ESTADÍSTICA I
Tema 1: Estadística descriptiva

- Estadística descriptiva e inferencia estadística
- Resúmenes numéricos de un conjunto de datos
- Herramientas gráficas para el análisis exploratorio de datos
- Datos bivariantes. Correlación. Recta de regresión
- Implementación práctica con el programa `R`

El significado de "Estadística"

- Análisis de datos (estadística descriptiva): técnicas para resumir la información proporcionada por una gran masa de datos.
- Inferencia estadística: Se propone un modelo para describir una característica de interés en una "población" (una "variable") y se trata de estimar alguna característica de dicho modelo (la media de la variable, su distribución, etc.). En este caso se supone que los datos bajo estudio proceden de la observación de una **muestra** de la v.a. X. Es decir, que los datos provienen de tomar observaciones **independientes e idénticamente distribuidas** (iid) con la misma distribución de X.
- En muchos casos (tanto en los problemas descriptivos como inferenciales) se observan varias variables e interesa estudiar, de diferentes formas, la relación entre ellas.

Algunas nociones básicas sobre análisis de datos

Tomamos como punto de partida un **conjunto de datos**

$$x_1, \ldots, x_n.$$

Típicamente estos datos corresponden a la observación reiterada de una magnitud en diferentes individuos, en diferentes lugares, etc.

Para fijar ideas consideremos el conjunto de datos `Datos-IMC` que corresponde a los índices de masa corporal de 125 personas.

Un primer objetivo natural es **resumir la información proporcionada por estos datos**.

Estadísticos de tendencia central: la media

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Es la medida de tendencia central más utilizada.

Es bastante sensible a los valores "atípicos" (*outliers*), por ejemplo, observaciones anormalmente grandes que aparecen en un conjunto de datos debido a errores de medición o transcripción.

Ejemplo (`Datos-IMC.txt`): Cálculo de la media de los datos de `Datos-IMC` usando el programa `R`

```
x = scan("Datos-IMC.txt")
Read 125 items
mean(x)
[1] 25.63005
```

Estadísticos de tendencia central: la mediana

En términos informales, la mediana es un valor que divide a los datos en dos mitades: la mitad de los datos son menores que la mediana y la otra mitad son mayores.

Para calcular la mediana se ordenan los datos de menor a mayor:

$$x_{(1)},\ x_{(2)},\ \ldots,\ x_{(n)}.$$

Si el tamaño muestral n es impar, la mediana es el dato $x_{\left(\frac{n+1}{2}\right)}$, que ocupa el lugar central. Si n es par, la mediana se define como el promedio de los dos datos centrales, $x_{\left(\frac{n}{2}\right)}$ y $x_{\left(\frac{n}{2}+1\right)}$.

Ejemplo (`Datos-IMC.txt`):

```
median(x)
[1] 24.73526
length(x)
[1] 125
```

```
xOrd = sort(x)
xOrd[63]
[1] 24.73526
```

En general, el valor de la mediana no coincide con el de la media, aunque ambos estadísticos dan una idea de la tendencia central de los datos. La mediana es menos sensible que la media a los valores "anómalos" (outliers) anormalmente grandes en valor absoluto.

Por ejemplo, si se cambia el dato mayor de la muestra por otro valor cualquiera, mucho mayor que él, el valor de la mediana no se altera pero el de la media puede aumentar mucho (dependiendo del número de datos). Se dice, en términos técnicos, que la mediana es una medida de tendencia central "más robusta" que la media.

Ejemplo (`SpeedofLight.txt`): Calcular la media y la mediana de las observaciones que utilizó Newcomb para estimar la velocidad de la luz (en millonésimas de segundo). Recalcular ambas medidas retirando la menor observación del conjunto de datos.

Medidas de dispersión

La **varianza** de los datos $x_1, \ldots, x_n$ se define mediante

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{o} \quad s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Propiedad importante que facilita el cálculo de la varianza:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

Una medida de dispersión expresada en las mismas unidades que los datos es la **desviación típica**,

$$s_n = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2}.$$

A mayor varianza mayor dispersión de los datos en torno a su media.

Ejemplo (`Datos-IMC.txt`):

```
var(x)
27.31611
```

```
sd(x)
5.226482
```

Ejemplo (Notas en 3 grupos):

	Nota obtenida								
	2	3	4	5	6	7	8	9	10
Nº alumnos grupo A	0	0	0	40	60	0	0	0	0
Nº alumnos grupo B	1	5	15	24	31	18	4	1	1
Nº alumnos grupo C	6	12	14	18	24	9	3	5	9

```
x = c(rep(5,40),rep(6,60))
y = c(2,rep(3,5),rep(4,15),rep(5,24),rep(6,31),rep
  (7,18),rep(8,4),9,10)
z = c(rep(2,6),rep(3,12),rep(4,14),rep(5,18),rep
  (6,24),rep(7,9),rep(8,3),rep(9,5),rep(10,9))
```

Los cuartiles y los cuantiles

De manera análoga a la mediana, se definen **los cuartiles**:

- El primer cuartil Q_1 es un valor que deja la cuarta parte de los datos "a la izquierda" cuando se ordenan de menor a mayor y las tres cuartas partes a la derecha.
- El segundo cuartil Q_2 es la mediana.
- El tercer cuartil Q_3 deja las tres cuartas parte de los datos "a la izquierda" cuando se ordenan de menor a mayor y la cuarta parte a la derecha.

En general, para $p \in (0, 1)$ se llama **"cuantil p"**, q_p, o **"percentil $100p$"** al valor que deja una proporción p de los datos "a la izquierda" (es decir, el $100p\%$ de los datos son menores que él) y una proporción $1 - p$ "a la derecha" (es decir, son mayores).

Con esta notación, $q_{0.25} = Q_1$, $q_{0.75} = Q_3$.

Hay varios métodos para calcular los cuantiles muestrales. Todos hacen una media ponderada de dos observaciones consecutivas $x_{(j)}$ y $x_{(j+1)}$ de la muestra ordenada que aproximadamente dejan una proporción p de los datos "a la izquierda".

Para tamaños muestrales grandes, los resultados de todos los métodos son parecidos. `R` es el programa que ofrece un mayor número (9) de maneras de calcular los cuantiles: el método por defecto es `type=7`.

Ejemplo (`Datos-IMC.txt`):

```
quantile(x,0.25)
22.01189
quantile(x,0.75)
27.79046
help(quantile)
quantile(x,0.75,type=4)
     75%
27.72319
```

Para calcular el cuantil p a "mano" podemos utilizar este método (`type 7` de R): descomponemos $p(n-1)+1$ en su parte entera y decimal

$$p(n-1)+1 = j + m \quad \text{con } j \text{ entero y } \ 0 \leq m < 1.$$

Entonces

$$q_p = (1-m)x_{(j)} + m\, x_{(j+1)}$$

Ejemplo (`Datos-IMC.txt`):

Representación gráfica de datos: diagrama de caja

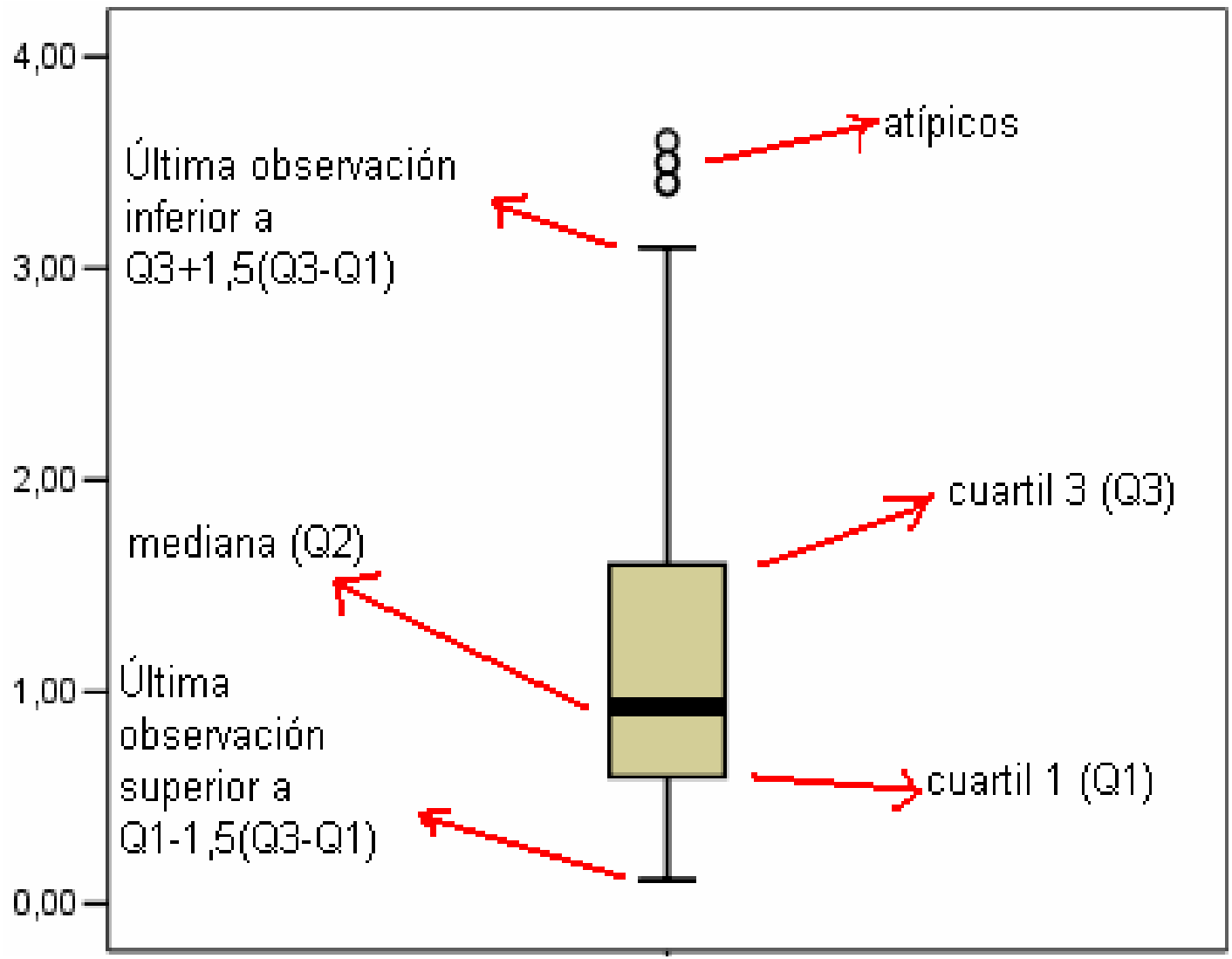
El **rango intercuartílico** (RI) es la diferencia entre el primer y el tercer cuartil: $\text{RI} = Q_3 - Q_1$.

Si separamos los datos ordenados en cuatro grupos con el mismo número de observaciones, el RI mide la distancia entre los dos grupos más extremos.

Para visualizar estas medidas de dispersión respecto a la mediana se utiliza el **diagrama de caja** (*box plot*).

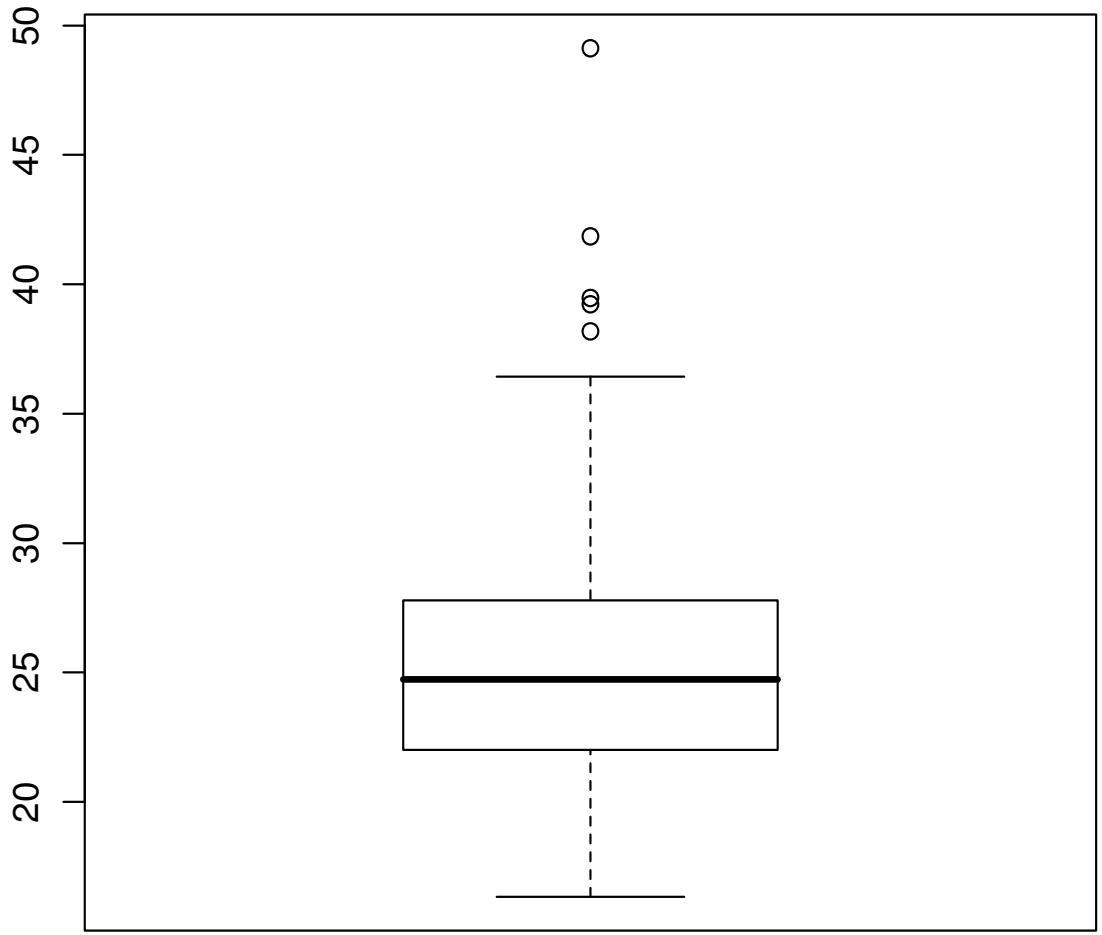
Para construir el diagrama de caja de la muestra, calculamos Q_1, Q_2, Q_3, RI y los límites inferior y superior del diagrama

$$\begin{array}{rcl} \text{LI} & = & \text{La menor observación en el intervalo} \\ & & [Q_1 - 1.5 \cdot \text{RI}, Q_3 + 1.5 \cdot \text{RI}] \\ \text{LS} & = & \text{La mayor observación en el mismo intervalo} \end{array}$$


4,00
3,00
2,00
1,00
0,00
Última observación inferior a Q3+1,5(Q3-Q1)
atípicos
cuartil 3 (Q3)
mediana (Q2)
Última observación superior a Q1-1,5(Q3-Q1)
cuartil 1 (Q1)

Ejemplo (`Datos-IMC.txt`):

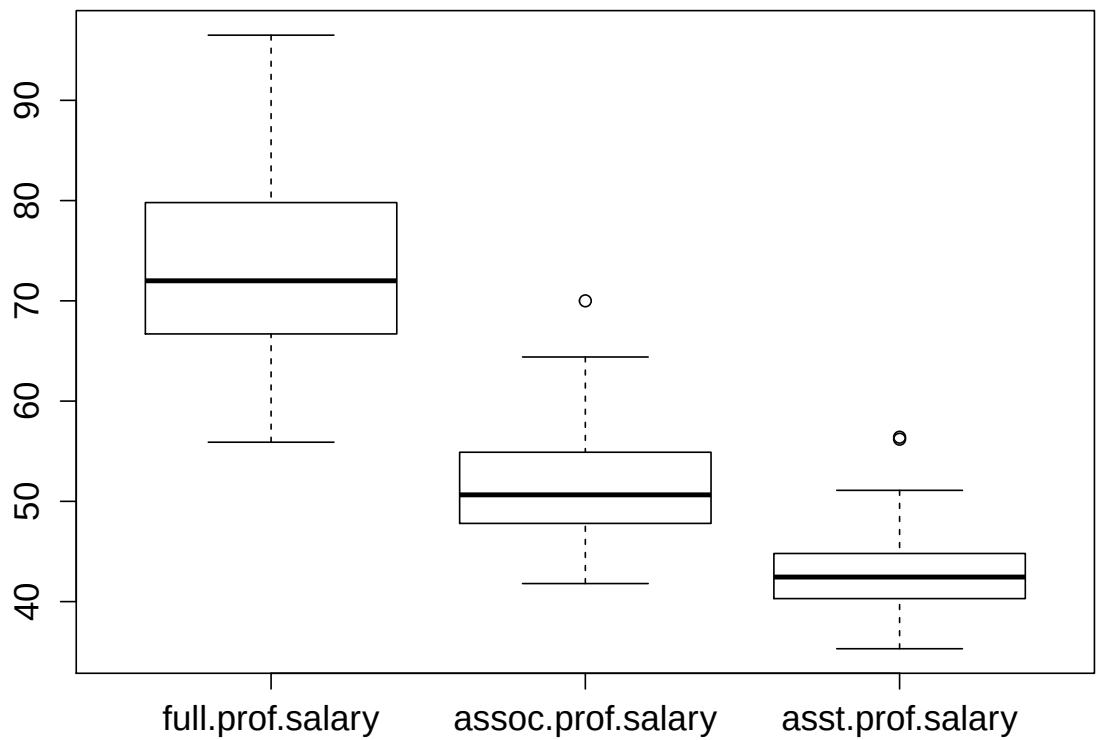
```
boxplot(x)
```



El diagrama de caja es especialmente útil para comparar grupos de datos entre sí.

Ejemplo (`FacultySalaries.txt`):

```
Datos = read.table("FacultySalaries.txt",header=TRUE)
boxplot(Datos$full.prof.salary,Datos$assoc.prof.salary,
  Datos$asst.prof.salary,names=c("full.prof.salary","assoc
  .prof.salary","asst.prof.salary"))
```



Representación gráfica de datos: el histograma

Fijada una sucesión $\ldots < a_i^{(n)} < a_{i+1}^{(n)} < \ldots$, con $h_n = a_{i+1}^{(n)} - a_i^{(n)}$ y dada la muestra $x_1, \ldots, x_n$, se define (siendo $\#C$ el cardinal del conjunto C)

$$\hat{f}_n(t; x_1, \ldots, x_n) \equiv \hat{f}_n(t) = \frac{\#\{i : x_i \in (a_j^{(n)}, a_{j+1}^{(n)}]\}}{nh_n},$$

para $t \in (a_j^{(n)}, a_{j+1}^{(n)}]$, $j = 0, \pm 1, \pm 2, \ldots$.

También se puede expresar esto con la (muy útil) notación de funciones indicatrices,

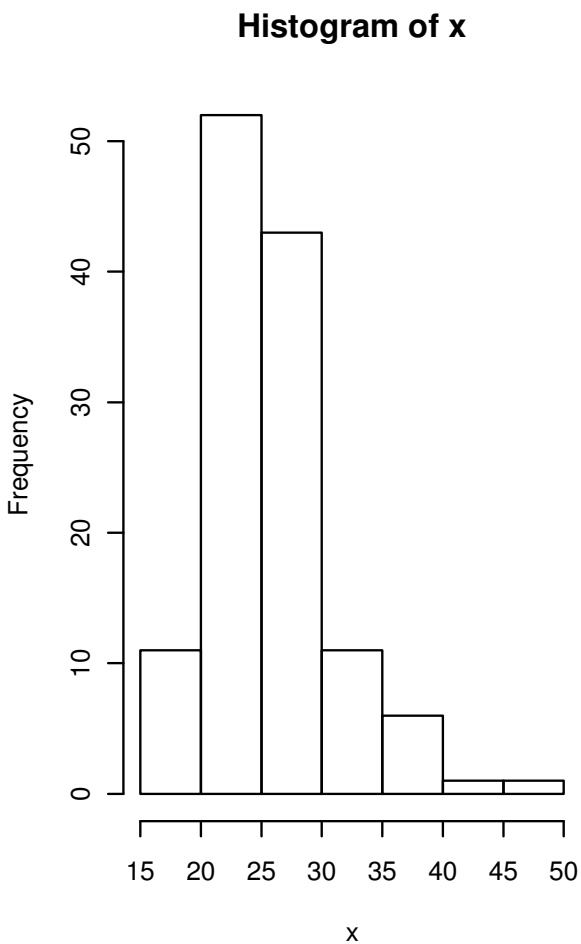
$$\hat{f}_n(t) = \frac{\sum_{i=1}^n \mathbb{1}_{(a_j^{(n)}, a_{j+1}^{(n)}]}(x_i)}{nh_n},$$

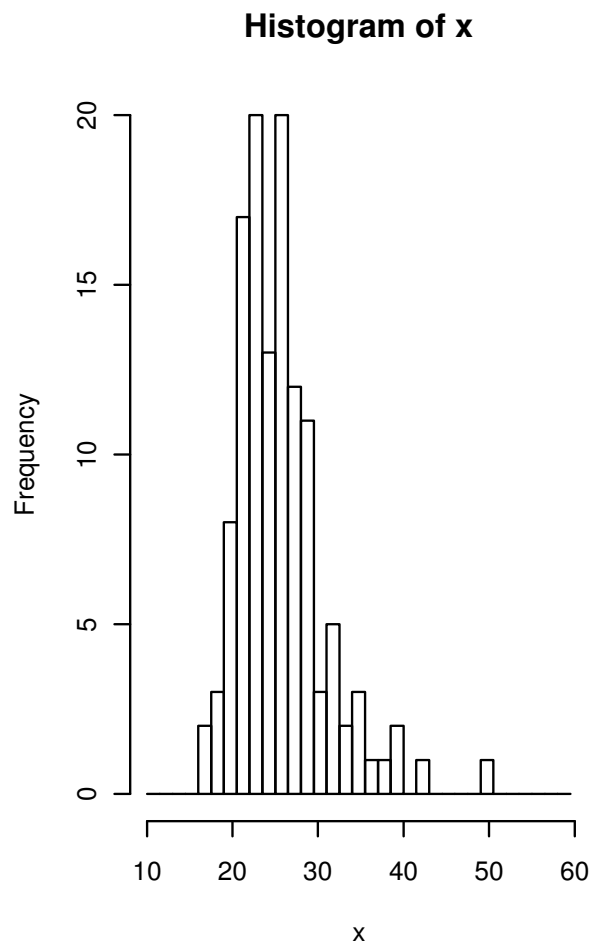
Ejemplo (Datos-IMC.txt):

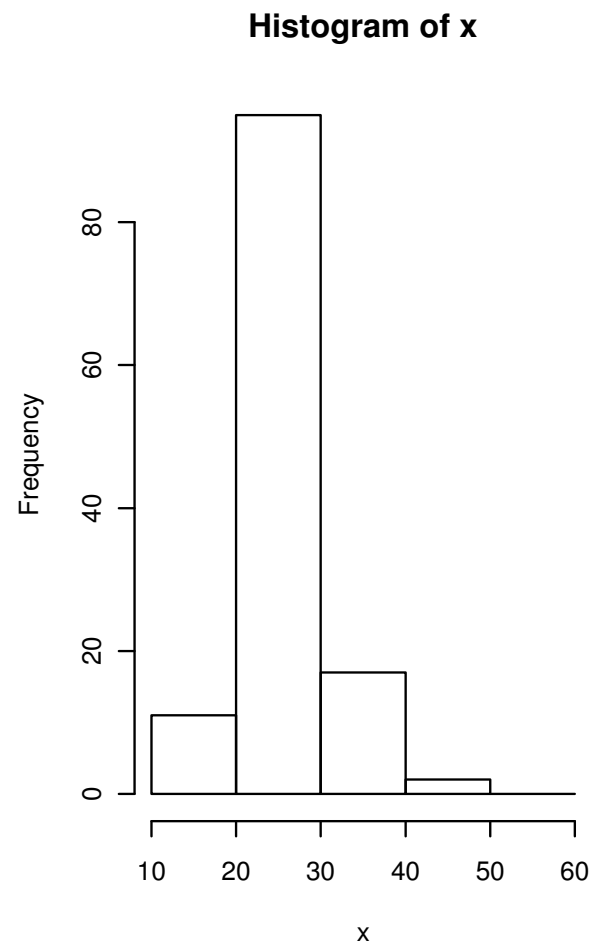
```
x = scan("Datos-IMC.txt")
H = hist(x,freq=FALSE)
H$breaks # nodos de la particion elegida por R
[1] 15 20 25 30 35 40 45 50
```

Estudiemos el efecto de variar la partición:

```
layout(matrix(1:3,1,3)) # Abrimos una ventana para
  tres graficos
# Ahora dibujamos los tres graficos.
# Con la opcion "por defecto" tenemos:
hist(x)
# Con una particion "mas fina":
hist(x,breaks=seq(10,60,1.5))
# Con una particion "menos fina":
hist(x, breaks=seq(10,60,10))
layout(1)
```


Histogram of x
Frequency
0
10
20
30
40
50
15
20
25
30
35
40
45
50
x


Histogram of x
Frequency
0
5
10
15
20
10
20
30
40
50
60
x


Histogram of x
Frequency
0
20
40
60
80
10
20
30
40
50
60
x

Estimadores núcleo o kernel

Otras formas de visualizar la distribución de los datos:

$$\hat{f}_n(t) = \frac{1}{n2h_n} \sum_{i=1}^{n} \mathbb{1}_{[t-h_n, t+h_n]}(x_i) = \frac{1}{n2h_n} \sum_{i=1}^{n} \mathbb{1}_{[-1,1]} \left(\frac{t - x_i}{h_n} \right)$$

Se puede generalizar este procedimiento (método de ventana móvil o *moving window method*) reemplazando la densidad uniforme $\mathbb{1}_{[-1,1]}/2$ por otra función de densidad K (llamada *kernel* o núcleo). Un ejemplo típico es $K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ (núcleo gaussiano). Obtenemos entonces el estimador *kernel*

$$\hat{f}_n(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K \left(\frac{t - x_i}{h_n} \right)$$

que es una versión "suavizada" del histograma.

Obsérvese que $\hat{f}_n$ es una función de densidad ($\hat{f}_n \geq 0$ y $\int \hat{f}_n = 1$).

Supongamos, por ejemplo, que K es el núcleo gaussiano.
Podríamos generar datos "artificiales" x_i^0 de la densidad $\hat{f}_n$ así:

$$x_i^0 = x_i^* + h_n Z_i, \ \ i = 1, \ldots, k$$

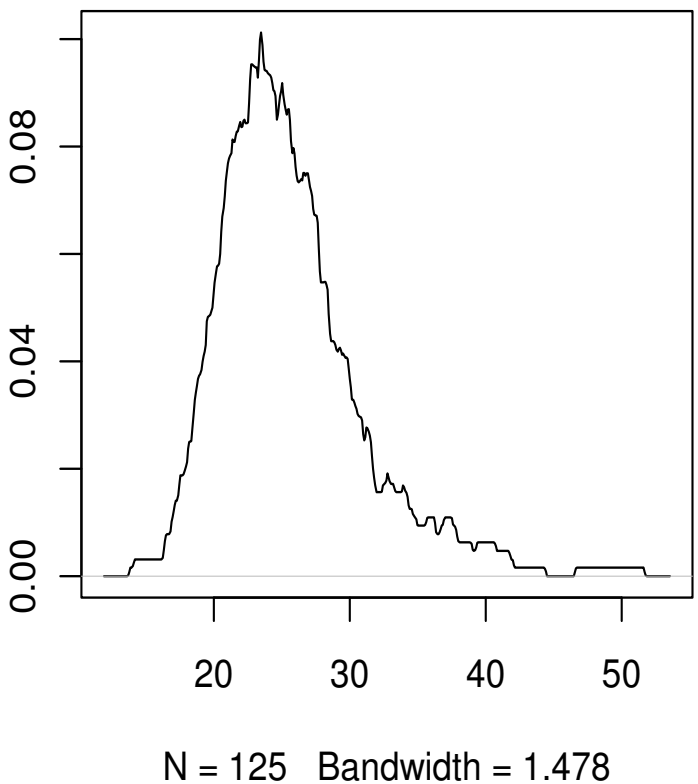
donde x_i^* es una observación elegida al azar con igual probabilidad entre los datos originales y Z_i es una observación aleatoria $N(0, 1)$.
Como Z_i está multiplicada por el parámetro de suavizado h_n, la observación $h_n Z_i$ sigue la distribución $N(0, h_n)$.
Es decir, tomar una observación aleatoria cuya densidad sea $\hat{f}_n$ equivale a tomar al azar uno de los datos originales $x_1, \ldots, x_n$ y sumarle una "pequeña perturbación aleatoria" $N(0, h_n)$.

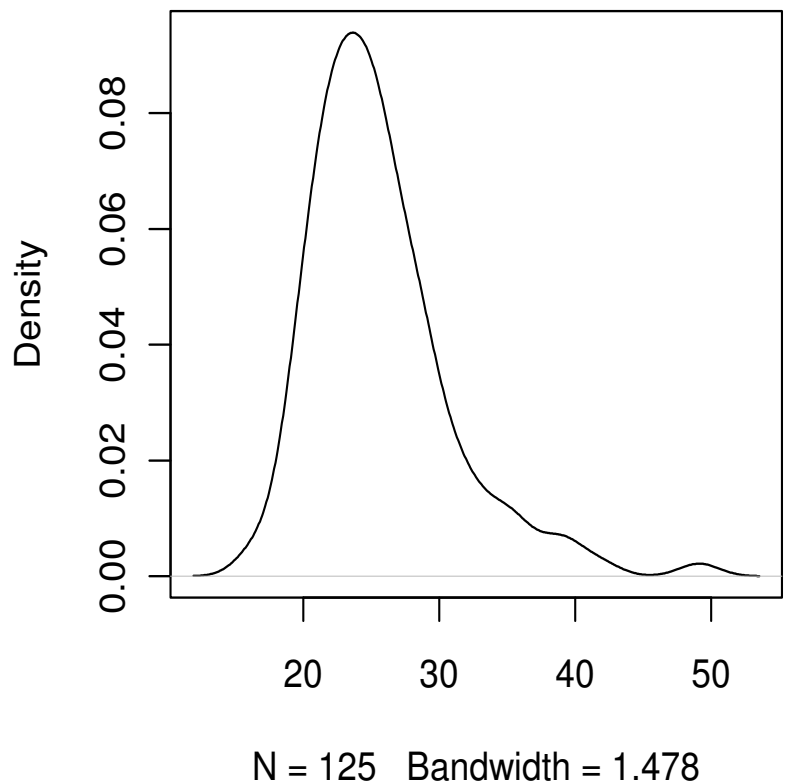
Intuitivamente, el estimador *kernel* $\hat{f}_n$ es una densidad "basada en los datos" $x_1, \ldots, x_n$ que, asintóticamente, aproxima la densidad "verdadera", f, que los generó (si es que proceden de una distribución continua). En el Tema 2 veremos esto con más detalle.

Ejemplo (`Datos-IMC.txt`):

El estimador kernel con núcleo uniforme (i.e. constante en un intervalo) y con núcleo gaussiano se obtienen respectivamente con

```
plot(density(x,kernel="rectangular"))
plot(density(x,kernel="gaussian"))
```

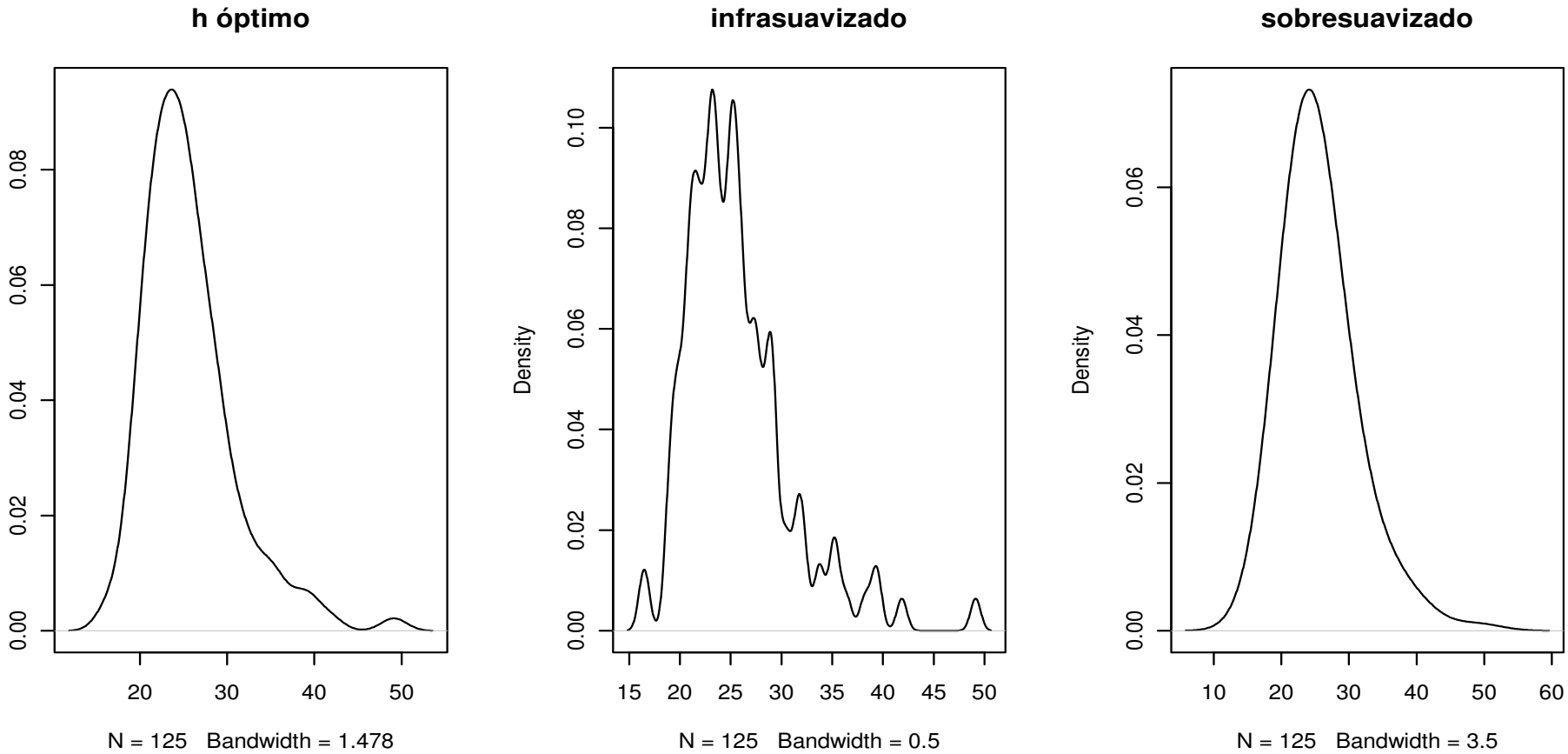




Ejemplo (`Datos-IMC.txt`):

Examinemos el efecto de variar el parámetro de suavizado (*bandwidth* o *smoothing parameter*) h.

```
layout(matrix(1:3,1,3))
plot(density(x,kernel="gaussian"),main="h optimo")
plot(density(x,kernel="gaussian", bw=0.5),main="
  infrasuavizado")
plot(density(x,kernel="gaussian", bw=3.5),main="
  sobresuavizado")
```



Existen diferentes criterios de selección de h, pero no se estudiarán aquí.

En general están basados en la idea de minimizar el error cometido por $\hat{f}_n(t)$ cuando se utiliza como estimador de la "verdadera densidad" $f(t)$ de la variable de la cual proceden los datos.

El coeficiente de asimetría (skewness)

El tercer momento respecto a la media

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

da una medida natural de la asimetría de los datos en torno a su media. Para obtener una medida adimensional podemos dividir por s_n^3. Se define entonces el coeficiente de asimetría como el momento de orden 3 de los datos estandarizados

$$\frac{1}{n\, s_n^3}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Ejemplo (`Datos-IMC.txt`):

```
library(moments)
skewness(x)
[1] 1.434180
```

Análisis exploratorio de datos bivariantes

En los problemas de regresión se observan DOS variables (x, y) en cada uno de los individuos de la muestra. Por ejemplo:

- En un estudio sobre problemas de nutrición puede ser $x =$ consumo de hidratos de carbono de una persona, $y =$ índice de masa corporal de esa persona.
- En un estudio de mercado, podemos considerar $x =$ gastos en publicidad de una pequeña empresa del sector informático, $y =$ facturación de esa empresa
- $x =$ concentración de colesterol en la sangre de una persona, $y =$ concentración de triglicéridos en esa misma persona.
- $x =$ ingresos anuales netos de una familia, $y =$ gasto en alimentación de esa familia.

En general, los objetivos de analizar observaciones multivariantes o vectores aleatorios son:

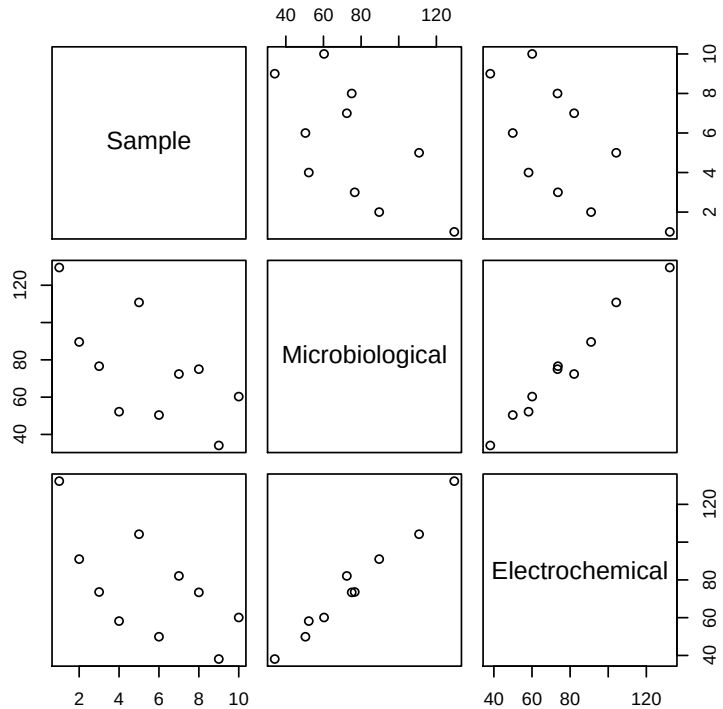
- Entender mejor la relación entre las dos variables
- Predecir o aproximar el valor de una de ellas (digamos las y) cuando se conoce el valor de la otra.

Ejemplo (monensina): Marecek *et al.* (1991) desarrollaron un nuevo método electroquímico para determinar rápidamente la concentración de monensina, un antibiótico poliéter, en las cubas de fermentación donde se produce. El método estándar, un análisis de actividad microbiológica, era complicado y consumía mucho tiempo. Se tomaron muestras en diez cubas de fermentación y se midió la concentración (en ppt) de monensina en cada una de ellas utilizando ambos métodos:

Muestra	Microbiológico	Electroquímico
1	129.5	132.3
2	89.6	91.0
3	76.6	73.6
4	52.2	58.2
5	110.8	104.2
6	50.4	49.9
7	72.4	82.1
8	75.0	73.4
9	34.1	38.1
10	60.3	60.1

El diagrama de dispersión

```
Datos = read.table("Monensin.txt",header=T)
pairs(Datos)
```



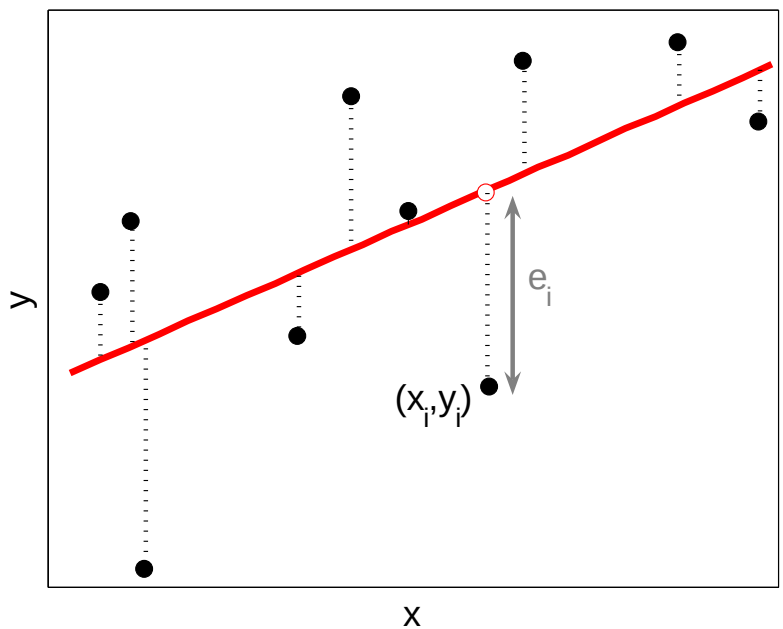
También

```
x = Datos$Microbiological
y = Datos$Electrochemical
plot(x,y)
```

La recta de regresión

La recta de regresión de y sobre x basada en los datos $(x_1, y_1), \ldots, (x_n, y_n)$ coincide con la recta de ajuste por mínimos cuadrados con ecuación $y = a + bx$. Los valores a y b se obtienen minimizando la suma de cuadrados de distancias verticales de los puntos a la recta:

$$(\hat{a}, \hat{b}) = \arg\min_{(a,b)} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$



Un cálculo elemental permite deducir:

$$\hat{b} = \frac{s_{x,y}}{s_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x},$$

donde

$$s_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n}\left(\sum_{i=1}^n x_i^2\right) - \bar{x}^2$$

es la varianza muestral de x y

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\left(\sum_{i=1}^n x_i y_i\right) - \bar{x}\bar{y}$$

es la covarianza muestral entre x e y.

$\hat{b}$ se denomina coeficiente de regresión lineal de y sobre x o, simplemente, el parámetro de la regresión.

La recta de ecuación $\hat{y} = \hat{a} + \hat{b}x$ se denomina recta de regresión de y sobre x. El valor pronosticado o previsto de y dado $x = x_i$ es $\hat{y}_i = \hat{a} + \hat{b}x_i$. Los valores $e_i = y_i - \hat{y}_i$ se denominan residuos.

Obsérvese que $\hat{a}$ y $\hat{b}$ han sido seleccionados para minimizar la suma de cuadrados residuales $\sum_{i=1}^{n} e_i^2$. Nótese también que $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$.

Una manera de medir el "error cuadrático" cometido en la aproximación dada por la recta de regresión es mediante la llamada varianza residual

$$s_R^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2.$$

El coeficiente de correlación lineal

El coeficiente de correlación lineal (de Pearson) evaluado a partir de los datos (x_i, y_i) es

$$r = \frac{s_{x,y}}{s_x s_y},$$

donde s_y^2 es la varianza muestral de y.

Observemos que

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 = s_y^2(1 - r^2). \qquad (1)$$

Como el error cuadrático medio (1) debe ser menor que s_y^2 (el error cometido al aproximar los y_i con la media muestral), se cumple que

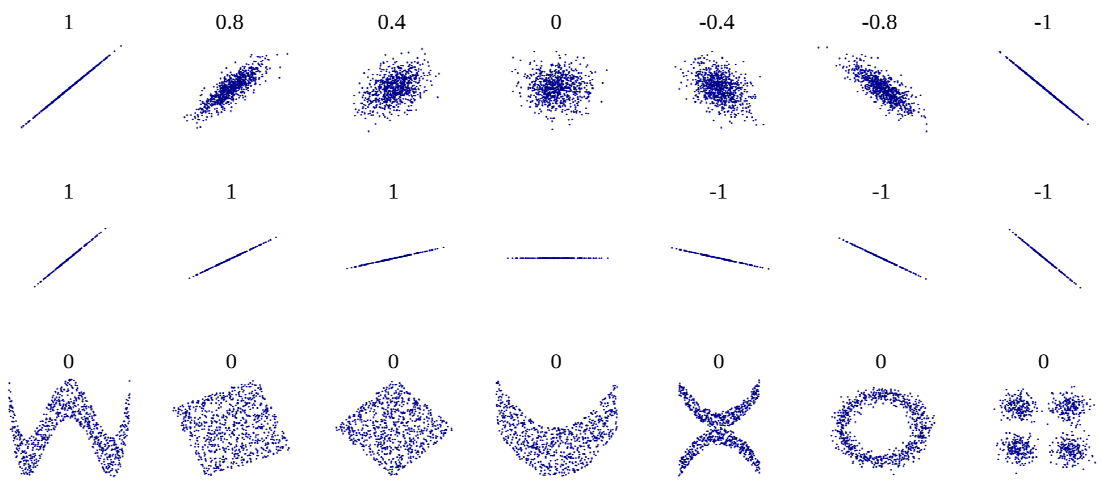
$$0 \leq r^2 \leq 1.$$

Obsérvese que $r = \hat{b}\,\frac{s_x}{s_y}$. En particular, r y $\hat{b}$ tienen el mismo signo.

Una correlación r cercana a 1 indica un alto grado de ajuste lineal de y en términos de x. Se dice que hay una alta "correlación positiva" o "relación lineal directa" entre ambas variables (en el sentido de que al aumentar los valores de una de ellas esperamos que aumenten los correspondientes valores de la otra).

Un r cercano a -1 indica también un alto grado de ajuste lineal de y en términos de x pero en este caso hay una "correlación negativa" o "relación lineal inversa" entre ambas variables.

Un r cercano a 0 se interpreta como una débil asociación lineal entre x e y.

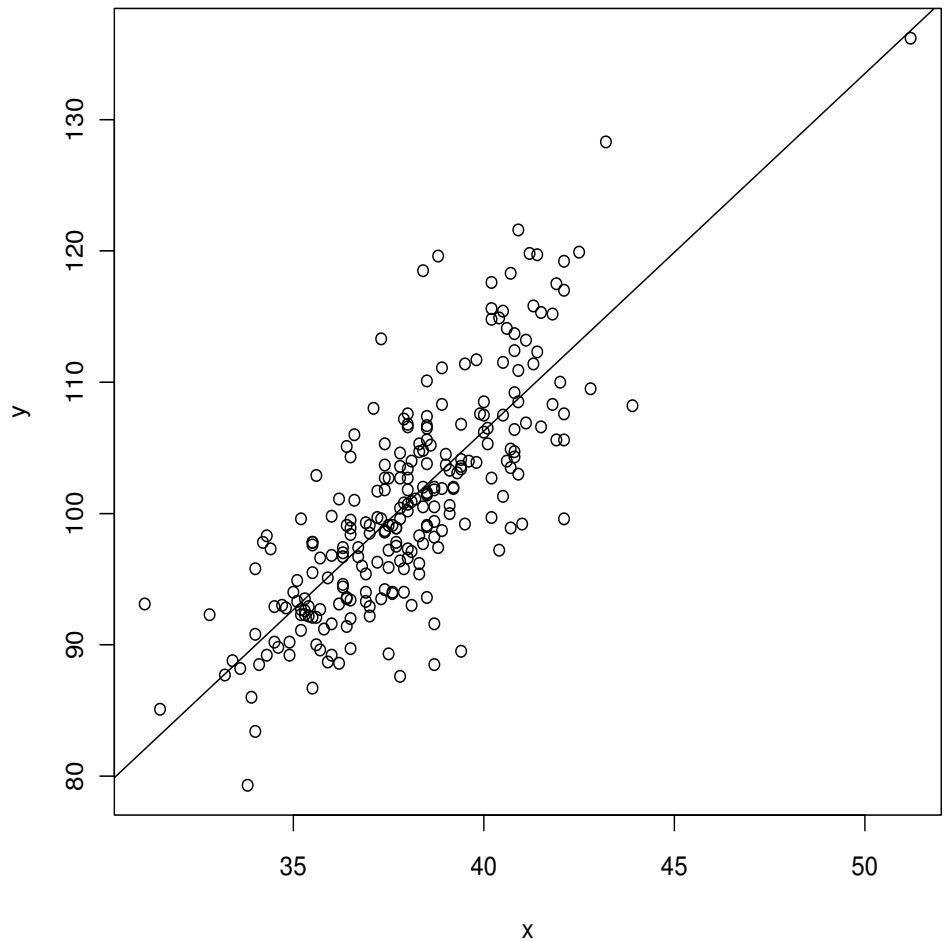


Ejemplo (Datos-bodyfat.txt):

```
xx = read.table("Datos-bodyfat.txt")
dim(xx)
[1] 252  15
# Estudiamos la relacion entre las variables
# x= "circunferencia del cuello"
# y= "circunferencia del pecho"
x = xx$V6
y = xx$V7
cor(x,y)
[1] 0.784835
# Recta de regresion (calculo de los coeficientes):
lm(y~x)
Coefficients:
(Intercept)            x
     -2.584        2.722
```

Ejemplo (Datos-bodyfat.txt):

```
# Para dibujar el grafico:
zz <- lm(y~x)
plot(x,y)
abline(zz)
```



Una página web

En el enlace

http://www.gapminder.org/

se pueden encontrar varios ejemplos interesantes de análisis exploratorio de datos con técnicas gráficas avanzadas.

Referencias

Dekking, F.M., Kraaikamp, C., Lopuhaa, H.P. y Meester, L.E. (2005). *A Modern Introduction to Probability and Statistics: Understanding Why and How*. Springer.

Grolemund, G. y Wickham, H. (2016). *R for Data Science*. O'Reilly Media.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.