

***UNIVERSIDAD CARLOS III DE MADRID
MASTER EN CALIDAD TOTAL***

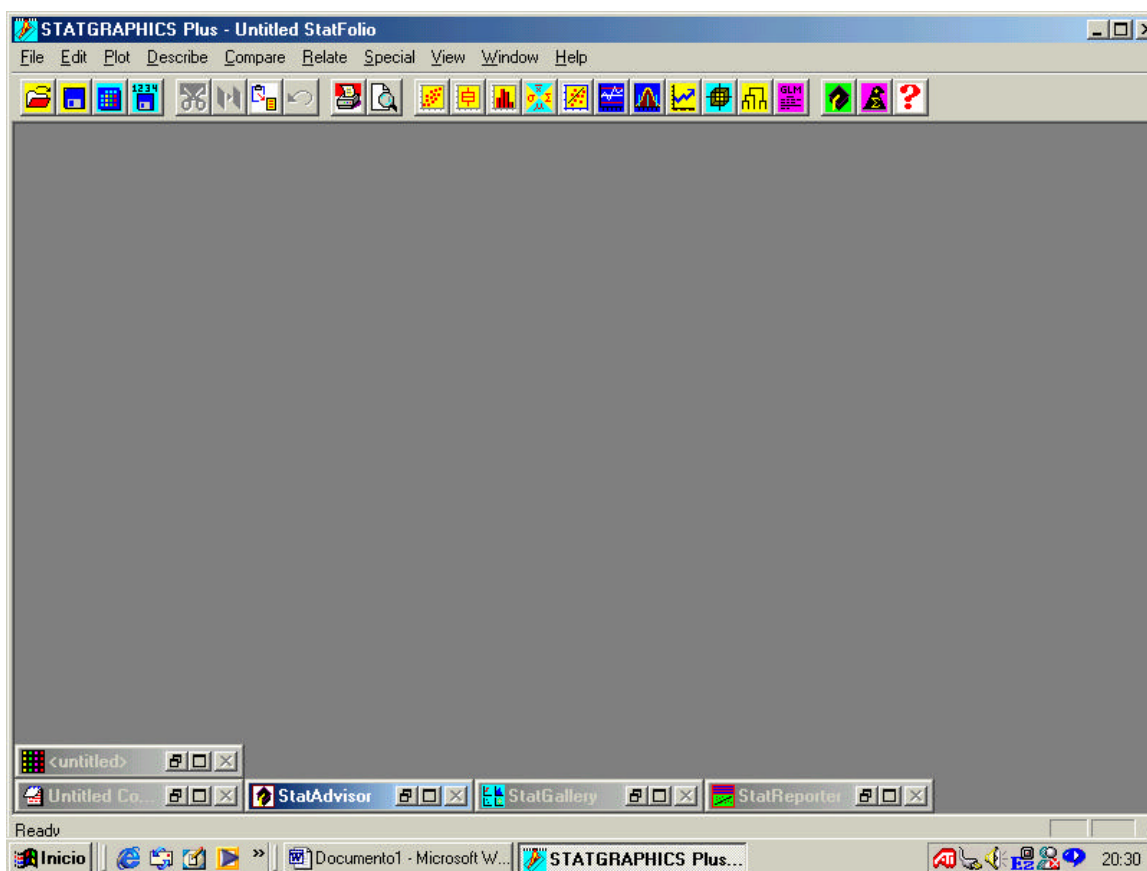
***PRÁCTICAS DE ESTADÍSTICA I
MANUAL DE STATGRAPHICS***

I. INTRODUCCIÓN Y MANEJO DE DATOS

INTRODUCCION

El programa Statgraphics es un software que está diseñado para facilitar el análisis estadístico de datos. Mediante su aplicación es posible realizar un análisis descriptivo de una o varias variables, utilizando gráficos que expliquen su distribución o calculando sus medidas características. Entre sus muchas prestaciones, también figuran el cálculo de intervalos de confianza, contrastes de hipótesis, análisis de regresión, análisis multivariantes, así como diversas técnicas aplicadas en Control de Calidad.

El programa trabaja en un entorno WINDOWS y su pantalla principal (a la que se accede ejecutando el programa SGWIN.EXE o directamente clickeando sobre el icono correspondiente, es la siguiente:



(Para salir del programa seleccionamos en la barra de menú FILE...EXIT STATGRAPHICS o simplemente se cierra la ventana principal de la aplicación)

En la pantalla principal de Statgraphics, podemos distinguir los siguientes elementos:

1. Barra de menú
2. Barra de herramientas
3. Barra de tareas

Analicemos ahora cada uno de los elementos que podemos encontrar en la ventana principal.

Barra de menú



La barra de menú siempre estará disponible al utilizar el programa, de forma que sea posible seleccionar el análisis deseado.

Al clickear con el ratón sobre cada una de las palabras que componen la barra, aparecerá un menú desplegable con otras opciones asociadas. Así tendremos:

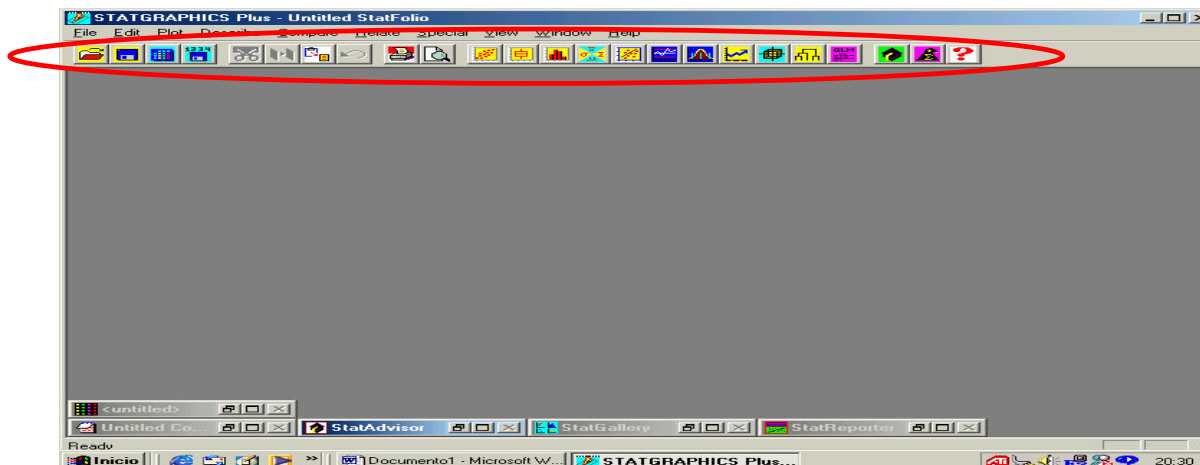
File: permite realizar operaciones de carácter general: abrir, cerrar o grabar ficheros, imprimir y salir de Statgraphics.

Edit: como en otras aplicaciones en entorno Windows, este menú está asociado a diversas opciones de edición: cortar, copiar, pegar, deshacer...

Plot, Describe, Compare, Special: al presionar con el ratón sobre ellos tendremos acceso a diversos menús de análisis de Statgraphics que se irán analizando a lo largo de este manual.

View, Window, Help: tienen disponible varias opciones de formato y ayuda, de forma similar a otras aplicaciones que trabajan en el mismo entorno.

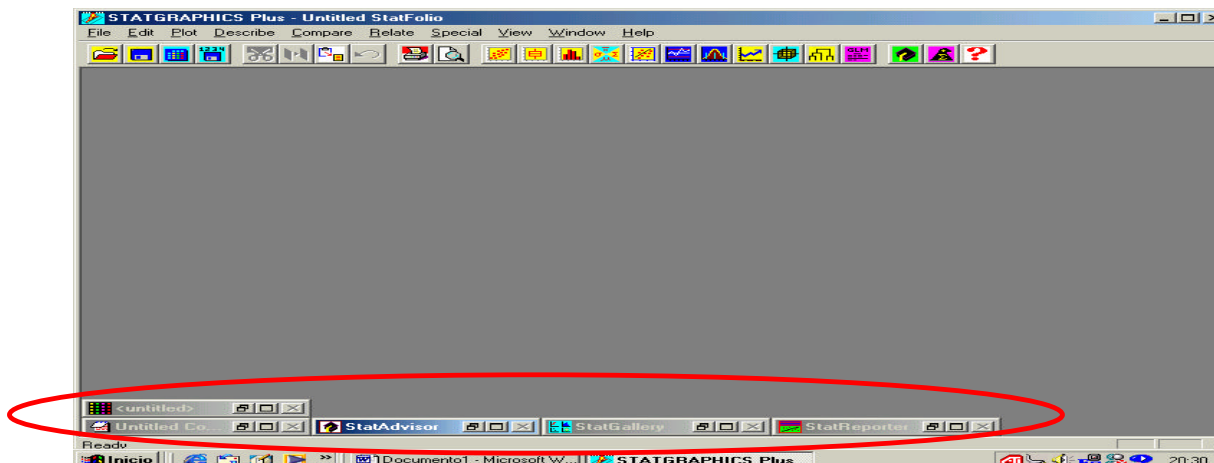
Barra de herramientas



La barra de herramientas tiene como función asociar iconos (botones rápidos) con algunas de las opciones mas frecuentemente utilizadas de la barra de menú. Si se señala con el ratón cualquier botón de la barra, aparecerá una breve descripción de la función asociada.

Barra de tareas

Incluye iconos asociados que contendrán los datos que se analizan, comentarios personales sobre el análisis, resultados del análisis efectuado y comentarios e interpretaciones del programa de los resultados obtenidos. El conjunto de estos elementos forma el Statfolio.



Statadvisor: herramienta incorporada al programa, que interpreta de forma sencilla los resultados obtenidos.

Statgalery: permite almacenar los resultados (gráficos incluidos) del análisis realizado. El realizar cualquier análisis estadístico, el sistema genera una ventana de análisis, que estará dividida en paneles conteniendo las diferentes partes del análisis. Clickeando con el botón derecho del ratón sobre cada uno de estos paneles y seleccionando *Copy to Galery* podremos incluir el panel en el Statgalery al utilizar

la opción de *Copiar* una vez posicionados con el ratón sobre el panel de destino. (La configuración de los paneles del Statgalery es seleccionable sin más que desplazar con el ratón las barras horizontales y verticales)

Untiled comments y Statreporter: opciones de Statgraphics que permiten introducir los comentarios de usuario para su posterior edición.

Ventana de datos: hoja de cálculo que contiene los datos que se van a analizar. Pueden introducirse directamente desde el teclado o recurrirse desde un fichero ya grabado. (FILE...OPEN...OPEN DATA FILE)

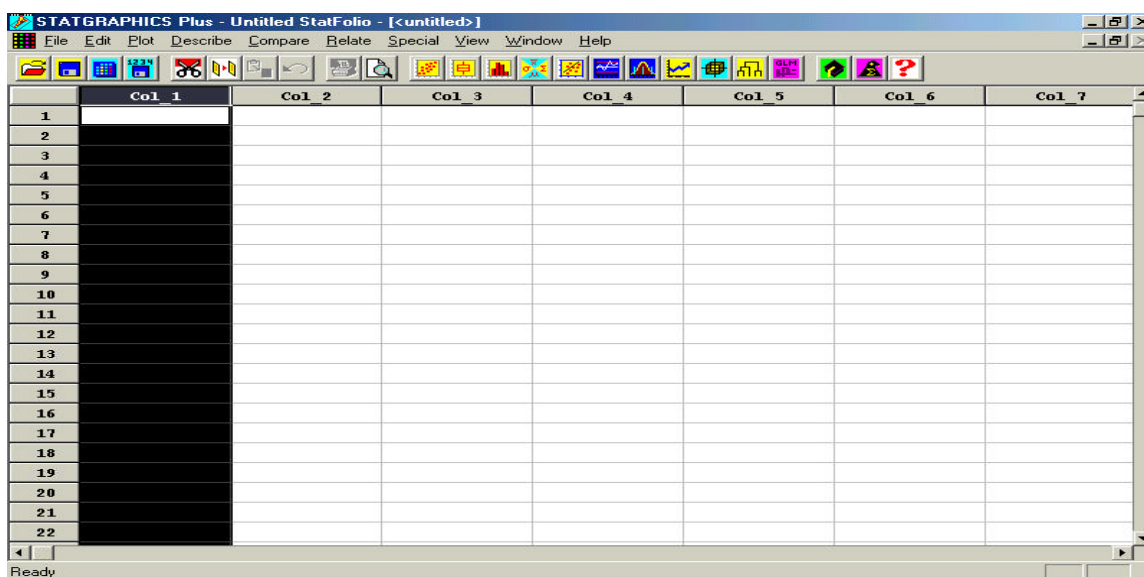
Al conjunto de los elementos anteriores se le denomina *Statfolio*, que puede almacenarse bajo un nombre único (fichero .spg) activando la opción FILE...SAVE...SAVE STATFOLIO. Si abrimos un Statfolio previamente guardado y continuamos con el análisis estadístico, cualquier modificación que se realice sobre los datos se transmitirá automáticamente sobre todos los análisis previamente realizados, por lo que la principal utilidad del Statfolio es repetir un análisis sistemáticamente sobre distintos conjuntos de datos.

TRABAJAR CON DATOS EN STATGRAPHICS

Los datos que van a analizarse mediante Statgraphics pueden introducirse directamente desde el teclado en la ventana de datos. Los datos pueden agruparse formando una variable (cada una de las columnas de la hoja de cálculo constituye la ventana de datos).

Para poder analizar una variable (es decir, los datos que contiene) es necesario definirla realizando las siguientes operaciones:

Seleccionamos la columna en la que queremos introducir los datos. Para ello clickeamos sobre la etiqueta de la columna (Inicialmente será Col_1)

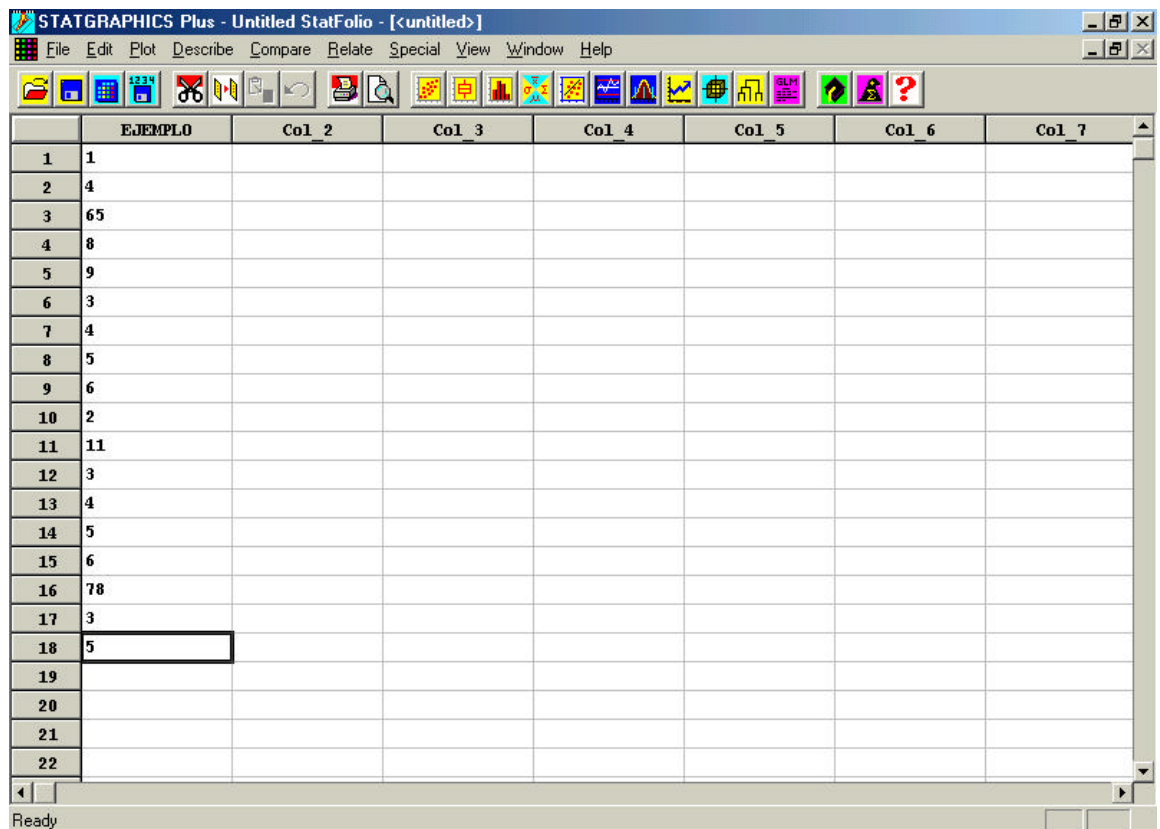


Pulsamos con el botón derecho del ratón sobre la columna seleccionada. Aparecerá un menú del que seleccionamos la opción *Modify Column*:

The 'Modify Column' dialog box is shown. It includes input fields for 'Name:', 'Comment:', and 'Width:' (currently 13). Action buttons 'OK', 'Cancel', 'Define...', and 'Help' are on the right. The 'Type' section offers radio button options: 'Numeric' (selected), 'Character', 'Integer', 'Date', 'Month', 'Quarter', 'Time (HH:MM)', 'Time (HH:MM:SS)', 'Fixed Decimal:' (with a sub-field of 2), and 'Formula'. An empty text box is at the bottom of the 'Type' section.

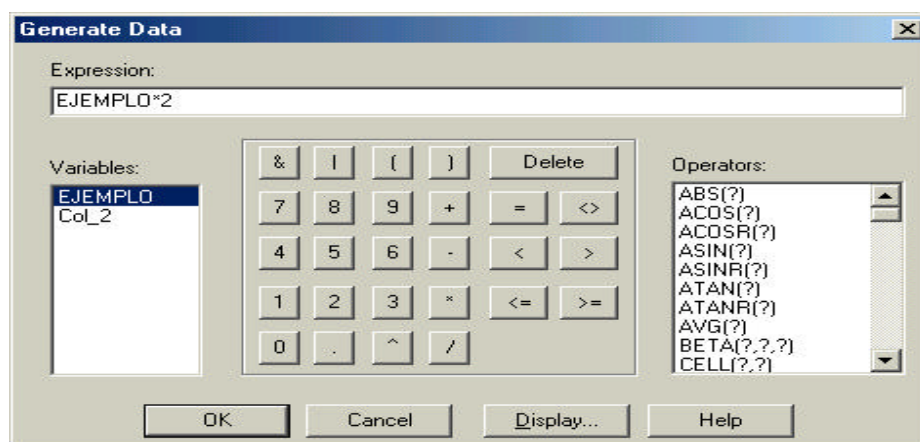
En esta pantalla escribiremos el nombre de la variable (máximo 32 caracteres, sin blancos ni signos especiales y utilizando siempre una letra como primer carácter), y el tipo de variable (*numeric* si vamos a analizar números). Tras pulsar OK ya estamos en condiciones de introducir los datos en las distintas celdas que componen la columna.

A continuación vemos como se han introducido un conjunto de datos agrupados en la variable *EJEMPLO*

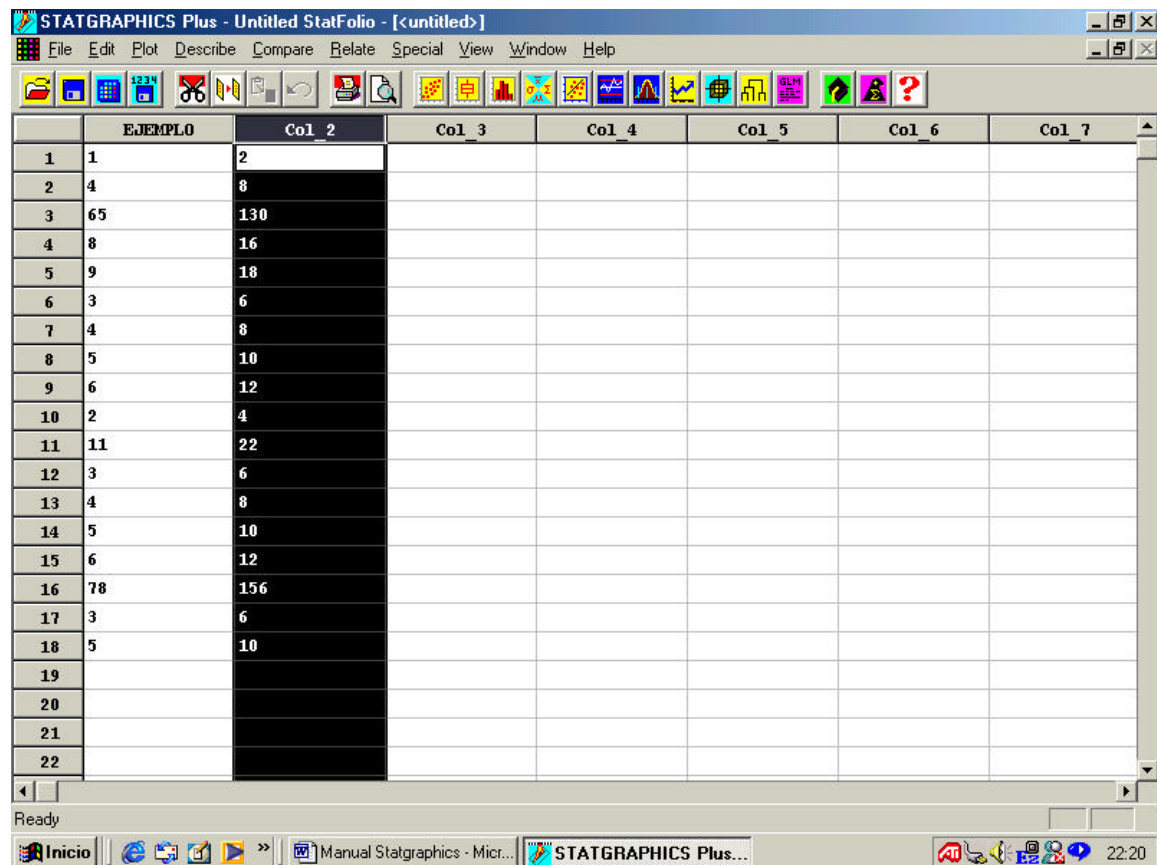


Statgraphics permite introducir columnas calculadas como una transformación de otras columnas previamente definidas. Para ello realizaremos las siguientes operaciones:

1. Seleccionamos la columna donde queremos que aparezcan los datos calculados
2. Clickeamos con el botón derecho del ratón y elegimos la opción *Generate Data* del menú que aparece
3. Componemos, en la ventana que aparece, la expresión para el cálculo de los nuevos datos: (en este caso multiplicaremos por 2 la variable *EJEMPLO*)



Al pulsar OK nos aparecerá en la ventana de datos el cálculo deseado:



	EJEMPLO	Col_2	Col_3	Col_4	Col_5	Col_6	Col_7
1	1	2					
2	4	8					
3	65	130					
4	8	16					
5	9	18					
6	3	6					
7	4	8					
8	5	10					
9	6	12					
10	2	4					
11	11	22					
12	3	6					
13	4	8					
14	5	10					
15	6	12					
16	78	156					
17	3	6					
18	5	10					
19							
20							
21							
22							

Los ficheros de datos generados pueden almacenarse para análisis posteriores. Para ello, en el menú FILE seleccionaremos SAVE DATA FILE AS... y elegiremos el nombre y la ubicación del archivo deseada. (Podrán recuperarse posteriormente con la opción OPEN DATA FILE del menú FILE)

UNIVERSIDAD CARLOS III DE MADRID
MASTER EN CALIDAD TOTAL

PRÁCTICAS DE ESTADÍSTICA I
MANUAL DE STATGRAPHICS

II. ESTADÍSTICA DESCRIPTIVA / GRÁFICOS DE DATOS

La Estadística Descriptiva se ocupa de presentar, de forma resumida, la información más importante de un conjunto de datos. Para ello se calculan sus medidas centrales (media, mediana...) y se da una medida de cómo están los datos dispersos en torno a esos valores centrales (varianza, desviación típica, rango...). Asimismo, tras un análisis descriptivo, se dispondrá de una representación de los datos en forma de gráficos, de forma que sea posible detectar valores atípicos, tendencias o agrupaciones.

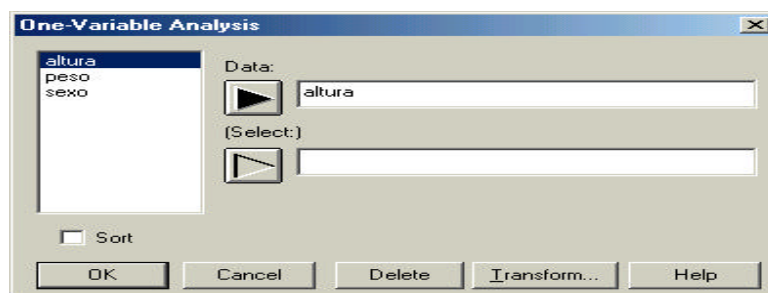
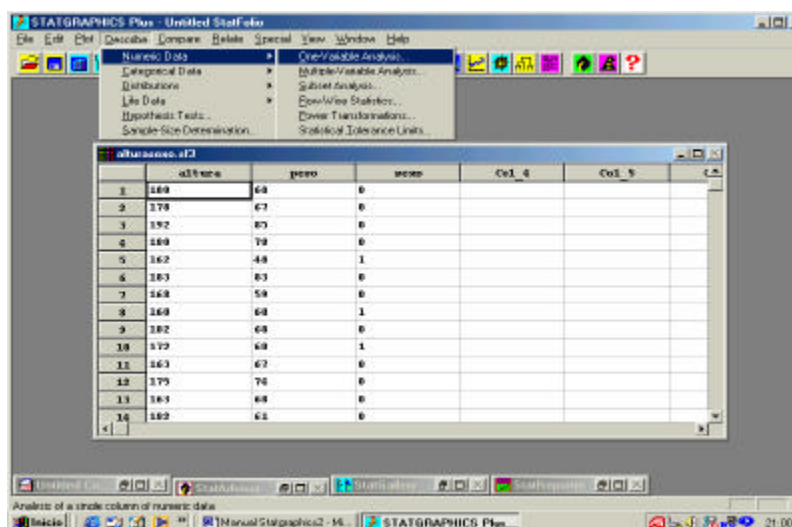
Las diferentes opciones de análisis descriptivo de las que dispone Statgraphics están incluidas en la opción DESCRIBE de la barra de menú.

A continuación se muestran las opciones más importantes de un análisis descriptivo de los datos.

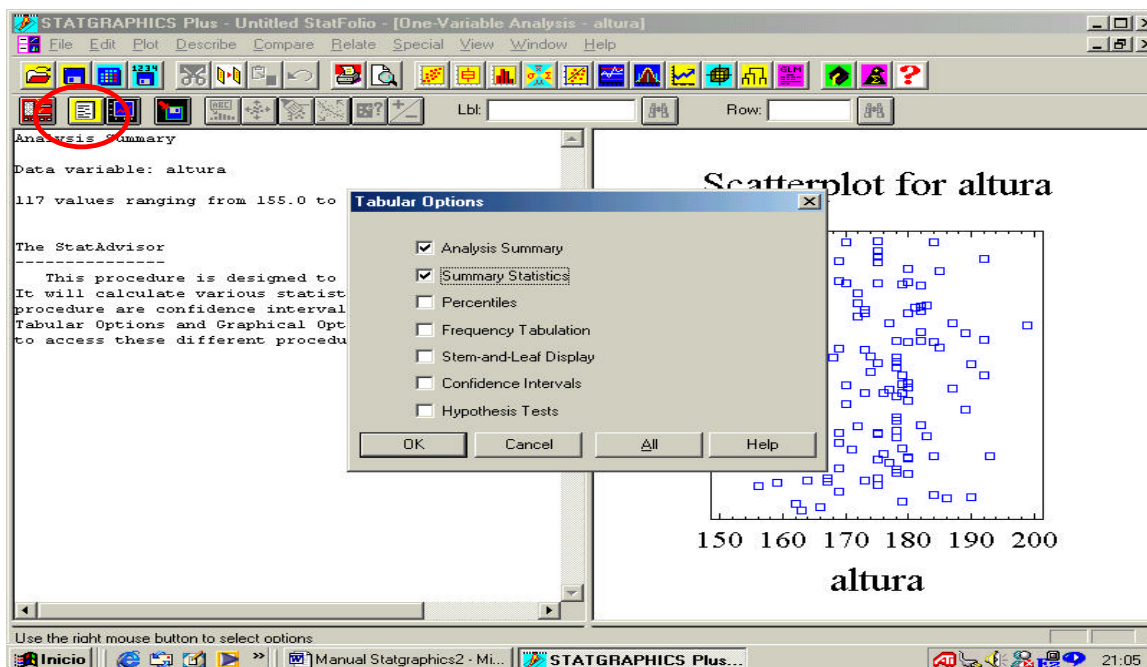
Resumen estadístico

El resumen estadístico (SUMMARY STATISTICS) nos reproduce hasta 19 estadísticos (valores numéricos característicos) de un conjunto de datos.

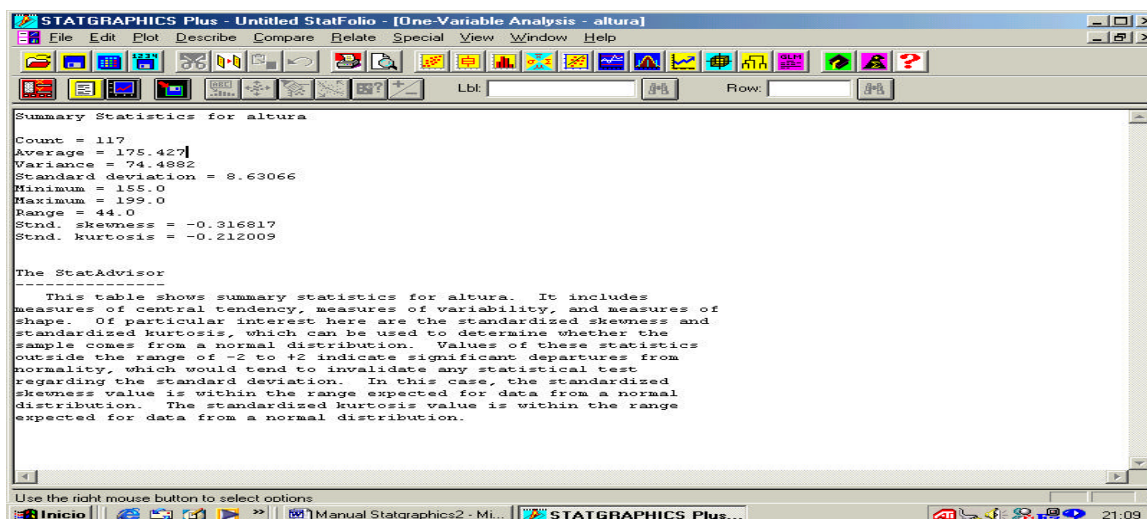
Para ello, en la pantalla de entrada de datos tendremos que introducir la variable que se quiere analizar, tal y como aparece a continuación:



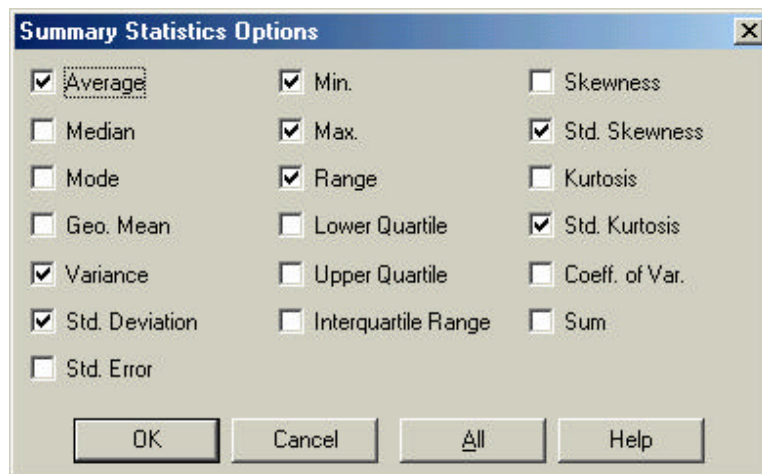
Una vez seleccionada la variable a analizar, debe seleccionarse la opción de SUMMARY STATISTICS en el menú de TABULAR OPTIONS:



Por defecto, aparecerán calculados los estadísticos de uso más común, como puede verse en la figura que sigue:



Sin embargo pueden seleccionarse otros estadísticos que Statgraphics calcula sin más que clicar con el botón derecho del ratón sobre el panel de SUMMARY STATISTICS y activar la opción de PANE OPTIONS:



Activando la opción de cualquiera de los estadísticos que están incluidos en la ventana que aparece, el resultado de su cálculo se mostrará inmediatamente por pantalla al clicar OK.

El SUMMARY STATISTICS puede obtenerse simultáneamente para varias variables, sin más que entrar en el análisis múltiple de variables: DESCRIBE...NUMERIC DATA...MÚLTIPLE VARIABLE ANÁLISIS.

Tabla de frecuencias

La tabla de frecuencias nos permite resumir la distribución de los datos contenidos en una variable. Al igual que el SUMMARY STATISTICS, la opción de la tabla de frecuencias (FREQUENCY TABULATION) se activa en el menú de TABULAR OPTIONS del análisis descriptivo de una variable. Como resultado del análisis, Statgraphics crea una serie de intervalos que constituyen una partición del rango de los datos estudiados; la tabla nos dará información del número de datos que tienen su valor dentro de cada intervalo.

STATGRAPHICS Plus - Untitled StatFolio - [One-Variable Analysis - altura]

File Edit Plot Describe Compare Relate Special View Window Help

Frequency Tabulation for altura

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below	150.0	150.0	150.0	0	0.0000	0	0.0000
1	150.0	157.5	153.75	3	0.0256	3	0.0256
2	157.5	165.0	161.25	16	0.1368	19	0.1624
3	165.0	172.5	168.75	20	0.1709	39	0.3333
4	172.5	180.0	176.25	51	0.4359	90	0.7692
5	180.0	187.5	183.75	18	0.1538	108	0.9231
6	187.5	195.0	191.25	8	0.0684	116	0.9915
7	195.0	202.5	198.75	1	0.0085	117	1.0000
8	202.5	210.0	206.25	0	0.0000	117	1.0000
above	210.0	210.0	210.0	0	0.0000	117	1.0000

Mean = 175.427 Standard deviation = 8.63066

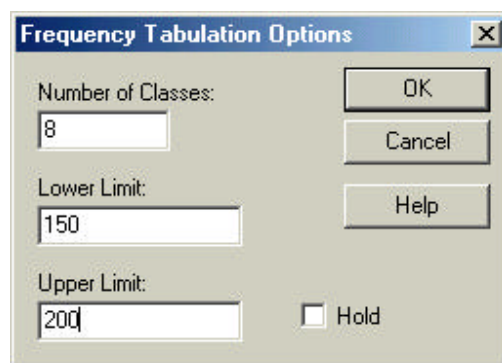
The StatAdvisor

This option performs a frequency tabulation by dividing the range of altura into equal width intervals and counting the number of data values in each interval. The frequencies show the number of data values in each interval, while the relative frequencies show the proportions in each interval. You can change the definition of the intervals by pressing the alternate mouse button and selecting Pane Options. You can see the results of the tabulation graphically by

Use the right mouse button to select options

Inicio Manual Statgraphics2 - Mi... STATGRAPHICS Plus... 21:33

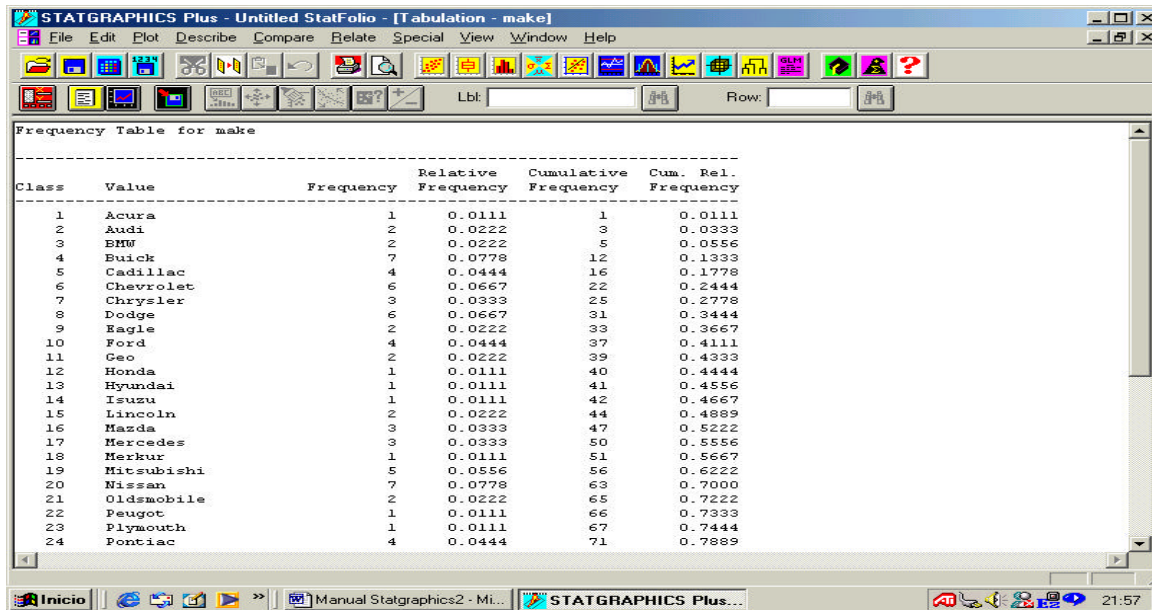
El número de observaciones en cada intervalo será la frecuencia absoluta, mientras que el porcentaje que esas observaciones representa frente al total se llama frecuencia relativa. (El programa presenta también las frecuencias acumuladas para cada una de los intervalos). El número de intervalos (también llamados clases) en los que se divide el rango de los datos puede modificarse clickeando con el botón derecho del ratón sobre la tabla y seleccionando la opción PANE OPTIONS:



La tabla de frecuencias no sólo puede aplicarse a datos numéricos, sino también a variables cualitativas. Así en el fichero *cardata.sf* se recogen diferentes variables de automóviles junto con el nombre de su fabricante:

	mpg	cylinders	displace	horsepower	weight	origin	make
1	20	4	147	106	3340	3	Nissan
2	29	4	98	90	2241	3	Nissan
3	25	4	121	97	2780	3	Nissan
4	20	4	144	107	3295	3	Mitsubishi
5	21	6	182	142	3085	3	Mitsubishi
6	23	6	182	160	3096	3	Nissan
7	27	4	98	90	2398	3	Nissan
8	29	4	91	81	2281	3	Mitsubishi
9	27	4	110	90	2250	3	Subaru
10	21	6	164	161	3174	1	Sterling
11	35	3	74	66	1755	3	Subaru
12	28	4	122	96	2443	1	Pontiac
13	30	4	99	74	2245	1	Pontiac
14	21	4	122	125	2745	2	Saab
15	22	4	122	125	3032	2	Saab
16	24	6	232	165	3371	1	Oldsmobile
17	25	6	174	130	3094	1	Oldsmobile
18	23	4	147	140	2700	3	Nissan
19	23	6	182	165	3149	3	Nissan
20	32	4	123	135	2420	1	Plymouth
21	23	4	117	110	2470	2	Peugot
22	27	4	152	110	2602	1	Pontiac

Veamos como podemos aplicar la tabla de frecuencias a la variable que contiene el fabricante del vehículo. Para ello se sigue DESCRIBE...CATEGORICAL DATA...TABULATION y se selecciona la opción FREQUENCY TABULATION del menú de *Tabular Options*. El resultado es el que continuación se muestra:

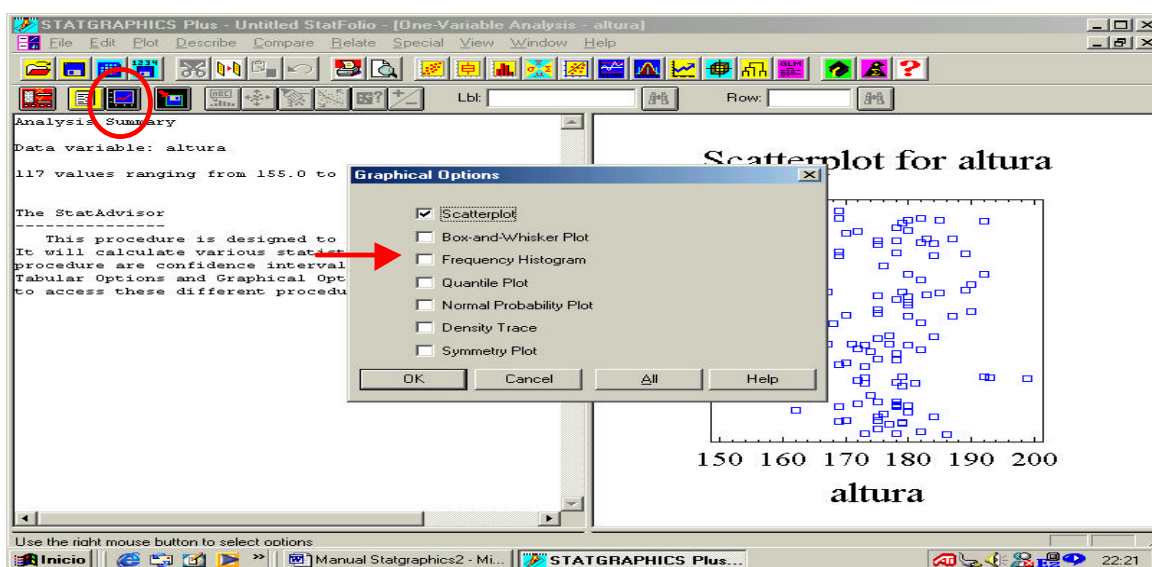


Donde obtenemos información sumaria de los vehículos que aporta cada fabricante a la muestra y de su frecuencia de aparición.

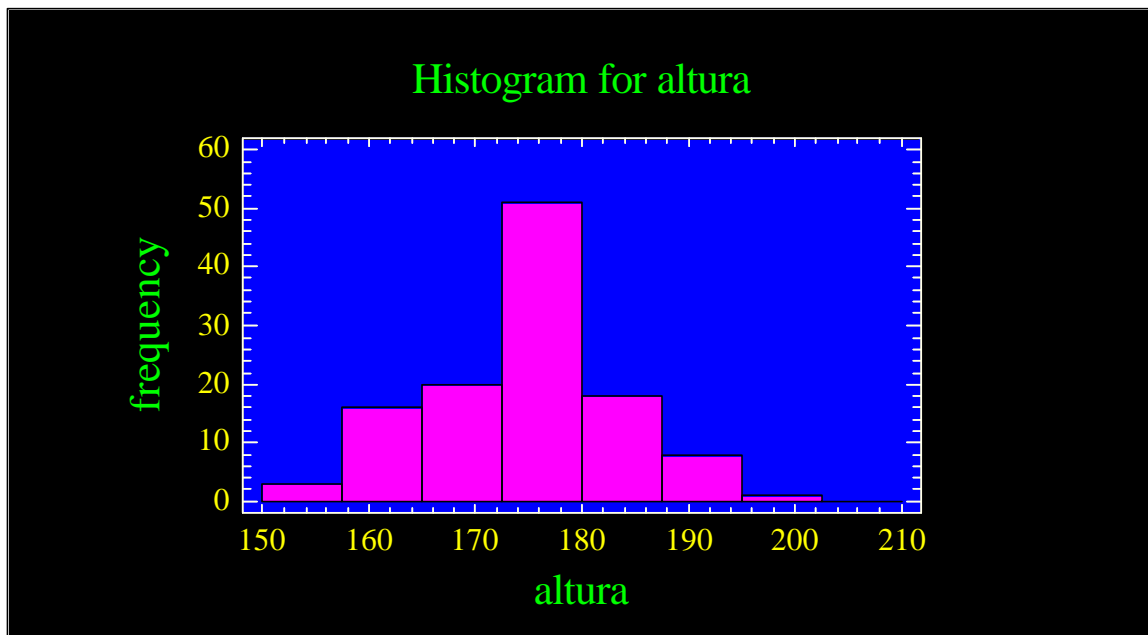
Histograma de frecuencias

Los histogramas de frecuencias son representaciones gráficas de las tablas de frecuencias estudiadas con anterioridad, donde a cada intervalo o clase en que se divide el rango de los datos, se le asigna una barra cuya altura es proporcional a la frecuencia de aparición de sus elementos.

El histograma se encuentra en las opciones gráficas del menú DESCRIBE... NUMERIC DATA... ONE VARIABLE ANÁLISIS, tal y como puede verse en la figura que sigue:



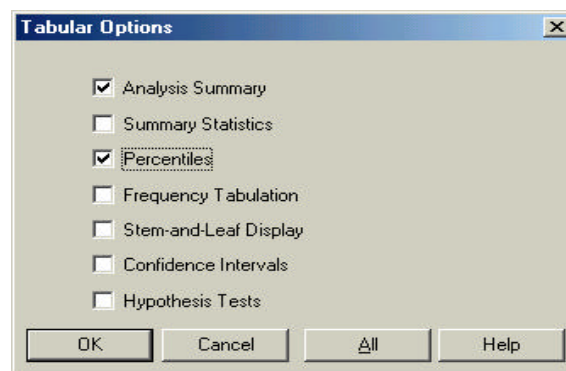
El resultado se muestra en la siguiente pantalla:



Donde podemos ver que el histograma presenta información sobre la variable analizada. En los datos analizados, la altura más frecuente entre los individuos analizados está entre 172 y 182 cms.

Percentiles

Los percentiles de una variable proporcionan información sobre como están distribuidos los datos estudiados. El percentil de orden k de una distribución es un valor que es mayor que el k% de los valores que toma la variable. Así el percentil 10 es aquel valor de los datos estudiados que es mayor que el 10% de las observaciones. Son importantes los percentiles 25 (cuartil inferior), 50 (mediana) y 75 (cuartil superior). Los percentiles pueden obtenerse en la opción *Tabular options* del menú DESCRIBE... NUMERIC DATA... ONE VARIABLE ANALYSIS.



El resultado es el siguiente:

Percentiles for altura

1.0% = 156.0
5.0% = 161.0
10.0% = 163.0
25.0% = 169.0
50.0% = 176.0
75.0% = 180.0
90.0% = 186.0
95.0% = 190.0
99.0% = 193.0

The StatAdvisor

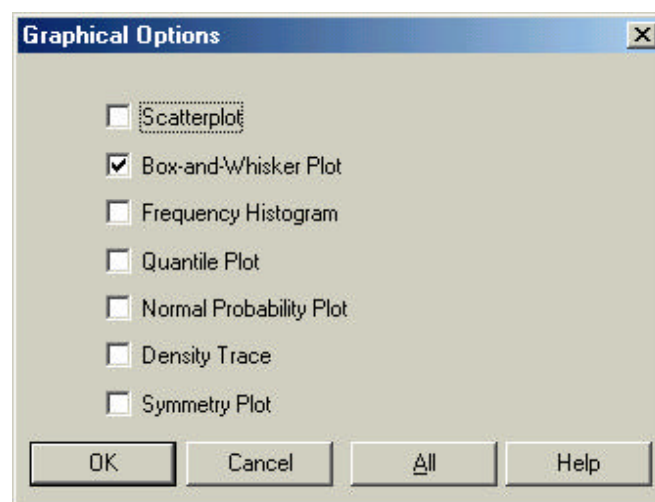
This pane shows sample percentiles for altura. The percentiles are values below which specific percentages of the data are found. You can see the percentiles graphically by selecting Quantile Plot from the list of Graphical Options.

Diagrama de la caja

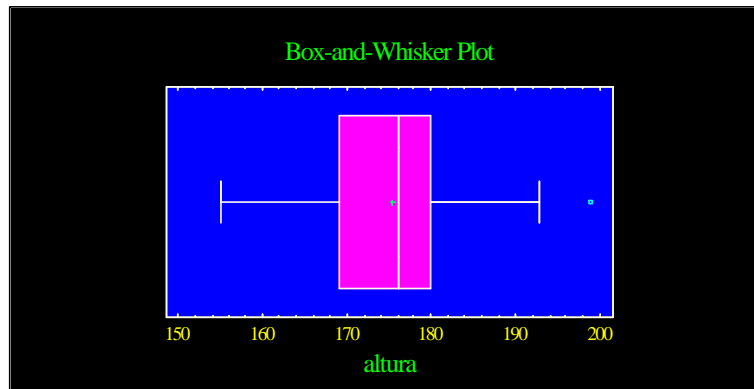
El diagrama de la caja es una representación gráfica de una variable en la que a partir de sus percentiles se obtiene información sobre la distribución de sus observaciones (concentración o dispersión de los datos o existencia de valores atípicos).

El diagrama de la caja se construye a partir de los percentiles 25%, 50% (mediana) y 75 %. Como medida de la dispersión se utiliza el rango intercuartílico (percentil 75 % - percentil 25%) de manera que cualquier dato que se aleje de los percentiles 25 ó 75% una distancia superior a 1,5 veces el rango intercuartílico se considera atípico.

Para obtener el diagrama de la caja de una variable se sigue la ruta DESCRIBE...NUMERIC DATA... ONE VARIABLE ANÁLISIS...y se selecciona BOX AND WHISKER PLOT en el menú de opciones gráficas.

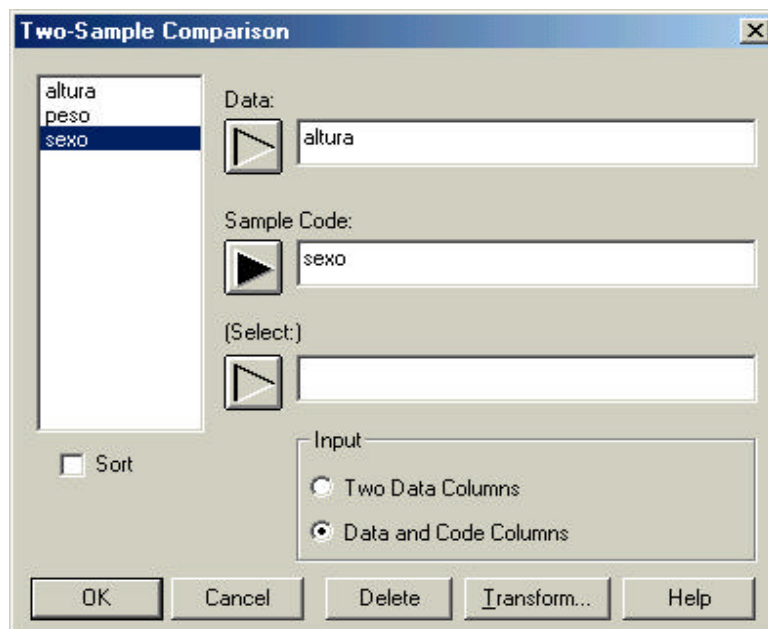


El resultado es el siguiente:

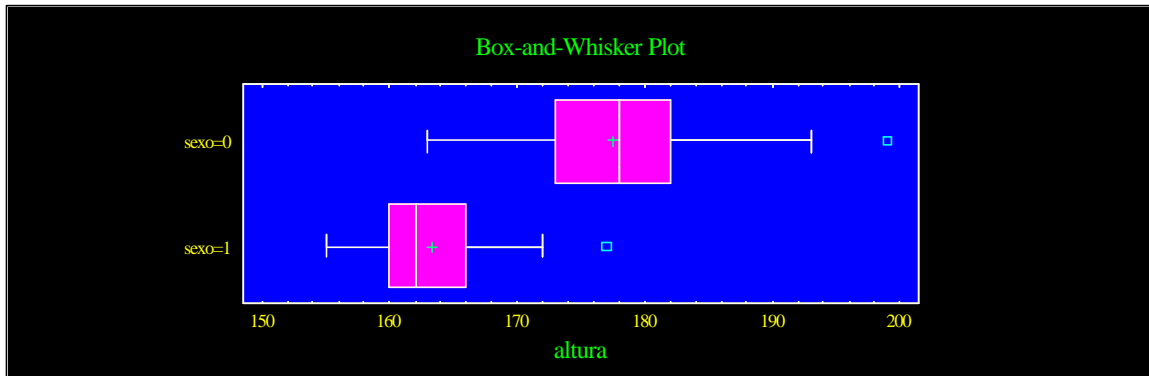


En el diagrama se debe observar: la forma de los rectángulos que forman la caja (cuanto más estrechos sean, indicarán una mayor concentración de datos); la posición de la media, marcada con una cruz, respecto de la mediana, línea central de la caja (la coincidencia de ambas indica simetría de la distribución), y la existencia de valores áticos (quedan fuera de los segmentos de longitud 1,5 veces el rango intercuartílico colocados a derecha a izquierda).

En ocasiones puede ser útil observar simultáneamente dos diagramas de la caja: por ejemplo para la variable altura en la que se separan los valores de las observaciones en función del diferente sexo de los individuos. Esta opción está disponible en el menú COMPARE...TWO SAMPLES...TWO SAMPLES COMPARISSON, seleccionando en la ventana que aparece de acuerdo con la disposición de nuestros datos.



El resultado obtenido (tras seleccionar la opción de BOX AND WHISKER PLOT en el *Graphical Options*) es el siguiente:



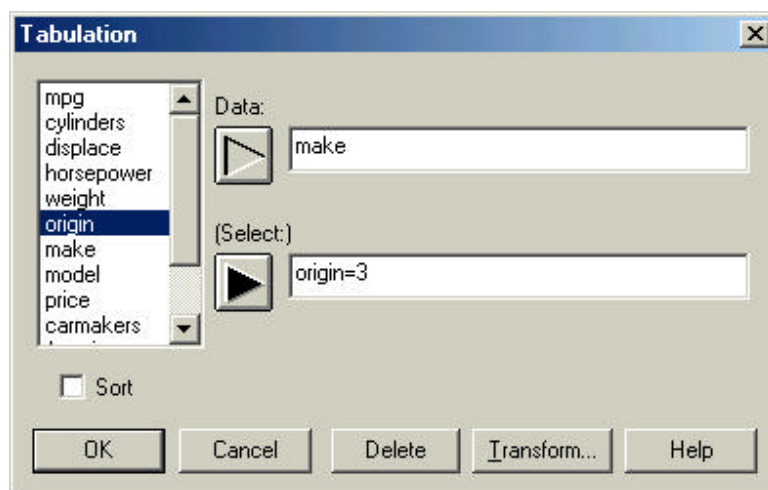
De forma que es posible analizar simultáneamente una variable discriminada según el criterio de selección.

(Esta misma representación simultanea de gráficos también está disponible cuando se quiera observar el histograma de una variable).

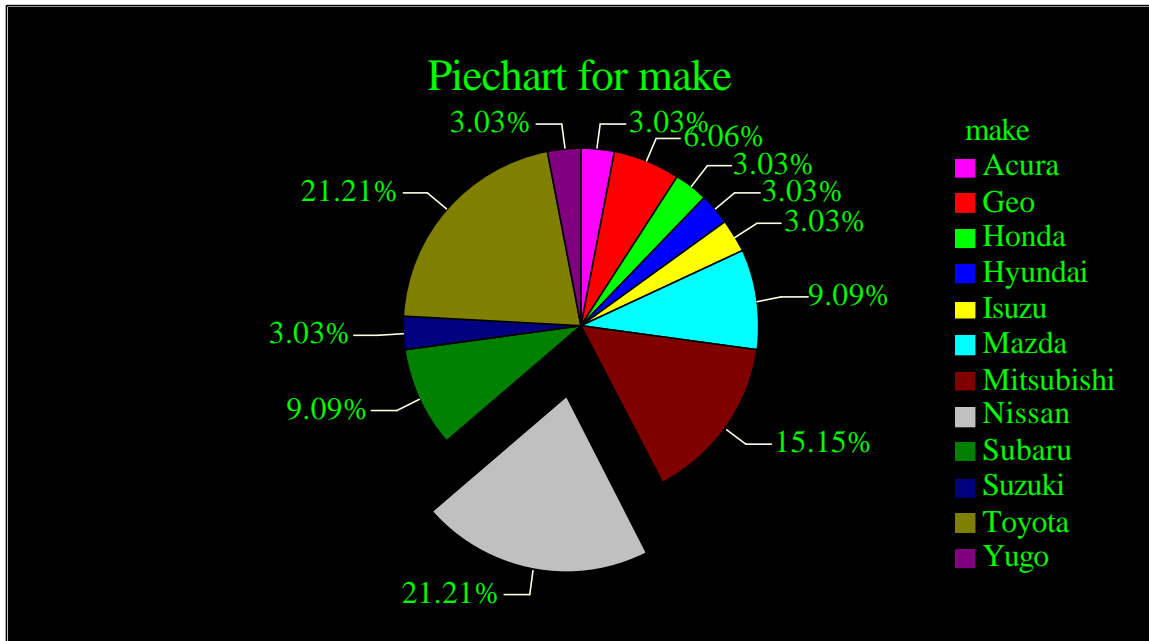
Diagrama de tarta o gráfico de sectores

El diagrama de tarta (Piechart) proporciona información sobre las categorías en que puede dividirse una variable (y la importancia relativa de las mismas)

Para ensayara su aplicación utilizaremos el fichero *cardata.sf* que contiene datos de diferentes automóviles fabricados en el mundo. Siguiendo el menú DESCRIBE...CATEGORICAL DATA... TABULATION



y activamos la opción de PIECHART en el menú del *Graphical options* veremos el diagrama de sectores que nos dará la distribución de las diferentes categorías en que puede dividirse la variable *make* (que contiene marcas de coches) cuando la variable *origin* toma el valor 3 (lo que equivale a estudiar únicamente coches fabricados en Japón)

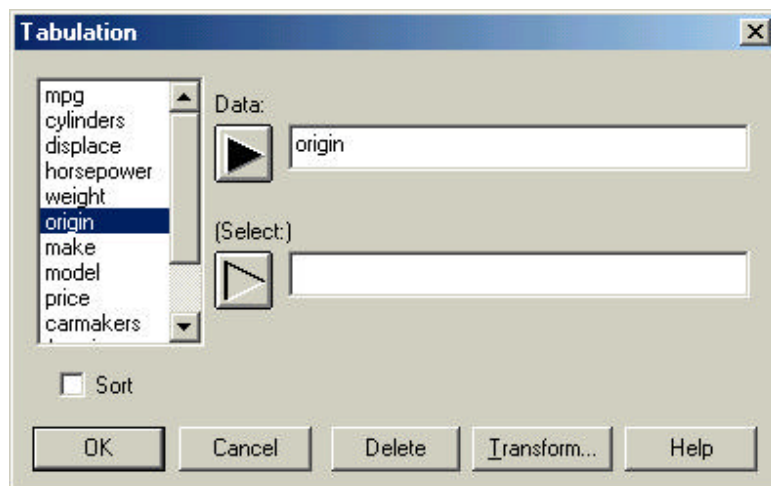


De modo que es posible analizar gráficamente la importancia relativa de los fabricantes de coches radicados en Japón.

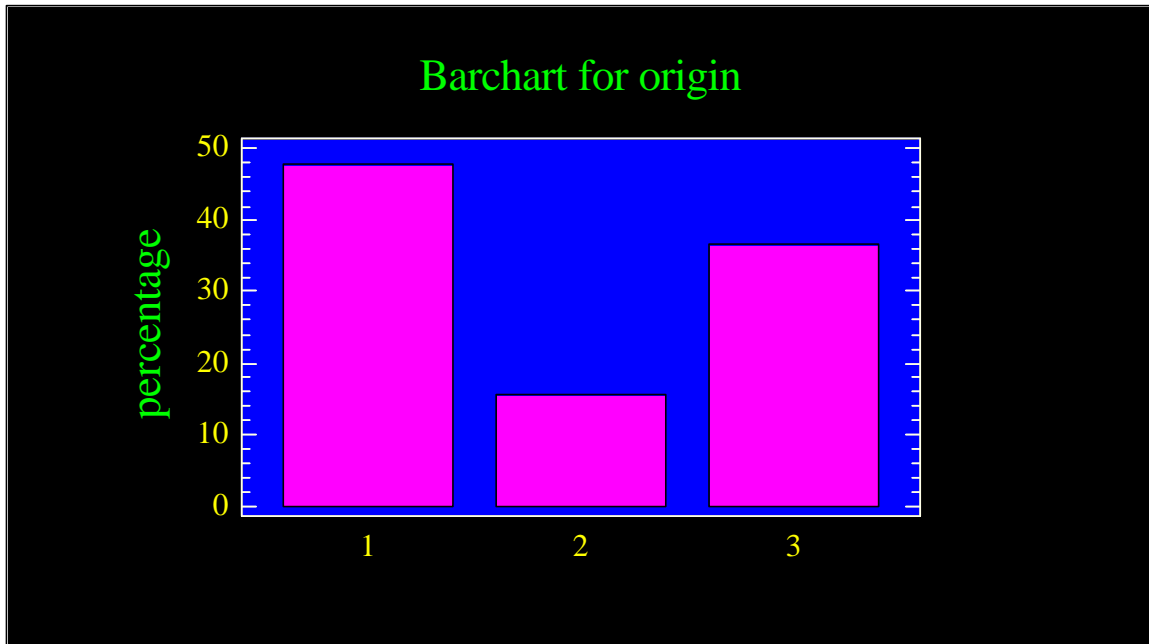
Diagrama de barras

Mediante esta gráfico es posible obtener información sobre las diferentes categorías en que puede dividirse una variable.

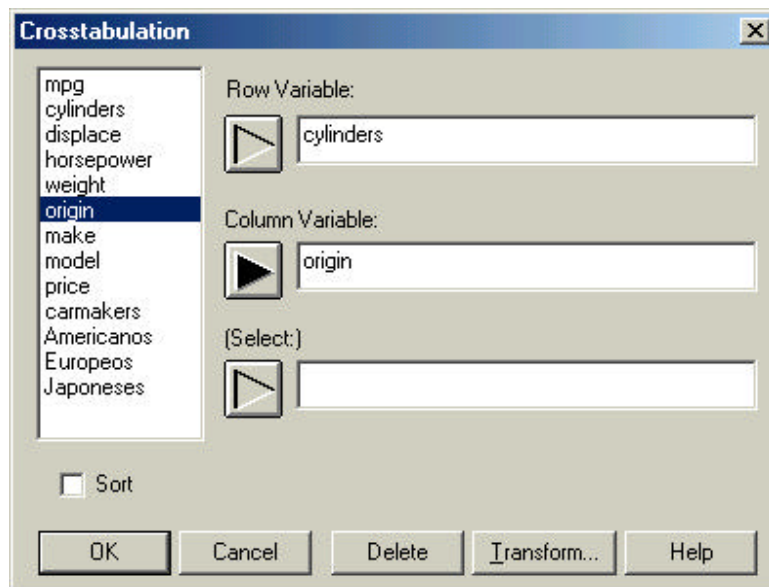
Así por ejemplo en el fichero *cardata.sf* podemos analizar los coches fabricados en América (*origin* = 1), en Europa (*origin* = 2) o en Japón (*origin* = 3) sin más que hacer DESCRIBE....CATEGORICAL DATA..TABULATION



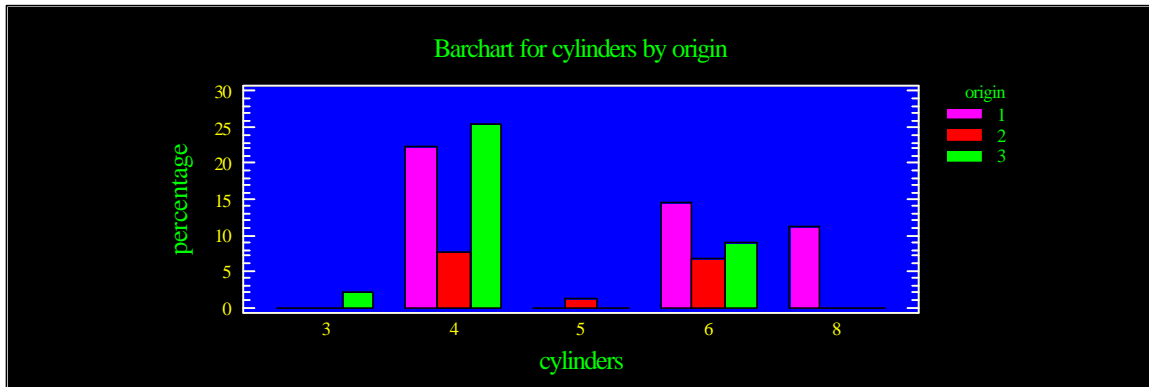
Y seleccionar la opción BARCHART del *Graphical Options*



La representación de gráfico de barras permite cruzar dos variables y analizar por ejemplo el número de cilindros del automóvil (variable *cylinder*) según su origen (variable *origin*). Para ello seleccionamos el menú DESCRIBE...CATEGORICAL DATA... CROSSTABULATION.



Activando la opción de BARTCHART del *Graphical Options*, se obtiene:

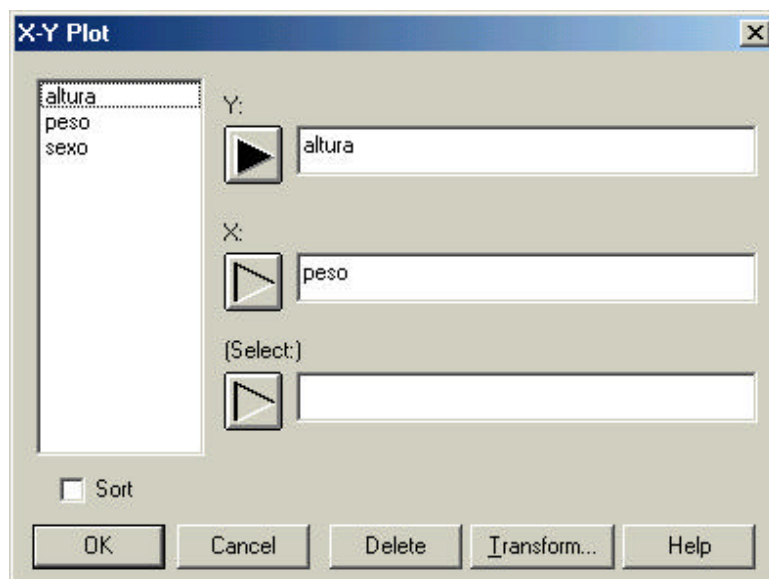


Lo que nos permite hacer un análisis de las dos variables: por ejemplo puede verse que coches con 8 cilindros sólo son fabricados en América.

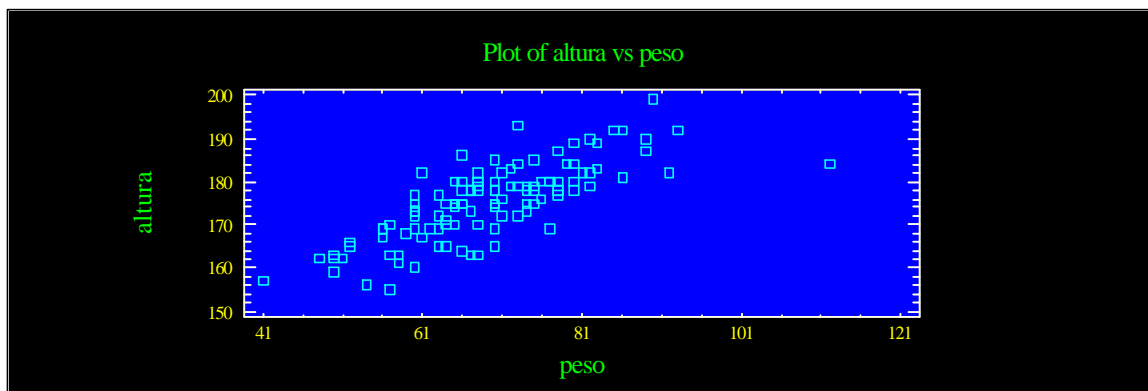
Gráficos de dispersión o Scatterplots

Los gráficos de dispersión proporcionan información acerca de la distribución de una variable. Son especialmente útiles los gráficos XY, pues permiten analizar la relación entre dos variables

Para visualizarlos se sigue el menú PLOT...SCATTER PLOT...X-Y PLOT

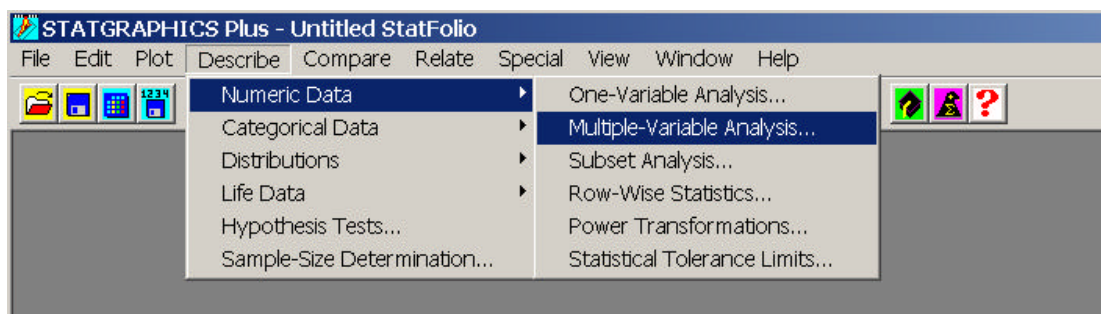


Como resultado obtenemos en diagrama que nos permite ver la distribución conjunta de ambas variables, y por tanto su relación lineal, en la que al aumentar la altura de una persona también lo hará su peso. (Como puede verse, también está permitida la selección de valores de las variables mediante una variable de selección, en el caso estudiado *sexo*)

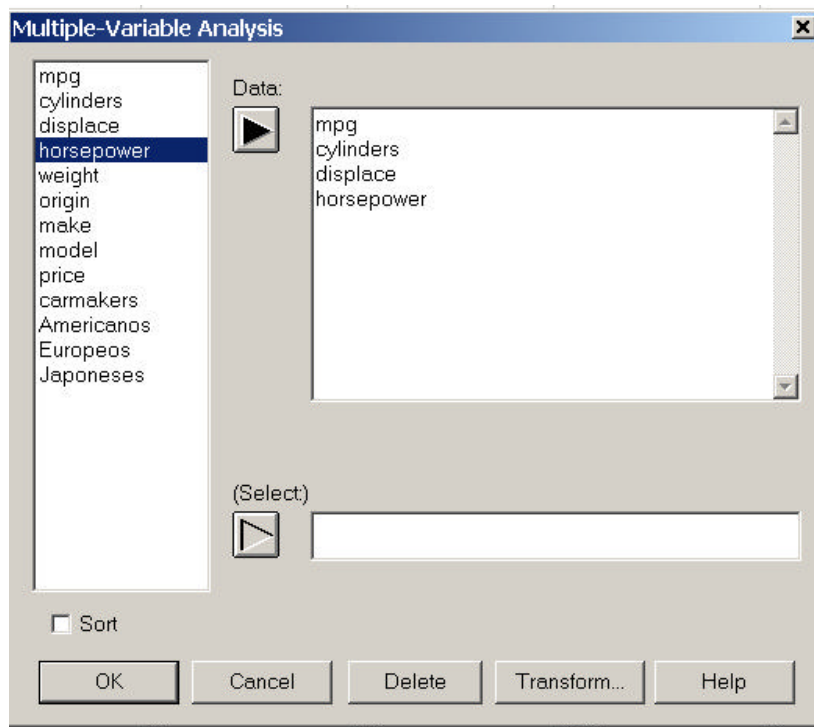


Correlaciones (Con CARDATA)

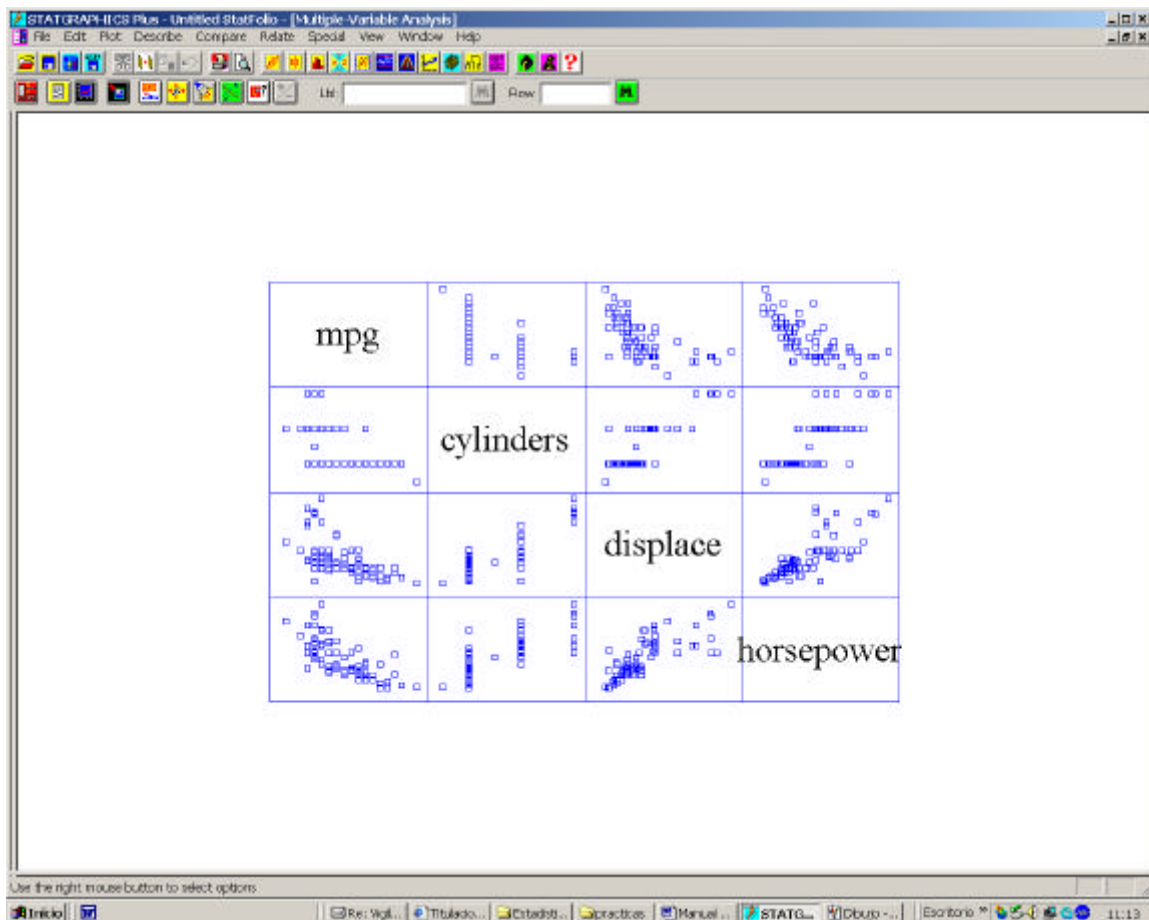
Vamos a



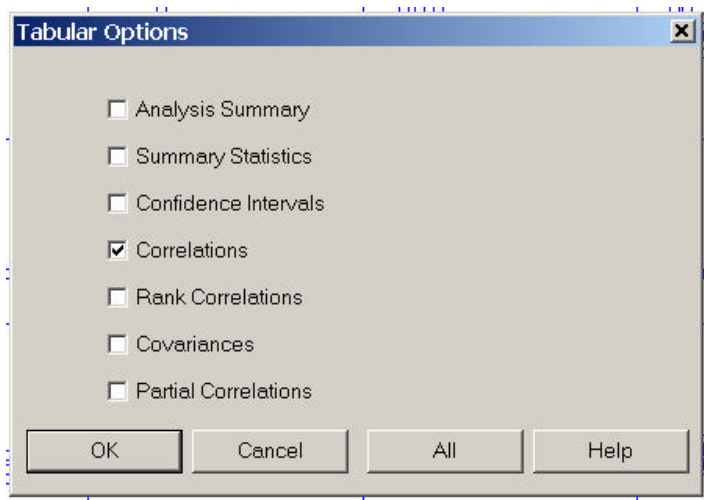
Y en el menú introducimos las variables que nos interesen pero siempre deben ser numéricas. Las categóricas no pueden usarse y darían error.



Al ejecutar este menú nos sale un gráfico matricial de dispersión, con los gráficos de dispersión de cada pareja de variables:



Y si se pincha en Tabular Options nos aparece la opción correlación



El resultado es

Correlations

	mpg	cylinders	displace
mpg		-0,6311 (86) 0,0000	-0,6302 (86) 0,0000
cylinders	-0,6311 (86) 0,0000		0,8770 (86) 0,0000
displace	-0,6302 (86) 0,0000	0,8770 (86) 0,0000	
horsepower	-0,7131 (86) 0,0000	0,7643 (86) 0,0000	0,7592 (86) 0,0000

	horsepower
mpg	-0,7131 (86) 0,0000
cylinders	0,7643 (86) 0,0000
displace	0,7592 (86) 0,0000
horsepower	

Que nos indica por ejemplo que entre la potencia en caballos de vapor y los cilindros hay una correlación positiva de 0.76