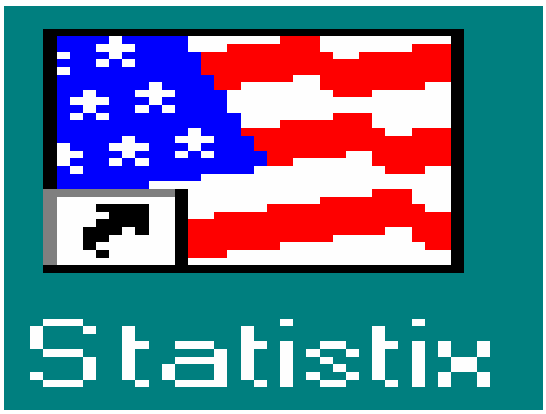
Cuadernillo de Prácticas con Statistix



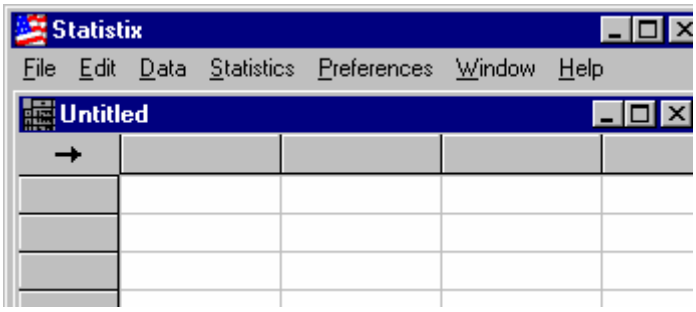
Departamento de Matemática Aplicada y Estadística
Área de Estadística e Investigación Operativa
Universidad Politécnica de Cartagena

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***.

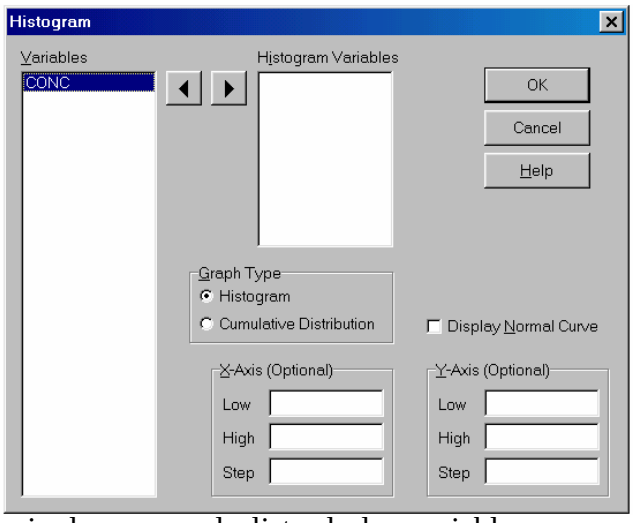
Se aconseja guardar la tabla de datos en un fichero después de la introducción de datos. Para ello, se usa la opción "***SAVE***" o "***SAVE AS***" del menú ***FILE***. Al igual que cualquier programa Windows, se puede recorrer las carpetas para decidir donde guardar el fichero.
Guardar la tabla de datos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

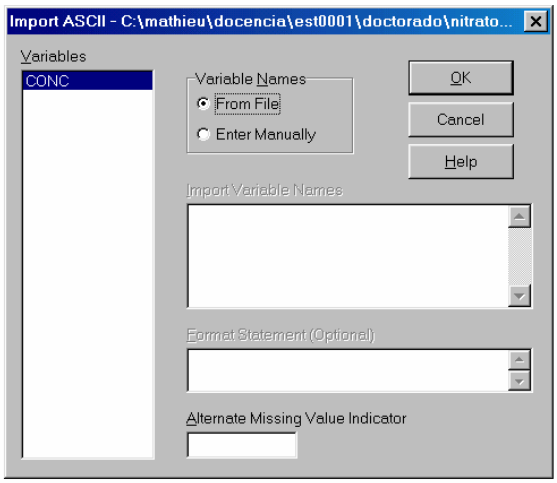
Ahora que hemos introducido los datos, podemos pasar a una primera exploración. Lo haremos con el menu STATISTICS.
Una buena idea es empezar por un histograma para tener una primera impresión visual. Para ello, seleccionamos la opción `HISTOGRAM` del submenú `SUMMARY STATISTICS.`



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

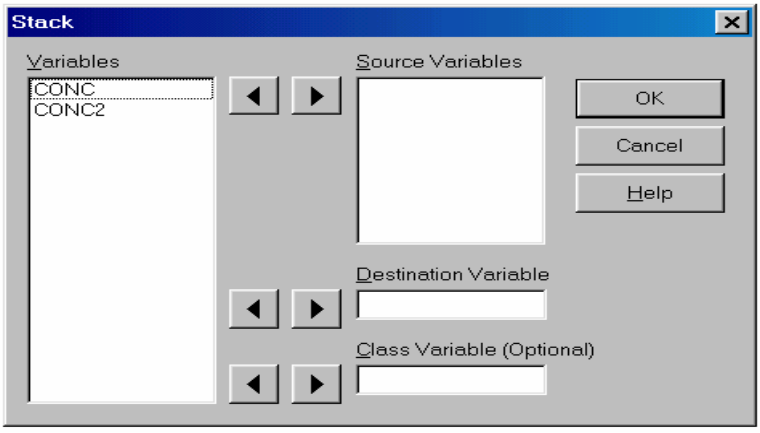
En muchas situaciones, se nos proporcionan los datos en forma de un fichero ASCII. Para trabajar con ellos, debemos importar los datos desde el fichero fuente. Supongamos por ejemplo que, en una segunda sesión, se han medido otras 20 veces el contenido en nitrato de la misma muestra de agua, y que los 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.**

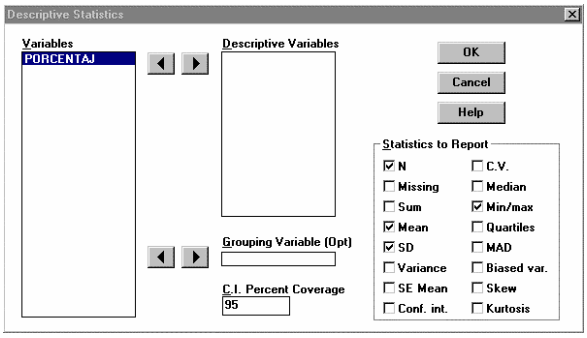
la mayoría de los cuales utilizaremos en las prácticas siguientes.

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

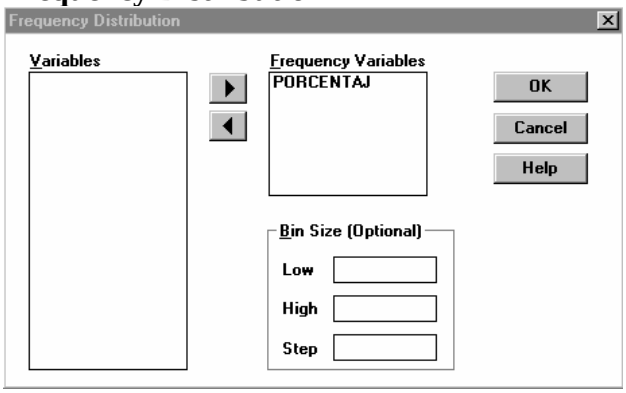


En primer lugar, tendremos que seleccionar las variables descriptivas, es decir, las variables de las que queremos obtener alguna medida descriptiva.

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.**

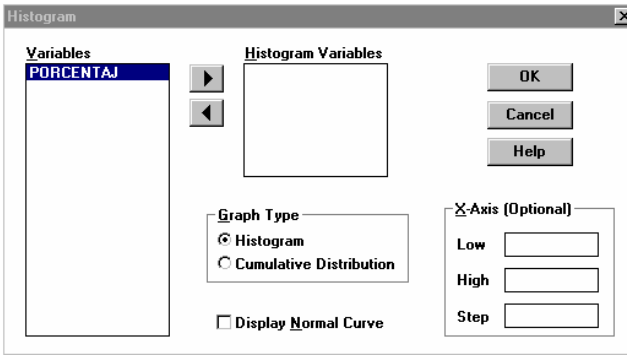
Además, se puede considerar una variable de índices que permite hallar estas medidas según los distintos casos de esta, es decir, sin desapilar los datos.

Frequency Distribution



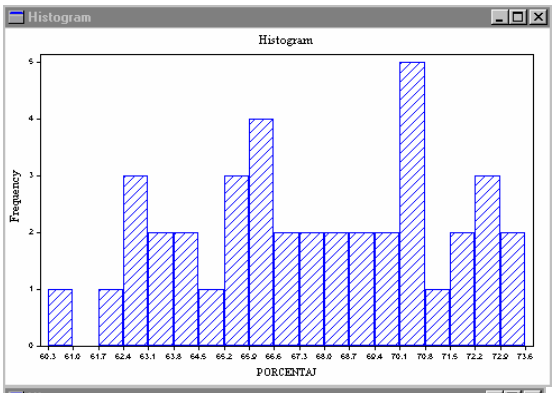
Esta sentencia obtiene las frecuencias absolutas y acumuladas de la variable o variables que se indiquen, bien considerando los datos sin agrupar (de tipo discreto) o agrupados por intervalos, necesitando, en este caso, introducir el recorrido de los datos o variable (Low: menor, High: mayor) y la amplitud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

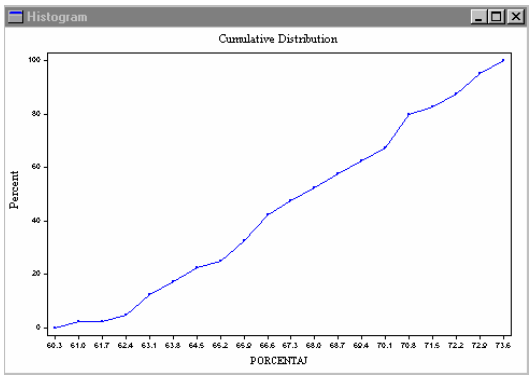
Histogram

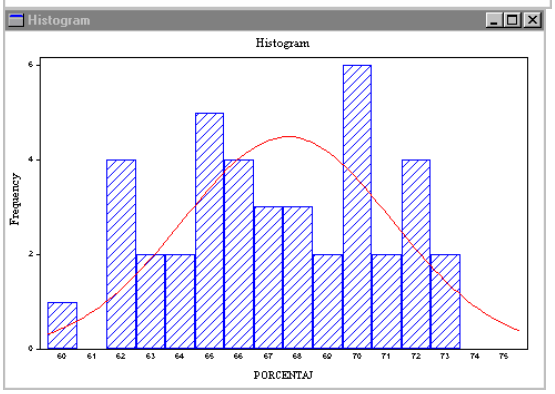


Con esta opción podemos representar los histogramas de las frecuencias de las variables, bien seleccionando los valores menor y mayor sobre el eje OX y la amplitud de los intervalos (como en el caso de las frecuencias), o bien, el programa selecciona el mínimo y máximo valor y una amplitud por defecto.

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

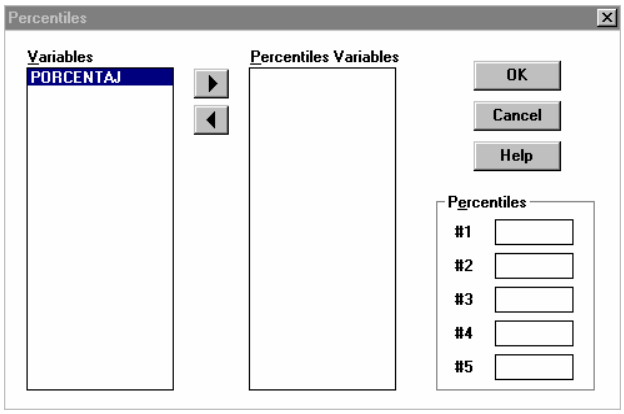






Además, nos permite seleccionar la representación de una curva correspondiente a la distribución normal en los datos de la variable, con parámetros las medidas muestrales de los mismos.

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.

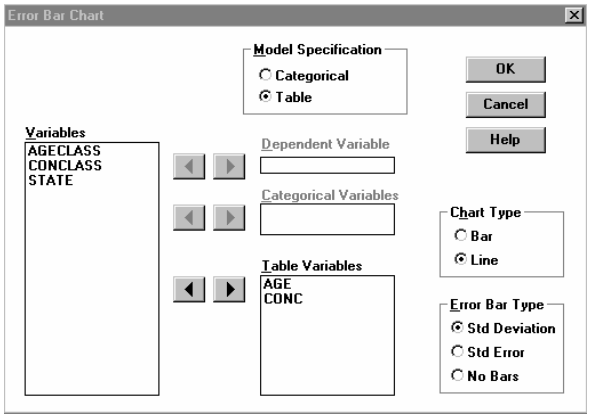
Como se presenta en los resultados, puede calcular hasta 5 percentiles de cada variable:

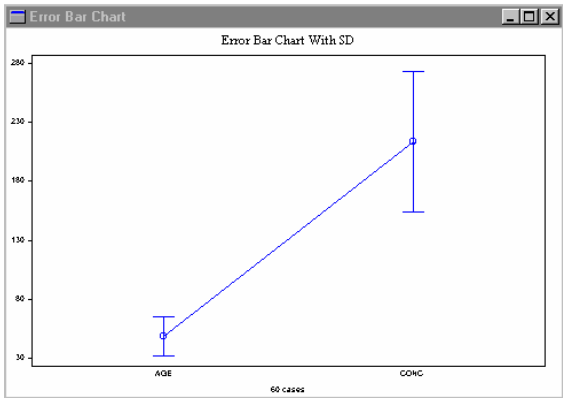
PERCENTILES

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

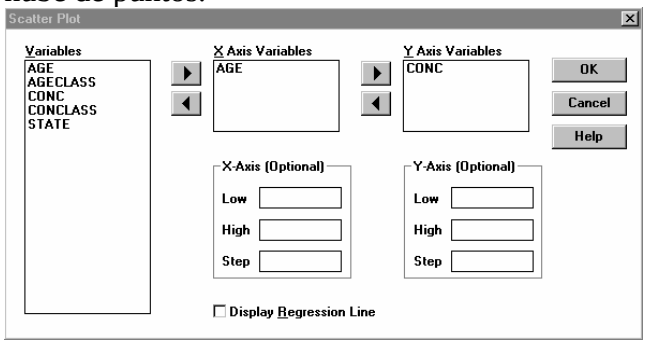
Esta opción obtiene gráficas que pueden utilizarse para comparar varias variables, señalizando la media de las variables por un círculo o una barra de altura igual a la media, y segmentos centrados en la media con amplitud determinada por la desviación estándar (llamada cuasidesviación típica) o por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

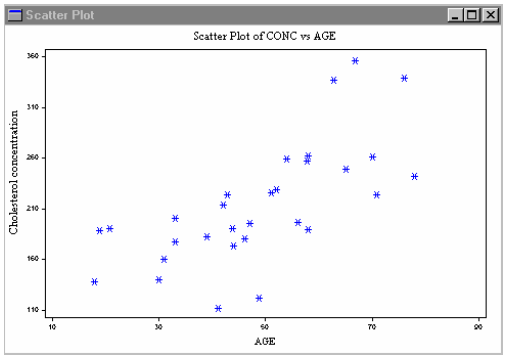




Scatter Plot

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.

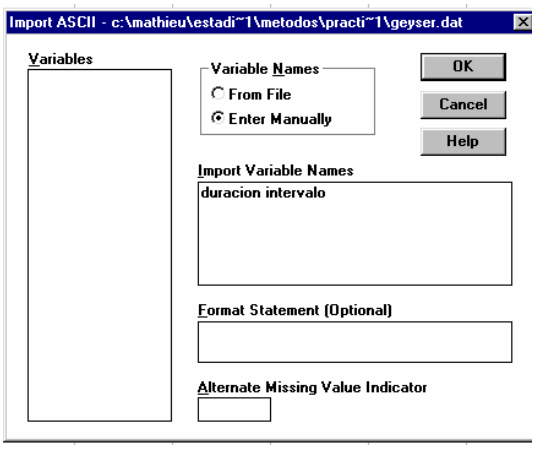
6.- Crear una nueva variable Y ($=X^2$). Para ello, se utilizará el comando `Transformations` del menu `Data`, escribiendo en el cuadro Y=X^2. Representar la nube de puntos asociada a los pares (X,Y).

Práctica 3. Explorando datos con Statistix

En esta práctica aprenderemos a explorar un conjunto de datos utilizando el menú Statistics de nuestro programa.

Un geyser es un nacimiento de agua hirviente que de vez en cuando se vuelve inestable y expulsa agua y vapor. El geyser "Old Faithful" en el parque de Yellowstone en Wyoming es probablemente el más famoso del mundo. Los visitantes del parque se acercan al emplazamiento del geyser intentando no tener que esperar demasiado para verlo estallar. Los servicios del Parque colocan un cartel donde se anuncia la próxima erupción. Es por lo tanto de interés estudiar los intervalos de tiempo entre dos erupciones conjuntamente con la duración de cada erupción.
En esta práctica analizaremos los intervalos entre erupciones sucesivas así como la duración de las mismas durante los meses de agosto 1978 y agosto 1979. Se observaron 222 erupciones y los datos de los que disponemos se presentan por pares: (duración de la erupción, intervalo hasta la siguiente). Las unidades de medición son mn.

Para importar los datos en nuestra hoja de cálculo, seleccionamos la opción ***import*** del menu ***File***. A continuación debemos recorrer las carpetas para encontrar el fichero **geyser.dat** (vuestro profesor os dirá en qué carpeta se encuentra). Una vez que hemos localizado el fichero, pulsamos ***OK*** y aparece la ventana siguiente

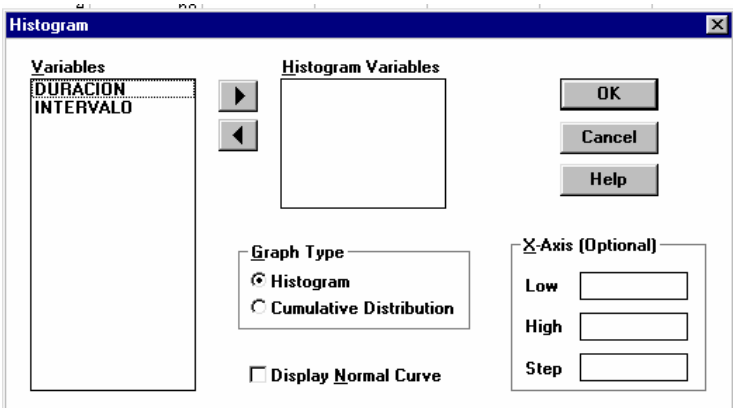


En el cuadro **Variable Names**, escogemos la opción "**Enter manually**" puesto que el fichero de datos no contiene los nombres de la variable, mientras que en el cuadro "**Import Variable Names**" introducimos **Duracion** e **Intervalo** que serán los nombres de nuestras variables. Después de pulsar **OK**, tenemos en nuestra hoja de cálculo los 444 datos.

Podemos empezar con la exploración de los datos: tal como se vio en clase, el primer paso cuando uno dispone de varias variables es empezar por estudiar cada una por separado.

1. Estudio individual de cada variable

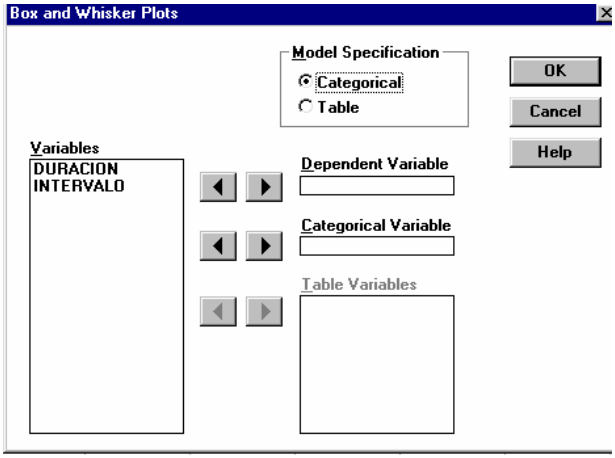
En particular, para hacerse una idea de la distribución de los datos, vamos a realizar un histograma tanto para la duración como para el intervalo. Para ello, seleccionamos la opción **`Statistics->Summary Statistics->Histogram`** y aparece la ventana siguiente:



Por ejemplo, pasamos la variable "***duracion***" desde la izquierda hasta el cuadro "**Histogram variables**". Si no especificamos nada más en el cuadro **X-axis**, el programa realizará de manera automática la elección de clases. Pulsamos por lo tanto **OK**, y aparece el histograma de la duración. Minimizamos la ventana y repetimos los pasos con la variable "**intervalo**".
Describir las características globales de cada histograma:

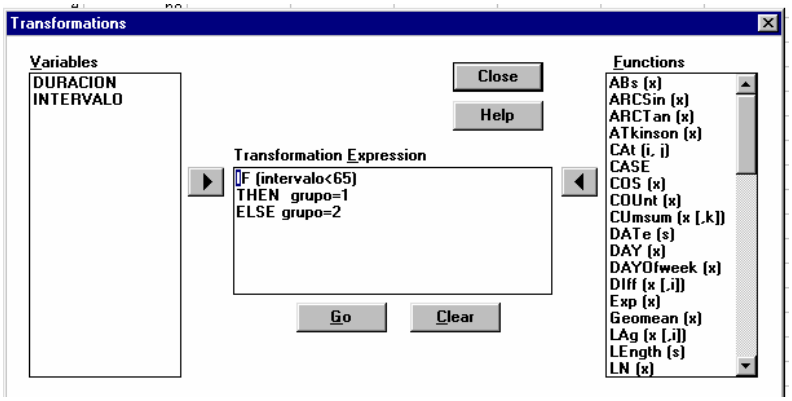
¿Puedes indicar medidas convenientes de centralización y de dispersión de los datos?

Pasamos a realizar un diagrama de caja-bigotes para, por ejemplo, los datos de la duración: Escogemos ***Statistics->Summary Statistics->Box and Whisker plots*** y aparece la ventana siguiente:

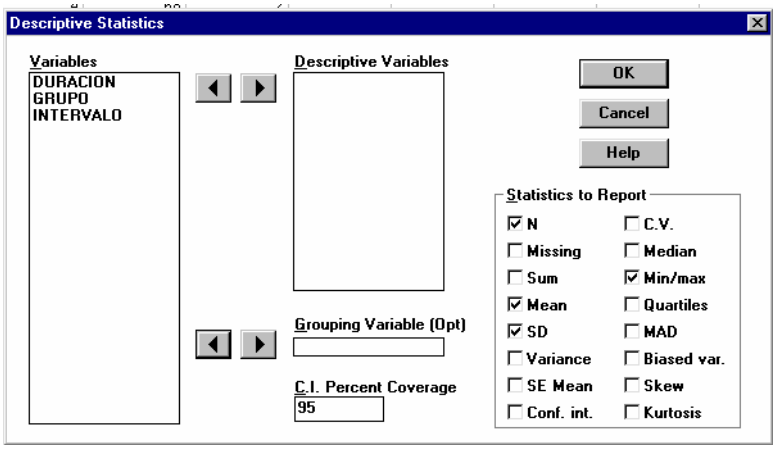


Pasamos la variable ***Duracion*** de la izquierda al cuadro "**Dependent Variable**" y pulsamos **OK**. Como podemos ver, en este caso, el diagrama de cajas y bigotes es un resumen muy expeditivo que oculta la estructura de los datos .

Decidimos estudiar con más detalle los intervalos entre dos erupciones y decidimos separar los dos subgrupos que hemos detectado en el histograma. Empezamos por decidir de un punto de corte, por ejemplo 65mn. Vamos a crear una variable llamada grupo, que valga 1 si el intervalo es menor de 65mn y 2 si es mayor o igual. Para ello, utilizamos la opción **Data->Transformations**, y rellenamos el cuadro "**Transformation expression**" con la siguiente expresión lógica:



Podemos ahora calcular medidas numéricas descriptivas para cada uno de los subgrupos. Para ello, seleccionamos **Statistics->Summary Statistics->Descriptive Statistics**, y nos aparece el cuadro:

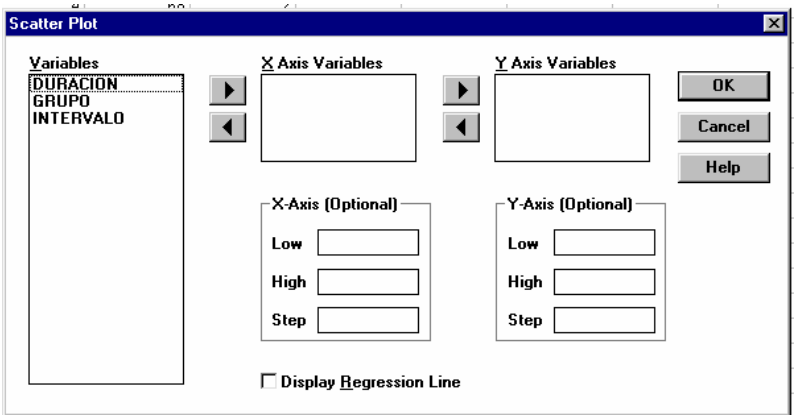


Queremos obtener medidas para dos grupos de valores de "**INTERVALO**". Pasamos por lo tanto la variable INTERVALO al cuadro "**Descriptive variables**" y la variable "**GRUPO**" al cuadro "**Grouping variable (Opt)**". (Si no especificáramos nada en ***Grouping variable***, obtendríamos las medidas para todos los valores de INTERVALO) Seleccionamos en el cuadro "***Statistics to report***", las medidas que nos interesan. Pulsamos **OK** y obtenemos estas últimas.
Para el grupo 1 de INTERVALO:
Media= Desviación típica=
Para el grupo 2 de INTERVALO:
Media= Desviación típica=

Podríamos por supuesto realizar algo parecido para la variable ***DURACION***, decidiendo un punto de corte etc...

2. Estudio conjunto de las dos variables

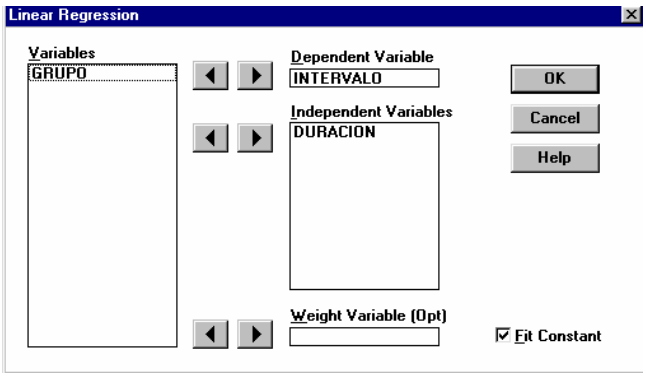
Queremos ahora estudiar la relación que pueda existir entre **DURACION** e **INTERVALO**, en particular, ¿podemos decir que una duración corta implica un intervalo corto o algún tipo de efecto contrario?
Una vez más, lo primero que hay que hacer es conseguir una impresión visual de los datos realizando una nube de puntos. Para ello, seleccionamos la opción ***Statistics-> Summary Statistics -> Scatter plot.***



y pasamos la ***duracion*** al cuadro ***X Axis Variable*** e ***INTERVALO*** al cuadro ***Y Axis Variables***. Pulsamos **OK**, y obtenemos la nube de puntos.

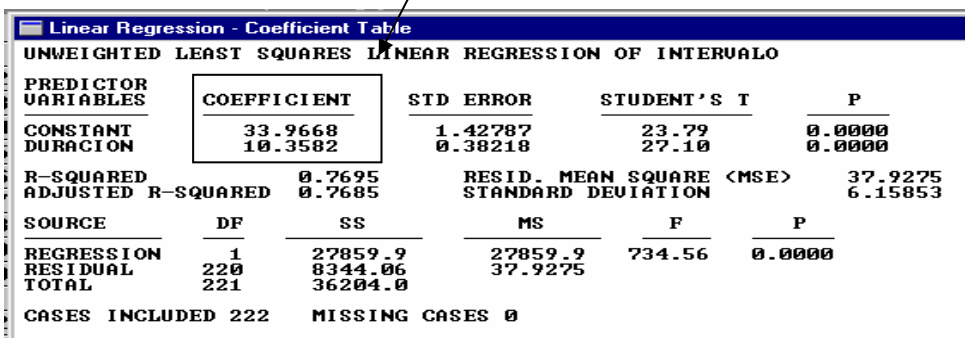
¿Qué tipo de relación existe entre el intervalo de tiempo hasta la siguiente erupción y la duración de la última? ¿Cuál podría ser un modelo teórico?

Si decidimos ajustar una recta a la nube de puntos, podemos conseguir de manera automática los coeficientes de la recta:
Seleccionamos ***Statistics->Linear Models->Linear regression***,



y pasamos ***INTERVALO*** como variable dependiente, y ***DURACION*** como variable independiente. La casilla "***Fit constant***" corresponde a si queremos que calcule la ordenada al origen o si forzamos la recta por el origen. En este caso, la mantenemos activada. Al pulsar **OK**, obtenemos la tabla siguiente:

Coeficientes de la recta de regresión



Linear Regression - Coefficient Table

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF INTERVALO

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	33.9668	1.42787	23.79	0.0000
DURACION	10.3582	0.38218	27.10	0.0000

R-SQUARED	0.7695	RESID. MEAN SQUARE (MSE)	37.9275
ADJUSTED R-SQUARED	0.7685	STANDARD DEVIATION	6.15853

SOURCE	DF	SS	MS	F	P
REGRESSION	1	27859.9	27859.9	734.56	0.0000
RESIDUAL	220	8344.06	37.9275		
TOTAL	221	36204.0			

CASES INCLUDED 222 MISSING CASES 0

Deducimos que podemos proponer como modelo teórico para explicar el intervalo de tiempo entre dos erupciones en función de la duración de la última erupción :

INTERVALO= 33.97+10.36*DURACIÓN.

En cuanto a la bondad del ajuste, se lee el coeficiente de determinación R^2 en frente de "R-SQUARED". En nuestro caso, encontramos un coeficiente de determinación igual a 0.77, lo que indica que nuestro ajuste es aceptable.

Gracias a un modelo como éste, los servicios del parque son capaces de predecir con una precisión satisfactoria, después de una erupción, cuándo será la siguiente.

Práctica 4 : Resolución de Problemas con Statistix: Ajuste por Mínimos Cuadrados.

¡Error! Marcador no definido.

Veamos como resolver el siguiente problema utilizando el programa Statistix:

Se ha realizado un estudio para investigar el efecto de un determinado proceso térmico en la dureza de una determinada pieza. Once piezas se seleccionaron para el estudio. Antes del tratamiento se realizaron pruebas de dureza para determinar la dureza de cada pieza. Después, las piezas fueron sometidas a un proceso térmico de templado con el fin de mejorar su dureza. Al final del proceso, se realizaron nuevamente pruebas de dureza y se obtuvo una segunda lectura. Se recogieron los siguientes datos (Kg. de presión):

¡Error! Marcador no	1	2	3	4	5	6	7	8	9	10	11
Dureza previa	**182**	**232**	**191**	**200**	**148**	**249**	**276**	**213**	**241**	**480**	**262**
Dureza posterior	**198**	**210**	**194**	**220**	**138**	**220**	**219**	**161**	**210**	**313**	**226**

a) Calcular la dureza media antes y después del proceso. Así como las desviaciones típicas en cada caso.
b) Realizar un diagrama de caja bigotes para la dureza previa y la dureza posterior.
c) Estudiar el ajuste de mínimos cuadrados del nivel posterior con respecto al nivel previo de dureza.
d) Estudiar la precisión del ajuste anterior.

Resolución:

Empezamos por definir gracias al menu ***Data->insert->Variables***, dos nuevas variables **Previa** y **posterior**. A continuación introducimos los valores de estas variables.

a) Para calcular medidas numéricas asociadas a cada variable, utilizamos las nociones vistas en la práctica anterior: seleccionamos la opción

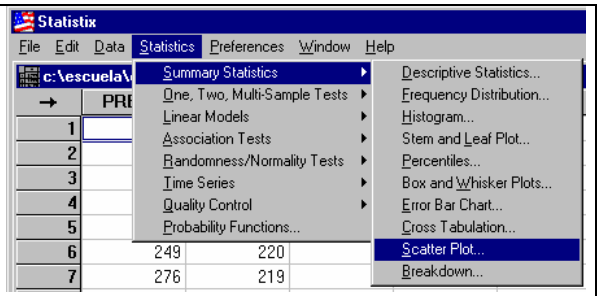
Statistics->Summary Statistics->Descriptive Statistics.

b) Igualmente, para realizar un diagrama de cajas-bigotes, seleccionamos, tal como lo vimos en la última práctica, la opción ***Statistics->Summary Statistics->Box and Whisker Plots***. Especificamos el modelo en forma de tabla ("***Table***") y pasamos las variables X e Y al cuadro "***Table variables***".

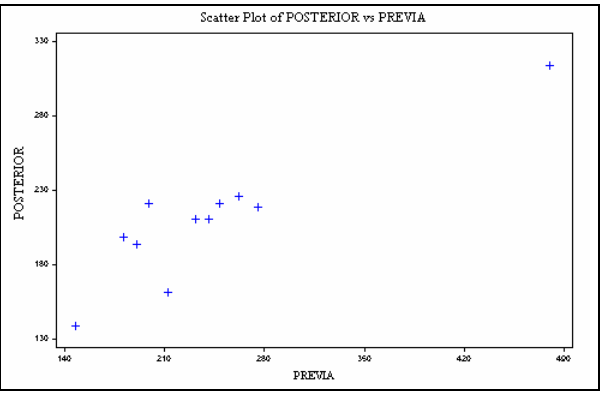
c)

Representaremos en primer lugar el gráfico de dispersión o también llamado nube de puntos con el fin determinar si existe una cierta tendencia lineal. Para ello seleccionaremos:

Statistics->Sum. Statistics->Scatter Plot.

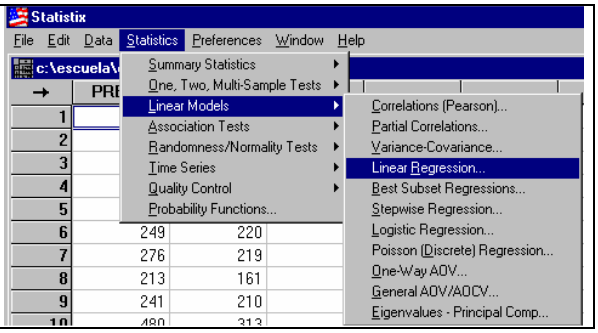


Como se observa en la gráfica, existe un punto demasiado alejado (se corresponde con los resultados de la pieza 10, (480,313)) que en principio puede dar un resultado engañoso sobre la dependencia lineal de ambas características.



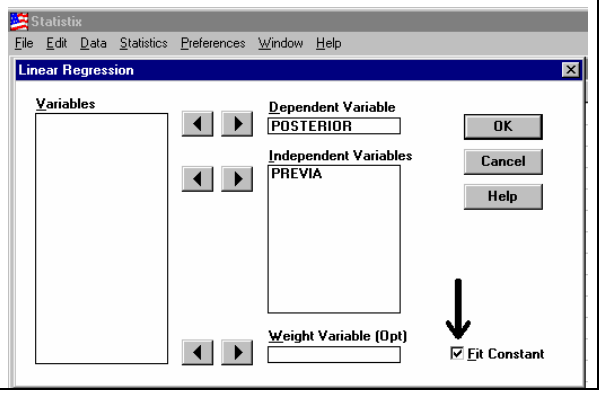
En cualquier caso, pasemos a calcular la ecuación de la recta de regresión, para lo cual seleccionaremos:

Statistics-> Linear models-> ->Linear Regresion



Indicaremos como variable independiente la variable **Previa** y como dependiente la variable Posterior puesto que pretendemos estudiar la dureza **Posterior** en función de la dureza previa.

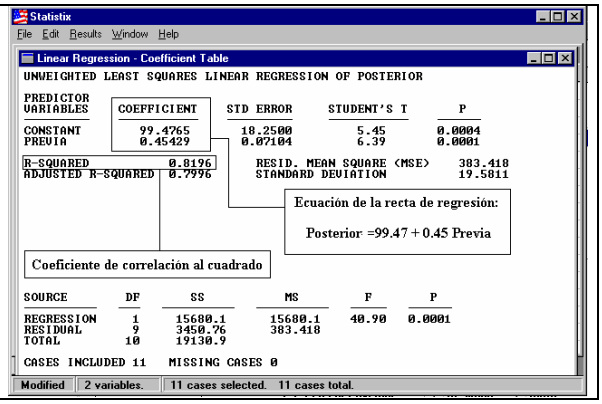
Observar que en la parte inferior aparece una casilla con el indicador Fit Constant**. Esta casilla** debe estar marcada **cuando se ajusta una recta de la forma y=ax+b, pero se debe desactivar para una recta forzada por el origen: y=ax.**



Como podemos observar, Statistix nos proporciona la ecuación de la recta ajustada y el coeficiente de correlación al cuadrado:

$$Y = 99.47 + 0.45\ X$$
$$R^2 = 0.81$$

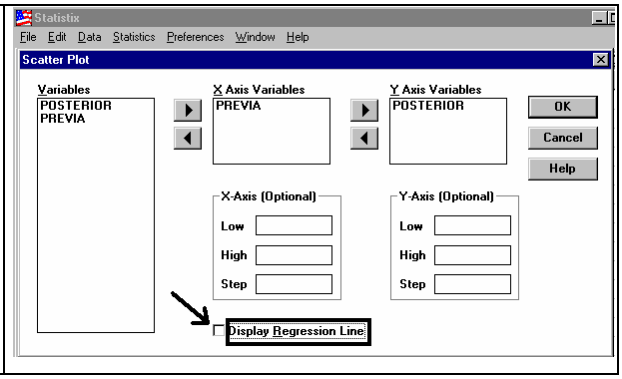
junto con otra información que comentaremos en posteriores prácticas.



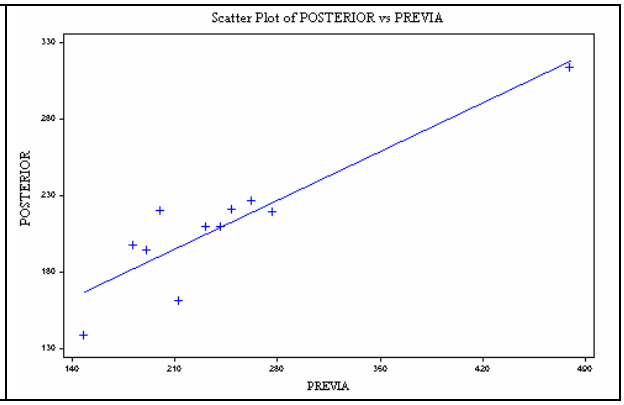
Si recordamos, al seleccionar:

Statistics-> Summary Statistics->Scatter Plot.

tenemos la posibilidad de presentar junto con el diagrama de dispersión la recta de regresión con el fin de observar si existe algún valor que presente un gran residuo.



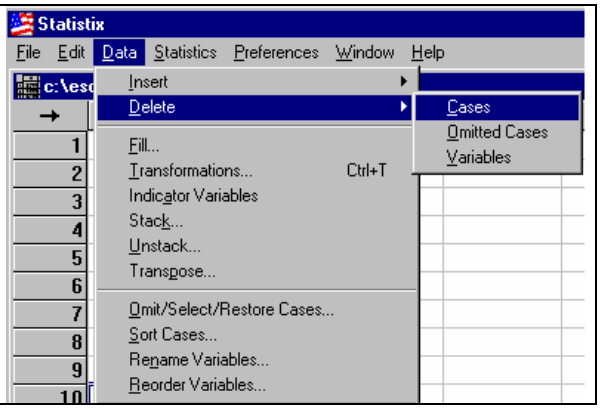
La gráfica que obtenemos es:



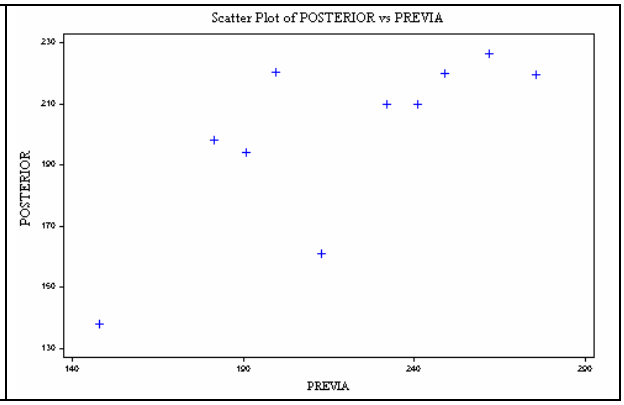
Veamos que ocurre si eliminamos el dato número 10, el nuevo gráfico de dispersión no presenta una clara tendencia lineal.

Para eliminar el dato seleccionaremos:

<u>Data -> Delete -> Cases</u>

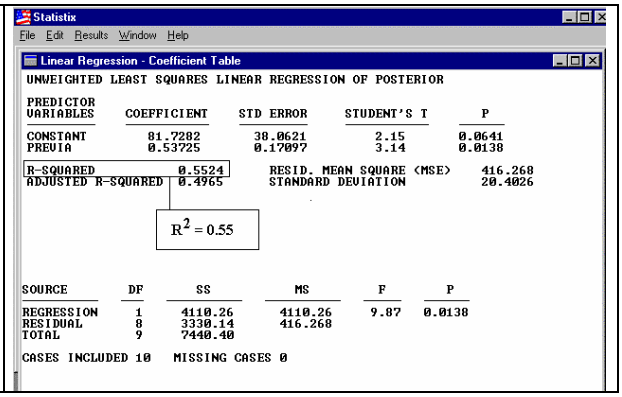


Una vez eliminado dicho valor, el gráfico de dispersión indica que no parece existir una clara tendencia lineal entre los puntos.



Como se observa, si realizamos el ajuste obtenemos un coeficiente de correlación bastante bajo:

R^2=0.55



Conclusión:

Si el valor correspondiente a la pieza 10 no se descarta, se obtiene un coeficiente de determinación de 81%, lo que indica un buen ajuste lineal. Sin embargo, si el valor obtenido para la pieza 10 resulta erróneo y lo descartamos, ,se obtiene un coeficiente de determinación de 55%, lo que pone en duda la validez de un ajuste lineal entre los resultados obtenidos antes y después del proceso de templado.

Ejercicios propuestos
Resolución con Statistix

1) Las materias primas empleadas en la producción de una fibra sintética son almacenadas en un lugar en donde no se tiene control de la humedad. La siguiente tabla refleja en porcentajes la humedad relativa del almacén X y la humedad observada en la materias primas Y durante un estudio que tuvo lugar durante 12 días.

X	41	53	59	65	71	78	50	65	74
Y	1.6	13.6	19.6	25.6	31.6	33.2	14.7	21.2	28.3

a) Realizar un ajuste de mínimos cuadrados entre ambas variables.
b) Estudiar la precisión del ajuste en función del valor obtenido por el coeficiente de correlación, representar gráficamente la recta hallada y comentar los resultados.

2) Con el fin de determinar si existe relación entre la cantidad de polímeros de látex incluida durante el proceso de mezclado de cemento Portland y su resistencia adhesiva a tensión, una empresa encargada de realizar certificaciones de obras toma una muestra de tamaño 10, obteniendo los siguientes resultados

X	13.5	11.0	13.0	11.2	12.0	13.2	12.0	13.5	11.2	13.0
Y	17.5	16.6	17.2	16.6	17.0	17.3	16.9	17.3	16.8	17.1

a) Calcular la media y varianza asociada a cada una de las variables.
b) Calcular la covarianza existente entre ambas variables así como el coeficiente de correlación.
c) Realizar un ajuste por mínimos cuadrados de la resistencia respeto a la cantidad de polímeros añadida en la mezcla.

3) La hidrólisis de un cierto éster tiene lugar en medio ácido según un proceso cinético de primer orden. Partiendo de una concentración inicial de $3.^{10\text{-}2}$ M del éster, se han medido las concentraciones del mismo a diferentes tiempos obteniéndose los resultados siguientes.

T (mn)	3	4	10	15	20	30	40	50	60	75	90
C 10^{-3}(M)	25.5	23.4	18.2	14.2	11	6.7	4.1	2.5	1.5	0.7	0.4

a) Realice una nube de puntos de las dos variables. ¿ Le parece adecuado un modelo lineal para escribir este conjunto de datos?
b) Defina una nueva variable Y' que sea Y'=ln (concentración) y realizar la nube de puntos Y' en función de t.
c) Realizar un ajuste por mínimos cuadrados de Y' sobre t con un modelo del tipo: y=ax+b. ¿Cuál es el modelo teórico que propone para C en función del tiempo?
d) Nos dan la información adicional de que se sabe con exactitud que la concentración inicial para T=0 era igual a 30.10^{-3}M. ¿Cómo podemos incluir esta información en nuestro modelo?

Práctica 5. Modelo Normal

Una cooperativa decide invertir en una calibradora automática de naranjas. Existen en el mercado diversos tipos de modelos que la cooperativa clasifica según el rango de diámetro D de naranjas que pueden recoger :

- Tipo 1 : 6cm $\leq$ D $\leq$ 8cm
- Tipo 2 : 6cm $\leq$ D $\leq$ 10cm
- Tipo 3 :4cm $\leq$ D $\leq$ 10cm

El precio de los modelos es mayor si el rango de diámetros posibles es mayor. La cooperativa quiere ser asesorada sobre qué tipo de calibradora comprar. El primer paso consiste en describir la situación de acuerdo con las nociones teóricas que hemos visto en clase: podemos introducir el experimento aleatorio *"se escoge al azar una naranja en la producción de las fincas de los miembros"*, el espacio muestral es por lo tanto toda la producción de las fincas. Se define la variable aleatoria continua D como el diámetro de la naranja escogida. Lo que quiere saber la cooperativa es qué proporción de naranjas de la producción tiene un diámetro comprendido entre 6 y 8cm, entre 6 y 10cm, y entre 4 y 10cm. Por lo tanto lo que necesita saber son las siguientes probabilidades: Pr(6 $\leq$D $\leq$8); Pr (6 $\leq$D $\leq$10) y Pr (4 $\leq$D $\leq$10).

La cooperativa realiza un estudio preliminar y coge una muestra de 1000 naranjas en el momento de la recogida en las fincas de los miembros. Eso se puede traducir de la manera siguiente : puesto que es imposible para la cooperativa el medir el diámetro de todas las naranjas de la producción, se contenta con realizar el experimento aleatorio 1000 veces y dispone de los valores de D para esas 1000 realizaciones del experimento. Su esperanza es que las conclusiones que se puedan sacar a partir de la muestra sigan válidas para toda la producción.

El objetivo de esa práctica es doble :

1. Realizar el estudio descriptivo de la muestra recogida.
2. Proponer un modelo para la distribución de la v.a. D en toda la producción, e investigar la validez de ese modelo a partir de los datos de la muestra.

Finalmente deberemos concluir la práctica asesorando a la cooperativa.

Resolución

1. Estudio descriptivo de la muestra recogida

a.- En primer lugar debemos cargar en nuestro hoja de cálculo de Statistix los datos de la muestra. Estos se encuentran en un fichero ASCII llamado:

diam.dat

ubicado en nuestro directorio de prácticas */prácticas/estadística/datos/*.

Antes de importar los datos desde el fichero ***diam.dat*** puede ser conveniente visualizar el fichero para asegurarse de que los datos están bien presentados. Para ellos seleccionamos:

File -> View Ascii File

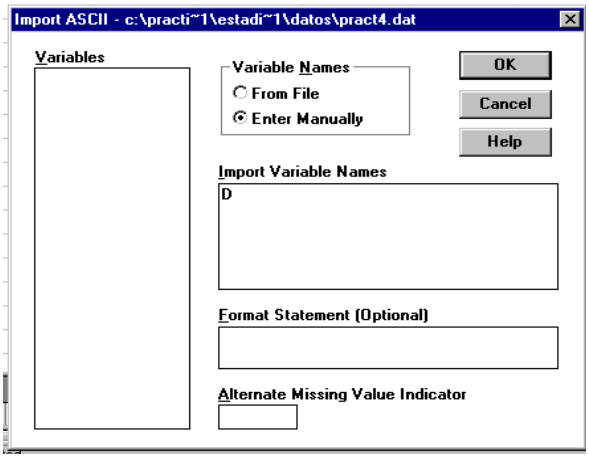
y en el cuadro de diálogo que aparece, nos movemos en las carpetas hasta llegar a */prácticas/estadística/datos/*, seleccionamos ***diam.dat*** y aceptamos. Al visualizar el fichero ***diam.dat***, podemos comprobar en particular que sólo contiene datos.

Podemos ahora pasar a importar los datos desde el fichero ***diam.dat***, para ello seleccionamos

File -> Import

Aparece un cuadro de diálogo y recorriendo la estructura de carpetas del disco duro, llegamos hasta la carpeta */prácticas/estadística/datos/*. Después de haber seleccionado ***diam.dat***, pinchamos **"Aceptar"** .

En el cuadro de diálogo, al contener el fichero ***diam.dat*** sólo los datos y no el nombre de la variable, debemos activar la opción "***Enter manually***" y especificar en el cuadro "***Import variable names***" el nombre de la variable a la que queremos asignar los datos importados, por ejemplo ***D***.



b.- Establecer la tabla de frecuencias para la variable D, utilizando como límite inferior de las clases 3cm, como límite superior 11cm y como amplitud de clase 0.5cm. ***Representar el histograma correspondiente.***

c.- ¿Cuál es el porcentaje de naranjas en la muestra cuyo diámetro está comprendido entre 6cm y 8cm?

¿entre 6cm y 10cm?

¿entre 4cm y 10cm?

d.- Basándose en los datos de la muestra, ¿qué tipo de modelo de recogedoras aconsejarías a la cooperativa?

2. Modelo para la distribución de la variable D

En este apartado estamos interesados, basándonos en la información recogida en la muestra, en proponer para la distribución de la v.a. D un modelo particular de distribución de los que hemos introducido en clase, es decir, para el comportamiento del diámetro para toda la producción.

a.- Calcular, para la muestra dada, la media y la desviación típica.

b.- Representar el histograma de las frecuencias absolutas, junto con la curva normal[1].

c.- Basándote en la muestra, ¿qué modelo de distribución continua te parece más adecuado para D?

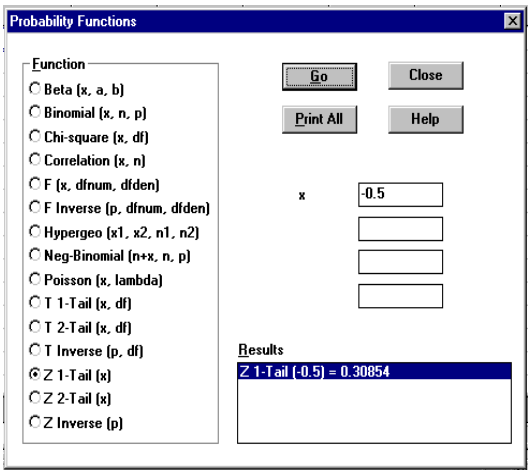
¿Con qué media y qué varianza?

d.- Suponiendo que D sigue una distribución normal N(8;1), vamos a calcular la probabilidad P(7 $\leq$D < 7,5):
Para ello, empezamos por considerar la variable tipificada:

$$\mathbf{Pr(7 \leq D < 7{,}5) = Pr\left(\frac{\quad}{} \leq Z < \frac{\quad}{} \right) = \Phi(-0{,}5) - \Phi(-1) =}$$

[1] Notar que, al activar la opción "Display Normal Curve", Statistix representa una densidad normal multiplicada por N * h, (N: tamaño muestral, h: amplitud de clase), de media y desviación típica calculadas a partir de la muestra

En el menú ***Statistics -> Probability functions***, la o pción ***Z1 Tail (x)*** nos permite calcular Pr(Z $\leq$x) si x < 0 y Pr(Z > x); si x $\geq$ 0:



También podéis utilizar la calculadora estadística NCSSCALC.

El resultado es P(7 $\leq$D < 7,5) = 0,30854 – 0,15866 = 0,15.

Completar la tabla siguiente

Clases	**muestra**	**Población**	**Clases**	**muestra**	**Población**
diámetro (cm)	*Frec.rel.*	*Pr ($\leq$D <)*	*diámetro (cm)*	*Frec.rel.*	*Pr ($\leq$D <)*
3,0 $\leq$D < 3,5		0.0000031	**7,5 $\leq$D < 8,0**		0.1915
3,5 $\leq$D < 4,0		0.0000283	**8,0 $\leq$D < 8,5**		
4,0 $\leq$D < 4,5			**8,5 $\leq$D < 9,0**		
4,5 $\leq$D < 5,0			**9,0 $\leq$D < 9,5**		
5,0 $\leq$D < 5,5			**9,5 $\leq$D < 10,0**		
5,5 $\leq$D < 6,0		0.0165	**10,0 $\leq$D <10,5**		
6,0 $\leq$D < 6,5		0.0441	**10,5 $\leq$D <11,0**		
6,5 $\leq$D < 7,0					
7,0$\leq$D < 7,5		0.1500			

e.- Utilizando el apartado anterior, ¿te parece que la muestra induce a pensar que la v.a D, diámetro de las naranjas en toda la producción, sigue una distribución N(8; 1)?

f.- En la producción entera, ¿qué porcentaje de naranjas tienen un diámetro comprendido entre 6cm y 8cm?

¿y entre 6cm y 10cm?

¿y entre 4cm y 10cm?

g.- ¿Qué tipo de modelo aconsejarías a la cooperativa?

Práctica 6. Ajuste de una serie temporal a un modelo Autorregresivo

En esta práctica aprenderemos a realizar el ajuste de un conjunto de datos a un modelo autorregresivo (AR), determinando el orden del modelo.

Introducción:

Una serie temporal es un conjunto de observaciones ordenadas en el tiempo, que pueden representar la evolución de una variable a lo largo de él. El objetivo del análisis de una serie temporal es el conocimiento de su patrón de comportamiento, para así prever su evolución futura, suponiendo que las condiciones no variarán.

Dado que no se trata de fenómenos deterministas, sino sujetos a una aleatoriedad, el estudio del comportamiento pasado ayuda a inferir la estructura que permita predecir su comportamiento futuro, pero es necesaria una gran cautela en la previsión debido a la inestabilidad del modelo.

La particular forma de la información disponible de una serie cronológica (se dispone de datos en periodos regulares de tiempo) hace que las técnicas habituales de inferencia estadística no sean válidas para estos casos, ya que nos encontramos ante n muestras de tamaño 1 procedentes de otras tantas poblaciones de características y distribución desconocidas.

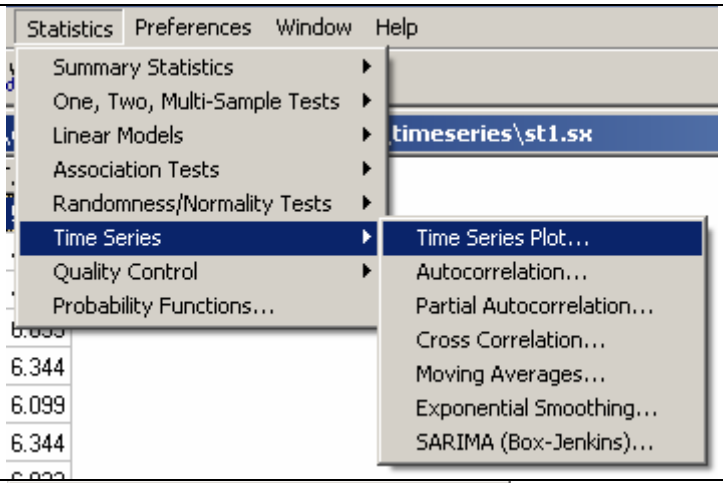
Normalmente, la mejor forma de comenzar a analizar los datos de una serie temporal es representar las observaciones vs. el tiempo a fin de detectar tendencias, patrones estacionarios, y outliers.

Desarrollo de la práctica

En primer lugar recuperaremos el fichero **st1.dat** que se encuentra en la ruta usual y almacenaremos los datos en la variable data.

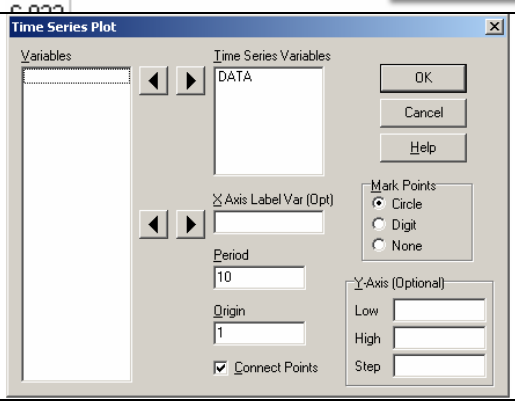
Una vez recuperados los datos representaremos gráficamente la serie obtenida. Para ello seleccionaremos:

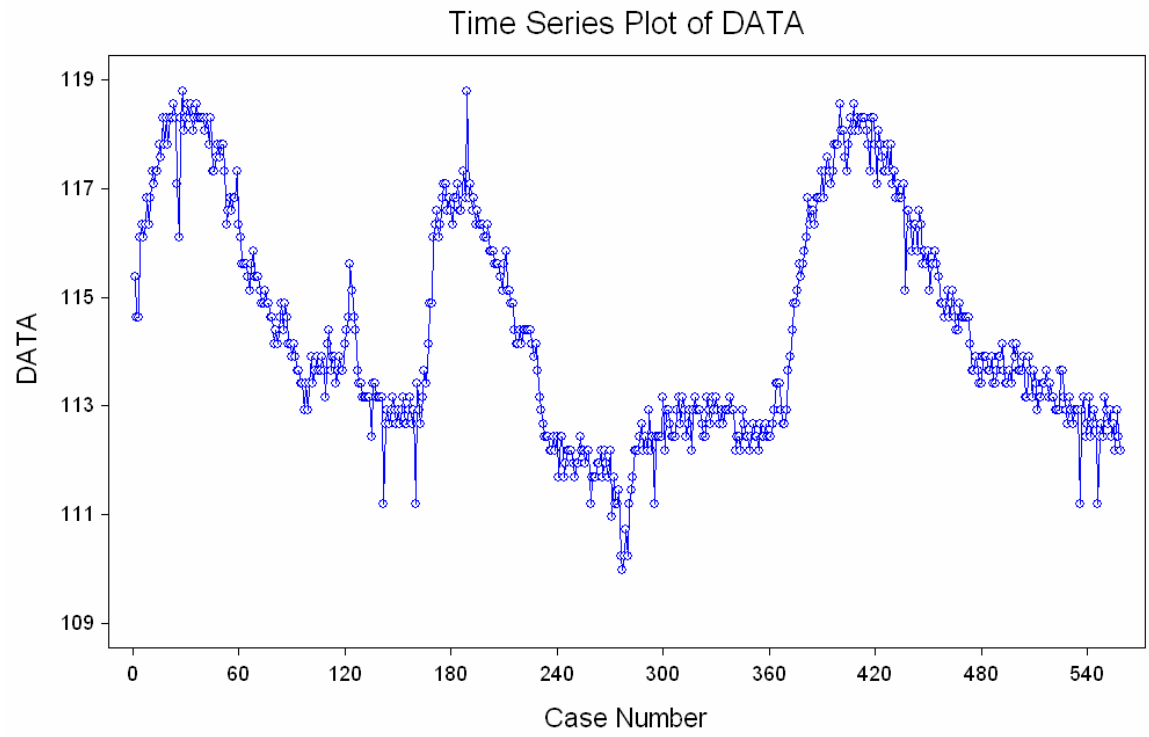
Statistics->Time Series->Time Series Plot



Y pasamos la variable DATA al cuadro:

Time Series Variables



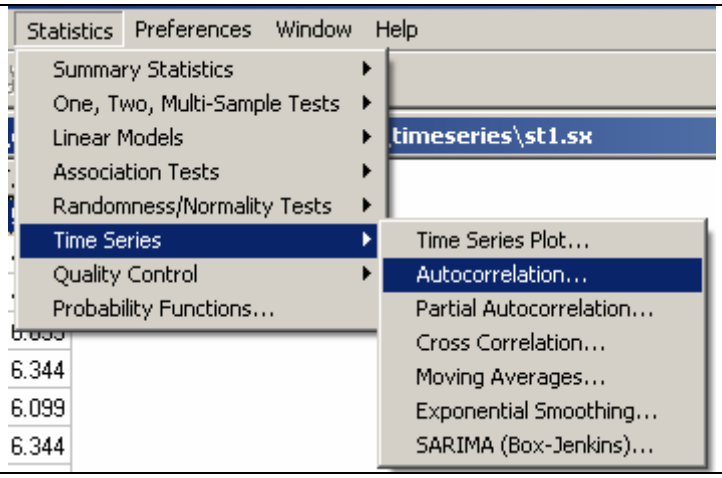


Como se observa no existe un patrón claro que permita predecir el comportamiento de la serie.

Con el fin de poder predecir su comportamiento, realizaremos las gráficas de autocorrelacion y autocorrelación parcial de la serie.

Para ello seleccionaremos:

- **Statistic-> Time Series -> Autocorrelation.**
- **Statistic-> Time Series -> Partial Autocorrelation**.



Autocorrelaciones

Autocorrelation Plot

```
STATISTIX FOR WINDOWS                                   ST1, 22/04/200

AUTOCORRELATION PLOT FOR DATA

              -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
 LAG    CORR.  +----+----+----+----+----+----+----+----+----+----+----+
   1    0.972                          > **>**********************
   2    0.963                        >   ****>**********************
   3    0.955                       >    *****>*********************
   4    0.944                       >    *****>*********************
   5    0.930                      >     ******>********************
   6    0.921                     >      *******>*******************
   7    0.911                     >      *******>*******************
   8    0.896                    >       ********>******************
   9    0.884                    >       ********>******************
  10    0.870                    >       ********>******************
  11    0.851                   >        *********>*****************
  12    0.837                   >        *********>*****************
  13    0.820                   >        *********>*****************
  14    0.800                  >         **********>****************
  15    0.784                  >         **********>****************
  16    0.766                  >         **********>****************
  17    0.748                 >          ***********>***************
  18    0.729                 >          ***********>***************
  19    0.705                 >          ***********>***************
  20    0.683                 >          ***********>**************
  21    0.660                 >          ***********>*************
  22    0.637                >           ************>*************
  23    0.615                >           ************>************
  24    0.591                >           ************>************
  25    0.569                >           ************>***********
  26    0.547                >           ************>***********
```

Se observa que decrecen conforme se aumentan los retardos

Autocorrelaciones Parciales

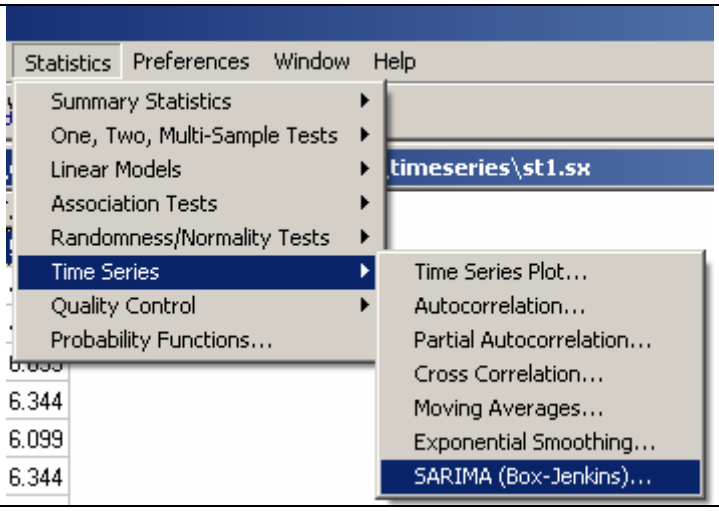
Partial Autocorrelation Plot

```
STATISTIX FOR WINDOWS                                   ST1, 22/04/200

PARTIAL AUTOCORRELATION PLOT FOR DATA

              -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
 LAG    CORR.  +----+----+----+----+----+----+----+----+----+----+----+
   1    0.972                          > **>**********************
   2    0.332                          > **>******
   3    0.159                          > **>**
   4   -0.002                          > * <
   5   -0.091                          <** <
   6    0.026                          > **<
   7    0.010                          > * <
   8   -0.084                          <** <
   9   -0.004                          > * <
  10   -0.058                          >** <
  11   -0.117                         *<** <
  12    0.005                          > * <
  13   -0.063                          >** <
  14   -0.068                          >** <
  15    0.004                          > * <
  16   -0.027                          >** <
  17    0.012                          > * <
  18   -0.027                          >** <
  19   -0.141                         **<** <
  20   -0.051                          >** <
  21   -0.052                          >** <
  22   -0.030                          >** <
  23    0.042                          > **<
  24   -0.057                          >** <
  25    0.006                          > * <
  26    0.032                          > **<
  27    0.048                          > **<
  28   -0.050                          >** <
```

En este caso se observa que sólo los dos primeros valores resultan significativos. Esto nos indica que nos encontramos frente a un proceso ***Autorregresivo de orden 2 (AR(2)).***

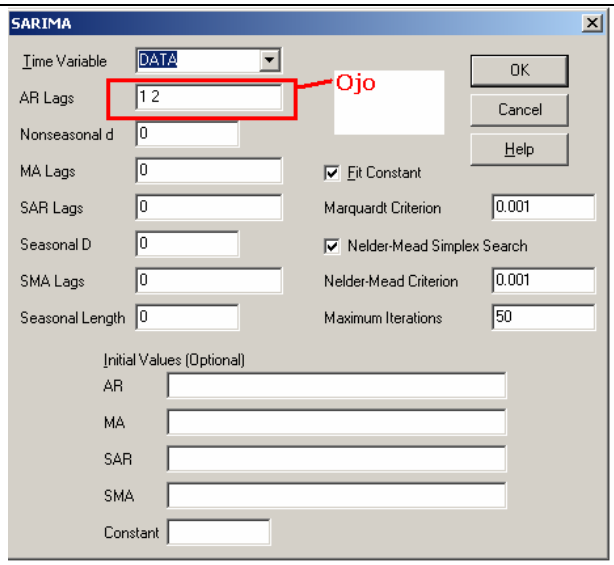
Con el fin de estimar los valores de los coeficientes de la autorregresión seleccionaremos:

- **Statistic-> Time Series -> SARIMA (Box-Jenkins)**



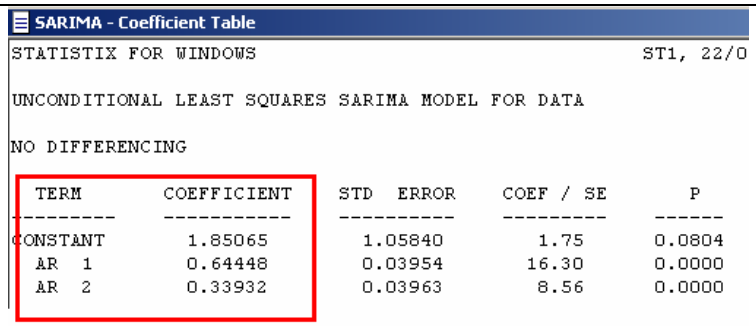
e introduciremos los datos correspondientes:

- La variable data en el cuadro **Time Variable**
- 2 coeficientes autorregresivos: **Pondremos 1 2 separados por un espacio.**
- **No modificaremos ningún parámetro más**.



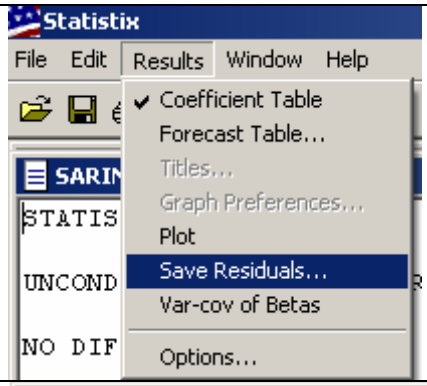
En la ventana que obtenemos destacamos la siguiente información:

- **Constante: 1.85065**
- **Coeficiente AR1: 0.64448**
- **Coeficiente AR2: 0.33932**



Seguidamente almacenaremos los valores ajustados y los residuos. Para ello seleccionaremos sobre la ventana anterior en el menú **Results**:

- **Save Residuals**

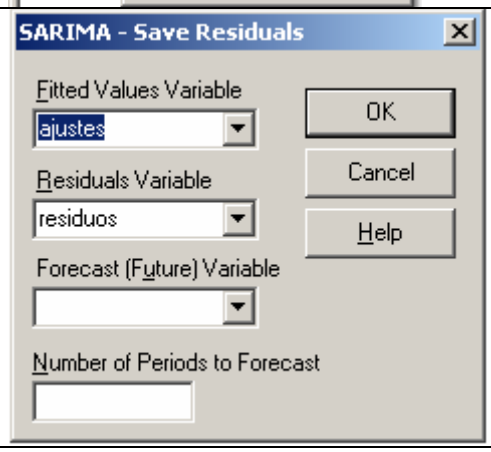


Almacenaremos los ajustes en la variable:

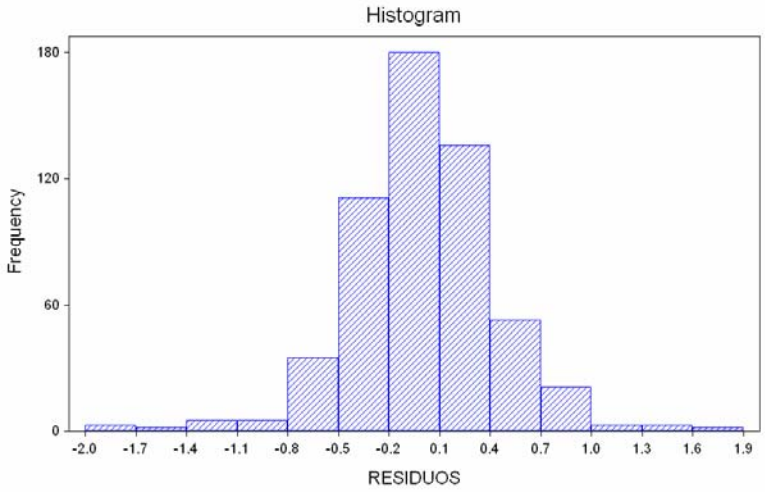
- **ajustes**

y los errores (residuos) en la variable:

- **residuos**



Seguidamente verificaremos que los residuos verifican las **hipótesis de normalidad** (haciendo el histograma), **independencia** (haciendo el gráfico de autocorrelaciones y autocorrelaciones parciales y observando que no existen valores significativos) y **estabilidad de la varianza y centrados en cero** (representando los residuos con un Time Series Plot)



Autocorrelación de los residuos

Autocorrelation Plot

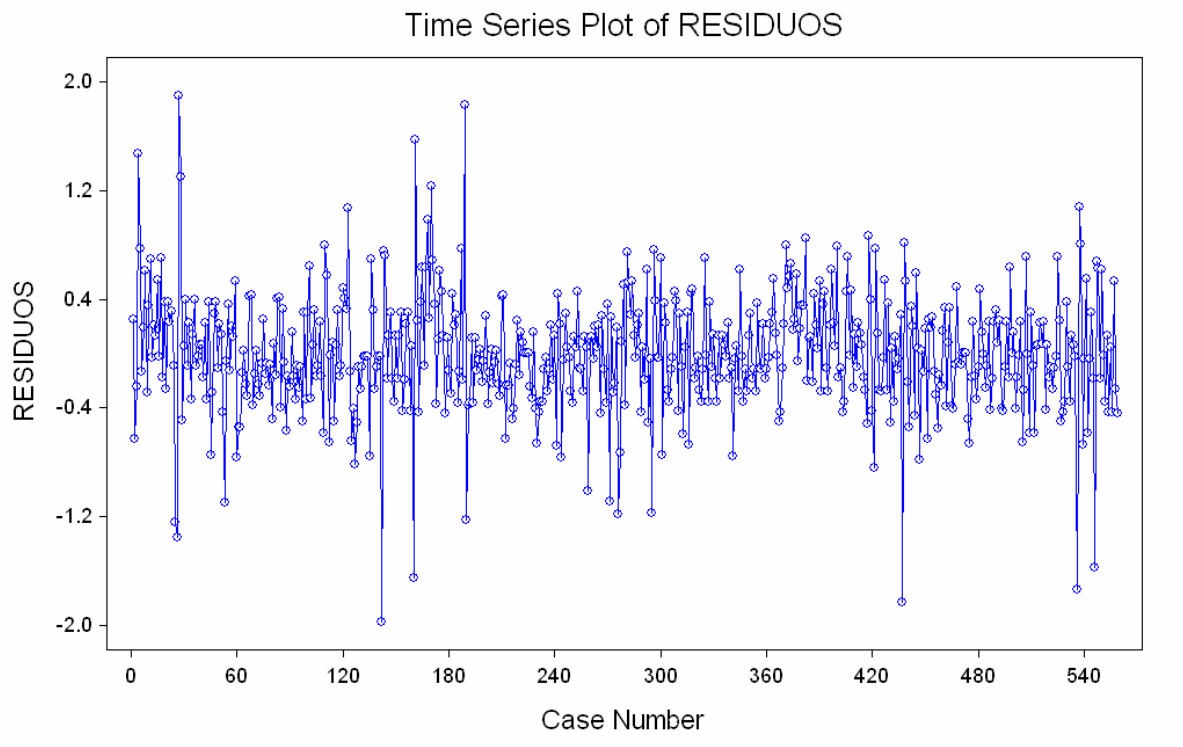
```
STATISTIX FOR WINDOWS                                   ST1, 22/04/2005,

AUTOCORRELATION PLOT FOR RESIDUOS

              -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
 LAG    CORR.  +----+----+----+----+----+----+----+----+----+----+----+
   1   -0.056                          >** <
   2   -0.139                         *<** <
   3    0.083                          > **<
   4    0.067                          > **<
   5   -0.079                          >** <
   6    0.040                          > **<
   7    0.095                          > **>
   8   -0.021                          >** <
   9    0.073                          > **<
  10    0.078                          > **<
  11   -0.046                          >** <
  12    0.054                          > **<
  13    0.037                          > **<
  14   -0.015                          > * <
  15    0.021                          > **<
  16    0.014                          > * <
  17    0.068                          > **<
  18    0.116                          > **>*
  19    0.005                          > * <
  20    0.013                          > * <
  21    0.009                          > * <
  22   -0.027                          >** <
  23    0.016                          > * <
  24    0.002                          > * <
  25   -0.017                          > * <
  26   -0.021                          >** <
  27    0.118                          > **>*
  28    0.045                          > **<
```

Autocorrelación Parcial de los residuos

Partial Autocorrelation Plot

```
STATISTIX FOR WINDOWS                                   ST1, 22/04/20

PARTIAL AUTOCORRELATION PLOT FOR RESIDUOS

              -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
 LAG    CORR.  +----+----+----+----+----+----+----+----+----+----+----+
   1   -0.056                          >** <
   2   -0.143                         **<** <
   3    0.068                          > **<
   4    0.057                          > **<
   5   -0.053                          >** <
   6    0.044                          > **<
   7    0.076                          > **<
   8    0.004                          > * <
   9    0.100                          > **>
  10    0.069                          > **<
  11   -0.021                          >** <
  12    0.072                          > **<
  13    0.009                          > * <
  14    0.001                          > * <
  15    0.027                          > **<
  16   -0.021                          >** <
  17    0.073                          > **<
  18    0.127                          > **>*
  19    0.010                          > * <
  20    0.043                          > **<
  21   -0.012                          > * <
  22   -0.050                          >** <
  23    0.023                          > **<
  24   -0.037                          >** <
  25   -0.041                          >** <
  26   -0.039                          >** <
  27    0.069                          > **<
```

Gráfico de los residuos



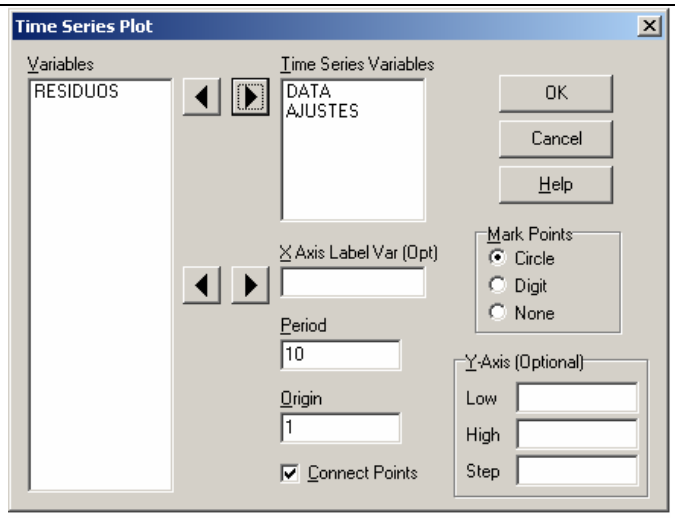
Si calculamos la media y la varianza de los residuos obtenemos:

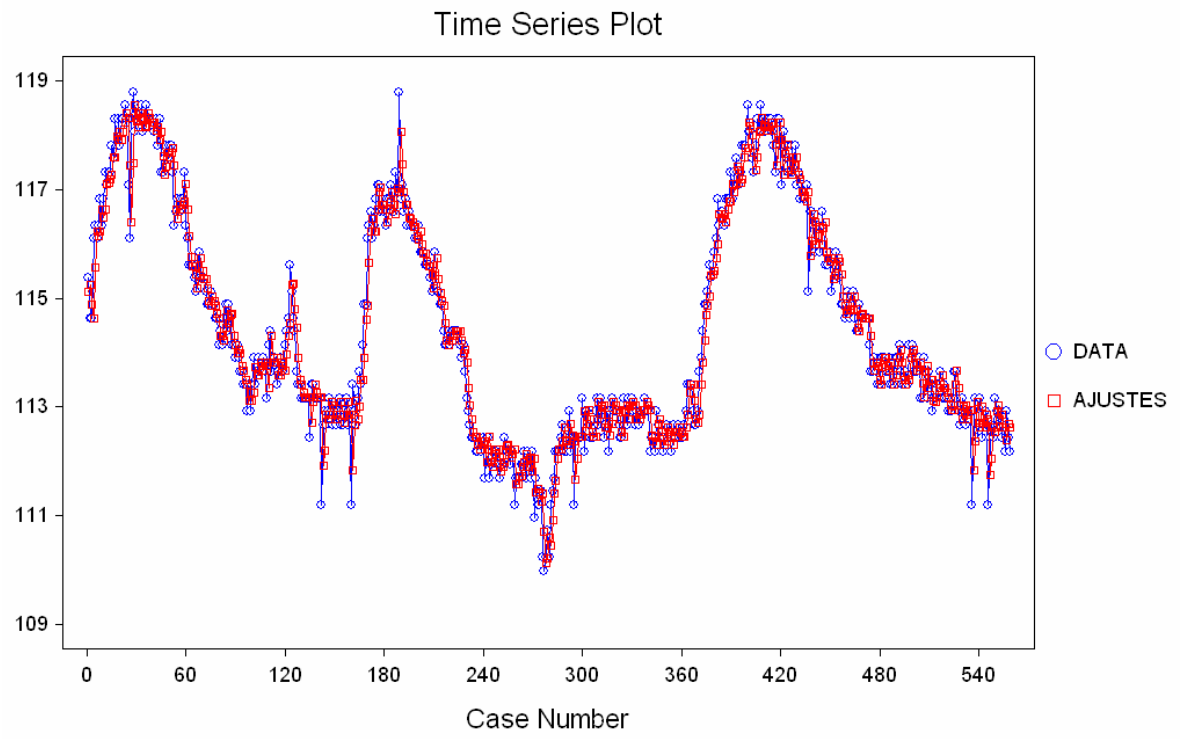
```
STATISTIX FOR WINDOWS
DESCRIPTIVE STATISTICS

VARIABLE           MEAN     VARIANCE
RESIDUOS     -3.784E-03       0.1999
```

Una vez que se han verificado las hipótesis sobre los residuos pasamos a representar de manera conjunta los valores observados (variable data) y las predicciones (ajustes). Para ello seleccionaremos:

- **Statistics->Time Series->Time Series Plot**

y pasaremos ambas variables para observarlas en el mismo gráfico.





Conclusión:

Como se observa, las predicciones coinciden prácticamente con los valores observados, por lo que proponemos el siguiente modelo explicativo para la serie observada:

- Constante: 1.85065
- Coeficiente AR1: 0.64448
- Coeficiente AR2: 0.33932

$$Y_t=1.85065+0.64448\ Y_{t-1}+ 0.33932\ Y_{t-2} + E_t$$

donde Et representa los errores que siguen una distribución normal de media cero y varianza constante igual a 0.2

Ejercicio:

Analizar los datos almacenados en el fichero **st2.dat** ubicado en la ruta usual.

Práctica 7. Muestreo

En la siguiente práctica, tenemos que estimar la media poblacional de un lote de 2000 rodamientos que necesitamos en el proceso de fabricación de nuestro producto. Intentaremos simular el proceso de selección de distintas muestras sobre este lote. Las piezas en el lote van numeradas de 1 a 2000. La práctica seguirá tres pasos

(1) Para formar nuestra muestra, generaremos de manera aleatoria los números de las piezas escogidas.

(2) Después de cargar en nuestra hoja de cálculo Statistix los diámetros que corresponden a nuestra muestra, realizaremos el tratamiento estadístico y construiremos intervalos de confianza con dos niveles de confianza distintos para la media poblacional.

(3) Acabaremos con un paso que NO se puede realizar en general en una situación real : estudiaremos la población entera, para comparar los resultados obtenidos en (2) con los resultados reales.

Repetiremos los pasos (1) y (2) para tres tamaños muestrales distintos : se tomará una muestra de tamaño 20, otra de tamaño 50 y por último una de tamaño 100.

Resolución:

En primer lugar, debemos simular el proceso de selección de 20 unidades. Para ello, generaremos 20 números aleatorios comprendidos entre 0 y 2000 de la siguiente manera:

Crearemos una nueva variable llamada:

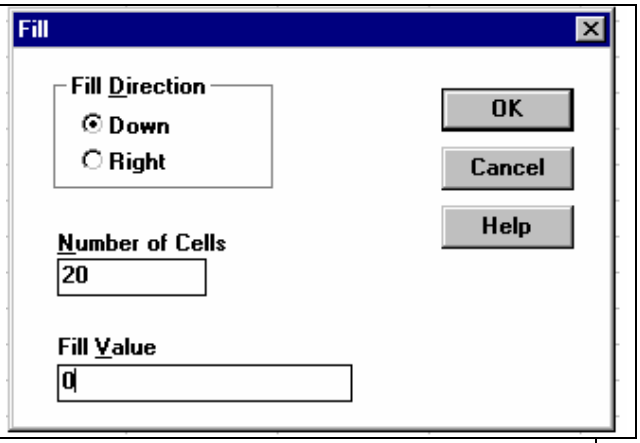
Seleccionadas

Utilizando:

Data-> Insert-> Variables.

Posteriormente la llenaremos 20 celdas con ceros o cualquier otro valor utilizando la opción **Fill** del menú Data:

Data-> Fill.



Seguidamente, generaremos los 20 números aleatorios que se almacenarán en la variable **SELECCIÓN** los cuales corresponden a los rodamientos que serán inspeccionados suponiendo que estos se encuentran ordenados.

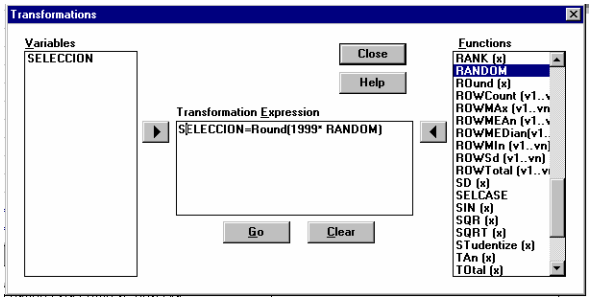
Para ello utilizaremos la opción:

Data-> Transformation.

tras lo cual introduciremos la siguiente instrucción:

SELECCION=1+Round(1999* RANDOM)

Como se observa, se genera un número aleatorio entre 0 y 1, posteriormente se multiplica por 1999, con lo cual obtendremos un número que se encuentra en (0,1999). Al efectuar el redondeo y sumar 1 obtendremos un número en [1,2000].

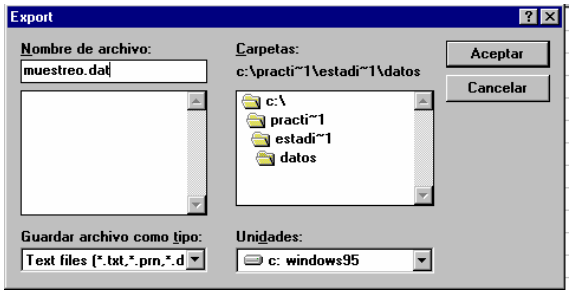


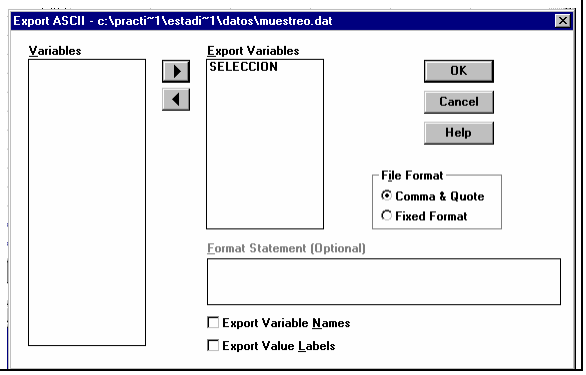
Estos 20 valores son los números de las piezas seleccionadas como muestra. Por consiguiente almacenaremos sus valores en un fichero llamado **muestreo.dat** dentro del directorio **PRACTI~1\ESTADI~1\datos**, esto es, seleccionaremos dentro del menú **File** la opción **Export**:

File-> Export

Y grabaremos como:

\PRACTI~1\ESTADI~1\datos\muestreo.dat





Una vez almacenado el fichero, ejecutaremos el programa:

Muestreo

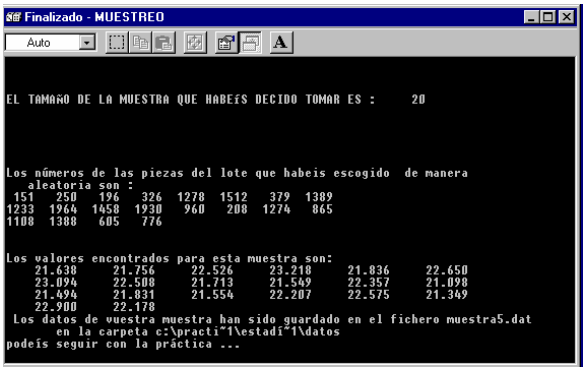
Picando sobre el icono correspondiente que se encuentra en el Escritorio de W95. El programo **Muestreo** se limita a conseguir del fichero que contiene los datos de todo el lote los que corresponden a la muestra que hemos seleccionado.

Tras ejecutar el programa, se creará un nuevo fichero llamado **muestra5.dat** dentro de:

\practi~1\estadi~1\datos

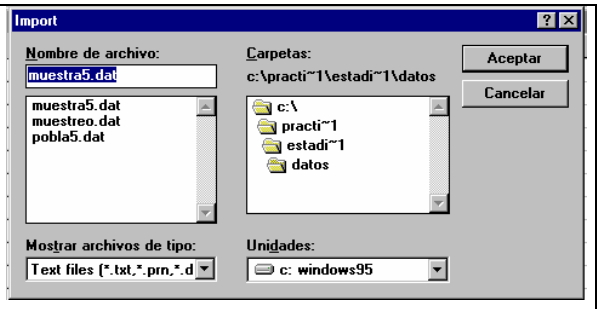
el cual contiene los valores de los diámetros de las piezas seleccionadas.

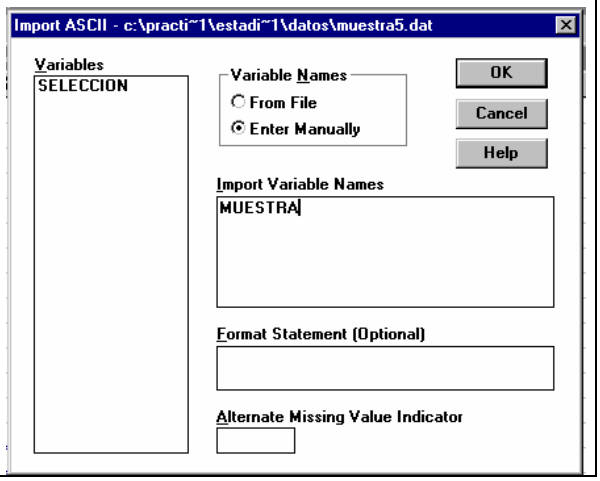
Debemos observar que los valores seleccionados por cada ordenador son distintos pues la idea es simular que cada alumno ha realizado un proceso de muestreo diferente.



Una vez que hemos generado la muestra, volveremos a Statistix y recuperaremos los valores obtenidos en una nueva variable llamada: MUESTRA **utilizando la opción** Import **del menú** FILE**:**

File->Import



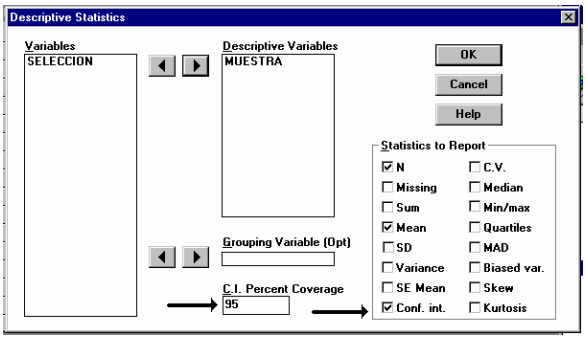


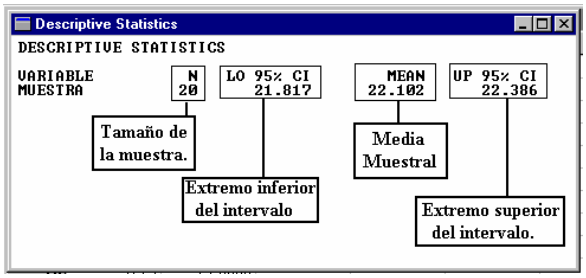
Seguidamente, calcularemos un intervalo de confianza para la media con una confianza del 95%, para ello, seleccionaremos del menú

Statistics-> Summary Statistics ->Descriptive Statistics.

Marcando el recuadro correspondiente al intervalo de confianza (*Conf. Int.*) y poniendo el valor 95 para el porcentaje de confianza (*C.I. Percent. Coverage*)

Nota: en clase sólo hemos visto intervalos de confianza si la varianza es conocida, si no es el caso, se estima ésta última de los datos, y se procede de manera similar, tal como lo veremos en un tema posterior



Tras lo cual obtendremos el intervalo de confianza, para la media poblacional, obtenido con esa muestra en concreto.	

Anota aquí los resultados obtenidos:

Nivel de confianza	Tamaño Muestral	Extr. Inferior	Media muestral	Extr. Superior
95%	20			
90%	20			

Para el proceso de fabricación, un rodamiento conviene si su diámetro está comprendido entre 20 y 24mm. ¿Cuál, es para tu muestra, la proporción de rodamientos de la muestra que convienen?

Por último, podemos examinar la población.

El fichero que contiene los 2000 datos de la población es pobla5.dat. Se puede comprobar sin dificultad que la media poblacional es

$$\mu=22$$

Vamos a representar gráficamente los intervalos para la media poblacional al 90% de confianza obtenidos con un tamaño muestral de 20 para todos los grupos de esta práctica.

¿Cómo se interpreta esta representación? ¿Cuántos intervalos han fallado?

Repite el proceso (pero sin representación gráfica) tomando muestras de tamaño 50 y 100.

Nivel de confianza	Tamaño Muestral	Extr. Inferior	Media muestral	Extr. Superior	Proporción muestral de defectuosos
95%	50				
90%	50				
95%	100				
90%	100				

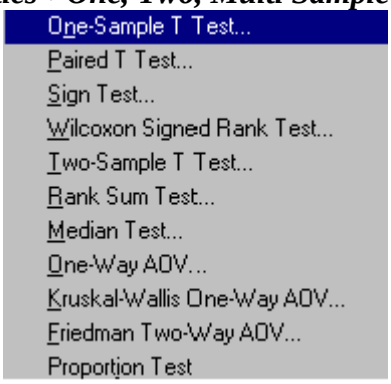
Nota.-

Como se observa, los intervalos de confianza y las proporciones muestrales obtenidos por cada grupo de alumnos resultan distintos debido a que cada uno de ellos ha trabajado con una muestra distinta de la población aunque todas son de igual tamaño.

Práctica 8: Contrastes paramétricos.

El programa Statistix puede realizar los cálculos correspondientes a varios contrastes paramétricos de hipótesis. Se encuentran los comandos en el menu:

Statistics->One, Two, Multi Sample tests :



Las instrucciones que nos pueden interesar de momento son:

a.- One sample t test: test de Student para una muestra, contraste sobre la media.

b.- Two sample t test: test de Student para dos muestras, contraste de igualdad de medias.

Procedimiento para One sample t test :

Ejemplo: Se obtuvieron los siguientes datos en la medición de la intensidad de una corriente :

3,823 3,844 3,762 3,871 3,762

Se supone que el error que se comete durante la medición sigue una distribución normal de media 0 de varianza desconocida σ^2 .

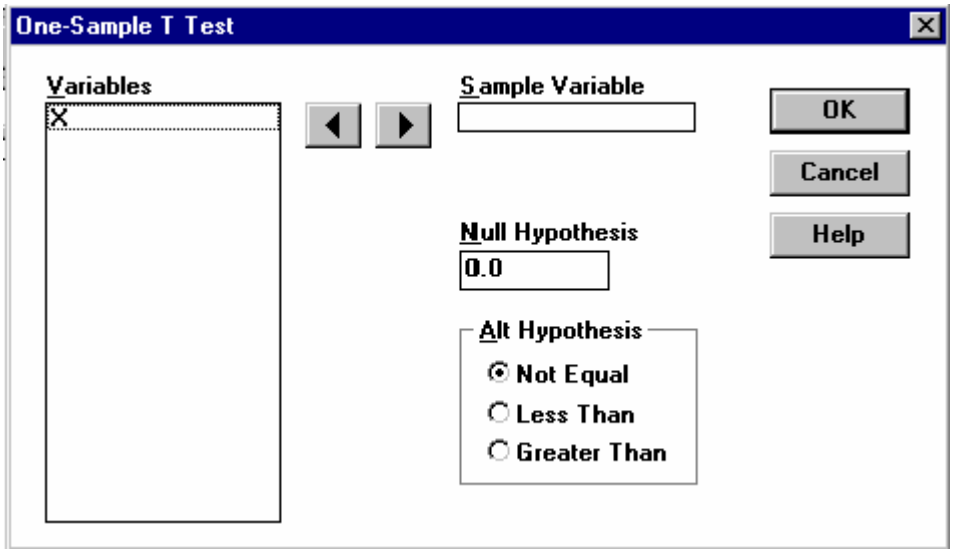
Construir un intervalo de confianza al nivel de 95% para la intensidad real y realizar el contraste de hipótesis para comprobar si la intensidad real es significativamente menor que 3.90.

Resolución:

Empezamos por introducir los datos (menu ***Data->insert variable***). Una vez que los tengamos en nuestra hoja de cálculo, podemos obtener de manera simultánea el intervalo de confianza y el contraste de hipótesis. En nuestro caso las hipótesis de interés son:

$$H_0 : \mu = 3{:}9$$
$$H_1 : \mu < 3{:}9$$

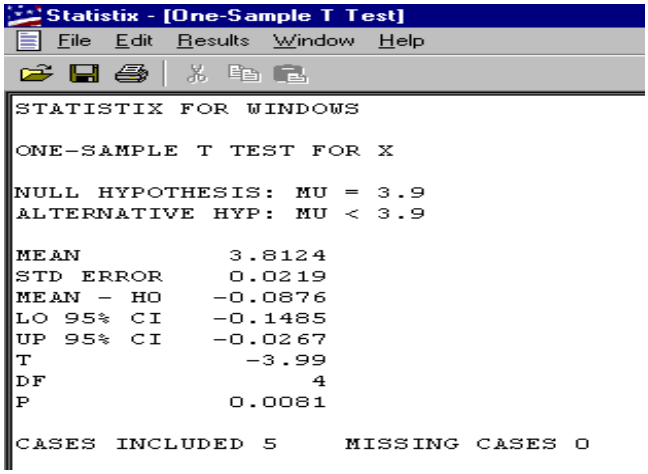
En el menu, ***Statistics->One, Two, Multi Sample tests*** , activamos la opción ***One Sample t test*** y aparece una ventana de diálogo



En esa ventana, encontramos, como es usual con Statistix, una lista de las variables que ya tenemos definidas y entre las cuales hay que escoger la variable cuyo parámetro queremos contrastar. Una vez seleccionada en esa lista la variable, la pasamos al cuadro " ***Sample variable***" con las flechas. En el cuadro "***Null hypothesis***" indicamos el valor μ_0 que queremos contrastar, es decir el que entra en la definición de $\mathbf{H_0}$: $\mu=\mu_0$, mientras que en el cuadro "***Alt hypothesis***" debemos seleccionar el tipo hipótesis alternativa:

not equal $H_1 : \mu \neq \mu_0$
less than $H1 : \mu < \mu_0$
greater than $H1 : \mu > \mu_0$

Al pinchar en OK, obtenemos una ventana que recoge los resultados:



donde:

- ***Mean, Std error:*** representan la media y desviación muestrales.
- ***Mean-H_0, LO 95% CI, UP 95% CI :*** representan la media, extremo inferior y superior del intervalo de confianza al 95 % de la media $\mu - \mu_0$.
- ***T*** : valor del estadístico de Student.
- ***D.F :*** grados de libertad (Degrees of Freedom).
- ***P :*** p-valor.

En este caso obtenemos un p-valor de 0.0081, por lo tanto rechazamos H_0 con gran confianza y afirmamos que la intensidad real es significativamente menor que 3.900.

Procedimiento para Two-sample t test :

Ejemplo: En el departamento de control de calidad de una empresa, se quiere determinar si habido un descenso significativo de la calidad de su producto entre las producciones de dos semanas consecutivas a consecuencia de un incidente ocurrido durante el fin de semana. Deciden tomar una muestra de la producción de cada semana, si la calidad de cada artículo se mide en una escala de 100, obtienen los resultados siguientes:

Semana 1 : 93 86 90 90 94 91 92 96
Semana 2 : 93 87 97 90 88 87 84 93

Construya un intervalo de confianza para la diferencia de medias al nivel de 95%. A continuación realice el contraste de hipótesis para probar si ha habido un descenso significativo de la calidad en la semana 1 y la semana 2.

Resolución:

Empezamos por introducir las variables, tenemos dos posibilidades:

(a) definimos una variable llamada "***Calidad"*** que contiene los 16 datos, y otra variable ***"Semana"*** que valga 1 si el dato de calidad corresponde a la primera semana y 2 si proviene de la segunda. Esto corresponde a una presentación categórica de los datos.
(b) definimos dos variables ***"Calidad1"*** y ***"Calidad2"*** que contienen los 8 datos de calidad de cada una de las dos semanas. Corresponde a una presentación de los datos en tabla.

Al activar la instrucción ***Two-sample t test***, aparece una ventana, con una lista de las variables definidas. Debemos escoger en el cuadro "***Model especification***" la manera en la que están presentadas nuestras variables. Si los datos que queremos analizar se presentan en dos variables distintas (opción (b) anterior), escogemos la opción "***Table***". En cambio si hemos escogido la opción (a) anterior activamos la opción "***Categorical***"

- **Con la opción (b)**: Después de especificar el modelo, pasamos las dos variables que queremos analizar desde la lista de variables al cuadro "***Table variables***".
- **Con la opción (a):** pasamos la variable "***Calidad***" al cuadro ***"Dependent variable***", y la variable "***Semana***" al cuadro "***Categorical variable***".

Para rellenar los cuadros "***Null Hypothesis***" y "***Alt Hypothesis***", debemos considerar que contrastamos la diferencia entre las dos medias, (por ejemplo, para igualdad de medias, el cuadro "**Null Hypothesis**" debe contener **0**). En nuestro problema las hipótesis son:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

La ventana de resultados es:

```
TWO-SAMPLE T TESTS FOR CALIDAD BY SEMANA

                          SAMPLE
  SEMANA         MEAN      SIZE       S.D.        S.E.
----------   ----------   ------   ---------   ---------
     1          91.500       8       3.0237      1.0690
     2          89.875       8       4.2237      1.4933
DIFFERENCE      1.6250

NULL HYPOTHESIS: DIFFERENCE = 0
ALTERNATIVE HYP: DIFFERENCE > 0

ASSUMPTION             T       DF        P     95% CI FOR DIFFERENCE
-----------------   ------   ------   ------   ---------------------
EQUAL VARIANCES       0.88     14     0.1956     (-2.3139, 5.5639)
UNEQUAL VARIANCES     0.88     12.7   0.1964     (-2.3527, 5.6027)

                        F      NUM DF   DEN DF      P
TESTS FOR EQUALITY   -------   ------   ------   ------
      OF VARIANCES     1.95       7        7     0.1988

CASES INCLUDED 16    MISSING CASES 0
```

Las primeras líneas contienen un resumen descriptivo de las dos variables de la muestra. A continuación aparecen las hipótesis que hemos contrastado (**difference=0** contra **difference <> 0**). El programa ha realizado el cálculo de dos estadísticos diferentes:

- El primero corresponde al caso en que suponemos que las dos variables tienen la misma varianza poblacional.
- Para el cálculo del segundo, las dos varianzas pueden ser distintas.

Para cada uno de esos dos casos, tenemos el valor del estadístico T, el número de grados de libertad, el p valor así como el intervalo de confianza al 95% para la diferencia de medias. **¿Cuál es su conclusión para el problema que nos interesa?**

Ejercicio 1: Cierto fabricante suministra determinado material siendo una característica de interés de este la resistencia en kg/cm^2 . El fabricante afirma que la media de resistencia es de 220 kg/cm^2. Una empresa que quiere comprar el material decide contrastar la afirmación anterior. Para ello toma una muestra de 9 elementos de ese material y obtiene los siguientes datos sobre la resistencia:

203 229 215 220 223 233 208 228 209

¿Puede concluirse a partir de los datos que la media es significativamente diferente de la que afirma el fabricante?

Ejercicio 2: En un estudio sobre el tiempo empleado a la hora de almorzar entre los empleados de una determinada empresa, se seleccionan 16 mujeres y 16 hombres, obteniéndose los siguientes resultados en minutos:

Mujeres 54 61 44 50 50 54 59 54 22 58 45 30 25 29 24 38

Hombres 61 46 50 17 45 31 20 54 27 38 30 42 58 44 58 38

¿Podemos concluir que la duración empleada es la misma en hombres y mujeres?