Índice general

Capítulo 2

Presentación del paquete estadístico Statgraphics. Estadística Descriptiva

2.1. Ventanas de Statgraphics

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, etc.

El programa trabaja en un entorno WINDOWS y su pantalla principal (a la que se accede ejecutando el programa SGWIN.EXE o directamente clicando sobre el icono Correspondiente), es la que aparece en la figura 2.1.

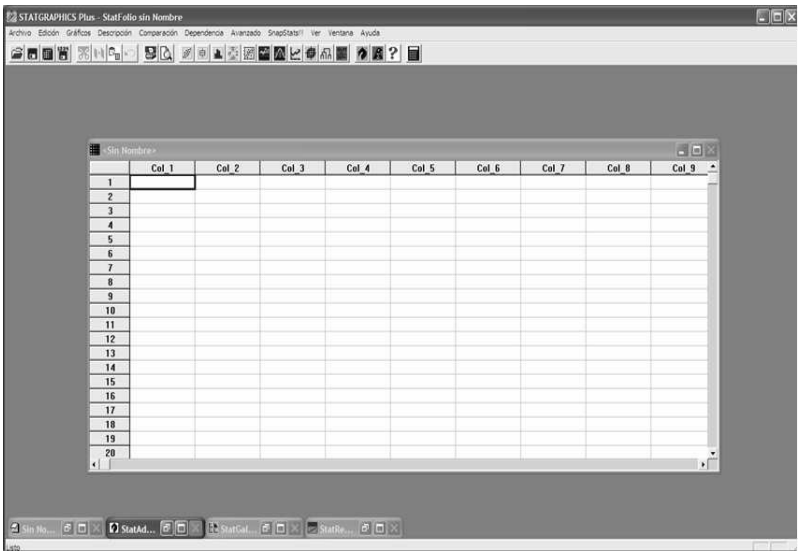

Figura 2.1: Ventana de Statgraphics

Para salir del programa seleccionamos en la barra de menú Archivo...Salir, la secuencia de teclas Alt+F4, 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.

2.1.1. Barras de menú, de herramientas y de tareas

Figura 2.2: Barra de menú

La barra de menú siempre estará disponible al utilizar el programa, de forma que sea posible seleccionar el análisis deseado. Al clicar con el ratón sobra cada una de las palabras que componen la barra, aparecerá un menú desplegable con otras opciones asociadas. Así tendremos:

- Archivo: permite realizar operaciones de carácter general: abrir, cerrar o grabar ficheros, imprimir y salir de Statgraphics.
- Edición: como en otras aplicaciones en entorno Windows, este menú esta asociado a diversas opciones de edición: cortar, copiar, pegar, deshacer, etc.
- Gráficos, Descripción, Comparación, Dependencia,Avanzado y SnapStats!!: al presionar con el ratón sobre ellos tendremos acceso a diversos menús de análisis de Statgraphics que se irán analizando a lo largo del capítulo.
- Ver, Ventana y Ayuda: tienen disponible varias opciones de formato y ayuda, de forma similar a otras aplicaciones que trabajan en el mismo entorno.

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.

La barra de tareas (figura 2.3 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.

Figura 2.3: Barra de tareas

- 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. Clicando 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).
- Comentarios (Sin nombre) y Statreporter: opciones de Statgrafics que permiten introducir los comentarios de usuario para su posterior edición. "Ventana de datos: hoja de cálculo que contiene los datos que se van analizar. Pueden introducirse directamente desde el teclado o recurarse desde un fichero ya grabado. (ARCHIVO...ABRIR...ABRIR DATOS ó crtl+F12)

Al conjunto de los elementos anteriores se le denomina Statfolio, que puede almacenarse bajo un nombre único (fichero .spg) activando la opción ARCHIVO...GUARDAR...GUARDAR STATFOLIO ó may+F11. 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.

2.2. Introducción de datos

Los datos que van 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 de 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 clicamos sobre la etiqueta de la columna (Inicialmente será Col_1), figura 2.4.
- Pulsamos con el botón derecho del ratón sobre la columna seleccionada. Aparecerá un menú del que seleccionamos la opción Modificar Columna (figura 2.6).

- 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 (Numérica si vamos a analizar números). Tras pulsar OK ya estamos en condiciones de introducir los datos en las distintas celdas que componen la columna.

2.3. Transformación de datos. Recodificación

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

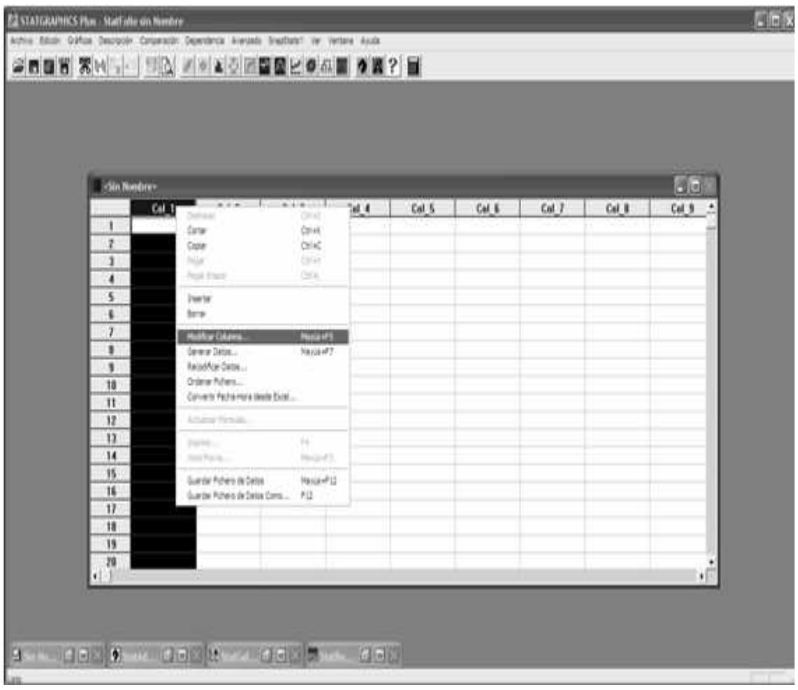

Figura 2.4: Modificar columna (ventana general)

Figura 2.5: Modificar columna (ventana específica)

1. Seleccionamos la columna donde queremos que aparezcan los datos calculado.
2. Clicamos con el botón derecho del ratón y elegimos la opción Generar Datos 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 los datos de una variable previamente introducida, llamada SALARIO).

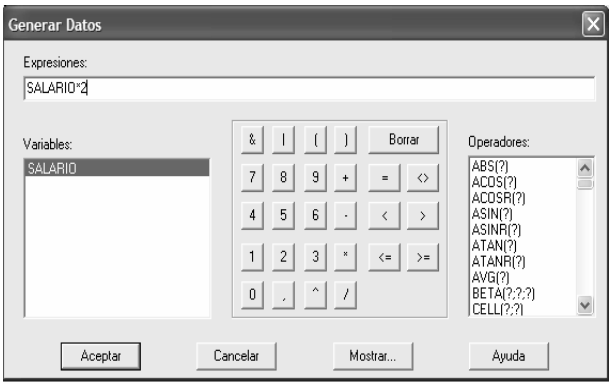

Figura 2.6: Transformar datos

Al pulsar OK nos aparecerá en la ventana de datos el cálculo deseado. Los ficheros de datos generados pueden almacenarse para análisis posteriores. Para ello, en el menú FILE seleccionaremos GUARDAR DATOS COMO... ó F12, y elegiremos el nombre y la ubicación del archivo deseada. (Podrán recuperarse posteriormente con la opción ABRIR DATOS del menú ARCHIVO)

Si queremos clasificar los datos de salario por categorías, habremos de activar la opción RECODIFICAR DATOS (figura 2.7).

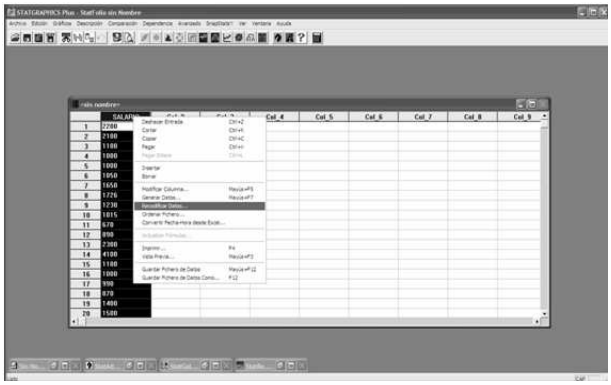

Figura 2.7: Recodificar datos

A continuación se activará un panel en el que habremos de definir los límites de las clases que queremos considerar. En nuestro caso, salarios de [500, 750), [750, 1200), [1200, 2200) y [2200, 8000]. Los límites

inferior y superior de la serie se han puesto a conveniencia, ya que no hay salarios menores ni mayores que los valores respectivos (figura 2.8).

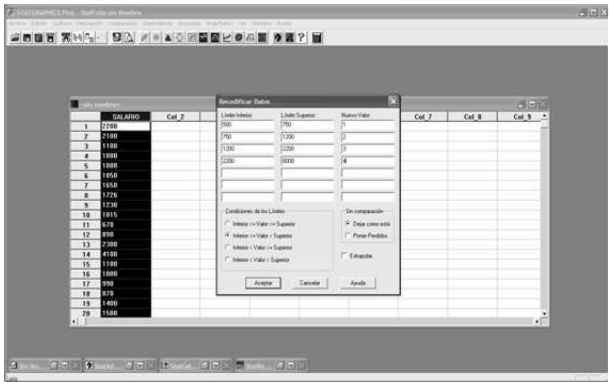

Figura 2.8: Panel de códigos

Los códigos de la variable generada son: 1, 2, 3 y 4. Hay que tener en cuenta que de esta forma se ELIMINA la variable original, con lo que se recomienda realizar la remodificaciones sobre variables-çopia"de las originales.

2.4. Importación y exportación de datos

Statgraphics puede importar y exportar datos en formato texto (txt) y en formato de EXCEL. Para ello, desde el menú ARCHIVO, podremos importar datos mediante la opción ABRIR DATOS, seleccionando el formato en el que estén los datos que queramos analizar (figura 2.9).

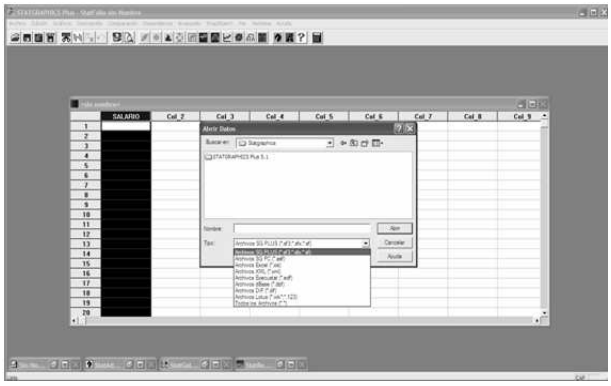

Figura 2.9: Importar datos

Como norma general, los datos estarán en formato matricial (fila y columna), y sin títulos ni comentarios. También es posible utilizar las funciones de copiado y pegado estándar de Windows, para incorporar datos a nuestra hoja de Statgraphics.

Para exportar datos accedemos al menú ARCHIVO y seleccionamos GUARDAR DATOS COMO (figura 2.10).

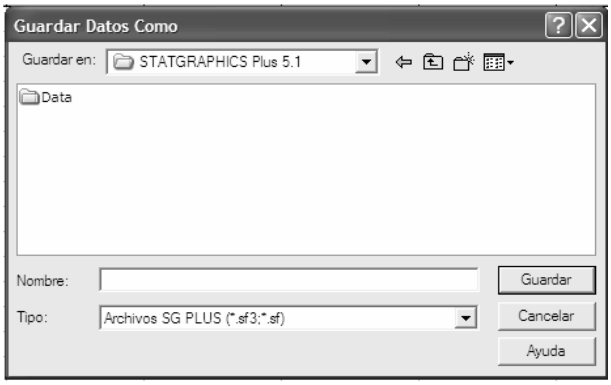

Figura 2.10: Exportar datos

Los formatos de exportación de que dispone Statgraphics son TXT y EXCEL, los cuales se podrán incorporar a otros programas sin mucha dificultad.

2.5. Tabla de frecuencias. Medidas descriptivas

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 DESCRIPCIÓN de la barra de menú. A continuación se muestran las opciones más importantes de un análisis descriptivo de los datos.

2.5.1. Resumen estadístico

El resumen estadístico (ANALISIS UNIDIMENSIONAL) produce hasta 19 estadísticos (media, desviación típica, varianza, etc.); cálculo de percentiles; diagrama de tallo y hojas; histograma, 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 (figura 2.11).

Una vez seleccionada la variable a analizar y pulsado el botón ACEPTAR, se obtiene la salida que aparece en la figura 2.12.

Figura 2.11: Panel de ANALISIS UNIDIMENSIONAL

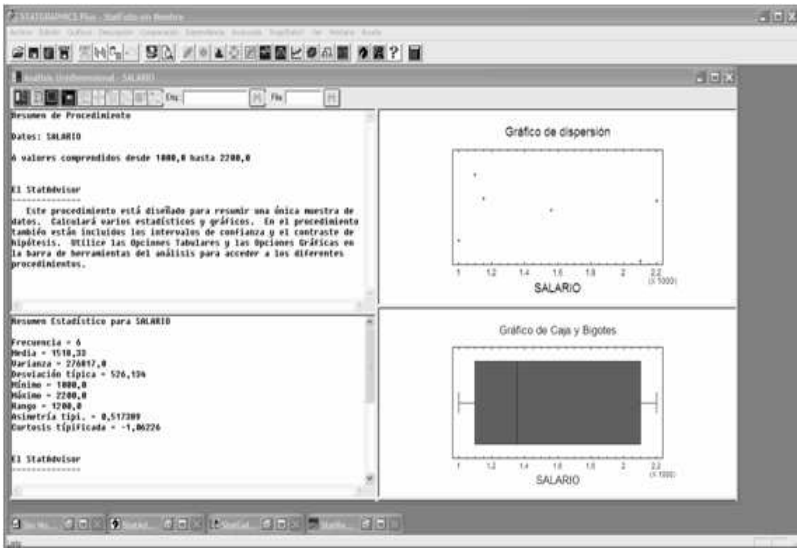

Figura 2.12: Resumen estadístico

El programa activa 4 ventanas de análisis, las dos de la izquierda dedicadas a texto y las de la derecha a gráficos. Si pulsamos las OPCIONES TABULARES (icono ), aparecen activas dos casillas, que son las que el programa proporciona por defecto (figura 2.13).

Figura 2.13: Opciones tabulares

Como vemos, por defecto, aparecen calculados los estadísticos de uso más común. Sin embargo pueden seleccionarse otros estadísticos que Statgraphics calcula sin más que clicar con el botón derecho del ratón sobre el panel de SUMMARY STATISTIS y activar la opción de OPCIONES DE VENTANA (figuras 2.14 y 2.15).

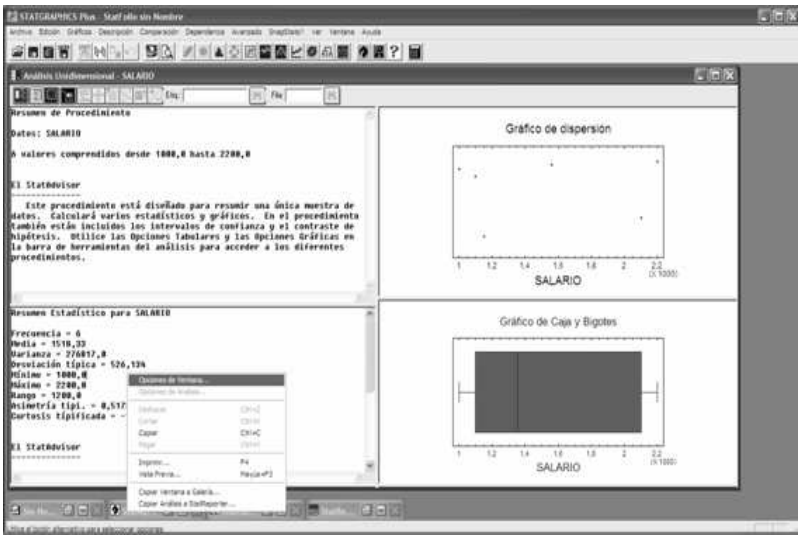

Figura 2.14: Opciones de ventana

Figura 2.15: Opciones de Resumen Estadístico

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. Estos estadísticos se pueden obtener fácilmente para varias variables sin más que entrar en el análisis múltiple de variables: DESCRIPCIÓN...DATOS NUMERICOS... ANALISIS MULTIDIMENSIONAL.

2.5.2. Tabla de frecuencias

La tabla de frecuencias nos permite resumir la distribución de los datos contenidos en una variable. Al igual que para el análisis anterior, la opción de la tabla de frecuencias (TABLA DE FRECUENCIAS) se activa en el menú de OPCIONES TABULARES 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.

El número de observaciones en cada intervalo es la frecuencia absoluta, mientras que el porcentaje que esas observaciones representan 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 clicando con el botón derecho del ratón sobre la tabla y seleccionando la opción OPCIONES DE VENTANA (figura 2.16).

La tabla de frecuencias no sólo puede aplicarse a datos numéricos, sino también a variables cualitativas. Así, si abrimos el fichero cardata.sf que se encuentra en el directorio DATA, se observan diferentes variables de automóviles junto con el nombre de su fabricante. Veamos como podemos aplicar la tabla de frecuencias a la variable que contiene el fabricante del vehículo. Para ello se sigue DESCRIPCIÓN...DATOS CUALITATIVOS... TABULACIÓN. El resultado es el que continuación se muestra en la figura 2.17, donde podemos ver el diagrama de barras, el diagrama de sectores y la tabla de frecuencias de la variable considerada (cylinders).

2.5.3. 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

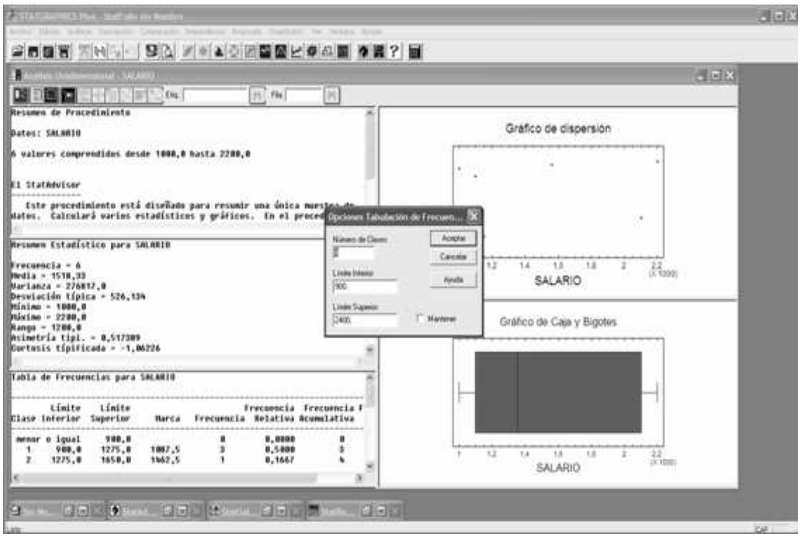

Figura 2.16: Modificar clases

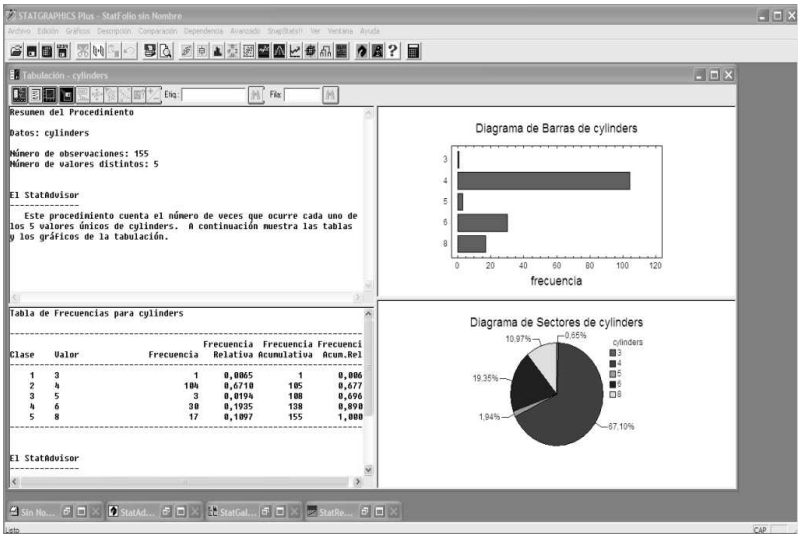

Figura 2.17: Tabla de frecuencias para cylinders

que el 10 % de las observaciones. Son importantes los percentiles 25 (cuartil inferior), 50 (mediana) y 75 (cuartil superior). Los percetiles pueden obtenerse en la opción OPCIONES TABULARES  del menú DESCRIPCIÓN... DATOS NUMÉRICOS... ANÁLISIS UNIDIMENSIONAL. Considerando los datos de la variable weight (peso) del fichero cardata, obtenemos:

```
Percentiles para weight

1,0% = 1760,0
5,0% = 1875,0
10,0% = 1975,0
25,0% = 2144,0
50,0% = 2620,0
75,0% = 3070,0
90,0% = 3530,0
95,0% = 3830,0
99,0% = 4080,0

El StatAdvisor
--------------
   Este cuadro muestra los percentiles de la muestra para weight.  Los
percentiles son valores bajo los cuales se encuentran porcentajes
especificos de datos.  Puede ver los percentiles graficamente
seleccionando Grafico Cuantil de la lista de Opciones Graficas.
```

Pulsando botón derecho del ratón, las OPCIONES DE VENTANA nos permiten definir otros percentiles.

2.6. Gráficos de distribuciones unidimensionales

2.6.1. 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 DESCRIPCIÓN... DATOS NUMÉRICOS... ANÁLISIS UNIDIMENSIONAL, tal y como puede verse en la figura 2.18.

Como podemos observar, el programa activa por defecto las opciones de Grafico de Dispersión y de Gráfico de Caja y Bigotes. Si queremos producir un Histograma (p.e. de la variable weight), basta con pulsar en la casilla correspondiente, y se habilitará un nuevo panel en la parte derecha (figura **??**).

Este histograma se puede modificar sin más que activar el panel y pulsar las OPCIONES DE VENTANA. Se puede cambiar el número de clases, el tipo de gráfico y si la frecuencia es relativa o acumulada (curva de distribución)(figura 2.20).

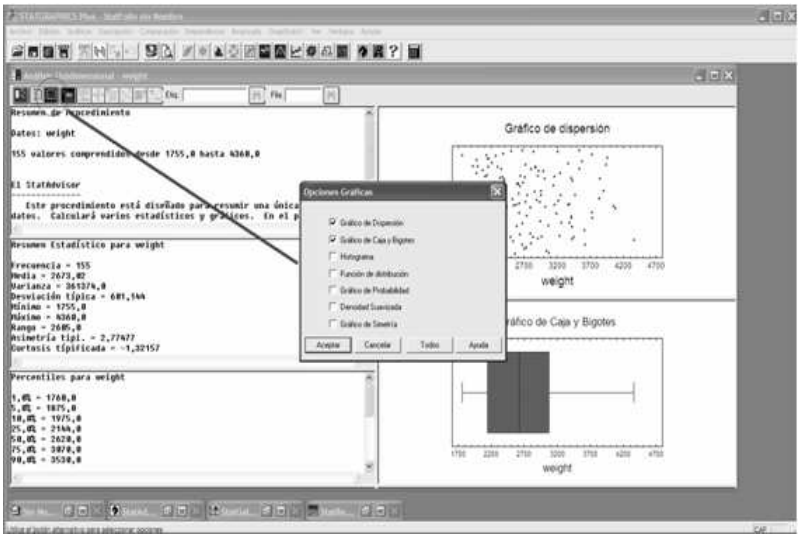

Figura 2.18: Opciones de gráficos

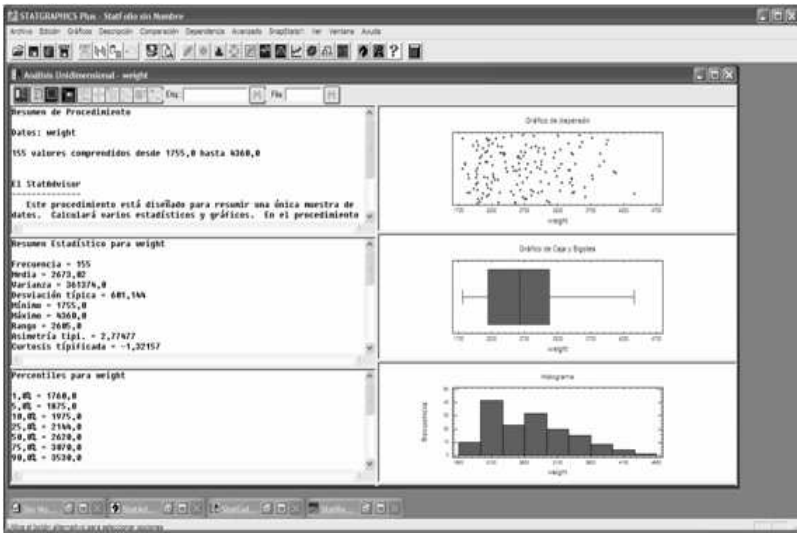

Figura 2.19: Gráficos obtenidos

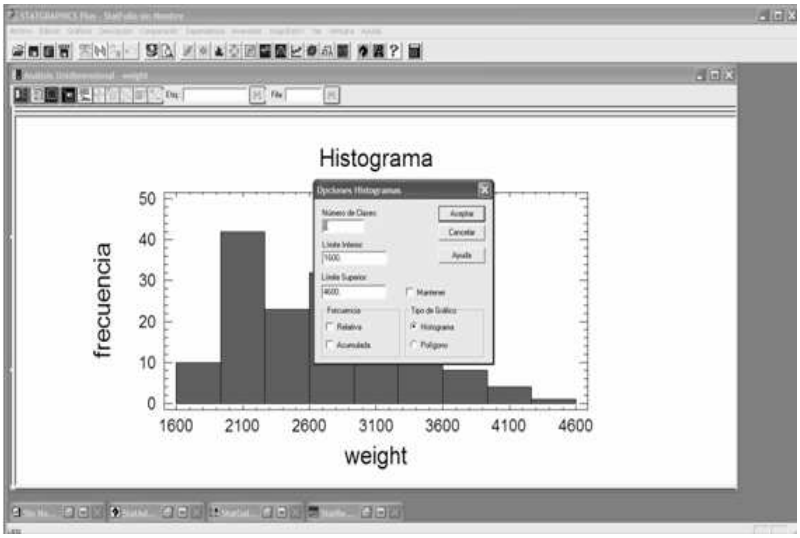

Figura 2.20: Modificación del histograma

2.6.2. Diagrama de caja con bigotes (simple y múltiple)

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.

Como hemos visto, el programa proporciona por defecto el diagrama en las opciones gráficas. Para la variable weight, el diagrama asociado aparece en la figura 2.21.

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), que en nuestro caso no existen.

En ocasiones puede ser útil observar simultáneamente dos diagramas de la caja: por ejemplo para la variable weight, varios gráficos en función del origen del vehículo. Esta opción está disponible en el menú GRÁFICOS... GRÁFICOS EXPLORATORIOS... GRÁFICO DE CAJA Y BIGOTES MÚLTIPLE (figura 2.22).

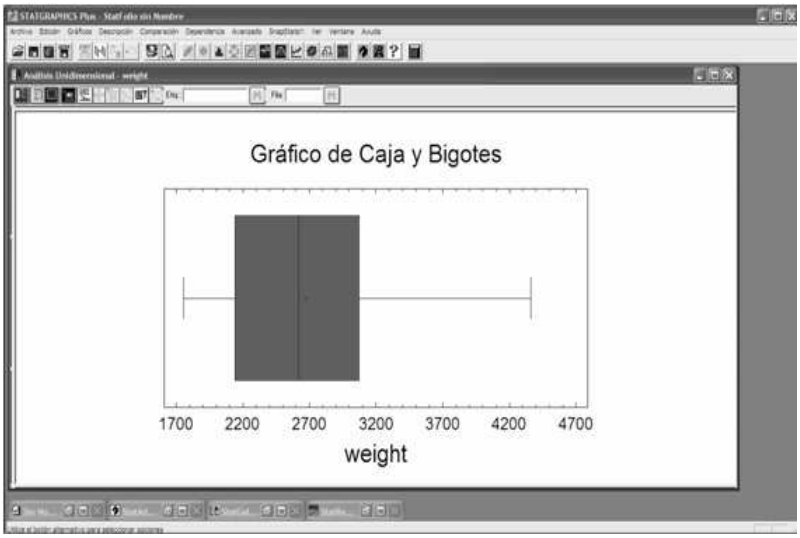

Figura 2.21: Diagrama de caja

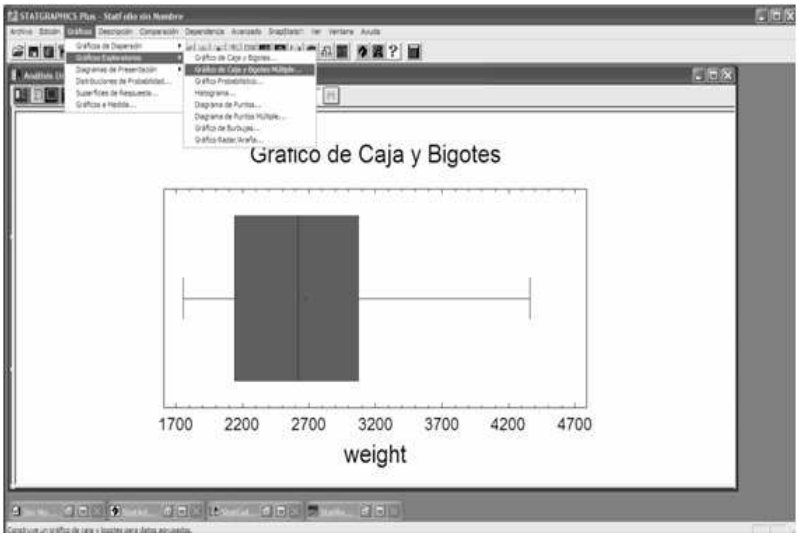

Figura 2.22: Menú de gráficos exploratorios

Se observa inmediatamente que también podíamos construir el gráfico anterior en este menú. En el caso múltiple, basta con incluir en el menú que se activa, la variable çategorizadora", que define los llamados Códigos de Nivel, y que en este caso es Origin. El resultado aparece en la figura 2.23, de forma que es posible analizar simultáneamente una variable discriminada según el criterio de selección.

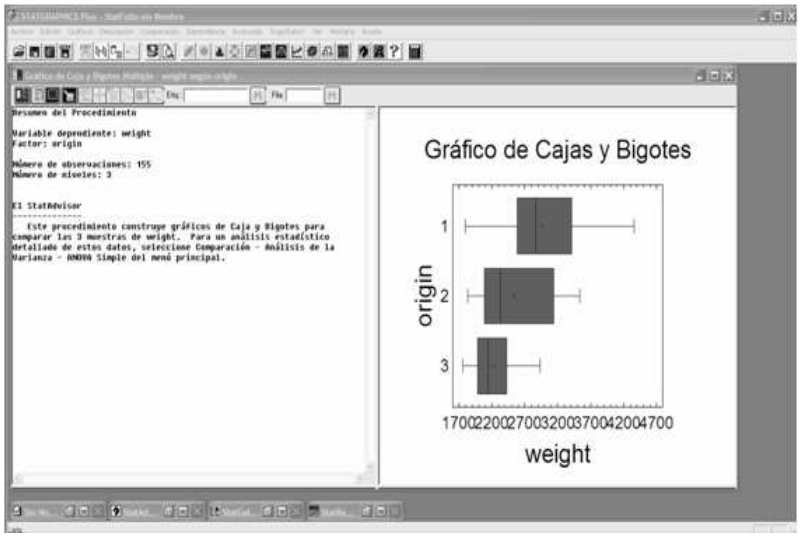

Figura 2.23: Diagrama de cajas múltiples

2.6.3. Diagrama de sectores

El diagrama de sectores proporciona información sobre las categorías en que puede dividirse una variable (y la importancia relativa de las mismas). Este tipo de gráfico es adecuado para datos categóricos, por lo que nos remitimos al epígrafe de TABLAS DE FRECUENCIAS, en el que aparece un diagrama para la variable cylinders.

2.6.4. Diagrama de barras (simple y múltiple)

Este gráfico es equivalente al anterior, y muestra la frecuencia de cada categoría respecto del resto de la variable. En el epígrafe de TABLA DE FRECUENCIAS, construimos un diagrama de barras para la variable cylinders. Es conveniente mencionar que el programa dispone de un menú específico llamado DIAGRAMAS DE PRESENTACIÓN para generar diagramas de barra simple y múltiples, pero los datos HAN DE ESTAR PREVIAMENTE TABULADOS (figura 2.24). Esta opción se verá en un capítulo posterior.

A modo de ejemplo, usando el menú DESCRIPCIÓN... DATOS CUALITATIVOS... TABULACIÓN CRUZADA para las variables year y origin, se obtienen el Gráficos de Barra Múltiple y el llamado Gráfico en Mosaico (figura 2.25).

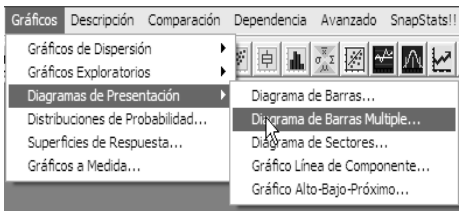

Figura 2.24: Menú de Diagramas de Presentación

Figura 2.25: Gráficos de mosaico

2.6.5. Gráficos de dispersión

Los gráficos de dispersión proporcionan información acerca de la distribución de una variable o varias variables. Son especialmente útiles los gráficos XY (dos variables), pues permiten analizar la relación entre dos variables. Para visualizarlos se sigue el menú GRÁFICOS... GRÁFICOS DE DISPERSIÓN... GRÁFICO XY. Si incluimos las variables weight y horsepower en el análisis, como aparece en la figura 2.26,

Figura 2.26: Panel de gráficos de dispersión

obtenemos como resultado un diagrama (figura 2.27) que nos permite ver la distribución conjunta de ambas variables, y por tanto su relación lineal, en la que al aumentar el peso del vehículo (weight) hace que se deba incrementar su potencia (horsepower) .

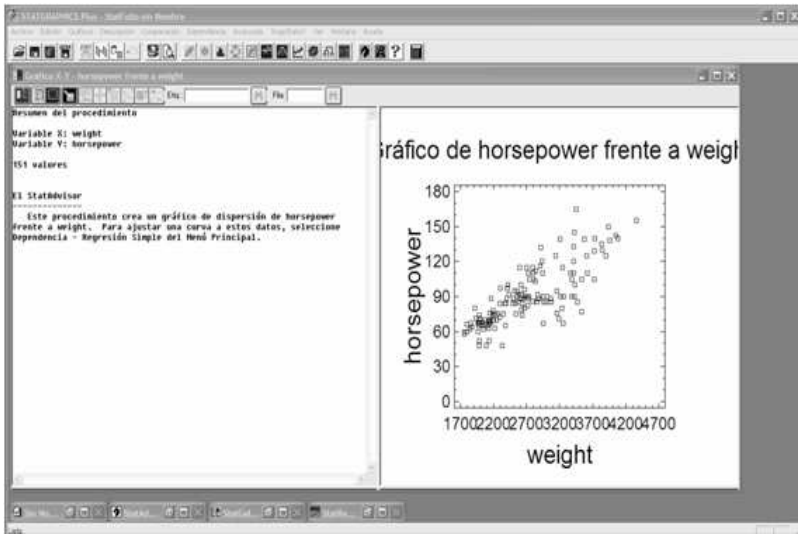

Figura 2.27: Gráfico de dispersión

Podemos modificar el gráfico anterior en las OPCIONES DE VENTANA incluyendo una variable categorizadora (Códigos de Nivel), en este caso origin. El resultado la figura 2.28.

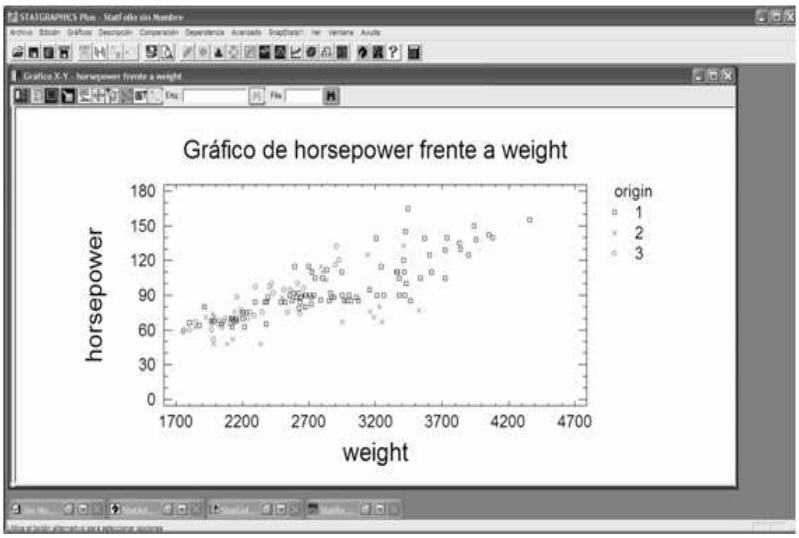

Figura 2.28: Gráfico de dispersión con marcas de nivel

En el gráfico puede observarse que los coches de origin=1 (EE.UU.) son los más pesados y de más potencia.

2.7. Gráficos de multidimensionales

Son de interés los gráficos que se pueden confeccionar con más de dos variables, entre los que presentaremos el diagrama de dispersión múltiple para dos variables XY, y el diagrama de dispersión tridimensional XYZ.

El primero en GRAFICOS... GRAFICOS DE DISPERSIÓN... GRÁFICOS XY MÚLTIPLE, presenta simultáneamente varios gráficos de dispersión para dos variables. Habremos de introducir varias variables (accel, horsepower y weight) en el menú de la figura 2.29.

Obsérvese que habrá dos variables en el eje Y (puede haber más) y una variable en el eje X El resultado aparece en la figura 2.30.

También podríamos haber construido un gráfico "tridimensional"para las tres variables anteriores, accediendo al menú GRAFICOS... GRAFICOS DE DISPERSIÓN... GRÁFICOS XYZ. Completando los tres campos requeridos con esta variables, obtenemos la figura 2.31.

El programa dispone de varias opciones de visualización de gráfico (cambio de perspectiva), en el apartado de OPCIONES DE VENTANA, una de ellas referida a identificar los puntos por Códigos de punto, como ya se realizó para diagramas de dispersión anteriores.

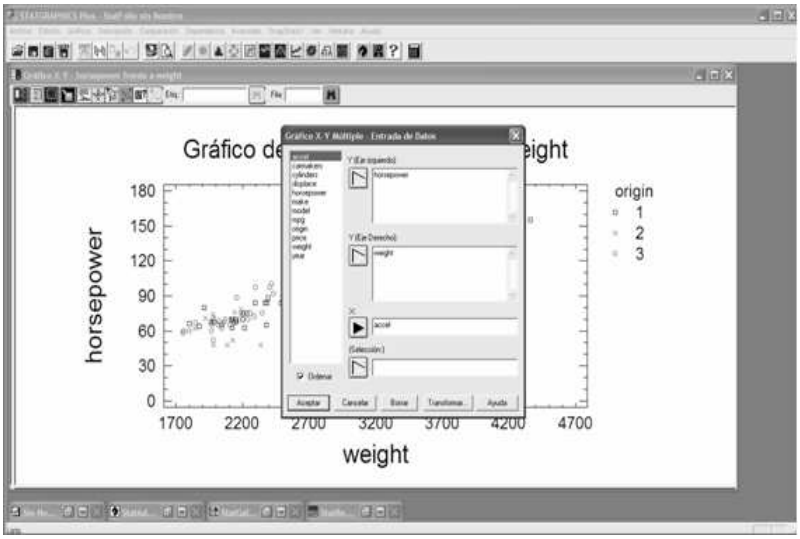

Figura 2.29: Menú para gráfico de dispersión multidimensionales

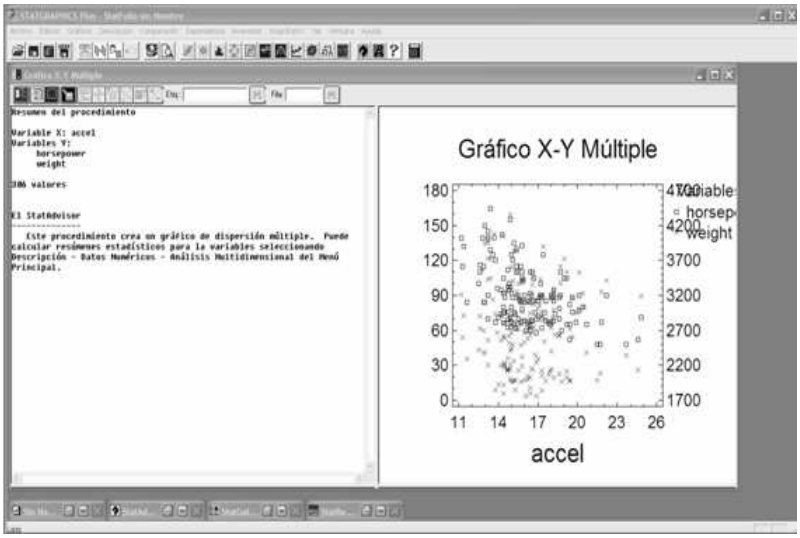

Figura 2.30: Menú para gráfico de dispersión multidimensionales

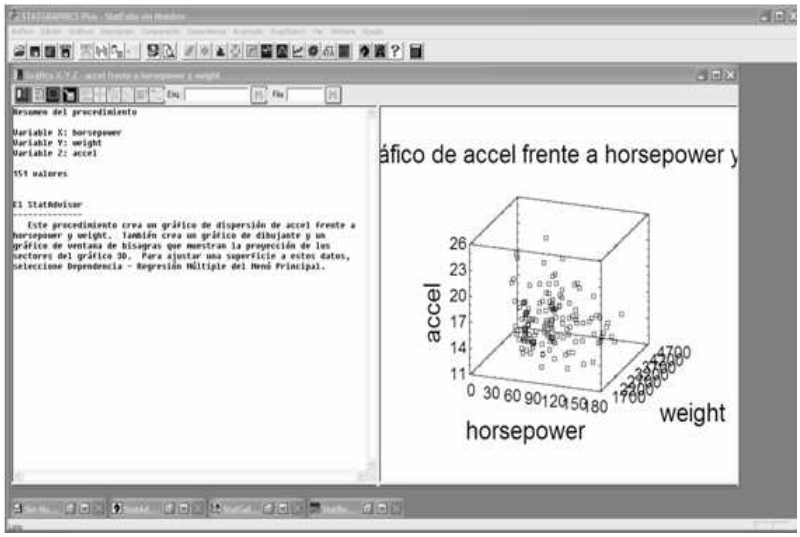

Figura 2.31: Menú para gráfico de dispersión multidimensionales