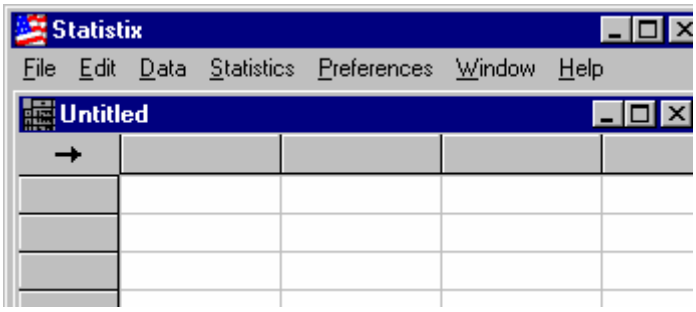
Departamento de Matemática Aplicada y Estadística
Universidad Politécnica de Cartagena

Práctica 1. Introducción al programa Statistix.

El **SX** o **STATISTIX** es un paquete estadístico del que vamos a usar la segunda versión en Windows.

Para ejecutarlo, se pulsa dos veces sobre el icono del programa, en el escritorio de Windows. La primera presentación es de una tabla de datos, donde se deberán introducir los datos de cada problema o leerlos de un fichero.



En el menú principal, se encuentran las opciones :

File **Edit** **Data** **Statistics** **Preferences**

·File :

Nos permitirá seleccionar las opciones de grabar en un fichero los datos introducidos, leer los datos de un fichero, imprimir, y otras opciones de manejo de ficheros (usuales en cualquier programa)

·Edit :

Nos permitirá copiar, cortar y pegar, uno o varios datos seleccionados con el ratón, de una o varias columnas.

·Data :

Nos permitirá introducir variables y sus datos, así como distintas opciones de manejo de los datos (seleccionar, omitir,..., para realizar los cálculos estadísticos con una parte de los mismos sin la pérdida de los restantes).

·Statistics :

Nos permitirá realizar los cálculos estadísticos que precisemos para la mayoría de las prácticas.

·Preferences :

Nos permitirá modificar las opciones por defecto del tratamiento de datos y gráficos.

I. Los primeros pasos.

Antes de todo, debemos introducir los datos. Para hacerlo, distinguiremos dos posibilidades: introducimos los datos manualmente o los importamos a nuestra hoja de cálculo desde un fichero externo. En el ejemplo ilustrativo que seguiremos a lo largo de esta primera sesión, veremos las dos situaciones.

I.1. Introducimos los datos manualmente:

En este caso debemos introducir, en primer lugar, los nombres de las variables y el tipo de cada una. Esto se realiza seleccionando la opción "***Insert -> variable***" del menu Datos. En el cuadro de diálogo, aparecen dos campos, en la ventana de la izquierda encontramos un listado de las variables que ya están definidas mientras que en la ventana de la derecha podemos introducir el nombre de la o las nuevas variables que deseamos definir. Junto con el nombre de la nueva variable podemos, si es necesario, introducir su tipo. Existen cuatro tipos de variables en Statistix: real, entero, fecha y caracteres. El tipo se especifica entre paréntesis directamente después de la variable con los códigos siguientes

(r) : real (opción por defecto)
(i): entero
(d):fecha (mes/día/año)
(s#) : # caracteres

Por ejemplo, queremos introducir los valores obtenidos en mediciones repetidas de contenido en nitratos de una muestra de agua que aparecen tabulados a continuación:

VALORES (µg/l)	FRECUENCIA	VALORES(µg/l)	FRECUENCIA
0.45	1	0.49	8
0.46	2	0.50	10
0.47	4	0.51	5
0.48	8	0.52	2

Definimos una única variable CONC, que tome valores reales, y empezamos a introducir los datos. Los valores de cada variable se introducen, colocándose con el ratón en la casilla deseada y desplazándose de casilla a casilla con las flechas del cursor.

En el caso en que debemos introducir repetidamente el mismo valor podemos utilizar la opción ***Fill*** del menú ***DATA***, que nos permite especificar el valor que queremos introducir junto con el número de casillas que debe ocupar.
Si queremos añadir algún comentario sobre el conjunto de datos, sobre alguna variable (sus unidades de medidas) o sobre algún valor en particular, podemos hacerlo a través de la opción ***LABELS*** del mismo menú ***DATA***.

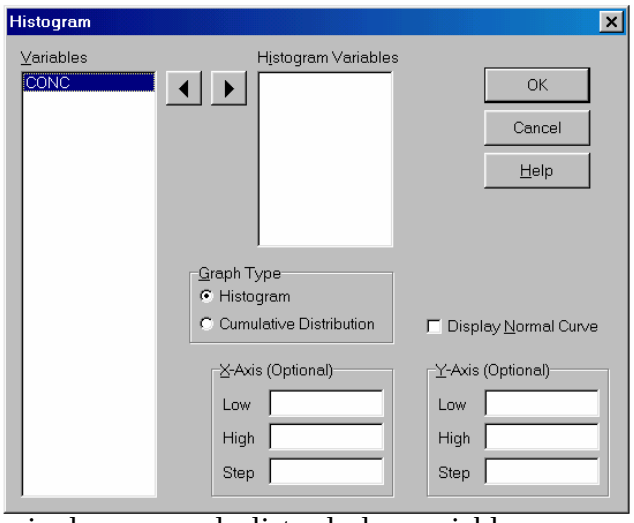
Se aconseja guardar la tabla de datos en un fichero después de la introducción de datos. Para ello, se usa la opción "***SAVE***" o "***SAVE AS***" del menú ***FILE***. Al igual que cualquier programa Windows, se puede recorrer las carpetas para decidir donde guardar el fichero.
Guardar la tabla de datos anterior en un fichero llamado ***nitrato.sx*** en la carpeta **\PRACTICAS\ESTADISTICA.**

Una vez que se han entrado unos datos, es posible añadir entre dos filas de una variable uno o varios datos nuevos usando la opción "***Insert->cases***" del menu ***Datos***. Tenemos que especificar el

número de la casilla que ocupará el primer dato nuevo y el número de casillas nuevas que hay que introducir.
La opción **DELETE** es utilizada para borrar datos por bloques, o bien para eliminar alguna o algunas variable. Para practicar, podéis introducir dos datos entre las filas 13 y 14, y volverlos a borrar.

I.2. Exploración de los datos

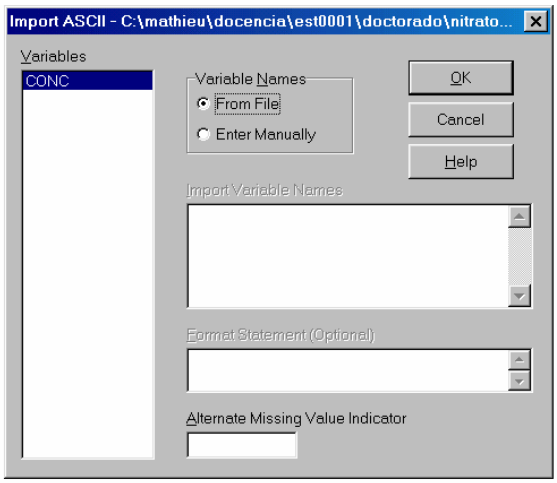
Ahora que hemos introducido los datos, podemos pasar a una primera exploración. Lo haremos con el menu STATISTICS.
Una buena idea es empezar por un histograma para tener una primera impresión visual. Para ello, seleccionamos la opción `HISTOGRAM` del submenú `SUMMARY STATISTICS.`



En el cuadro de la izquierda aparece la lista de las variables que ya tenemos definidas. Basta con seleccionar la variable que nos interesa y pasarla al cuadro ***Histogram Variables*** gracias a la flecha de la derecha. En primer lugar podemos dejar a Statistix la elección de las clases del histograma, y no rellenar el cuadro ***X-Axis.***
Si el resultado no es de nuestro agrado, podemos repetir los pasos especificando el rango de valores del eje de las abscisas (***Low*** y ***High***) y la amplitud de las clases utilizadas en el histograma (***Step***).
A continuación, realizamos un diagrama de cajas-bigotes de los datos. Para ello, seleccionamos la instrucción **Box** and ***Whisker Plot*** del submenú ***SUMMARY STATISTICS.*** Puesto que sólo tenemos una variable, pasamos ***CONC*** al cuadro ***Dependent Variable***, y pinchamos en ***OK***. Utilizamos en particular el diagrama para detectar datos atípicos.
Si nos hemos convencido de qué medidas de centralización y de dispersión son las adecuadas para nuestro conjunto de datos, podemos pedir un informe sobre las medidas numéricas que escojamos. Para ello, seleccionamos la instrucción ***Descriptive Statistics*** del submenú ***SUMMARY STATISTICS,*** pasamos las variables que nos interesan al cuadro ***Descriptive*** variables, y activamos en el cuadro ***Statistics to report*** las casillas correspondientes a las medidas deseadas.

I.3. Importar los datos desde un fichero

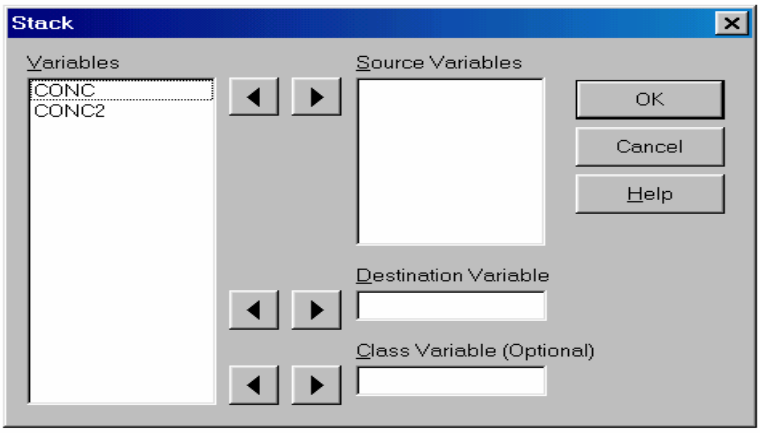
En muchas situaciones, se nos proporcionan los datos en forma de un fichero ASCII. Para trabajar con ellos, debemos importar los datos desde el fichero fuente. Supongamos por ejemplo que, en una segunda sesión, se han medido otras 20 veces el contenido en nitrato de la misma muestra de agua, y que los resultados están en el fichero ***nitrato2.txt***. Al escoger la opción ***IMPORT*** del menu ***FILE***, debemos recorrer las carpetas para encontrar el fichero que buscamos. Lo seleccionamos y aceptamos , aparece la ventana siguiente



En el cuadro de la izquierda aparecen las variables ya definidas. En el cuadro ***Variable Names***, debemos activar la opción que corresponde a nuestra situación:
1) ***From File:*** los nombres de las variables se pueden exportar de la primera línea del fichero fuente.
2) ***Enter Manually***: el fichero fuente no contiene los nombres de las variables sino sólo los datos, y hay que introducir los nombres en el cuadro ***Import Variable Names***.

Con el fichero **nitrato2.txt**, estamos en la segunda situación, activamos por lo tanto la opción **Enter manually**, y propongo que llamemos la nueva variable **Conc2**, pinchamos el botón **OK**. En nuestra hoja de cálculo aparece la nueva columna con sus 20 datos.

Conc y **Conc2** representan verdaderamente valores de la misma variable. Es muy razonable que queramos calcular la media, desviación típica, etc... para el conjunto de datos formado por los valores de las dos variables. Será por lo tanto útil apilar **Conc** y **Conc2** en una única columna, con un nuevo nombre. Lo conseguimos con la opción **`STACK`** del menú **`DATA`**. Aparece la ventana:



Pasamos al cuadro **`Source Variables`**, los nombres de las columnas que queremos apilar, en el cuadro **`Destination Variable`**, escribimos el nombre de la nueva variable que recibirá los valores apilados (¿por qué no llamarla **conctot**?) y en el cuadro de **`Class Variable`**, podemos escribir el nombre de una nueva variable que tomará valores enteros que corresponden al número de la columna original de la que proviene el dato de la variable destino. Propongo que llamemos esta última **Sesion**. Pinchamos en **OK** y observamos el resultado.

Si no os gustan los nombres de las variables que hemos definido, tenéis la posibilidad de renombrarlas con la instrucción **`Rename Variables`** del menu **`DATA`**.

I.4 Nueva exploración de los datos

Ahora que tenemos más datos, queremos repetir la exploración de datos de la primera parte. Realizamos el histograma, el diagrama de cajas-bigotes. ¿Aparecen algunos datos atípicos?

Supongamos que hemos identificado el dato 0.56 que proviene de la segunda sesión como un dato atípico y hemos decidido no tenerlo en cuenta para nuestro análisis. Podemos borrarlo sencillamente, o podemos omitirlo, lo que nos permitirá recuperarlo en cualquier momento. Para omitir datos, utilizamos la instrucción menú **DATA.** En el **cuadro Omit/Select/Restore** expresión, especificamos la condición lógica que debe satisfacer la casilla para que sea omitida. Podemos por ejemplo especificar (Omite todos las filas para las cuales conctot es mayor que 0.55)

```
OMIT(conctot>0.55)
```

O con el mismo resultado (Omite la fila número 46)

```
OMIT(CASE=46)
```

Ya podemos calcular la media, desviación típica etc... de los datos de **conctot** sin el dato atípico 0.56.
A la hora de estudiar nuestros datos, será interesante también comparar las dos sesiones. Para ello, basta con especificar, cuando queremos calcular medias etc... , que los cálculos se deben agrupar según los valores de ***Sesion***. Para ello, pasaremos ***Sesion*** al cuadro de **Grouping Variable**.
Si queremos comparar las dos sesiones con dos diagramas de cajas-bigotes, pasamos la variable **Sesion** al cuadro **Categorical Variable.**
¿Y si queremos hacer un histograma sólo de los datos de la segunda sesión?

Práctica 2. Estadística Descriptiva

En esta práctica vamos a utilizar la opción del menú STATISTICS, dentro de la cual tenemos diversos procedimientos estadísticos:

*** SUMMARY STATISTICS.**
*** ONE, TWO & MULTI-SAMPLE TEST.**
*** LINEAR MODELS.**
*** ASSOCIATION TEST.**
*** RANDOMNESS/NORMALITY TEST.**
*** TIME SERIES.**
*** QUALITY CONTROL.**
*** PROBABILITY DISTRIBUTIONS.**

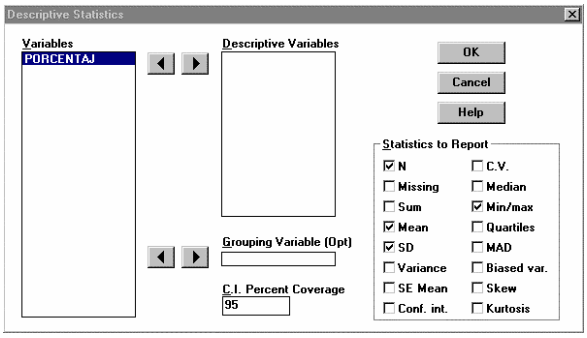
la mayoría de los cuales utilizaremos en las prácticas siguientes.

Para la resolución de problemas de ***Estadística Descriptiva***, nosotros vamos a ver el funcionamiento de la primera opción **SUMMARY STATISTICS**, que está compuesta por los procedimientos siguientes

- **DESCRIPTIVE STATISTICS**
- **FREQUENCY DISTRIBUTION**
- **HISTOGRAM**
- **STEM AND LEAF PLOT**
- **PERCENTILES**
- **BOX AND WHISKER PLOT**
- **ERROR BAR CHART**
- **CROSS TABULATION**
- **SCATTER PLOT**
- **BREAKDOWN**

Estos son los procedimientos implementados en el programa para realizar el análisis descriptivo de los datos. Nosotros sólo utilizaremos algunas de ellas.

Descriptive Statistics

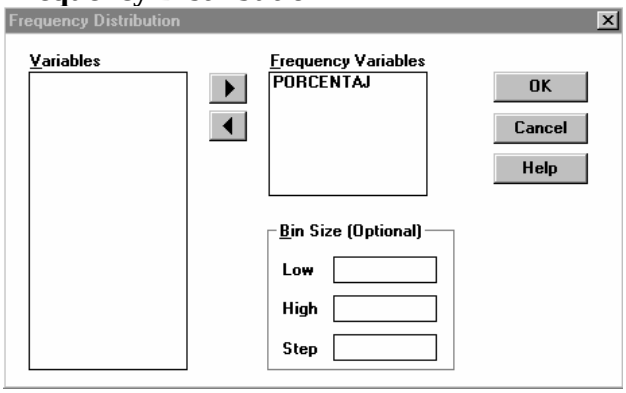


En primer lugar, tendremos que seleccionar las variables descriptivas, es decir, las variables de las que queremos obtener alguna medida descriptiva.

Observar que las medidas **VARIANCE** y **SD** (standard deviation) son las conocidas como cuasivarianza y cuasidesviación típica, es decir, se divide por **n-1** en lugar de por **n.**

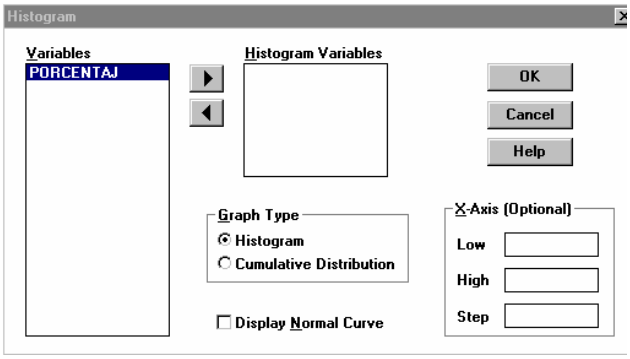
Además, se puede considerar una variable de índices que permite hallar estas medidas según los distintos casos de esta, es decir, sin desapilar los datos.

Frequency Distribution



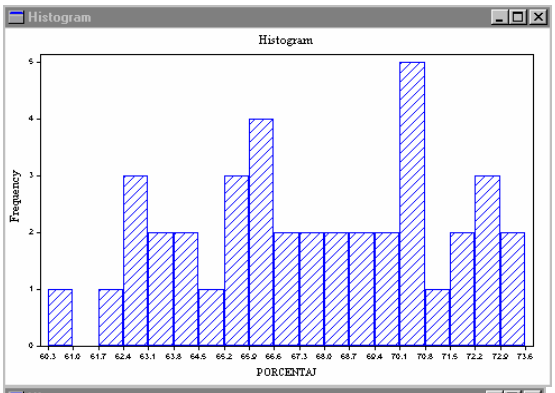
Esta sentencia obtiene las frecuencias absolutas y acumuladas de la variable o variables que se indiquen, bien considerando los datos sin agrupar (de tipo discreto) o agrupados por intervalos, necesitando, en este caso, introducir el recorrido de los datos o variable (Low: menor, High: mayor) y la amplitud de los intervalos (Step).

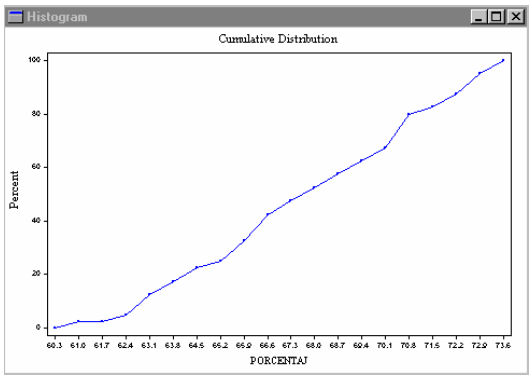
Histogram

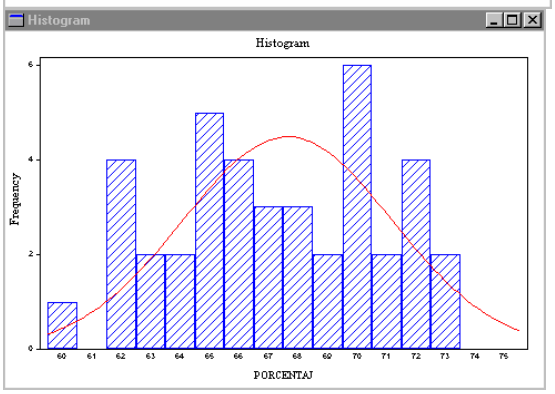


Con esta opción podemos representar los histogramas de las frecuencias de las variables, bien seleccionando los valores menor y mayor sobre el eje OX y la amplitud de los intervalos (como en el caso de las frecuencias), o bien, el programa selecciona el mínimo y máximo valor y una amplitud por defecto.

Como se observa en las gráficas, también permite la representación de las frecuencias acumuladas.

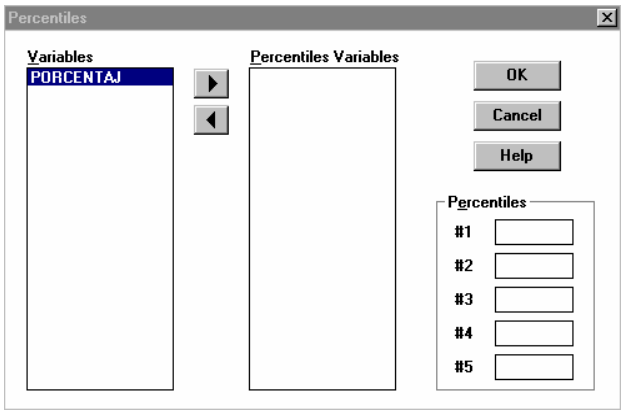






Además, nos permite seleccionar la representación de una curva correspondiente a la distribución normal en los datos de la variable, con parámetros las medidas muestrales de los mismos.

Percentiles



Esta opción permite calcular los percentiles de las variables que se deseen.

Recordar que el percentil 50 es la mediana, el primer cuartil es el percentil 25 y el tercer cuartil es el percentil 75.

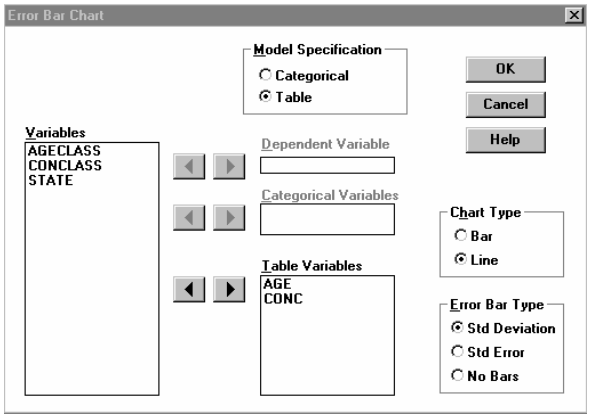
Como se presenta en los resultados, puede calcular hasta 5 percentiles de cada variable:

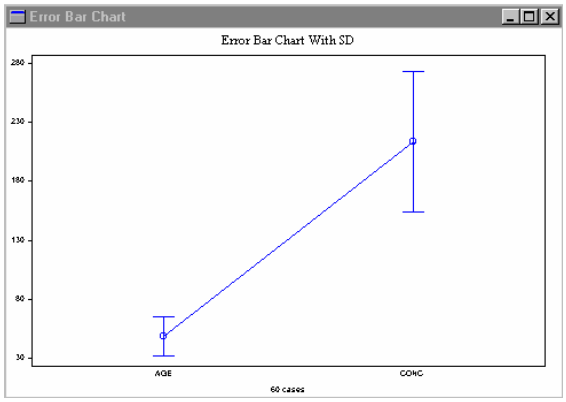
PERCENTILES

VARIABLE:	25.0	50.0	75.0	30.0	10.0
PORCENTAJ	65.050	67.650	70.350	65.290	62.810

Error Bar Chart

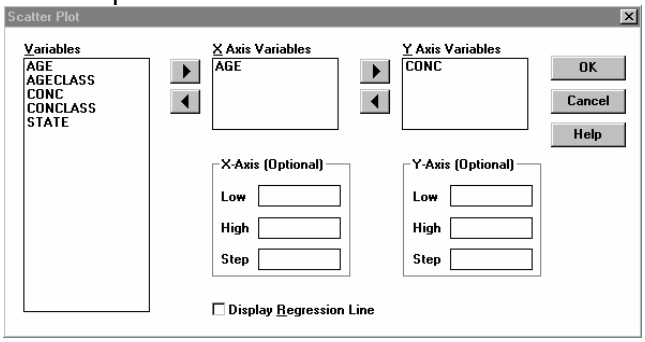
Esta opción obtiene gráficas que pueden utilizarse para comparar varias variables, señalizando la media de las variables por un círculo o una barra de altura igual a la media, y segmentos centrados en la media con amplitud determinada por la desviación estándar (llamada cuasidesviación típica) o por el error estándar de la media.

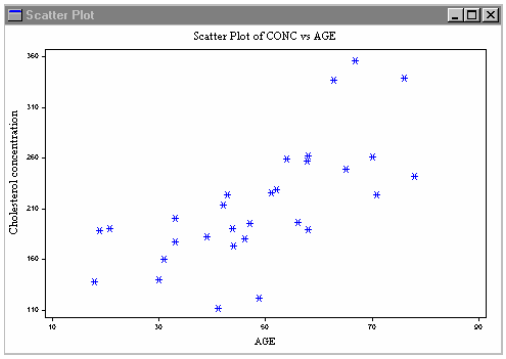




Scatter Plot

Esta opción permite representar gráficamente la nube de puntos en el plano de una variable bidimensional, es decir, seleccionar una variable como variable del eje X y otra como variable del eje Y, además puede indicarse los valores mínimos y máximos para utilizar en los ejes de la gráfica, así como la representación de la recta de regresión que mejor ajusta mediante mínimos cuadrados a la nube de puntos.





EJERCICIO:

Los siguientes datos corresponden con la realización de una muestra de tamaño 23 de una variable X:

105, 135, 148, 160, 194, 154, 183, 169, 196, 180, 150, 157,
131, 146, 211, 110, 190, 218, 171, 163, 121, 165, 178 .

1.- Introducir los datos en un fichero llamado **prac0.sx**

2.- Construir la tabla de frecuencias una vez seleccionadas las clases.

3- Realizar el histograma de frecuencias absolutas.

4.- Calcular las siguientes medidas de localización:

a) Media muestral.

b) Mediana muestral.

c) Primer y tercer cuartil.

d) Percentil 90.

5.- Calcular las siguientes medidas de dispersión:

a) Varianza muestral.

b) Desviación típica muestral.

6.- Crear una nueva variable Y ($=X^2$). Para ello, se utilizará el comando `Transformations` del menu `Data`, escribiendo en el cuadro Y=X^2. Representar la nube de puntos asociada a los pares (X,Y).