

# Estadística

# Contenido

- 1. Introducción
  - ¿Qué es la estadística?
  - Modelo estadístico
  - Método estadístico
  - Algunas herramientas de análisis de datos mediante estadística
- 2. Datos
  - Descripción de una sola variable
  - Descripción conjunta de varias variables
- 3. Modelos
  - Probabilidad y variables aleatorias
  - Modelos de distribución de probabilidad
  - Modelos Multivariantes
- 4. Inferencia estadística
  - Estimación puntual
  - Estimación por intervalos
  - Estimación bayesiana
  - Contraste de hipótesis

# Introducción

## ➤ Introducción

- ¿Qué es la estadística?
- Modelo estadístico
- Método estadístico
- Construir un Modelo estadístico
- Algunas herramientas de análisis de datos mediante estadística (además de Python): Matlab y SPSS

# Introducción

## ➤ ¿Qué es la estadística?

- Es la ciencia que estudia como obtener conclusiones de la investigación empírica mediante el uso de **modelos matemáticos**.
- Actúa como vínculo en los modelos matemáticos y los fenómenos reales.
- Un modelo matemático es una abstracción simplificada de algún aspecto de la realidad más compleja.
  - Siempre existen **discrepancias** entre el modelo y lo que representa este en la realidad.
  - La estadística proporciona una metodología para medir esas discrepancias (fundamental en ciencia aplicada, como tecnología, economía, sociología, medicina, y hoy en día en Big Data, etc.).
  - Es la **tecnología** del método científico experimental.

# Introducción

- ¿Que problemas puede resolver?
  - **Descripción de datos:** permite resumir la información contenida en los datos.
  - **Análisis de muestras:** No es posible en algunos casos estudiar todos los elementos de una población. Así dada una muestra representativa se pueden hacer inferencias respecto la población completa. La inferencia se puede realizar mediante un modelo probabilístico obtenido de la muestra. Hoy en día con Big Data, casi se puede tener una población.
  - **Contrastación de hipótesis:** En investigación empírica o experimental es fundamental la contrastación de hipótesis. Nos permite comparar las predicciones resultantes de la hipótesis con los datos observados.
  - **Medición de relaciones:** las relaciones que observamos entre variables físicas, sociales y técnicas son prácticamente siempre estadísticas, por los errores de medición.
  - **Predicción:** el estudio de la historia de variables estadísticas permite inferir valores futuros.

# Introducción: Método estadístico



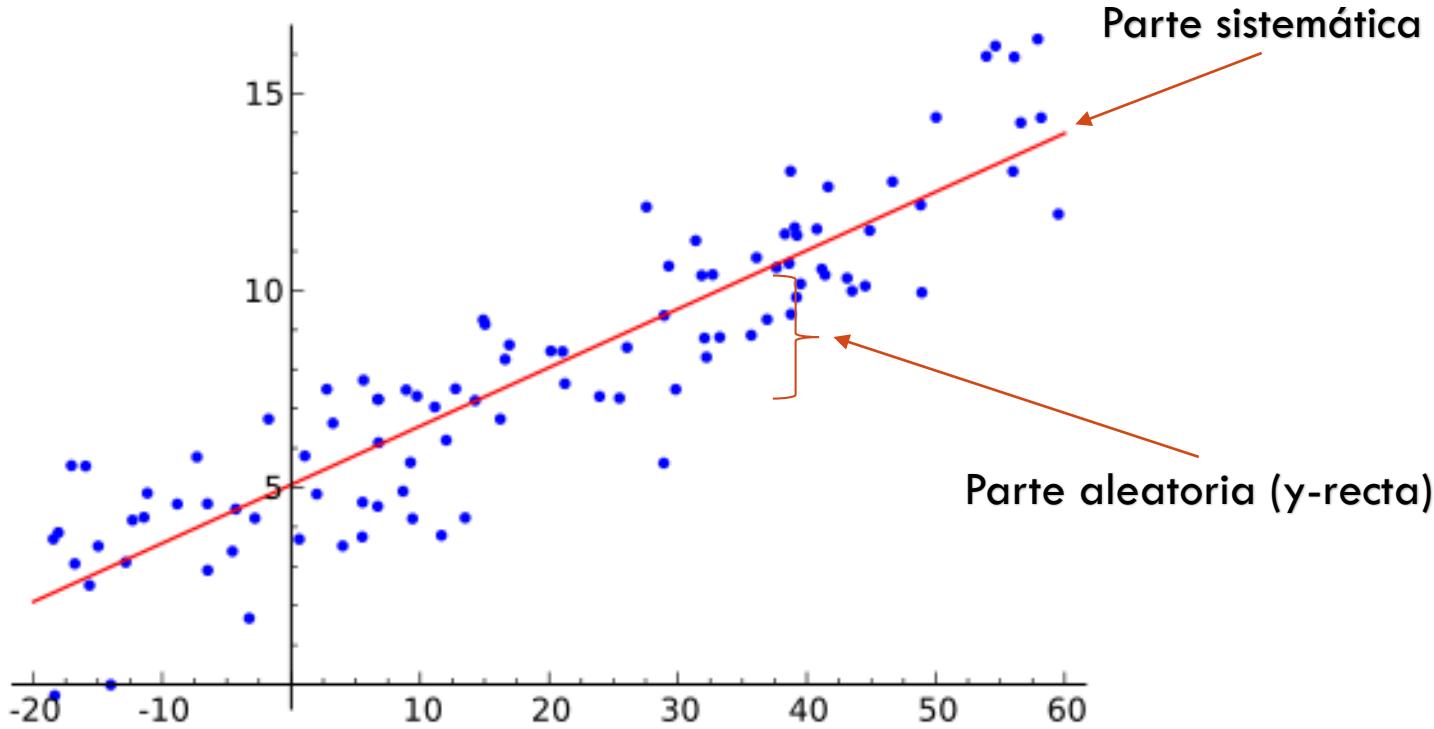
- **Estadística descriptiva:** recolecta, ordena, analiza y representa a un conjunto de datos, con el fin de describir apropiadamente las características de este y así resumirlos.
  - Calcula qué parámetros estadísticos del conjunto de datos mediante: gráficos, medidas de frecuencias, centralización, posición, dispersión, etc.
- **La estadística inferencial:** saca conclusiones generales para toda la población a partir del estudio de una muestra (para un grado de fiabilidad o significación de los resultados obtenidos).
  - Utiliza herramientas como el muestreo, la estimación de parámetros, el contraste de hipótesis, etc.

# Introducción: Modelo estadístico

- El **Modelo estadístico descompone** una variable respuesta de estudio en dos valores:
  - Parte predecible o sistemática
  - Parte aleatoria, impredecible o residual.
- Así el modelo estadístico define la forma de la **parte predecible** (respuesta media), y la **variabilidad de la impredecible** respecto a la respuesta media.
  - Observación=parte sistemática (predecible) + parte aleatoria (impredecible).
- En un problema real, en definitiva un variable respuesta  $y$  (o varias  $y_i$  en el caso multivariante), será función de otras variables de número desconocido (incluso no observables), siendo el modelo estadístico un acercamiento aproximación de esta realidad.
  - Esta aproximación tiene en cuenta las variables observables presumiblemente más importantes, y engloba en la parte aleatoria los efectos del resto de variables que no se tienen en cuenta.
- Para el cálculo de **modelos estadísticos** se suele utilizar **distribuciones de probabilidad**.

# Introducción: Método estadístico

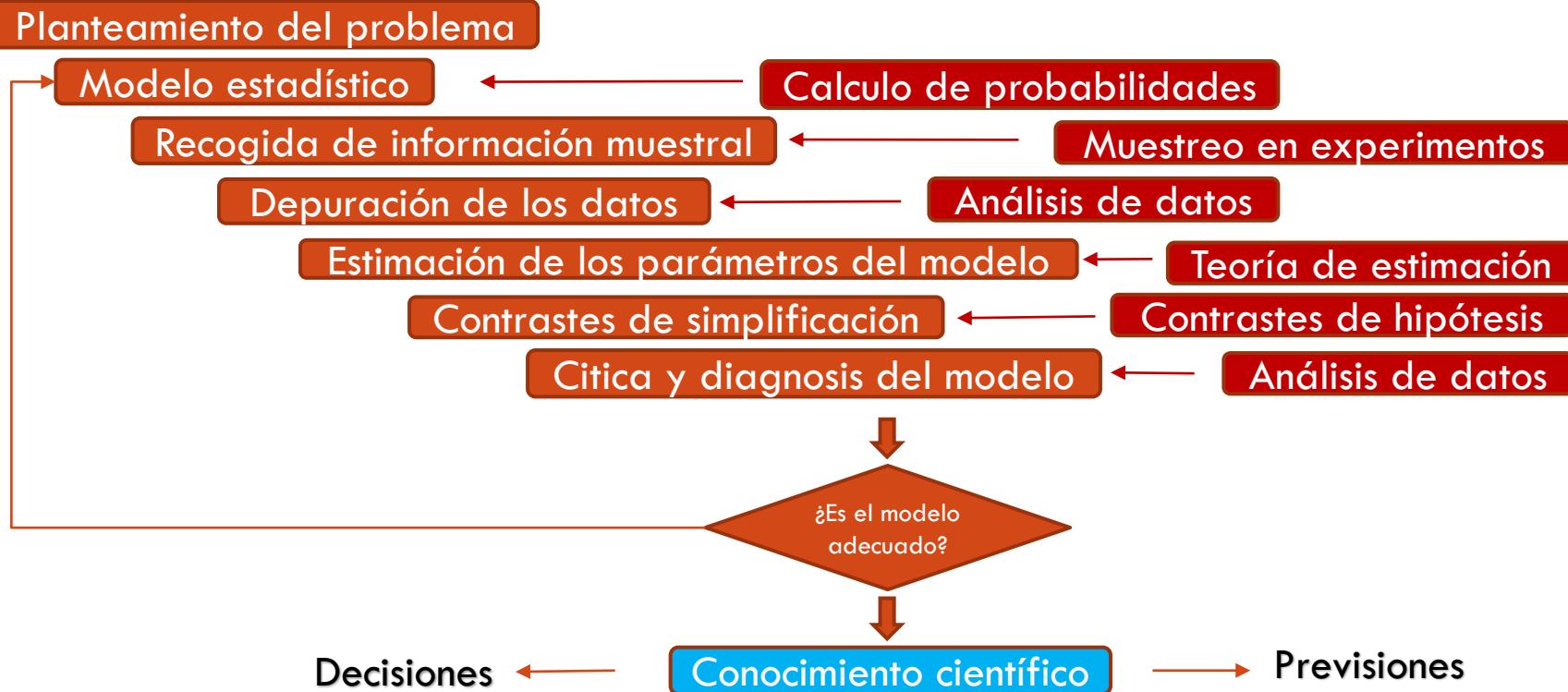
Parte de la foto sacada de  
[https://en.wikipedia.org/wiki/Linear\\_regression](https://en.wikipedia.org/wiki/Linear_regression)



# Introducción: Etapas para construir un Modelo estadístico

Construcción y refinamiento del modelo estadístico

## Problema Real



# Introducción: Metodología Estadística y Big Data

- El Big Data empieza a cambiar la metodología vista anteriormente de estadística:

- Según lo que hemos mostrado anteriormente, generalmente tenemos un problema real que queremos resolver, así empezamos a tomar datos mediante un muestreo de una población definida orientados a resolver ese problema. Es decir no tenemos los datos al principio.
- Sin embargo en Big Data, tenemos los datos al principio y empezamos a hacer un análisis exploratorio para aprender sus estructuras.

- Generalmente en estadística proponemos un modelo para hacer inferencia respecto unos parámetros del modelo que se ajustaron con la población.
- Con Big Data es muy difícil hacer esto, no sabemos como parametrizar en algunos casos.

- En estadística es muy importante como extraer de la muestra la máxima información (eficiencia estadística) que me pueda dar a conocer la población de la que proviene esa muestra.
- En Big Data no es tan importante la eficiencia estadística ya que en algunos casos tenemos casi la población.

# Introducción: Algunas Herramientas - Matlab

- Introducción
- Tipos de elementos
- Tipos de estructuras de datos
- Importar ficheros cvs
- Histogramas
- Medidas de centralidad y posición
- Medidas de forma
- Diagramas de caja (boxplot)
- Q-Qplots
- Diagramas de dispersión e histogramas

# Introducción (en Python pylab recuerda a Matlab)

- Matlab (abreviatura de MATrix LABoratory, "laboratorio de matrices") es un software matemático comercial, cuya función principal es análisis de datos:
  - Entorno de desarrollo integrado (IDE).
  - Lenguaje de programación propio (lenguaje M).
- ¿Qué puede hacer Matlab?
  - Manipulación de matrices
  - representación de datos y funciones
  - implementación de algoritmos
  - creación de interfaces de usuario (GUI)
  - Comunicación con programas en otros lenguajes y con otros dispositivos hardware.
- MATLAB dispone de dos herramientas adicionales:
  - Simulink (plataforma de simulación multidominio)
  - GUIDE (editor de interfaces de usuario - GUI)
  - Paquetes para ampliar las capacidades de Matlab:
    - Cajas de herramientas (toolboxes)
  - Paquetes para ampliar las capacidades de Simulink:
    - Paquetes de bloques (blocksets)
- Es un software muy usado en universidades, centros de investigación y desarrollo.

## Entorno pylab:

- **NumPy**
- **Matplotlib**
- **SciPy**
- **IPython**

## Equivalente Python

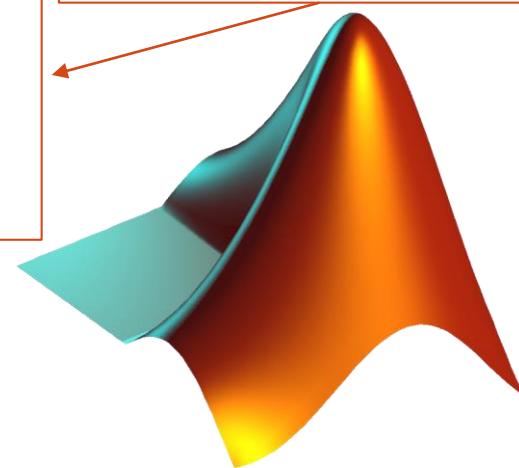


Imagen extraída de

<https://es.wikipedia.org/wiki/MATLAB>

# Introducción

- Por ejemplo, la Toolbox Statistics proporciona un conjunto completo de herramientas para evaluar e interpretar datos:
  - Organización y gestión de datos
  - Estadística descriptiva
  - Elaboración de gráficos estadísticos y visualización de datos
  - Distribuciones de probabilidades
  - Modelo lineal y no lineal
  - Contrastos de hipótesis
  - Estadística multivariante
  - Etc.

# Tipos de elementos (algunos ejemplos)

## ➤ Los tipos de elementos en Matlab para generar expresiones:

### ➤ Números

#### ➤ Decimales

#### ➤ Notación científica

#### ➤ Números imaginarios

d1 =

10

c3 =

0.3416

d2 =

2

d3 =

-15

i1 =

0.0000 + 3.0000i

d4 =

2.1000

i2 =

0.0000 + 3.0000i

d5 =

-2.5000

i3 =

0.0000 - 1.6000i

c1 =

4.1600

i4 =

2.0000e+00 + 2.1000e+04

c2 =

4.1600

% Números decimales

d1=10, d2=+2, d3=-15, d4=2.1, d5=-2.5

% Notación científica

c1=0.416e+1, c2=0.0000416e5, c3=4.16e-2

% Números imaginarios

i1=3i, i2=3j, i3=-1.6\*i, i4=2.1e4\*i

# Tipos de elementos (algunos ejemplos)

- Los tipos de elementos en Matlab para generar expresiones:
  - Variables: no requiere declaración del tipo de variable, ya que son consideradas como matrices de tamaño variable.
  - Por defecto se inicializan a 0.
  - El nombre de la variable siempre empieza por una letra seguida por letras, dígitos y “\_”, máximo 31 (son los que se leen).
  - Se distingue entre mayúsculas y minúsculas.
  - También se pueden usar como variables lógicas.

```
% Variables lógicas
if (0), disp('0 equivale a False'), end
if (1), disp('1 equivale a True'), end
a_1=21; if (a_1), disp('a_1 distinto 0 equivale a True'), end
if (~0), disp(~ 0 equivale a True'), end
```

Salida en la terminal:

1 equivale a True  
a\_1 distinto 0 equivale a True  
~ 0 equivale a True

# Tipos de elementos (algunos ejemplos)

- Los tipos de elementos en Matlab para generar expresiones:
  - Operadores
  - Operadores aritméticos: + , - , \*, / , ^

```
% Operadores aritméticos + , - , *, / , ^  
a = 10  
b = 10.1  
s=a + b          % Suma  
r=a - 3          % Resta  
p=b*3            % Multiplicación  
d=3.5/7          % División  
e=a^b            % Exponenciación
```

```
a =  
10  
b =  
10.1000  
s =  
20.1000  
r =  
7  
p =  
30.3000  
d =  
0.5000  
e =  
1.2589e+10
```

# Tipos de elementos (algunos ejemplos)

- Los tipos de elementos en Matlab para generar expresiones:
  - Funciones: realizan un conjunto de operaciones sobre los parámetros de las funciones. Las funciones están **vectorizadas**:

```
% Ayuda sobre funciones  
elementales  
help elfun
```

```
>> help elfun  
Elementary math functions.
```

Trigonometric.

sin	- Sine.
sind	- Sine of argument in degrees.
sinh	- Hyperbolic sine.
asin	- Inverse sine.
asind	- Inverse sine, result in degrees.
asinh	- Inverse hyperbolic sine.
cos	- Cosine.
cosd	- Cosine of argument in degrees.

.....

# Tipos de estructuras de datos (algunos ejemplos)

- Las estructuras fundamentales son las matrices:
  - Matrices de dimensión  $n \times 1$  (vectores), por ejemplo para organizar  $n$  observaciones de variables unidimensionales

% Vectores

```
vf1=[2 3 4 5 6] % vector fila (1,5)
vf2=[2, 3, 4, 5, 6] % vector fila (1,5)           vc =
vc=[2; 3; 4; 5; 6] % vector columna (5,1)
```

vf1 =	2
2    3    4    5    6	3
vf2 =	4
2    3    4    5    6	5
	6

# Tipos de estructuras de datos (algunos ejemplos)

- Las estructuras fundamentales son las matrices:
  - Matrices de dimensión  $n \times m$  (matrices), por ejemplo para organizar  $n$  observaciones multivariantes en filas y columnas

```
% Matrices
a=[2 3 4 5 6;2 3 4 5 6] % Matriz (2,5)
b=[1 2 3 4 5;1 2 3 4 5] % Matriz (2,5)
c1=[a;b] % Matriz (4,5)
c2=[a b] % Matriz (2,10)

a =
    2   3   4   5   6
    2   3   4   5   6
b =
    1   2   3   4   5
    1   2   3   4   5

c1 =
    2   3   4   5   6
    2   3   4   5   6
    1   2   3   4   5
    1   2   3   4   5

c2 =
    2   3   4   5   6   1   2   3   4   5
    2   3   4   5   6   1   2   3   4   5
```

- Cells para organizar datos heterogéneos con diferentes tamaños, unidades, etc,...

# Tipos de estructuras de datos (algunos ejemplos)

- La generación y manipulación de matrices se hace con el operador colon ":"

```
%Generación y manipulación de matrices
":"
a=1:10 %vector (1,10)
b=1:0.5:10 %vector (1,19)
c=[1:10; 11:20; 21:30] %matriz (3,10)
ct=c' %matriz (10,3)
a=[1 2 3; 4 5 6; 7 8 9] %matriz (3,3)
f1a=a(1,:) % fila 1 de la matriz a
f3a=a(3,:) % fila 3 de la matriz a
c1a=a(:,1) % columna 1 de la matriz a
c2a=a(:,2) % columna 2 de la matriz a
v=[] % matriz vacía
a=[1 2 3; 4 5 6; 7 8 9] %matriz (3,3)
r1=a(1:3,2:3) %matriz (3,2)
r2=a(1:3,[1 3]) %matriz (3,2)
```

a =

1 2 3 4 5 6 7 8 9 10

b =

Columns 1 through 12

1.0000	1.5000	2.0000	2.5000	3.0000
3.5000	4.0000	4.5000	5.0000	5.5000
6.0000	6.5000			

Columns 13 through 19

7.0000	7.5000	8.0000	8.5000	9.0000
9.5000	10.0000			

# Tipos de estructuras de datos (algunos ejemplos)

- La generación y manipulación de matrices se hace con el operador colon ":"

```
%Generación y manipulación de matrices
":"
a=1:10 %vector (1,10)
b=1:0.5:10 %vector (1,19)
c=[1:10; 11:20; 21:30] %matriz (3,10)
ct=c' %matriz (10,3)
a=[1 2 3; 4 5 6; 7 8 9] %matriz (3,3)
f1a=a(1,:) % fila 1 de la matriz a
f3a=a(3,:) % fila 3 de la matriz a
c1a=a(:,1) % columna 1 de la matriz a
c2a=a(:,2) % columna 2 de la matriz a
v=[] % matriz vacía
a=[1 2 3; 4 5 6; 7 8 9] %matriz (3,3)
r1=a(1:3,2:3) %matriz (3,2)
r2=a(1:3,[1 3]) %matriz (3,2)
```

ct =	1 11 21	c2a =	2
	2 12 22		5
	3 13 23		8
	4 14 24	v =	[]
	5 15 25	a =	1 2 3
	6 16 26		4 5 6
	7 17 27	r1 =	7 8 9
	8 18 28		2 3
	9 19 29	f1a =	5 6
	10 20 30		8 9
a =	1 2 3	r2 =	1 3
	4 5 6		4 6
	7 8 9	c1a =	7 9
f1a =	1 2 3		
	7 8 9		
f3a =	1 2 3		
	7 8 9		
c1a =	1		
	4		
	7		

# Importar ficheros cvs

- Para leer un fichero cvs, se hace mediante la función `cvsread()`

(spamND.csv es un fichero que contiene una serie de columnas de diferentes datos, mirar siguiente transparencia)

```
%Importando datos en cvs  
data = csvread('spamND.csv', 1);
```

- El parámetro 1 sirve para eliminar la fila de etiquetas.
- Podemos eliminar columnas también, por ejemplo las 4 primera columnas:
  - `data = csvread('spamND.csv', 1, 4);`



## Normal text file

length : 1.107.320 lines :

3 Ln : 1 Col : 741 Sel : 0

U

INS

# Histogramas

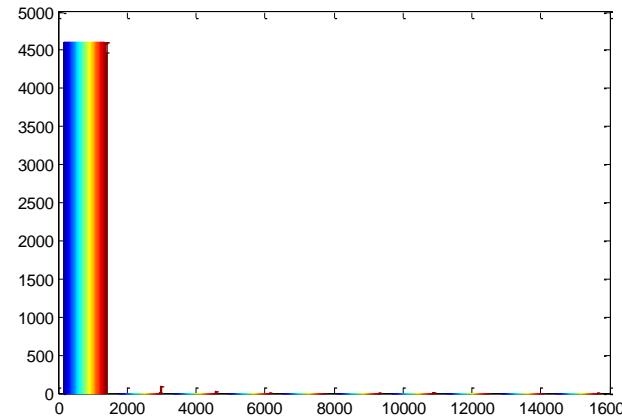
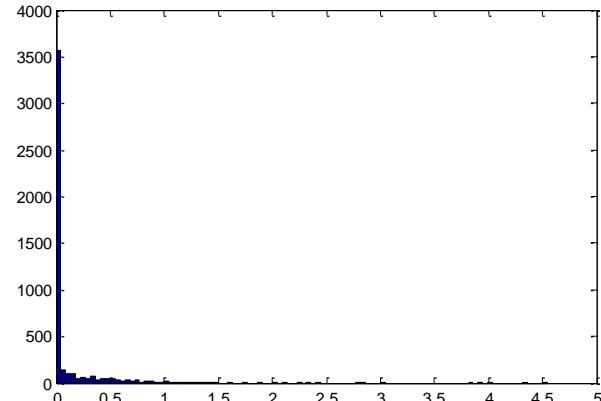
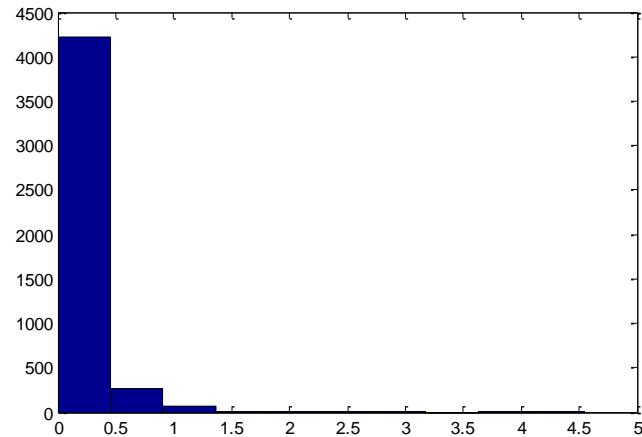
## ➤ Haciendo histogramas por defecto

```
%Calculando histogramas  
size(data) %filas columnas  
hist(data(:,1)) % histograma de la columna 1  
hist(data(:,1),10) % igual que el anterior  
hist(data(:,1),100) % con 100 bins  
[N,X] = hist(data(:,1),10) % no pinta nada  
hist(data) % pinta todos a la vez
```

```
ans =  
100 16  
N =  
4 9 10 15 22 17 13 7 2 1  
X =  
-1.9690 -1.4932 -1.0174 -0.5416 -0.0658 0.4100 0.8857 1.3615 1.8373 2.3131
```

# Histogramas

- Haciendo histogramas por defecto

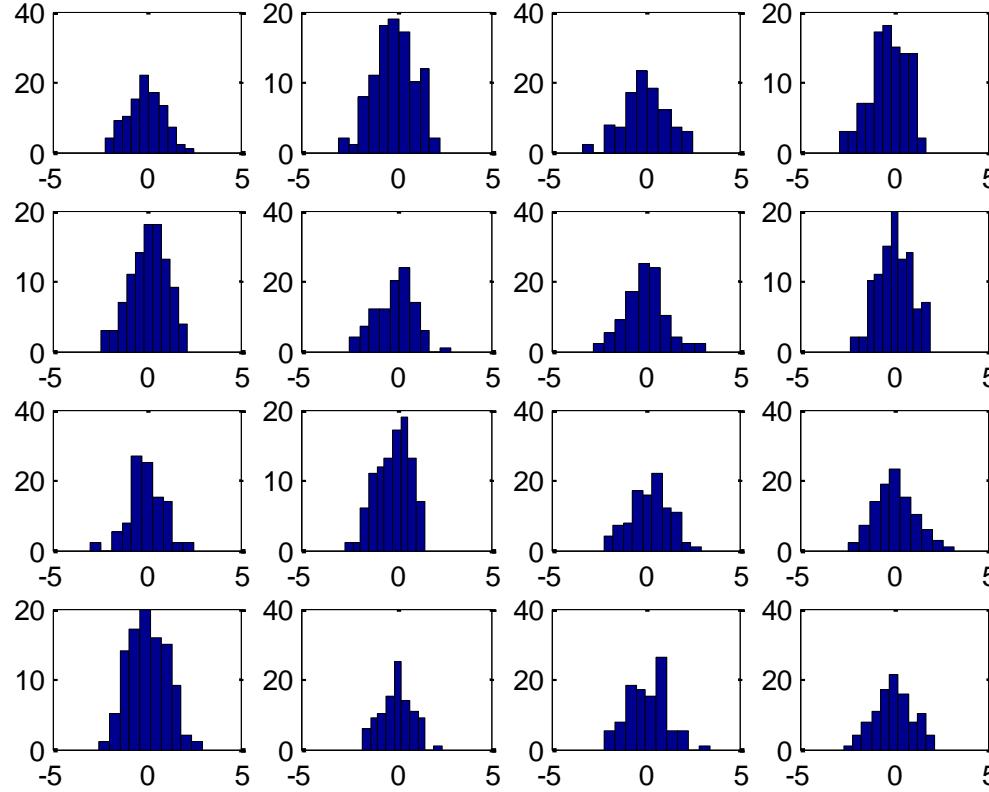


# Histogramas

- Pintando varios histogramas en una matriz:

```
%Pintando varios histogramas en una matriz  
data=randn(100,16); % matriz de datos  
aleatorios gaussianos (100,16)  
for(i=1:16)  
    subplot(4,4,i) % matriz de 4x4 subfiguras  
    hist(data(:,i))  
end
```

# Matriz de Histogramas



# Medidas de centralidad y posición

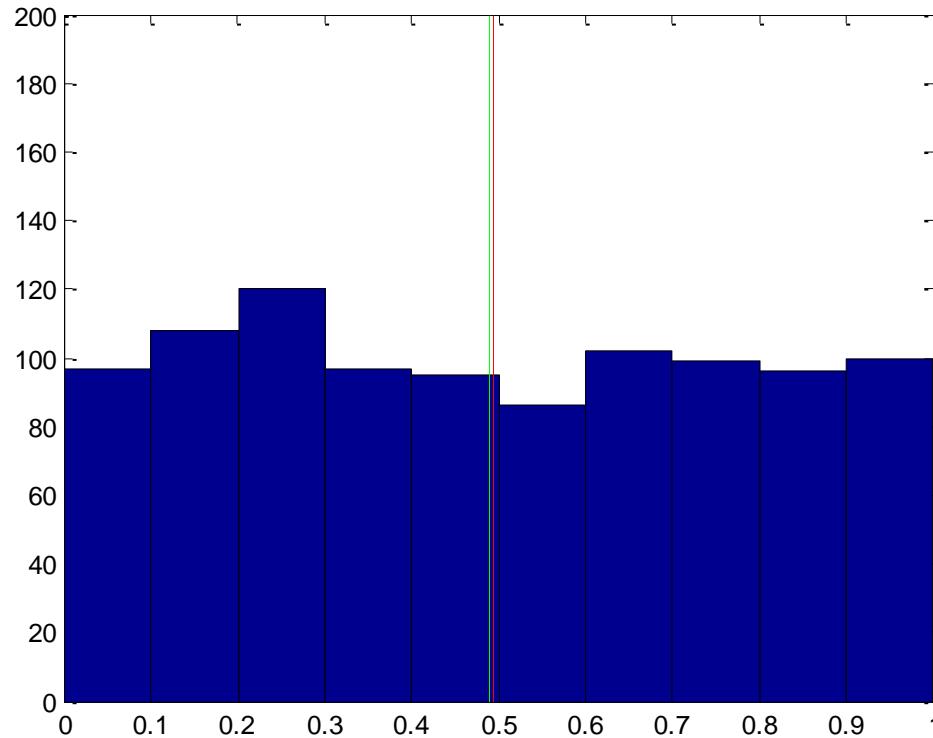
- Para el calculo la media aritmética podemos utilizar `mean()`
- Para el calculo de la mediana podemos utilizar `median ()`
- Para el calculo de percentiles podemos utilizar `prctile()`
  - `prctile(x,p)` devuelve un valor que sería mayor que el p% de los valores del vector x.
  - Así la mediana es el percentil 50% `median(x)=prctile(x,50)`
  - Los cuartiles se calculan como `prctile(x,[25,50,75])`

# Medidas de centralidad y posición

## ➤ Distribución uniforme de datos

```
%Medidas de centralidad y posición de datos uniformes  
x_uniforme=rand(1000,1); % 1000 datos distribuidos  
uniformemente entre 0 y 1  
m=mean(x_uniforme)  
med=median(x_uniforme)  
hist(x_uniforme)  
hold on  
plot([m m],[0 200],'r')  
plot([med med],[0 200],'g')  
hold off  
pcr=prctile(x_uniforme,50)  
q=prctile(x_uniforme,[25,50,75])
```

# Medidas de centralidad y posición



$$m = 0.4932$$

$$\text{med} = 0.4892$$

$$\text{pcr} = 0.4892$$

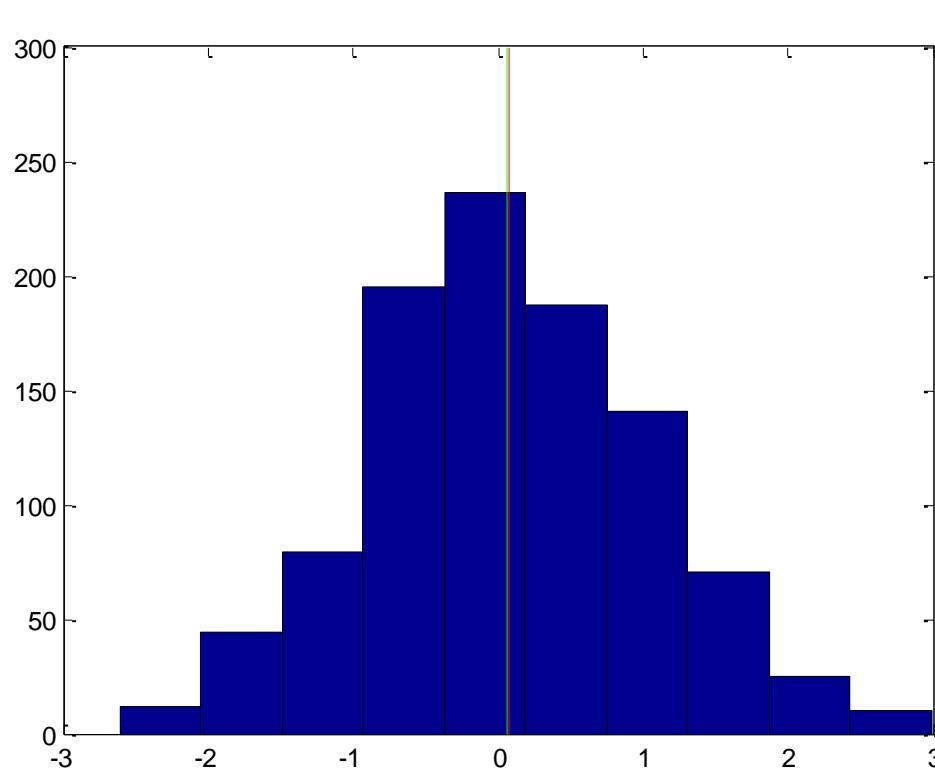
$$q = 0.2384 \quad 0.4892 \quad 0.7532$$

# Medidas de centralidad y posición

## ➤ Distribución gaussiana de datos

```
%Medidas de centralidad y posición de datos distribuidos
según una gaussiana
x_g=randn(1000,1); % 1000 datos distribuidos
uniformemente entre 0 y 1
m=mean(x_g)
med=median(x_g)
hist(x_g)
hold on
plot([m m],[0 300],'r')
plot([med med],[0 300],'g')
hold off
pcr=prctile(x_g,50)
q=prctile(x_g,[25,50,75])
```

# Medidas de centralidad y posición



$m =$

0.0745

$med =$

0.0621

$pcr =$

0.0621

$q =$

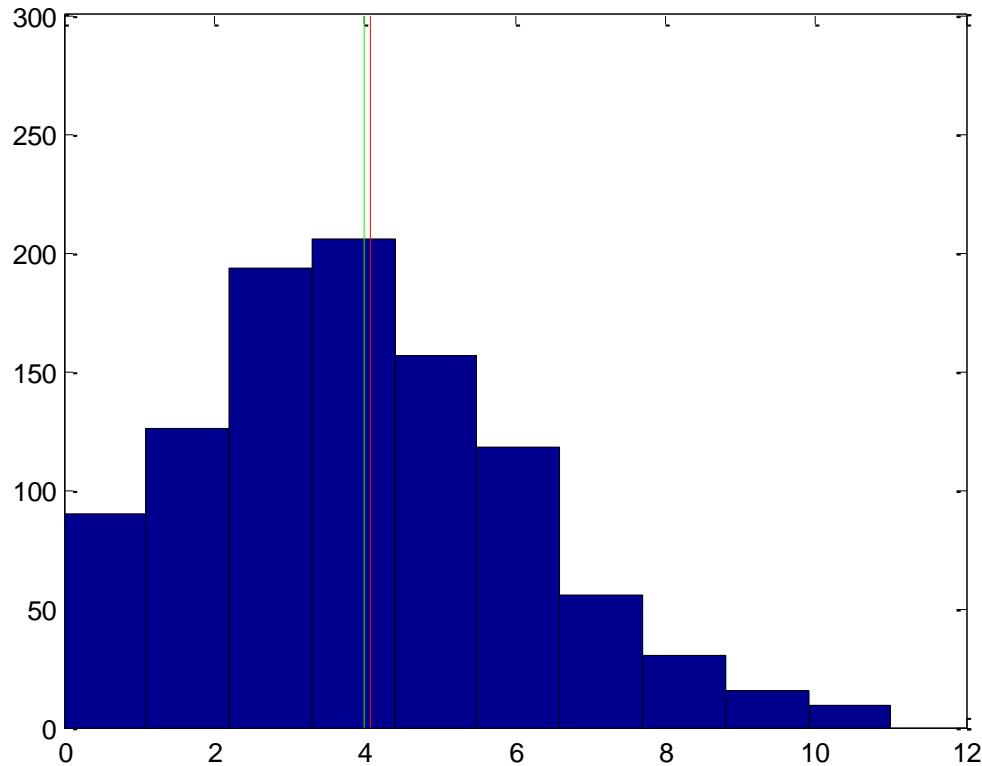
-0.5949 0.0621 0.7452

# Medidas de centralidad y posición

## ➤ Distribución de Poisson para los datos

```
%Medidas de centralidad y posición de datos  
distribuidos según una poisson  
x_p=poissrnd(4,1000,1); % 1000 datos distribuidos  
seún una poisson de lambda 4  
m=mean(x_p)  
med=median(x_p)  
hist(x_p)  
hold on  
plot([m m],[0 300],'r')  
plot([med med],[0 300],'g')  
hold off  
pcr=prctile(x_p,50)  
q=prctile(x_p,[25,50,75])
```

# Medidas de centralidad y posición



$m =$   
4.0800

$med =$   
4

$pcr =$   
4

$q =$   
3    4    5

# Medidas de dispersión

- Rango
  - `range(x)`
  - `max(x)-min(x)`
- Rango intercuartílico
  - `iqr(x)`
  - `prctile(x,75)-prctile(x,25)`
- Varianza
  - `sum((x-mean(x)).^2))/length(x)`
  - `var(x)` %normaliza por n - 1
  - `var(x,1)` %normaliza por n
- Desviación típica
  - `sqrt(sum((x-mean(x)).^2))/ length(x)`
  - `sqrt(var(x))` %normaliza por n - 1
  - `std(x)` %normaliza por n - 1
  - `sqrt(var(x,1))` %normaliza por n
  - `std(x,1)` %normaliza por n

# Medidas de dispersión

```
%Medidas de dispersión de datos distribuidos según  
una gaussiana  
x_g=randn(1000,1); % 1000 datos distribuidos según  
una gaussiana  
r1=range(x_g) %Rango  
r2=max(x_g)-min(x_g)  
i1=iqr(x_g) %Rango intercuartílico  
i2=prctile(x_g,75)-prctile(x_g,25)  
v1=sum(((x_g-mean(x_g)).^2))/length(x_g) %Varianza  
v2=var(x_g) % normaliza por n-1  
v3=var(x_g,1) % normaliza por n  
s1=sqrt(sum((x_g-mean(x_g)).^2)/length(x_g))  
%Desviación típica  
s2=sqrt(var(x_g,1)) % normaliza por n  
s3=std(x_g,1) % normaliza por n  
s4=sqrt(var(x_g,1))% normaliza por n  
s5=std(x_g,1) % normaliza por n
```

$$r1 = 6.8901$$

$$r2 = 6.8901 \quad s1 = 1.0108$$

$$i1 = 1.4083 \quad s2 = 1.0108$$

$$i2 = 1.4083 \quad s3 = 1.0108$$

$$v1 = 1.0217 \quad s4 = 1.0108$$

$$v2 = 1.0227 \quad s5 = 1.0108$$

$$v3 = 1.0217$$

# Medidas de forma

- **Coeficiente de asimetría:**
  - $CA = \sum(x_i - \bar{x})^3 / ns^3$
  - $\text{skewness}(x)$
- **Coeficientes de apuntamiento o curtosis:**
  - $CA_p = \sum(x_i - \bar{x})^4 / ns^4$
  - $\text{kurtosis}(x)$

# Medidas de forma

```
%Medidas de forma  
x_u=rand(1000,1); % 1000 datos  
distribuidos uniformemente entre 0 y 1  
figure(1); hist(x_u)  
x_g=randn(1000,1); % 1000 datos  
distribuidos según una gaussiana  
figure(2); hist(x_g)  
x_p=poissrnd(4,1000,1); % 1000 datos  
distribuidos según una Poisson de lambda  
4  
figure(3); hist(x_p)  
CA_u=skewness(x_u), CA_g=skewness(x_g),  
CA_p=skewness(x_p) % Coeficientes de  
asimetría  
CAp_u=kurtosis(x_u), CAp_g=kurtosis(x_g),  
CAp_p=kurtosis(x_p) % Coeficientes de  
curtosis
```

$$CA_u = 0.0281$$

$$CA_g = -0.0526$$

$$CA_p = 0.6886$$

$$CAp_u = 1.8104$$

$$CAp_g = 2.9335$$

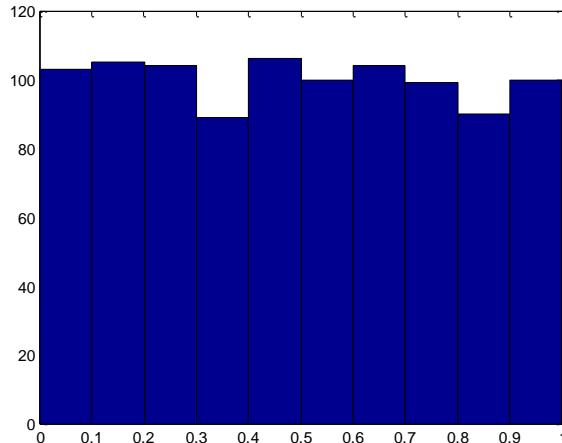
$$CAp_p = 3.8057$$

# Medidas de forma

CA\_u =  
0.0281

CAp\_u =  
1.8104

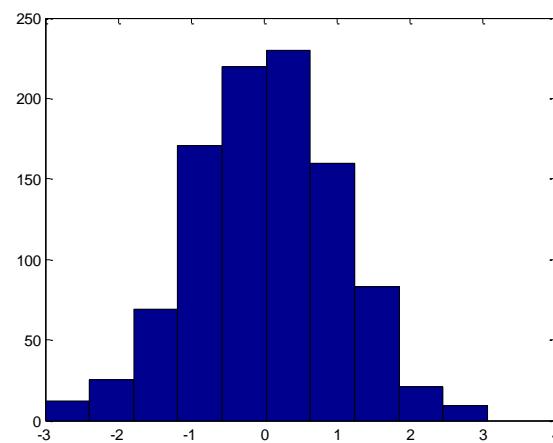
figure(1); hist(x\_u)



CA\_g =  
-0.0526

CAp\_g =  
2.9335

figure(2); hist(x\_g)

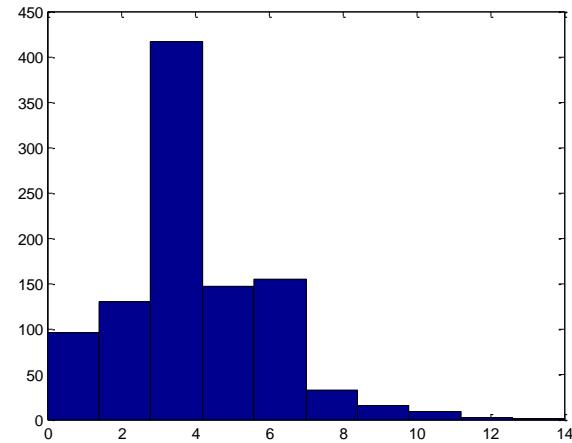


CA\_p =  
0.6886

(positiva queda más masa a la izquierda de la media)

CAp\_p =  
3.8057

figure(3); hist(x\_p)



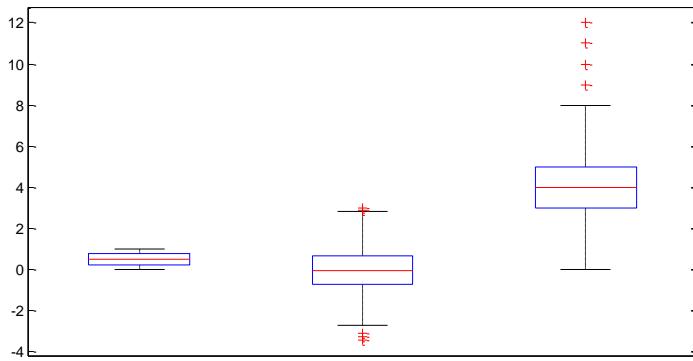
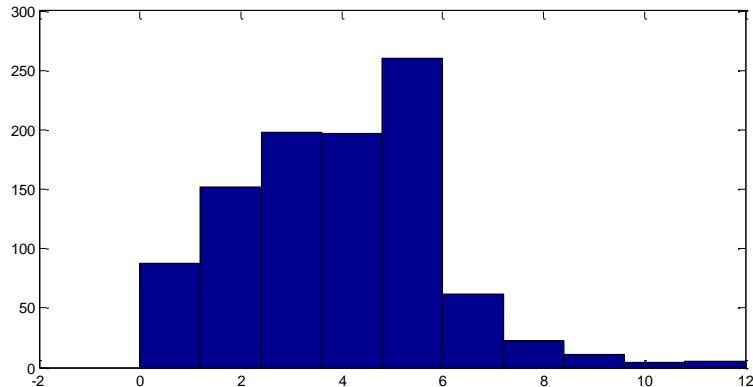
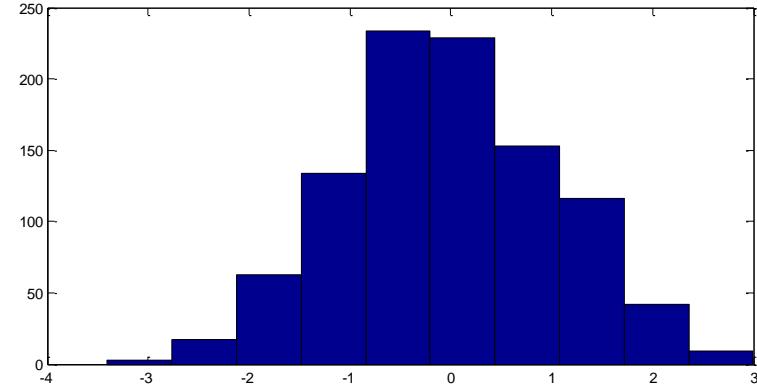
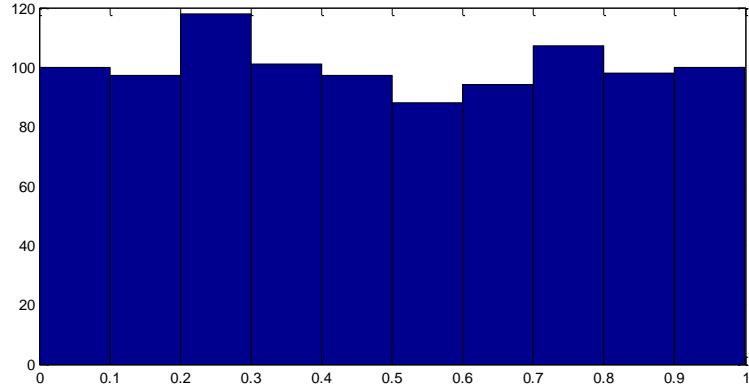
# Diagramas de caja (boxplot)

- La información obtenida a partir de las medidas de centralización, dispersión y forma se pueden resumir en un único diagrama: boxplots.
- El diagrama de caja consta de una caja central que está delimitada por la posición de los cuartiles Q3 y Q1. La caja está dividida por la mediana de los datos.
- Se suele considerar atípico un valor fuera de 1.5 veces el rango intercuartílico.
- Así de los extremos de la caja salen unas líneas (“bigotes”) que se extienden hasta los puntos inferior y superior:
  - $LI = \max\{ \min(x_i), Q1 - 1.5 \text{ RIC} \}$
  - $LS = \min\{ \max(x_i), Q3 + 1.5 \text{ RIC} \}$
- Los “bigotes” son las líneas que van desde cada extremo del rectángulo central hasta el valor más alejado no atípico, es decir, que está dentro del intervalo (LI, LS).
- Los datos fuera del intervalo (LI, LS) se consideran atípicos y se pintan en el gráfico.
- La función de Matlab es `boxplot(x)`.

# Diagramas de caja (boxplot)

```
%Boxplots
x_u=rand(1000,1); % 1000 datos distribuidos
uniformemente entre 0 y 1
subplot(2,2,1); hist(x_u)
x_g=randn(1000,1); % 1000 datos distribuidos según
una gaussiana
subplot(2,2,2);hist(x_g)
x_p=poissrnd(4,1000,1); % 1000 datos distribuidos
según una Poisson de lambda 4
subplot(2,2,3);hist(x_p)
subplot(2,2,4);boxplot([x_u x_g x_p])
```

# Diagramas de caja (boxplot)



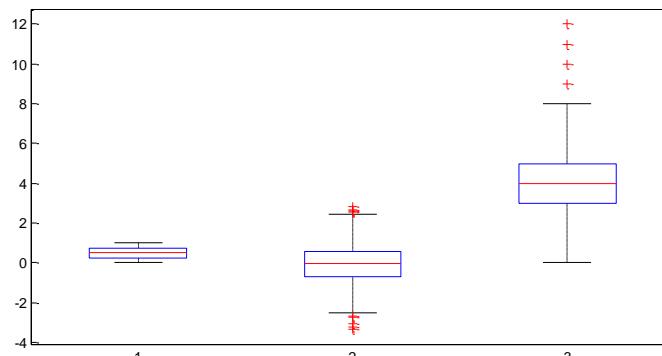
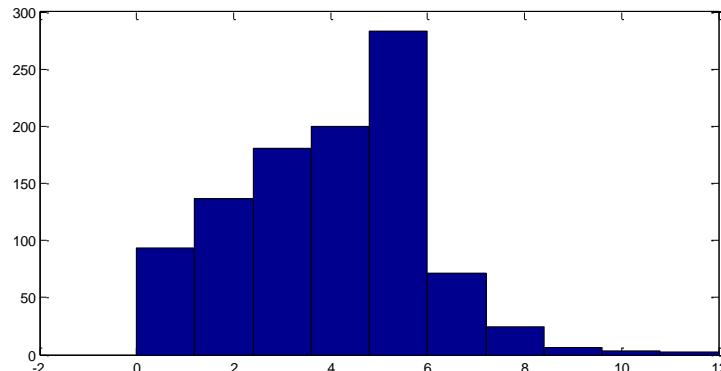
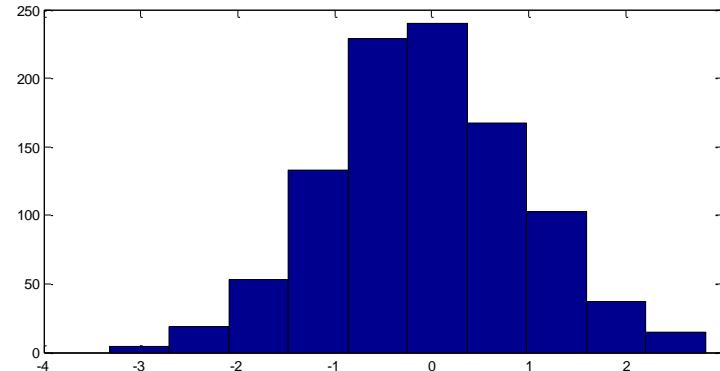
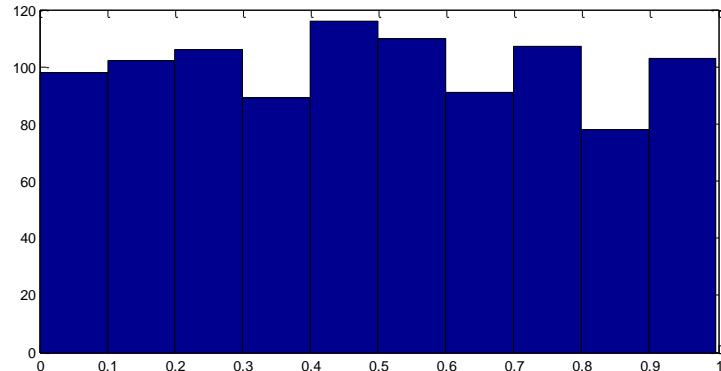
# Visualización datos con medidas de posicionamiento: Gráficos Q-Q

- Los gráficos Q-Q (Q-Q plots) se caracterizan por visualizar de una manera muy rápida y sencilla como se diferencian los datos de dos distribuciones de observaciones.
- Se basan en representar enfrentados en un gráfico x-y los cuantiles de ambas distribuciones. El “Q” viene de cuantil en inglés.
- Si todos los cuantiles son iguales aparecerá la recta  $x=y$  en el gráfico, y significará los dos conjuntos de datos se distribuyen de manera idéntica.
- Generalmente una de la distribuciones es conocida (por ejemplo una normal en Matlab por defecto), para contrastar si los datos observados se ajustan a la distribución conocida.
- La función en Matlab es `qqplot()`.

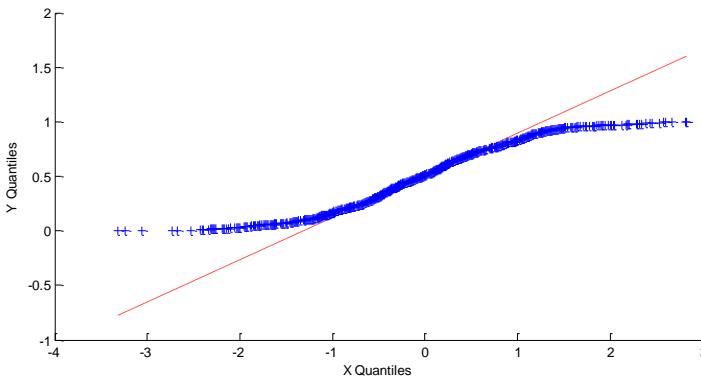
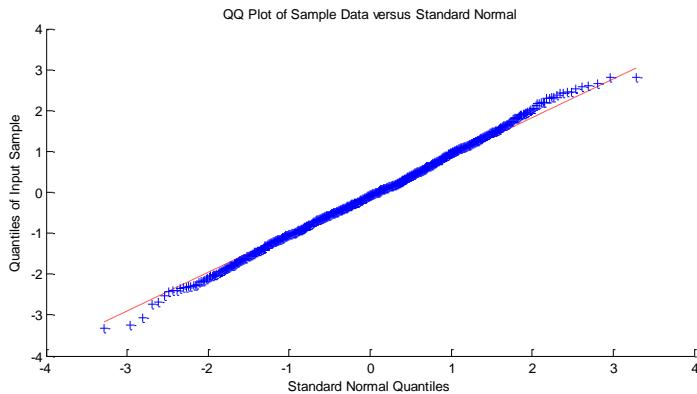
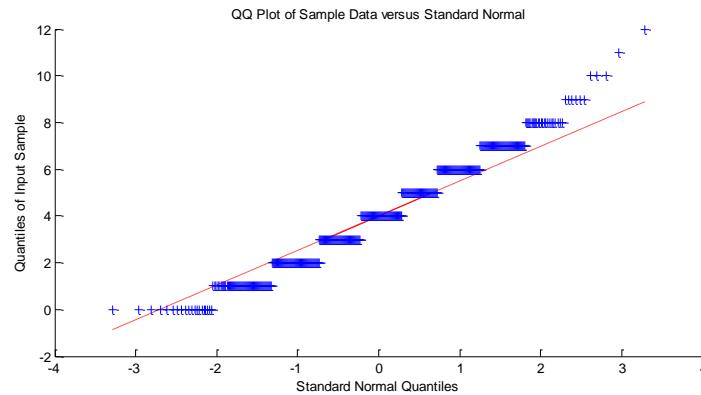
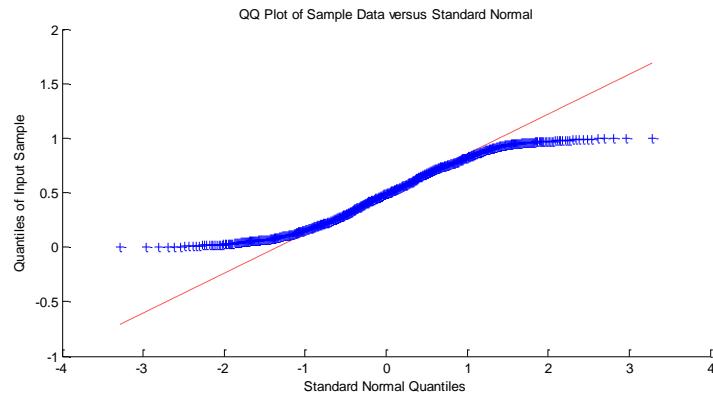
# Visualización datos con medidas de posicionamiento: Gráficos Q-Q

```
%QQplots
x_u=rand(1000,1); % 1000 datos distribuidos uniformemente entre
0 y 1
subplot(2,2,1); hist(x_u)
x_g=randn(1000,1); % 1000 datos distribuidos según una
gaussiana
subplot(2,2,2);hist(x_g)
x_p=poissrnd(4,1000,1); % 1000 datos distribuidos según una
Poisson de lambda 4
subplot(2,2,3);hist(x_p)
subplot(2,2,4);boxplot([x_u x_g x_p])
figure(2)
subplot(2,2,1); qqplot(x_u)
subplot(2,2,2);qqplot(x_p)
subplot(2,2,3);qqplot(x_g)
subplot(2,2,4);qqplot(x_g,x_u)
```

# Visualización datos con medidas de posicionamiento: Gráficos Q-Q



# Visualización datos con medidas de posicionamiento: Gráficos Q-Q



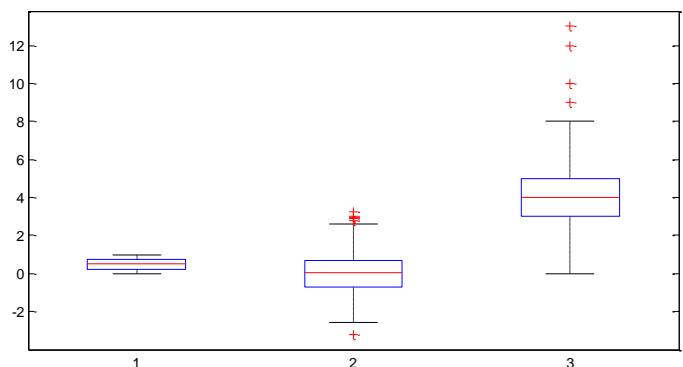
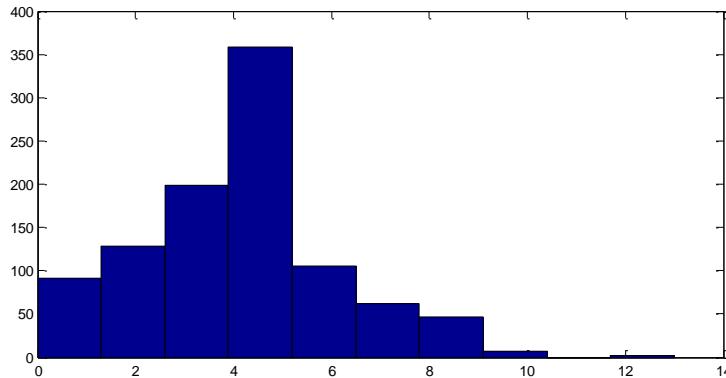
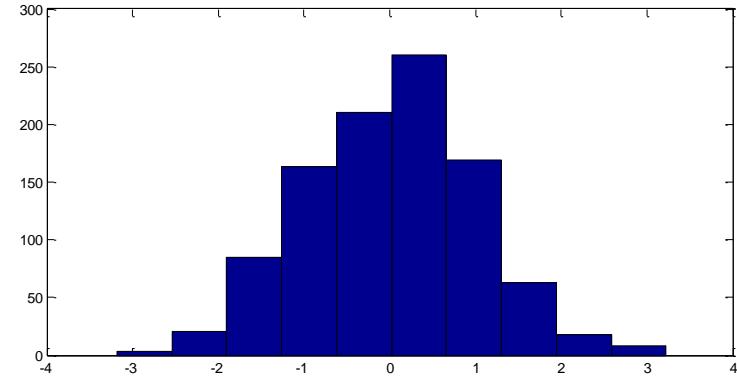
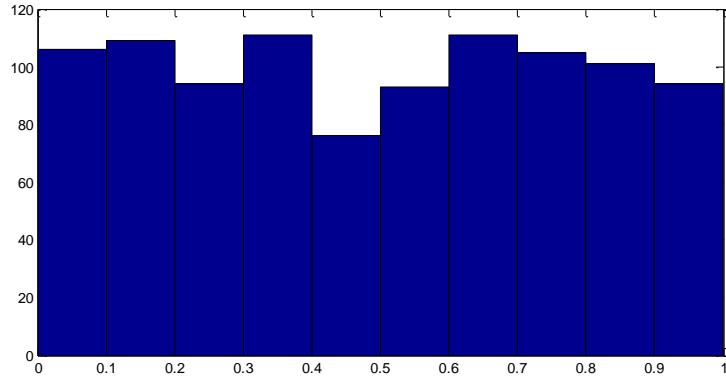
# Diagramas de dispersión e histogramas

- Se suelen combinar en estadística multivariante los diagramas de dispersión de las diferentes variables junto con los histogramas de las variables individuales.
- Un diagrama de dispersión es la representación de las observaciones de dos o tres variables enfrentadas.
- Existen varios tipos en matlab:
  - `scatterhist()`
  - `gplotmatrix()`

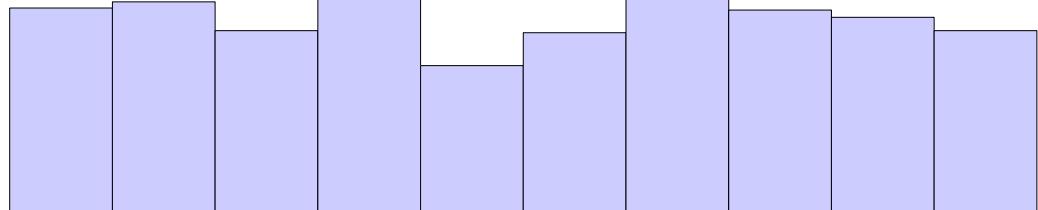
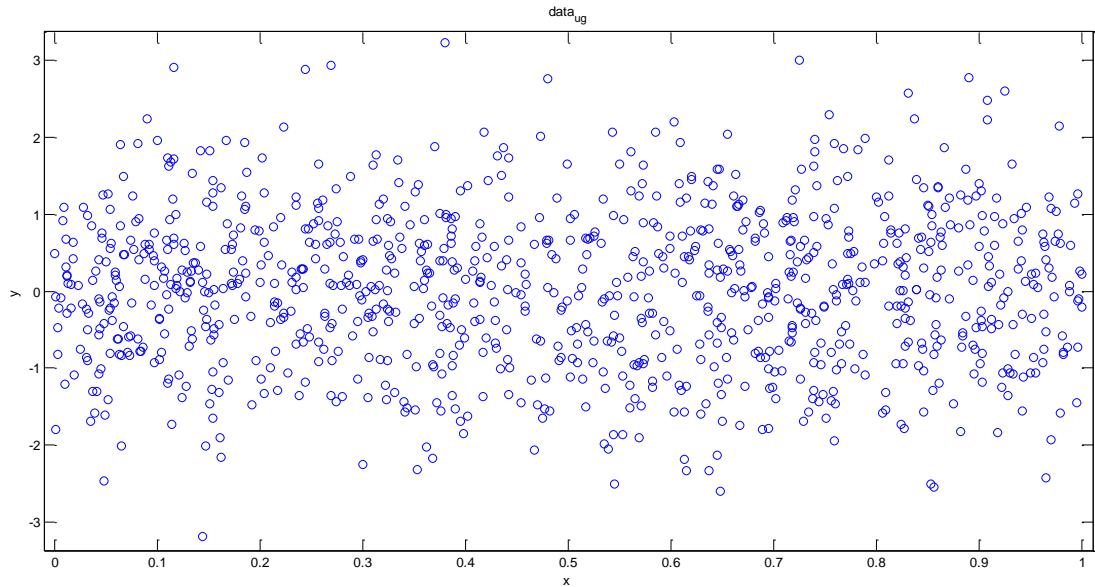
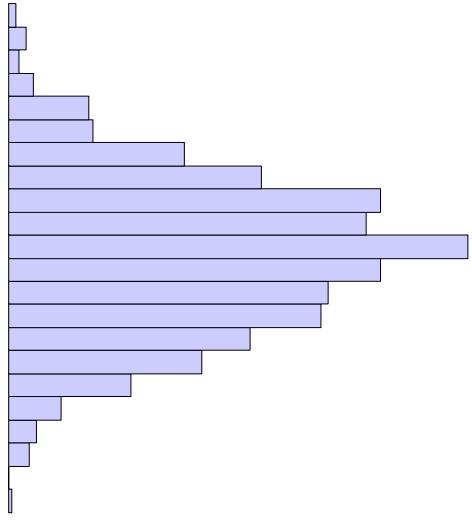
# Diagramas de dispersión e histogramas

```
%Diagramas de dispersión e histogramas
figure(1);
x_u=rand(1000,1); % 1000 datos distribuidos uniformemente entre 0 y 1
subplot(2,2,1); hist(x_u)
x_g=randn(1000,1); % 1000 datos distribuidos según una gaussiana
subplot(2,2,2);hist(x_g)
x_p=poissrnd(4,1000,1); % 1000 datos distribuidos según una Poisson de
lambda 4
subplot(2,2,3);hist(x_p)
subplot(2,2,4);boxplot([x_u x_g x_p])
data_ug=[x_u x_g];
data_gp=[x_g x_p];
data_pu=[x_p x_u];
figure(2); scatterhist(data_ug(:,1),data_ug(:,2)); title('data_{ug}')
figure(3); scatterhist(data_gp(:,1),data_gp(:,2)); title('data_{gp}')
figure(4); scatterhist(data_pu(:,1),data_pu(:,2)); title('data_{pu}')
data=[x_u x_g x_p];
figure(5); gplotmatrix(data);
```

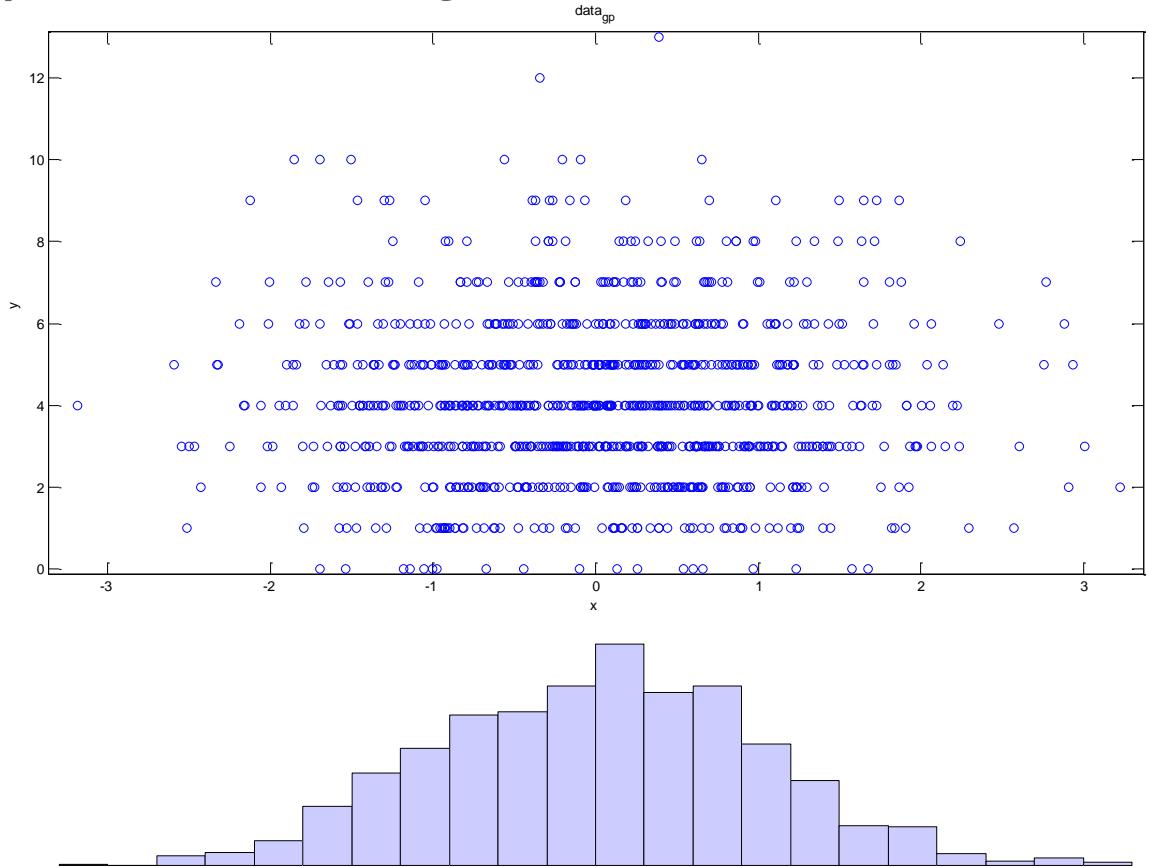
# Diagramas de dispersión e histogramas



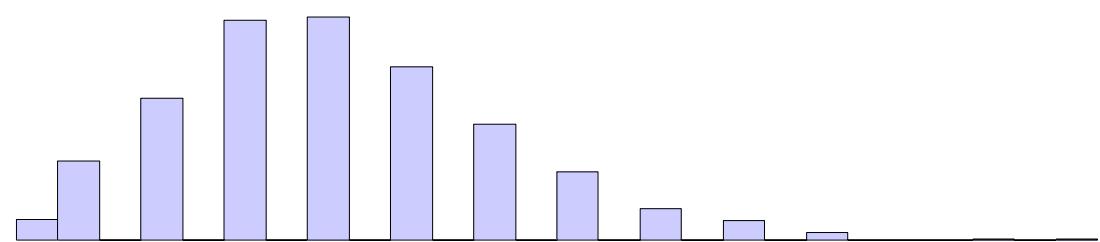
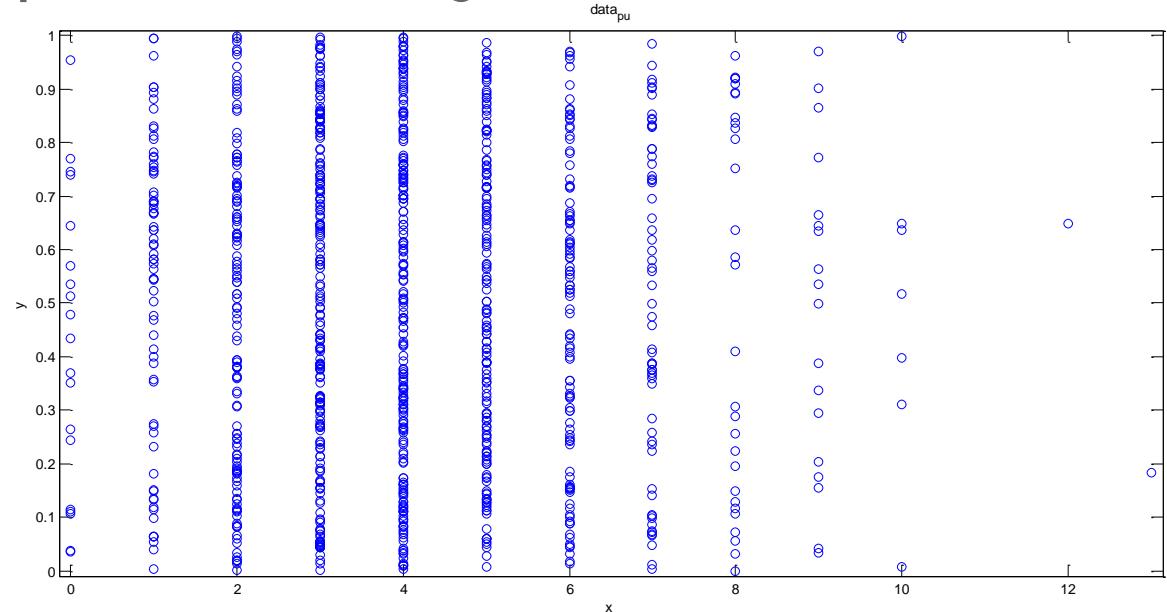
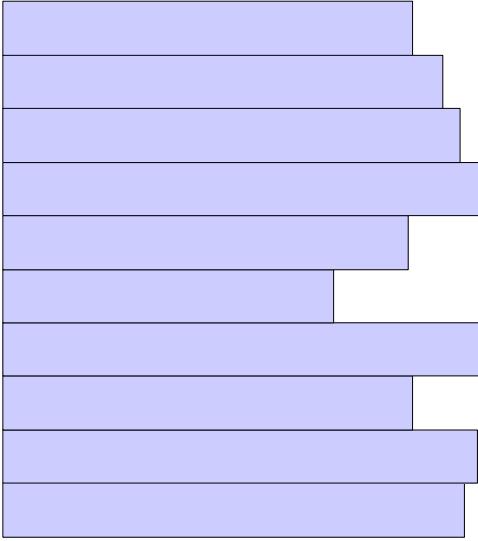
# Diagramas de dispersión e histogramas



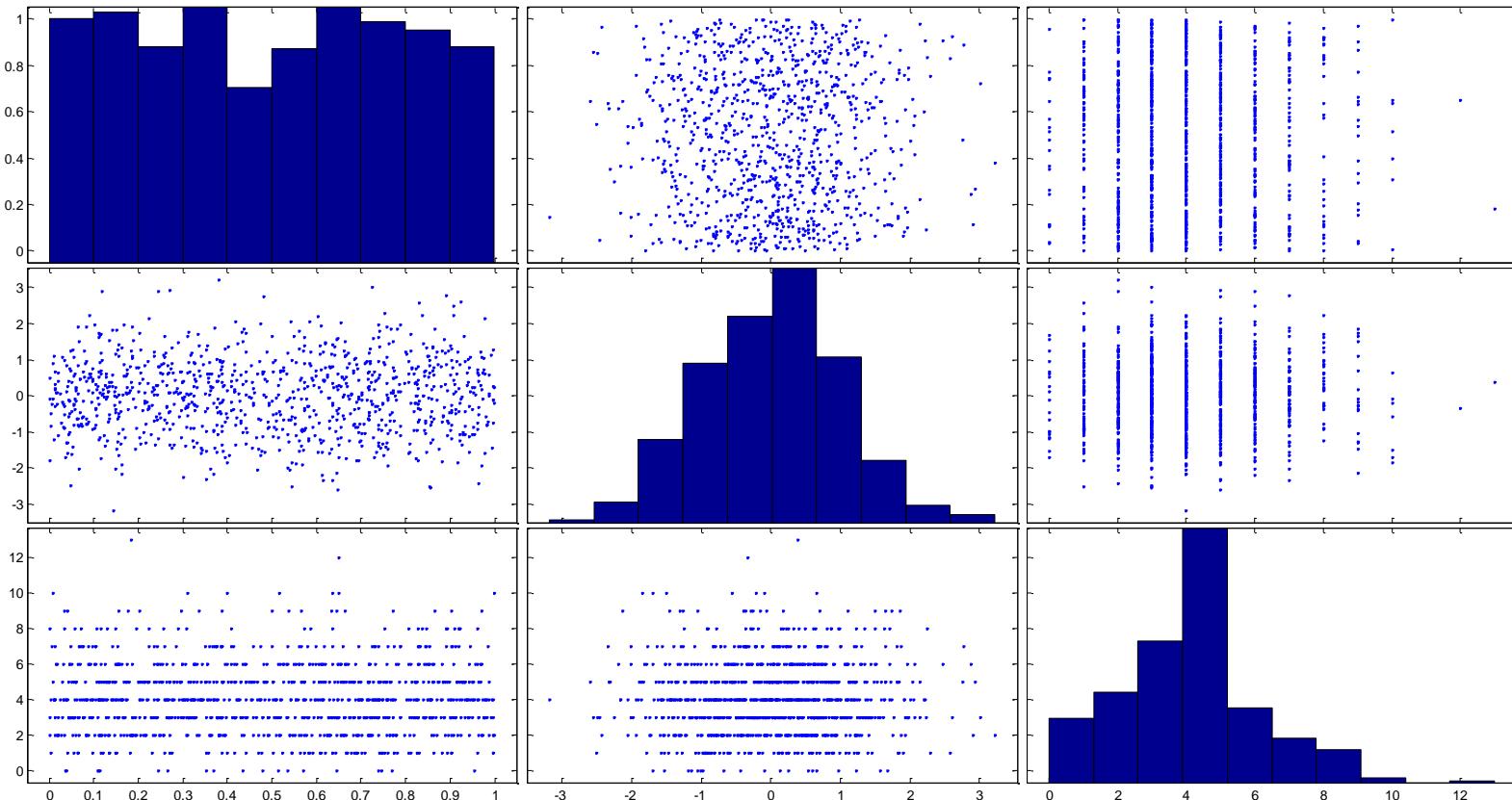
# Diagramas de dispersión e histogramas



# Diagramas de dispersión e histogramas



# Diagramas de dispersión e histogramas



# Bibliografía y lecturas relacionadas:

- <http://es.mathworks.com/help/>
- <http://es.mathworks.com/products/statistics/>
- [http://www.pi.ingv.it/~longo/CORSO Matlab/OriginalManuals/  
stats.pdf](http://www.pi.ingv.it/~longo/CORSO Matlab/OriginalManuals/stats.pdf)
- Fundamentos de estadística. Daniel Peña Sánchez Ribera.  
Alianza Editorial, 2001 o 2008.
- Computational statistics handbook with MATLAB. Martinez,  
Wendy L. Chapman & Hall/CRC, 2008.
- Statistics in MATLAB A Primer. Cho, MoonJung. Chapman and  
Hall/CRC, 2014.

# Introducción: Algunas Herramientas - SPSS

- Introducción
- Las cuatro ventanas
- Análisis Básico
- Frecuencias
- Descriptivos
- Análisis de regresión lineal
- Varias representaciones Visuales

# Introducción

- Originalmente es un acrónimo del Paquete Estadístico para las Ciencias Sociales (Statistical Package for the Social Science), Sin embargo, en la actualidad la parte SPSS del nombre completo del software (IBM SPSS) no es acrónimo de nada.
- Es uno de los paquetes estadísticos hoy en día más populares que puede realizar la manipulación de datos de gran complejidad y un análisis de los mismos con instrucciones muy simples.



Imagen extraída de  
<https://es.wikipedia.org/wiki/SPSS>

# Las cuatro ventanas

- En SPSS puedes manejar lo básico con cuatro ventanas.
- Las cuatro ventanas:
  - Editor de datos
  - Una ventana de salida de resultados
  - Un editor de sintaxis
  - Y una ventana para ejecutar secuencias de comandos (ventana de Script)
- Estas cuatro ventanas pueden ser invocadas desde
  - Si son nuevas: archivo>nuevo y aparecen las cuatro opciones.
  - Si ya existen como fichero: archivo>abrir y aparecen las cuatro opciones

# Las cuatro ventanas: Editor de datos

- Editor de datos: Sistema de hoja de cálculo para definir, introducir, editar y visualizar datos. La extensión del archivo guardado será "SAV", esto es lo que se llama archivo muestral.
- Vamos a utilizar demo.sav: es un estudio ficticio de varios miles de personas que contiene información básica demográfica y de consumo (viene en el directorio de "samples" de SPSS con otros muchos ejemplos).
- demo.sav es un fichero binario, y para abrirlo archivo>abrir>datos....
- Descripción de todos los archivos muestrales de SPSS

# Las cuatro ventanas: Editor de datos

- La ventana por defecto tendrá el editor de datos
- Hay dos hojas en la ventana:
  - Vista de datos
  - Vista de variables
- La ventana Vista de datos es visible cuando se abre por primera vez el Editor de datos
  - Esta hoja contiene los datos que se han cargado
- Se pueden marcar datos con el ratón y hacer análisis estadísticos.

# Las cuatro ventanas: Editor de datos

The screenshot shows the IBM SPSS Statistics Editor de datos window with the following details:

- Title Bar:** demo.sav [Conjunto\_de\_datos2] - IBM SPSS Statistics Editor de datos
- Menu Bar:** Archivo, Editar, Ver, Datos, Transformar, Analizar, Gráficos, Utilidades, Ventana, Ayuda.
- Toolbar:** Includes icons for opening files, saving, printing, transforming, analyzing, and more.
- Data View:** A large grid showing 37 rows of data. The columns are labeled: edad, mantal, direcc, ingreso, ingcat, coche, cocheat, educ, empleo, retirado, empcat, satlab, genero, residen, inalam, multiline, voz, busca, internet, idllam, espera, and tv. The first row is highlighted in yellow.
- Status Bar:** Visible: 29 de 29 variables
- Bottom Navigation:** Buttons for Vista de datos (selected) and Vista de variables.
- System Taskbar:** Shows icons for File Explorer, Google Chrome, Microsoft Word, and Microsoft Excel. Status bar: IBM SPSS Statistics Processor está listo, Unicode:ON, 13:50, 30/09/2015.

# Las cuatro ventanas: Editor de datos

The screenshot shows the SPSS Data View window. The data consists of 9 rows (labeled 29 to 37) and 7 columns. The columns represent numerical values ranging from 0 to 272,00. Below the table, the status bar shows the number '1'. At the bottom, there are two tabs: 'Vista de datos' (highlighted in yellow) and 'Vista de variables' (highlighted with a red oval). A toolbar with various icons is also visible.

	29	52	0	20	272,00	4,00	74,90
30		53	1	29	41,00	2,00	19,90
31		34	0	10	20,00	1,00	10,00
32		47	1	6	22,00	1,00	11,40
33		58	0	2	60,00	3,00	29,70
34		25	1	0	58,00	3,00	28,40
35		57	1	28	92,00	4,00	45,50
36		30	1	7	21,00	1,00	10,50
37		21	0	0	13,00	1,00	6,30

# Las cuatro ventanas: Editor de datos

The screenshot shows the IBM SPSS Statistics Editor de datos window. The menu bar includes Archivo, Editar, Ver, Datos, Transformar, Analizar, Gráficos, Utilidades, Ventana, and Ayuda. The title bar indicates "demo.sav [Conjunto\_de\_datos2] - IBM SPSS Statistics Editor de datos". The main area displays a data grid with 41 rows and 13 columns. The columns are: #, Nombre, Tipo, Ancha, Decimales, Etiqueta, Valores, Perdidos, Columnas, Alineación, Medida, and Rol. The data includes various variables like "ingres", "ingcat", "coche", etc., with their respective properties such as type (Numérico), width (8), decimal places (2), and role (Entrada). The bottom navigation bar has tabs for "Vista de datos" (selected) and "Vista de variables". The status bar at the bottom right shows "IBM SPSS Statistics Processor está listo", "Unicode ON", the date "30/09/2015", and the time "14:57".

#	Nombre	Tipo	Ancha	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
4	ingres	Numérico	8	2	Ingresos del ho...	Ninguna	8		Derecha	Escala	Entrada
5	ingcat	Numérico	8	2	Categoría de in...	{1,00, Men...}	Ninguna	8	Derecha	Ordinal	Entrada
6	coche	Numérico	8	2	Precio del coch...	Ninguna	8		Derecha	Escala	Entrada
7	cocheat	Numérico	8	2	Categoría del pr...	{1,00, Econ...}	Ninguna	8	Derecha	Ordinal	Entrada
8	educ	Numérico	4	0	Nivel educativo	{1, No comp...}	Ninguna	8	Derecha	Escala	Entrada
9	empleo	Numérico	4	0	Años con la em...	Ninguna	8		Derecha	Escala	Entrada
10	retirado	Numérico	4	0	Retirado	{0, No...}	Ninguna	8	Derecha	Escala	Entrada
11	empcat	Numérico	4	0	Años con la em...	{1, Menos d...	Ninguna	8	Derecha	Ordinal	Entrada
12	sattab	Numérico	4	0	Satisfacción la...	{1, Muy ins...	Ninguna	8	Derecha	Escala	Entrada
13	genero	Cadena	1	0	Género	{h, Hombre...}	Ninguna	8	Izquierda	Nominal	Entrada
14	residen	Numérico	4	0	Número de per...	Ninguna	8		Derecha	Escala	Entrada
15	inalan	Numérico	4	0	Servicio inalám...	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
16	multiline	Numérico	4	0	Múltiples líneas	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
17	voz	Numérico	4	0	Buzón de voz	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
18	busca	Numérico	4	0	Buscapersonas	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
19	internet	Numérico	4	0	Internet	{0, Sí...}	8, 9	8	Derecha	Escala	Entrada
20	idllam	Numérico	4	0	ID llamadas	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
21	espera	Numérico	4	0	Llamada en es...	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
22	tv	Numérico	4	0	Tiene TV	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
23	video	Numérico	4	0	Tiene Video	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
24	cd	Numérico	4	0	Tiene Hi-Fi/CD	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
25	pda	Numérico	4	0	Tiene PDA	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
26	pc	Numérico	4	0	Tiene ordenador	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
27	fax	Numérico	4	0	Tiene fax	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
28	noticias	Numérico	4	0	Suscríto a un p...	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
29	request	Numérico	4	0	Respuesta	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada
30											
31											
32											
33											
34											
35											
36											
37											
38											
39											
40											
41											
42											

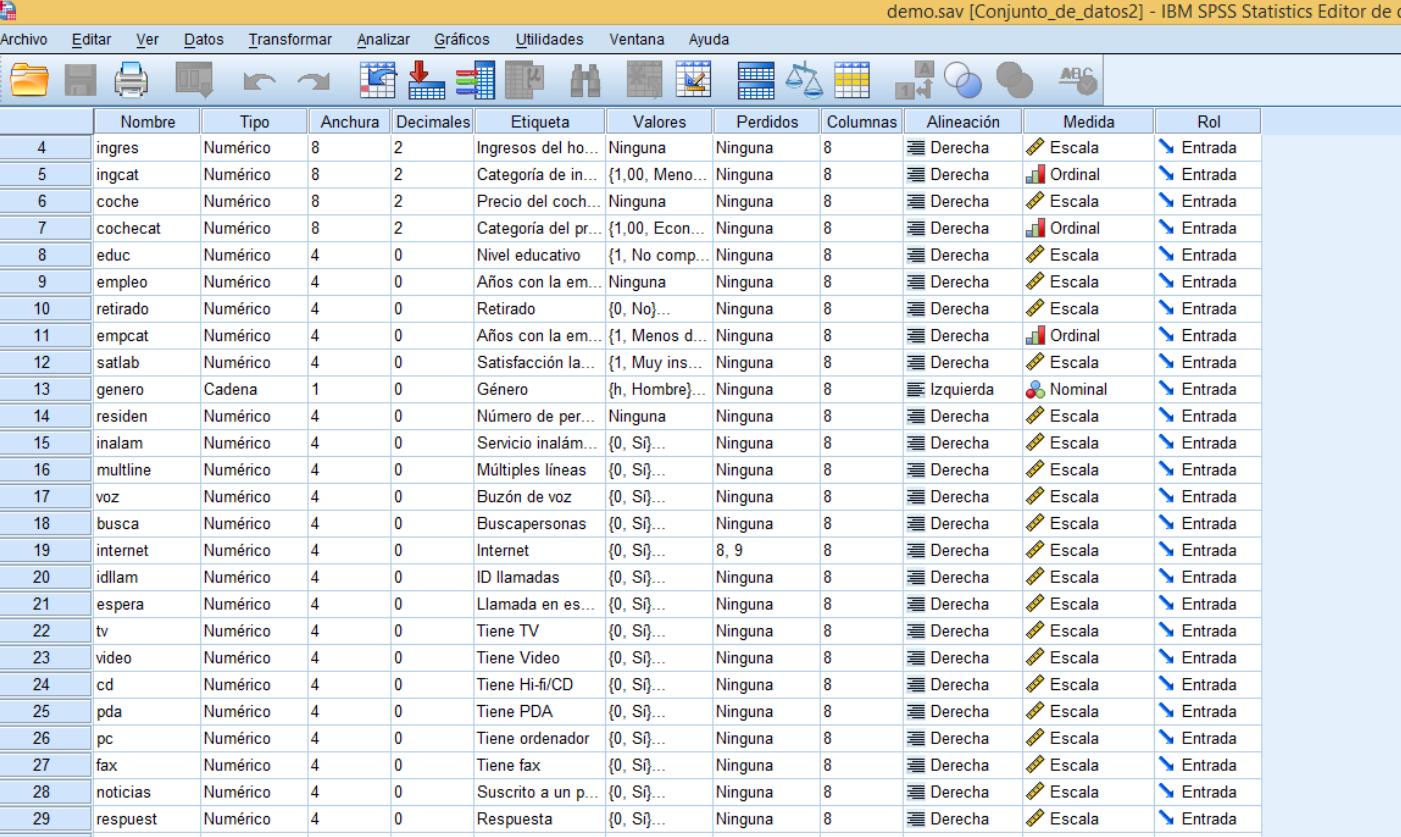
# Las cuatro ventanas: Editor de datos

- Si hacemos clic en la pestaña Vista de variables, pasamos a vista de variables.
- Esta hoja contiene información sobre el conjunto de datos que se almacena en el conjunto de datos: nombre, tipo, anchura, ...
- Nombre: el primer carácter del nombre de la variable debe ser alfabético
  - Los nombres de variables deben ser únicos, y tiene que ser inferior a 64 caracteres.
  - No se permiten espacios.
- Tipo: indica el tipo de variable.
  - Haga clic en la casilla de «tipo». Los dos tipos básicos de las variables que se utilizarán son numéricas y de cadena. Esta columna le permite especificar el tipo de variable.

# Las cuatro ventanas: Editor de datos

- **Anchura:** el número de dígitos para valores numéricos o la longitud de una variable de cadena.
- **Decimales:** número de decimales
  - Tiene que ser menor o igual a 16
- **Etiqueta:** puede especificar los detalles de la variable
  - Se puede escribir caracteres hasta 256.
- **Valores:** esto se utiliza para sugerir que números representan la categoría de la variable, cuando esta representa una categoría.
  - Definición de las etiquetas de valor:
    - Haga clic en la celda de la columna los valores.
    - Para el valor, y la etiqueta, puede poner hasta 60 caracteres.
    - Después de definir los valores haga clic en Agregar y, a continuación, haga clic en Aceptar.

# Las cuatro ventanas: Editor de datos



The screenshot shows the IBM SPSS Statistics Editor interface. The title bar reads "demo.sav [Conjunto\_de\_datos2] - IBM SPSS Statistics Editor de da". The menu bar includes Archivo, Editar, Ver, Datos, Transformar, Analizar, Gráficos, Utilidades, Ventana, and Ayuda. Below the menu is a toolbar with various icons for file operations like Open, Save, Print, and View. The main area is a data grid with 29 rows and 13 columns. The columns are: #, Nombre, Tipo, Anchura, Decimales, Etiqueta, Valores, Perdidos, Columnas, Alineación, Medida, and Rol. The data rows represent variables such as "ingres", "ingcat", "coche", etc., with their respective properties like type (Numérico, Cadena), width, decimals, and scale level (Escala, Ordinal, Nominal).

#	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol	
4	ingres	Numérico	8	2	Ingresos del ho...	Ninguna	Ninguna	8	Derecha	Escala	Entrada	
5	ingcat	Numérico	8	2	Categoría de in...	{1,00, Menor...	Ninguna	8	Derecha	Ordinal	Entrada	
6	coche	Numérico	8	2	Precio del coch...	Ninguna	Ninguna	8	Derecha	Escala	Entrada	
7	cochecat	Numérico	8	2	Categoría del pr...	{1,00, Econ...}	Ninguna	8	Derecha	Ordinal	Entrada	
8	educ	Numérico	4	0	Nivel educativo	{1, No comp...}	Ninguna	8	Derecha	Escala	Entrada	
9	empleo	Numérico	4	0	Años con la em...	Ninguna	Ninguna	8	Derecha	Escala	Entrada	
10	retirado	Numérico	4	0	Retirado	{0, No...}	Ninguna	8	Derecha	Escala	Entrada	
11	empcat	Numérico	4	0	Años con la em...	{1, Menos d...}	Ninguna	8	Derecha	Ordinal	Entrada	
12	satlab	Numérico	4	0	Satisfacción la...	{1, Muy ins...}	Ninguna	8	Derecha	Escala	Entrada	
13	genero	Cadena	1	0	Género	{h, Hombre...}	Ninguna	8	Izquierda	Nominal	Entrada	
14	residen	Numérico	4	0	Número de per...	Ninguna	Ninguna	8	Derecha	Escala	Entrada	
15	inalam	Numérico	4	0	Servicio inalám...	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
16	multiline	Numérico	4	0	Múltiples líneas	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
17	voz	Numérico	4	0	Buzón de voz	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
18	busca	Numérico	4	0	Buscapersonas	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
19	internet	Numérico	4	0	Internet	{0, Sí...}	8, 9	8	Derecha	Escala	Entrada	
20	idllam	Numérico	4	0	ID llamadas	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
21	espera	Numérico	4	0	Llamada en es...	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
22	tv	Numérico	4	0	Tiene TV	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
23	video	Numérico	4	0	Tiene Video	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
24	cd	Numérico	4	0	Tiene Hi-fi/CD	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
25	pda	Numérico	4	0	Tiene PDA	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
26	pc	Numérico	4	0	Tiene ordenador	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
27	fax	Numérico	4	0	Tiene fax	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
28	noticias	Numérico	4	0	Suscripto a un p...	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	
29	respuest	Numérico	4	0	Respuesta	{0, Sí...}	Ninguna	8	Derecha	Escala	Entrada	

# Las cuatro ventanas: Salida de resultados

- Una ventana de salida de resultados: visualiza la salida de los resultados.
- Puede grabar los resultados en ficheros de extensión “spv”.
- Aparece los comandos que se introducen a SPSS en la parte de arriba.
- Aparece un árbol de operaciones a la izquierda.
- Puedes modificar las etiquetas de la salida.
- p. ej. marcar columnas con el ratón>botón derecho>estadísticos descriptivos.

# Las cuatro ventanas: Salida de resultados

The screenshot displays the IBM SPSS Statistics software interface with four open windows:

- Data Editor:** Shows a table with 37 rows and 5 columns. The columns are labeled: edad, marital, direcc, and tv. The data includes various numerical values and some missing data (represented by empty cells).
- Output Viewer:** Shows the results of a FREQUENCIES command. It includes:
  - FREQUENCIES VARIABLES=edad marital /STATISTICS=RANGE MINIMUM MAXIMUM STDDEV MEAN MEDIAN /FORMAT=LIMIT(50) /ORDER=ANALYSIS.**
  - Frecuencias:** A table showing frequencies for marital status (Valido, Perdidos), mean age (42.06), median age (41.00), standard deviation (5.50), range (59), minimum (18), and maximum (77).
  - Tabla de frecuencia:** A table showing the distribution of marital status (Valido, Sin casar, Casado) across the population (Total). The data is as follows:

	Frecuencia	Porcentaje válido	Porcentaje acumulado
Valido	3224	50,4	50,4
Sin casar	3176	49,6	49,6
Casado			100,0
Total	6400	100,0	100,0
- Syntax Editor:** Shows the command entered: FREQUENCIES VARIABLES=edad marital /STATISTICS=RANGE MINIMUM MAXIMUM STDDEV MEAN MEDIAN /FORMAT=LIMIT(50) /ORDER=ANALYSIS.
- Results window:** Shows the output of the analysis, including the frequency table and the marital status distribution table.

# Las cuatro ventanas: Editor de sintaxis

- Es un editor para composición de comandos y lenguaje de SPSS y posteriormente ejecutarlo.
- Se pueden grabar las hojas de composición de expresiones en SPSS con la extensión “sps”. Son archivos de texto.
- Puede **ahorrar y automatizar** muchas **tareas** comunes mediante el eficaz lenguaje de comandos.
- El lenguaje de comandos también proporciona algunas **funcionalidades no incluidas** en los **menús** y cuadros de **diálogo**.
- El lenguaje de comandos también permite guardar los trabajos en un archivo de sintaxis, con lo que podrá **repetir los análisis** en otro momento.

# Las cuatro ventanas: Editor de sintaxis

The screenshot displays the IBM SPSS Statistics software interface with four main windows:

- Data Editor:** Shows a dataset named "ingcat" with 29 variables and 37 rows of data.
- Syntax Editor:** Displays the syntax command: `GET  
DATASET NAME  
FREQUENCIES  
/STATISTICS=RANGE MINIMUM MAXIMUM STDDEV MEAN MEDIAN  
/FORMAT=LIMIT(50)  
/ORDER=ANALYSIS.`
- Output Viewer:** Shows the output for the "FREQUENCIES" command, including a frequency distribution table for the variable "ingres".
- Results viewer:** Shows the results of the "FREQUENCIES" command, including descriptive statistics for "ingres": N=6400, Media=69,4748, Mediana=45,0000, Desviación estándar=78,1856, Rango=1107,00, Mínimo=9,00, and Máximo=1116,00.

# Las cuatro ventanas: Ventana de Script

- La ventana de Script permite escribir programas, en un lenguaje muy similar al BASIC.
- La ventana es un simple editor de texto para la composición de sintaxis. La extensión del archivo guardado será “sbs”.
- El fichero se puede ejecutar en el SPSS.
- Esta ventana es para usuarios más avanzados.
- Algunos ejemplos y documentación en :
  - <ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/ScriptingAutomation.pdf>

# Las cuatro ventanas: Ventana de Script

The screenshot shows the IBM SPSS Statistics interface with four main windows:

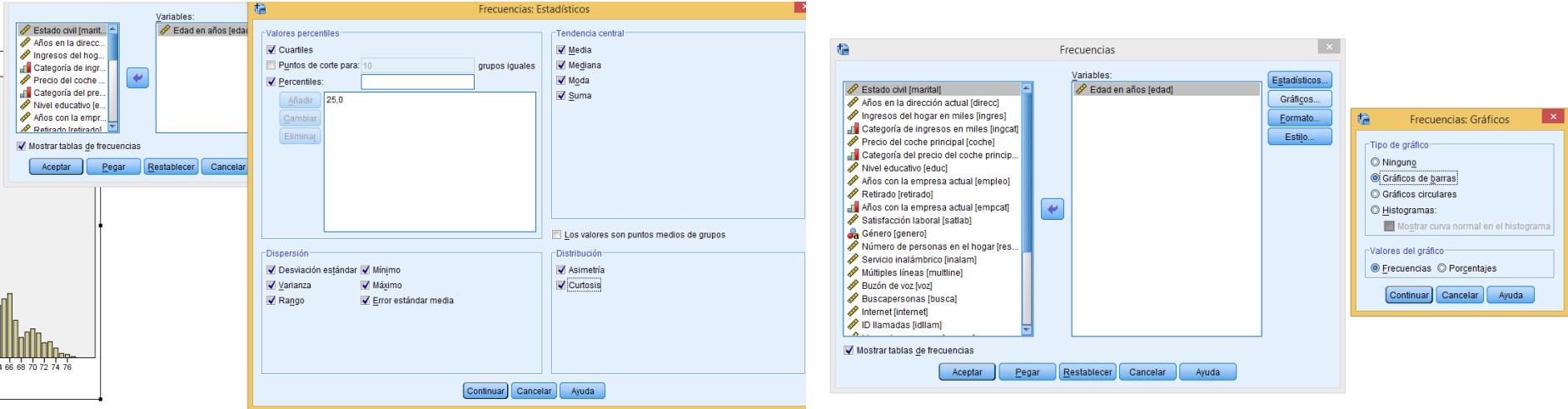
- Data View:** A large window on the right displaying a dataset with 37 rows and 29 columns. The columns include variables like edad, marital, direcc, ingreso, ingcat, coche, cocecat, educ, empleo, retirado, empcat, satlab, genero, residen, inalam, multiline, voz, busca, internet, idllam, espera, and tv.
- Variable View:** A smaller window below the Data View showing variable details for each column.
- Script Editor:** A window titled "Script11 \*(script) - SPSS Statistics Basic Script Editor [design]" containing VBA-like code for file processing.
- System Log:** A window at the bottom titled "IBM SPSS Statistics Processor está listo" (IBM SPSS Statistics Processor is ready).

# Análisis Básico

- **Frecuencias**
  - Este análisis genera tablas de frecuencias que muestran recuentos de frecuencias y porcentajes de los valores de las variables individuales.
- **Descriptivos**
  - Este análisis muestra el máximo, mínimo, media y desviación estandar de las variables
- **Análisis de regresión lineal**
  - La regresión lineal estima los coeficientes de la ecuación lineal

# Análisis Básico: Frecuencias

- Haga clic en Analizar>estadísticos descriptivos>Frecuencias
- Haga clic en la variable de estudio y muévala a la derecha
  - En estadísticos: se puede seleccionar lo que quieras presentar.
  - En gráficos puedes elegir barras, histograma, etc.
- En aceptar presenta los cálculos y gráficas.



# Análisis Básico: Frecuencias

\*Resultado3 [Documento3] - IBM SPSS Statistics Visor

Archivo Editar Ver Datos Transformar Insertar Formato Gráficos Utilidades Ventana Ayuda

N Válido Perdidos  
Media 42,06  
Error estándar de la media ,154  
Mediana 41,00  
Moda 39  
Desviación estándar 12,290  
Varianza 151,032  
Asimetría ,299  
Error estándar de asimetría ,031  
Curtosis ,602  
Error estándar de curtosis ,061  
Rango 59  
Mínimo 18  
Máximo 77  
Suma 269175  
Percentiles 25 33,00  
50 41,00  
75 51,00

Edad en años

Frecuencia

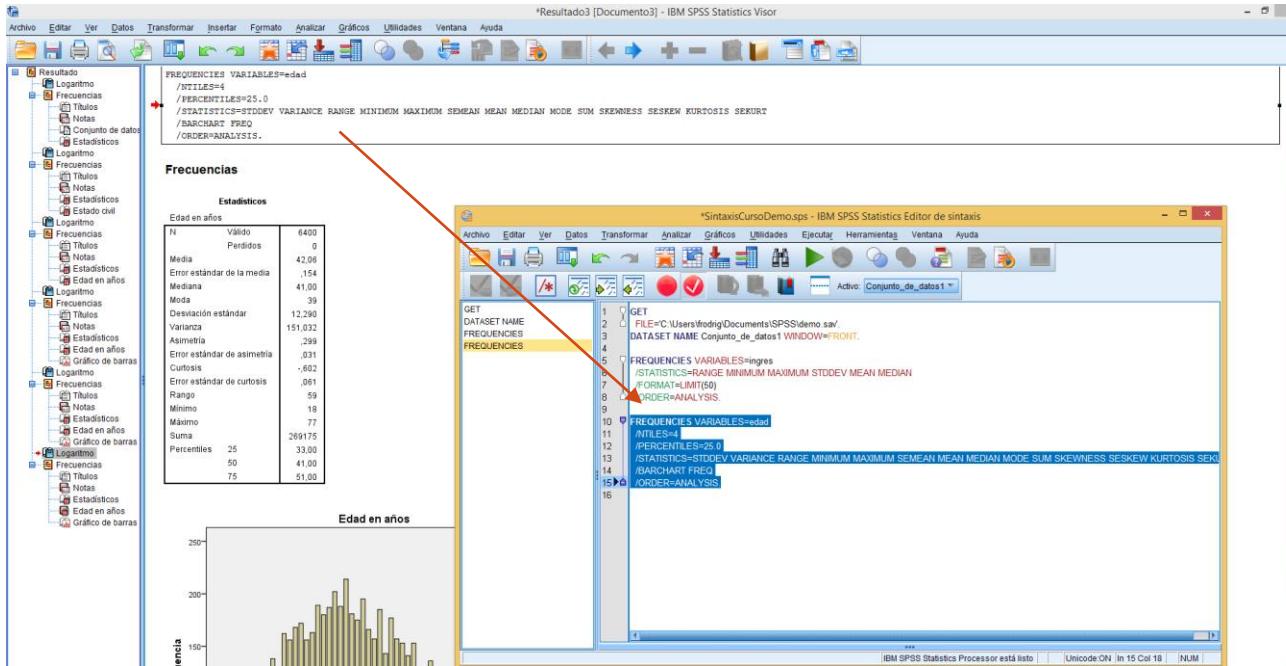
Edad en años

IBM SPSS Statistics Processor está listo 19:41 Unicode ON 30/09/2015

Edad (Años)	Frecuencia
10	10
12	20
14	30
16	40
18	50
20	60
22	70
24	80
26	90
28	100
30	110
32	120
34	130
36	140
38	150
40	160
42	150
44	140
46	130
48	120
50	110
52	100
54	90
56	80
58	70
60	60
62	50
64	40
66	30
68	20
70	10
72	5
74	2
76	1

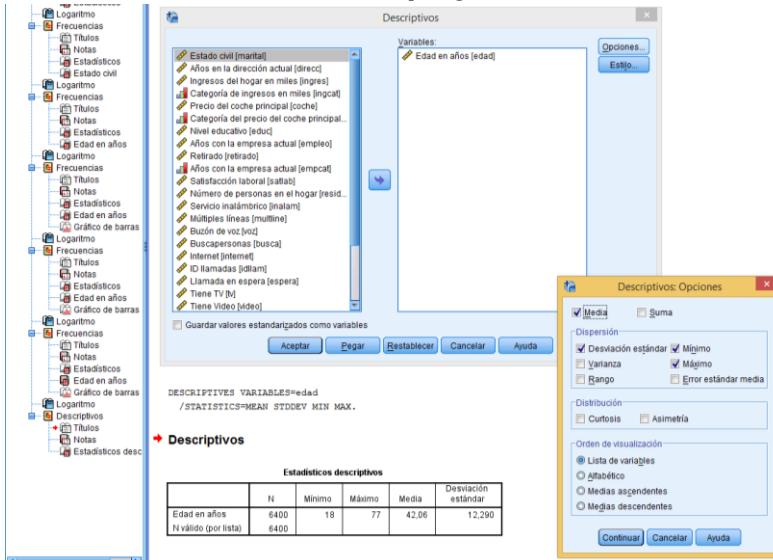
# Análisis Básico: Frecuencias

- Todos procedimientos utilizando la GUI de SPSS, se pueden hacer mediante el editor de sintaxis, copiando los comandos que aparecen al principio de los cálculos del visor de resultados:



# Análisis Básico: Descriptivos

- Este análisis muestra el máximo, mínimo, media y desviación estándar de las variables.
- Haga clic en Analizar>estadísticos descriptivos>Descriptivos
- Haga clic en la variable de estudio y muévala a la derecha
- En aceptar presenta los cálculos y gráficas.



# Análisis Básico: Análisis de regresión lineal

- La regresión lineal estima los coeficientes de la ecuación lineal
- Haga clic en Analizar>Regresiones>Lineales y por ejemplo:

The screenshot shows the IBM SPSS Statistics Visor interface. On the left, there's a navigation tree with various analysis options like Logaritmo, Descriptivos, and ANOVA. The main area displays the output from a regression analysis.

**REGRESIÓN**

```
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT edad
/METHOD=ENTER ingresos.
```

**Resumen del modelo**

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	.335 <sup>a</sup>	.112	.112	11.579

a. Variable dependiente: Edad en años  
b. Todas las variables solicitadas introducidas.

**ANOVA<sup>b</sup>**

Modelo	Suma de cuadrados	gl	Media cuadrada	F	Sig.
1	Regresión 108590,209	1	108590,209	809,871	,000 <sup>b</sup>
	Residuo 957864,818	6398	134,083		
	Total 966455,027	6399			

a. Variable dependiente: Edad en años  
b. Predictores: (Constante), Ingresos del hogar en miles

**Coefficientes<sup>a</sup>**

Modelo	Coeficientes no estandarizados			t	Sig.
	B	Error estándar	Beta		
1 (Constante)	38,423	,193	199,020	,000	
Ingresos del hogar en miles	,052	,002	,335	28,458	,000

**Resolución lineal**

Dependientes: Edad en años [edad]

Bloque 1 de 1

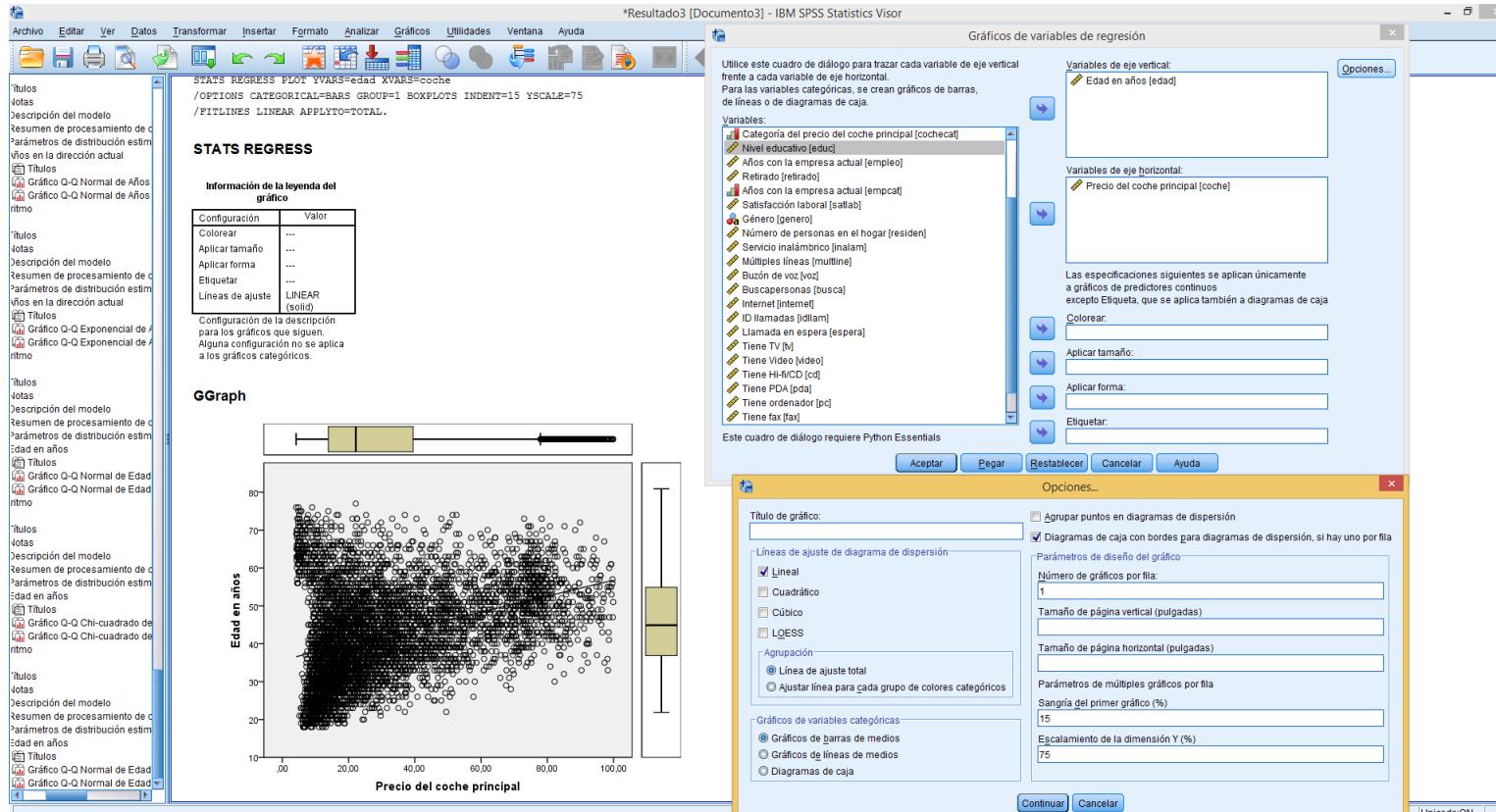
Independientes: Ingresos del hogar en miles [ingresos]

Método: Intro

Acceptar Pagar Restablecer Cancelar Ayuda

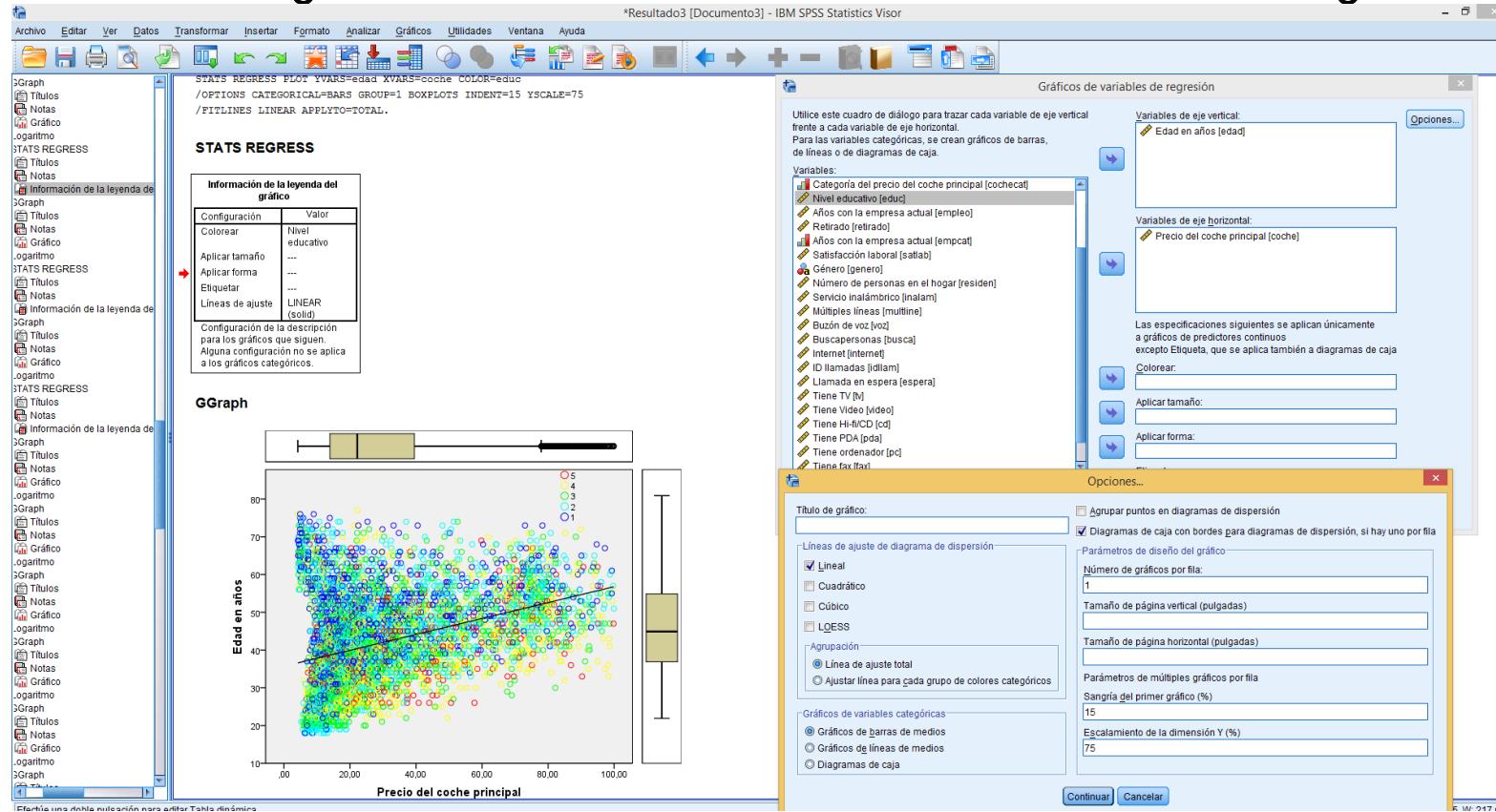
# Análisis Básico: Análisis de regresión lineal

- Pintar la regresión: Gráficos>Gráficos de variables de regresión.



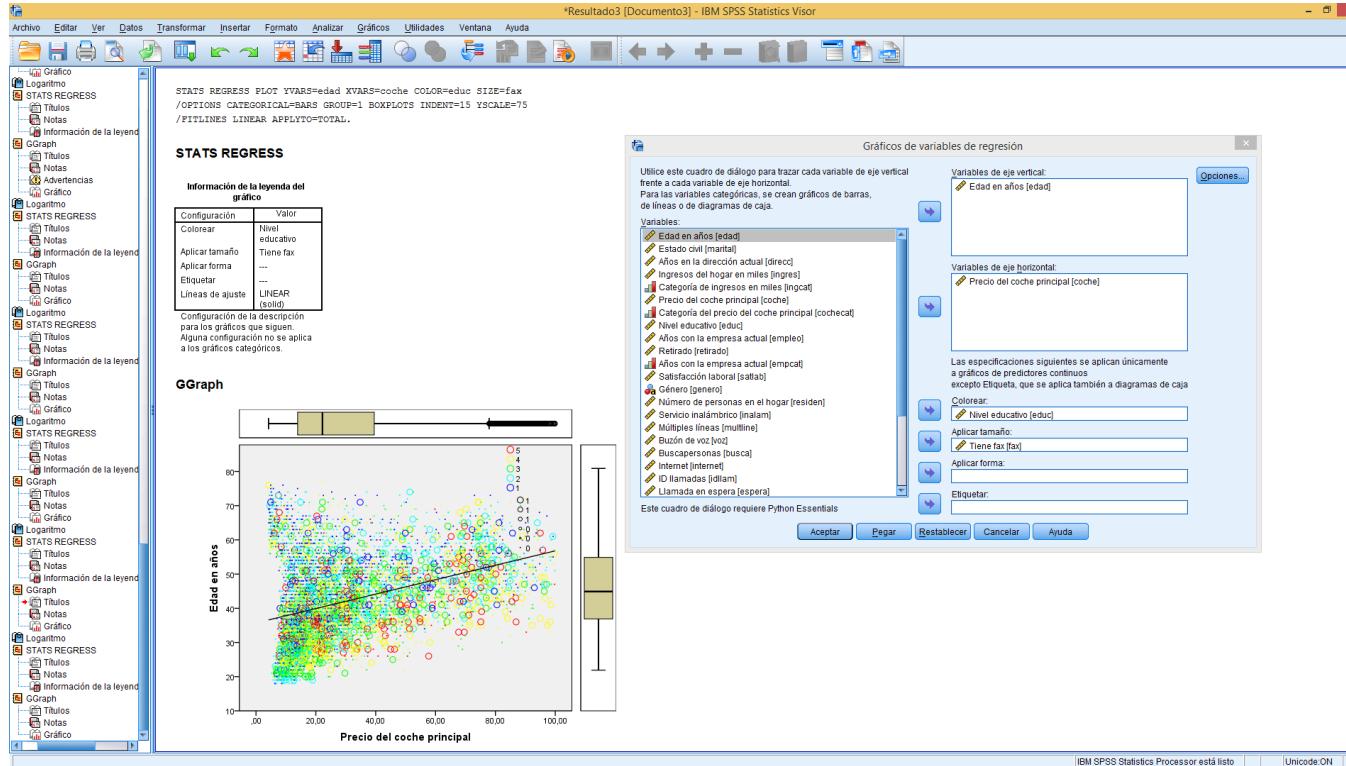
# Análisis Básico: Análisis de regresión lineal

➤ Pintar la regresión: Gráficos>Gráficos de variables de regresión.

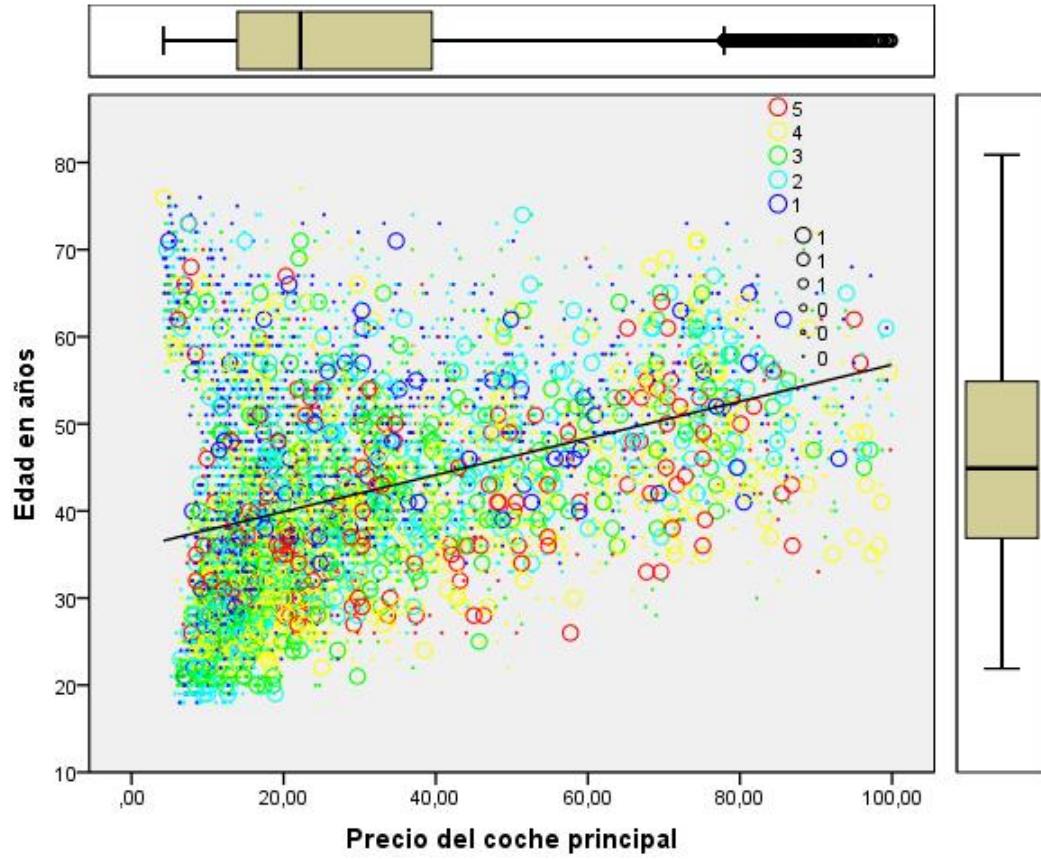


# Análisis Básico: Análisis de regresión lineal

- Pintar la regresión: Gráficos>Gráficos de variables de regresión.



# Análisis Básico: Análisis de regresión lineal



**Información de la leyenda del gráfico**

Configuración	Valor
Colorear	Nivel educativo
Aplicar tamaño	Tiene fax
Aplicar forma	---
Etiquetar	---
Líneas de ajuste	LINEAR (solid)

Configuración de la descripción para los gráficos que siguen.  
Alguna configuración no se aplica a los gráficos categóricos.

# Representaciones Visuales: Histogramas

\*Resultado3 [Documento3] - IBM SPSS Statistics Visor

Archivo Editar Ver Datos Transformar Insertar Formato Analizar Gráficos Utilidades Ventana Ayuda

Edad en años

\* Generador de gráficos.

GGGRAPH

/GRAPHDATASET NAME="graphdataset" VARIABLES=edad MISSING=LISTWISE REPORTMISSING=NO

/GRAPHSPEC SOURCE=INLINE.

BEGIN GPL

SOURCE: s=userSource(id("graphdataset"))

DATA: edad=col(source(s), name("edad"))

GUIDE: axis(dim1), label("Edad en años")

GUIDE: axis(dim2), label("Frecuencia")

ELEMENT: interval(position(summary.count(bin.rect(edad))), shape.interior(shape.square))

END GPL.

Edad en años

Frecuencia

Media = 42.06  
Desviación estándar = 12.29  
N = 6.400

Edad en años

Generador de gráficos

Variables:

- Edad en años (e...)
- Estado civil [mant...]
- Años en la direcc...
- Ingresos del hog...
- Categoría de ingr...
- Precio del coche ...
- Categoría del pre...
- Nivel educativo (e...
- Años con la empr...
- Retirado (retirado)
- Años con la empr...

No categories (scale variable)

Histograma

Edad en años

Galería Elementos básicos Grupos/ID de puntos Títulos/notas al pie

Propiedades de elemento... Opciones...

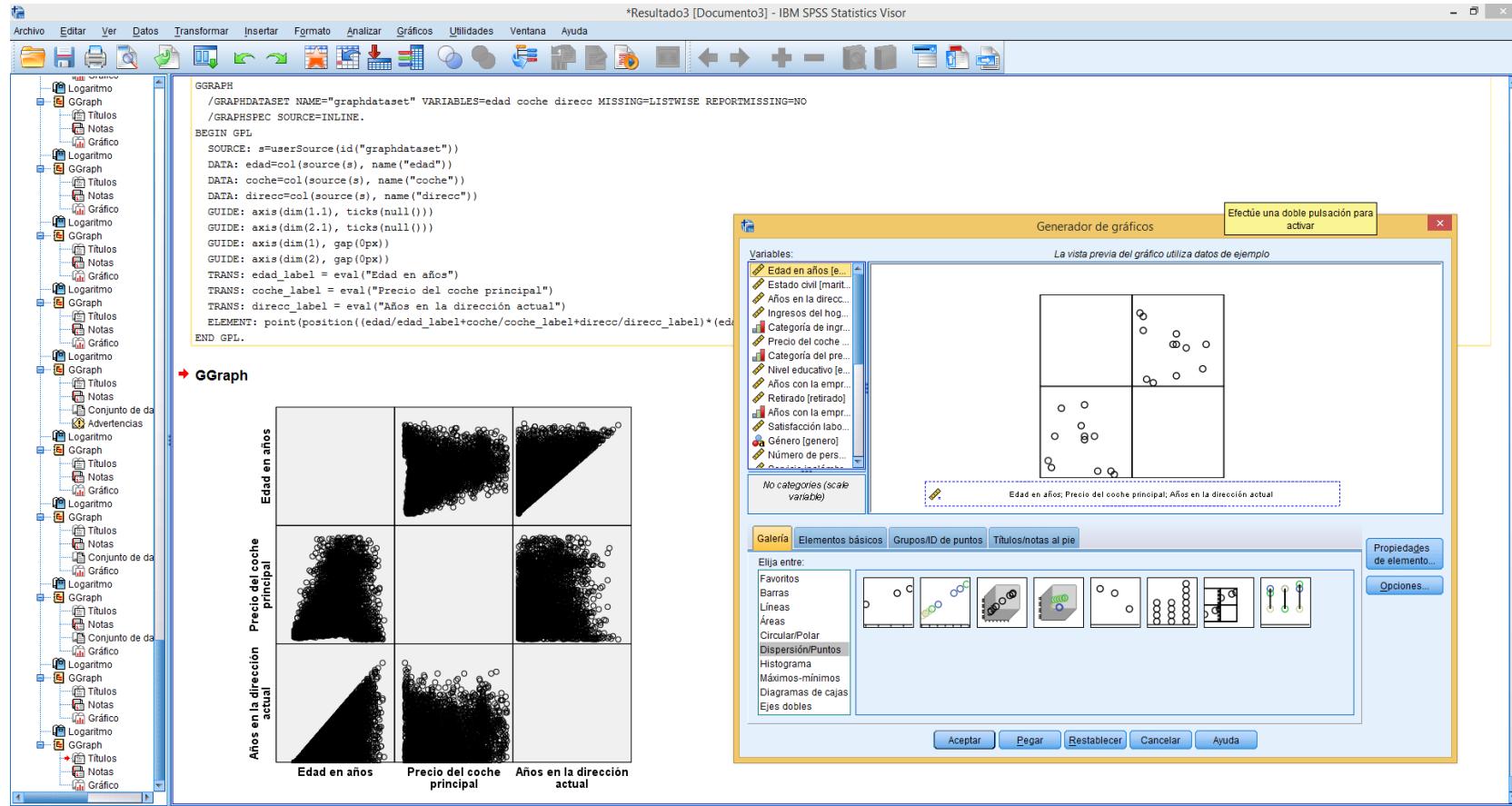
Elige entre:

- Favoritos
- Barra
- Líneas
- Áreas
- Circular/Polar
- Dispersión/Puntos
- Histograma
- Máximos-mínimos
- Diagramas de cajas
- Ejes dobles

Aceptar Pegar Restablecer Cancelar Ayuda

Universidad Autónoma de Madrid

# Representaciones Visuales: Matriz de dispersión



# Representaciones Visuales: Boxplots

The screenshot shows the IBM SPSS Statistics interface with the title bar "Resultado3 [Documento3] - IBM SPSS Statistics Visor". The menu bar includes Archivo, Editar, Ver, Datos, Transformar, Insertar, Formato, Analizar, Gráficos, Utilidades, Ventana, and Ayuda. The toolbar has various icons for file operations, selection, and analysis.

The left pane displays a hierarchical tree structure of saved graphs and their components. A red arrow points from the tree to a box plot in the main workspace.

The main workspace contains a box plot titled "Años en la dirección actual". The y-axis ranges from 0 to 60. The box plot shows a median around 10, a box from approximately 5 to 15, and whiskers extending to about 20 and 55. Outliers are labeled with their values: 1998, 5078, 6259, 517, 153, 1870, 1748, 153, 215, 103, 468, 116, 665, 668, and 1064.

The right pane shows the "Generador de gráficos" dialog box. It displays the generated GGRAPH command:

```
* Generador de gráficos.  
GGRAPH  
/GRAPHDATASET NAME="graphdataset" VARIABLES=direcc MISSING=LISTWISE REPORTMISSING=NO  
/GRAPHSPEC SOURCE=INLINE.  
BEGIN GPL  
SOURCE: s=userSource(id("graphdataset"))  
DATA: direcc=col(source(s), name("direcc"))  
DATA: id=col(source(s), name("$CASENUM"), unit.category())  
GUIDE: axis(dim(2), label("Años en la dirección actual"))  
ELEMENT: schema(position(bin.quantile.letter(1*direcc)), label(id))  
END GPL.
```

A yellow box highlights the instruction "Efectúe una doble pulsación para activar".

The "Generador de gráficos" dialog box also shows the "Variables" list, which includes "Edad en años (edad)", "Estado civil (mant.)", "Años en la direcc...", "Ingresos del hogar...", "Categoría de ingr...", "Precio del coche...", "Categoría del pre...", "Nivel educativo (e...)", "Retirado (retirado)", "Años con la empr.", "Retirado (retirado)", and "Años con la empr.". The "Años en la dirección actual" variable is selected.

The "Galería" tab is selected in the dialog, showing various chart options like "Favoritos", "Barras", "Líneas", etc. The "Diagramas de cajas" option is highlighted.

Buttons at the bottom of the dialog include "Aceptar", "Pegar", "Restablecer", "Cancelar", and "Ayuda".

# Representaciones Visuales: Histogramas

\*Resultado3 [Documento3] - IBM SPSS Statistics Visor

Archivo Editar Ver Datos Transformar Insertar Formato Analizar Gráficos Utilidades Ventana Ayuda

Generador de gráficos.

```
GRAPH  
/GRAPHDATASET NAME="graphdataset" VARIABLES=direcc MISSING=LISTWISE REPORTMISSING=NO  
/GRAPHSPEC SOURCE=INLINE.  
BEGIN GPL  
SOURCE: s=userSource(id("graphdataset"))  
DATA: direcc=col(source(s), name("direcc"))  
GUIDE: axis(dim(1), label("Años en la dirección actual"))  
GUIDE: axis(dim(2), label("Frecuencia"))  
ELEMENT: interval(position(summary.count(bin.rect(direcc))), shape.interior(shape.square))  
END GPL.
```

**GGraph**

Edad en años (e... Estado civil [mar... Años en la direc... Ingresos del hog... Categoría de ingr... Precio del coche... Categoría del pre... Nivel educativo (e... Años con la empr... Retirado [retirado] Años con la empr...

No categories (scale variable)

Efectúe una doble pulsación para activar

La vista previa del gráfico utiliza datos de ejemplo

Histograma

Años en la dirección actual

Frecuencia

Media = 11,56 Desviación estándar = 9,938 N = 6,400

Elige entre:

- Favoritos
- Barras
- Líneas
- Áreas
- Circular/Polar
- Dispersión/Puntos
- Histograma**
- Máximos-mínimos
- Diagramas de cajas
- Ejes dobles

Galería Elementos básicos Grupos/ID de puntos Títulos/notas al pie

Propiedades de elemento... Opciones...

Aceptar Pegar Restablecer Cancelar Ayuda

The screenshot shows the IBM SPSS Statistics Visor interface. On the left, there's a tree view of saved graphs. The main area displays a histogram titled "Años en la dirección actual" with a frequency axis ranging from 0,0 to 500,0. The histogram has approximately 40 bins. A context menu is open over the histogram, listing various chart types like "Favoritos", "Barras", "Líneas", etc., with "Histograma" currently selected. To the right of the histogram is a preview window titled "Generador de gráficos" showing a smaller version of the histogram with a legend for "Años en la dirección actual". The preview window also includes tabs for "Galería", "Elementos básicos", "Grupos/ID de puntos", and "Títulos/notas al pie". At the bottom of the preview window are buttons for "Aceptar", "Pegar", "Restablecer", "Cancelar", and "Ayuda".

# Representaciones Visuales: QQplot

\*Resultado3 [Documento3] - IBM SPSS Statistics Visor

Archivo Editar Ver Datos Transformar Insertar Fórmato Analizar Gráficos Utilidades Ventana Ayuda

Informes Estadísticos descriptivos Tablas Comparar medias Modelo lineal general Modelos lineales generalizados Modelos mixtos Correlaciones Regresión Loglineal Clasificar Reducción de dimensiones Escala Pruebas no paramétricas Previsiones Supervivencia Respuesta múltiple Simulación... Control de calidad Curna COR...

Método de estimación de rango fijo Rango asignado a empates Aplicando las especificaciones d...

Resumen de procedimientos

Longitud de serie o secuencia Número de valores Perdidos en el gráfico Los casos no están ponderados...

Distribución normal Ubicación Escala Los casos no están ponderados...

Edad en años

Gráfico Q-Q Normal de Edad en años

Valor Normal esperado  
80  
60  
40  
20  
0

0 20 40 60 80 100

Valor observado

Gráficos Q-Q

Variables: Edad en años [edad]

Distribución de prueba: Normal

g: 1

Parámetros de distribución

Estimar a partir de los datos

Posición: 0 Escala: 1

Transformar

Estandarizar valores

Diferencia:

Diferenciar ciclo: 1

Periodicidad actual: Ninguna

Fórmula de estimación de la proporción

De Blom Ranit De Tukey De Van der Waerden

Rango asignado a empates

Media Mayor Menor Romper los empates arbitrariamente

Aceptar Pegar Restablecer Cancelar Ayuda

Universidad Autónoma de Madrid

# Representaciones Visuales: QQplot

\*Resultado3 [Documento3] - IBM SPSS Statistics Visor

Archivo Editar Ver Datos Transformar Insertar Formato Analizar Gráficos Utilidades Ventana Ayuda

Títulos  
Notas  
Descripción del modelo  
Resumen de procesamiento de c  
Parámetros de distribución estim  
ños en la dirección actual

Gráfico Q-Q Normal de Años  
ritmo

Títulos  
Notas  
Descripción del modelo  
Resumen de procesamiento de c  
Parámetros de distribución estim  
ños en la dirección actual

Gráfico Q-Q Exponencial de A  
ritmo

Títulos  
Notas  
Descripción del modelo  
Resumen de procesamiento de c  
Parámetros de distribución estim  
Edad en años

Gráfico Q-Q Normal de Edad  
ritmo

Títulos  
Notas  
Descripción del modelo  
Resumen de procesamiento de c  
Parámetros de distribución estim  
Edad en años

Gráfico Q-Q Chi-cuadrado de  
ritmo

Títulos  
Notas  
Descripción del modelo  
Resumen de procesamiento de c  
Parámetros de distribución estim  
Edad en años

Gráfico Q-Q Normal de Edad  
ritmo

Edad en años

Gráfico Q-Q Normal de Edad en años

Gráfico Q-Q Normal de Edad en años sin tendencia

Valor Normal esperado

Valor observado

Desviación de Normal

Valor observado

Gráficos Q-Q

Variables: Edad en años [edad]

Distribución de prueba: Normal

gl: 1

Estimar a partir de los datos

Posición: 0

Escala: 1

Transformar

Estandarizar valores

Diferencia:

Diferenciar ciclo:

Fórmula de estimación de la proporción

De Blom Rankit De Tukey

De Van der Waerden

Rango asignado a empates

Media Mayor Menor

Romper los empates arbitrariamente

Aceptar Pegar Restablecer Cancelar Ayuda

The screenshot shows the IBM SPSS Statistics interface with two QQ plots displayed. The top plot is titled 'Gráfico Q-Q Normal de Edad en años' and shows the relationship between 'Valor observado' (X-axis, 0-100) and 'Valor Normal esperado' (Y-axis, 0-100). The data points closely follow a diagonal line, indicating approximate normality. The bottom plot is titled 'Gráfico Q-Q Normal de Edad en años sin tendencia' and shows the relationship between 'Valor observado' (X-axis, 10-80) and 'Desviación de Normal' (Y-axis, -15 to 15). This plot shows the deviation from the normal distribution, with points generally centered around zero. To the right of the plots is the 'Gráficos Q-Q' dialog box, which allows for specifying variables, distribution types, and various transformation options like standardizing values or handling ties.

# Bibliografía y lecturas relacionadas:

- [SPSS White Papers](#)
- [SPSS Demos and Tutorials](#)
- **Repositorio de IBM: <ftp://public.dhe.ibm.com/>**
  - <ftp://public.dhe.ibm.com/software/analytics/spss/>
- [Descripción de todos los archivos muestrales de SPSS](#)
- [Discovering Statistics Using IBM SPSS Statistics. Field, Andy. Sage Publications Ltd, 2012.](#)