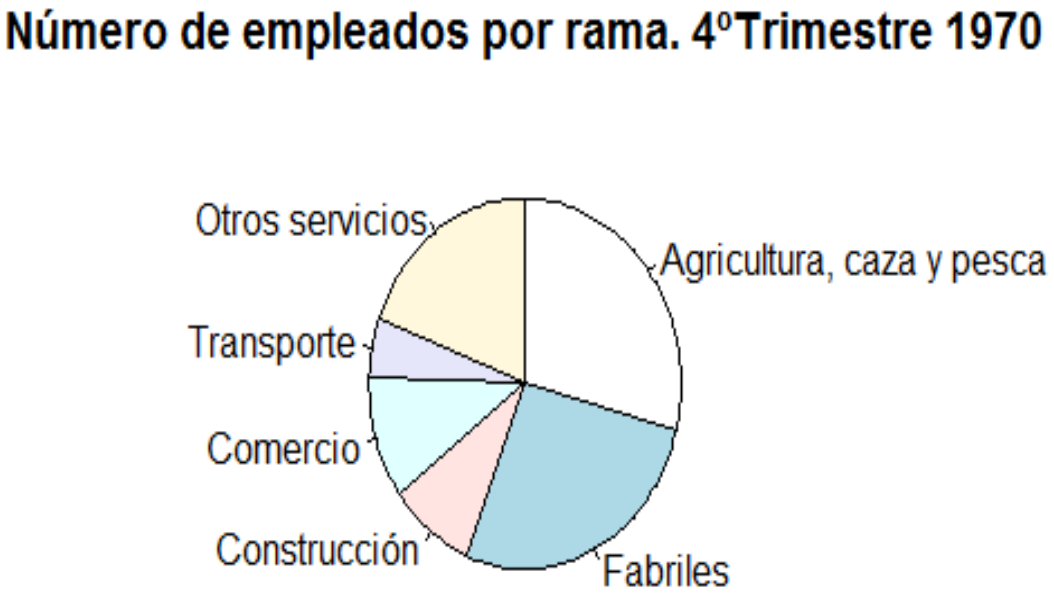
```
> Sector <- c(3706.3, 3437.8, 1096.3, 1388.3, 648.7, 2454.8)
```

Y posteriormente, le asignaremos la Rama de Actividad al vector Sector mediante la función names():

```
>names(Sector) <- c("Agricultura, caza y pesca", "Fabriles",
"Construcción", "Comercio", "Transporte", "Otros servicios")
```

Gráfica de sectores

```
pie(Sector, clockwise=TRUE, main="Número de empleados por rama.
4°Trimestre 1970", col=c(2,3,4,5,6,7))
```



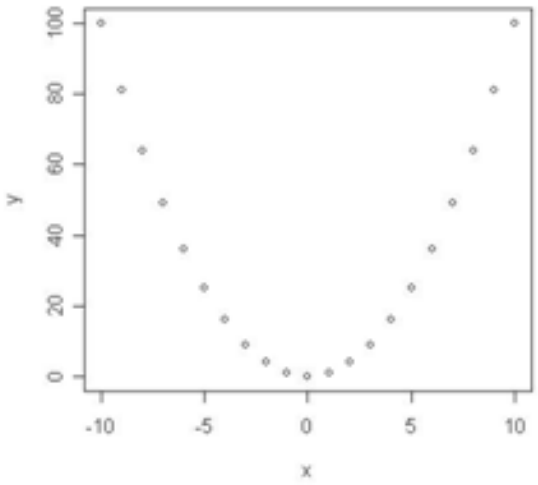
Gráfica XY

Permite comparar datos de dos variables cuantitativas

La función plot()

El comando plot se utiliza para crear una nueva figura.

```
> x = seq(-10,10)          # Generamos los números -10, -9,....,9, 10
> y = x^2                  # Generamos los cuadrados de dichos números
> plot(x,y)                # Graficamos
```



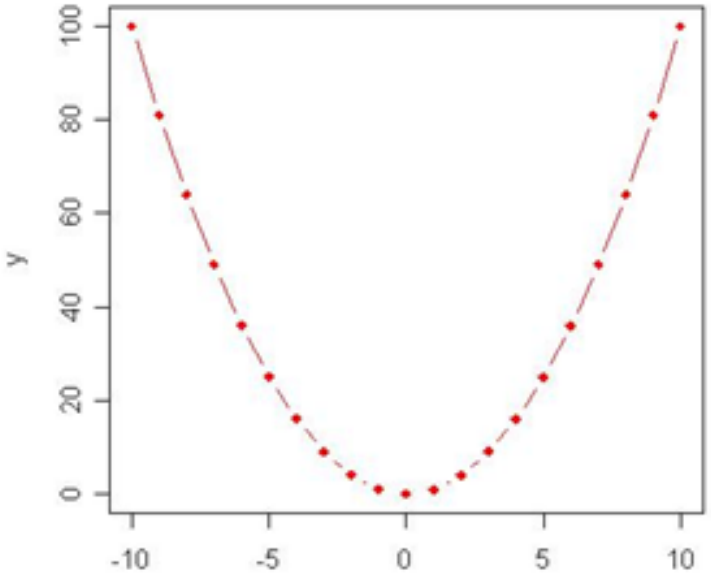
Gráfica XY

axes=F	Suprime la generación de los ejes
log="x" log="y" log="xy"	Hace que alguno de los ejes se tome en escala logarítmica
type="p" type="l" type="b" type="o" type="h" type="s" type="S" type="n"	Dibuja puntos individuales (opción por defecto) Dibuja líneas Dibuja puntos y líneas Dibuja puntos atravesados por líneas Dibuja con líneas verticales Dibuja a base de funciones escalera Casi lo mismo No dibuja nada. Pero deja marcados los puntos para manejos posteriores
xlab="cadena" ylab="cadena" main="cadena" sub="cadena"	Etiqueta para el eje de las x Etiqueta para el eje de las y Título del gráfico Subtítulo del gráfico
pch="simbolo" pch=18 pch="x" pch="P"	Se dibuja con el simbolo especificado. Por ejemplo:
col= numero entero col=2 col=3	Color para dibujar Color rojo Color verde

Gráfica XY

Algunos ejemplos:

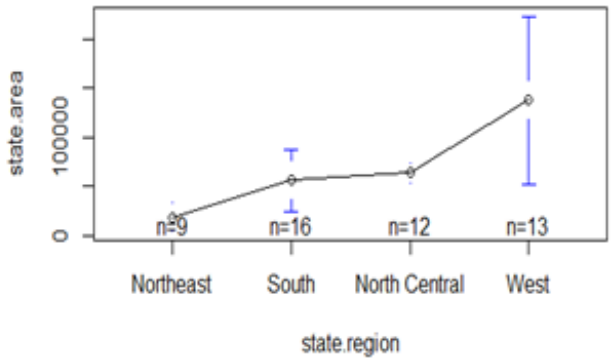
```
> x = seq(-10,10)
> y = x^2
> plot(x,y,type="l",xlab="eje de las x",ylab="eje de las y",
main="Parabola")
> plot(x,y,type="h",xlab="eje de las x",ylab="eje de las y",
main="Parabola",axes=F)
> plot(x,y,pch=18,col=2,type="b")
```



Gráfica de las medias

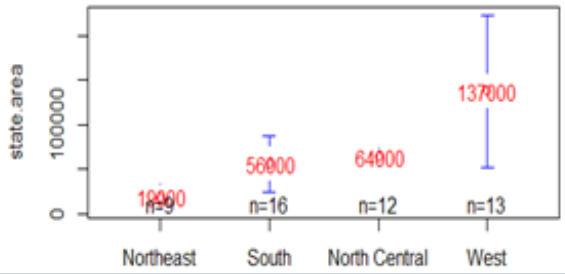
Permite comparar el efecto de los niveles de uno o dos factores en el comportamiento de una variable cuantitativa. Junto a las medias se añade a cada lado una desviación típica muestral que se ha elegido en las opciones.

```
> data(state)
> plotmeans(state.area ~ state.region)
```



Muestra el valor medio en otro color

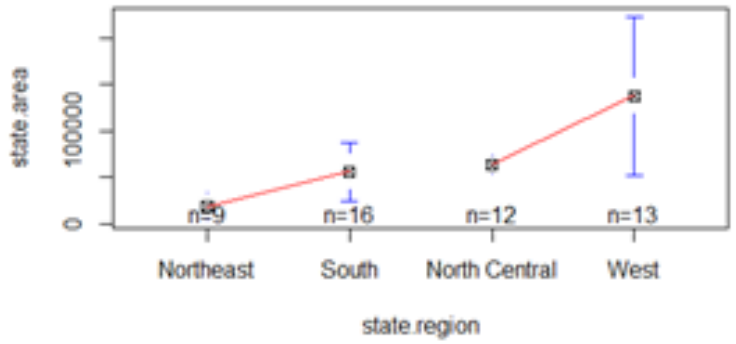
```
> plotmeans(state.area ~ state.region, mean.labels=TRUE, digits=-3,
col="red", connect=FALSE)
```



Gráfica de las medias

Conexión entre algunos de los valores medios

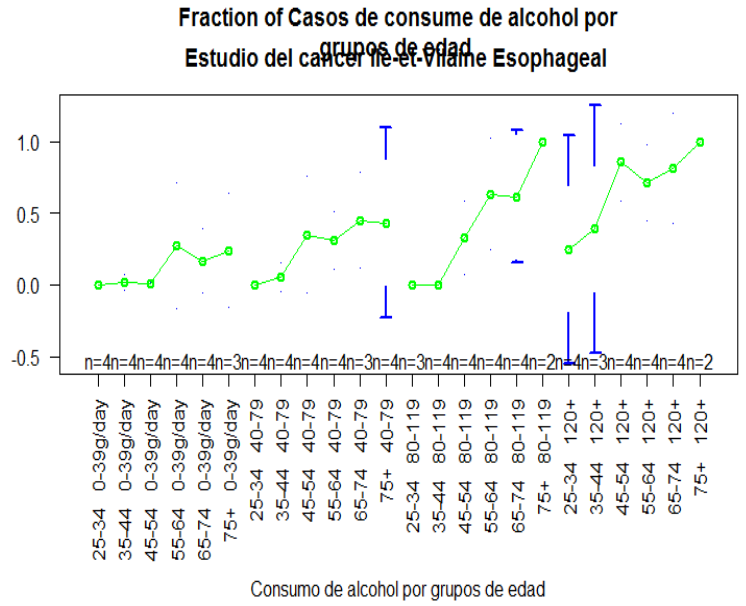
```
> plotmeans(state.area ~ state.region, connect=list(1:2, 3:4),  ccol="red",
pch=7 )
```



Gráfica de las medias

Y un ejemplo más complicado

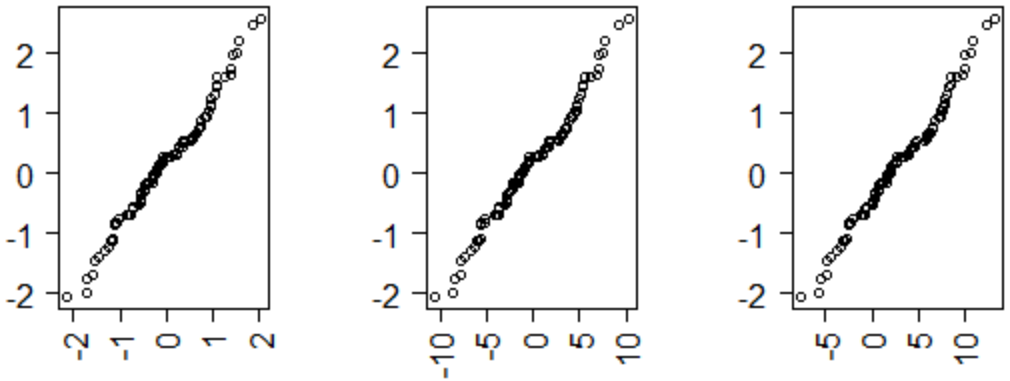
```
>  data(esoph)
>  par(las=2,                    # use perpendicular axis labels
      mar=c(10.1,4.1,4.1,2.1),   # create enough space for long x labels
      mgp=c(8,1,0)               # move x axis legend down to avoid overlap
      )
>  plotmeans(ncases/ncontrols ~ interaction(agegp , alcgp, sep ="    "),
           connect=list(1:6,7:12,13:18,19:24),
           barwidth=2,
           col="green",
           data=esoph,
           xlab="Consumo de alcohol por grupos de edad",
           ylab="# Casos / # Controles",
           main=c("Fraction of Casos de consumo de alcohol por
                       grupos de edad ",
                  "Estudio del cancer Ile-et-Vilaine Esophageal
           )
  abline(v=c(6.5, 12.5, 18.5), lty=2)
```



Matriz de diagramas de dispersión QQ

En una matriz de gráficas representa por parejas los datos asociados a un conjunto de variables cuantitativas. Extiende los Diagramas de dispersión a más de 2 variables. Permite considerar un factor cualitativo asociado a las variables cuantitativas.

```
x <- rnorm(100)
y <- rnorm(100)
par(mfrow=c(1,3))
qqplot(x,y ,cex.lab=2,cex.axis=1.5)
qqplot(5*x,y,cex.lab=2,cex.axis=1.5 )
qqplot(5*x+3,y,cex.lab=2,cex.axis=1.5)
```



x 5 * x 5 * x + 3